Using satellite-based measurements to explore spatiotemporal scales and variability of drivers of new particle formation


1Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA
2COMET, School of Civil Engineering and Geosciences, Newcastle University, Newcastle upon Tyne, UK
3Storm Peak Laboratory, Desert Research Institute, Steamboat Springs, CO, USA and Atmospheric Science Department, University of Utah, Salt Lake City, UT, USA
4Spectroscopie de l'Atmosphère, Service de Chimie Quantique et Photophysique, Université Libre de Bruxelles (ULB), Bruxelles, Belgium
5Environment Canada, Toronto, Ontario, Canada
6U.S. Environmental Protection Agency, Office of Research and Development, National Risk Management Research Laboratory, Durham, NC, USA
7Desert Research Institute, Reno, NV, USA
8Pervasive Technology Institute, Indiana University, Bloomington, IN, USA

*Corresponding author (rcs365@cornell.edu)

Key Points
1. NPF at five sites across North America exhibit positive one-day autocorrelation
2. NPF frequency, formation and growth rates, survival probabilities, and ultrafine particle concentrations are spatially variable
3. Similarities and variability in NPF characteristics can be partly explained by satellite-based measurements of atmospheric composition
Abstract

New particle formation (NPF) can potentially alter regional climate by increasing aerosol particle (hereafter particle) number concentrations and ultimately cloud condensation nuclei. The large scales on which NPF is manifest indicate potential to use satellite-based (inherently spatially averaged) measurements of atmospheric conditions to diagnose the occurrence of NPF and NPF characteristics. We demonstrate the potential for using satellite-measurements of insolation (UV), trace gas concentrations (sulfur dioxide (SO2), nitrogen dioxide (NO2), ammonia (NH3), formaldehyde (HCHO), ozone (O3)), aerosol optical properties (aerosol optical depth (AOD), Ångström exponent (AE)), and a proxy of biogenic volatile organic compound emissions (leaf area index (LAI), temperature (T)) as predictors for NPF characteristics: formation rates, growth rates, survival probabilities, and ultrafine particle (UFP) concentrations at five locations across North America. NPF at all sites is most frequent in spring, exhibits a one-day autocorrelation, and is associated with low condensational sink (AOD×AE) and HCHO concentrations, and high UV. However, there are important site-to-site variations in NPF frequency and characteristics, and in which of the predictor variables (particularly gas concentrations) significantly contribute to the explanatory power of regression models built to predict those characteristics. This finding may provide a partial explanation for the reported spatial variability in skill of simple generalized nucleation schemes in reproducing observed NPF. In contrast to more simple proxies developed in prior studies (e.g. based on AOD, AE, SO2, UV), use of additional predictors (NO2, NH3, HCHO, LAI, T, O3) increases the explained temporal variance of UFP concentrations at all sites.
1 Introduction and motivation

New particle formation (NPF) events generate large concentrations of ultrafine particles (UFP; particle diameter (Dp) < 100 nm), often occur on regional scales, and exhibit high temporal autocorrelation (multi-day persistence) [Hussein et al., 2009; Jeong et al., 2010; Crippa and Pryor, 2013]. Therefore NPF may substantially increase the concentration of particles with Dp ≥ 100 nm and cloud condensation nuclei (CCN) [Spracklen et al., 2008b; Merikanto et al., 2009; Yu and Luo, 2009; Pierce et al., 2012, 2014], and thus impact regional climates [Spracklen et al., 2008a; Paasonen et al., 2013]. However, the magnitude of this effect remains uncertain [Carslaw et al., 2013].

Although the precise atmospheric conditions conducive to NPF are not fully understood [Boy et al., 2007] and may vary in space and time [Kulmala et al., 2004; Yu and Luo, 2009; Westervelt et al., 2013; Yu and Hallar, 2014; Yu et al., 2015], most observational studies are consistent with ternary nucleation involving sulfuric acid (H₂SO₄), water vapor, and some other low-volatility and/or stabilizing condensable species (e.g. oxidation products of biogenic volatile organic compounds (BVOCs) or ammonia (NH₃)) [Kulmala et al., 2000; Metzger et al., 2010; Sipilä et al., 2010; Zhang et al., 2010; Kirkby et al., 2011; Pryor et al., 2011; Riccobono et al., 2014]. Further, the intensity and probability of NPF appears to be positively associated with high insolation and a reduction of the condensational sink (CS) (and thus competition for semi-volatile species) [O’Dowd et al., 2002; Zhang et al., 2004a; Sipilä et al., 2010; Pryor et al., 2011; Almeida et al., 2013; Pierce et al., 2014]. In situ studies indicate that although initial growth of the recently nucleated particles is largely due to coagulation and condensation of the nucleating gases, growth beyond diameters of 10’s nm may exhibit an increased contribution from condensation of semi-volatile, low-volatility, and extremely low volatility organic gases [Zhang et al., 2004a; Smith et al., 2005, 2008; Knol et al., 2009; Pryor et al., 2011; Pierce et al.,
These commonalities coupled with the occurrence of NPF on regional scales has led to suggestions that nucleation mode and UFP concentrations can be predicted using satellite-based, and thus spatially averaged, atmospheric properties such as: accumulation mode particle properties (aerosol optical depth (AOD) and Ångström exponent (AE)), trace gas concentrations (sulfur dioxide (SO₂) and nitrogen dioxide (NO₂)), and ultra-violet irradiance (UV) \cite{Kulmala2011, Crippa2013, Sundstrom2015}. More recent work at a single site in the Midwestern USA indicated potential to diagnose not only the total UFP concentrations from remote sensing observations, but also the probability of NPF and descriptors of NPF events: particle formation rates ($J_n$, where ‘n’ is determined by the minimum detectable particle diameter for each instrument), growth rates (GR), and survival probabilities (SP) \cite{Sullivan2016}.

Based on this prior research, we postulate that satellite-derived observations of key properties known, or theorized to, determine the frequency and temporal persistence of NPF, $J_n$, GR, SP, and UFP concentrations ($N_{n-100nm}$), may also be used diagnostically to explain some of the observed spatial variability in these characteristics of NPF \cite{Pierce2014, Qi2015, Rose2015}, and reported variations in the closure between models based on simplified nucleation schemes (with fixed nucleation rate coefficients) and observations \cite{Spracklen2008, Zhang2010, Lee2013}. Given the likely impact of NPF on climate, and substantial uncertainty and model-to-model variability in simulating NPF, improved treatment of NPF in global models is critical for improved understanding of aerosol-climate impacts. Herein we use in situ particle size distribution (PSD) measurements from five sites distributed across North America (Figure 1; Table 1) to address the following research questions:

1) Can satellite-based measurements of parameters known to be important to NPF and
subsequent growth (or proxies for those variables) be used to explain site-to-site
variations in NPF frequency and one-day autocorrelation? For example, is there a
relationship between the inherent spatial scales of coherence of the satellite-based
measurements of the drivers of NPF and the one-day autocorrelation in NPF
occurrence, and if so which one of the drivers appears to limit the persistence of NPF?

2) Do satellite-based observations offer insights into the causes of variability in NPF
characteristics (probability of NPF, Jn, GR, SP, and Nn-100nm) at the five sites?

3) Do proxy algorithms wherein the predictands are the NPF characteristics at the five
sites and the predictors are drawn from the suite of remote sensing parameters exhibit
commonalities in terms of the most important predictors and the variance explained?

4) Does a proxy model of UFP concentrations using a larger suite of predictors exhibit
more explanatory power than the prior satellite-based proxies, which have employed
AOD and AE, SO₂ (or NO₂), and UV?

Additionally, given in situ PSD measurements are time consuming and expensive, and thus are
typically made for limited time periods, we perform a statistical analysis to quantify how using a
limited sample of environmental conditions impacts the generalizability of inferences drawn
from data collected during time-limited campaigns.

2 Methods

2.1 Particle size distribution data

The PSD measurements used herein all have durations of a year or more, and derive from
locations distributed across North America with different land use and proximity to major point
source emissions (Table 1; Figure 1). While all PSD measurements are taken at/near the surface,
the high elevation of Storm Peak Laboratory (SPL) renders it representative of free tropospheric
air on a near daily basis [Yu and Hallar, 2014]. Prior to presenting the derived NPF descriptors it
is important to note that the individual sites used different instrumentation (and thus have
different minimum Dp detection limits, Table 1) and sampling protocols that may confound
inter-comparison across the five sites. For example, archiving of data from Southern Great Plains
(SGP) at 30-minute resolution reduces the confidence in the calculated NPF metrics for that site.
However, these are the longest records of UFP PSD measurements currently available for North
America. Further, analysis of data collected using two different instruments (an FMPS and
SMPS) at Morgan Monroe State Forest (MMSF) can be used to partly evaluate the impact of
instrumentation versus spatial variability in determining NPF characteristics and drivers.

2.2 Remote sensing measurements

Once daily satellite-based observations used herein as predictors of NPF and their
associated uncertainties are summarized in Table 2. The justification for the selection of these
variables is as follows:

1. The cross product of aerosol optical depth (AOD) and Ångström exponent (AE) is used as a
proxy for CS following Crippa et al. [2013], and is anticipated to be negatively associated
with NPF occurrence, $J_n$, GR, and SP.

2. SO$_2$ is used as a proxy for H$_2$SO$_4$ following Crippa et al. [2013], Kulmala et al. [2011], and
Sundström et al. [2015], although it must be noted that the retrievals exhibit a low signal to
noise ratio, except near large emissions [Krotkov et al., 2008; Fioletov et al., 2011], and
many negative concentrations are reported in the OMI SO$_2$ product. Although nucleation
rates may be $\propto$ [SO$_2$]$^n$ [Kuang et al., 2008] or dependent on H$_2$SO$_4$ production ($\propto$ SO$_2$$\times$UV),
we do not include exponential or compound predictor variables due to the low sensitivity of
satellite-based measurements of SO$_2$ and to avoid overfitting the regression models. SO$_2$ is
anticipated to be positively associated with NPF occurrence, $J_n$, GR, and SP.

3. UV is used as a proxy for photochemical production of oxidants (e.g. the hydroxyl radical
(OH)) and thus oxidation of SO$_2$ to H$_2$SO$_4$ and BVOCs to low-volatility products following

Crippa et al. [2013], Kulmala et al. [2011], and Sundström et al. [2015], and is anticipated to

be positively associated with NPF occurrence, J$_n$, GR, and SP.

4. NO$_2$ is used as a proxy for air masses influenced by anthropogenic emissions (including

primary emitted particles and condensable vapors) [Russell et al., 2012], following Kulmala

et al. [2011] and Sundström et al. [2015]. Additionally, satellite-based measurements of NO$_2$

have been observed to correlate better with in situ SO$_2$ measurements than satellite-based

retrievals of SO$_2$ [Sundström et al., 2015]. NO$_2$ is anticipated to be negatively associated with

NPF occurrence and SP due to its association with primary particle emissions and thus

increased CS, but positively associated with J$_n$ and GR due to increased precursor

concentrations.

5. NH$_3$ may play a role in enhancing NPF by acting as a stabilizing base for nucleating clusters,

and is used here following Crippa et al. [2013] (daily NH$_3$ estimates are used herein, versus

seasonal averages in Crippa et al. [2013]), and is anticipated to be positively associated with

NPF occurrence, J$_n$, GR, and SP.

6. Formaldehyde (HCHO) is a product of oxidation of VOCs and one of the few organic species

retrievable from satellite-based measurements. It is used as a proxy for the abundance of low-

volatility VOCs [Chance et al., 2000; Henze and Seinfeld, 2006]. The dominant source of

HCHO estimated from satellite-based measurements appears to be the oxidation of isoprene

because of the short lifetime of isoprene and production of HCHO in the initial oxidation

steps, while most anthropogenic VOC emissions require more oxidation steps prior to HCHO

production and thus are diluted prior to HCHO production [Millet et al., 2008]. Organics play

a role in NPF and/or growth of newly formed particles at least in some environments
[O’Dowd et al., 2002; Zhang et al., 2004a, 2004b; Henze and Seinfeld, 2006; Metzger et al., 2010; Paasonen et al., 2010; Pryor et al., 2011; Pierce et al., 2012; Riipinen et al., 2012; Kulmala et al., 2013], but uncertainty remains regarding whether isoprene products contribute to, or suppress NPF [Surratt et al., 2006; Kiendler-Scharr et al., 2009]. Thus the expected association between HCHO and NPF occurrence is uncertain, but is anticipated to be positively associated with Jn, GR, and SP.

7. The cross product of leaf area index (LAI) and skin temperature (T) is used as an additional proxy of BVOC emissions [Guenther et al., 1993], and is anticipated to be positively associated with NPF occurrence, Jn, GR, and SP.

8. Ozone (O3) is both a key atmospheric oxidant [Helmig, 1997; Seinfeld and Pandis, 2006] and a proxy for atmospheric stagnation [Valente et al., 1998]. Total column O3 as retrieved from the Ozone Monitoring Instrument (OMI) is naturally dominated by stratospheric concentrations, but the presence of a temporal mode of variance at synoptic time scales (see below) indicates that these measurements are also responsive to tropospheric variability. O3 is anticipated to be positively associated with NPF occurrence, Jn, GR, and SP.

Prior research has demonstrated the potential for using satellite-based measurements as proxies for daily ultrafine or nucleation mode particle concentrations [Kulmala et al., 2011; Crippa et al., 2013; Sundström et al., 2015] using AOD and AE, SO2, and UV as predictors. We use a proxy based solely on variables 1-3 (‘simple model’) as a benchmark against which to evaluate whether a model including additional predictor variables: NO2, NH3, HCHO, LAI×T, and O3 (‘full model’) exhibits improved performance.

All observations as obtained from the respective retrieval teams are subject to the following post-processing:
1. For the spectral and spatial correlation analyses, spatially consistent time series are required. Thus, the remotely-sensed measurements are spatially averaged to a 0.5° × 0.5° grid. This resolution was selected to remove some noise through spatial averaging, without removing important mesoscale variability [Anderson et al., 2003]. Due to the lower temporal resolution of the Moderate Resolution Imaging Spectroradiometer (MODIS) LAI measurements (1 in 8 days), LAI×T is excluded from these analyses. For all remaining analyses (Wilcoxon rank sum test, regression trees, and multiple linear regression), all valid retrievals within 100 km of each PSD measurement site are averaged for each observation day to reduce noise, particularly in the trace gas measurements [Krotkov et al., 2008; Fioletov et al., 2011] and to enhance data availability for these predictands.

2. For all variables, days without valid measurements are filled using a weighted mean of the nearest preceding and succeeding measurement days, which will likely reduce the explanatory power of the regression models built herein. The percentages of missing data are given in Table 2.

3. AE is calculated from AOD at 470 and 660 nm after spatial averaging using the Ångström power law [Ångström, 1964].

4. NH₃ measurement availability begins in 2008 (Table 2) and thus they are only available for the full duration of PSD measurements at SGP, SPL, and MMSFb, and for a portion of the PSD measurements at Egbert and MMSFa (Table 1), but many days do not have coincident measurements (e.g. all of 2007) and are filled with a mean NH₃ value. There are no coincident NH₃ and PSD measurements at Duke, and thus NH₃ is excluded from the regression analysis at this site. Satellite-based measurements of NH₃ can have a high associated uncertainty related to unfavorable atmospheric conditions and/or low NH₃
abundances [Whitburn et al., 2016]. Thus, only NH$_3$ pixels with relative uncertainty < 100 % of the retrieved concentration or an absolute error < $5 \times 10^{15}$ molecules cm$^{-2}$ are used.

5. All OMI pixels impacted by the row anomalies [OMI Team, 2012] are treated as missing data.

6. All OMI and Infrared Atmospheric Sounding Interferometer (IASI) retrievals are filtered using a cloud screen to remove retrieval with cloud fractions $>0.3$ [Fioletov et al., 2011; McLinden et al., 2014; Vinken et al., 2014]. The MODIS aerosol retrieval algorithm filters out cloud contaminated pixels prior to averaging spectral reflectances and deriving spectral AOD [Levy et al., 2013], and therefore, no additional cloud screening is applied here.

7. Prior to regression analyses, NO$_2$ and HCHO are log transformed to more closely approximate Gaussian distributions and all predictors are converted to standard normal scores. Therefore any systematic bias in satellite retrievals should not impact the analyses as only relative concentrations are considered. Random errors in the retrievals will propagate through the analyses and are expected to reduce the association between the regression predictors and predictands, but not sign or slope of the relationship.

2.3 Event classification and characteristics, and statistical methods

To quantify similarities and differences in NPF frequency, persistence, and seasonality across North America, an automated methodology is applied to each of the PSD data sets to identify event occurrence and estimate J$_n$, GR, and SP (an earlier version of the approach was described in Sullivan and Pryor [2016]). In brief, a NPF event is reported, and included in the analysis, if:

1. The minimum nucleation mode geometric mean diameter (< 30 nm; D$_{gNuc}$) occurs within 10 hours of the peak nucleation mode number concentration and while the difference in the geometric mean diameter for Dp < 100 nm (D$_g$) and D$_{gNuc}$ is less than or equal to 10 nm,
2. The $r^2$ of the regression fit for the GR calculation (tracking $D_{g\text{nuc}}$ from event start to + 3 hrs.) is $\geq 0.5$,

3. And the event metrics can be reasonably calculated (e.g. GR and $J_n > 0$; eq. 1-6):

$$J_n = \frac{dN_{\text{nuc}}}{dt} + F_{\text{coag}} + F_{\text{growth}} \ldots (1)$$

where $F_{\text{coag}} = \sum_{i=30 \text{ nm}}^{d_{\text{min}(D_p)}} \frac{1}{2} K_{i \text{, } N_i} + \sum_{j=\text{max}(D_p)}^{\text{max}(D_p)} (K_{ij} N_j) \ldots (2)$

and $F_{\text{growth}} = \sum_{i=30 \text{ nm}}^{d_{\text{min}(D_p)}} \left[ \frac{N_i}{30 \text{ nm} - D_{p_{i}}} \right] \cdot GR \ldots (3)$

$$SP = \prod_{i=100 \text{ nm} - \Delta D_p}^{d_{\text{min}(D_p)}} \exp \left( - \frac{t_{\text{growth}}}{\tau_{\text{coag}}} \right) \ldots (4)$$

where $\tau_{\text{growth}} = \frac{(D_{p_{i}+\Delta D_p} - D_{p_{i}})}{GR} \ldots (5)$

and $\tau_{\text{coag}} = \frac{1}{2} K_{i \text{, } N_i} + \sum_{j=\text{max}(D_p)}^{\text{max}(D_p)} K_{ij} N_j \ldots (6)$

where $K$ is the Fuch’s form Brownian coagulation coefficient, $N_{\text{nuc}}$ is the number concentration of particles $\leq 30$ nm, $N_i$ and $D_{p_{i}}$ are the number concentration and median diameter of bin $i$, $\Delta D_p$ is the bin width, and $J_n$ is averaged from event start to + 3 hrs.

The classification algorithm is thus designed to capture unambiguous ‘A’ type (appearance of nucleation mode particles, followed by clear, sustained growth) NPF events and will classify all days not meeting all of the above criteria as a ‘non-event’ day even though new particles may be forming (e.g. ‘B’ or ‘C’ type events (no appearance of particles in the smallest diameters measured or lack of clear, sustained growth) [Pryor et al., 2010]), thereby reducing NPF frequencies relative to a subjective approach.

Prior research indicates that the dominant NPF mechanism may vary seasonally [Yu and Hallar, 2014; Yu et al., 2015]. Thus, the analyses described below are conducted by
climatological season, or for leaf-active (defined here as LAI×T ≥ 50th percentile) and leaf-
dormant (LAI×T < 50th percentile) periods.

The persistence of NPF events is characterized using the conditional probability of events
with a lag of one day (i.e. p(1|1)) relative to the probability of an event on any day (p(1)), thus:

1. p(1) ≈ p(1|1): No autocorrelation
2. p(1) < p(1|1): Positive autocorrelation
3. p(1) > p(1|1): Negative autocorrelation

Once the sites have been characterized in terms of the NPF occurrence, intensity, and
persistence, we then seek to determine if inherent scales in the NPF predictors (satellite
observations) around each site can be used to diagnose and explain the observed consistencies
and site-to-site differences in NPF frequency and event characteristics. To identify the dominant
temporal scales of variability in the satellite-based predictors of NPF and the spatial scale on
which they exhibit coherence, the time series of remotely sensed parameters at each
measurement site were subject to a fast Fourier transformation and used to compute power
spectra. The spatial coherence is defined as the distance from the PSD measurement site at which
the mean correlation coefficient (Pearson’s r) between the time series of the predictors at that site
and each surrounding grid cell drops below an arbitrary threshold of 0.3.

Finally we focus on assessing the potential to extend satellite-based proxies of UFP
concentrations by expanding both the number of predictor variables used (i.e. the suite of
satellite observations included) and the range of characteristics of NPF events considered. We
begin by applying the non-parametric Wilcoxon rank sum test to quantify whether the remote
sensing predictors exhibit different values on NPF event vs. non-event days at each site. This
tests the null hypothesis that it is equally probable that a given observation from one sample is
either greater than or less than a given observation from a second sample (different populations across the range of observations, not solely mean or median). We then build regression trees [Hyvönen et al., 2005] to recursively partition the predictors based on the occurrence (or not) of NPF. In this way we can determine which predictors (and predictor interactions) are most important in terms of predicting whether an individual day as described using the remote sensing variables will be characterized by NPF or not. The predictors that are most important lie closer to the root node and can be used to interpret how the dependence of NPF on a given predictor variable is conditional on other predictor variables. Finally, multiple linear regression models (eq. 7) are fit in which the predictands are $J_n$, GR, SP, and $N_{n-100nm}$ and the predictors are all of the remote sensing variables, and the variance explanation are compared with those from a smaller suite of previously used predictors (i.e. AOD×AE, SO$_2$, and UV [Kulmala et al., 2011; Crippa et al., 2013; Sundström et al., 2015]):

$$y_j = \beta_1 x_{1,j} + \cdots + \beta_i x_{i,j} + \text{constant} \ldots (7)$$

where $y_j$ is the predictand, $\beta_i$ is the coefficient weighting, and $x_{i,j}$ is the standard normal score of the predictor variable, ‘i’, on day, ‘j’.

The predictor variable coefficient weights ($\beta_i$) are used to diagnose which predictors control each event characteristic, and the degree to which they differ among the measurement sites. The results of this analysis are interpreted cautiously because multiple linear regression assumes a linear relationship between the predictors and predictands, multivariate normality, and no multicollinearity amongst the predictors. To investigate the impact of using finite temporal sample, the multiple linear regression is conducted as a Monte Carlo experiment (1000 iterations) in which we sub-sample the PSD datasets to train the regression models using a k-fold (k = 5 folds) cross-validation with 20% of the data withheld from the training model. This is
designed to quantify how model skill and coefficient weightings depend on precise time period of field measurements.

3 Results

3.1 NPF characteristics at the five sites

NPF is frequently observed at all five sites with highest NPF frequency, total sub-100 nm particle concentrations ($N_{n-100\text{nm}}$), and highest GR in spring, with a secondary peak in NPF frequency in fall (except at SPL, where prior analyses have indicated a secondary fall peak [Hallar et al., 2016]) (Figure 2). The discrepancy with prior research at SPL may be due to the lower data availability in fall (due to limited site access) and that while nucleation mode particle formation is observed it is not frequently followed by clear, sustained growth to larger particle sizes (requisite for classification of an event day). Similarly, the low NPF frequency in summer at SGP may be due to missing data in the summer of 2011 and/or the lower temporal frequency (30 min$^{-1}$) of the PSD measurements at the site resulting in fewer days meeting the strict criteria for an event day.

All sites exhibit an overall positive one-day autocorrelation of NPF ($p(1) < p(1|1)$) indicating a higher probability of an event if one occurred on the prior day, although this is not observed for all seasons (Figure 2a). Despite these commonalities, there are also differences in NPF characteristics among the five sites. For example, NPF event frequency is substantially higher at Duke than the other sites (Figure 2a). Also, seasonally averaged GR, particle formation rates ($J_n$), and $N_{n-100\text{nm}}$ are highest at MMSF in both periods (MMSFa and MMSFb) suggesting that the impact of differences in instrumentation during the two sampling periods on NPF metrics is modest compared to the spatial variability. The observation that GR, $J_n$, and $N_{n-100\text{nm}}$ concentrations are highest at MMSF may be due, at least in part, to the lowest min(Dp) at this
site, but given the min(Dp) at Duke differed from that at MMSF by only 1 nm, some of the site-to-site variability in these metrics may also reflect spatial variability in NPF events.

3.2 Spatiotemporal scales of the predictors

The proposed satellite predictors of NPF exhibit similar dominant scales of temporal variability at all sites (Figure 3). SO$_2$ exhibits highest variance at synoptic timescales (3-10 days) at all sites except SPL, where the variance is focused on the annual mode. The dominance of the annual mode at SPL may reflect reduced upwind power plant emissions [Mast et al., 2005] or the fact that this site is frequently in the free troposphere, while the SO$_2$ product is designed to represent SO$_2$ in the planetary boundary layer (Figure 3b) [Krotkov et al., 2006; Fioletov et al., 2011]. Despite large spatial gradients in isoprene emissions and HCHO concentrations across the study area [Millet et al., 2006, 2008], HCHO variability is dominated by the synoptic scale at all sites (Figure 3f), while NO$_2$ concentrations exhibit high variance on both synoptic and seasonal timescales (Figure 3d). O$_3$ concentrations and UV are naturally dominated by the annual cycle (Figure 3c and g) and exhibit the highest spatial scales of coherence (Figure 4), but O$_3$ concentrations also exhibit a variance peak at synoptic time scales (though it is smallest at SPL). The short atmospheric lifetime of NH$_3$ [Clarisse et al., 2009], and seasonality in NH$_3$ fertilizer application and emissions from other agricultural activities [Aneja et al., 2003; Goebes et al., 2003], is reflected in NH$_3$ variability being characterized by the synoptic and seasonal modes (Figure 3e), and that seasonality may be in part responsible for the observed seasonality in GR and J$_n$ (Figure 2). AOD×AE varies on the annual time scale at all sites and strongly at the synoptic time scale at Egbert and SGP, moderately at SPL and MMSF, and weakly at that time scale at Duke. Many of the variables show a minimum in variance at the ~180 day period,
supporting separation of the data into leaf-active and leaf-dormant periods for the regression
analysis below.

In accord with previous analyses that suggest NPF occurs at the regional scale and
exhibits temporal autocorrelation, at least for lags of one day, the predictors also exhibit
relatively large scales of spatial coherence around all PSD measurement sites (Figure 4). Further,
consistent with the highest overall probability of NPF and p(1|1) at Duke, all predictors (except
NH₃) exhibit comparatively large scales of spatial coherence there (Figure 4a). At all sites the
gas phase concentrations (particularly SO₂, NH₃, and HCHO) exhibit smaller scales of coherence
than AOD×AE and UV (Figure 4), potentially indicating that they may play a greater role in
determining the likelihood of NPF events. NH₃ concentrations exhibit greater spatial coherence
in the spring and summer, particularly at Egbert, SPL, and MMSF (Figure 4b, c, and e). This
finding is in conjunction with the higher NH₃ concentrations on event days during the leaf-active
period (Table 3), and may contribute to the spring peak in NPF frequency and high observed GR
at these sites during these seasons [Zhang et al., 2010; Pryor et al., 2011]. NO₂ and AOD×AE
exhibit similar (and large) spatial scales at MMSF and SGP implying anthropogenic primary
particle emissions may dominate accumulation mode concentrations and thus the CS (Figure 4c
and e). HCHO exhibits much larger scales of coherence than SO₂ and NH₃ at Duke due to the
large regional isoprene emissions in the southeastern US [Millet et al., 2008] as reflected in the
high leaf-active HCHO concentrations at this site (Table 3). These local differences in the
temporal variability and spatial coherence of the predictor variables may thus offer partial
explanations for the site-to-site variations in NPF occurrence and characteristics.
3.3 Association between satellite-based measurements and NPF occurrence and event characteristics at the five sites

Consistent with prior research [Yu et al., 2015; Sullivan and Pryor, 2016], NPF frequency and characteristics exhibit marked seasonality (Figure 2) as does the dependence on the satellite-based predictors. For example, NPF is more frequent when LAI×T is lower (higher) during leaf-active (leaf-dormant) measurement days (Table 3), due to the higher frequency of events during spring and fall (Figure 2). The dominant difference in the satellite-based predictors on NPF event versus non-event days during leaf-dormant periods is insolation receipt (UV) and the resultant production of atmospheric oxidants, while the differences in the predictors during the leaf-active periods are more complex. Consistent with the expectation that a higher CS will tend to suppress NPF, non-event days are characterized by higher AOD×AE at all sites during the leaf-active period (Table 3). However, this is not the case during the leaf-dormant season when particle loading is generally lower (except SPL; Table 3) [Sullivan et al., 2015], indicating that another parameter(s) (e.g. availability and/or oxidation of NPF precursors) may limit NPF during leaf-dormant periods. Despite the large uncertainty in NH3 retrievals, consistent with a priori expectations of ternary nucleation, higher concentrations are observed on NPF event days at the two sites located near high NH3 emissions, SGP (significant difference leaf-active; α = 0.1) and MMSFb (higher p25 and p75 leaf-active; significantly difference leaf-dormant) [Goebes et al., 2003; US Environmental Protection Agency, 2011]. Higher HCHO concentrations are observed on leaf-active non-event days at all sites (not significant at Egbert, p-value = 0.11), supportive of the postulate that isoprene (likely the major source of remotely-sensed HCHO) tends to quench available oxidants, reduce H2SO4 production, and suppress NPF. There was no significant difference in HCHO on event and non-event days during leaf-dormant periods, supporting the
assertion that satellite-based measurements of HCHO are primarily indicative of BVOC emissions. Significantly higher O$_3$ concentrations are observed on NPF days during the leaf-active season (though not at Duke or Egbert) possibly indicating that high non-isoprene VOC concentrations are associated with both high O$_3$ production and an increased likelihood of NPF.

Regression trees constructed to ‘predict’ event occurrence illustrate the importance of predictor interactions. For example, in the MMSFa data set the overall NPF frequency is 20%, but increases to 30% when UV $> 26$ mW m$^{-2}$ nm$^{-1}$ and to 39% when UV $> 26$ mW m$^{-2}$ nm$^{-1}$ and LAI×T $< 580$ m$^2$ m$^{-2}$ K (Figure 5e). At Duke, the first node is LAI×T $\approx$ median (Table 3), with an increase in NPF frequency from 26% to 64% conditional on AOD×AE $< 0.23$ if LAI×T $> 670$ m$^2$ m$^{-2}$ K and from 54% to 68% when UV $> 28$ mW m$^{-2}$ nm$^{-1}$ if LAI×T $< 670$ m$^2$ m$^{-2}$ K (cf. 41% for all days; Figure 5a). This is consistent with the event versus non-event day conditions described above, where AOD×AE and UV are important discriminators between event and non-event days during leaf-active and leaf-dormant period, respectively (Table 3). AOD×AE is the 1$^{st}$ or 2$^{nd}$ level node variable at all sites except MMSFa, and in all cases low AOD×AE is associated with increased probability of NPF (Figure 5a-d, and f). LAI×T and UV are also important discriminators of event and non-event days, each being the 1$^{st}$ or 2$^{nd}$ level node at three of the six sites. Higher UV is typically associated with increased probability of NPF, while the relationship with LAI×T is less clear because, as discussed above, NPF frequency is highest at moderate LAI×T (i.e. in spring and fall). O$_3$ is the 1$^{st}$ node and NO$_2$ is a 2$^{nd}$ level node at SPL, where higher (lower) O$_3$ (NO$_2$) favors NPF, indicating that the presence of high concentrations of stabilizing organics coinciding with low anthropogenic emissions is favorable for NPF (Figure 5d). MMSF is in a location of high BVOC and NH$_3$ emissions. Accordingly, NH$_3$ and HCHO are the 2$^{nd}$ level nodes with higher (lower) NH$_3$ (HCHO) being associated with increased probability
of NPF at MMSFb (NH$_3$ measurements are not available for most of MMSFa), again emphasizing the role of a stabilizing base (such as NH$_3$) in promoting NPF and supporting the postulate that high isoprene emissions can suppress NPF (Figure 5f).

In general, multiple linear regression models constructed using the satellite-derived variables as predictors and NPF characteristics as predictands explain more of the variability in GR, J$_n$, SP, and N$_{n-100nm}$ at each site than a random model with equal sample size and number of predictors, indicating that the satellite-based predictors exhibit some explanatory skill in characterizing NPF events over North America (Figure 6). The multiple linear regression models exhibit higher explanatory power for J$_n$, SP, and N$_{n-100nm}$ than for GR, indicating that the proxy variables are better able to capture the intensity of NPF events than the growth rates (Figure 6), potentially because the species that participate in nucleation and subsequent growth may differ [Kulmala et al., 2004]. Further, although there is some site-to-site consistency in terms of which predictor variables have significant coefficients in the models, the absolute form of the regression models is variable from site-to-site, and generally the r$^2$ of the regression models is higher in the leaf-dormant periods (Figure 6 and 7).

At all sites UV has a positive $\beta_i$ for regression equations of J$_n$ (and N$_{n-100nm}$) particularly in leaf-dormant periods. Consistent with higher UV on leaf-dormant NPF event days (Table 3), this indicates that UV not only controls whether NPF occurs, but also the intensity (J$_n$). Conversely, increased UV is associated with decreased SP, consistent with higher formation rates increasing particle loss through self-coagulation and reducing survival probability of individual particles. In general, $\beta_i$ for AOD$\times$AE are negative in equations for J$_n$ and N$_{n-100nm}$, particularly during the leaf-active period. AOD$\times$AE has a significant positive $\beta_i$ for SP during leaf-active events at MMSFa, but a negative $\beta_i$ for SP at SGP and MMSFb, indicating the
controls on survival probabilities may vary both in space and time. Higher AOD×AE is expected
to reduce SP by increasing coagulation loss, but as discussed above AOD×AE appears to be
driven by anthropogenic emissions at MMSF and SGP (Figure 4c and e), and thus may also
indicate the presence of high precursor concentrations. Increased precursor concentrations can
increase GR (e.g. AOD×AE exhibits positive βi for GR at MMSFa) and therefore increase SP,
which may explain the positive association between AOD×AE and SP at MMSFa and the lack of
AOD×AE dependence in the MMSFa regression tree (Figure 5). The sign and significance of the
βi weights on SO2, NO2, NH3, HCHO, and O3 are highly variable by site and leaf activity, which
may reflect differential NPF mechanisms in space and time, and thus explain the site-to-site
differences in NPF frequency and characteristics. For example: SO2 has a positive, significant βi
for Nn-100nm at SPL (leaf-active) and Duke (leaf-dormant), but generally negative βi elsewhere;
NO2 has a negative, significant βi for Nn-100nm at Duke (leaf-active), but positive, significant βi at
Egbert (leaf-active), and variable sign elsewhere; and NH3, HCHO, and O3 are generally split
between positive and negative βi across the sites for all NPF metrics. As each of the predictor
variables are significant at at least one site, we retain them all, but site-to-site variability in which
predictor variable(s) are most important suggests a key challenge in building a generalizable
model.

3.4 Impact of sub-sampling on stability of analyses

The regression trees described in Section 3.3 were built using all measurement days, but
were also built after withholding 20% of the data, for cross validation analysis. The full model
regression trees had re-substitution accuracies of 85-91%. When the testing data is withheld from
the training models, the cross validation accuracies averaged 76-90% (mean standard deviation
of accuracies ~< 1%) across 1000 iterations of cross validation (Figure 5). Thus, sub-sampling
only moderately reduces the models’ accuracy, and the results are relatively stable independent of the specific sub-sampled days.

Running the multiple linear regressions as Monte Carlo experiments, and iteratively cross validating the analysis, provides insights into the models’ stability. Without cross validation, multiple linear regression models explain a significant amount of the observed variability in the NPF characteristics (see cyan triangles in Figure 6). However, when the cross validation is performed, it is not uncommon that the training datasets show poor performance in predicting characteristics (particularly for GR; Figure 6) when some of the validation data were withheld from training the models. This implies the models are not generalizable. Conversely, the models for $J_n$ and $N_{n-100nm}$ at Duke, Egbert (leaf-dormant), and SPL (leaf-dormant), and SP at Duke (leaf-dormant) and SPL (leaf-dormant) seem robust, independent of the data set sub-sampling, indicating the precise measurement dates may not significantly impact the inferences drawn herein.

### 3.5 Improved satellite-based proxy for ultrafine particle concentrations

Both the simple proxy model, where UFP total number concentrations ($N_{n-100nm}$) are predicted using only AOD×AE, SO$_2$, and UV as predictors, and the full models, which include additional predictor variables: NO$_2$, NH$_3$, HCHO, LAI×T, and O$_3$, have much higher explanatory skill when only NPF event days are considered (Figure 7). The full model improves explanatory skill over the simpler model at all sites during both leaf-active and leaf-dormant periods (Figure 7), and is associated with variance explanation ($r^2$) on NPF event days of 29-46% during the leaf-dormant period and 4-37% during leaf-active periods (Figure 7). The increase in variance explanation with the addition of extra predictors is particularly large at Duke, SGP, and MMSF. These are locations influenced by high organic emissions (quantified using HCHO, LAI×T, and
O$_3$), anthropogenic emissions (quantified using NO$_2$), and ternary nucleation precursor emissions (e.g. NH$_3$) (Figure 1) [Goebes et al., 2003], and may be indicative of an enhanced role of these species in dictating UFP concentrations at these sites.

4 Discussion and conclusions

We examine the frequency, persistence, and characteristics of NPF events at five locations across North America, and employ statistical analysis of satellite-based measurements of atmospheric composition to explain spatial similarities and variability in NPF frequency, autocorrelation, formation rates, growth rates, survival probabilities, and daily mean ultrafine particle concentrations. Despite large geographic separation, and vastly different local land use and point source pollution emissions between the sites, NPF is observed at all sites with peak frequencies in spring and fall, and exhibits positive one-day autocorrelation. Accordingly, the temporal modes of variability and spatial scales of coherence of the remotely-sensed variables thought to control NPF also show considerable site-to-site consistency, and are typically coherent on larger scales at sites with larger one-day autocorrelation in NPF occurrence. There is also broad agreement in terms of the conditions associated with NPF events between the sites: NPF is more frequent during moderate LAI×T, low AOD×AE and low HCHO during leaf-active periods, and high UV in leaf-dormant periods. The spatial consistencies in the primary drivers of NPF may explain why simplified nucleation schemes can be used with some skill in global models to characterize the impact of NPF on particle size distributions and CCN concentrations [Spracklen et al., 2008b]. However, event characteristics ($J_n$, GR, SP, and $N_{n-100}$) exhibit greater site-to-site variability in terms of their dependence on the remote-sensing predictors. Site-to-site variability in NPF characteristics and the corresponding variability in satellite-based measurements of the drivers of GR, $J_n$, SP, and $N_{n-100}$ may explain the spatial variability in the performance of simplified NPF schemes [Lee et al., 2013]. Generalized schemes with a single
NPF mechanism and set of coefficients may not be able to capture the variability in precise nucleation mechanisms [Yu et al., 2015] and/or importance of specific precursor species at the different sites. Proxy models of total UFP concentrations that expand the suite of remote sensing predictors exhibit improved variance explanation relative to simpler models that have been previously proposed [Kulmala et al., 2011; Crippa et al., 2013; Sundström et al., 2015] (Figure 7). However, the model coefficients and hence the magnitude and even sign of the dependencies of \( N_{n-100\text{nm}} \) on the suite of predictors considered (AOD×AE, SO2, UV, NO2, NH3, HCHO, \( \text{LAI}\times T, \text{O}_3 \)) (Figure 6) implies great challenges to generating a single generalizable proxy. We recommend future NPF schemes try to reproduce the (spatially variable) relationships between NPF and its drivers presented here, and connect existing theory with ground and satellite observations to evaluate new model treatments of NPF before accepting them, and ultimately improve understanding of regional to global scale impacts of NPF on climate.

Future research is necessary to examine how satellite-based measurement error impacts the explanatory skill (and variability) of the proxies and to further examine the feasibility of developing global proxies for NPF occurrence and characteristics. Further, it may be appropriate to develop the proxies using non-linear techniques or additional (or compound) variables. Given the high uncertainty in direct satellite measurements, the proxies may benefit from use of variables from reanalysis products such as MERRA-2 [Bosilovich et al., 2016]. If such a proxy could be found it may provide computationally efficient first order estimates of the impact of NPF on particle size distributions, CCN concentrations, and ultimately the potential impact of NPF on climate.
5 Acknowledgments

Funding was supplied by: NASA Earth and Space Science Fellowship Program - Grant “14-EARTH14F-0207” (RCS), L’Oréal-UNESCO UK and Ireland Fellowship For Women In Science (PC), Belgian F.R.S-FNRS (LC), Boursier FRIA (SW), NSF (1102309, 1517365) and NASA (NNX16AG31G) (SCP). We also gratefully acknowledge computational resources provided by the Lilly Endowment, Inc. to the Indiana University Pervasive Technology Institute and the Indiana METACyt Initiative. The MODIS and OMI satellite data were collected under the NASA Earth-Sun System Division and Science Mission Directorate, SGP PSD data were acquired under the DoE Atmospheric Radiation Measurement Program. SPL is a permittee of the Medicine-Bow Routt National Forests, and we acknowledge assistance from Randolph Borys and Ian McCubbin. The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Data are available from: MODIS and OMI satellite data (NASA, http://reverb.echo.nasa.gov/reverb), IASA NH₃ (L. Clarisse, lclariss@ulb.ac.be), PSD from Duke (J. T. Walker, Walker.Johnt@epa.gov), Egbert (W. R. Leaitch, richard.leaitch@canada.ca), SGP (US DoE, http://www.archive.arm.gov/), SPL (A. G. Hallar, gannet.hallar@dri.edu), and MMSF (S. C. Pryor, sp2279@cornell.edu).

6 References


Kulmala, M., J. Kontkanen, H. Junninen, K. Lehtipalo, H. E. Manninen, T. Nieminen, T. Petäjä,

Land Processes Distributed Active Archive Center (LP DAAC) (2014), Land Cover Type Yearly L3 Global 0.05Deg CMG, V051. NASA EOSDIS LP DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (https://lpdaac.usgs.gov), accessed 11/13/2015, at http://dx.doi.org/10.5067/MODIS/MCD12C1.006.


Chem. Phys., 10(22), 11223–11242.


### Table 1. Description of particle size distribution measurement sites and instrumentation. Locations of the sites relative to land use and point source emissions are given in figure 1.

<table>
<thead>
<tr>
<th>Site</th>
<th>Location</th>
<th>Elevation (m)</th>
<th>Dates</th>
<th>Instrumenta</th>
<th>Dp range</th>
<th>Bins</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duke Forest, NC</td>
<td>35.98 N 79.09 W</td>
<td>179</td>
<td>11/2005 - 09/2007</td>
<td>SMPS</td>
<td>7 - 289 nm</td>
<td>103</td>
<td>[Pillai et al., 2013]</td>
</tr>
<tr>
<td>Egbert, ON, CAN</td>
<td>44.23 N 79.78 W</td>
<td>251</td>
<td>05/2007 - 05/2008</td>
<td>SMPS</td>
<td>11 - 398 nm</td>
<td>26</td>
<td>[Pierce et al., 2014]</td>
</tr>
<tr>
<td>Storm Peak Laboratory, CO (SPL)</td>
<td>40.46 N 106.74 W</td>
<td>3210</td>
<td>03/2012 - 07/2014</td>
<td>SMPS</td>
<td>9 – 346 nm</td>
<td>104</td>
<td>[Halar et al., 2011]</td>
</tr>
<tr>
<td>Morgan Monroe State Forest, IN (MMSFa)</td>
<td>39.32 N 86.42 W</td>
<td>275</td>
<td>01/2007 - 03/2009</td>
<td>SMPS</td>
<td>6 - 100 nm</td>
<td>81</td>
<td>[Pryor et al., 2010]</td>
</tr>
<tr>
<td>Morgan Monroe State Forest, IN (MMSFb)</td>
<td>39.32 N 86.42 W</td>
<td>275</td>
<td>03/2012 - 12/2013</td>
<td>FMPS</td>
<td>6 - 523 nm</td>
<td>32</td>
<td>[Pryor et al., 2014]</td>
</tr>
</tbody>
</table>

*aSMPS: Scanning Mobility Particle Sizer, FMPS: Fast Mobility Particle Sizer, TDMA: Tandem Differential Mobility Analyzer.

bThese data are acquired as part of the US Department of Energy’s Atmospheric Radiation Measurement program, and to our knowledge have not been published elsewhere; the program and the site are described in this reference.
Table 2. Description of daily (1 in 8 days for LAI) satellite-based measurements used herein.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Satellite/Instrument</th>
<th>Overpass (LST)</th>
<th>Available beginning (% missing(^b::\ total, ≤ 2\ days)</th>
<th>Version</th>
<th>Resolution (at nadir)</th>
<th>Uncertainty/Accuracy(^c)</th>
<th>Proxy for</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Dark target&quot; Aerosol optical depth at 550nm (AOD)</td>
<td>Terra/MODIS</td>
<td>1030</td>
<td>2000 (49 %, 71 %)</td>
<td>Collection 6</td>
<td>10 km</td>
<td>± 0.05 ± 0.15 × AOD</td>
<td>AOD×AE = Condensational sink</td>
<td>[Levy et al., 2013]</td>
</tr>
<tr>
<td>Ångström exponent 470-660 nm (AE)</td>
<td>Terra/MODIS</td>
<td>1030</td>
<td>2000 (49 %, 71 %)</td>
<td>Collection 6</td>
<td>10 km</td>
<td>±0.4(^d)</td>
<td></td>
<td>[Levy et al., 2013]</td>
</tr>
<tr>
<td>Sulfur Dioxide (SO(_2); DU)</td>
<td>Aura/OMI</td>
<td>1345</td>
<td>2004 (49 %, 75 %)</td>
<td>Version 3</td>
<td>13 km × 24 km</td>
<td>Greatest of 1.1 DU, 50%</td>
<td>H(_2)SO(_4) production</td>
<td>[Chance, 2002; Brinksma et al., 2003]</td>
</tr>
<tr>
<td>Local solar noon spectral irradiance at 310 nm (UV; mW m(^-2) nm(^-1))</td>
<td>Aura/OMI</td>
<td>1345</td>
<td>2004 (20 %, 99 %)</td>
<td>Version 3</td>
<td>13 km × 24 km</td>
<td>10%</td>
<td>Production of oxidants</td>
<td>[Chance, 2002; Brinksma et al., 2003]</td>
</tr>
<tr>
<td>Nitrogen Dioxide (NO(_2); molec. cm(^-2))</td>
<td>Aura/OMI</td>
<td>1345</td>
<td>2004 (43 %, 79 %)</td>
<td>Version 3</td>
<td>13 km × 24 km</td>
<td>2 ×10(^{14}) molec. cm(^-2) (30%) background (polluted)(^e)</td>
<td>Anthropogenic emissions</td>
<td>[Chance, 2002; Brinksma et al., 2003]</td>
</tr>
<tr>
<td>Ammonia (NH(_3); molec. cm(^-2))</td>
<td>MetOp/IASI</td>
<td>930(^a)</td>
<td>2008 (58 %, 63 %)</td>
<td>NN Version 1</td>
<td>12 km</td>
<td>&lt; 100% or &lt; 5 × 10(^{15}) molec. cm(^-2)</td>
<td>Ternary nucleation</td>
<td>[Whitburn et al., 2016]</td>
</tr>
<tr>
<td>Formaldehyde (HCHO; molec. cm(^-2))</td>
<td>Aura/OMI</td>
<td>1345</td>
<td>2004 (44 %, 78 %)</td>
<td>Version 3</td>
<td>13 km × 24 km</td>
<td>35%</td>
<td>Production of low-volatility vapors from BVOC</td>
<td>[Chance, 2002; Brinksma et al., 2003]</td>
</tr>
<tr>
<td>Leaf area index (LAI; m(^2) m(^-2))</td>
<td>Terra&amp;Aqua/MODIS</td>
<td>1030&amp; 1330</td>
<td>2002</td>
<td>Version 5</td>
<td>1 km</td>
<td>1 m(^2) m(^-2)</td>
<td>LAI×T = BVOC emissions</td>
<td>[Fang et al., 2012]</td>
</tr>
<tr>
<td>Daytime land surface temperature (T; K)</td>
<td>Terra&amp;Aqua/MODIS</td>
<td>1030&amp; 1330</td>
<td>2002</td>
<td>Version 5</td>
<td>1 °</td>
<td>1 K</td>
<td></td>
<td>[Wan, 2008]</td>
</tr>
<tr>
<td>Ozone (O(_3); DU)</td>
<td>Aura/OMI</td>
<td>1345</td>
<td>2004 (43 %, 79%)</td>
<td>Version 3</td>
<td>13 km × 24 km</td>
<td>Greatest of 10 DU, 5%</td>
<td>Oxidant and stagnation</td>
<td>[Chance, 2002; Brinksma et al., 2003]</td>
</tr>
</tbody>
</table>

\(^a\)9:30 is the overpass time at the equator. Only AM retrievals are used.
bPercentage of missing days (averaged across the five sites and entire satellite observation period), and percentage of missing days with a duration of less than or equal to two consecutive days. MODIS LAI is an 8 day composite product, and thus missing days are not shown for LAI or T.

OMI Accuracy: "root sum of the square of all errors, including forward model, inverse model, and instrument errors" [Brinksma et al., 2003].

MODIS AE is typically bimodal in nature and thus uncertainty is ambiguous [Levy et al., 2010].

When averaged to 26 km × 48 km [Brinksma et al., 2003].
Table 3. Median [p25-p75] conditions from satellite-based measurements on event and non-event days (n) during leaf-active and leaf-dormant measurements. **Bold** indicate rejection of the null hypothesis that the samples are from the same population (α=0.1, Wilcoxon rank sum test). Note NH₃ measurements are available beginning in 2008 and are thus available for only a portion of the PSD measurements at Egbert and MMSFa (no coincident measurements at Duke), and days without coincident measurements (e.g. all of 2007) and are filled with a mean NH₃ value. Thus the distributions of NH₃ on event versus non-event days can be significantly different, despite the median values being strongly driven by missing data.

<table>
<thead>
<tr>
<th></th>
<th>Duke</th>
<th>Egbert</th>
<th>Southern Great Plains</th>
<th>Storm Peak Laboratory</th>
<th>Morgan Monroe State Forest a</th>
<th>Morgan Monroe State Forest b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Event</td>
<td>Non-event</td>
<td>Event</td>
<td>Non-event</td>
<td>Event</td>
<td>Non-event</td>
</tr>
<tr>
<td>n</td>
<td>64</td>
<td>144</td>
<td>55</td>
<td>122</td>
<td>83</td>
<td>655</td>
</tr>
<tr>
<td>AODxAE</td>
<td>0.47</td>
<td>0.68</td>
<td>0.26</td>
<td>0.32</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>[0.14-0.73]</td>
<td>[0.43-0.92]</td>
<td>[0.14-0.37]</td>
<td>[0.22-0.55]</td>
<td>[0.03-0.18]</td>
<td>[0.07-0.25]</td>
</tr>
<tr>
<td>SO₂ (DU)</td>
<td>0.06</td>
<td>-0.05</td>
<td>0.08</td>
<td>-0.05</td>
<td>-0.02</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>[-0.19-0.02]</td>
<td>[-0.21-0.06]</td>
<td>[-0.06-0.21]</td>
<td>[-0.22-0.11]</td>
<td>[-0.20-0.14]</td>
<td>[-0.15-0.11]</td>
</tr>
<tr>
<td>UV (mW m² nm⁻¹)</td>
<td>96</td>
<td>97</td>
<td>82</td>
<td>75</td>
<td>84</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>[74-106]</td>
<td>[86-105]</td>
<td>[69-90]</td>
<td>[50-91]</td>
<td>[67-109]</td>
<td>[75-111]</td>
</tr>
<tr>
<td>NO₂ (×10¹⁵ molec. cm⁻²)</td>
<td>2.91</td>
<td>2.64</td>
<td>2.53</td>
<td>2.11</td>
<td>1.72</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>[2.13-3.94]</td>
<td>[2.22-3.14]</td>
<td>[1.37-3.93]</td>
<td>[1.33-3.53]</td>
<td>[1.30-2.11]</td>
<td>[1.28-2.05]</td>
</tr>
<tr>
<td>NH₃ (×10¹⁵ molec. cm⁻²)</td>
<td>N/A</td>
<td>N/A</td>
<td>2.32</td>
<td>2.32</td>
<td>8.06</td>
<td>6.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[2.32-2.32]</td>
<td>[2.32-2.32]</td>
<td>[4.21-13.4]</td>
<td>[2.68-10.9]</td>
</tr>
<tr>
<td>HCHO (×10¹⁶ molec. cm⁻²)</td>
<td>1.53</td>
<td>1.79</td>
<td>1.00</td>
<td>1.11</td>
<td>1.05</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>[1.17-2.18]</td>
<td>[1.49-2.25]</td>
<td>[0.82-1.26]</td>
<td>[0.86-1.35]</td>
<td>[0.89-1.32]</td>
<td>[0.94-1.43]</td>
</tr>
<tr>
<td>LAIxT (m² m⁻² K)</td>
<td>1053</td>
<td>1140</td>
<td>594</td>
<td>682</td>
<td>322</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>[665-1197]</td>
<td>[1035-1207]</td>
<td>[401-769]</td>
<td>[394-784]</td>
<td>[248-405]</td>
<td>[268-423]</td>
</tr>
<tr>
<td>O₃ (DU)</td>
<td>315</td>
<td>311</td>
<td>311</td>
<td>316</td>
<td>313</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>[301-328]</td>
<td>[303-326]</td>
<td>[294-336]</td>
<td>[294-337]</td>
<td>[298-331]</td>
<td>[295-322]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

34
<table>
<thead>
<tr>
<th>n</th>
<th>108</th>
<th>99</th>
<th>48</th>
<th>130</th>
<th>79</th>
<th>657</th>
<th>86</th>
<th>197</th>
<th>29</th>
<th>203</th>
<th>60</th>
<th>146</th>
</tr>
</thead>
<tbody>
<tr>
<td>AODxAE</td>
<td>0.11</td>
<td>0.10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.07</td>
<td>0.09</td>
<td>0.23</td>
<td>0.23</td>
<td>0.14</td>
<td>0.14</td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>[0.06-0.23]</td>
<td>[0.05-0.19]</td>
<td>[0.10-0.27]</td>
<td>[0.14-0.25]</td>
<td>[0.03-0.16]</td>
<td>[0.21-0.26]</td>
<td>[0.09-0.27]</td>
<td>[0.11-0.20]</td>
<td>[0.02-0.17]</td>
<td>[0.05-0.17]</td>
<td>[0.05-0.17]</td>
<td>[0.05-0.17]</td>
</tr>
<tr>
<td>SO₂ (DU)</td>
<td>0.08</td>
<td>-0.03</td>
<td>-0.09</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.03</td>
<td>-0.15</td>
<td>-0.13</td>
<td>-0.13</td>
<td>-0.20</td>
<td>-0.05</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td>[-0.10-0.26]</td>
<td>[-0.23-0.19]</td>
<td>[-0.53-0.12]</td>
<td>[0.29]</td>
<td>[-0.24-0.17]</td>
<td>[-0.27-0.18]</td>
<td>[-0.38-0.05]</td>
<td>[-0.43-0.23]</td>
<td>[-0.46-0.18]</td>
<td>[-0.40-0.07]</td>
<td>[-0.36-0.20]</td>
<td>[-0.36-0.20]</td>
</tr>
<tr>
<td>UV (mW m⁻² nm⁻¹)</td>
<td>47</td>
<td>29</td>
<td>19</td>
<td>12</td>
<td>37</td>
<td>35</td>
<td>78</td>
<td>50</td>
<td>29</td>
<td>15</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>[36-64]</td>
<td>[21-42]</td>
<td>[10-35]</td>
<td>[7-22]</td>
<td>[31-57]</td>
<td>[25-104]</td>
<td>[47-74]</td>
<td>[28-41]</td>
<td>[10-23]</td>
<td>[27-38]</td>
<td>[18-38]</td>
<td>[18-38]</td>
</tr>
<tr>
<td>NO₂ (×10¹⁵ molec. cm⁻²)</td>
<td>5.23</td>
<td>5.46</td>
<td>3.36</td>
<td>3.95</td>
<td>2.53</td>
<td>2.49</td>
<td>0.39</td>
<td>0.56</td>
<td>5.01</td>
<td>4.79</td>
<td>3.86</td>
<td>4.17</td>
</tr>
<tr>
<td></td>
<td>[4.15-6.71]</td>
<td>[4.32-6.85]</td>
<td>[1.49-5.25]</td>
<td>[2.25-6.65]</td>
<td>[2.01-3.31]</td>
<td>[1.93-5.07]</td>
<td>[0.35-1.00]</td>
<td>[3.32-7.04]</td>
<td>[3.06-5.89]</td>
<td>[3.14-5.39]</td>
<td>[3.25-5.31]</td>
<td>[3.25-5.31]</td>
</tr>
<tr>
<td>NH₃ (×10¹⁵ molec. cm⁻²)</td>
<td>2.32</td>
<td>2.32</td>
<td>3.89</td>
<td>4.17</td>
<td>1.23</td>
<td>1.62</td>
<td>0.18</td>
<td>0.18</td>
<td>4.26</td>
<td>1.75</td>
<td>1.75</td>
<td>6.68</td>
</tr>
<tr>
<td></td>
<td>[1.85-2.32]</td>
<td>[1.85-2.32]</td>
<td>[1.92-9.00]</td>
<td>[1.49-9.71]</td>
<td>[-0.06-3.35]</td>
<td>[-0.05-3.91]</td>
<td>[0.18-0.76]</td>
<td>[0.18-0.76]</td>
<td>[0.87-9.77]</td>
<td>[0.87-9.77]</td>
<td>[0.87-9.77]</td>
<td>[0.87-9.77]</td>
</tr>
<tr>
<td>HCHO (×10¹⁶ molec. cm⁻²)</td>
<td>0.96</td>
<td>0.96</td>
<td>0.84</td>
<td>0.84</td>
<td>1.20</td>
<td>1.04</td>
<td>1.04</td>
<td>0.89</td>
<td>1.06</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>[0.82-1.12]</td>
<td>[0.83-1.12]</td>
<td>[0.64-1.17]</td>
<td>[0.64-1.17]</td>
<td>[0.92-1.20]</td>
<td>[0.90-1.31]</td>
<td>[0.91-1.27]</td>
<td>[0.70-1.11]</td>
<td>[0.91-1.17]</td>
<td>[0.84-1.18]</td>
<td>[0.84-1.18]</td>
<td>[0.84-1.18]</td>
</tr>
<tr>
<td>LAIₓT (m² m⁻² K)</td>
<td>288</td>
<td>245</td>
<td>23</td>
<td>19</td>
<td>149</td>
<td>141</td>
<td>49</td>
<td>49</td>
<td>86</td>
<td>74</td>
<td>163</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>[223-312]</td>
<td>[236-321]</td>
<td>[13-165]</td>
<td>[15-165]</td>
<td>[139-170]</td>
<td>[110-160]</td>
<td>[36-97]</td>
<td>[40-118]</td>
<td>[62-85]</td>
<td>[54-85]</td>
<td>[115-287]</td>
<td>[83-141]</td>
</tr>
<tr>
<td>O₃ (DU)</td>
<td>299</td>
<td>299</td>
<td>319</td>
<td>336</td>
<td>281</td>
<td>294</td>
<td>308</td>
<td>295</td>
<td>315</td>
<td>310</td>
<td>292</td>
<td>296</td>
</tr>
<tr>
<td></td>
<td>[280-324]</td>
<td>[264-326]</td>
<td>[300-366]</td>
<td>[300-366]</td>
<td>[260-302]</td>
<td>[274-318]</td>
<td>[287-336]</td>
<td>[275-336]</td>
<td>[279-336]</td>
<td>[278-336]</td>
<td>[275-314]</td>
<td>[275-314]</td>
</tr>
</tbody>
</table>
Figure 1. Locations at which the particle size distribution measurements analyzed herein were taken (see Table 1 for details). Background denotes the land cover classification from the MODIS combined Terra and Aqua dataset using the LAI/FPAR scheme (type 3) for 2012 [Land Processes Distributed Active Archive Center (LP DAAC), 2014]. Overlaid circles are the 100 largest point source emissions for PM$_{2.5}$ (blue), SO$_2$ (red), NO$_x$ (black), and NH$_3$ (magenta) from the EPA 2011 National Emissions Inventory [US Environmental Protection Agency, 2011], arbitrarily scaled as a fraction of the largest single emission source for each pollutant. Note, point sources do not clearly reflect the spatial patterns of NH$_3$ emissions from animal and fertilizer sources.
Figure 2. a) Probability of a NPF event occurring ($p(1)$; stars) and the probability of an event occurring given that an event occurred on the prior day ($p(1|1)$; terminal point of arrows) by season. The instrument’s minimum particle $D_p$ measured at each site is given in parenthesis in the legend. b) Event metrics (growth rates ($GR$), formation rate ($J_n$, where ‘n’ is determined by the minimum detectable particle diameter for each instrument), and survival probabilities ($SP$)) and daily mean ultrafine particle concentration ($D_p < 100$ nm; $N_{n-100nm}$) by season and location.
Figure 3. Normalized variance of once daily a) AOD×AE, b) SO2, c) UV, d) NO2, e) NH3, f) HCHO, and g) O3 for the grid cells containing each PSD measurement site (indicated by different colors). The variance at each waveperiod is normalized by the maximum variance at any frequency (waveperiod) and the power spectra have been smoothed to emphasize the primary modes of variability. For this analysis the entire satellite measurement period is used. The OMI measurements (SO2, UV, NO2, HCHO, O3) are from 2004 – 2014, MODIS measurements (AOD×AE) are from 2000 – 2014, and IASI measurements (NH3) are from 2008 – 2014. Due to low temporal resolution of MODIS LAI (1 in 8 days), LAI×T is omitted.
Figure 4. Mean correlation ($r$; binned in 100 km distances) between the time series of each of the predictors for the grid cell containing each PSD measurement site and all other grid cells expressed as a function of separation distance. The mean correlation is thus computed over all azimuth directions. The PSD measurement sites are a) Duke Forest, b) Egbert, c) Southern Great Plains, d) Storm Peak Laboratory, and e) Morgan Monroe State Forest. The distance at which $r < 0.3$ by season is indicated by the abscissa location of ‘W’ = winter, ‘Sp’ = spring, ‘Su’ = summer, ‘F’ = fall, and ‘A’ = all measurements, with the ordinate location selected solely for visibility. For this analysis the entire satellite measurement period is used. The OMI measurements (SO$_2$, UV, NO$_2$, HCHO, O$_3$) are from 2004 – 2014, MODIS measurements (AOD×AE) are from 2000 – 2014, and IASI measurements (NH$_3$) are from 2008 – 2014. Due to low temporal resolution of MODIS LAI (1 in 8 days), LAI×T is omitted.
Figure 5. Regression trees for predicting NPF event occurrence at a) Duke, b) Egbert, c) SGP, d) SPL, e) MMSFa, and f) MMSFb. The branch upward (downward) is for all days above (below) the variable threshold given at the node. The colors of the boxes correspond to the variable colors from Figure 4, with the addition that purple is used for LAI×T. The probability of an event is given at each node and the sample size is given in the parenthesis (note the far left nodes are the entire dataset at each site). Also given is the resubstitution accuracy (R) and mean (standard deviation) cross validation accuracy (V) after withholding 20% of measurements days, over 1000 iterations. The trees are built using a maximum of 10 nodes and minimum leaf size of 5, but have been truncated here for legibility.
Figure 6. P-value of the regression coefficients on each of the predictor variables averaged across the 5 training/validation cross-validation data sub-sets and across the 1000 iterations of sub-sampling, used in the multiple linear regression (eq. 7) to predict event metrics (growth rate, formation rate, and survival probability) and daily mean ultrafine particle (Dp < 100 nm; N_{100nm}, where n is the instrument minimum Dp detection limit) concentrations. Satellite-based measurements of NH₃ are not available prior to 2008, and thus are not available for Duke (black fill) and portions of the Egbert and MMSFa PSD measurement days. For each site, the first (second) row is leaf-active (leaf-dormant) period. Red (blue) indicates a positive (negative) coefficient weighting, with the opacity indicating the significance (mean p-value) of the weighing. A cyan triangle on the ordinate indicates a significant p-value (α = 0.1) for the regression model trained on the complete dataset. The shading of the second (black) triangles (upward = leaf-active, downward = leaf-dormant; abscissa offset solely for visibility) are the percentage of data sub-sets that show significantly (α = 0.1) higher r² in predicting the NPF characteristics than expected by random chance (F-statistic for the sample size and number of predictors).
Figure 7. Variance of daily mean UFP particle concentration \( (N_{n-100\text{nm}}) \) explained \( (r^2) \) for each data set using a simple model (stars; where in the predictors are: AOD×AE, SO₂, and UV \cite{Kulmala11, Crippa13, Sundstrom15}) and the full model developed herein (triangles; wherein the predictors are: AOD×AE, SO₂, UV, NO₂, NH₃, HCHO, LAI×T, and O₃), during all measurement days (open symbols) and during only event days (filled symbols).