Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-364-RC1, 2020

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Increased new particle yields with largely decreased probability of survival to CCN size at the summit of Mt. Tai under reduced SO_2 emissions" *by* Yujiao Zhu et al.

Anonymous Referee #3

Received and published: 15 August 2020

The manuscript analyzes seven field campaigns where particle number size distributions (PNSD) and sulfur dioxide were measured at the summit of a mountain site in the North China plain. Supporting measurements of time-resolved PM2.5, O3, and oxides of nitrogen were taken. And each campaign included 1-h time resolution ions in PM2.5 using water extractive methods (URG-AIM or MARGA). The most recent campaign was in 2018. Across the 7 campaigns, a little over 100 particle formation and growth events were detected, with the analysis focused on the size range of 10-300 nm size range. From the earliest to most recent campaign, SO2 emissions and concentrations have dropped dramatically, and the paper tries to analyze whether the particle formation and growth activity has changed in ways that are expected from the sulfur dioxide

decrease. A large number of metrics are computed and then analyzed for each particle formation and growth event (PFGE). The metrics include, but are not limited to, the apparent formation rate of 10-25 nm particles (FR), the growth rate, the absolute increase in N10-25 particle concentration from the start to the peak of the PFGE (this is the NMINP variable), the PFGE duration, the PFGE frequency, the size to which the growth event reaches (Dpgmax variable in the manuscript), and particle counts which are used as surrogates for the change in CCN concentrations at low, medium, and high supersaturations (N100-300, N80-300, and N50-300). The paper includes values for and discussion of total VOC during the campaigns.

Complicating the analysis is that the field campaigns were in different months of the year: April 2007 (∼30 d), June 2009 (∼20 d), Aug 2014 (∼30 d), Oct/Nov 2014 (∼70 d), Jul 2014 (∼40 d), Dec 2017 (∼35 d), and Mar 2018 (∼30 d).

The paper's abstract makes five claims: a. The formation rate in 2018 is 2-3 times higher than the formation rate in 2007. b. Net maximum increase in nucleation mode number concentration is 2-3 times higher in 2018 than in 2007. c. The occurrence of events where the mode of the growth event goes above 50 nm is lower in 2018 than it was in 2007. d. A surrogate for CCN production at high supersaturation (N50-300 at its peak during each growth event minus N50-300 before the event) decreased from 3703 per cm3 (before 2015) to 1026 (2017-2018). e. The authors argue availability of organic precursors has increased in the most recent campaigns, allowing more particle production and initial growth; furthermore, they argue that the lack of later growth is from reduction of "anthropogenic precursors" (presumably SO2).

The paper requires substantial revision before it is suitable for publication. The key issue, to this reviewer, is that making accurate claims about year-on-year trends and variability in PFGE is difficult. The requirements to make the claims defensible are: (1) take a sufficient number of samples to reduce random variability and give sufficient statistical power; (2) take steps to minimize, test for, and quantify campaign-specific systematic instrument bias (also known as "instrument drift"); (3) take steps to enforce

consistency in any subjective data interpretation steps, such as classification of PFGE into “types” and the determination of the start and end times of events; (4) use statistical methods designed for trend analysis, time series analysis, and combined analysis of seasonal and interannual variability.

Each requirement needs to be met in order for the claims about trends to be defensible. And for peer review and reproducibility purposes, things need to be documented for the peer-review and scientific communities.

I think the current work fails to meet all four of the requirements. While some of the conclusions are likely accurate (in that they would not change if all the requirements were met) – others would change, or require extensive qualification.

1. Statistical power:

PFGE exhibit substantial seasonal variation, due to changes in temperature, relative humidity, biogenic activity, atmospheric chemistry, soil moisture, preexisting aerosol concentration and chemistry, radiation, cloudiness, boundary layer structure, land cover/vegetation canopy structure, synoptic meteorology, anthropogenic emissions, and atmospheric ion levels. Local meteorological features (i.e. orographic meteorology) and local sources may also have month-to-month variability. And at the 20-30 d time scale, large scale persistent geophysical features can cause a whole campaign of measurements to be atypically high or low for a number of PFGE variables. To accommodate all these sources of variability, large sample sizes are required for analysis of seasonal variation and interannual trends. In the absence of large sample sizes, careful pairing of events and analysis of alternate sources of variability / alternate hypotheses are needed isolate cause-effect relationships on specific PFGE variables.

With each campaign at a slightly different time of the year, some campaigns as short as $\sim$20 d, and no discussion of whether air pollution levels, air pollution meteorology, and climate variables were at climatologically representative levels, the reader has to apply great skepticism to any claims of interannual trends and cause-effect relationships for

those interannual trends. See for example (Birmili and Wiedensohler 2000) who do take into account air mass characteristics.

The size distributions shown in Figure S3 are suggestive of insufficient number of days sampled in the dataset. Telling whether the system shifted from unimodal to bimodal behavior between 2015 and 2017 (all unimodal for 2015 and prior) vs. this occurring through some instrument drift vs. this occurring through sampling non-climatological conditions due to small samples sizes is difficult.

Given the decrease (Table 2) in PM2.5, sulfate, and SO2 between spring 2007 and 2018 (PM2.5 60 vs. 30 ug/m3; sulfate 17 vs. 4 ug/m3, SO2 18 vs. 3 ppb), more discussion is needed of the large increase in condensation sink in 2018 (Figure S3) and in the large increase in the height of the size distribution function at 100 and 150 nm between 2007 and 2015.

The discontinuity in the slope of the size distribution function at 200 nm also indicates there may be some drift in the size-specific performance of the WPS (Figure S3). The discontinuity in slope is not really evident until 2017, but then appears in 2017 and 2018.

2. Minimize, test for, and quantify campaign-specific instrument drift:

Achieving consistency in PNSD in long-term measurements is difficult. And it is not sufficient to state that each individual campaign had sufficient quality assurance, referring the reader to the campaign specific papers. There needs to be a presentation of data and discussion of how comparable the instrument responses are from campaign to campaign. What steps were taken to make sure instruments were not drifting. Aging of components can cause variation in flows, sizing accuracy, counting accuracy, particle losses, CPC supersaturations, and in the effective lower size limit of the instrumentation of the particle number spectrometer system. The detection efficiency as a function of size at the lower range of the instrument (5-25 nm), at the upper range of the mobility analyzer, at the lower end of the optical particle devices, and at the upper end

of the optical particle analyzer – these are all difficult to maintain at stable levels over long periods of time. The total particle counts, height of the size distribution function, sensitivity at the lower and upper ranges of size distributions – these vary from year to year and require careful intercomparison, quality assurance, and maintenance procedures to deal with. See for example the results of intercomparison studies (Pfeifer, Müller et al. 2016) and papers focusing on quality assurance, calibration, and harmonization (Pitz, Birmili et al. 2008, Wiedensohler, Birmili et al. 2012, Wiedensohler, Wiesner et al. 2018, Gaie-Levrel, Bau et al. 2020). Comparison to other instruments for total particle counts, size distribution functions in overlapping regions, checks with monodisperse particles are some of the techniques that can be used to establish more confidence and quantify campaign-to-campaign comparability.

Consistency in inlet dimensions, inversion algorithms (including multiple charge correction), use of impactors to manage multiple charge issues, corrections for inlet transmission efficiency, – these can all be issues in campaign-to-campaign comparability. They need to be discussed.

While being able to reproduce time-resolved PM2.5 measurements from the WPS size distribution is not sufficient to show accuracy in the nucleation and Aitken ranges – it is probably necessary. At least showing consistency from campaign to campaign in the volume of particles measured by the WPS and the mass of particles by time-resolved mass measurements can help to demonstrate stability in instrumentation and data processing algorithms.

The fact that the authors are using an instrument with nominal lower cutoff of 5 nm, but discarding data between 5-10 nm indicates that there may be a problem with sensitivity at the lower size limit, or (more likely) variability in the sensitivity at the lower size limit. There is further evidence in Figures 1 and S6 – of a problem. In all the bursts shown save one, the particle size distribution function slopes down from a peak at about 13 nm to a lower value at 10 nm. If the instrument is biased low in the 10-13 nm range, then the statistics developed in the work will also be biased. If that bias varies from

campaign to campaign, then that creates additional interpretation difficulties.

At line 180, it is implied that at times the WPS was collocated with instruments with lower limit of 3 nm. Therefore, the actual performance in the 5-15 nm range could (and should) be determined though comparison to such collocated instruments.

3. Consistency in subjective data interpretation/classification steps:

It is not clear which of the variables used for analysis involve human classification. Sometimes, human classification is used for PFGE types (often using how smooth the growth event is in time); human classification is used sometimes for establishing times (start of event, end of the event). The end time is described. From line 113 of manuscript, “The end time of an NPF event was defined as the time when the particle number concentrations approached the background levels observed before the NPF event. The NPF event duration was defined as the time duration between the start time and end time of an NPF event.” This seems like the end of event was a subjective determination of when background was approached. Thus the end time, duration, and any rate that has the duration in the denominator may be subjective.

For subjective (human) event classifications, were the events uniformly reclassified for this paper, or were prior classifications adopted from 2007 and 2009 and mixed with new classifications done for the more recent campaigns. See (Dal Maso, Kulmala et al. 2005) for best practices on human classification.

4. Statistical methods appropriate to analysis of combined seasonal and interannual variability

Statistical procedures for evaluating trends in seasonally varying time series need to be followed in order to state claims that trends exist. These can be found in a number of textbooks, papers, and government reports. See for example Statistical Methods for Environmental Pollution Monitoring by Gilbert https://www.osti.gov/servlets/purl/7037501/. And (Asmi, Coen et al. 2013, Collaud

Coen, Andrews et al. 2013, Squizzato, Masiol et al. 2019). Many other good models for seasonally adjusted trend detection can be found in the O3, NOx, PM2.5, and hydrology/climatology literature. Squizzato et al. (2019) for example have the statistical procedures necessary to detect turning points (see line 236 where manuscript discusses turning points)

See for example line 290 “the CS still increased in 2018 compared with that in 2007.” That implies annual average condensational sink increased, and this is a season or month specific result – and it is not clear there is enough statistical confidence to state this. Many other locations in the paper have broad statements about PFGE behavior in one year vs. another, or imply a long term trend where it has not really been shown.

Interpretation of PFGE data from this site seems more complicated than most, because of two issues: (1) it is sometimes influenced by boundary layer and other times by free troposphere; (2) very long PFGE events (see for example Figure 1a, where a 3-d long event is shown) are being compared with shorter (midday + afternoon) growth events. See Figure 7 which has events ranging from 3-h duration to 85-h duration. The flow patterns and chemistry required to sustain a 3-h event and an 80+ h event are likely very different, and would require more thoughtful comparison metrics than used in the paper.

The paper acknowledges this difficulty in interpretation (line 295) but more needs to be done than just acknowledge the difficulty. See analysis papers from PFGE studies at other high altitude sites. They do attempt to determine the degree of FT influence and the impact of polluted boundary layer air. And there are many papers that factor in air mass characteristics and/or back trajectory in analysis of PFGE.

See for example Figure 1a where on 25-Dec 2017 there were simultaneously occurring a short PFGE (category 1) and evolution of the category 3 event that started on 24-Dec 2017. This raises a number of questions on how such a dataset can be analyzed to determine trends.

How much of the variability in data is that some campaigns had more free tropospheric influence and others have less. How much of the conclusions of the paper are driven by switches (during PFGE) in air mass influence to/from FT influence. In other words, PFGE events that have their evolution dynamics controlled by airflows, and not by chemistry – hence the authors observed lack of influence or counterintuitive effects of SO2.

As for statistical procedures, I think it would be much more appropriate to put 95% confidence intervals on means rather than standard deviations on the plots. (Most figures have standard deviations)

Some of the variables appear to NOT be normally distributed (see figure S4) and thus use of statistical tests designed for normally distributed data are inappropriate.

Another weakness of the approaches used are that changes in boundary layer height are not accounted for. This weakness cannot really be addressed without additional measurements, but it should be noted.

Other issues:

5. The abstract overstates the conclusions of the work. The conclusions have significant caveats, are based on limited number of sampled days, but the abstract makes it seem like the trends are well established, statistically significant, and based on a complete multi-year time series.

6. There are a number problems with Figure 6. It is not appropriate to grey out datasets that are not correlated. Data are data, and data points should not be deemphasized visually just because they do not fit a linear correlation. The datasets should be clearly labeled so that each symbol type can be connected back to its underlying study and land cover type. Having a linear correlation shown and then a change in the tick mark spacing is not a fair way of graphing in my opinion. The size ranges in question should be included in the axis labels and/or the caption. I believe this is the formation rate at

10 nm, and the NMINP at 10-25 nm? Is that consistent for all the datasets? If not, then I don't think this is a fair plot to put in. I don't think having regression equations and correlation coefficients on graphs is effective or appropriate (see additional comments on this later).

7. If a p-value appears in a figure or in the paper, then the statistical test needs to be discussed. What are the null and alternative hypothesis. And why is each hypothesis test implied by each p value important, scientifically interesting, novel, or useful?

8. If a regression equation (e.g., y=12.5x+5.6) appears in a figure or in the paper, then its use – either for scientific or engineering purposes – needs to be discussed. The paper has 9 regression equations in it. Are they of any use?

9. I believe all r values can be deleted from the paper without any loss.

10. Is the size range covered sufficient for calculating the condensational sink? Or stated differently, how much of the condensational sink is being missed by focusing on 10 to 150 or 250 nm.

11. Line 138 "can be calculated" or was calculated?

12. Are variables that are sensitive to the upper size limit (CCN concentrations that are based on the number of particles greater than size X, condensation sink) consistent given the change in the upper size limit shown in Figure S3, from campaign to campaign.

13. Line 282 – climate change typically requires 30-y averaging. Interannual variability may be much more likely at the time scales studied here.

14. Line 293 – "data size was small" is vague. A more detailed description of what aspects of the dataset are too small is needed.

15. Line 294 – there are two issues: spatial representativeness, and sparsity of the record in time. In my opinion these create two different problems for the work. "the

data size was small, and we should be cautious in extending the conclusion to a large spatiotemporal scale"

16. Line 299 – this shows the authors are thinking of these events as perfect Lagrangian experiments, where sampling at the mountain site is equivalent to sampling along a 0-D Lagrangian air mass trajectory. Vertical and horizontal mixing are not accounted for in this conceptual model. And the possibility that back trajectories evolve over the course of the PFGE is neglected. In reality, as the event evolves, winds will bring air with a variety of histories (chemical, emissions, radiation, accumulation mode particles, interaction with precipitation and clouds, etc.). The survival probabilities over 100% (Figure 4) are likely a symptom of the fact that reality has complex flows and spatial heterogeneity and does not fit the idealized box model concept.

17. Figure 7 is of low resolution. Difficult to see some of the symbols, and symbols are of different sizes in different plots.

18. The discussions of biogenic and total VOCs throughout the paper are problematic. What species are these? How were they measured? Were the measurements collocated with the PFGE measurements and matched in time? The amount of oxidation needed to grow from 3 to 10 nm or 10 to 20 nm, is quite small, so making broad generalizations about seginficant changes in entire classes of VOCs or in specific compounds, and then connecting them to PFGE is not scientifically valid.

19. Rather than making the data available "on request", the data should be publicly posted in machine readable formats at the time of publication in order to allow replication.

Birmili, W. and A. Wiedensohler (2000). "New particle formation in the continental boundary layer: Meteorological and gas phase parameter influence." Geophysical Research Letters 27(20): 3325-3328.

Dal Maso, M., M. Kulmala, I. Riipinen, R. Wagner, T. Hussein, P. P. Aalto and K. E. J.

Lehtinen (2005). "Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland." Boreal Environment Research 10(5): 323-336. Gaie-Levrel, F., S. Bau, L. Bregonzio-Rozier, R. Payet, S. Artous, S. Jacquinot, A. Guiot, F. X. Ouf, S.

Bourrous, A. Marpillat, C. Foulquier, G. Smith, V. Crenn and N. Feltin (2020). "An intercomparison exercise of good laboratory practices for nano-aerosol size measurements by mobility spectrometers." Journal of Nanoparticle Research 22(5): 13.

Pfeifer, S., T. Müller, K. Weinhold, N. Zikova, S. Martins dos Santos, A. Marinoni, O. F. Bischof, C. Kykal, L. Ries, F. Meinhardt, P. Aalto, N. Mihalopoulos and A. Wiedensohler (2016). "Intercomparison of 15 aerodynamic particle size spectrometers (APS 3321): uncertainties in particle sizing and number size distribution." Atmospheric Measurement Techniques 9(4): 1545-1551.

Pitz, M., W. Birmili, O. Schmid, A. Peters, H. E. Wichmann and J. Cyrys (2008). "Quality control and quality assurance for particle size distribution measurements at an urban monitoring station in Augsburg, Germany." Journal of Environmental Monitoring 10(9): 1017-1024.

Wiedensohler, A., W. Birmili, A. Nowak, A. Sonntag, K. Weinhold, M. Merkel, B. Wehner, T. Tuch, S. Pfeifer, M. Fiebig, A. M. Fjaraa, E. Asmi, K. Sellegri, R. Depuy, H. Venzac, P. Villani, P. Laj, P. Aalto, J. A. Ogren, E. Swietlicki, P. Williams, P. Roldin, P. Quincey, C. Huglin, R. Fierz-Schmidhauser, M. Gysel, E. Weingartner, F. Riccobono, S. Santos, C. Gruning, K. Faloon, D. Beddows, R. M. Harrison, C. Monahan, S. G. Jennings, C. D. O'Dowd, A. Marinoni, H. G. Horn, L. Keck, J. Jiang, J. Scheckman, P. H. McMurry, Z. Deng, C. S. Zhao, M. Moerman, B. Henzing, G. de Leeuw, G. Loschau and S. Bastian (2012). "Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions." Atmospheric Measurement Techniques 5(3): 657-685.

Wiedensohler, A., A. Wiesner, K. Weinhold, W. Birmili, M. Hermann, M. Merkel, T. Muller, S. Pfeifer, A. Schmidt, T. Tuch, F. Velarde, P. Quincey, S. Seeger and A. Nowak (2018). "Mobility particle size spectrometers: Calibration procedures and measurement uncertainties." Aerosol Science and Technology 52(2): 146-164.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2020-364, 2020.