# Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale

Brian Nathan, David J Lary

https://doi.org/10.1007/s10661-019-7429-9

# Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale

**Brian J. Nathan · David J. Lary**

Received: 22 March 2017 / Accepted: 20 March 2019


**Abstract** For the period of the Barnett Coordinated Campaign, October 16–31, 2013, hourly concentrations for 46 volatile organic compounds (VOCs) were recorded at 14 air monitoring stations within the Barnett Shale of North Texas. These measurements are used to identify and analyze multi-species hydrocarbon signatures on a regional scale through the novel combination of two techniques: domain filling with Lagrangian trajectories and the machine learning unsupervised classification algorithm called a self-organizing map (SOM). This combination of techniques is shown to accurately identify concentration enhancements in the lightest measured alkane species at and downwind of the locations of active-permit oil and gas facilities, despite the model having no a priori knowledge of these source locations. Site comparisons further identify the SOM's ability to distinguish between signatures with differing influences from oil- and gas-related processes and from urban processes. A random forest (a machine learning supervised classification) analysis is conducted to further probe the sensitivities of the SOM classification in response to changes in any hydrocarbon species' concentration values. The random forest analysis of four representative classes finds that the SOM classification is appropriately more sensitive to changes in certain urban-related species for urban-related classes, and to changes in oil- and gas-related species for oil- and gas-related classes.

**Keywords** Hydrocarbons · Volatile organic compounds · Regional signatures · Self-organizing maps · Machine learning · Environmental health impacts

This article is part of the Topical Collection on *Topical Collection on Geospatial Technology in Environmental Health Applications*

Brian J. Nathan (✉)
Institut Méditerranéen de Biodiversité et d'Ecologie (IMBE), Aix-Marseille Université, Site Arbois, 13290 Aix-en-Provence, France
e-mail: brian.nathan@mio.osupytheas.fr

David J. Lary
William B. Hanson Center for Space Sciences, The University of Texas at Dallas, 800 W. Campbell Road WT-15, Richardson, TX 75080, USA

## Introduction

The Barnett Shale, located in North Central Texas, is one of the most productive natural gas regions in the USA, accounting for 8.3% of the total natural gas production of the USA in 2013 (Railroad Commission of Texas 2014a; United States Energy Information Adminisration (EIA) 2014). Between October 16 and

31, 2013, the Barnett Coordinated Campaign (BCC) was undertaken to help quantify emissions related to these activities (Harriss et al. 2015), with techniques ranging from airplane and UAV measurements (Karion et al. 2015; Lavoie et al. 2015; Nathan et al. 2015; Smith et al. 2015) to mobile laboratory and on-site measurements (Lan et al. 2015; Rella et al. 2015; Townsend-Small et al. 2015; Yacovitch et al. 2015). Most of these studies focused on methane, specifically, which is the most abundant component of natural gas. Other volatile organic compounds (VOCs) are known to be related to oil and gas emissions, as well, especially ethane (Miller et al. 2013; Townsend-Small et al. 2015; Vinciguerra et al. 2015; Yacovitch et al. 2014) and other light alkanes (Pétron et al. 2014; Townsend-Small et al. 2015).

This study investigates a broad range of VOCs ($n = 46$) to better understand the multi-species regional signatures within the Barnett during the period of the BCC. Aside from some VOCs being useful tracers to oil- and gas-related activity, VOCs are also of interest because they are precursors to tropospheric ozone (Atkinson 2000; National Research Council (NRC) 1991) and because they play an important role in the formation of other secondary air pollutants including aldehydes and secondary organic aerosols (Leuchner and Rappenglück 2010).

The regional-scale VOC investigation presented here is achieved through applying a novel two-step analysis approach to ambient mixing ratio measurements. The first step is the use of multiple Lagrangian trajectories through a meteorological model to create a domain filling. This domain filling provides a wider context to in situ measurements made at 14 air monitoring stations in the domain by characterizing where the measured air parcels were in a 4-day window (+/− 48 h). The second step is the implementation of a self-organizing map (SOM), which is an unsupervised classification technique. The SOM objectively classifies and groups together similar records in terms of both wind regimes and multi-species hydrocarbon signatures involving 46 different hydrocarbons (the full list of which can be seen in Fig. 4). Deeper insights into the key distinguishing factors of some representative SOM signatures are then provided using a random forest.

Through the combination of these two methodologies, regional multi-species signatures are created across the Barnett, then classified into groups with similar characteristics. As validation of the accuracy of this approach, the model is shown to properly identify enhancements in light alkane concentrations at and downwind of locations of active-permit oil and gas facilities, compared to the BCC domain average. This is especially noteworthy given that the joint methodology operated with no a priori knowledge of these source locations. Site comparisons further show this approach's ability to accurately distinguish between varying degrees of oil- and gas-related signatures and urban signatures. The veracity of the SOM classification's use in this capacity, in particular, is further verified through a random forest analysis of the SOM's sensitivities for several distinct representative classes.

## Environmental health impacts

A comprehensive characterization of the health effects of hydraulic fracturing is in its infancy (Allen 2014; Field et al. 2014; Gordalla et al. 2013; Hultman et al. 2011; Jackson et al. 2014; Myers 2012; United States Energy Information Administration 2017). Environmental impacts can occur at all stages of the oil and gas supply chain. In the shale gas extraction phase alone, failure of the structural integrity of the well cement and casing, surface spills and leakage from above-ground storage, emissions from gas-processing equipment, and the large numbers of heavy transport vehicles involved are the most important factors that contribute to environmental contamination and exposure. Environmental exposures include air pollutants such as volatile organic compounds, ozone, and diesel particulate matter (Field et al. 2014; Gordalla et al. 2013; Jackson et al. 2014) and pollutants in both ground and surface water (e.g., benzene, hydrocarbons, endocrine-disrupting chemicals, and heavy metals) (Gordalla et al. 2013; Myers 2012). Known occupational hazards include airborne silica exposure at the well pad. Toxicological data for the chemicals injected into wells show that many of them have known adverse effects on health (Gordalla et al. 2013), with no toxicological data available for some. Assessment of potential risks has been difficult because drilling operators in most states are not required to declare which chemicals are used, and only in recent years has the released data been easily accessible (e.g. https://fracfocus.org). In addition to local health and environment threats, an important consideration is the

contribution of shale gas extraction to greenhouse gas emissions and, thus, to climate change. Shale gas has rapidly become an added source of fossil fuel, leading to an increase in cumulative global greenhouse gas emissions (Allen 2014; Hultman et al. 2011). In this first study of its kind, we focus on providing a comprehensive characterization of 46 different hydrocarbons across the Barnett Shale that could be used on a routine basis for characterization of the environmental health effects. Namely, the presented multi-species hydrocarbon classification allows for identification of subtle spatial and temporal changes in local-scale hydrocarbon signatures. This would be of particular use to monitors (perhaps air quality companies or policymakers) with an interest in those changes for a subset of the hydrocarbon species, either for health-related or environmentally related reasons.

## Methodology

There are two main goals of this study: to identify multi-species hydrocarbon signatures on a regional scale and to verify their accuracy. The first is achieved through a domain filling approach, and the latter through the use of a self-organizing map (SOM).

### Domain filling

The first step in achieving a domain filling is to identify the domain which is to be filled. In this study, that domain is the Barnett Shale, the constituent counties of which are defined by the Texas Railroad Commission (Railroad Commission of Texas 2014b). Inside of this domain were situated 14 air monitoring stations owned by the Texas Commission on Environmental Quality (TCEQ), which recorded hourly average concentration measurements of 46 non-methane hydrocarbon species using automated gas chromatographs (auto-GCs) (Texas Commission on Environmental Quality (TCEQ) 2014a).

The filling of this domain is achieved in a manner similar to that originally described in Sutton et al. (1994). This domain filling requires two major steps: simulating multiple Lagrangian trajectories, then interpolating the trajectory data onto a grid which encompasses that domain.

In this analysis, the hybrid single-particle Lagrangian integrated trajectory (HYSPLIT) model (Draxler and Hess 1998) is used to extend the resolution of the average hourly hydrocarbon concentration measurements. Because this is a Lagrangian trajectory model, included species are assumed to be non-reactive during their advection through the domain. For every hour in the desired 2-week period, two trajectories are run from each measurement site: one forwards and one backwards, each for 48 h. One spreadsheet dataset is then created and saved for every hydrocarbon species for every hour. Each spreadsheet contains all of the relevant information for every air parcel whose 4-day window overlaps that time. This information includes the latitude, longitude, altitude, age (the number of hours since the start of the air parcel's simulation time), mixing layer depth, which measurement station the parcel originated from, whether it is a forwards or a backwards trajectory, and the measured concentration value for that hydrocarbon species at the start of the simulation. Effectively, for any given hydrocarbon species, each hourly spreadsheet is a comprehensive snapshot of all air parcels that either had passed through or would pass through an air monitoring station within $\pm$ 48 h of that hour.

The next step to complete the domain filling is to interpolate the trajectories' measured concentrations onto a regularly spaced grid. This data is interpolated onto the same grid that the TCEQ use for their air quality modeling domains: the comprehensive air-quality model with extensions (CAMx) at 4-km resolution as projected onto a Lambert conformal cone, with the specifications that match the National Regional Planning Organization's (RPO's) domains (Texas Commission on Environmental Quality (TCEQ) 2014b). This grid extends well beyond the Barnett's boundaries, but the superfluous grid cells are ignored and have no impact on the rest of the analysis.

Each hourly dataset is binned and averaged into the grid cells of the CAMx regular grid, where any air parcels whose altitude exceeds the mixing layer height are excluded from the analysis. One average concentration grid is thus created for every hour of the BCC for each species.

It should be noted that the domain filling technique described here is very close, in principle, to potential source contribution functions (PSCFs). First developed by Ashbaugh et al. (1985), PSCFs are receptor models that use back trajectories to create spatial probability distributions (with respect to source locations)

for a desired species, based on the fraction of total measured air parcels in a grid cell that had been measured as above some threshold criteria (typically the mean of measurements) (Hopke 2003). PSCFs have been used in single-species analyses for source location and preferred transport pathways (e.g., Cheng et al. 1993 and Polissar et al. 1999), as well as in multivariate studies (e.g., Poirot et al. 2001, Xie et al. 1999, and Xie and Berkowitz 2007).

Unlike with PSCFs, however, the measured concentration values in this study remain tied to each individual air parcel which contributes to the average signature of a grid cell, rather than this signature being an indication of the probability of "high" measurements based on some concentration threshold criterion. In this way, subtle changes in concentrations of species which have relationships with one or more other measured species (e.g., subtle changes in the ethane/propane ratio, indicating different degrees of "wetness" of the oil and gas emissions) have the best chance of being detected and exploited by the self-organizing map (SOM) classification, as detailed later.

The domain-filled grids created as stated here can have their validity tested through analyzing for well-known species or ratios of species to determine if the expected spatial patterns are observed. This approach is used as an initial analysis step, although the large number of hydrocarbon species, and the larger number of potential ratios, makes an exhaustive analysis of this sort prohibitive. A machine learning classification, such as a self-organizing map, is instead an ideal approach for this high-dimensional comparison.

## Classification using self-organizing maps

Self-organizing maps have been known for decades to be an effective tool for detecting and separating out patterns from even very noisy input signals (Kohonen 1990). They are such a robust tool for identifying structures in high-dimensional datasets that by 1998 (8 years after their first use), a literature survey by Kaski et al. (1998) found over 3300 papers either analyzing it or using it for tasks in fields ranging from the financial to the speech pathological to the physical. Included in their survey, Mangiameli et al. (1996) compared the SOM against seven heirarchical clustering methods for their ability to classify "messy" data (which is more apt for real-world, empirical data). They found that, not only is an SOM a robust method for the classification, but it is superior to all seven heirarchical methods, where the "messier" the data got (indicating the presence of wide data dispersion, outliers, irrelevant information, etc.), the more "dominant" the self-organizing map's performance was.

SOM analyses have been used for atmospheric classification of synoptic patterns (e.g., Huth et al. 2008), towards diagnosing air quality events (e.g., Pearce et al. 2011), and for various analyses of VOCs. VOC SOM analyses have included *Calliphora* age estimation (Moore et al. 2016), analysis of oil spill emissions (Fernández-Varela et al. 2010), aiding in sensor detection of petroleum at low ppb (Sugimoto et al. 1999) or of VOCs in sea water (Tonacci et al. 2015), for validation of tropospheric VOC degradation models (Papa and Gramatica 2008) and even in an urban benzene spatial map analysis (Strebel et al. 2013). We believe that we are presenting the SOM's first use combining a meteorological model with a multi-species VOC spatial assessment.

By their nature, self-organizing maps are a dimension reduction tool. An SOM classifies high-dimensional, multivariate datasets by grouping like signatures together, and it outputs a low-dimensional (here, one-dimensional) list of classes which has been organized by the dataset's similarities.

For the study presented here, a self-organizing map is used for two separate parts of the analysis. It is used to classify the mean hydrocarbon concentration signatures for the dataset containing all grid cell signatures from all hourly grids. It is also used to classify the wind conditions at each of the measurement stations during the course of the 2-week campaign.

The SOM for wind conditions is run on a dataset containing hourly average wind speed and wind direction measurements as recorded and reported at each of the 14 air monitoring stations. By grouping like with like, in this case, the SOM organizes the hours of the 2-week measurement period into groups (classes) of nonconsecutive hours that had similar wind speed and wind direction conditions. This affords the opportunity to create more robust domain-filled grids that are representative ensemble averages of all the hourly grids which fit into a desired wind class (so an average of all grids created under similar wind conditions).

The hourly mean hydrocarbon concentration grids are classified in a similar manner. Rather than looking at all 46 domain-filled grids for a particular hour,

it would be preferable to look at one grid that has information about all 46 hydrocarbon species. The SOM classification allows this. Before classification, one could imagine there to be one "master" grid per hour, where each filled grid cell has a signature array of 46 hydrocarbon concentration values. After SOM classification, each grid cell instead is assigned a one-dimensional class number based on these signatures. For this study, the actual classification was performed on a dataset containing every grid cell signature from every hour's master grid. It is important to note here that no a priori knowledge about source locations are included in this SOM hydrocarbon classification; only the concentration values for all grid cells are being sorted.

The wind SOM and the hydrocarbon SOM contain 6 classes and 200 classes, respectively. Note that these are referring to the sizes of our output layer (detector grid), which we chose to be one-dimensional to facilitate analysis of the differences between these classes. These numbers were determined in a trial-and-error fashion. In any search for the appropriate number of classes, a number is sought that has adjacent classes neither being overly similar to each other, which would indicate a need for fewer classes, nor overly different from each other, which would imply a need for more classes (Vesanto and Alhoniemi 2000).

In both classification cases, a simple general approach is used to implement the SOM:

1. A large dataset is created with every measurement being used, where each column corresponds to a different variable. For the wind data, the first 14 columns are the wind speed measurements from each TCEQ station, the second 14 columns are the corresponding wind direction measurements, and every row is an hourly measurement during the BCC. For the mean hydrocarbon concentration maps, there is one column for each of the 46 species, and one row for each non-empty grid cell from every hourly grid (172,618, in total).
2. A self-organizing map (SOM) network is created with a designated number of classes, using built-in functions of MATLAB.
3. The SOM network is trained with the corresponding dataset for the default 200 iterations.

As stated, the SOM is created and trained for this analysis using the built-in functions in the MATLAB toolbox, as described at length in Vesanto et al. (1999). Some of the default values include learning rates of 0.9 and 0.02 for the ordering phase and tuning phase, respectively, a neighborhood distance set at 1, with a hexagonal topology. Future investigations may be interested in adjusting these parameters to see if there is a noticeable improvement in results. The deformation of the self-organizing map during the training process uses competitive learning following the Kohonen rule (Kohonen 1990).

After the SOM network has been trained, it acts as a mapping function, with the ability to take any signature measurement relating to the relevant dataset and output the corresponding class number to which that signature belongs. For this analysis, the wind data SOM is used first to divide the hours of the measurement campaign according to their wind characteristics. To avoid confusion, these wind classes are hereafter referred to as "wind regimes." Then, a hydrocarbon grid is constructed which is the mean of all hourly hydrocarbon grids for every hour in a given wind regime. This mean hydrocarbon grid is the master grid for that wind regime, and each filled grid cell has a corresponding 46-dimensional signature. The master grid for the wind regime is simulated through the hydrocarbon SOM network to produce one grid filled with appropriate class numbers representative of their hydrocarbon signatures. Further details are provided with the results in the "Results and discussion" section.

The accuracy of the hydrocarbon classification is verified in a multi-stage process, since there are no comparable regional-scale multi-species estimates to perform a validation against. The first test is through looking at the characteristics of the class number most frequently assigned to grid cells containing a known strong source emitter of particular hydrocarbon species, to see if expected enhancements in the appropriate hydrocarbons are observed. The second test is through a site comparison of selected representative air monitoring stations. Finally, a random forest is performed on the hydrocarbon classification to better understand the sensitivity of the SOM classification to changes in less abundant species.

The strong source emitters chosen for the first validation test of the SOM are oil and gas facilities that had active operating permits during the period of interest. The emissions for any such facility are known to have increased concentrations of the lightest alkanes. A coordinate list for all oil and gas

facilities with active operating permits was obtained from the TCEQ, and the class numbers assigned to these spots are scrutinized to see if enhancements in light alkane concentrations are detected. We note that this coordinate list included 427 active-permit oil and gas facilities within the Barnett, 377 of which are classified as "Oil and Gas Production Facilities" (106.352) (Texas Commission on Environmental Quality (TCEQ) 2012, 2017) with the remainder being a combination of "Non-Rule Oil and Gas Facilities" (6002), "Oil and Gas Facilities" (6002-116.620), and "Salt Water Disposal (Petroleum)" (106.351), as defined by the TCEQ (Texas Commission on Environmental Quality (TCEQ) 2017).

For the site comparison (the second test), 4 of the 14 TCEQ measurement stations are singled out to verify the accuracy of this methodology. The Eagle Mountain Lake, Flower Mound Shiloh, and Dallas Hinton stations are selected to be consistent with the analysis done in Zavala-Araiza et al. (2014), which focused on data from April 19, 2010, through December 31, 2011. Additionally, the Godley FM2331 station is included here, because of its proximity to the largest number of oil and gas facilities. For brevity, these stations are hereafter referred to as the EML, FMS, DH, and GLY stations, respectively. Figure 1 shows the relative positions of the TCEQ stations and oil and gas facilities in the Barnett. For the site comparisons, as with the first test, attention is paid only to the light alkane species (for now) to see if the expected enhancements or lack of enhancements are observed when these stations are or are not downwind of the aforementioned facilities.

As a deeper analysis of the full hydrocarbon classification, which looks beyond just the light alkane species, a random forest is conducted. First developed

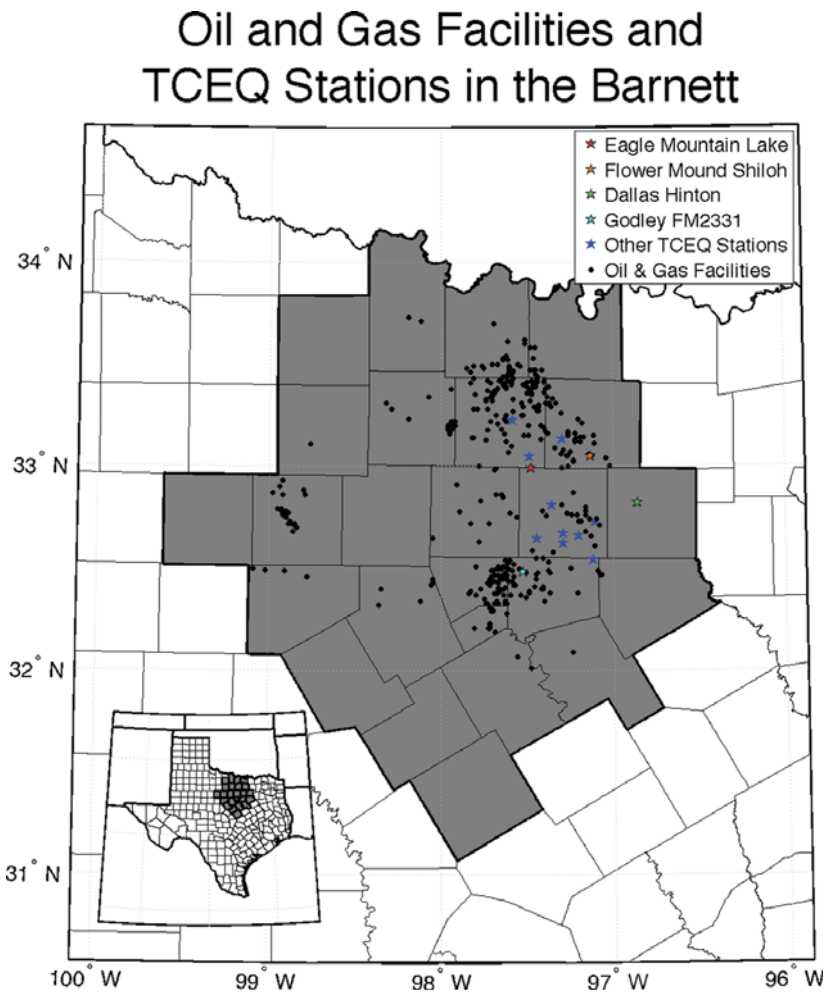


**Fig. 1** A map showing the relative positions of the active-permit oil and gas facilities in the Barnett and the TCEQ air monitoring stations. The four stations that were singled out for site comparison, EML, FMS, DH, and GLY, are individually color-coded. The thick black line outlines the gray-shaded counties that the Texas Railroad Commission designates as defining the Barnett

by Breiman (2001), random forest is a robust technique which uses an ensemble learning with decision trees. The specific ensemble learning technique used in the random forest is called "bagging" or "bootstrap aggregation" (Breiman 1996a). Bagging decision trees describes the process of creating a multitude of trees (a forest) from random subsets of the full dataset. The best predictor is then determined by aggregating the "votes" of each tree for how to split each node. Random forests take this one step beyond what happens with bagged trees by changing what happens at the nodes. Rather than choosing what happens at a given node based on what is best for all predictor variables, a random forest chooses based on what is best for some random subset of those predictor variables (Liaw and Wiener 2002). This technique is robust enough to avoid the problem of overfitting that single decision trees have (that is, where the trees end up fitting random signals or noise, too), and it has performed well compared to discriminant analysis, support vector machines, and neural network classifiers (Breiman 2001).

Random forests can be used to develop a quantification for the relative importances of predictor variables in a classification. This is possible since the bootstrapping leaves approximately one-third of the data "out-of-bag," on average, which allows this additional data to be used to gauge the accuracy of predictions (or, inversely, the prediction error) (Breiman 1996b). Thus, the knowledge of the accuracy of predictions can lead to an assessment of the relative importance of the descriptors. For this study, this is used to determine the relative importances of all 46 hydrocarbons in the definitions of some of the key classes presented. That is to say that, since the analysis looking at just the light alkane species generally identifies class numbers associated with strong or moderate oil- and gas-related emissions or with urban signatures, these assessments can be corroborated through this random forest investigation, which will also shed light on previously unseen nuances in the classification.

The quantity used to describe the "relative importance" is unitless, and is derived as follows:

1. For a given hydrocarbon species, its concentration values are permuted across any trees in the forest for which this species is out-of-bag.
2. The change in prediction error is calculated for every tree.
3. The average of the change-in-prediction-error calculations is computed across the entire ensemble.
4. This average is divided by the standard deviation of the prediction errors over the entire ensemble.

This is an accepted technique, the beginning analytical ideas for which were outlined in Breiman (2001), and which is now a built-in function of MATLAB called "OOBPermutedVarDeltaError".

## Results and discussion

### Wind classification

Following the procedure outlined in the "Methodology" section, the average wind speed and direction measurements for all TCEQ air monitoring stations are classified by a self-organizing map. In this way, the 2-week period of the BCC is divided into wind regimes of similar conditions. The output wind regime definitions are shown in Table 1, as well as the values for the total BCC. Of the 360 total hours of the campaign, 348 have valid measurements at all 14 active stations, and these were the hours that were sorted by the SOM for the wind classification.

The remainder of this analysis will focus on the hours of wind regime 2. Wind regime 2 was found to best demonstrate the validity of the domain filling and self-organizing map classification, because under these conditions, nearly all air monitoring stations are downwind of active-permit oil and gas facilities.

**Table 1** The average wind speed, average wind direction, and total number of hours contained in each wind regime

| Wind regime | Wind speed (m/s) | Wind direction (deg) | Hours contained |
|---|---|---|---|
| 1 | 3.2 ± 0.8 | 333.2 ± 8.0 (NNW) | 33 |
| 2 | 2.0 ± 0.5 | 301.2 ± 7.4 (WNW) | 44 |
| 3 | 1.9 ± 0.5 | 270.7 ± 35.2 (W) | 45 |
| 4 | 3.4 ± 1.0 | 170.6 ± 8.5 (S) | 111 |
| 5 | 2.7 ± 0.7 | 131.6 ± 10.8 (SE) | 78 |
| 6 | 3.1 ± 0.8 | 36.2 ± 5.1 (NE) | 37 |
| Total BCC | 2.8 ± 0.7 | 159.7 ± 17.6 (SSE) | 348 |

The corresponding values for all hours of the BCC with valid data are also included. The included standard deviations are for the mean values of all stations

Wind regime 2

As shown in Table 1, wind regime 2 is characterized by a mean wind speed of 2.0 $\pm$ 0.5 $\frac{m}{s}$ and a mean wind direction of 301.2 $\pm$ 7.4 degrees (WNW). The representative domain-filled grid for this wind regime is a mean of all of the hourly domain-filled grids for the hours that comprise this wind regime.

*Domain filling*

Since the domain filling procedure achieved the goal of creating regional-scale spatial estimates of hydrocarbon concentrations, the relative accuracy of these estimates can now be tested. For the hours of wind regime 2, a representative domain-filled grid is created by taking the mean of each hourly domain-filled grid which comprise the wind regime.

The accuracy of this representative grid is tested by looking for expected signatures from large sources. The light alkane species are universally the most abundant species by concentration of all of the 46 measured hydrocarbons. During this wind regime, across the whole Barnett grid, ethane averages 28.7 ppbv, propane averages 12.1 ppbv, and *n*-butane averages 4.6 ppbv. These mean concentrations account for 52.4%, 22.1%, and 8.5% of the measured non-methane hydrocarbon concentrations, respectively, making up 83.0% of the total. Since the light alkane species are all tracers for oil- and gas-related activity, it would be expected to see spikes in their concentrations at and downwind of active-permit facilities.

The leftmost map in Fig. 2 shows the log10 spatial ethane estimates for wind regime 2. The circles labeled "A" and "B" indicate where the two largest general clusters of oil and gas facilities are located. The shaded region is included solely to aid the reader's comprehension of the figures by delineating the area of the Barnett for which the authors hold little or no confidence in the domain-filled values. Although it is placed over areas deemed to contain too few trajectory points or to be too far upwind from a measurement station, these quantities are not rigorously defined. It is believed that future studies may be able to define a reasonable uncertainty threshold based on both the local meteorology and the model statistics, though this is outside the scope of this present study.

In the log10 ethane plot of Fig. 2, clear enhancements are observed at and downwind of the two central clusters of oil and gas facilities ("A" and "B").

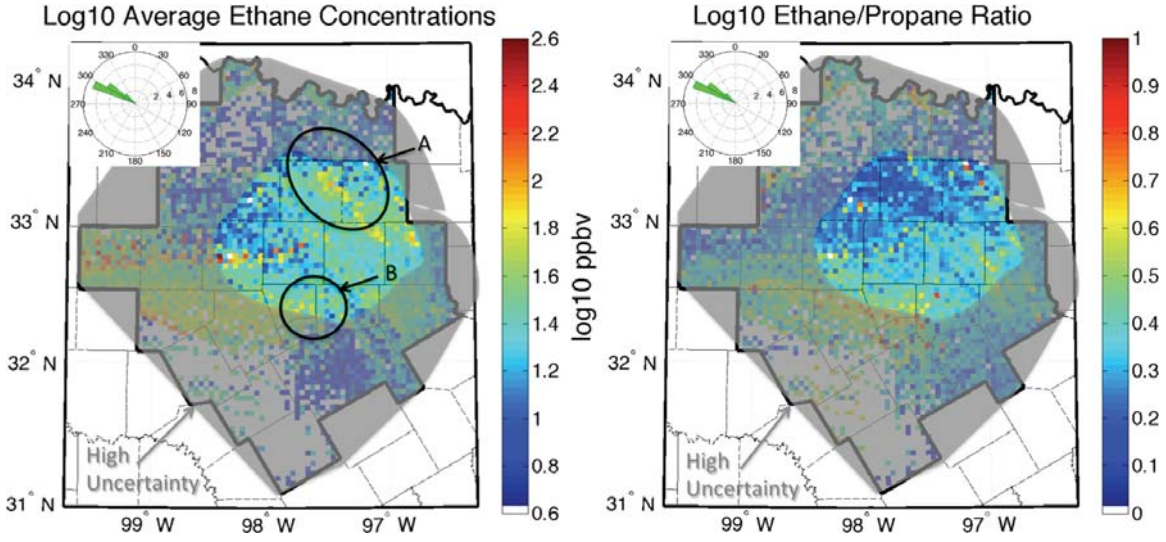


**Fig. 2** Two domain-filled grids are shown, created from the ensemble average of the hours of wind regime 2. On the left are the log10 mean ethane concentration estimates, and on the right are the log10 ethane/propane volume ratio estimates. "A" and "B" indicate the relative positions of clusters of active-permit oil and gas facilities. The shaded region is included to aid the reader's comprehension of the figures, and is a qualitative estimate of the region where the domain filling values become untrustworthy based loosely on having a low number of trajectory points or being a large distance from a measurement station. It is believed that future studies may be able to quantitatively define an uncertainty threshold based on both the local meteorology and the model statistics, though this is outside the scope of this study

Only the southern portion of the northernmost cluster ("A") and the northern portion of the southernmost cluster ("B") have stations downwind to sufficiently characterize them. These enhancements serve as validation of the domain-filling's accuracy and showcase its ability to identify hydrocarbon emissions from sources on a regional scale.

Beyond the domain filling's ability to look at any particular hydrocarbon's regional-scale spatial estimates, it also affords the opportunity to analyze the spatial estimates for any meaningful ratios. On the right side of Fig. 2 is the log10 ethane/propane spatial plot. This ratio is an indicator of "wetter" gas regions versus "drier" gas regions. "Wet" gas refers to gas which is less mature, and thus has a higher proportion of heavier alkane elements. (When the heavier alkanes reach the surface, they are still in liquid form, thus making things "wet.") Conversely, "dry" gas describes gas which has a higher relative proportion of the lighter alkane elements. Thus, the ethane/propane ratio will be smaller for wetter regions of the shale and higher for drier regions.

For the Barnett, it would be expected to see a general wet-to-dry gradient moving through the central portion of the Barnett from the west and northwest to the east and southeast (Montgomery et al. 2005). Although this is generally what is seen in the right panel of Fig. 2, the gradient does appear to be shifted more north to south for these hours of wind regime 2. This is a joint consequence both of the mean wind direction for this regime being out of the west-northwest (so the air parcels filling in the domain were traveling virtually horizontally) and of the lack of measurement stations in the western portion of the Barnett (especially the south-western portion). Although the east/west resolution of the domain filling is lacking for this wind regime, it is still able to properly account for the north-to-south component of the wet-to-dry gradient. This gives more confidence in the domain filling's spatial estimates, while leaving room for potential improvements, as discussed further in the "Conclusions" section.

Overall, the domain filling results allow for a standard-style analysis, which would involve looking at spatial distributions for concentrations of one species of interest (such as ethane) or a ratio of interest between two important species (such as ethane/propane). The achieved domain filling presented here still permits this, and the analysis of Fig. 2 shows that it yields reasonable results. However, these approaches are insufficient to compare the full 46-dimensional signature. These approaches may result in more subtle or unanticipated relationships between hydrocarbon species going unnoticed, especially if the relationships are nonlinear. To overcome this limitation, a self-organizing map is implemented in this study.

*Classification of hydrocarbon signatures*

As described in the "Methodology" section, the hydrocarbon self-organizing map is trained using every hydrocarbon signature from every grid cell of every grid in the 2-week period of interest. Each hydrocarbon signature is the mean of the concentration values for any air parcel that was in that hour-grid's grid cell (and under the top of the mixing layer).

For wind regime 2, the representative domain-filled grid was mapped through the trained hydrocarbon self-organizing map to yield class numbers for each filled grid cell. Since the SOM groups like with like, it would be expected that similar hydrocarbon signatures would be grouped into either the same or adjacent classes. Thus, since oil and gas facilities emit strong, similar signatures, one way to test the accuracy of the classification is to compare the classes assigned to the grid cells that house these facilities.

Figure 3 shows the classification numbers assigned to the grid cells that house active-permit oil and gas facilities for the hours of wind regime 2. An immediate take-away from the figure is the predominance of assigned classes from the blue portion of the colorscale. Since, by nature of the self-organizing map, similar class numbers should indicate similar signatures, this is strong evidence that the SOM accurately identified similar emissions from the active-permit facilities.

The fact that these grid cells all have similar signatures does not guarantee on its own that they are proper expected signatures for strong oil- and gas-related emissions. This accuracy can be investigated through an analysis of the class definitions. Each class number is defined by its own unique hydrocarbon signature, which is comprised of the mean of the concentration signatures for each grid cell that was assigned to that class.

The class number assigned most frequently to oil and gas facilities during wind regime 2 is 49. If the

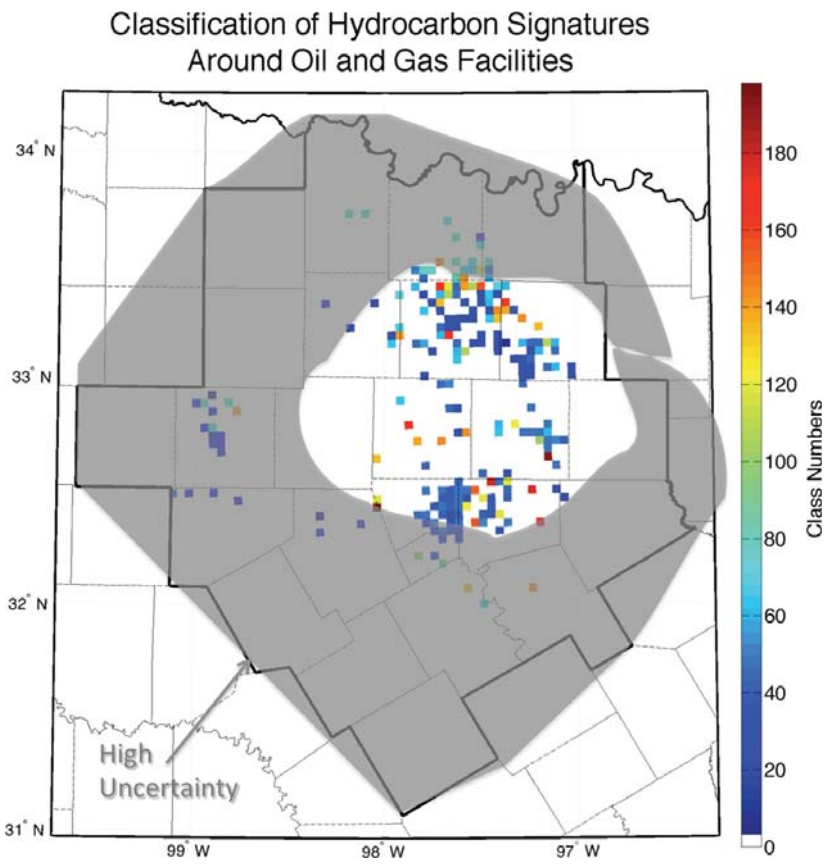


**Fig. 3** The hydrocarbon classification grid for the grid cells that house active-permit oil and gas facilities during the hours of wind regime 2. The predominance of blue indicates that these grid cells were classified into similar classes, indicating similar hydrocarbon signatures. No a priori knowledge of source locations was given to the SOM during the classification process

SOM properly captured emissions from these facilities, it would be expected that the light alkane concentrations associated with class 49 are higher than the average values across the Barnett for the duration of the BCC. As are expressed in Table 2, class 49 is associated with 24.1 ppbv of ethane, 9.8 ppbv of propane, and 3.7 ppbv of *n*-butane. These are each indeed higher than the average values across the Barnett for the whole BCC of 22.2 ppbv of ethane, 9.1 ppbv of propane, and 3.5 ppbv of *n*-butane.

Further validation of the classification can be achieved through site comparisons. As outlined in the "Methodology" section, four of the 14 TCEQ air monitoring stations are singled out for this. Zavala-Araiza et al. (2014) had chosen the Eagle Mountain Lake (EML) station because it is located in the approximate middle of the production zone. They had chosen the Flower Mound Shiloh (FMS) station to be the suburban signature. And they had chosen the Dallas Hinton (DH) station to be the urban signature. For consistency, these stations are also included in this site comparison, with the addition of the Godley FM2331 (GLY) station, because of its proximity to a large cluster of oil and gas facilities.

It would be expected that enhancements in the light alkane concentrations would occur at these stations when they are downwind of the oil and gas facilities. Indeed, for the hours of wind regime 2, the classes assigned to the grid cells housing these stations do show light alkane enhancements for all four stations when compared to their class values for the total BCC. This is also shown in Table 2.

The hydrocarbon concentrations for the classes assigned to the grid cells around these stations also align with expectations regarding these stations' differences. The DH station would be expected to have the weakest light alkane enhancements, because it is the designated urban station and is farthest from any active-permit oil and gas facilities. Conversely, the GLY station during wind regime 2 should be assigned a class with very strong light alkane enhancements, given its position in relation to the southernmost cluster of oil and gas facilities. Both of these expectations are met by the SOM classification.

The relative proximity of the class numbers being assigned to the different stations in wind regime 2

**Table 2** A comparison of classes is presented through looking at light alkane concentrations to identify oil-and-gas-emissions-related enhancements

| Class number | Assigned to (period) | Ethane concentration (ppbv) | Propane concentration (ppbv) | *n*-Butane concentration (ppbv) |
|---|---|---|---|---|
| N/A | Whole Barnett (total BCC) | 22.2 | 9.1 | 3.5 |
| 49 | O&G Facilities (WR 2) | 24.1 | 9.8 | 3.7 |
| 156 | EML (total BCC) | 15.4 | 6.5 | 2.3 |
| 47 | EML (WR 2) | 29.2 | 11.0 | 3.9 |
| 117 | FMS (total BCC) | 18.5 | 7.5 | 2.4 |
| 44 | FMS (WR 2) | 29.6 | 13.6 | 5.5 |
| 137 | DH (total BCC) | 11.1 | 6.2 | 2.8 |
| 121 | DH (WR 2) | 20.4 | 9.9 | 4.0 |
| 118 | GLY (total BCC) | 20.6 | 7.9 | 2.7 |
| 38 | GLY (WR 2) | 38.1 | 14.9 | 5.3 |
| 48 | O&G Facilities (total BCC) | 25.9 | 10.9 | 4.2 |
| 83 | DH (WR 4) | 8.7 | 4.5 | 2.5 |

The Barnett averages for the total BCC are included as a point of reference. Relevant classes for comparing the mean oil- and gas-related facilities' signature and relevant station signatures for site comparisons in different wind regimes are also included

should be taken note of. In addition to class 49 being assigned most frequently to oil and gas facilities, class numbers 47, 44, and 38 are assigned to the EML, FMS, and GLY stations, respectively. Based on the associated concentrations for the lightest alkanes, these all are apparently strong signatures for oil- and gas-related emissions. It is no coincidence that these class numbers are so similar. The SOM does not just group "like with like" within the same class, but it does so in an incremental fashion so that neighboring classes are most similar to each other. In fact, it is clearly seen that the light alkane concentrations apparently get stronger as the classes decrease from the high 40s down through the high 30s. Looking back at Fig. 3, these classes all seem to fall in the blue portion of the colorscale, which was previously identified as predominantly associated with the active-permit oil and gas facilities. Thus, it can now be said that the SOM has performed well in accurately classifying these signatures in this wind regime.

Besides the limitations of the model, the variations between classes with similar signatures (i.e., the reason these signatures are assigned to nearby classes instead of identical classes) may be a result of fluctuations in the concentrations of hydrocarbons not related to oil and gas activities, since all 46 hydrocarbons were involved in the classification. This demands a deeper investigation of the SOM classification.

*Using a random forest to probe the classification*

As described in the "Methodology" section, a random forest analysis allows for a quantitative assessment of the relative importances for variables in a classification. In the case of the hydrocarbon classification, this is used to determine which species are the most important in defining some particular class. The metric for this measurement, as defined above, may appropriately be thought of as the sensitivity any given class has to (concentration) value changes in any given predictor (species) variable.

The analysis of variable importances for four representative classes is presented here: classes 47, 48, 121, and 83. The average concentrations for the lightest measured alkanes associated with these four classes are included in Table 2. Class 47 was already identified as having been assigned to the grid cell housing the EML station during wind regime 2 and is considered as having a strong oil- and gas-related signature. Class 48 was assigned most frequently to the average signatures of the oil and gas facilities for the entire BCC and is also considered to be a strong oil- and gas-related signature. Class 121 was assigned to the grid cell housing the Dallas Hinton station during wind regime 2. The Dallas Hinton station was considered to be the urban signature station, but this wind regime allowed for some enhancements in the light alkane concentrations from upwind oil and gas facilities. Thus, this class is expected to have a mixed urban/oil and gas signature. Although this investigation focuses on wind regime 2, a purer urban signal is seen at the Dallas Hinton station during wind regime 4. This is because, for wind regime 4, the wind is blowing from the south, as seen in Table 1, and there are no oil and gas facilities upwind. We are therefore including the associated class number, 83, in this presented random forest analysis, to fully demonstrate its ability to distinguish between urban, oil- and gas-dominated, and mixed hydrocarbon signatures in the classification. Figure 4 shows the relative variable importances for classes 47, 48, 121, and 83.

For classes 47 and 48, the top three most important species are identical, as would be expected given the incremental nature of the SOM. On the surface, it also serves as a confidence boost in the SOM to see that two of these three species are ethane and propane. These would be expected here, because they are the two most abundant measured light alkane species, and these classes were previously identified as having strong oil- and gas-related signatures. However, of the two, only the relative importance position of propane is meaningful here. Ethane's substantially larger concentration values over any other measured species led it to be the most important variable in all but three classes in the entire hydrocarbon classification.

For class 121, propane has dropped to being the ninth most important variable associated with that class. Since class 121 is expected to showcase a signature that has both urban and oil and gas influences, this appears to align with expectations, pending investigation to the potential sources of some of the hydrocarbons which usurped it, such as the trimethylbenzenes, acetylene, and isoprene.

Trimethylbenzenes naturally occur in petroleum deposits, and so most human exposure comes from gasoline or mixed aromatic hydrocarbon solvents (Jones et al. 2006). For 1,2,4-trimethylbenzene, in particular, an estimated 32 billion pounds is produced annually from oil refineries as part of their distillation fraction, and most of this is usually eventually added to gasoline (United States Environmental Protection Agency 1994). These are, then, potential urban influences, especially 1,2,4-trimethylbenzene. It should be noted, though, that 1,2,3-trimethylbenzene was determined to be important in the strong oil-and-gas-signature classes, as well.

Isoprene is also shown to be a more important descriptor for class 121 than the measured non-ethane light alkanes. Over 90% of isoprene emissions are believed to come from plant foliage, with the remaining sources being microbes, animals, and aquatic organisms (Guenther et al. 2006). Isoprene is the most important biogenic nonmethane hydrocarbon, though its emissions have been difficult to quantify, because it is a very reactive compound which has a short atmospheric lifetime on the order of minutes to hours (Guenther 1999). Because of isoprene's near-ubiquity and its volatility, it would make sense that the SOM would be sensitive to its fluctuations in nearly all classes. Indeed, when averaging the relative importances across all classes, isoprene ranks fourth overall, behind only ethane, propane, and 1,2,3-trimethylbenzene, respectively. Thus, further evidence is gathered that the SOM was accounting for many potential sources, including biogenic ones, during the classification. This helps explain some of

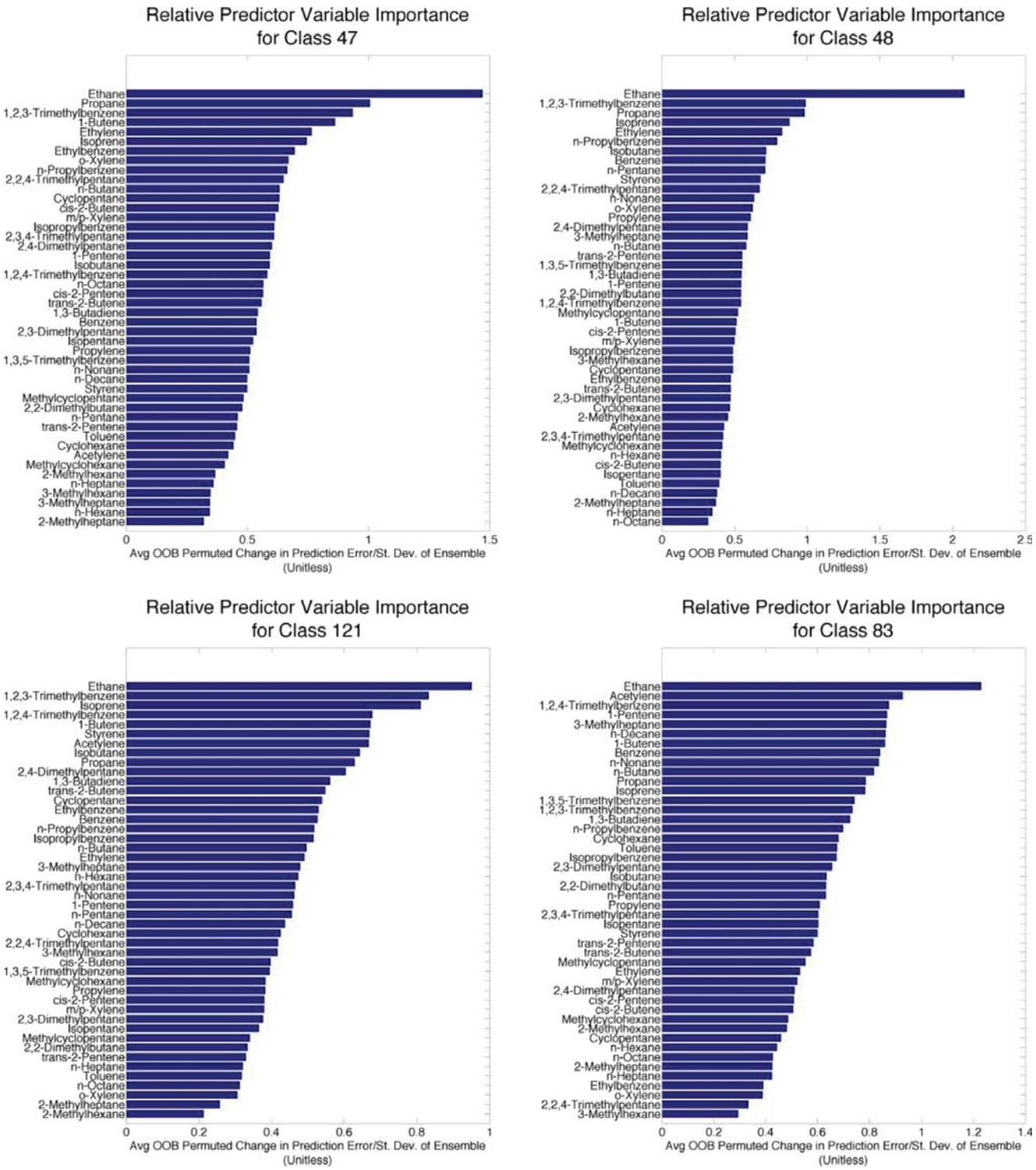


**Fig. 4** The relative variable importances for four classes of interest: classes 47, 48, 121, and 83. These can also be thought of as the sensitivity the classification has to changes in these variables. Classes 47 and 48 are adjacent classes with strong oil- and gas-related signatures, class 121 is a class with both oil- and gas-related and urban contributions, and class 83 is a more pure urban signature

the variations between classes with dominant signatures (such as those with dominant oil-and-gas-related source signatures).

Acetylene may be ranked above all measured light alkanes besides ethane in class 121 for the same reasons as the trimethylbenzenes. Acetylene, also known

as ethyne, occurs in the atmosphere as a byproduct of the combustion of fossil fuels, biofuels, or biomass (Xiao et al. 2007). It has been accepted for decades that the overwhelmingly dominant source, specifically, is gasoline combustion from automobiles (Whitby and Altwicker 1978). Thus, acetylene is also an excellent urban tracer, and its relative importance rank through this random forest analysis is consistent with what would be expected, giving further credence to the SOM classification.

Overall, the variable importances for class 121 do align with the expected blend of source influences. Propane is still high on the list, though no longer in the top 3, and some urban tracers have gotten higher up instead. It is expected that this trend will be strengthened in a purer urban signature class, such as class 83.

Figure 4 also includes the variable importances for class 83. As stated, this is expected to be a strong urban signature, because it corresponds to the signature for the Dallas Hinton station during wind regime 4. Wind regime 4 had winds blowing out of the south (as seen in Table 1), so there are no upwind oil and gas facilities of this station.

Indeed, the variable importances confirm what was expected for class 83. Propane has dropped farther down (to eleventh) compared to the strong oil- and gas-related classes or the mixed class. Acetylene and 1,2,4-trimethylbenzene, already detailed as strong urban tracers, are now second and third on the list of importance, respectively. The fourth most important variable for class 83 is 1-pentene, which is a byproduct from petroleum or other hydrocarbon fractions that is most often blended into gasoline (Jones and Pujadó 2006). As a result, almost all of its atmospheric abundance can be traced to automobiles, diesel, or gasoline vapors (Graedel 2012). This too, then, is a good tracer for urban-related emissions. Having the top three most important non-ethane variables for class 83 be strong urban tracers like this gives even further evidence to the reliability of the original SOM hydrocarbon classification.

## Conclusions

A novel joint methodology is presented. The two-step approach of a domain filling and a self-organizing map classification is shown to be an accurate technique for regional-scale multi-species characterization of hydrocarbons.

The accuracy is demonstrated through the identification of expected enhancement signatures from oil and gas facilities near and downwind of those facilities despite having no a priori knowledge of source locations. Its veracity is further confirmed through site comparisons, where the self-organizing map classification is shown to accurately distinguish subtle differences between areas with varying degrees of urban and oil- and gas-related signatures.

The subtleties inherent in the SOM classification are explored by a random forest analysis. Gasoline-related tracers, urban tracers, and biogenic tracers are all shown to be influential in the separation between class numbers associated with strong oil- and gas-related signatures and those associated with mixed or urban emissions. Most importantly, for the classes with expected strong sources, the corresponding expected tracers were identified as having high importance. This gives strong confidence in the accuracy of the SOM and paves the way for its use in identifying source signatures in future multi-species studies, as well as for its use in monitoring by regulators and policymakers.

Building off of the encouraging results, future studies may look to improve the accuracy of this model. Some ideas include extending the spatial or temporal coverage of the data, factoring in diffusion/dispersion, and running multiple classifications and using the mean, for robustness. From the spatial coverage standpoint, increasing the number of measurement stations would yield obvious and immediate improvements to the model's accuracy. To this end, a quantitative assessment of the spatial sensitivity of stations-versus-sources would be worthy of investigation (e.g., a variogram analysis with real data or an observation system simulation experiment (OSSE) approach in the theoretical framework). Similarly, greater temporal coverage of the data could help marginalize the influence of large, stochastic emission events, while adding robustness to the spatial estimates. This may also open the door to investigating seasonal or annual variability. Factoring in diffusion could increase the accuracy of spatial estimates, assuming that a logical background value for each species can be established. Finally, through using multiple classifications and reporting the average, the robustness of the neural network itself can be improved, and

error statistics can then be more readily calculated and communicated.

Ultimately, the characterization framework presented may offer a straightforward method for monitoring and assessing changes in the multi-species hydrocarbon signatures of the Barnett Shale region. Further, the domain filling technique should be easily applicable to any region with sufficient measurement density, and the simplicity afforded by the dimension-reduction inherent in self-organizing maps should facilitate multi-species monitoring in any such region. This multi-species monitoring could be especially useful to regulators or policymakers with a vested interest in subtle transitions brought about by any local change in activity. We expect that the ease and simplicity should especially lend itself to monitoring regional progress, especially in the time period following the implementation of new mitigation policy.

**Acknowledgements** The authors acknowledge the help of David Lyon, Ramon Alvarez, and Daniel Zavala-Araiza from the Environmental Defense Fund for their help in acquiring the auto-GC data and their perspectives on the early and intermediate stages of the project. We also thank the Texas Commission on Environmental Quality for graciously providing their datasets.

**Funding information** This investigation was funded by the Environmental Defense Fund as part of the Barnett Coordinated Campaign. Funding for EDF's methane research series, including this work, is provided by Fiona and Stan Druckenmiller, Heising-Simons Foundation, Bill and Susan Oberndorf, Betsy and Sam Reeves, Robertson Foundation, Alfred P. Sloan Foundation, TomKat Charitable Trust, and Walton Family Foundation.

## References

Allen, D.T. (2014). Atmospheric emissions and air quality impacts from natural gas production and use. *Annual Review of Chemical and Biomolecular Engineering*, *5*, 55–75.

Ashbaugh, L., Malm, W., Sadeh, W. (1985). A residence time probability analysis of sulfur at Grand Canyon National Park. *Atmospheric Environment*, *19*(8), 1263–1270.

Atkinson, R. (2000). Atmospheric chemistry of VOCs and NO(x). *Atmospheric Environment*, *34*(12-14), 2063–2101.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *140*, 123–140.

Breiman, L. (1996b). Out-of-bag estimation. Technical report. Department of Statistics, UC Berkeley.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Cheng, M.D., Hopke, P.K., Barrie, L., Rippe, A., Olson, M., Landsberger, S. (1993). Qualitative determination of source regions of aerosol in Canadian high arctic. *Environmental Science & Technology*, *27*(10), 2063–2071.

Draxler, R., & Hess, G. (1998). An overview of the HYSPLIT_4 modelling system for trajectories, dispersion and deposition. *Australian Meteorological Magazine*, *47*(1997), 295–308.

Fernández-Varela, R., Gómez-Carracedo, M.P., Ballabio, D., Andrade, J.M., Consonni, V., Todeschini, R. (2010). Self organizing maps for analysis of polycyclic aromatic hydrocarbons 3-way data from spilled oils. *Analytical Chemistry*, *82*(10), 4264–4271.

Field, R., Soltis, J., Murphy, S. (2014). Air quality concerns of unconventional oil and natural gas production. *Environmental Science: Processes & Impacts*, *16*(5), 954–969.

Gordalla, B.C., Ewers, U., Frimmel, F.H. (2013). Hydraulic fracturing: a toxicological threat for groundwater and drinking-water? *Environmental Earth Sciences*, *70*(8), 3875–3893.

Graedel, T.E. (2012). Chemical compounds in the atmosphere. New York, NY: Academic Press, Inc.

Guenther, A. (1999). Modeling biogenic volatile organic compound emissions to the atmosphere. In *Reactive hydrocarbons in the atmosphere* (pp. 98–116).

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P.I., Geron, C. (2006). Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). *Atmospheric Chemistry and Physics Discussions*, *6*, 3181–3210.

Harriss, R., Alvarez, R.A., Lyon, D., Zavala-Araiza, D., Nelson, D., Hamburg, S.P. (2015). Using multi-scale measurements to improve methane emission estimates from oil and gas operations in the Barnett Shale Region, Texas. *Environmental Science & Technology*, *49*(13), 7524–7526.

Hopke, P.K. (2003). Recent developments in receptor modeling. *J Chemometrics*, *17*(5), 255–265.

Hultman, N., Rebois, D., Scholten, M., Ramig, C. (2011). The greenhouse impact of unconventional gas for electricity generation. *Environmental Research Letters*, *6*(4), 044008.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J., Tveito, O.E. (2008). Classifications of atmospheric circulation patterns: recent advances and applications. *Annals of the New York Academy of Sciences*, *1146*, 105–152.

Jackson, R.B., Vengosh, A., Carey, J.W., Davies, R.J., Darrah, T.H., O'sullivan, F., Pétron, G. (2014). The environmental costs and benefits of fracking. *Annual Review of Environment and Resources*, *39*, 327–362.

Jones, D.S., & Pujadó, P.R. (2006). *Handbook of petroleum processing*. Dordrecht, The Netherlands: Springer Science & Business Media.

Jones, K., Meldrum, M., Baird, E., Cottrell, S., Kaur, P., Plant, N., Dyne, D., Cocker, J. (2006). Biological monitoring for trimethylbenzene exposure: a human volunteer study and a practical example in the workplace. *Annals of Occupational Hygiene*, *50*(6), 593–598.

Karion, A., Sweeney, C., Kort, E.A., Shepson, P.B., Brewer, A., Cambaliza, M., Conley, S.A., Davis, K., Deng, A., Hardesty, M., Herndon, S.C., Lauvaux, T., Lavoie, T., Lyon, D., Newberger, T., Pétron, G., Rella, C., Smith, M., Wolter, S., Yacovitch, T.I., Tans, P. (2015). Aircraft-based estimate

of total methane emissions from the Barnett Shale region. *Environmental Science & Technology*, *49*(13), 8124–8131.

Kaski, S., Kangas, J., Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, *1*(3&4), 102–350.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Lan, X., Talbot, R., Laine, P., Torres, A. (2015). Characterizing fugitive methane emissions in the Barnett shale area using a mobile laboratory. *Environmental Science & Technology*, *49*(13), 8139–8146.

Lavoie, T.N., Shepson, P.B., Cambaliza, M.O.L., Stirm, B.H., Karion, A., Sweeney, C., Yacovitch, T.I., Herndon, S.C., Lan, X., Lyon, D. (2015). Aircraft-based measurements of point source methane emissions in the Barnett Shale Basin. *Environmental Science & Technology*, *49*(13), 7904–7913.

Leuchner, M., & Rappenglück, B. (2010). VOC source-receptor relationships in Houston during TexAQS-II. *Atmospheric Environment*, *44*(33), 4056–4067.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*, 18–22.

Mangiameli, P., Chen, S.K., West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, *93*(96), 402–417.

Miller, S.M., Wofsy, S.C., Michalak, A.M., Kort, E.A., Andrews, A.E., Biraud, S.C., Dlugokencky, E.J., Eluszkiewicz, J., Fischer, M.L., Janssens-Maenhout, G., Miller, B.R., Miller, J.B., Montzka, S.A., Nehrkorn, T., Sweeney, C. (2013). Anthropogenic emissions of methane in the United States. *Proceedings of the National Academy of Sciences*, *110*(50), 20018–20022.

Montgomery, S.L., Jarvie, D.M., Bowker, K.A., Pollastro, R.M. (2005). Mississippian Barnett Shale, Fort Worth basin, north-central Texas: gas-shale play with multi-trillion cubic foot potential. *AAPG Bulletin*, *89*(2), 155–175.

Moore, H.E., Butcher, J.B., Adam, C.D., Day, C.R., Drijfhout, F.P. (2016). Age estimation of Calliphora (Diptera: Calliphoridae) larvae using cuticular hydrocarbon analysis and artificial neural networks. *Forensic Science International*, *268*, 81–91.

Myers, T. (2012). Potential contaminant pathways from hydraulically fractured shale to aquifers. *Groundwater*, *50*(6), 872–882.

Nathan, B.J., Golston, L.M., O'Brien, A.S., Ross, K., Harrison, W.A., Tao, L., Lary, D.J., Johnson, D.R., Covington, A.N., Clark, N.N., Zondlo, M.A. (2015). Near-field characterization of methane emission variability from a compressor station using a model aircraft. *Environmental Science & Technology*, *49*, 7896–7903.

National Research Council (NRC) (1991). *Rethinking the Ozone Problem in Urban and Regional Air Pollution / Committee on Tropospheric Ozone Formation and Measurement*. Washington, D.C.: National Academy Press.

Papa, E., & Gramatica, P. (2008). Externally validated qspr modelling of voc tropospheric oxidation by no3 radicals. *SAR and QSAR in Environmental Research*, *19*(7–8), 655–668. PMID: 19061082.

Pearce, J.L., Beringer, J., Nicholls, N., Hyndman, R.J., Uotila, P., Tapper, N.J. (2011). Investigating the influence of synoptic-scale meteorology on air quality using self-organizing maps and generalized additive modelling. *Atmospheric Environment*, *45*(1), 128–136.

Pétron, G., Karion, A., Sweeney, C., Miller, B.R., Montzka, S.A., Frost, G., Trainer, M., Tans, P., Andrews, A., Kofler, J., Helmig, D., Guenther, D., Dlugokencky, E., Lang, P., Newberger, T., Wolter, S., Hall, B., Novelli, P., Brewer, A., Conley, S., Hardesty, M., Banta, R., White, A., Noone, D., Wolfe, D., Schnell, R. (2014). A new look at methane and non-methane hydrocarbon emissions from oil and natural gas operations in the Colorado Denver-Julesburg Basin. *Journal of Geophysical Research: Atmospheres*, *2*(303).

Poirot, R.L., Wishinski, P.R., Hopke, P.K., Polissar, A.V. (2001). Comparative application of multiple receptor methods to identify aerosol sources in northern Vermont. *Environmental Science and Technology*, *35*(23), 4622–4636.

Polissar, A.V., Hopke, P.K., Paatero, P., Kaufmann, Y.J., Hall, D.K., Bodhaine, B.A., Dutton, E.G., Harris, J.M. (1999). The aerosol at Barrow, Alaska: long-term trends and source locations. *Atmospheric Environment*, *33*(16), 2441–2458.

Railroad Commission of Texas (2014a). Texas Barnett Shale total natural gas production 2000 through April 2014.

Railroad Commission of Texas (2014b). Texas RRC - Barnett Shale Information.

Rella, C.W., Tsai, T.R., Botkin, C.G., Crosson, E.R., Steele, D. (2015). Measuring emissions from oil and natural gas well pads using the mobile flux plane technique. *Environmental Science & Technology*, *49*(7), 4742–4748.

Smith, M.L., Kort, E.A., Karion, A., Sweeney, C., Herndon, S.C., Yacovitch, T.I. (2015). Airborne ethane observations in the Barnett Shale: quantification of ethane flux and attribution of methane emissions. *Environmental Science & Technology*, *49*(13), 8158–8166.

Strebel, K., Espinosa, G., Giralt, F., Kindler, A., Rallo, R., Richter, M., Schlink, U. (2013). Modeling airborne benzene in space and time with self-organizing maps and Bayesian techniques. *Environmental Modelling and Software*, *41*, 151–162.

Sugimoto, I., Seyama, M., Nakamura, M. (1999). Detection of petroleum hydrocarbons at low ppb levels using quartz resonator sensors and instrumentation of a smart environmental monitoring system. *Journal of Environmental Monitoring: JEM*, *1*(2), 135–142.

Sutton, R., Maclean, H., Swinbank, R., O'Neill, A., Taylor, F. (1994). High-resolution stratospheric tracer fields estimated from satellite observations using Lagrangian trajectory calculations. *Journal of the American Meteorological Society*, *51*(20), 2995–3005.

Texas Commission on Environmental Quality (TCEQ) (2012). Oil and gas handling and production facilities. https://www.tceq.texas.gov/assets/public/permitting/air/NewSourceReview/oilgas/106-352sub1.pdf.

Texas Commission on Environmental Quality (TCEQ) (2014a). Automated Gas Chromatographs (AutoGCs) Barnett Shale Monitoring Network.

Texas Commission on Environmental Quality (TCEQ) (2014b). Texas state and local air quality planning group - modeling domains.

Texas Commission on Environmental Quality (TCEQ) (2017). Keyword index to air permits by rule. https://www.tceq.texas.gov/permitting/air/permitbyrule/pbr_index.html#g.

Tonacci, A., Corda, D., Tartarisco, G., Pioggia, G., Domenici, C. (2015). A smart sensor system for detecting hydrocarbon volatile organic compounds in sea water. *Clean - Soil, Air, Water*, *43*(1), 147–152.

Townsend-Small, A., Marrero, J.E., Lyon, D.R., Simpson, I.J., Meinardi, S., Blake, D.R. (2015). Integrating source apportionment tracers into a bottom-up inventory of methane emissions in the Barnett shale hydraulic fracturing region. *Environmental Science & Technology*, *49*(13), 8175–8182.

United States Energy Information Adminisration (EIA) (2014). Natural gas production, transmission, and consumption, by state, 2013.

United States Energy Information Administration (2017). Annual energy outlook 2017, with projections to 2050. Technical report.

United States Environmental Protection Agency (1994). Chemicals in the Environment: OPPT Chemical Fact Sheets.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. (1999). Self-organizing map in Matlab: the SOM Toolbox. In *Proceedings of the Matlab DSP conference* (pp. 35–40): Espoo.

Vinciguerra, T., Yao, S., Dadzie, J., Chittams, A., Deskins, T., Ehrman, S., Dickerson, R.R. (2015). Regional air quality impacts of hydraulic fracturing and shale natural gas activity: evidence from ambient VOC observations. *Atmospheric Environment*, *110*, 144–150.

Whitby, R., & Altwicker, E. (1978). Acetylene in the atmosphere: sources, representative ambient concentrations and ratios to other hydrocarbons. *Atmospheric Environment*, *12*(6–7), 1289–1296.

Xiao, Y., Jacob, D.J., Turquety, S. (2007). Atmospheric acetylene and its relationship with CO as an indicator of air mass age. *Journal of Geophysical Research: Atmospheres*, *112*(February), 1–14.

Xie, Y., & Berkowitz, C.M. (2007). The use of conditional probability functions and potential source contribution functions to identify source regions and advection pathways of hydrocarbon emissions in Houston, Texas. *Atmospheric Environment*, *41*(28), 5831–5847.

Xie, Y.-L., Hopke, P.K., Paatero, P., Barrie, L.A., Li, S.M. (1999). Locations and preferred pathways of possible sources of Arctic aerosol. *Atmospheric Environment*, *33*, 2229–2239.

Yacovitch, T.I., Herndon, S.C., Pétron, G., Kofler, J., Lyon, D., Zahniser, M.S., Kolb, C.E. (2015). Mobile laboratory observations of methane emissions in the Barnett Shale Region. *Environmental Science & Technology*, *49*, 7889–7895.

Yacovitch, T.I., Herndon, S.C., Roscioli, J.R., Floerchinger, C., McGovern, R.M., Agnese, M., Pétron, G., Kofler, J., Sweeney, C., Karion, A., Conley, S.A., Kort, E.A., Nähle, L., Fischer, M., Hildebrandt, L., Koeth, J., McManus, J.B., Nelson, D.D., Zahniser, M.S., Kolb, C.E. (2014). Demonstration of an ethane spectrometer for methane source identification. *Environmental Science and Technology*, *48*, 8028–8034.

Zavala-Araiza, D., Sullivan, D.W., Allen, D.T. (2014). Atmospheric hydrocarbon emissions and concentrations in the Barnett shale natural gas production region. *Environmental Science & Technology*, (2).