Forecasting ultrafine particle concentrations from satellite and in situ observations

P. Crippa, S. Castruccio, and S. C. Pryor

1COMET, School of Civil Engineering and Geosciences, Newcastle University, Newcastle Upon Tyne, UK, 2School of Mathematics and Statistics, Newcastle University, Newcastle Upon Tyne, UK, 3Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York, USA

Abstract Recent innovations in remote sensing technologies and retrievals offer the potential for predicting ultrafine particle (UFP) concentrations from space. However, the use of satellite observations to provide predictions of near-surface UFP concentrations is limited by the high frequency of incomplete predictor values (due to missing observations), the lack of models that account for the temporal dependence of UFP concentrations, and the large uncertainty in satellite retrievals. Herein we present a novel statistical approach designed to address the first two limitations. We estimate UFP concentrations by using lagged estimates of UFP and concurrent satellite-based observations of aerosol optical properties, ultraviolet solar radiation flux, and trace gas concentrations, wherein an expectation maximization algorithm is used to impute missing values in the satellite observations. The resulting model of UFP (derived by using an autoregressive moving average model with exogenous inputs) explains 51 and 28% of the day-to-day variability in concentrations at two sites in eastern North America.

1. Introduction

Atmospheric ultrafine particles (UFPs, i.e., particles with diameters < 100 nm) significantly impact human health [Penttinen et al., 2001; Utell and Frampton, 2000] and the Earth’s climate by providing a potential source of cloud condensation nuclei [Pierce et al., 2014]. New particle formation (NPF) is a major source of UFP [Kirkby et al., 2011; Kulmala et al., 2013], often occurs on regional scales, and exhibits high temporal autocorrelation [Crippa and Pryor, 2013; Hussein et al., 2009; Jeong et al., 2010]. However, the subsequent growth of those particles to climate-relevant sizes, their removal, the importance relative to direct emissions of UFP number concentrations, and their role in the current and future radiation budget are incompletely understood [Boucher et al., 2013; Merikanto et al., 2009; Myhre et al., 2013], partly because of the limited availability, in both space and time, of measurements of the particle size distribution (PSD) and composition [Kulmala et al., 2011b; Kulmala et al., 2012]. Thus, there is interest in finding alternative approaches to estimate UFP concentrations in order to better understand their current spatiotemporal scales and infer their role in a changing climate.

Recent innovations in remote sensing technologies and data retrievals offer the potential for predicting UFP concentrations from satellite-borne radiometer measurements [Crippa et al., 2013; Kulmala et al., 2011a; Sullivan et al., 2016; Sullivan and Pryor, 2016; Sundstrom et al., 2015]. The major methodological assumptions and primary findings from previous research in this field can be summarized as follows:

1. Kulmala et al. [2011a] were the first to propose that boundary layer nucleation mode number concentrations could be predicted based on transfer functions, wherein the predictors were drawn from remotely sensed properties. In their analysis the predictors were once-daily satellite-based observations of ultraviolet radiation (UV), sulfur dioxide (SO$_2$), and aerosol optical depth (AOD) and their predictand was 30 min mean particle number concentrations in the diameter range (Dp) of 3–25 nm. They obtained correlations (R) of 0.25 (R$^2$ = 0.06) between nucleation mode number concentrations observed at the Hyvittilä site in Finland and the satellite proxy.

2. Crippa et al. [2013] extended that initial work by developing a proxy for daily mean total UFP number concentrations for Dp 6–100 nm based on satellite observations of SO$_2$ UV, and ammonia (NH$_3$) (as proxies of ternary nucleation); AOD; and the Ångström exponent (AE) (as proxies for the condensational sink). They developed and applied the proxy only to days where all remote sensing observations were available and also considered only days when NPF was observed. The transfer function was developed by using multiple linear regression (MLR), and all predictors were found to have coefficients significantly different to 0.
(i.e., associated \( P < 0.05 \)) except for \( \text{SO}_2 \). Given the importance of sulfuric acid to nucleation, the latter finding was ascribed to the high retrieval uncertainty for \( \text{SO}_2 \). For the two measurement sites included (Morgan Monroe State Forest (MMSF) and Egbert) the \( R^2 \) between \( \log_{10} \) number concentrations of UFP from the measurements and proxy were 0.359 (sample size \( n = 63 \)) and 0.172 (\( n = 53 \)), respectively.

3. **Sundstrom et al.** [2015] sought to overcome the issues with highly uncertain satellite observations of \( \text{SO}_2 \) by instead using remote sensing estimates of columnar nitrogen dioxide (\( \text{NO}_2 \)) as a proxy for sulfuric acid. Their proxy was developed and applied over Southern Africa and for nucleation mode concentrations (\( \text{Dp} < 30 \text{ nm} \)). A proxy based on \( \text{NO}_2/\text{AOD} \) exhibited an \( R^2 \) with observations at two sites of 0.24 and 0.09 (\( n = 49 \) and 124), respectively. A key finding in this research was the substantial degradation in proxy skill when a direct measurement of the condensational sink at the site is replaced by the satellite-derived AOD in the denominator.

4. **Sullivan and Pryor** [2016] sought to expand the suite of predictors drawn from remote sensing observations to include a measure of biological activity and thus potential for release of biogenic volatile organic compounds (the cross-product of leaf area index (LAI) and surface temperature), \( \text{SO}_2 \), \( \text{NO}_2 \), formaldehyde (HCHO) (as a proxy for production of semivolatile organic oxidation products), UV, the cross product of \( \text{AE} \) and \( \text{AOD} \) (as a proxy for the condensational sink), and the cross product of \( \text{UV} \) and \( \text{SO}_2 \) (as a proxy for sulfuric acid production). They also sought to quantify other descriptors of nucleation events: the occurrence (modeled by using logistic regression), formation rates, growth rates, and survival probabilities (modeled by using MLR) using data from the MMSF site. They found that the proxy performance was poorer during leaf-on than leaf-off conditions and that many predictors exhibited regression coefficients that are not significantly different from zero. Nevertheless, when the data are conditionally sampled to extract only NPF event days, the \( R^2 \) for proxy algorithms of new particle formation rates ranged from 0.13 to 0.67 (for sample sizes of 18 to 55).

5. **Sullivan et al.** [2016] extended this earlier work by including in situ data from five sites distributed across North America. They found that, relative to more simple proxies developed in the prior studies (e.g., based on AOD, \( \text{AE} \), \( \text{SO}_2 \), and UV), use of additional predictors (i.e., \( \text{NO}_2 \), as a proxy for primary emitted particles, and \( \text{NH}_3 \), HCHO, LAI, \( T \), and \( \text{O}_3 \)) increases the explained temporal variance of UFP concentrations at all sites. The \( R^2 \) for observed and predicted UFP number concentrations ranged from 0.05 to 0.38 during the leaf-on period and from 0.01 to 0.32 in the leaf-off season. However, the regression coefficients and hence the magnitude and even sign of the dependencies of \( N_n/100 \text{ nm} \) (where \( n \) is the lower limit of measured particle diameter) on the suite of predictors considered (\( \text{AOD} \times \text{AE}, \text{SO}_2, \text{UV}, \text{NO}_2, \text{NH}_3, \text{HCHO}, \text{LAI} \times \text{T}, \text{and O}_3 \)) imply great challenges to generating a single generalizable proxy.

This prior work indicates the following major challenges to development of robust proxies of nucleation metrics and total UFP number concentrations:

1. The large number of missing data in the remote sensing observations (e.g., due to cloud screening). As discussed in **Sullivan et al.** [2016], typically 50% of each variable obtained from satellite-borne observations are missing, even when spatially averaged to a 0.5° × 0.5° grid. This reduces the statistical power (leads to biased estimates of variance explanation, \( R^2 \)) and restricts the analysis to a biased sample (representing only cloud-free conditions).

2. Large uncertainties in the retrievals of some atmospheric properties from the radiometer measurements degrade the explanatory power of the proxies (this is particularly true for \( \text{SO}_2 \)).

3. The satellite measurements are columnar estimates, and thus, there is an implicit assumption of strong coupling between concentrations in the boundary layer where the particle size distribution measurements are made and total column (as represented by the satellite observations). Typically, the proxies are considerably less skillful when satellite-borne data instead of in situ measurements are used as predictors.

In this work, we develop a novel statistical model that is able both to account for the temporal variability of UFP and to predict UFP concentrations near the surface based on prior (temporally lagged) UFP concentrations and concurrent satellite retrievals of aerosol optical properties, nucleation precursors, and solar radiation. While a proxy based solely on satellite-based predictors could potentially be used to generate temporal and spatial fields of UFP concentrations (as in **Crippa et al.** [2013] and **Kulmala et al.** [2011a]), an algorithm that also employs UFP concentration measurements from the previous day might have applications to go-no-go flight decisions conducted as part of field campaigns to examine the spatiotemporal variability in nucleation events [**Pryor et al.**, 2011] and/or their impacts on cloud conditions [**Clarke and Kapustin**, 2010].
The proposed UFP proxy algorithm is innovative because it
1. exploits all the information from satellite retrievals by including days when only part of the predictor suite
   is available and is more sophisticated than past work that has “infilled” missing observations by using a
   simple running mean;
2. provides a unifying framework to describe the temporal variability of UFP by accounting for past UFP
   observations and concurrent satellite retrievals;
3. exhibits high skill in capturing the temporal variability in daily mean UFP concentrations on all days (either
   with or without NPF) and provides more robust estimates of UFP than previous studies that had smaller
   (and biased) data samples and focused solely on conditions during NPF events; and
4. can be used to forecast UFP from satellite data either with or without past UFP observations and provides
   a measure of their relative contribution to predictive skill.

The paper is structured as follows. In section 2, we introduce the ground- and satellite-based observations used
in this study. In section 3, we describe the expectation maximization (EM) algorithm used to impute missing
values and the statistical model developed to capture the temporal structure of data. In section 4, we quantify
the model performance and examine its generalizability by using data from two sites. Section 5 concludes with
remarks on the applicability of this methodology to real-time monitoring of UFP from space.

2. Data

2.1. Ground-Based Observations of Particle Size Distribution

In this work we focus on PSD measurements from two sites in eastern North America: Morgan Monroe State
Forest (MMSF, 39.317°N, 86.417°W) in southern Indiana and the Center for Atmospheric Research
Experiments in Egbert, Canada (44.23°N, 79.78°W). The two sites differ in terms of emission background:
MMSF is a deciduous forest located in a regionally polluted environment, but that is not directly exposed
to local primary anthropogenic sources, whereas Egbert is classified as semirural and is located around
70 km north of Toronto. These sites were selected to examine the flexibility of the proposed method to
different emission contexts, to provide continuity with previous work, and to select sites with a substantial
overlap in the measurement period and because there are relative few long-term measurements of UFP
PSD measurements available in eastern North America [Sullivan et al., 2016]. PSD measurements were
undertaken by using Scanning Mobility Particle Spectrometers at MMSF during January 2007 to March
2009 [Pryor et al., 2010] and at Egbert during May 2007 to May 2008 [Riipinen et al., 2011]. In this work we
focus on the daily mean number concentration of UFP, where UFPs are defined as particles with diameters
in the range of 6–100 nm at MMSF and 10–100 nm for Egbert.

2.2. Satellite Observations of Aerosol Optical Properties and NPF Drivers

Nucleation frequency, intensity, and likelihood of occurrence are dictated by emissions of gaseous precursors
(e.g., SO2 and NH3), drivers of photochemical reactions (e.g., ultraviolet solar irradiance), and of the amount of
preexisting particles [Kirkby et al., 2011]. Based on the prior research described in section 1, the statistical
proxy is designed to include the main variables indicative of secondary formation, which are crucial in the
presence of regional NPF phenomena [Crippa and Pryor, 2013]: AOD and AE as indicators of preexisting
particles and SO2, NH3, and solar radiation as key players in new particle formation and growth processes
in the atmosphere (Table 1).

AOD and AE: Given the critical importance of the condensational sink [Dal Maso et al., 2002] in dictating the
formation and growth of particles in the atmosphere, in our model, we include once-daily AOD and AE level 3
(1° × 1°) observations from the Moderate Resolution Imaging Spectroradiometer (MODIS; Collection 5), as
indicators of aerosol atmospheric columnar loading (AOD at wavelength $\lambda = 550$ nm) and size distribution
(AE, at $\lambda = 470–660$ nm) [Levy et al., 2010]. We focus on observations from the Sun-synchronous Terra satellite
(local equatorial overpass time ~10:30 local solar time (LST)), which are thus indicative of conditions prior to
the occurrence of highest UFP concentrations (typically 10:00–12:00 LST at MMSF [Pryor et al., 2010]). Use of
Collection 5 is dictated by the absence of AE in the Collection 6 released products. The uncertainty in the
satellite retrieval is $\pm 0.05 \pm 0.15 \times$ AOD for AOD and $\pm 0.4$ for AE [Levy et al., 2013].

SO2: Sulfuric acid (H2SO4) has been identified as one of the major players in regulating the rate of NPF
[Sipilä et al., 2010]; thus, herein, we include SO2 as a proxy for H2SO4 concentrations. Monitoring SO2 from
space is challenging due to the high-ozone UV absorption in the UV wavelengths from which SO₂ abundance is derived and to the relatively low concentrations over most developed countries [Krotkov et al., 2006; Vehkamäki et al., 2002]. However, given the importance of H₂SO₄ (and thus SO₂) as a precursor of NPF, we use level 3 daily planetary boundary layer SO₂ at a spatial resolution of 0.25°×0.25° from the Ozone Monitoring Instrument on board NASA’s Aura satellite (OMI, OMSO2e product [Krotkov et al., 2006; Lee et al., 2009]). OMI-retrieved [SO₂] correlates well with ground-based and aircraft measurements, although the measurement uncertainty is largest when small spatial and temporal scales are considered [Lee et al., 2011; Lee et al., 2009]. The accuracy of nonvolcanic SO₂ retrievals is ~50% [Brinksma et al., 2003].

NH₃: There is evidence that ammonia is the dominant stabilizing molecule in nucleation within eastern North America [Pryor et al., 2011]. Thus, here we use level 2 NH₃ measured by the Tropospheric Emission Spectrometer (TES) on board the Aura satellite. It must be acknowledged that TES exhibits highest sensitivity in the lower to middle troposphere (700–900 hPa, i.e., ≈1–3 km above sea level) and that the measurements are subject to relatively large and variable uncertainty (~10%), since no a priori averaging kernel is used in the retrieval [Pinder et al., 2011; Shephard et al., 2011]. Further, the measurements are sparsely distributed in both space and time; thus, [NH₃] at MMSF and Egbert are derived by using an inverse squared distance weighted interpolation technique.

UV: We also include level 3 (resolution = 1°×1°) OMI-Aura local noontime surface UV irradiances at 310 nm (UVB) due to its role as a main driver of photochemical reactions to produce nucleation precursors [Leibensperger et al., 2012; Zhang et al., 2012]. The accuracy of UV retrieval is 10% [Brinksma et al., 2003].

Although in previous work over eastern North America [e.g., Sullivan et al., 2016] the remote sensing data are spatially averaged to 0.5°×0.5°, in the current work, level 3 (gridded at 1°×1° resolution) satellite retrievals are used in the proxy. Given the large scale of spatial coherence in nucleation and also the predictors (e.g., over 1000 km for mean AOD and over 150 km for extreme AOD over most of eastern North America [Sullivan et al., 2015]), this difference in the predictor spatial averaging is not expected to substantially impact the analysis.

### 3. Methodology

UFP measurements in eastern North America exhibit strong autocorrelation [Crippa and Pryor, 2013; Jeong et al., 2010; Sullivan et al., 2016; Sullivan and Pryor, 2016]; thus, we propose a functional form that allows use of only concurrent satellite observations as predictors and also one in which past particle number concentrations are also included in the proxy algorithm to allow an investigation of the dependence of particle concentration from its past values. More specifically, the proxy algorithm is designed to estimate UFP at time \( t \) based on UFP up to time \( t - 1 \) and all five satellite-retrieved properties at time \( t \), separately for the two sites. We further assume a linear dependence of UFP on the predictors to facilitate an easy examination of their contribution to the UFP concentration and to ensure continuity with past research. We thus propose a

### Table 1. Summary Statistics of the Observations Used in This Work and Percentage of Valid Data Available for MMSF and Egbert

| Variable | UFP (cm⁻³) | AOD | AE | UV (mW m⁻² nm⁻¹) | SO₂ (DU) | NH₃ (ppb) |
|----------|------------|-----|----|------------------|---------|-----------|
| Instrument (satellite) | SMPS | MODIS (Terra) | MODIS (Terra) | OMI (Aura) | OMI (Aura) | TES (Aura) |
| Uncertainty/accuracy | N/A | ±0.05 ± 0.15 × AOD | ±0.4 | 10% | 0.5–1 DU | ±0.07 (10%) |
| Reference | [Pryor et al., 2010; Riipinen et al., 2011] | [Levy et al., 2010] | [Levy et al., 2010] | [Brinksma et al., 2003] | [Krotkov et al., 2006] | [Shephard et al., 2011] |
| MMSF | % available | 84 | 50 | 52 | 94 | 25 | 41 |
| min | 4.05 × 10⁵ | 0 | 0.55 | 3.94 | 0.01 | 0.005 |
| mean | 7.19 × 10⁵ | 0.21 | 1.33 | 57.35 | 0.85 | 0.37 |
| max | 7.59 × 10⁵ | 1.04 | 1.78 | 128.70 | 3.39 | 3.75 |
| σ | 5.29 × 10⁵ | 0.18 | 0.43 | 37.67 | 0.67 | 69.5 |
| Egbert | % available | 90 | 45 | 43 | 96 | 26 | 42 |
| min | 8.87 × 10 | 0.002 | 0.57 | 2.58 | 0.01 | 0.01 |
| mean | 2.90 × 10⁶ | 0.23 | 1.48 | 55.63 | 0.83 | 0.35 |
| max | 1.54 × 10⁶ | 2.02 | 1.78 | 121.36 | 1.97 | 3.66 |
| σ | 2.14 × 10⁶ | 0.23 | 0.38 | 37.81 | 0.5 | 69.5 |

*Details on measurement uncertainty/accuracy are provided for each variable. The uncertainty of TES and OMI retrieval is expressed as 1σ (standard deviation) [Krotkov et al., 2006; Shephard et al., 2011].

*N/A, not applicable.
functional form that allows to incorporate both satellite observations and past particle number concentrations and allows data imputation with the expectation maximization (EM) algorithm [Dempster et al., 1977] detailed below.

Since all variables are positive and display a highly skewed distribution, we denote with UFP, AOD, AE, UV, SO2, and NH3, the logarithm (in base 10, but equivalent results would hold for a different base) of the six quantities measured at time \( t \) and with \( Y_t = \text{UFP}_{t-1}, \text{AOD}_{t}, \text{AE}_t, \text{UV}_t, \text{SO}_2, \text{NH}_3 \).

A primary limitation of previous studies on the use of satellite data for UFP estimation is the very small number of occurrences where the vector \( Y_t \) is fully observed. Indeed, for the satellite-derived predictors used herein, data availability is 25% to 94% at MMSF (26% to 96% at Egbert) with [NH3] and [SO2] exhibiting lowest data availability (Table 1). When a data completeness criterion of 90% of hours is applied to represent a valid daily mean UFP concentration, 84% of days are available for MMSF and 90% for Egbert. Similarly, UV measurements are almost complete, with 94% data availability for MMSF and 96% for Egbert. Crippa et al. [2013] proposed a list-wise deletion approach to develop a satellite-based proxy for UFP in which the model was built by using only days when NPF was observed and all satellite-retrieved predictors were available. Such approach uses only 4% of measurement days at MMSF and 9% of days at Egbert. However, the total percentage of available data including vectors with partially available information is considerably higher: 58% for MMSF and 56% for Egbert. Our novel statistical model is able to fully exploit the partially available information from an incomplete suite of predictors. Also, it accounts for the daily temporal dependence of the data and thus naturally includes both nucleation and non-nucleation days.

The predictor suite is initially completed by filling in the missing observations under the assumption that \( Y_t \) follows a vector autoregressive process of order 1:

\[
Y_t = AY_{t-1} + \epsilon_t. \tag{1}
\]

where \( A \) is a 6 × 6 matrix of unknown coefficients and \( \epsilon_t \sim N(0, \Sigma) \). A model selection procedure has shown that higher order autoregressive processes lead to severe overparameterization, as each new order of autoregression requires 6 × 6 new parameters. We impute \( Y_t \) via an EM algorithm [Dempster et al., 1977], consisting of the following steps:

1. All missing values are initially set to the corresponding mean. Denote with \( \hat{Y}_t \) the imputed vector.
2. Plug in \( \hat{Y}_t \) in equation (1) and estimate \( A \) and \( \Sigma \) with least squares.
3. For each time \( t \), whenever at least one entry of \( Y_t \) is missing, the one-step-ahead predictor is computed:

\[
\hat{Y}_t = \hat{A}\hat{Y}_{t-1}. \tag{2}
\]

4. and the missing values are imputed from the predictor; i.e., if an element of \( Y_t \) is missing, then it is a set equal to the corresponding element in \( \hat{Y}_t \).
5. Iterate from point 2 until the maximum absolute difference between consecutive missing values estimates is smaller than a threshold (in practice, after three iterations, this difference is smaller than the epsilon machine).

Figure 1 shows an example of the results of the imputation process for UFP and UV. While the imputation technique captures some of the temporal features present in the time series for short periods of missing data, it is not very informative and generates values very close to the sample mean for long sequences of missing values (e.g., UFP from MMSF during March 2008).

Once the data imputation is performed, we assume the following ARMAX (autoregressive moving average model with exogenous inputs) [Shumway and Stoffer, 2011] for UFP:

\[
\text{UFP}_t = \epsilon_t + \alpha_1 \text{UFP}_{t-1} + \alpha_2 \text{UFP}_{t-2} + \beta_1 \text{AOD}_t + \beta_2 \text{AE}_t + \beta_3 \text{UV}_t + \beta_4 \text{SO}_2 + \beta_5 \text{NH}_3 + \epsilon_t. \tag{3}
\]

where \( \epsilon_t \sim N(0, \sigma^2) \). The model assumes an autoregressive dependence of UFP concentrations from the previous 2 days, as well as a dependence on the pseudosimultaneous retrievals of the five satellite-based variables. Figures 2a–2c show the original time series of daily mean UFP concentrations at both sites with the corresponding fitted values. The fitted values are less variable as they seek to capture the mean and not the intrinsic variability of the process, which should be reproduced via conditional simulations.
4. Results

4.1. Fit and Prediction

Forecasting $UFP_{t+1}, UFP_{t+2}, \ldots$ based on observations up to time $t$ in model (3) requires knowledge of both the past trajectory of UFP and the concurrent values of satellite observations. It is, however, of practical interest to also consider the case when no ground-based measurement of UFP is available for the current period, i.e., when $\alpha_1 = \alpha_2 = 0$ in equation (3). Thus, we consider two models: a full version as in equation (3) which we denote as ALL and a model based only on satellite retrievals (SAT). Excluding the terms that represent the autocorrelation of UFP concentrations naturally degrades the model performance. The coefficient of determination ($R^2$) for ALL at MMSF is 51%, while for SAT it is 13%, and the $R^2$ for ALL is 28% and for SAT is 13% at Egbert (Figures 2b and 2d). The degraded fit is also reflected in performance measures that use some of the data to condition the model, and some data are withheld for model evaluation. In particular, we computed the mean square prediction error (MSPE) for PSD observations in January–December 2008 at MMSF. For each of the first 5 days at the beginning of each calendar month, the model detailed in the previous section is fit by using all data available up to that day and the 1 day ahead prediction $UFP_t$ is performed with both ALL and SAT. The mean square prediction error is then computed as

$$MSPE = \frac{1}{5} \sum_{t=5}^{M} (UFP_t - \hat{UFP}_t)^2,$$

where $M$ is the first day of each month in 2008 (see Table 2). On 2 September 2008 a very high value of UFP was observed at MMSF which caused the 5 day MSPE to be much higher than all the other values for both SAT and ALL (Table 2). Removing this day produces results in line with the other months. ALL produces more accurate predictions for 9 (10 when considering the outlier day) out of 12 calendar months, and the error distribution (median and interquartile range) is considerably closer to zero, a result that highlights the crucial role of the back trajectory (i.e., the autoregressive terms) in the model.

4.2. Comparison of the Sites

At both MMSF and Egbert, the past values of UFP (at time $t - 1$ for Egbert and up to $t - 2$ for MMSF) appear to be the main determinants of current UFP concentrations, which highlights the strong temporal autocorrelation of UFP and supports the postulate of regional-scale NPF events [Crippa and Pryor, 2013; Sullivan and Pryor, 2016]. Besides the intercept, other significant predictors of UFP are UV and AOD, which are the predictors that are characterized by higher accuracy (Table 1). The proposed model is able to capture 51% and 28%
of the variability in daily mean UFP concentrations at MMSF and Egbert, respectively, as indicated by the coefficient of determination ($R^2$) between the observed and predicted UFP (Figure 2 and Table 3), against 36% and 17% of the model in Crippa et al. [2013]. However, it is important to note that the previous work considered only days with NPF and had a much smaller sample size (for the analysis of data from MMSF [Crippa et al., 2013] used 63 days), while this research is based on 821 consecutive days and considers both events with NPF and those without. Besides, the proposed new approach exploits all information available from satellite observations and considers the temporal structure of the data.

The statistical model is also physically consistent with our a priori expectations and prior research: daily mean UFP concentrations are positively associated with UV and past UFP at both sites, whereas a negative association is found for AE and AOD at Egbert (only AE has a negative coefficient at MMSF), thus indicating the role played by photochemical reactions in enhancing particle formation and of preexisting particles in acting as a condensational sink (and thus suppressing NPF). The significant positive association between AOD and UFP at MMSF is consistent with prior work showing that a high AOD × AE is not necessarily associated with nonevent days during the leaf-dormant season and that other parameters, such as the availability of nucleation precursors, may play a major role in controlling NPF and therefore UFP at that site [Sullivan et al., 2016].

| Month   | SAT × 10 | ALL × 10 |
|---------|----------|----------|
| Jan 2008 | 1.35     | 0.86     |
| Feb 2008 | 3.04     | 1.66     |
| Mar 2008 | 0.69     | 1.63     |
| Apr 2008 | 0.11     | 0.02     |
| May 2008 | 1.85     | 1.69     |
| Jun 2008 | 1.21     | 0.41     |
| Jul 2008 | 1.68     | 1.32     |
| Aug 2008 | 0.84     | 1.07     |
| Sep 2008 | 3.21 (15.93) | 3.98 (13.08) |
| Oct 2008 | 0.35     | 0.24     |
| Nov 2008 | 2.26     | 1.77     |
| Dec 2008 | 1.94     | 1.34     |
| Median (Q1, Q3) | 1.11 (0.13, 2.22) | 0.42 (0.04, 1.85) |

Supplementary Table 2. Mean Square Prediction Error (MSPE, in log$_{10}$(UFP)) for 1 Day Ahead Prediction, Averaged for the First 5 Days of a Month at MMSF, According to Model (3) for Both Prediction Approaches ALL and SAT$^a$

$^a$The median and first and third quartiles of MSPE are reported in the last row. In September, two results are shown: the MSPE excluding the outlier of the second and the one including it (in parentheses).
A key question in the development of satellite-based UFP proxy algorithms is the degree to which they are generalizable. Thus, the parameter coefficients from models fitted independently by using data from MMSF and Egbert were compared (Table 3). The coefficients for SO$_2$ and NH$_3$ vary in sign between the sites but are not identified as significant predictors, likely as a result of the high retrieval uncertainty. We tested if the coefficient estimates differ between the sites and found that, among the significant predictors, only the coefficients for UV are not significantly different between MMSF and Egbert. Additionally, we computed error magnitude when the estimated parameters from MMSF are used to predict conditions at Egbert (denoted by $\beta_{MMSF}$ and $\beta_{Egbert}$ were compared (Table 3). The coefficient values that are significant at a confidence level of 5% and 10%, respectively.

5. Discussion and Concluding Remarks

We propose a new approach for estimating daily near-surface UFP concentrations based on satellite retrievals of aerosol optical properties, trace gas concentrations, and solar irradiance which explicitly accounts for the temporal associations between the predictors and allows use of partial information from satellite retrievals by estimating missing values by using an EM approach. This new methodology is not computationally demanding and enables more efficient use of sparse (i.e., discontinuous) satellite observations and can thus be used to make projections for UFP concentrations over a wider array of atmospheric conditions. When the proxy includes temporally lagged UFP number concentrations the proposed model is able to capture 51% and 28% of the variability in daily mean UFP concentrations at MMSF and Egbert, respectively, and is also shown to be able to forecast UFP concentrations. When applied by using only satellite-based predictors the variance explanation ($R^2$) drops to 0.13 (which is still higher than, or equivalent to, the values reported in previous studies).

The approach we propose opens up important new directions for research in reconstructing UFP spatial scales and their associated uncertainty since it could be extended to also consider spatial interpolation. A spatiotemporal model would require robust definition of spatial dependence across all satellite observations and could allow production of near real-time predictions of UFP over eastern North America. Specifically, equation (3) could be generalized to account for a spatially varying linear model as

$$UFP_t(s) = \alpha_0(s) + \alpha_1(s)UFP_{t-1}(s) + \alpha_2(s)UFP_{t-2}(s) + \beta_1(s)AOD_t(s) + \beta_2(s)AE_t(s) + \beta_3(s)UV_t(s) + \beta_4(s)SO2_t(s) + \beta_5(s)NH3_t(s) + \epsilon_t,$$

where $s$ indicates a generic site. The linear coefficients could be modeled as spatially varying coefficients [Gelfand et al., 2003], while the uncertainty could be modeled with a Matérn covariance function [Stein, 1999] or some more sophisticated nonstationary constructions in the presence of heterogeneous geographical features.

| Quantity | MMSF | Egbert | $P$ value (in %) |
|----------|------|--------|-----------------|
| Intercept ($\beta_0$) | 1.43 | 1.94 | 2.5 |
| $UFP_{t-1}$ ($\beta_1$) | 0.54 | 0.35 | 0.3 |
| $UFP_{t-2}$ ($\beta_2$) | 0.08 | −0.03 | 6.1 |
| AOD$_t$ ($\beta_3$) | 0.09 | −0.08 | 0.07 |
| AEl$_t$ ($\beta_4$) | −0.05 | −0.11 | 74.1 |
| UV$_t$ ($\beta_5$) | 0.12 | 0.19 | 23.7 |
| SO$_2$ ($\beta_6$) | −0.01 | 0.04 | 51.2 |
| NH$_3$ ($\beta_7$) | 0.05 | −0.03 | 7.0 |
| $R^2$ | 51% | 28% | |

In the last column, the $P$ value for the $t$ test of equality of the two estimated parameters across sites is reported. In bold and underlined are coefficient values that are significant at a confidence level of 5% and 10%, respectively.
Acknowledgments
This research was supported in part by a L’Oréal-UNESCO UK and Ireland Fellowship for Women in Science (to P.C.), the Natural Environmental Research Council (NERC) through the LIC’s project (ref. NE/K010794/1), and grants to S.C.P. from NSF (1517365) and NASA (NNX16AG31G). The MODIS and OMI satellite data were obtained from the GIOVANNI website (http://giovanni.gsfc.nasa.gov/giovanni/), and the TES data were accessed from the NASA Atmospheric Science Data Center (https://eosweb.larc.nasa.gov/project/tes/tes_t2nh3n_table). The PSD data from Egbert were kindly provided by W.R. Leath (and can be obtained from him on request: richard.leath@canada.ca), and those for MMSF can be downloaded from http://www.geos.cornell.edu/eas/PeoplePlaces/Faculty/pryor/ or can be requested by e-mail (sp2279@cornell.edu). The clarity of this manuscript was substantially improved by the constructive criticism of two reviewers.

References
Boucher, O., et al. (2013), Clouds and Aerosols, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by T.F. Stocker et al., pp. 33–115, Cambridge Univ. Press, Cambridge, U.K., and New York, doi:10.1017/CBO9781107415324.005.

Brinksma, E. J., K. F. Boersma, P. F. Levelt, and R. D. McPeters (2003), OMI validation requirements document, version 1 Rep, 66 pp, Rep. RS-OMIE-404N-345.

Clarke, A., and V. Kapustin (2010), Hemispheric aerosol vertical profiles: Anthropogenic impacts on optical depth and cloud nuclei, Science, 329(5998), 1488–1492, doi:10.1126/science.1198883.

Crippa, P., and S. C. Pryor (2013), Spatial and temporal scales of new particle formation events in eastern North America, Atmos. Environ., 75, 257–264, doi:10.1016/j.atmosenv.2013.04.051.

Crippa, P., D. V. Spracklen, and S. C. Pryor (2013), Satellite-derived estimates of ultrafine particle concentrations over eastern North America, J. Geophys. Res. Atmos., 118, 9968–9981, doi:10.1002/jgrd.50707.

Dal Maso, M., M. Kulmala, K. E. I. Lehtinen, J. M. Makelä, P. Aalto, and C. D. O’Dowd (2002), Condensation and coagulation sinks and formation of nucleation mode particles in coastal and boreal forest boundary layers, J. Geophys. Res., 107(D19), PAR 2–1–PAR 2–10, doi:10.1029/2001JD001053.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B, 39(1), 1–38.

Gelfand, A. E., H.-J. Kim, C. F. Sirmans, and S. Banerjee (2003), Spatial modeling with spatially varying coefficients, J. Am. Stat. Assoc., 98(462), 387–396, doi:10.1198/0162145030000170.

Hussein, T., H. Junninen, P. Tunved, A. Kristensson, M. Dal Maso, I. Riipinen, P. P. Aalto, H. C. Hansson, E. Swietlicki, and M. Kulmala (2009), Time span and spatial scale of regional new particle formation events over Finland and Southern Sweden, Atmos. Chem. Phys., 9(14), 4699–4716, doi:10.5194/acp-9-4699-2009.

Lee, C., R. V. Martin, A. van Donkelaar, H. Lee, R. R. Dickerson, J. C. Hains, N. Krotkov, A. Richter, K. Vinnikov, and J. J. Schwab (2011), SO2 emissions and lifetimes: Estimates from inverse modeling using in situ and global, space-based (SCIAMACHY and OMI), Atmos. Chem. Phys., 11(21), 10,763–10,789, doi:10.5194/acp-11-10743-2011.

Leibensperger, E., L. J. Mickley, D. J. Jacob, W.-T. Chen, J. Seinfeld, A. Nenes, P. Adams, D. Streets, N. Kumar, and D. Rind (2012), Climatic effects of 1950–2050 changes in US anthropogenic aerosols, Atmos. Chem. Phys., 12(6), 3034–3049, doi:10.5194/acp-12-3034-2012.

Levy, R. C., S. Mattoo, L. A. Munchak, L. A. Remer, A. M. Sayer, F. Patadia, and N. C. Hsu (2013), The Collection 6 MODIS aerosol products over land and ocean, Atmos. Meas. Tech., 6, 8601–8616, doi:10.5194/amt-6-8601-2013.

Levy, R. C., L. A. Remer, R. G. Kleidman, S. Mattoo, C. Ichoku, R. Kahn, and T. F. Eck (2010), Global evaluation of the Collection 5 MODIS dark-target aerosol products over land, Atmos. Chem. Phys., 10, 14,075–14,100, doi:10.5194/acp-10-14075-2010.

Levy, R. C., L. A. Remer, R. G. Kleidman, S. Mattoo, L. A. Munchak, L. A. Remer, A. M. Sayer, F. Patadia, and N. C. Hsu (2013), The Collection 6 MODIS aerosol products over land and ocean, Atmos. Meas. Tech., 6(1), 2989–3034, doi:10.5194/amt-6-2989-2013.

Myhre, G., et al. (2013), Anthropogenic and natural radiative forcing, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by T.F. Stocker et al., pp. 659–740, Cambridge Univ. Press, Cambridge, U.K., and New York, doi:10.5194/50317014758.

Pierce, J. R., D. M. Westervelt, S. A. Atwood, E. A. Barnes, and W. R. Leaitch (2014), New-particle formation, growth and climate-relevant particle production in Egbert, Canada: Analysis from 1 year of size-distribution observations, Atmos. Chem. Phys., 14(16), 8647–8663, doi:10.5194/acp-14-8647-2014.

Pinder, R. W., J. T. Walker, J. O. Bash, K. E. Cady-Pereira, D. K. Henze, M. Luo, G. B. Osterman, and M. W. Shephard (2011), Quantifying spatial and seasonal variability in atmospheric ammonia with in situ and space-based observations, Geophys. Res. Lett., 38, L04802, doi:10.1029/2010GL046146.

Pryor, S. C., A. M. Spaulding, and R. J. Barthemlie (2010), New particle formation in the Midwestern USA: Event characteristics, meteorological context and vertical profiles, Atmos. Environ., 44(35), 4413–4425, doi:10.1016/j.atmosenv.2010.07.045.

Pryor, S. C., R. J. Barthemlie, L. L. Sørensen, J. G. McGrath, P. Hopke, and T. Petäjä (2011), Spatial and vertical extent of nucleation events in the Midwestern USA: Insights from the Nucleation in Forests (NiFty) experiment, Atmos. Chem. Phys., 11(4), 1641–1657, doi:10.5194/acp-11-1641-2011.

Riipinen, I., et al. (2011), Organic condensation: A vital link connecting aerosol formation to cloud condensation nuclei (CCN) concentrations, Atmos. Chem. Phys., 11(8), 3865–3878, doi:10.5194/acp-11-3865-2011.

Shephard, M. W., et al. (2011), TES ammonia retrieval strategy and global observations of the spatial and seasonal variability of ammonia, Atmos. Chem. Phys., 11(20), 10,743–10,763, doi:10.5194/acp-11-10743-2011.
Shumway, R. H., and D. S. Stoffer (2011), Time Series Analysis and Its Applications—With R Examples, Springer, New York, doi:10.1007/978-1-4419-7865-3.

Sipilä, M., et al. (2010), The role of sulfuric acid in atmospheric nucleation, Science, 327(5970), 1243–1246, doi:10.1126/science.1180315.

Stein, M. L. (1999), Interpolation of Spatial Data: Some Theory for Kriging, pp. 247, Springer, New York.

Sullivan, R. C., and S. C. Pryor (2016), Dynamic and chemical controls on new particle formation occurrence and characteristics from in situ and satellite-based measurements, Atmos. Environ., 127, 316–325, doi:10.1016/j.atmosenv.2015.12.050.

Sullivan, R. C., R. C. Levy, and S. C. Pryor (2015), Spatiotemporal coherence of mean and extreme aerosol particle events over eastern North America as observed from satellite, Atmos. Environ., 112, 126–135, doi:10.1016/j.atmosenv.2015.04.026.

Sullivan, R. C., P. Crippa, A. G. Hallar, L. Larisse, S. Whitburn, M. Van Damme, W. R. Leaitch, J. Walker, A. Khlystov, and S. C. Pryor (2016), Using satellite-based measurements to explore spatiotemporal scales and variability of drivers of new particle formation, J. Geophys. Res. Atmos., 121, 12,217–12,235, doi:10.1002/2016JD025568.

Sundstrom, A. M., et al. (2015), Characterization of satellite-based proxies for estimating nucleation mode particles over South Africa, Atmos. Chem. Phys., 15(9), 4983–4996, doi:10.5194/acp-15-4983-2015.

Utell, M. J., and M. W. Frampton (2000), Acute health effects of ambient air pollution: The ultrafine particle hypothesis, J. Aerosol Med., 13(4), 355–359, doi:10.1089/jam.2000.13.355.

Vehkamäki, H., M. Kulmala, I. Napari, K. E. J. Lehtinen, C. Timmreck, M. Noppel, and A. Laaksonen (2002), An improved parameterization for sulfuric acid-water nucleation rates for tropospheric and stratospheric conditions, J. Geophys. Res., 107(D22), AAC 3-1–AAC 3-10, doi:10.1029/2002JD002184.

Zhang, R., A. Khalizov, L. Wang, M. Hu, and W. Xu (2012), Nucleation and growth of nanoparticles in the atmosphere, Chem. Rev., 112(3), 1957–2011, doi:10.1021/cr2001756.