Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

Status Update: Is smoke on your mind? Using social media to determine smoke exposure

Bonne Ford[1], Moira Burke[2], William Lassman[1], Gabriele Pfister[3], and Jeffrey R. Pierce[1]

[1]Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, CO 80523

[2]Facebook, Menlo Park, CA 94025

[3]National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, CO 80301

Correspondence to: Bonne Ford (bonne@atmos.colostate.edu)

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

Abstract.

Exposure to wildland-fire smoke is associated with negative effects on human health. However, these effects are poorly quantified. Accurately attributing health endpoints to wildland-fire smoke requires determining the locations, concentrations, and durations of smoke events. Most current methods for determining these smoke-event properties (ground-based measurements, satellite observations, and chemical-transport modeling) are limited temporally, spatially, and/or by their level of accuracy. In this work, we explore using social-media posts regarding smoke, haze, and air quality from Facebook to determine population-level exposure for the summer of 2015 in the western US. We compare this de-identified, aggregated Facebook data to several other datasets that are commonly used for estimating exposure, such as satellite observations (MODIS aerosol optical depth and Hazard Mapping System smoke plumes), surface particulate-matter measurements, and model (WRF-Chem) simulated surface concentrations. After adding population-weighted spatial smoothing to the Facebook data, this dataset is well-correlated (R^2 generally above 0.5) with these other methods in smoke-impacted regions. Removing days with considerable cloud coverage further improves correlations of Facebook data to traditional exposure datasets, which implies that the population is less aware of smoke on cloudy days relative to sunny days. The Facebook dataset is better correlated with surface measurements of $PM_{2.5}$ at a majority of monitoring sites (163 of 293 sites) than the satellite observations and our model simulation are. We also present an example case for Washington state in 2015, where we combine this Facebook dataset with MODIS observations and WRF-Chem simulated $PM_{2.5}$ in a regression model. We show that the addition of the Facebook data improves the regression model's ability to predict surface concentrations. This high correlation of the Facebook data with surface monitors and our Washington state example suggests that this social-media-based proxy can be used to estimate smoke exposure in locations without direct ground-based particulate-matter measurements.

1 Introduction

Exposure to poor air quality is associated with negative impacts on human health (Dockery et al., 1993; Pope, 2007). As such, the Environmental Protection Agency (EPA) has set air-quality standards to limit concentration levels of pollutants in the United States, which has led to reductions in anthropogenic emissions. However, particulate matter (PM) also has natural and transboundary sources, which are more difficult to control. A large natural source of PM in the western US is from wildland fires, which are comprised of wildfires and prescribed burning on natural lands. Wildfire smoke drives much of the interannual variability in total $PM_{2.5}$ (PM with an aerodynamic diameter < 2.5 µm, Jaffe et al., 2008). The 2011 National Emission Inventory (NEI2011, epa.gov) attributes ~20 % of the primary $PM_{2.5}$ emissions in the US to wildfires, 15 % to prescribed fires, and 1.5 % to agricultural fires (epa.gov). Lelieveld et al. (2015) used concentration-response functions derived from previous studies of total ambient PM (and smoking and household air pollution) to estimate that ~2500 premature mortalities are attributable to exposure to biomass-burning (a broad category that includes wildland, prescribed, and agricultural fires) $PM_{2.5}$ per year in the US. However, the assumed toxicity and dose associated with wildland-fire PM were assumed to be the same as all other PM sources. Thus, it is important to determine the health responses to wildland-fire-specific smoke.

Accurately attributing health outcomes to wildland-fire smoke requires a determination of the exposed population. Studies of health impacts often rely on (I) fixed-site monitors (e.g. 3), which are sparse in much of the western US; (II) atmospheric model simulations (Alman et al., 2016; Fann et al., 2012; Johnston et al., 2012; Rappold et al., 2012), which can provide concentration estimates at high spatial resolution, but may be biased by their emission inventories (Davis et al., 2015; Zhang et al., 2014), spatial resolution (Misenis and Zhang, 2010; Punger and West, 2013; Thompson et al., 2014; Thompson and Selin, 2012), or input meteorological fields (Cuchiara et al., 2014; Srinivas et al., 2015; Žabkar et al., 2013); or (III) satellite products (Henderson et al., 2011; Rappold et al., 2011), which do not on their own provide surface-level concentrations. Thus, there is a growing effort to include multiple datasets (e.g. 14, 15) and create blended products (e.g. 16–19) that can exploit the strengths of each dataset. However, all of these methods still only provide estimates of ambient concentration levels and not of actual exposure. Additionally, attributing health effects specifically to wildland-fire-smoke exposure can be difficult as it requires separating the contribution of smoke from total $PM_{2.5}$ (Liu et al., 2015).

In this work, we propose the use of de-identified, aggregated Facebook data to determine population-level exposure for the summer of 2015, which was a particularly smoky year in the US (See Supplementary Figure 1 for number of fire and smoke days). We show that, region-wide, this dataset is better correlated with surface measurements of $PM_{2.5}$ than other traditional mean of estimate exposure, suggesting that it has the potential to be used to estimate smoke exposure in locations without direct ground-based particulate-matter measurements. We also present a test case for Washington state, where we demonstrate that a regression model that includes our Facebook dataset is better able to predict surface $PM_{2.5}$ than a regression model that only has model-simulated $PM_{2.5}$ and satellite AOD. We also compare our results to another measurement of internet behavior, Google Trends, as a proxy for air quality exposure.

The use of social media in risk exposure assessment is a growing field. In the past decade, data mining of social media has provided a wealth of information to news outlets, marketing firms, and the social sciences (Burke and Kraut, 2016; Golder and Macy, 2011; Kosinski et al., 2013; Masedu et al., 2014; Youyou et al., 2015). Only recently has social media and internet behavior been used for research in both the natural sciences and public health. Social media and internet behavior have been proposed to track epidemics and earthquakes (e.g., 14–16), fires (Abel et al., 2012; Bedo et al., 2015; De Longueville et al., 2009; Kent and Jr, 2013), and poor air quality (Jiang et al., 2015; Mei et al., 2014; Tao et al., 2016), and to predict hospitalizations (Ram et al., 2015). A paper by Sachdeva et al. (2016) also proposed the use of Twitter content and geographic information to estimate wildfire smoke concentrations. In this paper, we show how Facebook posting trends "track" significant changes in air quality, such as is associated with dense smoke plumes from large wildfires. Furthermore, we show that Facebook posting trends could also improve estimates of smoke exposure by serving as an extra constraint on more traditional methods for estimating exposure.

2 Methods and Datasets

2.1 Internet Behavior Datasets

Our dataset is the percentage of Facebook posters in each US city that used any of the following words: "smoke", "smoky", "smokey", "haze", "hazey", or "air quality", while attempting to filter out reference to cigarette smoking and other phrases not related to air quality (see Supporting Information). The search generates de-identified and aggregated data; no individual's text was viewed by researchers. Our goal was to focus on wildfire smoke because wildfire smoke often leads to extreme air quality degradation over broad regions of the US in the summertime. However, because this list includes "air quality" and "haze," this search criterion could also highlight trends in Facebook posts discussing air quality degradation due to other emissions, such as fossil-fuel combustion. Geographic location at the city level is determined by IP address. Data was provided for 5 June through 27 October 2015.

We analyzed this dataset of the de-identified, aggregated percent of Facebook posters that matched our search criteria at the city, town, or other municipality level (See Supplementary Figure 2a for location centroids, referred to as "raw" throughout text). In order to estimate the magnitude of the response between these specific locations, we generate a fixed 0.25° grid using an inverse distance weighting with a scale distance (or search neighborhood, d_s) of 20 km and a power of six. The scale distance and power are set to sharply reduce the influence of more-distant observations and were chosen based on the grid resolution in order to maintain the regional variability from the Facebook posting dataset. Additionally, some of these locations are in areas with small populations (and potentially few posters on Facebook), which could skew our results. Therefore, we translate the raw Facebook data to a standard latitude/longitude grid using an area-smoothing procedure with data weighted by the population of the municipality (See Supplementary Figure 2 for example). Our resulting gridded data is determined using the following formula:

$$f_i = \frac{\Sigma\left\{f_c \times \frac{P_c}{\left[1+\left(d_{i,c}/d_s\right)^6\right]}\right\}}{\Sigma\left\{\frac{P_c}{\left[1+\left(d_{i,c}/d_s\right)^6\right]}\right\}} \qquad (1)$$

Where the percent of Facebook posts (f_i) at a grid location (i) is the sum of all of the products of the population (P_c) and the original percent of Facebook posts (f_c) at each "Facebook municipality" (c), weighted by the inverse of the distance (d) between location (i) and the Facebook municipality (c).

We also analyzed Google Trends data (google.com/trends/) as a proxy for exposure. We searched for "air quality", "wildfire", "smoke", "pollution", "haze", "smog", and "ozone" for the time period of 1 May – 31 October 2015 for every designated media area (DMA) in the western U.S. Google Trends results are determined from a random sample of searches with location determined by IP address and duplicates (when the same person searches for the same term multiple times) removed. Results are aggregated and de-identified, but limited to popular terms, with low values appearing as zero (highest values are 100). Therefore, the popularity of a search term impacts the spatial resolution available of the aggregated results (country, DMA, or city). To compare to surface measurements, we identified the DMA in which each measurement site is located.

2.2 Surface Measurements and Satellite Products

We compare these datasets to several other datasets that are commonly used for estimating exposure to wildland-fire smoke. We use 24-hour average concentrations of total $PM_{2.5}$ mass from EPA Air Quality System (AQS, data from epa.gov/aqs) and Interagency Monitoring of Protected Visual Environments (IMPROVE, data from views.cira.colostate.edu/fed/) sites. At IMPROVE network sites, surface measurements of atmospheric composition are taken over a 24-hour period every third day (Malm et al., 1994). Depending on the measurement method at the site, 24-hr average concentrations are provided daily, every third day, or every sixth day at EPA-AQS sites. To maximize our data availability, we are using measurements from Federal Reference Method and Federal Equivalent Method (FRM/FEM, 88101) sites and from non-FRM/FEM (88502) sites (both are also used by the EPA for AQI summaries).

We also use the Hazard Mapping System (HMS) fire and smoke analysis product, which is produced routinely by the National Oceanic and Atmospheric Administration's (NOAA) National Environmental Satellite and Data Information Service (NESDIS) for the purpose of identifying fires and smoke emissions (satepsanone.nesdis.noaa.gov). The HMS product uses observations from both geostationary and polar-orbiting satellites. Polygons determined from satellite visible image analysis are currently categorized as light, moderate and heavy smoke and have assigned numerical values to estimate surface smoke concentrations (5, 16, 27 $\mu g\ m^{-3}$). This product is only available for daylight hours and each polygon is considered valid for a specific time period. We created a gridded surface from all the polygons valid for each day with the surface-concentration values suggested at the same 0.25° grid resolution as our gridded Facebook posts. In grids where there is more than one polygon valid for a day, we take the maximum value in each grid location during that day. Data files were available for every day during our analysis period except 20 August 2015, although sub-daily smoke plume analysis periods could also be missing.

For aerosol optical depth (AOD) from satellites, we use the Collection 6, MODerate resolution Imaging Spectroradiometer (MODIS) Level 2 10-km aerosol optical depth (AOD)

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

products from the Terra and Aqua platforms. Terra has a morning overpass (~10:30 AM LT) and Aqua has an afternoon overpass (~1:30 PM LT). With a swath width of 2,330 km, the instruments provide almost daily coverage of the globe in cloud-free conditions. The MODIS algorithm can have difficulty distinguishing thick smoke from cloud (van Donkelaar et al., 2011), causing some instances of heavy smoke to be erroneously filtered out (although Collection 6 has made improvements to the algorithm to minimize this misclassification, see (Levy et al., 2013)). We average the AOD observations from both instruments on the same 0.25° grid and use all quality levels for better coverage. We additionally use the MODIS cloud fraction (CF) products (“Cloud_Fraction_Land” and “Cloud_Fraction_Ocean,”) in order to determine the presence of clouds and to determine if cloudiness impacts Facebook postings on smoke.

Several models and model frameworks are also routinely used to estimate smoke exposure. Here, we use a chemical transport model, the Weather Research and Forecasting model with Chemistry (WRF-Chem). The simulation was completed for 5 June – 1 October 2015. We use Global Forecast System (GFS) meteorology, biogenic emissions from the Model of Emissions of Gases and Aerosols from Nature (MEGAN, 8), National Emission Inventory 2011 (NEI) anthropogenic emissions, FINN biomass-burning emissions (Wiedinmyer et al., 2011), MOZCART aerosol species and chemistry, and (MOZART) chemical boundaries (Emmons et al., 2010). Horizontal resolution is 15 km and there are 27 vertical levels. Concentrations are output for each model hour, which we then average to provide daily 24-hour average surface concentrations.

2.3 Regression Model

We also present a test case to determine the feasibility and usefulness of including this Facebook dataset in a statistical model. We compare two geographically weighted regression (GWR) models that use MODIS AOD and WRF-Chem $PM_{2.5}$ with and without the Facebook dataset. GWR has previously been used in a several different studies to predict surface air quality (Hu et al., 2013; Song et al., 2014; You et al., 2016). For our test case, we focus on Washington state because of the extensive network of surface $PM_{2.5}$ measurements available for validating results. In our regression model, we determine the dependent variable (surface $PM_{2.5}$ at each measurement site) from a linear combination of these different predictor variables (MODIS AOD, WRF-Chem $PM_{2.5}$, and gridded Facebook dataset). A separate set of regression coefficients is determined at each surface monitor location, which are then interpolated across the domain. We use the leave one out cross validation (LOOCV) method to test our models, in which the regression coefficients determined at a single monitor are removed from the interpolation scheme, and then the resulting $PM_{2.5}$ predicted by the regression model is compared to the observed $PM_{2.5}$ concentrations.

3 Results and Discussion

3.1 Comparison of Facebook Dataset to Observations

An example of the data used in this study is given in Figure 1 for 29 June 2015, which shows a dense smoke plume from wildfires in Canada causing degraded air quality over the Midwestern US and smoke from local fires in the Northwest over Washington, Oregon, and Idaho. The impact of this smoke plume is evident in the HMS smoke product, the anomalously high surface $PM_{2.5}$ concentrations, the elevated AOD values from MODIS, and in the WRF-Chem simulated surface $PM_{2.5}$. The spatial pattern of Facebook postings is somewhat consistent with regions of degraded surface air quality, suggesting some people were aware of the degraded air quality. The extent of the “Facebook plume” does not extend as far east or as far south as the smoke plume observed by the satellite products (MODIS AOD and HMS smoke plumes), and hotspots in the Facebook surface are centered around the eastern Montana/Canada border. To note, the surface measurements also do not show a strong increase in surface concentrations as far south (Missouri and Arkansas), suggesting that the plume observed by the satellites might have been lofted above the surface. Additionally, while the HMS smoke product suggests only light smoke over northeastern Montana and AOD from MODIS is only moderately higher than the surrounding region, surface $PM_{2.5}$ concentrations are elevated, which is in agreement with the spatial pattern in Facebook posts. In cases where the plume is lofted or smoke is concentrated at the surface, this new dataset might be more representative of surface air-quality changes than these satellite products.

In Figure 2, we also show example timeseries of Facebook posts (both the raw and population-weighted) and other datasets (surface $PM_{2.5}$ measurements, MODIS AOD, MODIS CF, HMS smoke) used in this study for four different locations in the western US: Fort Collins, CO; Pinehurst, ID; Bellingham, WA; and Great Falls, MT. From these timeseries, we see the main two fire event periods that impacted large areas of the US during the summer of 2015: (1) the Canadian wildfires in late June through early July and (2) the wildfires in the northwestern

US (mainly Washington and Idaho) in August. The magnitude of impact on these different metrics for estimating smoke varies by location and event. From the timeseries at Pinehurst, ID, where the population was ~1600 in 2015, we can see that population-weighting the Facebook timeseries improves the correlation with surface measurements (R^2= 0.55 for gridded and R^2= 0.00 for raw). In more populated regions, such as Fort Collins, CO (pop. ~161,000), Bellingham, WA (pop. ~85,000), and Great Falls, MT (pop. ~60,000); population-weighting the Facebook data has little impact on the timeseries and resulting correlation with the surface measurements (as shown in Figure 2, where these symbols overlap). Further discussion of these timeseries is presented throughout this result section.

In order to assess how well changes in the fraction of people posting about smoke and air quality in Facebook statuses represent actual changes in surface air quality, we compare timeseries of the percentage of Facebook posts matching our criteria to timeseries of $PM_{2.5}$ measured at all of the different surface sites across the summer of 2015, such as shown in the example timeseries of Figure 2. The coefficients of determination for all surface $PM_{2.5}$-measurement sites with the gridded, population-weighted Facebook posts are shown in Figure 3a, which suggests that the best agreement between the two datasets is in regions that experienced heavy smoke and/or anomalously high $PM_{2.5}$ concentrations during the summer, which is to be expected based on our search criteria. For example, the Mt. Hood IMPROVE site in Oregon (Figure 3) had 39 measurement days (June 5-September 30) and had 14 days when the HMS product suggested smoke over the location. This site provides the best R^2 between the Facebook posting and measured surface $PM_{2.5}$ with a value of 0.97.

We also compare agreement of the Facebook dataset against simulated concentrations from a chemical transport model simulation (WRF-Chem, Figure 3b), which again shows the highest correlation in the Northwest US, which was impacted by wildfire smoke for many days in the summer of 2015. We would expect this as our Facebook search criteria is aimed at smoke and poor air quality and would likely only show changes in postings in regions where air quality was noticeably degraded.

Agreement between MODIS AOD and Facebook posting trends are shown in Figure 3c. Because thick smoke can occasionally be classified as cloud in the MODIS algorithm (van Donkelaar et al., 2011), we filter out AOD observations where the cloud fraction was > 75 %. The impact of this filtering is shown in the timeseries of Figure 2. The criterion reduced our number of useable observations but improved correlations at the majority of sites (Supplementary Figure 3). Comparisons between Facebook and MODIS AOD are similar spatially to WRF-Chem $PM_{2.5}$ and surface measurements, but coefficients for MODIS AOD and Facebook posts are generally worse. However, this satellite product is derived for the full atmospheric column and is not necessarily directly relatable to surface concentrations. Smoke plumes (and transported pollution from other sources) can be lofted above the surface and may not impact surface-level exposure where astute Facebook posters would take notice.

Finally, we also show R^2 values between HMS estimated values and our Facebook dataset in Figure 3d. Again we see similar trends, where the best agreement occurs in regions which experienced numerous smoke days. The correlation values are not as high as for MODIS AOD or WRF-Chem-simulated $PM_{2.5}$. The HMS product only provides estimates for smoke, which is the primary focus of our search criteria although it also includes phrases related to general air quality degradation. Additionally, as with satellite AOD, the HMS product may not be representative of actual surface-level exposure. Finally, the HMS product only provides categorical estimates of “heavy,” “moderate,” or “light” smoke and likely cannot represent subtle changes in exposure concentration levels as compared to AOD.

While we have shown that our new dataset often correlates well with more-traditional datasets that have been used to estimate smoke concentrations/exposure, we also investigate whether Facebook can be used to improve estimates when combined with these other datasets. In Figure 4, we compare how good of a predictor each dataset is at estimating $PM_{2.5}$. We show the coefficients of determination for Facebook posts (4a, similar to 3a but for days where CF < 0.75), MODIS AOD (with CF < 0.75, Figure 4b), the HMS smoke product (Figure 4c), and WRF-Chem simulated $PM_{2.5}$ (Figure 4d) with the surface monitors. From Figure 4, we can determine which dataset best correlates with surface measurements in different regions of the western US.

We included the CF criterion for this analysis for all datasets. We noted that filtering out days with high cloud fraction (CF) improved agreement of Facebook posts and AOD (Figure 2 and Supplementary Figure 4). This led us to also hypothesize that people may have difficulty distinguishing poor air quality on cloudy days, especially farther downwind of a source. To test this, we also sampled our Facebook and surface measurement timeseries at each site with filtering using the MODIS cloud fraction. Compared to correlations between surface measurements and Facebook posts for the full time period, using only the days with CF < 0.75 improved correlations most noticeably at sites that were generally more than 500 km downwind of fires (such as in Colorado, Wyoming, and Utah, Supplementary Figure 4) but had less impact at sites closer to the 2015 wildfires (Oregon, western Montana, Washington, and Idaho, see Supplementary Figure 1a for fire locations). Cloudiness possibly impacting awareness on Facebook is seen in the timeseries for Fort Collins, Colorado in Figure 2a, where, although concentrations were greater during the July event than the August event, the response on Facebook was much less. Bellingham, WA was also impacted by smoke during the same period in July, and although lower surface concentrations were measured, the response in Facebook posts was greater. We noted however, that during the July event, the MODIS product reported a cloud cover of 100 % over Fort Collins. For the full time period, filtering out days with a CF > 0.75, improved the R^2 between Facebook posts and surface measurements in Fort Collins from 0.33 to 0.54. Alternatively, in Great Falls, MT, which had many nearby fires, filtering only changed the R^2 from 0.77 to 0.79, even though roughly the same number of days met the 0.75 criteria for exclusion.

We summarize these initial findings in Figure 4e, which shows the dataset that was best correlated with the surface measurement at each site (and the R^2 had to be greater than 0.5). This figure shows that our Facebook posting dataset is better correlated with actual surface measurements at the majority of sites in our domain for the given time period (5 June – 30 September 2015) compared to other datasets that are typically used to estimate exposure. We do find that MODIS AOD and WRF-Chem $PM_{2.5}$ are better predictors in regions with low populations, such as North Dakota, eastern Montana, and eastern Washington. Additionally, WRF-Chem and MODIS AOD are better predictors over much of the eastern US (not shown, R^2 values all less than 0.5), which is dominated by anthropogenic emissions during the time period, as these "normal" day-to-day changes in anthropogenic pollution may be less likely to be picked up by our Facebook posting search criteria. To note, we did not optimize the configuration of our WRF-Chem simulation to match surface observations. Changing emissions, meteorology, parameterization choices, grid resolution or time steps may have improved surface-concentration estimates, but the optimal configuration would likely differ by region and time period. However, our results shown in Figure 4 suggest that Facebook posting could be used to help estimate exposure in conjunction with these other datasets.

However, if Facebook posts are used to estimate exposure, there may be a few limitations; because, while trends in Facebook posting seem to represent well the variability in surface air quality over our study period at many sites, there is not a simple relationship between posting and $PM_{2.5}$. For one, there did not appear to be a threshold $PM_{2.5}$ concentration at which it was guaranteed that people would start posting, region-wide or at an individual city (e.g. there were cases with high smoke but little posting such as the July event in Fort Collins, CO). There are several potential reasons for this. (1) As noted, on cloudy days, people may not be able to distinguish poor air quality, especially if it is from long-range transport where residents are not aware of a nearby fire. (2) There could be a point of saturation or response fatigue wherein people who have experienced multiple days of smoke may find it less interesting to post about it, or they could have a cognitive bias that causes them to think that air quality has improved in comparison to air quality previously experienced. To test this, we looked also at the timeseries of the ratio of % of Facebook posts to surface concentrations and this ratio does appear to decrease over time during periods of smoke events lasting several days. A decrease throughout the season is only evident at a few sites, although this is difficult to compare since the major smoke event at most sites occurred in late August and early September with few-to-no smoke events occurring afterwards. (3) We noted that occasionally regions with a high Facebook-posting percent was centered over areas where the population had experienced poor air quality on preceding days rather than the current regions of poor air quality. This time shift could suggest that there could sometimes be a lag in either individual's awareness or in the time it takes to spread information among community-level social networks. Additionally, there could also be persistence in Facebook posts, where air quality might improve in a location but people are still posting about it. Conversely, awareness of events could spread through social network more quickly than an air quality event (such as a smoke plume) is transported such that individuals are discussing an event before it actually impacts them. Quantitatively, this is difficult to assess as it may be more event related than season-specific. We did compare +/- 1-day lag correlations between Facebook posts and surface measurements for all sites that had daily measurements (as opposed to every third day). Using the same day provided the best correlation at ~90 % of sites. Slightly better correlations were found using the previous day's measurement at several sites in Utah, and using the following day produced better estimates at several sites in Washington and Oregon, where there were broad regions and extended periods of degraded air quality due to local fires.

3.2 Google Trends comparison with surface observations

We also compared Google trends data to surface measurements of $PM_{2.5}$. Because of the coarse resolution of the aggregated Google Trends data (DMA-level), we chose to compare only to surface measurements and not the other gridded datasets. Our results are shown in Figure 5 for each search term. As with the Facebook dataset, correlations are best in the northwestern U.S., specifically, Washington, Montana, and Oregon, states that were heavily impacted by smoke in 2015. Although we are comparing to total $PM_{2.5}$, the best correlations were found for not only "air quality", but also "wildfire" and "smoke", which, as with the Facebook dataset, we might expect since wildfire smoke was the source of the most variability in surface $PM_{2.5}$ during this time period. The search terms that are more related to urban pollution ("pollution", "smog", "haze" and "ozone") have much lower correlations, and sites that do have R^2>0.1 are generally in urban areas or far downwind of smoke. "Ozone" in particular was not well-correlated with $PM_{2.5}$ measurements (all $R^2 < 0.22$), which should be expected since ozone concentrations and $PM_{2.5}$ concentrations are not always well-correlated (e.g. Reisen et al., 2011).

We also used the Google Trends data to analyze our Facebook search term criteria since we were not able to do this within the Facebook dataset. We chose several words that might be associated with "air quality" and determined the correlations between each word for each DMA as shown in Figure 8. As with the actual concentrations of $PM_{2.5}$, we find that "air quality" is, in general, more associated with "smoke" and "wildfire" than words more commonly associated with urban sources like "smog", "haze", "pollution", and "ozone". In Sachdeva et al. (2016), the authors found that distance from the fires impacted the content of postings about the fire, we also note some differences in our correlation maps based on distance. For example, closer to the fires (WA, OR, ID, MT), "air quality" is more associated with "smoke", while farther away (CO, NV, UT, WY), "air quality" is more associated with "wildfire". At these sites, "air quality" is also better correlated with "wildfire" than "smoke", which may suggest that people are aware of the impact of the wildfires on air quality, but not able to see smoke. However, Google Trends are scaled by popularity in each region and data on only very popular terms are available. This could lead to a discrepancy in that the same amount of people may be searching for these terms in different regions, but the relative popularity may be very different compared to other search terms, especially if there are other physical sources of "smoke" or "air quality" in a region. "Ozone", "smog", and "pollution" (terms that may be more associated with urban air pollution), are not well-correlated with “air quality”, “smoke”, or “wildfires” over our study period; however, “haze” is moderately correlated in WA, OR, and CO (Figure 6).

3.3 Geographically Weighted Regression Test Case

As a first case test to determine the usefulness of this Facebook dataset in a statistical model, we compared two geographically weighted regression model estimates using MODIS AOD and WRF-Chem simulated $PM_{2.5}$ with and without the Facebook dataset. From Figure 4, we see that WRF-Chem, MODIS AOD, and this Facebook dataset are all correlated with surface $PM_{2.5}$ in Washington state, and the best correlated variable varies between surface sites. Therefore, a regression model could allow us to leverage the strengths from each dataset to create an improved estimate.

In Figure 7, we show the results for our regression models with and without the Facebook dataset. We see that including the Facebook dataset in the regression model leads to improved R^2 values at many of the sites in Washington (only one site shows a decrease, Figure 7e). Additionally, for the full dataset (of all sites and all days), there is an improved R^2 (0.66 compared to 0.58), slope (0.60 compared to 0.52), and a smaller error. While, these improvements may be small; we find this is in part due to the fact that the Facebook dataset explains much of the same variability as WRF-Chem simulated surface $PM_{2.5}$ (and better explains variability in the urban region around Seattle, WA).

4 Conclusions

In this paper, we introduced a novel concept of using de-identified, aggregated counts of Facebook posts mentioning smoke, haze, or air quality to determine exposure by comparing to traditional datasets and in a regression model. We also looked at Google Trends data for the same time period and compared it to surface observations. The Facebook dataset was useful in regions meeting two conditions: (1) the region was impacted by wildland-fire smoke, and (2) there was a large-enough population posting to Facebook.The Google Trends data was also best correlated in regions impacted by smoke, however, it is aggregated at a much coarser resolution (DMA-level), therefore the impact of population density is unclear. For regions that meet these two criteria, the Facebook dataset agreed well with more-traditional datasets routinely used for estimating smoke concentrations. In fact, the dataset was often a better predictor of surface $PM_{2.5}$ than several of these other methods and/or datasets (MODIS AOD, HMS smoke plumes, WRF-

Chem simulated $PM_{2.5}$). Therefore, this Faceboook dataset could be useful in determining spatial extent of exposure between surface monitors.

In further investigating regions and time periods of poor agreement, we noted that the cloud cover negatively impacted our correlations, suggesting that some environmental factors might impact people's awareness. We also found that in some regions, correlation improved when comparing to the previous or following day, possibly suggesting some influence of social media on awareness. Some of the disagreement could also be due to our search criteria, which could be further refined. Other studies, which have relied on Twitter messages, have been able to optimize this process by examining subsets of individual posts ("Tweets") to test for false positives. However, again, because this dataset does not provide information on individual posts, this is difficult to do solely within this dataset, but we do plan to test different search criteria in the future to aid in optimizing our dataset.

Even with some of these limitations, we demonstrated that this Facebook dataset has strong potential to be used to estimate exposure to poor air quality. Sachdeva et al. (2016) has shown similar results with Twitter data, but only for a single fire in California. We believe that Facebook could provide some specific advantages over Twitter. Facebook is the most widely used social-media site in the US, with 70 % of its participants active daily (Duggan et al., 2015), compared to 36 % for Twitter. Additionally, only 1 % of Twitter posts are geo-referenced (Thom et al., 2013), and Google Trends relies on a subset of searches for a large region. In Sachdeva et al. (2016), the actual analysis only included 1297 tweets from a 45-day period covering a region of 40,000 km^2 in California and Nevada, and their statistical model was built from 705 tweets for a 37-day period covering a 7,500 km^2 area. With a broader user-base, Facebook could potentially provide better spatial resolution over a broader region. Therefore, this dataset of de-identified, aggregated counts of posts, could be very useful for actually estimating population-level exposure. While we showed that Google Trends data was also moderately well-correlated with surface $PM_{2.5}$ in the Northwest, results were only available for DMAs, of which there are only 210 in the U.S., leading to significantly less spatial information in the Google Trends data than with our Facebook dataset (which has results for >20,000 cities in the U.S.). In 2015, there was a broad region of smoke over much of the U.S.; therefore, correlations with Google Trends may be much higher than if we compared to years with only localized smoke events. Finally, we presented a first test case using Facebook in a statistical model to predict surface concentrations in Washington state for June – September 2015, showing improvements in slope and R^2 values and a reduced error in predicted $PM_{2.5}$. We plan to extend this work in order to provide improved estimates of smoke exposure for the whole western U.S. for the 2015 summer, which will then be used to quantify the health responses associated with exposure to wildfire smoke. Improving the understanding of these specific health effects can potentially aid the public and decision-makers on when and how to take measures to reduce exposure. While social media will not be able to completely replace traditional methods of estimating exposure, social media datasets could currently improve estimates without the costly investment of computer modeling or additional surface monitors. Using social media datasets as a proxy for exposure, also lends itself to analysis of people's response and understanding of smoke exposure (e.g. 42), which cannot be measured by traditional exposure methods.

5 Data Availability

The 24-hour average concentrations of total $PM_{2.5}$ mass are available from the EPA Air Quality System at epa.gov/aqs, and the IMPROVE $PM_{2.5}$ data are also available at views.cira.colostate.edu/fed/. The Collection 6, MODIS Level 2 10-km AOD products from the Terra and Aqua platforms are available at ladsweb.nascom.nasa.gov. The HMS fire and smoke analysis product is available through satepsanone.nesdis.noaa.gov. Google trends data are available at google.com/trends. Our WRF-Chem model output (daily, 24-hour average surface concentrations) is available at http://hdl.handle.net/10217/177042. To preserve the privacy of Facebook users we are unable to provide raw Facebook data, but we provide daily maps of the raw and gridded Facebook dataset at http://hdl.handle.net/10217/177043.

Acknowledgements

This work was funded by NASA Applied Science grant NNX15AG35G.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R. and Tao, K.: Twitcident: Fighting Fire with Information from Social Web Streams, in Proceedings of the 21st International Conference on World Wide Web, pp. 305–308, ACM, New York, NY, USA., 2012.

Alman, B., Pfister, G., Hao, H., Stowell, J., Hu, X., Liu, Y. and Strickland, M. J.: The association of wildfire smoke with respiratory and cardiovascular emergency department visits in Colorado in 2012: a case crossover study, Environ. Health, 15(64), 9, 2016.

Bedo, M., Blanco, G., Oliveira, W., Cazzolato, M., Costa, A., Rodrigues, J., Traina, A. and Traina Jr, C.: Techniques for effective and efficient fire detection from social media images, ArXiv150603844 Cs [online] Available from: http://arxiv.org/abs/1506.03844 (Accessed 29 November 2016), 2015.

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. V., Dentener, F., Dingenen, R. van, Estep, K., Amini, H., Apte, J. S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P. K., Knibbs, L. D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J. L. T., Shaddick, G., Anderson, H. R., Vos, T., Forouzanfar, M. H., Burnett, R. T. and Cohen, A.: Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013, Environ. Sci. Technol., doi:10.1021/acs.est.5b03709, 2015.

Broniatowski, D. A., Paul, M. J. and Dredze, M.: National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, PLOS ONE, 8(12), e83672, doi:10.1371/journal.pone.0083672, 2013.

Burke, M. and Kraut, R. E.: The Relationship between Facebook Use and Well-Being depends on Communication Type and Tie Strength, J. Comput.-Mediat. Commun., 21(4), 265–281, doi:10.1111/jcc4.12162, 2016.

Crooks, A., Croitoru, A., Stefanidis, A. and Radzikowski, J.: #Earthquake: Twitter as a Distributed Sensor System, Trans. GIS, 17(1), 124–147, doi:10.1111/j.1467-9671.2012.01359.x, 2013.

Cuchiara, G. C., Li, X., Carvalho, J. and Rappenglück, B.: Intercomparison of planetary boundary layer parameterization and its impacts on surface ozone concentration in the WRF/Chem model for a case study in Houston/Texas, Atmos. Environ., 96, 175–185, doi:10.1016/j.atmosenv.2014.07.013, 2014.

Davis, A. Y., Ottmar, R., Liu, Y., Goodrick, S., Achtemeier, G., Gullett, B., Aurell, J., Stevens, W., Greenwald, R., Hu, Y., Russell, A., Hiers, J. K. and Odman, M. T.: Fire emission uncertainties and their effect on smoke dispersion predictions: a case study at Eglin Air Force Base, Florida, USA, Int. J. Wildland Fire, doi:10.1071/WF13071, 2015.

De Longueville, B., Smith, R. S. and Luraschi, G.: “OMG, from Here, I Can See the Flames!”: A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires, in Proceedings of the 2009 International Workshop on Location Based Social Networks, pp. 73–80, ACM, New York, NY, USA., 2009.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G. and Speizer, F. E.: An association between air pollution and mortality in six US cities, N. Engl. J. Med., 329(24), 1753–1759, 1993.

van Donkelaar, A., Martin, R. V., Levy, R. C., da Silva, A. M., Krzyzanowski, M., Chubarova, N. E., Semutnikova, E. and Cohen, A. J.: Satellite-based estimates of ground-level fine particulate matter during extreme events: A case study of the Moscow fires in 2010, Atmos. Environ., 45(34), 6225–6232, doi:10.1016/j.atmosenv.2011.07.068, 2011.

van Donkelaar, A., Martin, R. V., Spurr, R. J. D. and Burnett, R. T.: High-Resolution Satellite-Derived PM2.5 from Optimal Estimation and Geographically Weighted Regression over North America, Environ. Sci. Technol., 49(17), 10482–10491, doi:10.1021/acs.est.5b02076, 2015.

Duggan, M., Elison, N. B., Lampe, C., Lenhart, A. and Madden, M.: Social Media Update 2014, Pew Research Center. [online] Available from: http://www.pewinternet.org/2015/01/09/social-media-update-2014/, 2015.

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L. and Kloster, S.: Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4), Geosci Model Dev, 3(1), 43–67, doi:10.5194/gmd-3-43-2010, 2010.

Fann, N., Lamson, A. D., Anenberg, S. C., Wesson, K., Risley, D. and Hubbell, B. J.: Estimating the National Public Health Burden Associated with Exposure to Ambient PM2.5 and Ozone, Risk Anal., 32(1), 81–95, doi:10.1111/j.1539-6924.2011.01630.x, 2012.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L.: Detecting influenza epidemics using search engine query data, Nature, 457(7232), 1012–1014, doi:10.1038/nature07634, 2009.

Golder, S. A. and Macy, M. W.: Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures, Science, 333(6051), 1878–1881, doi:10.1126/science.1202775, 2011.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I. and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmos Chem Phys, 6(11), 3181–3210, doi:10.5194/acp-6-3181-2006, 2006.

Henderson, S. B., Brauer, M., MacNab, Y. C. and Kennedy, S. M.: Three measures of forest fire smoke exposure and their associations with respiratory and cardiovascular health outcomes in a population-based cohort, Environ. Health Perspect., 119(9), 1266, 2011.

Hu, X., Waller, L. A., Al-Hamdan, M. Z., Crosson, W. L., Estes, M. G., Estes, S. M., Quattrochi, D. A., Sarnat, J. A. and Liu, Y.: Estimating ground-level PM(2.5) concentrations in the southeastern U.S. using geographically weighted regression, Environ. Res., 121, 1–10, doi:10.1016/j.envres.2012.11.003, 2013.

Jaffe, D., Hafner, W., Chand, D., Westerling, A. and Spracklen, D.: Interannual variations in PM2.5 due to wildfires in the Western United States., Environ. Sci. Technol., 42, 2812–2818, doi:10.1021/es702755v, 2008.

Jiang, W., Wang, Y., Tsou, M.-H. and Fu, X.: Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter), PLOS ONE, 10(10), e0141185, doi:10.1371/journal.pone.0141185, 2015.

Johnston, F. H., Henderson, S. B., Chen, Y., Randerson, J. T., Marlier, M., DeFries, R. S., Kinney, P., Bowman, D. M. J. S. and Brauer, M.: Estimated Global Mortality Attributable to Smoke from Landscape Fires, Environ. Health Perspect., 120(5), 695–701, doi:10.1289/ehp.1104422, 2012.

Kent, J. D. and Jr, H. T. C.: Spatial patterns and demographic indicators of effective social media content during theHorsethief Canyon fire of 2012, Cartogr. Geogr. Inf. Sci., 40(2), 78–89, doi:10.1080/15230406.2013.776727, 2013.

Kosinski, M., Stillwell, D. and Graepel, T.: Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci., 110(15), 5802–5805, doi:10.1073/pnas.1218772110, 2013.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525(7569), 367–371, doi:10.1038/nature15371, 2015.

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F. and Hsu, N. C.: The Collection 6 MODIS aerosol products over land and ocean, Atmos Meas Tech, 6(11), 2989–3034, doi:10.5194/amt-6-2989-2013, 2013.

Liu, J. C., Pereira, G., Uhl, S. A., Bravo, M. A. and Bell, M. L.: A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke, Environ. Res., 136, 120–132, doi:10.1016/j.envres.2014.10.015, 2015.

Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A. and Cahill, T. A.: Spatial and seasonal trends in particle concentration and optical extinction in the United States, J. Geophys. Res. Atmospheres, 99(D1), 1347–1370, doi:10.1029/93JD02916, 1994.

Masedu, F., Mazza, M., Di Giovanni, C., Calvarese, A., Tiberti, S., Sconci, V. and Valenti, M.: Facebook, quality of life, and mental health outcomes in post-disaster urban environments: the l'aquila earthquake experience, Front. Public Health, 2, 286, doi:10.3389/fpubh.2014.00286, 2014.

Mei, S., Li, H., Fan, J., Zhu, X. and Dyer, C. R.: Inferring air pollution by sniffing social media, in 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), pp. 534–539., 2014.

Misenis, C. and Zhang, Y.: An examination of sensitivity of WRF/Chem predictions to physical parameterizations, horizontal grid spacing, and nesting options, Atmospheric Res., 97(3), 315–334, doi:10.1016/j.atmosres.2010.04.005, 2010.

Pope, C. A.: Mortality Effects of Longer Term Exposures to Fine Particulate Air Pollution: Review of Recent Epidemiological Evidence, Inhal. Toxicol., 19(s1), 33–38, doi:10.1080/08958370701492961, 2007.

Pope III, C. A., Ezzati, M. and Dockery, D. W.: Fine-particulate air pollution and life expectancy in the United States, N. Engl. J. Med., 360(4), 376–386, 2009.

Punger, E. M. and West, J. J.: The effect of grid resolution on estimates of the burden of ozone and fine particulate matter on premature mortality in the USA, Air Qual. Atmosphere Health, 6(3), 563–573, doi:10.1007/s11869-013-0197-8, 2013.

Ram, S., Zhang, W., Williams, M. and Pengetnze, Y.: Predicting Asthma-Related Emergency Department Visits Using Big Data, IEEE J. Biomed. Health Inform., 19(4), 1216–1223, doi:10.1109/JBHI.2015.2404829, 2015.

Rappold, A. G., Stone, S. L., Cascio, W. E., Neas, L. M., Kilaru, V. J., Carraway, M. S., Szykman, J. J., Ising, A., Cleve, W. E., Meredith, J. T., Vaughan-Batten, H., Deyneka, L. and Devlin, R. B.: Peat bog wildfire smoke exposure in rural North Carolina is associated with cardiopulmonary emergency department visits assessed through syndromic surveillance, Env. Health Perspect, 119(10), 1415–1420, doi:10.1289/ehp.1003206, 2011.

Rappold, A. G., Cascio, W. E., Kilaru, V. J., Stone, S. L., Neas, L. M., Devlin, R. B. and Diaz-Sanchez, D.: Cardio-respiratory outcomes associated with exposure to wildfire smoke are modified by measures of community health, Env. Health, 11(1), 71, doi:10.1186/1476-069x-11-71, 2012.

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M. and Balmes, J. R.: Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning, Environ. Sci. Technol., 49(6), 3887–3896, doi:10.1021/es505846r, 2015.

Reisen, F., Meyer, C. P. (Mick), McCaw, L., Powell, J. C., Tolhurst, K., Keywood, M. D. and Gras, J. L.: Impact of smoke from biomass burning on air quality in rural communities in southern Australia, Atmos. Environ., 45(24), 3944–3953, doi:10.1016/j.atmosenv.2011.04.060, 2011.

Sachdeva, S., McCaffrey, S. and Locke, D.: Social media approaches to modeling wildfire smoke dispersion: spatiotemporal and social scientific investigations, Inf. Commun. Soc., 0(0), 1–16, doi:10.1080/1369118X.2016.1218528, 2016.

Song, W., Jia, H., Huang, J. and Zhang, Y.: A satellite-based geographically weighted regression model for regional PM2.5 estimation over the Pearl River Delta region in China, Remote Sens. Environ., 154, 1–7, doi:10.1016/j.rse.2014.08.008, 2014.

Srinivas, C. V., Prasad, K. B. R. R. H., Naidu, C. V., Baskaran, R. and Venkatraman, B.: Sensitivity Analysis of Atmospheric Dispersion Simulations by FLEXPART to the WRF-Simulated Meteorological Predictions in a Coastal Environment, Pure Appl. Geophys., 173(2), 675–700, doi:10.1007/s00024-015-1104-z, 2015.

Tao, Z., Kokas, A., Zhang, R., Cohan, D. S. and Wallach, D.: Inferring Atmospheric Particulate Matter Concentrations from Chinese Social Media Data, PLOS ONE, 11(9), e0161389, doi:10.1371/journal.pone.0161389, 2016.

Thom, D., Jankowski, P., Fuchs, G., Ertl, T., Bosch, H., Andrienko, N. and Andrienko, G.: Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics, Comput. Sci. Eng., 15(3), 72–82, 2013.

Thompson, T. M. and Selin, N. E.: Influence of air quality model resolution on uncertainty associated with health impacts, Atmos Chem Phys, 12(20), 9753–9762, doi:10.5194/acp-12-9753-2012, 2012.

Thompson, T. M., Saari, R. K. and Selin, N. E.: Air quality resolution for health impact assessment: influence of regional characteristics, Atmos Chem Phys, 14(2), 969–978, doi:10.5194/acp-14-969-2014, 2014.

Wiedinmyer, C., Akagi, S. K., Yokelson, R. J., Emmons, L. K., Al-Saadi, J. A., Orlando, J. J. and Soja, A. J.: The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning, Geosci. Model Dev., 4(3), 625–641, doi:10.5194/gmd-4-625-2011, 2011.

Yao, J. and Henderson, S. B.: An empirical model to estimate daily forest fire smoke exposure over a large geographic area using air quality, meteorological, and remote sensing data., J. Expo. Sci. Environ. Epidemiol., (July), 1–8, doi:10.1038/jes.2013.87, 2013.

Yao, J., Brauer, M. and Henderson, S. B.: Evaluation of a wildfire smoke forecasting system as a tool for public health protection, Environ. Health Perspect., 121(10), 1142–1147, doi:10.1289/ehp.1306768, 2013.

You, W., Zang, Z., Zhang, L., Li, Y. and Wang, W.: Estimating national-scale ground-level PM25 concentration in China using geographically weighted regression based on MODIS and MISR AOD, Environ. Sci. Pollut. Res., 23(9), 8327–8338, doi:10.1007/s11356-015-6027-9, 2016.

Youyou, W., Kosinski, M. and Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans, Proc. Natl. Acad. Sci., 112(4), 1036–1040, doi:10.1073/pnas.1418680112, 2015.

Žabkar, R., Koračin, D. and Rakovec, J.: A WRF/Chem sensitivity study using ensemble modelling for a high ozone episode in Slovenia and the Northern Adriatic area, Atmos. Environ., 77, 990–1004, doi:10.1016/j.atmosenv.2013.05.065, 2013.

Zhang, F., Wang, J., Ichoku, C., Hyer, E. J., Yang, Z., Ge, C., Su, S., Xiaoyang Zhang, Kondragunta, S., Kaiser, J. W., Wiedinmyer, C. and Silva, A. da: Sensitivity of mesoscale modeling of smoke direct radiative effect to the emission inventory: a case study in northern sub-Saharan African region, Environ. Res. Lett., 9(7), 75002, doi:10.1088/1748-9326/9/7/075002, 2014.

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

Atmospheric Chemistry and Physics
Open Access
Discussions

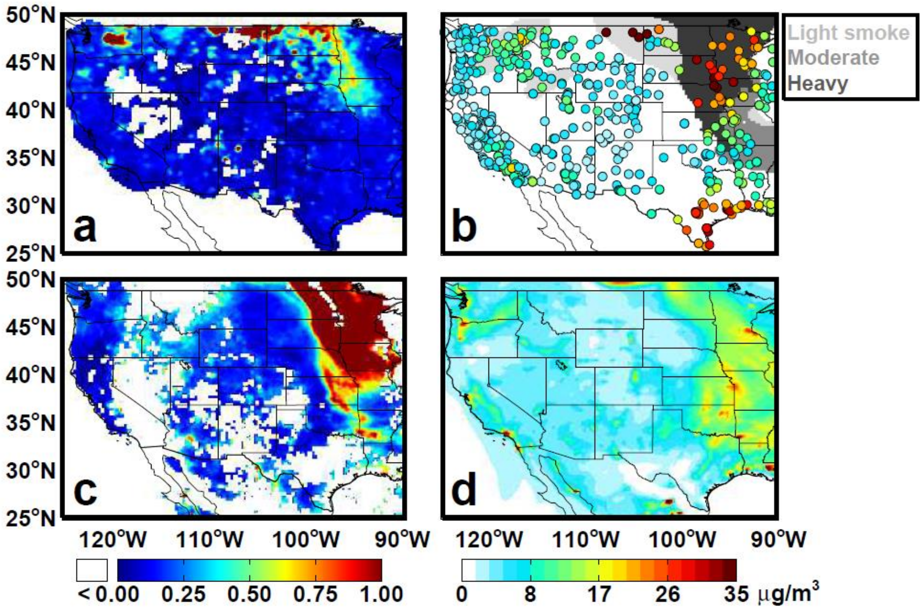

Figure 1. Example of datasets for 29 June 2015. a.) Population-weighted (Equation 1) percent of Facebook posts meeting criterion (white signifies regions with weighted population < 10), b.) gridded HMS smoke plumes overlaid with 24-hr average surface $PM_{2.5}$ concentrations from surface measurement sites, c.) gridded, unfiltered MODIS-Aqua and MODIS-Terra AOD (white signifies no valid observation), and d.) WRF-Chem simulated 24-hr average surface $PM_{2.5}$ concentrations.

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

Atmospheric Chemistry and Physics
Discussions

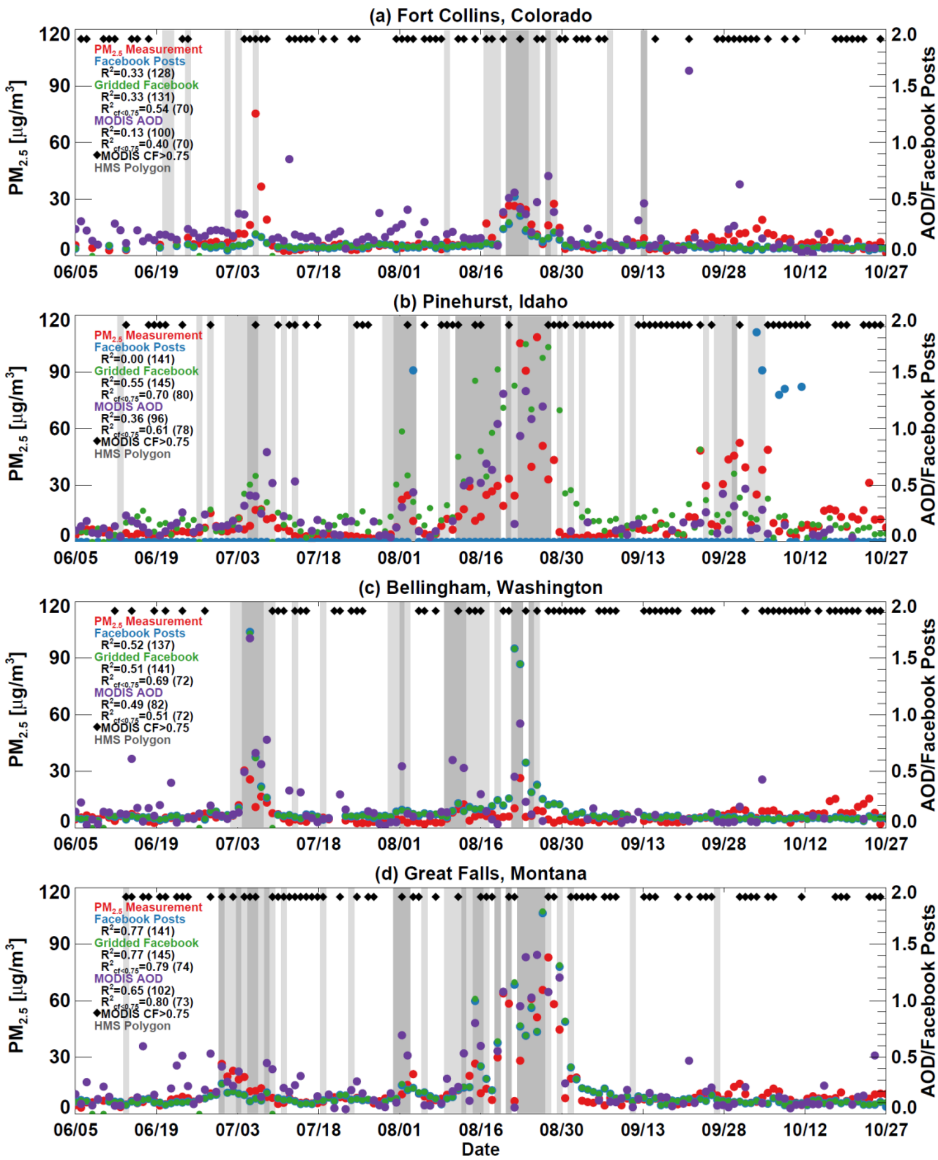

Figure 2. Timeseries of measured surface $PM_{2.5}$ concentrations (red), "raw" percent of Facebook posts matching criteria (blue), gridded and population-weighted percent of Facebook posts (green), MODIS AOD (purple), days with HMS-denoted light (light gray) and moderate/thick (dark gray) smoke, and days where the MODIS cloud fraction was greater than 0.75 (black diamonds) at (a) Fort Collins, CO; (b) Pinehurst, ID; (c) Bellingham, WA; and (d) Great Falls, MT for 5 June – 27 October 2015. R^2 values for each dataset with the surface measurement are given along with the number of days available for the calculation noted in parentheses.

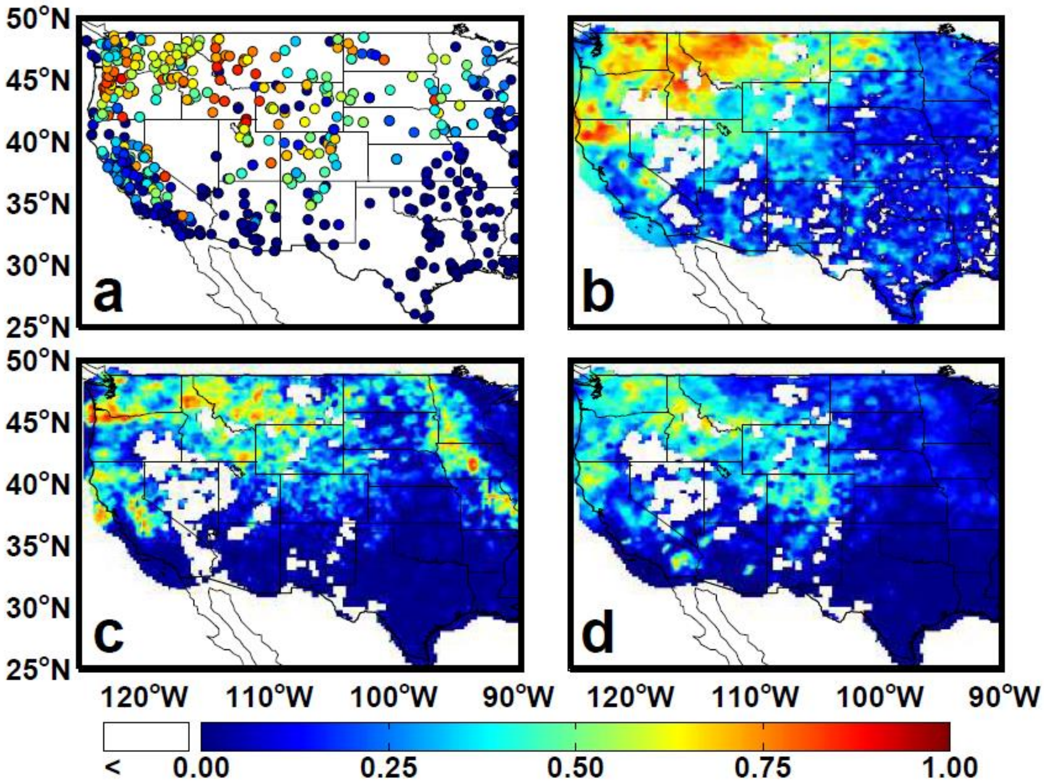

Figure 3. R^2 values for % Facebook posts and a.) IMPROVE and EPA-AQS surface measurements of $PM_{2.5}$ (for sites with > 35 days of measurements), b.) WRF-Chem simulated $PM_{2.5}$, c.) MODIS AOD when cloud fraction was below 0.75 and d.) HMS smoke for the period of 5 June – 30 September 2015.

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

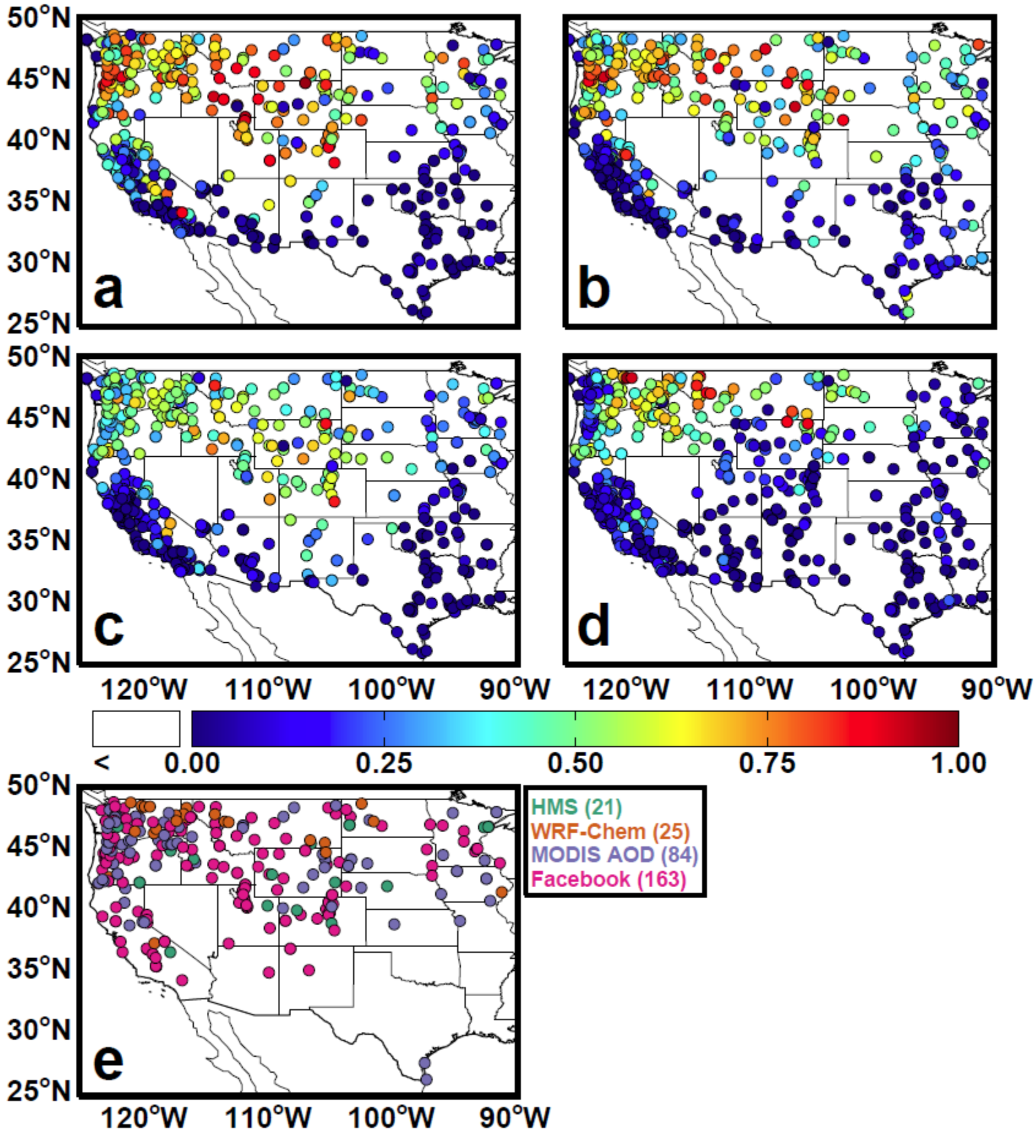

Figure 4. R^2 values for surface measurements of $PM_{2.5}$ with a.) Facebook (CF<0.75), b.) MODIS AOD (CF<0.75), c.) HMS smoke, and d.) WRF-Chem simulated $PM_{2.5}$, for the period of 5 June – 30 September 2015. e.) Product (HMS Smoke, WRF-Chem $PM_{2.5}$, MODIS AOD, or Facebook posts) that has the highest R^2 compared to surface measurements for the time period of 5 June – 30 September 2015 (sites are shown only if the resulting $R^2 > 0.5$). Number of sites in western US (domain shown) where product has highest R^2 (and $R^2 > 0.5$) is given in parentheses.

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

Atmospheric Chemistry and Physics
Open Access
Discussions

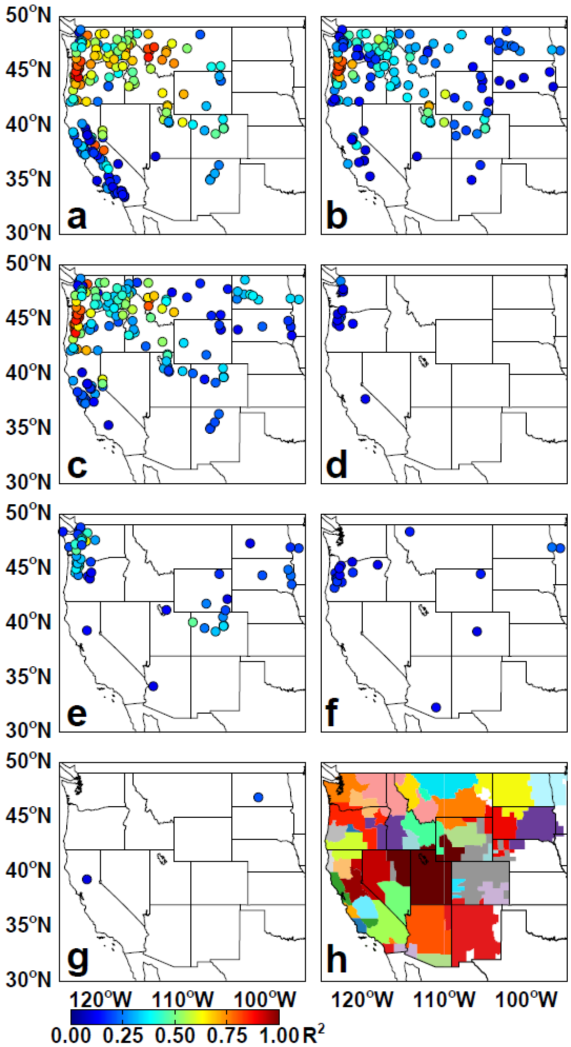

Figure 5. R^2 values at each measurement site for surface measurement and Google Trend search trend (a) "air quality", (b) "wildfire", (c) "smoke", (d) "pollution", (e) "haze", (f) "smog" and (g) "ozone." Only sites where R^2 > 0.1 are shown. The 48 DMAs considered are shown in (h).

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017
Manuscript under review for journal Atmos. Chem. Phys.
Published: 19 January 2017

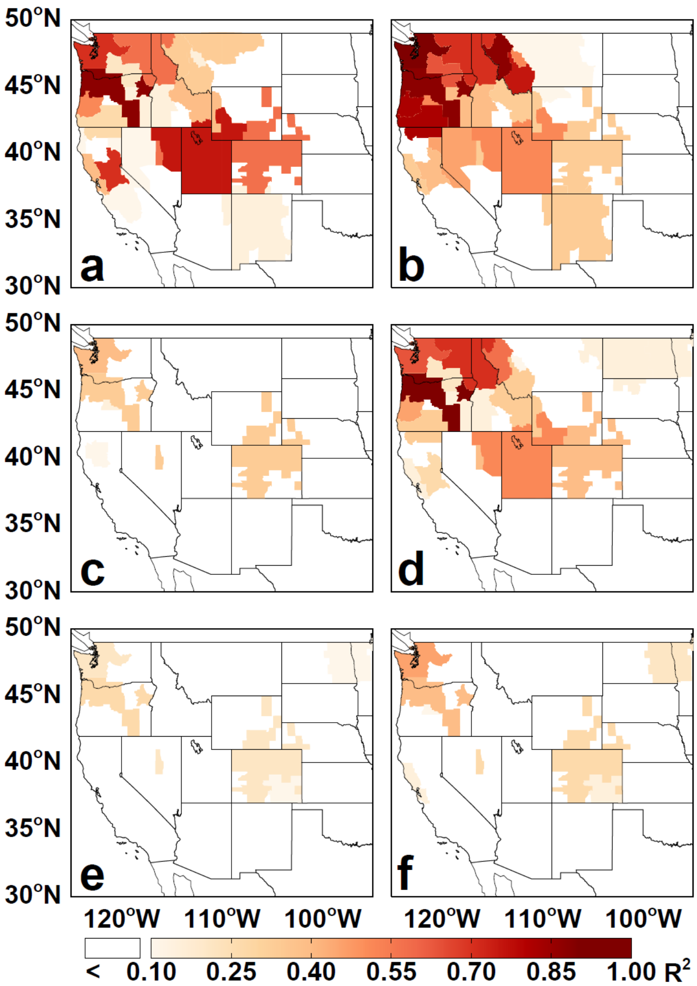

Figure 6. R^2 values for pairs of Google Trends search terms (a) "air quality" and "wildfire", (b) "air quality" and "smoke" (c) "air quality" and "haze", (d) "wildfire" and "smoke", (e) "wildfire" and "haze" and (f) "smoke" and "haze" for June –September 2015.

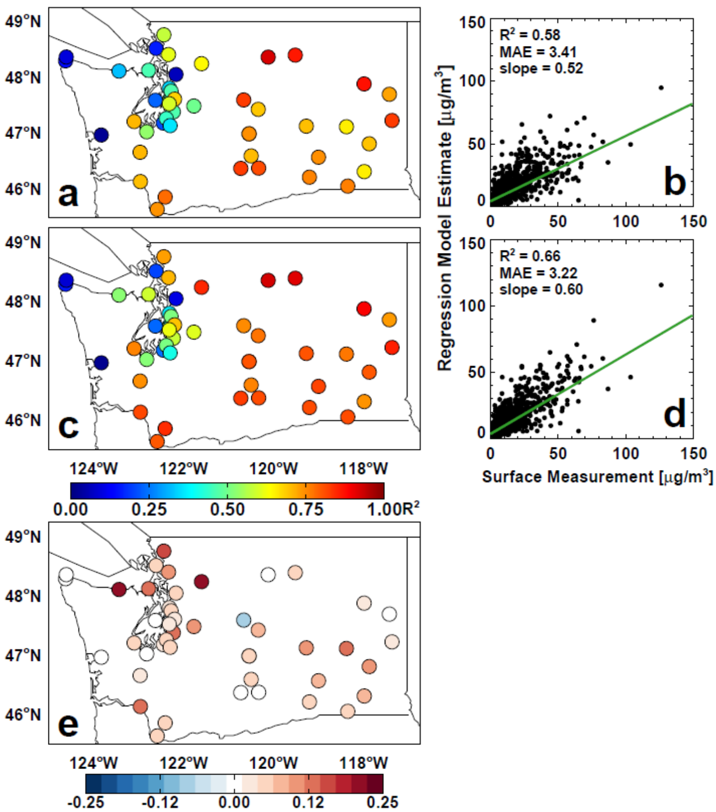

Figure 7. R^2 values at each measurement site for surface $PM_{2.5}$ and regression model estimate (a) using MODIS AOD and WRF-Chem $PM_{2.5}$ and (c) using MODIS AOD, WRF-Chem $PM_{2.5}$, and Facebook dataset for 5 June – 30 September, 2015 and the difference in R^2 between the two regression models (with Facebook data- without Facebook data). Also, scatterplots for all daily measured $PM_{2.5}$ and corresponding regression model estimates in the domain (b) using MODIS AOD and WRF-Chem $PM_{2.5}$ and (d) using MODIS AOD, WRF-Chem $PM_{2.5}$, and Facebook dataset.