Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: chemometric dimension reduction techniques for the deconvolution of complex atmospheric datasets

by

K. P. Wyche et al.

The authors are grateful to both referees for their excellent, thoughtful and insightful reviews, their comments were most welcome and indeed were very useful for manuscript improvements to be implemented. The following document addresses all of the points raised by both reviewers and explicitly details all of the correction made to the manuscript (revised manuscript attached separately with “track changes” documenting alterations) in accordance with both of the referee’s suggestions.

Author response key:

Plain, green text denotes referee comments

Italicised, back text, prefixed with “Author Response:” denotes direct author response to preceding referee comment

Plain, blue text, inside “” marks, denotes author changes to the manuscript in response to requests or recommendations from the reviewer
The manuscript of Wyche et al. represents an exciting step forward in the analysis of complex data sets, relevant to the formation of secondary organic aerosol (SOA). The authors use three statistical analysis approaches (PCA, HCA, and PLS-DA) for dimension reduction. The methods are applied to data obtained from a number of chamber studies involving different precursors grouped into four categories; application of the methods reproducibly resulted in successful classification of the gas- and particle-phase composition spectra by precursor category. While the authors suggested the potential for such approaches in ambient data mining, the results were not overstated and potential limitations were acknowledged. The paper was well written and the methodology was easy to follow. This paper should be of great interest to ACP readers involved in acquisition and interpretation of such data sets, as well as to readers interested in improving model representation of SOA.

**Author Response:** The authors thank the referee for their summary and recommendation.

---

**Content**

Studies have shown that limonene for example typically has higher SOA yields than other cyclic monoterpenes (Lee et al., 2006, JGR 111 D17305; Fry et al., 2014, ES&T 48: 11944-11953). Though the oxidation products of the two cyclic monoterpenes were separable from the other categories by the statistical approaches, did the results also suggest differences between the two precursors that may help explain such observations?

**Author Response:** To a certain degree the monoterpenes, limonene and α-pinene were separable within the monoterpane group, with three out of the four α-pinene experiments located to the upper and right region of the cluster. The referee is correct, this distribution within the statistical space results from the differences in respective gas-phase organic compositions of the two precursor types, which then in turn impact SOA yield and composition. As such the text has been amended in
Section 5, such that there are now 3 distinct discussion sections: 5.1 Mapping chemistry (the original text), 5.2 Mapping within a class (i.e. the following text supplied in response to this specific comment by referee 1) and 5.3 Mapping reactivity (i.e. additional text supplied in response to the following comment from referee 1 – see below – and comment 14 from referee 2). The additional text has been inserted on page 1676, after line 2, to answer the referees’ question and to include the references that the referee has suggested:

“5.2 Mapping within a class

Within the monoterpene group there is a small degree of separation between the limonene and α-pinene experiments, with three out of the four α-pinene experiments located to the upper and right region of the monoterpene cluster. This distribution/separation within the group may be a consequence of precursor-specific reaction pathways; for instance, although structurally similar, α-pinene and limonene react at somewhat different rates with respect to both OH and O3 (Atkinson and Arey, 2003). Over a fixed time period, such system reactivity will govern the degree of oxygenated content present within a closed analyte matrix and may facilitate the isolation of specific reaction pathways. Furthermore, the separation of such similar gas-phase precursors within a class cluster may help us to elucidate differences in resultant SOA yield and composition (e.g. limonene tends to have a larger SOA yield than α-pinene; Lee et al., 2006; Fry et al., 2014). It therefore may be possible with the use of larger and more detailed data sets to employ loading information to determine the importance of certain products to SOA composition. However, additional data to those reported here would be required to fully test this hypothesis.”

Along the same lines, within a category, does the spread/distance of clusters/specificity indicate dependence on VOC/NO levels, RH, or other factors thought to influence SOA formation?

Author Response: This is another interesting point, which we were originally intending to reserve for future work. However, following this comment and comment
13) from referee 2, we have now included a small separate analysis of the toluene data, which was conducted over a reasonable range of VOC/NOx conditions (low, medium, high NOx), to demonstrate how the experiments of one specific precursor type distribute/separate within the original class on account of system driving specifics, in this case, VOC/NOx ratio. This new information has been included in a new section, Section 5.3, following the original text on page 1676, after line 2, along with an additional Figure (10), shown here:

“5.3 Mapping reactivity

In order to explore how the PCA technique can be used to investigate product distributions driven by certain starting conditions, a separate analysis was conducted on the five toluene experiments. In this instance we investigate the product distribution dependency on initial VOC/NOx ratios. The VOC/NOx ratios employed nominally represent “low”, “medium” and “high” NOx conditions, with values of roughly 11 (i.e. low NOx, NOx limited; two experiments), 4 (i.e. moderate NOx, two experiments) and 1 (i.e. high NOx, VOC limited; one experiment), respectively (e.g. see work by Wagner et al., 2003). The resultant PCA loadings bi-plot (produced using the methodology described in Section 3) is given in Figure 10.

From inspection of the PCA loadings bi-plot in Figure 10, it is clear that the toluene oxidation spectra distribute in statistical space according to their respective VOC/NOx ratios. Figure 10 shows the low NOx, high VOC/NOx ratio experiments grouped in the lower right-hand quadrant of the PCA space, principally influenced by loadings representing toluene (m/z 93 and 77, parent and fragment ions, respectively; note m/z 93 off-scale in Figure 10) and cresol (m/z 109). Summed spectra containing larger quantities of precursor would suggest the presence of a less reactive environment, which is the case here, where low NOx levels in the NOx limited regime, result in low [O3] (via NO2 → O(3P) photolysis) and low [OH] (via O3 → O(1D) photolysis). Similarly, the relatively large contribution from cresol to the low NOx summed spectra, originates from a larger net cresol concentration across the experiment on account of low system reactivity (i.e. loss via reaction with OH).
The moderate NOₓ, medium VOC/NOₓ experiments group uniquely in the lower left-hand quadrant of the PCA space in Figure 10, principally on account of loadings representing benzaldehyde (m/z 107) and the ring-opening products, citraconic anhydride (m/z 113), 4-oxo-2-pentenal, maleic anhydride and/or angelicalactone (m/z 99) and methyl glyoxal (m/z 73). The greater abundance of higher generation, ring-opening products implies a more reactive environment than that formed under low NOₓ conditions. Larger net benzaldehyde concentrations originate from greater system reactivity and greater abundance of NO to fuel the RO₂ + NO reaction.

The high NOₓ, low VOC/NOₓ ratio experiment is sited in the left-hand half of Figure 10, on account of it possessing higher system reactivity (with respect to the low NOₓ experiments) and the resultant presence of ring-opening product ions (as the case for the moderate NOₓ experiments). However, the low VOC/NOₓ ratio experiment is uniquely displaced into the upper region of the PCA space owing to a large contribution from 2-butenedial and/or 2(SH)-furanone (m/z 85, off scale in Figure 10) to the summed spectra (the yields of both of which are likely to be important under high NOₓ conditions, owing to reaction through the RO₂ + NO channel).
Figure 10: PCA loadings bi-plot of the second vs. first principal components derived from the PCA analysis of the toluene experiments. Experiments were conducted under low NOx, high VOC/NOx ratio (red diamonds), moderate NOx, medium VOC/NOx ratio (green squares) and high NOx, low VOC/NOx ratio (blue triangle) conditions. For clarity, the scale has been set to show the bulk of the data, hence m/z 93 and 85 are not shown.

Lee et al. reported a number of m/z fragments for oxidation products of isoprene, monoterpenes, and a sesquiterpene detected by PTR-MS. It would be interesting to know if the same conclusions could be ascertained from the reported m/z fragments in Lee et al.; namely that there are significant relationships between the fragments of “like” precursors. There may not be enough data to perform a full analysis as was done in the subject manuscript, but the authors may think about whether their results are similar/supported by Lee et al.

Author Response: The referee is indeed correct, there are similarities between the fragments of products of “like” precursors. Here, the main similarities are between the monoterpenes limonene and α-pinene, and myrcene and linalool. However, we have highlighted this information for the major oxidation products in Section 4.2, on page 1666, line 15 onwards, e.g. "The monoterpene groupings are influenced by the presence of m/z 107, 151 and 169 (primary aldehydes, piononaldehyde and limononaldehyde) and 139 (primary ketone, limonaketone) ions in their mass spectra" and "Helping to separate the straight chain from cyclic monoterpenes are m/z 95 and 93, dominant features in both the myrcene and linalool spectra (relative abundance 10–24% for m/z 93). m/z 93 has previously been identified as a major fragment ion of first generation myrcene and linalool products 4-vinyl-4-pentenal and 4-hydroxy-4-methyl-5-hexen-1-al, respectively (Shu et al., 1997; Lee et al., 2006)"

However, following this comment from referee 1 and Technical Comment 5) from referee 2, the text has been amended as follows on page 1666, line 28, to link the findings of this work to that of Lee at al., 2010:
“A list of major ions contributing to the separation of spectra in statistical space is given in Table 3 along with potential identities and precursors. It is worthy of note here that these ions and the overall fragmentation patterns observed in this study are largely in-line with those reported by Lee et al. (2006), in their comprehensive PTR-MS analysis of a wide range of BVOC precursors and their associated oxidation products.”

It is suggested that the authors carefully review their citations. There are several places in which widely cited previous work is omitted in favor of self-citations.

Author Response: We thank the referee for pointing this out and we have amended our citations as follows, however we would like to highlight that in some instances we include numerous citations to our previous work in order to explain specifics of the experiments that are employed here.

One example can be found on p. 1655, line 22, regarding the importance of findings from chamber studies over the past decade. The authors do not cite the work coming out of the Caltech chamber (as early as the late 1990’s); the parameters from Hoffman et al., 1997 and Griffin et al., 1999 still are widely used in chemical transport models (see also papers by Cocker, D.).

Restricted to the last 10 years, lead authors on Caltech chamber publications include: Ng., N.; Kroll, J.; Surratt, J. and others.

Much influential work has also been conducted over the last 10 years in the Carnegie Mellon chamber; lead authors include: Presto, A., Hildebrandt, L., Grieshop, A., and others.

Author Response: The referee is indeed correct, this is an oversight on our part and the following citations and references have been added to the amended text on Page 1655, line 22:

“A significant proportion of the findings gained regarding SOA over the last decade and more have come from atmospheric simulation chamber experiments,
intermediate in complexity between classical single-process experiments and the fully open system (for various different chamber systems and VOC precursors, see for example, Pandis et al., 1991; Odum et al., 1996; Hoffmann et al., 1997; Griffin et al., 1999; Glasius et al., 2000; Cocker et al., 2001; Jaou and Kamens, 2003; Kleindienst et al., 2004; Presto et al., 2005; Bloss et al., 2005; Rohrer et al., 2005; Ng et al., 2006, 2007; Dommen et al., 2006; Surratt et al., 2006; Grieshop et al., 2007; Chan et al., 2007; Wyche et al., 2009; Hildebrandt et al., 2009; Rickard et al., 2010; Camredon et al., 2010; Chhabra et al., 2011; Hennigan et al., 2011; Jenkin et al., 2012).”

Page 1658, line 6:
“all have previously been shown to form SOA under simulation chamber conditions (e.g. Hoffmann et al., 1997; Griffin et al., 1999; Glasius et al., 2000; Jaou and Kamens, 2003; Presto et al., 2005; Ng et al., 2006; Surratt et al., 2006; Dommen et al., 2006; Lee et al., 2006; Hallquist et al., 2009; Alfara et al., 2013, and references therein).”

Page 1665, line 6:
“Full details describing the underlying chemical and physical mechanisms at play within such experiments can be found elsewhere (e.g. Larsen et al., 2001; Bloss et al., 2005; Paulsen et al., 2005; Surratt et al., 2006 and 2010; Wyche et al., 2009; 2014; Camredon et al., 2010; Rickard et al., 2010; Eddingsas et al., 2012b; Hamilton et al., 2011; Jenkin et al., 2012; Alfara et al., 2012, 2013; and references therein).”

Page 1670, line 9:
“For further details regarding β-caryophyllene oxidation products, see for example Lee et al. (2006), Winterhalter et al. (2009), Hamilton et al. (2011), Chan et al. (2011), Li et al. (2011) and Jenkin et al. (2012) and references therein, and Sect. 5.”

Page 1670, line 18:
“For further details regarding α-pinene and limonene oxidation products, see for example Larsen et al. (2001), Jaou et al. (2003), Capouet et al. (2004), Jenkin (2004),
Jaoui et al. (2006), Lee et al. (2006), Ng et al. (2006), Camredon et al. (2010) and Hamilton et al. (2011) and references therein.”

Page 1675, line 12:
“β-caryophyllene readily forms particulate matter on oxidation (e.g. Jaoui et al., 2003; Lee et al., 2006; Winterhalter et al., 2009; Alfarra et al., 2012; Chen et al., 2012)”

References added to manuscript list:


Editorial

p 1655, line 18: “Atmospheric chemistry” as a system is awkward. One suggestion: “The chemistry of the atmospheric system is highly nonlinear…”

Author Response: The text has been altered accordingly

p 1656, line 9: The “CIR-TOF-MS” abbreviation can be introduced on line 4 and used exclusively here.

Author Response: The text has been altered accordingly

p. 1658, section 2 heading: The adjective “Experimental “ needs a verb, or needs to
be changed to a noun.

**Author Response:** The text has been altered to read “Experimental Details”

**p. 1661, line 8:** Remove “/” prior to photolytic

**Author Response:** The text has been altered accordingly

**p. 1666, line 19:** Change “caryophyllon” to “caryophyllene”

**Author Response:** The text has been altered accordingly

**p. 1667, line 26:** It is not clear what is meant by “oxidized atmospheres”.

**Author Response:** For clarity, the text has been altered to read,

“As a further test of the technique to distinguish between and to classify VOCs and the matrix of oxidized organic compounds that may derive from their atmospheric chemistry, test data from an anthropogenic system was introduced into the model.”

**p. 1669, line 10:** It is suggested the sentence be reworded to more clearly indicate that data are available for only two experiments per each of the two noncyclic monoterpenes. As written the focus is on “only two types of precursor”, which is the same as the cyclic monoterpenes. In summary, it is the number of experiments, not the number of precursors that is likely affecting the result.

**Author Response:** the text has been altered to read,

“The greater spread in confidence of the noncyclic monoterpane group is once again likely to result to some extend from the low number of repeat experiments employed (i.e. only two each for myrcene and linalool).”

Can the panel and text sizes be increased for Figs. 7 and 8? They are hard to read (even w/significant expansion-175%)?
Author Response: Figures 7a, 7b, 8a and 8b have been altered as requested, with increased panel and text sizes. We will also work with the editing team to ensure in typesetting that all figures are sufficiently clear.

The authors may consider more clearly indicating the difference between Figs. 7 and 8 in the figure panels (e.g., just adding a AMS and LC-MS header).

Author Response: Figures 7a, 7b, 8a and 8b have been altered as requested, with additional headers to indicate LC-MS/MS or cTOF-AMS data.

The authors may consider using consistent colors for the different categories in each of the plots (generally the same in PCA plots, but not in dendrograms).

Author Response: We thanks the referee for pointing this out, but we would prefer to leave the dendrogram colors as there are, as they represent the best combination for the clearest description of the data, i.e. the dendrogram colors were chosen specifically to optimize clarity of view within the figure. The colors and symbols are the same for all PCA and PLS-DA plots, e.g. isoprene is always represented by a blue diamond.

Caption Figure 10: Remove “/” after abundance

Author Response: The text has been altered accordingly
Wyche et al. describe development of a chemometric mapping of gas-phase and particle-phase matrices from oxidation of several relevant compounds: isoprene, terpenes (i.e. linalool, myrcene, limonene, a-pinene, B-caryophyllene), toluene, as well as the oxidation products of fig and birch trees. This is done via principal component analysis (PCA), hierarchical cluster analysis (HCA), and positive least squares-discriminant analysis (PLS-DA) on the mass spectra of the oxidation products from several systems. PCA is performed on gas-phase observations of chamber data using chemical ionization reaction time-of-flight mass spectrometry for the gas-phase to separate distinct regions associated with oxidation of isoprene, cyclic monoterpenes, B-caryophyllene, single chain monoterpenes, and toluene. Addition of mesocosm data (plant emission oxidation experiments) shows expected mapping of oxidation products from fig and birch trees as related to the expected emissions being isoprene-dominated and cyclic monoterpene-dominated, respectively. HCA analysis is performed, supporting the relation of the mesocosm oxidation products with associated precursor oxidation schemes represented in chamber data. Similar PCA and HCA analysis applied to the particle-phase data (liquid chromatography-ion trap mass spectrometry and compact time-of-flight aerosol mass spectrometry) also shows separation by precursor type. Using a zero-dimensional box model simulation based on the Master Chemical Mechanism for alpha-pinene photooxidation, it is argued that the employed statistical deconvolution techniques could be applicable for determining precursor type and potential mechanisms from ambient data on the basis that “model mass spectra” simulated under increasingly complex (closer to ambient conditions) are generally well-captured by “model mass spectra” under conditions more like a typical chamber experiment.

Reviewer’s recommendation:

The article overall is well-written and is novel in that it seems to provide great...
potential for use as identifying chemistry from a particular precursor type given complex sets of mass spectra from oxidation products in the gas and particle phases. The weakest argument is that made for use of this technique to elucidate chemical mechanisms in addition to identifying precursor type, as the specificity for separating mechanistic pathways is not fully demonstrated in the current analyses. Still, the content is appropriate for ACP readers and I would recommend publication after the following comments are addressed.

Author Response: The authors thank the referee for their summary and recommendation.

General Comments:
1) Abstract: The reader would benefit from more specific result statements included here, similar to the lines included in the Introduction section p. 1657, lines 7-15. 

Author Response: The following text has been added as requested,

Abstract, line 18, page 1653,
“Results show that “model” biogenic oxidative systems can be successfully separated and classified according to their gaseous oxidation products.”

Abstract, line 26, page 1653,
“More specifically, the addition mesocosm data from fig and birch tree experiments shows that isoprene and monoterpenes emitting sources, respectively, can be mapped onto the statistical model structure and their positional vectors can provide insight into their biogenic sources and controlling oxidative chemistry.”

2) P. 1656, lines 20-25: Discussion of currently used statistical techniques is rather cursory. There is not even mention of the commonly used positive matrix factorization technique widely used for AMS data analysis. Additional details of why the presented technique is novel/necessary should be discussed.

Author Response: We have amended the introduction as follows: to include a more
detailed discussion of other more commonly used statistical techniques, including positive matrix factorization. The additional details regarding why the technique presented here is novel are addressed in reply to Referee 2, point 3) below:

“Similar approaches using statistical analyses have been recently applied to both detailed and broad ambient aerosol composition data (e.g. (Heringa et al., 2012; Paglione et al., 2014)), particularly in the context of source apportionment (e.g. (Alier et al., 2013)). Different methods have been attempted by several groups to deconvolve organic aerosol spectra measured by the Aerosol Mass Spectrometer (AMS) in particular (e.g. Zhang et al., 2005, 2007; Marcolli et al., 2006; Lanz et al., 2007). Zhang et al. (2005) applied a custom principal component analysis (CPCA) method to extract two distinct sources of organic aerosols in an urban environment using linear decomposition of AMS spectra and later applied a Multiple Component Analysis technique (MCA, an expanded version of the CPCA) to separate more than two factors in datasets from 37 field campaigns in the Northern Hemisphere (Zhang et al., 2007). Marcolli et al. (2006) applied a hierarchical cluster analysis method to an ambient AMS data set, and reported clusters representing biogenic VOC oxidation products, highly oxidised organic aerosols and other small categories. Receptor modelling techniques such as Positive Matrix Factorization (PMF) employ similar multivariate statistical methods in order to deconvolve a time series of simultaneous measurements into a set of factors and their time-dependent concentrations (Paatero and Tapper, 1994; Paatero, 1997). Depending on their specific chemical and temporal characteristics, these factors may then be related to emission sources, chemical composition and atmospheric processing. For example, Lanz et al. (2007) and Ulbrich et al., (2009) applied PMF to the organic fraction of AMS datasets and were able to conduct source apportionment analysis identifying factors contributing to the composition of organic aerosol at urban locations. Slowik et al. (2010), combined both particle-phase AMS and gas-phase proton transfer reaction mass spectrometry data for the PMF analysis of urban air, and were able to successfully obtain “regional transport, local traffic, charbroiling and oxidative process” factors. By combining the two datasets, Slowik and colleagues were able to acquire more in-depth information regarding the urban atmosphere than could be derived from the
analysis of each of the sets of measurements on their own.

Because receptor models require no a priori knowledge of meteorological conditions or emission inventories, they are ideal for use in locations where emission inventories are poorly characterised or highly complicated (e.g. urban areas), or where atmospheric processing plays a major role. However, because all of the values in the profiles and contributions are constrained to be positive, the PMF model can have an arbitrary number of factors and the user must select the “best” solution that explains the data. This subjective step of PMF analysis relies greatly on the judgment and skill of the user.”

3) The authors mention p. 1657, lines 16-25, the potential for this technique to be used on ambient data sets, but the paper would be stronger references were provided for which similar statistical analyses are being done to map out oxidation chemistry related to certain precursors as is done here, or utilize available and published field data in the described analyses of the paper to prove the point. For example, why would this technique be useful over positive matrix factorization techniques, which now includes a way of identifying SOA formation contributions from isoprene?

Author Response: The referee is indeed correct; the paper would be stronger if we could have included ambient data in the statistical analyses in the same way that we included the mesocosm data. Unfortunately, we do not have such data available and have tried to address the issue to some extend using the MCM simulations to model a representatively complex system. It should be noted that we could find no useable published field data (as the referee suggests). A large amount of such data comprises solely the compounds of interest and not the detailed mass spectra, and where mass spectra were available, they were incompatible/not fit for purpose, either being e.g. flux data and/or comprising PTR-Quadruple-MS data recorded in selected ion mode.

We feel (as was pointed out by referee 1) that we have been fully open regarding the potential limitations of this methodology and have attempted to address them to the
We feel that taking the technique from first principle, to single precursor and then to mesocosm scenarios, achieves the goals that were set out and that application to the “real world” would represent the next logical step, for our research and for other colleagues; given the results presented in this proof of concept study, using this approach successfully in the field is the ultimate goal.

Regarding Positive matrix factorization (PMF) and the uniqueness of the method presented here; PMF is an extremely valuable technique now used successfully for PM source apportionment. The ensemble methodology described here however is not intended as a replacement, far from it, it simply presents a different approach to isolate composition (and potentially therefore, sources) and moreover to identify and map different chemical pathways. PMF generally utilizes data solely from particle phase measurements (however, PMF with both gas and particle phase data have been reported, i.e. Slowik et al., ACP, 2010) and hence usually from one analytical technique. However, the ensemble method discussed here utilizes data from both the gas- and particle-phases and data from different analytical techniques. Furthermore, it is not designed simply to comprise a type of source-receptor/source apportionment model; the objective is rather compositional isolation to instruct on potential sources AND on underlying chemical processes.

The methodologies described here are fundamentally simpler to use and require fewer base assumptions (there is less importance placed on user skill to determine the correct input parameters, number of factors to employ etc); the techniques work on fewer data, require no reference datasets and no prior knowledge of sources is required.

We thank the referee for this useful comment and have altered the introduction to highlight these distinctions, with the addition of the following text to line 6, page 1657:

“Unlike other statistical techniques such as PMF, the ensemble methodology presented here does not require the use of additional external databases (comprising information regarding different environments/reference spectra), is
simpler to use and less labour intensive, and places less importance on user skill in
the production of accurate and meaningful results. Moreover, the primary focus of
techniques such as PMF is on source identification/separation, whereas here the
focus is placed on compositional isolation.”

4) P. 1659, lines 9-12: While the precursors were reacted to near completion in the
chamber experiments, how can one verify that the oxidation scheme went to
completion to match, for example, major oxidation products in the MCM schemes
presented in Figure 9. That is, should not theoretically the PCA analysis for the
chamber experiments ideally match the MCM PCA if the chamber experiments are
covering the same range as the simulated chemistry?

First of all, we must reiterate as was pointed out in the manuscript, the MCM
simulations are not idealized, and their inclusion is not for the purpose of making a
direct model-measurement comparison. Owing to a lack of in field ambient spectra,
the simulations are designed to exemplify how the gas phase spectra could change if
the system complexity was increased. The model is not included for direct
comparison with the measurements, it is intended as a guide to help bridge the gap
between what we have reported from “model” and “mesocosm” experiments and the
real world. This is stated on page 1677, line 16.

However, we can verify that precursor oxidation went to completion to match the
measurements, because the model was purposefully constrained to measured α-
pinene concentration, i.e. the primary reactivity in the model was constrained to
match the primary reactivity in the measurements. Furthermore, from our previous
work (see Camredon et al., 2010) the modeled and measured temporal profiles of the
precursor and first (and later) generation products are in reasonably good
agreement, suggesting that the reactivities of the systems are comparable (i.e. that
the model system is chemically dynamic in a similar way as you would see in the
chamber). Use of the MCM under such single precursor chamber conditions has been
evaluated in details on several previous occasions, for example see work by Bloss et
al., 2005 (toluene), Camredon et al., 2010 – supplementary information (α-pinene),
Rickard et al., 2010 (1,3,5-TMB) and Jenkin et al., 2012 (β-caryophyllene). The reader
has now been informed of this with the addition of the following text to page 1677, line 18:

“For work regarding the evaluation of the MCM with respect to single VOC precursor chamber experiments (including model-measurement intercomparison), see for example, Bloss et al., 2005 (toluene), Metzger et al., 2008 and Rickard et al., 2010 (1,3,5-TMB), Camredon et al., 2010 (α-pinene) and Jenkin et al., 2012 (β-caryophyllene).”

5) P. 1659, Section 2.2: What impacts on the comparison of the chemical oxidation would be expected for the systems, considering that these systems were run across several environmental chambers?

**Author Response:** Each (Teflon) chamber system will have its own specific background wall chemistry, which arguably changes by various degrees from day-to-day and even between experiments (i.e. cleanliness of chamber will vary). For certain chambers, e.g. EUPHORE (Bloss et al., 2005; Zador et al., J. Atmos. Chem., 55:147-166, 2006) and PSISC (Metzger et al., 2008), detailed auxiliary mechanisms have been constructed and tested. These chemical mechanisms are based on similar reactions and fundamental properties and go some way towards explaining the intricacies of environment chamber experiments. From our experience, where rates of reaction (influenced by radical chemistry associated with the chamber walls) can change between chambers for a give set of experimental conditions, the overall composition of the gas phase does not vary widely for the types of experiments studied. This is inherent in the data reported here, as the spectra collected are compositionally similar between chambers. The primary additives generated from the chamber walls that could influence organic gas phase composition are nitrogen-containing species such as HONO. However, for the experiments reported here (with the exception of the mesocosm experiments), initial [NO_x] added > [NO_x] liberated from chamber walls.

In addition to the wall chemistry, the light intensity will vary from chamber to chamber, however, EUPHORE employs natural solar radiation and the MAC and PSISC...
are designed to approximate the solar spectrum, as detailed in the work referenced appropriately in the experiment section 2.2, page 1659.

In summary, we expect the impacts to be fairly minimal and within acknowledged uncertainty bounds.

6) P. 1667, lines 2-4: Aromatic compounds are also observed to be emitted from the biosphere, (e.g. aromatics like toluene as well as aromatic monoterpenes; see Guenther et al., 2012 and references within; doi:10.5194/gmd-5-1471-2012). Would the authors expect addition of these compounds to map onto the same region as toluene/monoterpenes, and would this not change the interpretation that the analysis can separate anthropogenic and biogenic precursor sources vs. just precursor structure? Do the authors have potential ideas for why there is overlap with toluene and the single chain monoterpenes? Would single chain sesquiterpenes (e.g. farnasene) also map out separately from B-caryophyllene as the cyclic and straight chain monoterpenes do?

Author Response: Guenther et al., (2012) estimated global annual emissions of ~ 535 Tg yr⁻¹ for isoprene and a sum of ~ 147 Tg yr⁻¹ for non-aromatic terpenes; total sesquiterpene emissions were of the order 29 Tg yr⁻¹. In contrast aromatic terpene emissions were estimated in a group of ~ thirty “other” monoterpenes with total annual emissions ~ 15 Tg yr⁻¹, i.e. only 2 % of Σ(isoprene, major non-aromatic terpenes and sesquiterpenes). Estimated annual toluene emissions are not individually reported by Guenther et al., but are assumed to be relatively minor, reported with an “additional 11 stress VOC(s)”, with a total emissions value of 7.8 Tg yr⁻¹. Sindalevora et al. (2014) report mean isoprene, Σ(monoterpenes), Σ(sesquiterpenes) and toluene annual global emissions to be of the order 594, 95, 20 and 1.5 Tg yr⁻¹, respectively, i.e. toluene emissions are only ~ 0.2 % of Σ(isoprene, major non-aromatic terpenes and sesquiterpenes). In summary, biogenic aromatic compounds comprise only a small fraction of total average terpene emissions and it is likely therefore that their oxidation products would be present in such low abundance compared to those discussed within our work, that they would not cause
severe spectral/statistical interference.

Under exceptional circumstances where aromatic terpenes are present in significant abundance, it is possible that their compositional spectra could group in the vicinity of those of other monoterpenes, on account of them possessing spectra with some similar features, e.g. Lee et al. (2006; DOI: 10.1029/2006JD007050) investigated the PTR-MS spectra of methyl chavicol and reported the presence of a parent ion of m/z 149 and major product ions of m/z 151, 137, 121 and 109. Of these ions m/z 149, 151 and 137 do not appear in toluene PTR-MS spectra and moreover, the ions of m/z 151, 137 and 109 are also significant contributors to monoterpene PTR-MS oxidation spectra.

Experiments conducted with aromatic terpenes did not form the focus of this initial work and as such no data is available here to test the exact positioning of aromatic terpenes in the statistical space, however, in addition to the findings of Guenther, Sindalevora and Lee et al., the hypothesis that aromatic “like” BVOCs would not group with toluene was tested here with data from α-terpinene photooxidation. α-terpinene is not an aromatic terpene, but unlike α-pinene and limonene it does contain a 6-member carbon ring containing 2 C=C double bonds.

Results: α-terpinene oxidation produces a significant amount of α-terpinaldehyde, which like pinonaldehyde (from α-pinene photooxidation) possesses a spectral fingerprint including ions of m/z 169, 151, 123 and 107. As with other non-aromatic terpenes, α-terpinene photooxidation PTR-MS spectra also contains ions of m/z 139 (c.f. limonaketone from limonene) and 171 (also observed in α-pinene photooxidation spectra), as well as more unique features, such as an m/z 143 ion of significant abundance (also see Lee et al., 2006, DOI: 10.1029/2006JD007050). In combination, such features help to group the α-terpinene photooxidation spectrum in the vicinity of other terpenes, separate in statistical space from the aromatic toluene photooxidation spectra.
The above figure and a short description matching the above text have been added to a new Supplement document and section 4.3 text has been amended accordingly in order to direct the reader appropriately:

“A brief discussion regarding aromatic BVOCs is provided separately in the supplementary material.”

We do not believe that there is significant overlap between the straight chain terpenes and toluene; the overlap observed is from the high (95%) confidence limits, the magnitude of which is discussed in the main text of the paper. Also as discussed in the paper, the toluene cluster is located close to the straight chain monoterpene cluster owing to similar (but not identical) ion content (but not abundance), including primarily m/z 93. Toluene separates from the straight chain monoterpenes owing to unique spectral features such as m/z 99 and 113, resulting from ring-cleavage and the subsequent production of e.g. angelica lactone and hexenedione (and isomers thereof), respectively.

We have not yet performed experiments with single chain sesquiterpenes such as farnasene, however it is likely that, as with the cyclic and non-cyclic monoterpenes, single chain sesquiterpene spectra would group in the vicinity of cyclic
sesquiterpenes, owing to their similar composition and chemistry, but uniquely displaced as a consequence of their structure (e.g. unsaturated, double bond content).

7) P. 1669, lines 17-20: Why are the birch trees sensitivity cited to be low due to only two repeat experiments, but the fig trees experiments also only having two repeat experiments have high sensitivity and specificity?

**Author Response:** As stated in the main text, page 1669, Section 4.4, lines 17 – 20, the low birch sensitivity may result from a combination of (i) a low number of repeat experiments AND (ii) the aging trees producing lower emissions during the last experiment. As we state in the current text, the low sensitivity for birch trees is likely to result from a combination of these two factors.

8) Section 4.5: Why do the authors choose to do separate PCA analyses on the gas-phase data versus the particle-phase data? Considering that there would be overlap for semi-volatiles (e.g. B-caryophyllinc acid as stated p. 1670, line 10), should not a PCA analysis (though mindful of the different analysis technique) be more ideal (as well as for interpreting ambient data) to map regions of any chemistry associated with various precursor types? Should it be inferred that the separation of analysis on gas vs. particle-phase data is the recommended approach to apply this type of analysis to ambient data? I would be curious to see how the analyses would compare between gas-phase only, particle-phase only, and gas and particle-phase together.

**Author Response:** What the referee suggests here is an excellent idea, and one that we were hoping to execute in future work, however, such an undertaking will be extremely complex and a step beyond what we are trying to present here. To begin with, this task would require knowledge of (i) ALL of the species present in both the gas and aerosol phases (which the community does not possess yet) and (ii) the EXACT fragmentation patterns and ionisation routes for all of these species under each of the ionisation regimes used. If these data were available, one would then have to (re-) construct mass spectra representing the gas and particle phase matrices
for each of the experiments employed, with no detection method specific alteration
to the mass of each parent compound (e.g. protonation, adduct formation, fragmentation etc) or instrument biases. At present it is not possible to do this accurately, however we are steadily working towards obtaining such knowledge.

Moreover, methodology such as Parallel Factor Analysis would perhaps be more suitable to perform such an investigation, which again is a whole new piece of work in itself. The purpose of this work is to introduce these new methods of working with large atmospheric data sets, partially the aim of which is to reduce computational load, not extend it, which would be required in this instance.

9) Figure 7a: The plot is produced using LC-MS particle data, but what biases might be at work here based on this technique’s sensitivity to certain organic aerosol products?

Author Response: In any analytical method there will be a degree of bias towards particular species. The LC-MS was operated in negative ionisation mode and so here we are looking at the species with acidic protons that can be removed easily. This leads to certain biases, as the ionisation efficiency is different for different structures. However, this plot shows the key ions that allow the mass spectral patterns of the SOA from different precursors to be separated based on the observed chemical speciation. The same analysis in positive mode (which targets different functional groups) will likely lead to some different ions being important for differentiating the precursors. However all samples were analysed using the same methods and so the impact of any systematic biases are minimised.

10) Figure 8a and accompanying discussion p. 1671, lines 15- p. 1672, line 3: The meaning of M43, M44 should be discussed in context as to whether the relative aging of the oxidation systems were actually comparable. Although the precursors were near complete reaction, did the oxidation develop enough such that B-caryophyllene schemes could take on additional spectral contributions from M43, M44? That is, was their sufficient oxidant such that secondary reactions (producing
characteristic oxidation products and associated ions) could occur equally across precursor systems?

Author response: the meaning of M43 and M44 has been discussed in the text (e.g. see page 1671, line 15 onwards). As can be seen from the Experimental section (and Table 1) and references therein, the MAC/AMS experiments were designed to be as comparable as possible, i.e. experimental conditions were kept similar between individual experiments, with light conditions, relative humidity and initial VOC/NOx ratios kept roughly the same from experiment-to-experiment. The reader is guided to a full discussion of these experiments throughout Sections 2 and 4 (e.g. Alfarra et al., 2012 and 2013) to ensure that they are fully apprised of the experiments employed.

The main difference between the oxidation systems investigated was in terms of reactivity of the various precursor compounds with respect to O₃ and OH. Indeed, we purposefully chose a sufficient span of precursors to cover a range of reactivities and as such, each system exhibited a different overall reactivity, but on the time scales of our experiments, the systems were kept as comparable as possible. The references to our previous work have been made in part to account for/explain this.

There was indeed sufficient oxidant throughout all experiments. As described in Alfarra et al., throughout each of the experiments discussed, O₃ was produced through system photochemistry. Once formed O₃ was able to react appreciably with each of the target BVOC precursors. During transit through each respective O₃-alkene reaction mechanism and in the presence of water vapor, OH was also produced. As can be seen in Alfarra et al. (2013), with the exception of β-caryophyllene, each experiment was characterized by constant net O₃ production (and hence concomitant OH production), meaning there was always sufficient oxidant to allow the formation and reaction of second generation products; this can also be seen in the VOC temporal profiles presented in Alfarra et al., 2013. β-caryophyllene was observed to consume O₃ as it was produced and net O₃ production was observed only during the latter stages of the experiments; however during the net O₃ production stage, there was again sufficient oxidant to allow the formation and reaction of secondary products, this can be clearly seen in our companion work
on β-caryophyllene (Jenkin et al., 2012).

Furthermore, from measurement of the m/z 44 and 43 ions as a function of time in the AMS data, SOA chemical transformation was observed and fully reported, again in Alfarra et al., 2013.

However for further clarity, the reader has been directed to the work by Hamilton et al. Jenkin et al. and Alfarra et al. again, following the text describing particle phase results, in section 4.5, page 1672:

“For full details regarding the chemical evolution of the particle phase and the particle-phase specific experiments in general, see Hamilton et al. (2011), Jenkin et al., (2012) and Alfarra et al. (2012 and 2013)”.

11) P. 1677, lines 25-27: Can the authors give a statistic that relates the similarity of the “model mass spectra” to the experimental data set from the chambers? If this model mass spectra generated from the MCM is similar enough to the experimental MS obtained, why not run the PCA on MCM generated model mass spectra for each system and see if the results are similar to the PCA for the gas and particle-phase data? Or at least, in Discussion section, when trying to relate ions characteristic of particle-phase oxidation products from grouped precursors, can a metric be used for testing similarity of the MCM model mass spectra expected and the weight of ions affecting the various B-caryophyllene and myrcene mappings for example?

Author Response: A comparison between the model and measured mass spectra is not straightforward and would first require knowledge of all of the species present in the gas phase and the exact fragmentation patterns and ionisation routes for each of these species. Furthermore, one would need to understand all of the isobaric instances in the measured mass spectra; a difficult task when poorly characterised ions fragment and overlap. One would then have to (re-) construct mass spectra representing the gas phase matrices for each of the experiments employed, with no detection method specific alteration to the mass of each parent compound (e.g. protonation, adduct formation, fragmentation etc) or instrument biases (a simplified,
first approximation attempt at this was made for α-pinene in Camredon et al., 2010).

As we did with Wyche et al., 2009 and Rickard et al., 2010, we plan to work on a follow up/companion modelling paper to the current work, in which we perform/present more complex and in-depth work than is possible to include here, including the statistical analyses of modelled mass spectra as suggested by the referee. Unfortunately, to include all of this in one piece of work would make the product far too large and also divert from the original aims of the paper. The purpose of the modelling exercise performed here is simply to help us take steps towards understanding how useful these techniques are for application in a real world environment, and the discussion presented does just that. What the referee proposes here is an excellent idea, but a substantial and separate piece of work entirely. (Note, that the MCM scheme currently does not yet include mechanisms for myrcene or linalool, and is not directly coupled to partitioning code).

12) P. 1678, lines 26-29: This argument would be strengthened if the authors did similar MCM analyses as in Fig. 10 for the other precursor types represented in Fig. 9 and provided metrics for the similarity of the model mass spectra with that observed in the chamber experiments.

Author Response: As already stated, the purpose of performing the MCM analysis was to help us move the technique towards a “real-world” environment and to give some indication as to whether the techniques successfully employed in chamber and mesocom environments could be useful in the chemically more complex real world atmosphere. As pointed out by the referees, we are fully open regarding potential limitations in moving towards the real atmosphere, yet we believe at this stage of our work, the non-idealized simulations reported, provide a sufficiently chemically complex model with which to support the arguments made.

We feel that performing the simulations using the α-pinene mechanism was the most appropriate of the four options available, i.e. α-pinene, limonene, isoprene and β-caryophyllene (note, as yet no mechanisms are available in the MCM for myrcene and linalool oxidation). α-pinene was chosen primarily because it is the most well
characterised of the available options and moreover both limonene and β-caryophyllene are recent additions to the framework and therefore less well understood. Moreover, to perform the analyses proposed by the referee would be a substantial undertaking and extend the current work too greatly. It is the intention of the authors to work on a follow up/companion modelling paper (as we did with Wyche et al., 2009 and Rickard et al., 2010), where we perform/present more complex and in-depth work than is possible to include here, including statistical analyses on the modelled mass spectra.

As described in point 8), a comparison between the model and measured mass spectra would first require knowledge of all of the species present in the gas-phase and the exact fragmentation patterns and ionisation routes for each of these species. Furthermore, one would need to understand all of the isobaric instances in the measured mass spectra; a difficult task when poorly characterised ions fragment and overlap. However, this is not the current focus of this work, indeed one of the major points of the presented work is to employ the “mass spectral fingerprints” rather than speciated detail.

13) P. 1679, line 12: Authors claim that the gas-phase oxidation products of each structural type can be grouped “according to the controlling chemistry and the products formed.” While it is evident that the products formed (as interpreted by the accompanying MS) informs this grouping, what chemical controls are derived from the PCA analysis? I think of e.g. varying NO<sub>x</sub> and O<sub>3</sub> levels as different chemical regimes/controls on the chemistry, but this point is not emphasised in the PCA mapping of the various experiments grouped by precursor. Where is this shown/taken from the figures? Or, please clarify what is meant here.

Author Response: The referee is indeed correct, under the current analysis we have not demonstrated this and the text has been amended accordingly. To this end we have included a small separate analysis of the toluene data, which was conducted under a reasonable range of VOC/NO<sub>x</sub> conditions (low, medium, high NO<sub>x</sub>). This new information has been included in a new section, Section 5.3 along with an additional
Figure (10), shown here. Both the additional text and the new Figure 10 were provided following the second comment made by referee 1 (refer to Response to Referee 1).

14) P. 1679, lines 13-15: In line with comment above, this sentence is rather vague. No description thus far has really emphasized “decoding of mechanisms” by starting with an ensemble of summed mass spectra. I agree you can get to the precursor compounds based on tracer ions indicative of chemistry from a particular precursor, but what in the PCA analysis maps to particular mechanistic pathways? Do the authors envision additional mappings in the PCA plots within the precursor groups to show oxidation from one oxidant versus another, for example, which would be evident in unique tracer compounds from associated pathways?

Author Response: The referee is correct that this requires some clarification and we believe that the additional section regarding toluene/NO “mapping”, as discussed above under point 14), goes some way towards achieving this.

In a sense this sentence was pointing to the questions of whether, if you have similar product ions, would the PCA more closely cluster the different mechanistic pathways? The referee is correct in the summation that we envisage additional mappings. The text has been clarified as follows to point to this:

“Indeed, a major potential strength of the data analysis methodology described here, could lie in the decoding of mechanisms into pathways (i.e. separation within a group on account of different underlying chemistry) and consequently linking chemical pathways to precursor compounds.”

Furthermore, following similar comments made by referee 1, a small additional subsection (5.2) was added on page 1676, after line 2, (see Reply to Referee 1, first comment):

“5.2 Mapping within a class...”

Technical Comments:
1) P. 1659, Experiment design: It is not clear in Table 1 which experiments were conducted in which chamber.

**Author Response:** Table 1 and the accompanying caption have been altered to indicate which experiments were conducted in which chamber.

2) P. 1661, lines 14-16: Since it is not clear which experiments were conducted in which chamber, as in the proceeding comment, please also make clear here which experiments actually had particle-phase data. Later inspection of Figures 7-8 warrants the reader questioning of why the mesocosm experiments and isoprene experiments are not included in the analysis of the particle-phase data.

**Author Response:** As Table 1 and the accompanying caption have now been altered to indicate which experiments were conducted in which chamber (as per Referee 2, Technical Comment 1), the reader can now easily see which experiments were conducted in the MAC and therefore which had accompanying particle-phase data.

No data was available for certain mesocosm experiments, owing to the low detection limits imposed by the experimental set up and for those experiments with particle-phase data, insufficient repeat experiments were available to use in the statistical analysis presented here.

3) Figure 2 would benefit from caption description of the mass spectral ions populating the figure.

**Author Response:** Figure 2 caption has been altered accordingly to include a description of the major mass spectral ions populating the figure.

4) P. 1669, line 13: Change “been” to “be”

**Author Response:** The text has been altered accordingly.

5) For aid in interpretation of the ions mapped onto figures 2 and 7a, it may be
helpful to include a table that lists all ions (LC-MS/MS, CIR-Tof-MS, and AMS), their chemical assignment, and precursor type, so the information is more readily available than filtering through the text for these ion assignments.

**Author Response:** The requested table has been produced (now Table 3) and inserted and the relevant main text and captions have been altered accordingly.

6) Figures 7b and 8b: It would be more helpful if the color scheme used for the cyclic monoterpenes was consistent with that used in Figure 5.

**Author Response:** We thank the referee for pointing this out, but we would prefer to leave the dendrogram colors as there are, as they represent the best combination for the clearest description of the data, i.e. the dendrogram colors were chosen specifically to optimize clarity of view within the figure.
Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: chemometric dimension reduction techniques for the deconvolution of complex atmospheric datasets


1. Air Environment Research Group, School of Environment and Technology, University of Brighton, BN2 4GJ, UK
2. Department of Chemistry, University of Leicester, Leicester, LE1 7RH, UK
3. Wolfson Atmospheric Chemistry Laboratories, Department of Chemistry, University of York, York, YO10 5DD, UK
4. School of Earth, Atmospheric and Environmental Sciences, University of Manchester, M13 9PL, UK
5. National Centre for Atmospheric Science, University of Manchester, M13 9PL, UK
6. National Centre for Atmospheric Science, University of York, York, YO10 5DD, UK
7. Atmospheric Chemistry Services, Okehampton, Devon, EX20 1FB, UK
8. School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK
9. Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK
10. Birmingham Institute of Forest Research, University of Birmingham, B15 2TT, UK

*For correspondence:

K. P. Wyche

Air Environment Research Group, University of Brighton, Brighton, BN2 4GJ, UK.
Tel: +44 (0)1273 643306
Email: k.p.wyche@brighton.ac.uk
Abstract

Highly non-linear dynamical systems, such as those found in atmospheric chemistry, necessitate hierarchical approaches to both experiment and modeling in order, ultimately, to identify and achieve fundamental process-understanding in the full open system. Atmospheric simulation chambers comprise an intermediate in complexity, between a classical laboratory experiment and the full, ambient system. As such, they can generate large volumes of difficult-to-interpret data. Here we describe and implement a chemometric dimension reduction methodology for the deconvolution and interpretation of complex gas- and particle-phase composition spectra. The methodology comprises principal component analysis (PCA), hierarchical cluster analysis (HCA) and positive least squares-discriminant analysis (PLS-DA). These methods are, for the first time, applied to simultaneous gas- and particle-phase composition data obtained from a comprehensive series of environmental simulation chamber experiments focused on biogenic volatile organic compound (BVOC) photooxidation and associated secondary organic aerosol (SOA) formation. We primarily investigated the biogenic SOA precursors isoprene, α-pinene, limonene, myrcene, linalool and β-caryophyllene. The chemometric analysis is used to classify the oxidation systems and resultant SOA according to the controlling chemistry and the products formed. Results show that “model” biogenic oxidative systems can be successfully separated and classified according to their gaseous oxidation products. Furthermore, a holistic view of results across both the gas- and particle-phases shows the different SOA formation chemistry, initiating in the gas-phase, proceeding to govern the differences between the various BVOC SOA compositions. The results obtained are used to describe the particle composition in the context of the oxidized gas-phase matrix. An extension of the technique, which incorporates into the statistical models data from anthropogenic (i.e. toluene) oxidation and “more realistic” plant mesocosm systems, demonstrates that such an ensemble of chemometric mapping has the potential to be used for the classification of more complex spectra of unknown origin. More specifically, the addition of mesocosm data from fig and birch tree experiments shows that isoprene and monoterpene emitting sources, respectively, can be mapped onto the statistical model structure and their positional vectors can provide insight into their biological
The potential to extend the methodology to the analysis of ambient air is discussed using results obtained from a zero-dimensional box model incorporating mechanistic data obtained from the Master Chemical Mechanism (MCMv3.2). Such an extension to analysing ambient air would prove a powerful asset in assisting with the identification of SOA sources and the elucidation of the underlying chemical mechanisms involved.

Keywords: Volatile organic compounds, secondary organic aerosol, environmental simulation chamber, photooxidation, principal component analysis, cluster analysis, positive least-squares discriminant analysis, chemometrics, terpenes, mesocosm
1. Introduction

Biogenic Volatile Organic Compounds (BVOCs) are ubiquitous in the global troposphere, being emitted primarily from terrestrial plant life (Kanakidou et al., 2005). It is estimated that the total annual emission rate of all (non-methane) BVOCs is roughly ten times that of all anthropogenic volatile organic compounds, being around 750 Tg C yr\(^{-1}\) (Sindelarova et al., 2014). With the exception of methane, the most dominant species of BVOCs in terms of emission strength, reactivity and their impact upon the atmosphere, are terpenes (Reinnig et al., 2008) a subdivision of BVOCs that primarily comprise the hemiterpene, isoprene (C\(_5\)), monoterpenes (C\(_{10}\)) and sesquiterpenes (C\(_{15}\)) (e.g. Atkinson and Arey, 2003a; Kanakidou et al., 2005).

Within the troposphere terpenes are able to react with OH, O\(_3\) and NO\(_3\) at appreciable rates (e.g. Calvert et al., 2000; Koch et al., 2000; Fantechi et al., 2002; Capouet et al., 2004; Kroll et al., 2006) such that their atmospheric lifetimes are in the order of minutes – hours (e.g. Calogirou et al., 1999). Because of their large emission rates and high reactivities, terpenes have a strong impact upon the chemistry of the troposphere at the local, regional and global scales (e.g. Jaoui and Kamens, 2001; Paulot et al., 2012; Surratt, 2013). For instance, terpenes have high photochemical ozone creation potentials (Derwent et al., 2007) and extensive photochemical oxidation pathways that lead to the production of a complex array of oxygenated and nitrated products, some of which are able to form secondary organic aerosol (SOA) (e.g. Calvert et al., 2000; Capouet et al., 2004; Jenkin, 2004; Baltensperger et al., 2008; Kanakidou et al., 2005; Surratt et al., 2006; Kroll and Seinfeld, 2008; Hallquist et al., 2009).

Aerosol particles are natural components of the Earth’s atmosphere responsible for a range of well-documented impacts, ranging from visibility impairment on the local scale to climate change, with suspended particles being able to perturb the Earth’s radiative budget via both direct and indirect mechanisms (IPCC, 2007). Furthermore, fine airborne particles have been shown to have numerous detrimental effects on...
human health, particularly in vulnerable members of the population (Harrison et al., 2010; Heal et al., 2012).

Biogenic SOA (BSOA) has been estimated to account for a significant fraction of total global SOA. Modelling studies suggest the annual global production rate of BSOA is of the order 16.4 Tg Yr\(^{-1}\) (Henze and Seinfeld, 2006). However, despite its importance and the significant amount of investigation conducted upon it, the formation mechanisms and chemical composition of BSOA are still not well characterised (e.g. Librando and Tringali, 2005; Wang et al., 2013). Indeed under certain conditions as much as 80 – 90 % of analysed SOA mass is unknown (Limbeck et al., 2003; Kalberer et al., 2006). In particular, there remains a significant lack of information regarding the composition and evolution of the complex organic gas-phase matrix during aerosol formation, and its linkage to SOA (Kroll et al., 2005; Librando and Tringali, 2005). Indeed, in the many studies conducted on BSOA, very few oxidation products of the precursor are routinely identified and reported.

The chemistry of the atmospheric system is highly non-linear and can be studied by experiments ranging from highly controlled laboratory studies of a single process, to field studies of the whole complex system. A significant proportion of the findings gained regarding SOA over the last decade and more have come from atmospheric simulation chamber experiments, intermediate in complexity between classical single-process experiments and the fully open system (for various different chamber systems and VOC precursors, see for example, Pandis et al., 1991; Odum et al., 1996; Hoffmann et al., 1997; Griffin et al., 1999; Glasius et al., 2000; Cocker et al., 2001; Jaoui and Kamens, 2003; Kleindienst et al., 2004; Presto et al., 2005; Bloss et al., 2005; Rohrer et al., 2005; Ng et al., 2006, 2007; Dommenn et al., 2006; Surrat et al., 2006; Grieshop et al., 2007; Chan et al., 2007; Wyche et al., 2009; Hildebrandt et al., 2009; Rickard et al., 2010; Camredon et al., 2010; Chhabra et al., 2011; Hennigan et al., 2011; Jenkin et al., 2012). Chamber experiments produce a large amount of data, the interpretation of which can often be highly complex and time consuming even though the set-up of the chamber constrains the complexity to a large degree.
In the current “big data” age, advanced monitoring techniques are producing increasingly larger, more complex and detailed data sets. Modern chamber experiments, monitored by state-of-the-art gas- and particle-phase instrumentation, often yield so much data that often only a fraction is subsequently used in a given analysis. For example, during a typical six-hour environmental simulation chamber experiment, VOC monitoring chemical ionisation reaction time-of-flight mass spectrometry (CIR-TOF-MS) will produce roughly $1.1 \times 10^7$ data points. In order to keep pace with instrument development and maximise the information extracted from sometimes-complex experiments, it is crucial that we advance our data analysis methods and introduce new data mining techniques.

The work reported here focuses on detailed organic gas-phase and particle-phase composition data, recorded during SOA atmospheric simulation chamber experiments, using CIR-TOF-MS and liquid chromatography-ion trap mass spectrometry (LC-MS/MS), respectively, as well as broad (i.e. generic composition “type”; oxygenated organic aerosol, nitrated, sulphated etc) aerosol composition data, recorded by compact time-of-flight aerosol mass spectrometry (cTOF-AMS).

The goal of this paper is to demonstrate and evaluate the application of an ensemble reductive chemometric methodology for these comprehensive oxidation chamber datasets, to be used as a model framework to map chemical reactivity from mesocosm systems, thus providing a link from model systems to more “real” mixtures of organics. The intermediate complexity offered by simulation chamber experiments makes them an ideal test-bed for the methodology. Application of the methodology to resultant particle-phase data also aims to provide a level of particle composition classification in the context of gas-phase oxidation.

Similar approaches using statistical analyses have been recently applied to both detailed and broad ambient aerosol composition data (e.g. (Heringa et al., 2012; Paglione et al., 2014)), particularly in the context of source apportionment (e.g. (Alier et al., 2013)). Different methods have been attempted by several groups to deconvolve organic aerosol spectra measured by the Aerosol Mass Spectrometer (AMS) in particular (e.g. Zhang et al., 2005, 2007; Marcolli et al., 2006; Lanz et al.)
Zhang et al. (2005) applied a custom principal component analysis (CPCA) method to extract two distinct sources of organic aerosols in an urban environment using linear decomposition of AMS spectra and later applied a Multiple Component Analysis technique (MCA, an expanded version of the CPCA) to separate more than two factors in datasets from 37 field campaigns in the Northern Hemisphere (Zhang et al., 2007). Marcolli et al. (2006) applied a hierarchical cluster analysis method to an ambient AMS data set, and reported clusters representing biogenic VOC oxidation products, highly oxidised organic aerosols and other small categories. Receptor modelling techniques such as Positive Matrix Factorization (PMF) employ similar multivariate statistical methods in order to deconvolve a time series of simultaneous measurements into a set of factors and their time-dependent concentrations (Paatero and Tapper, 1994; Paatero, 1997). Depending on their specific chemical and temporal characteristics, these factors may then be related to emission sources, chemical composition and atmospheric processing. For example, Lanz et al. (2007) and Ulbrich et al., (2009) applied PMF to the organic fraction of AMS datasets and were able to conduct source apportionment analysis identifying factors contributing to the composition of organic aerosol at urban locations. Slowik et al. (2010), combined both particle-phase AMS and gas-phase proton transfer reaction mass spectrometry data for the PMF analysis of urban air, and were able to successfully obtain “regional transport, local traffic, charbroiling and oxidative process” factors. By combining the two datasets, Slowik and colleagues were able to acquire more in-depth information regarding the urban atmosphere than could be derived from the analysis of each of the sets of measurements on their own.

Because receptor models require no a priori knowledge of meteorological conditions or emission inventories, they are ideal for use in locations where emission inventories are poorly characterised or highly complicated (e.g. urban areas), or where atmospheric processing plays a major role. However, because all of the values in the profiles and contributions are constrained to be positive, the PMF model can have an arbitrary number of factors and the user must select the “best” solution that
explains the data. This subjective step of PMF analysis relies greatly on the judgment and skill of the user.

The central methodology employed, is based around the application of principal component analysis (PCA), hierarchical cluster analysis (HCA) and positive least squares-discriminate analysis (PLS-DA) of single-precursor oxidant chemistry in environmental simulation chambers. Colloquially, we can describe these three approaches as providing dimensions along which the data are separable (PCA), tests of relatedness (HCA) and checks for false-positives (PLS-DA).

Such dimension reduction techniques can be very powerful when used in chemometrics, enabling large and often complex datasets to be rendered down to a relatively small set of pattern-vectors provide an optimal description of the variance of the data (Jackson, 1980; Sousa et al., 2013; Kuppusami et al., 2014). Unlike other statistical techniques such as PMF, the ensemble methodology presented here does not require the use of additional external databases (comprising information regarding different environments/reference spectra), is simpler to use and less labour intensive, and places less importance on user skill in the production of accurate and meaningful results. Moreover, the primary focus of techniques such as PMF is on source identification/separation, whereas here the focus is placed on compositional isolation.

The analysis conducted in this work shows that “model” biogenic oxidative systems can be clearly separated and classified according to their gaseous oxidation products, i.e. isoprene from β-caryophyllene from non-cyclic monoterpenes and cyclic monoterpenes. The addition of equivalent mesocosm data from fig and birch tree experiments shows that large isoprene and large monoterpane emitting sources, respectively, can be mapped onto the statistical model structure and their positional vectors can provide insight into the oxidative chemistry at play. The analysis is extended to particle-phase data to show further classifications of model systems based on both broad and detailed SOA composition measurements.
The methodology described and the results presented (supported by findings obtained from zero-dimensional box modelling), indicate that there is some potential that the approach could ultimately provide the foundations for a framework onto which it would be possible to map the chemistry and oxidation characteristics of ambient air measurements. This could in turn allow “pattern” typing and source origination for certain complex air matrices and provide a snapshot of the reactive chemistry at work, lending insight into the type of chemistry driving the compositional change of the contemporary atmosphere. There are similarities between this approach to discovery science in the atmosphere and metabolomics strategies in biology (e.g. (Sousa et al., 2013; Kuppusami et al., 2014)).

2. Experimental details

2.1 Choice of precursors

Six different BVOCs and one anthropogenic VOC were chosen for analysis. The target compounds, their structures and reaction rate constants with respect to OH and O₃ are given in Table 1. The BVOCs were chosen according to their atmospheric prevalence, structure and contrasting photooxidative reaction pathways; all have previously been shown to form SOA under simulation chamber conditions (e.g. Hoffmann et al., 1997; Griffin et al., 1999; Glasius et al., 2000; Jaoui and Kamens, 2003; Presto et al., 2005; Ng et al., 2006; Surratt et al., 2006; Dommen et al., 2006; Lee et al., 2006; Hallquist et al., 2009; Alfarra et al., 2013, and references therein).

Isoprene is a C₅ diene that accounts for around 62 % (~ 594 Tg yr⁻¹) of total annual non-methane BVOC emissions (Sindelarova et al., 2014). After isoprene, monoterpenes (C₁₀H₁₆) have the next largest annual emission rate, they account for around 11 % (~ 95 Tg yr⁻¹) of total annual non-methane BVOC emissions (Sindelarova et al., 2014). α-pinene and limonene were chosen for analysis here alongside isoprene, the former acting as a model system to represent bicyclic monoterpenes, the later to represent monocyclic diene terpenes. In this work, α-pinene and limonene together generically represent (and are referred to hereafter as) “cyclic” monoterpenes (i.e. monoterpenes that contain one six-member carbon ring). In order to explore the chemistry of non-cyclic monoterpenes, myrcene, an acyclic triene monoterpane, was also included, as was the structurally similar acyclic diene
OVOC, linalool. In this work, myrcene and linalool together generically represent (and are referred to hereafter as) “straight chain” monoterpenes/BVOCs (note: linalool is not technically a monoterpene, but does contain the same carbon backbone as myrcene, consequently it is expected to exhibit similar photooxidative chemistry). Finally, β-caryophyllene was included to represent sesquiterpenes, which have annual emissions of the order 20 Tg yr\(^{-1}\) (Sindelarova et al., 2014). In order to test the ability of the methodology to distinguish between biogenic and anthropogenic systems, toluene was also included. Toluene is often used as a model system to act as a proxy for aromatic species in general (Bloss et al., 2005). For contrasting plant mesocosm systems, *Ficus benjamina* and *Ficus cyathistipula* (fig) and *Betula pendula* (birch) species were chosen to represent tropical rainforest and European environs, respectively.

In general, the VOC precursors employed have roughly similar reaction rate constants with respect to OH and O\(_3\), e.g. limonene, myrcene, linalool and β-caryophyllene all have atmospheric lifetimes with respect to OH of the order 40 – 50 minutes (Alfarra et al., 2013; Atkinson and Arey, 2003b). β-caryophyllene has the shortest lifetime with respect to O\(_3\) (ca. 2 minutes) and isoprene and α-pinene have the longest lifetimes with respect to both OH and O\(_3\), e.g. isoprene and α-pinene have atmospheric lifetimes with respect to OH of the order 1.4 – 2.7 hours (Alfarra et al., 2013; Atkinson and Arey, 2003b). In order to ensure the various systems had progressed sufficiently down their respective photooxidative reaction pathways, the experiment duration was set to be sufficiently long that the majority of the precursor had been consumed by the conclusion of the experiment.

### 2.2 Chamber Infrastructure

Experiments were carried out across three different European environmental simulation chamber facilities over a number of separate campaigns. The chambers used, included (1) The University of Manchester Aerosol Chamber (MAC), UK (Alfarra et al., 2012); (2) The European Photoreactor (EUPHORE), ES (Becker, 1996) and (3) The Paul Scherrer Institut Smog Chamber (PSISC), CH (Paulsen et al., 2005). A brief technical description of each facility is given in Table 2.
2.3 Experiment Design

Table 2 provides a summary of the experiments conducted, which can be divided into three separate categories, (1) photooxidation, indoor chamber (Wyche et al., 2009; Alfarra et al., 2012; Alfarra et al., 2013), (2) photooxidation, outdoor chamber (Bloss et al., 2005; Camredon et al., 2010) and (3) mesocosm photooxidation, indoor chamber (Wyche et al., 2014). In each case the reaction chamber matrix comprised a temperature (T = 292 – 299 K) and humidity (49 – 84 % for photooxidation, indoor chamber and < 2 – 6 % for photooxidation, outdoor chamber) controlled synthetic air mixture. For all experiments the chamber air matrix also contained a pre-defined initial quantity of NO and NO₂ (VOC/NO₂ ratios in the range 0.6 – 20, but typical ~ 2). The VOC precursor was introduced into the reaction chamber in liquid form via a heated inlet. In the case of the mesocosm photooxidation experiments, a known volume of air containing the precursor VOCs was transferred to the reaction chamber from a separate, illuminated plant chamber, which contained several tree specimens. For the indoor chamber systems, the experiments were initiated, after introduction of all reactants, by the switching on of artificial lights. For the outdoor chamber systems, the opening of the chamber cupola marked the start of the experiment. Experiments were typically run for 4 – 6 hours.

2.4 Instrumentation

CIR-TOF-MS was used to make real-time (i.e. 1 minute) measurements of the complex distribution of volatile organic compounds (ΣVOC, i.e. the sum of VOCs, oxygenated VOCs – OVOCs and nitrated VOCs – NVOCs) produced in the gas-phase during oxidation of each parent compound. In brief, the CIR-TOF-MS comprises a temperature controlled (T = 40 °C) ion source/drift cell assembly coupled to an orthogonal time-of-flight mass spectrometer equipped with a reflectron array (Kore Technology, UK). Proton Transfer Reaction (PTR) from hydronium (H₃O⁺) and hydrated hydronium (H₂O⁺(H₂O)ₙ) was employed as the ionisation technique during all experiments (Jenkin et al., 2012). Further details regarding the CIR-TOF-MS can be found in Blake et al. (Blake et al., 2004) and Wyche et al. (Wyche et al., 2007).
Aerosol samples were collected on 47 mm quartz fibre filters at the end of certain experiments and the water-soluble organic content was extracted for analysis using LC-MS/MS. Reversed phase LC separation was achieved using an HP 1100 LC system equipped with an Eclipse ODS-C18 column with 5 µm particle size (Agilent, 4.6 mm × 150 mm). Mass spectrometric analysis was performed in negative ionisation mode using an HCT-Plus ion trap mass spectrometer with electrospray ionisation (Bruker Daltonics GmbH). Further details can be found in Hamilton et al. (Hamilton et al., 2003).

For several experiments, real-time broad chemical characterisation of the SOA was made using a cTOF-AMS (Aerodyne Research Inc., USA). The cTOF-AMS was operated in standard configuration, taking both mass spectrum (MS) and particle time-of-flight (PTOF) data; it was calibrated for ionisation efficiency using 350 nm monodisperse ammonium nitrate particles, the vapouriser was set to ~ 600 °C and a collection efficiency value of unity was applied (Alfarra et al., 2006). For further details, refer to Drewnick et al. (Drewnick et al., 2005) and Canagaratna et al. (Canagaratna et al., 2007).

Each chamber was additionally instrumented with on-line chemiluminescence/photolytic NO₂ NO, analysers, UV photometric O₃ detectors, and scanning mobility particle sizers and condensation particle counters for aerosol size and number concentration, as well as temperature, pressure and humidity monitors. For full details regarding the various instrument suites employed at each chamber see Alfarra et al. (Alfarra et al., 2012), Paulsen et al. (Paulsen et al., 2005), Camredon et al. (Camredon et al., 2010) and references therein.

Filter and cTOF-AMS data were collected only during photooxidation experiments conducted at the MAC. Repeat experiments conducted at the MAC were carried out under similar starting conditions (e.g. VOC/NOₓ ratio (Alfarra et al., 2013)).

2.5 Model construction
In order to aid analysis, the composition and evolution of the gas-phase components of the α-pinene chamber system were simulated using a chamber optimised photochemical box model incorporating the comprehensive α-pinene atmospheric oxidation scheme extracted from the Master Chemical Mechanism website (Jenkin et al., 1997; Saunders et al., 2003; Jenkin et al., 2012; http://mcm.leeds.ac.uk/MCM). The α-pinene mechanism employed (along with an appropriate inorganic reaction scheme) contained approximately 313 species and 942 different reactions. The box model employed also incorporated a series of “chamber specific” auxiliary reactions adapted from Bloss et al. (Bloss et al., 2005), Zador et al. (Zador et al., 2006) and Metzger et al., (Metzger et al., 2008) in order to take into account background chamber reactivity. Photolysis rates were parameterised for the PSI chamber and constrained using measured values of (j(NO₂)). All simulations were run at 295 K and 50 % relative humidity. NO, NO₂, HONO and α-pinene were either initialised or constrained, depending on the scenario investigated. For further details see Rickard et al. (Rickard et al., 2010).

3. Data Analysis
3.1 Data Processing
All CIR-TOF-MS data were recorded at a time resolution of 1-minute. In order to remove the time dimension and simultaneously increase detection limit, the individual mass spectra were integrated over the entire experiment; as such no account is taken of overall reaction time in the CIR-TOF-MS analysis. Removing the time dimension acts to reduce the dimensionality of the data, whilst maintaining the central characteristic spectral fingerprints produced by the photooxidation process. On average across all experiments studied, 98 % of the precursor had been consumed by the conclusion of the experiment; hence it is assumed that sufficient reaction took place in each instance to provide summed-normalised mass spectra that fully capture first- and higher-generation product formation.

The resultant summed spectra were normalised to 10⁶ primary reagent ion counts (i.e. Σ(H₂O⁺ + H₂O⁺(H₂O)ₙ)). Similarly normalised background spectra (recorded prior to injection of the precursor) were then subtracted from the summed-and-
normalised experiment spectra. The 65 < m/z < 255 channels of the background
removed spectra were extracted to comprise the region of interest. These ions tend
to carry the most analyte-specific information, with lower m/z features tending to
comprise either generic fragment ions that provide little chemical information (Blake
et al., 2006) and/or small compounds emitted from illuminated chamber walls (e.g.
Bloss et al., 2005; Zador et al., 2006; Metzger et al., 2008). These extracted data
were refined further by the application of a Mann-Whitney test (see Statistical
Analysis for details), leaving residual spectra that comprised only the integrated-
over-time signals corresponding to the VOC precursor and any reactive intermediate
and product VOCs formed within the chamber during the experiment. Finally, the
signal counts (in units of normalised counts per second; ncps) in each mass channel
of the residuals, were expressed as a percentage of the total ion count in the refined
region of interest.

The LC-MS/MS signal intensity data for the region 51 < m/z < 599 were extracted for
analysis. For the AMS data, a 10-minute average was produced at 4 hours after
lights on (roughly around the time when SOA mass had reached a peak and towards
to the end of the experiment) and the region 40 < m/z < 150 (again the region
carrying the most information; Alfarra et al., 2006) was extracted. Similar to the gas-
phase data sets, the LC-MS/MS and AMS data were filtered using a Mann-Whitney
test. Finally, for each data set all signal counts were expressed as a percentage of
the total ion count in the respective m/z region of interest.

3.2 Statistical Analysis
Before any multivariate analysis was conducted, the processed CIR-TOF-MS, LC-
MS/MS and AMS spectra were first filtered to remove unwanted data that were
deemed to not be statistically significant. In order to do this, the mass spectra were
initially grouped by structure of the precursor employed, giving seven separate
groups for the CIR-TOF-MS data and three groups (owing to the smaller number of
precursor species investigated) for the LC-MS/MS and AMS data, respectively. A
two-sided Mann-Whitney test was then used to assess whether signals reported in
individual mass channels were significantly different from the corresponding signals
measured during a blank experiment. SPSS V20 (IBM, USA) was used for the
analysis. A p value of < 0.05 was considered statistically significant. The final
summed-normalised and filtered spectra were then subjected to a series of
multivariate statistical analysis techniques in order to probe the underlying chemical
information. PLS-Toolbox (Eigenvector Research Inc., USA) operated in MatLab
(Mathworks, USA; PLS-Tool Box) was used for the analysis.

To begin with, to reduce the data and identify similarities between the precursor
oxidation systems, a PCA was conducted on the BVOC dataset and the model
generated was then employed to map the reactivity of fig and birch tree mesocosm
systems and to investigate the fit of a typical anthropogenic system (toluene) into
the PCA space (both introduced into the model as test datasets). An unsupervised
pattern recognition, hierarchical cluster analysis was also conducted on the data and
a dendrogram produced to test relatedness, support the PCA and help interpret the
precursor class separations achieved. The dendrogram was constructed using PCA
scores, the centroid method and Mahalanobis distance coefficients. Finally, a
supervised pattern recognition PLS-DA analysis was employed as a check for false-
positives and as a quantitative classification tool to test the effectiveness of
classification of the various systems in the model.

For the superposition of “classification” confidence limits onto the results of the PCA
and HCA and for classification discrimination in the PLS-DA, prior to analysis the
experiments were grouped according to the structure of the precursor investigated.
Group 1 = isoprene (hemiterpene) and group 2 = α-pinene and limonene (both cyclic
monoterpenes with an endocyclic double bond). Although limonene also has an
exocyclic double bond in a side chain, we justify this classification on account of the
endocyclic double bond in limonene being much more reactive towards ozone and
slightly more reactive towards OH (Calvert et al., 2000). Group 3 = β-caryophyllene
(sesquiterpene) and group 4 = myrcene (straight chain monoterpane) and linalool
(straight chain OVOC). Strictly speaking, linalool is an OVOC (structure C_{10}H_{18}O) and
not a monoterpane (structure C_{10}H_{16}), however we justify this grouping on account
of both myrcene and linalool comprising primary BVOCs (often co-emitted; (Bouvier-
Brown et al., 2009; Kim et al., 2010; Wyche et al., 2014)) with certain structural similarities.

4. Results

4.1 Experiment overview

The temporal evolution of various key gas-phase (a) and particle-phase (b) parameters measured during a typical photooxidation experiment, are shown in Figure 1 in order to provide background context. In this instance the precursor was myrcene and the facility employed was the MAC. Full details describing the underlying chemical and physical mechanisms at play within such experiments can be found elsewhere (e.g. Larsen et al., 2001; Bloss et al., 2005; Paulsen et al., 2005; Surratt et al., 2006 and 2010; Wyche et al., 2009, 2014; Camredon et al., 2010; Rickard et al., 2010; Eddingsaas et al., 2012b; Hamilton et al., 2011; Jenkin et al., 2012; Alfarra et al., 2012, 2013; and references therein).

4.2 Mapping gas-phase composition

Of the 191 different mass channels extracted from the CIR-TOF-MS data for analysis (i.e. 65 < m/z < 255), the Mann-Whitney test identified 151 as significant for one or more of the terpene precursor groups tested. These data were subsequently subjected to PCA. From inspection of the Eigenvalues derived, four principal components (PCs) were selected for analysis, which collectively accounted for 96% of the variance within the data, with PCs 1 and 2 accounting for the vast majority, i.e. 63 and 18 %, respectively. This step, therefore, reduced the temporal traces of 191 mass-spectrum peaks to 4 composite and orthogonal dimensions.

Figure 2 shows a loadings bi-plot of PC2 vs. PC1. It is clear from Figure 2, that the model is able to successfully separate the four different classes of biogenic systems investigated. β-caryophyllene mass spectra are grouped in the upper left-hand quadrant of Figure 2, the monoterpenes in the lower left-hand quadrant and isoprene to the centre right. Moreover, the principal component analysis is able to distinguish between the cyclic monoterpane experiments of limonene and α-pinene (grouped into one class), and the straight chain monoterpane experiments of...
myrcene and linalool (grouped into a second class), albeit with the latter having a
greater spread in confidence.

The m/z loadings of the PCA allow us to understand how the spectral fingerprints of
the different terpene oxidation systems are grouped/-separated by the PCA model.
The first set of ions that contribute to separation of the different terpene systems
comprises the protonated parent ions (MH') of the precursors themselves (and
major fragments thereof), i.e. m/z 69 for isoprene, 137 (and fragment 81) for all
monoterpenes (regardless of structure) and 205 for β-caryophyllene. Important
contributions are to be expected from the respective parent-ions (being the basis for
the use of chemical-ionisation mass spectrometry as an analyser
(Blake et al., 2009)). Our purpose here goes beyond identification of precursor and
intermediate VOCs to an interpretation of reaction pathways in complex mixtures
and potential linkages to SOA. In doing this, a certain amount of disambiguation of
isobaric compounds becomes possible; indeed, as discussed in more detail below,
Figure 2 clearly shows separation between cyclic and non-cyclic monoterpene
oxidation groups, both of which have precursors of molecular weight (MW) 136 g
mol⁻¹. Note, for clarity within Figure 2, the scale has been set to show the bulk of the
data, hence precursor parent ions and m/z 71 are not shown.

Moving past the precursors into the detailed chemical information provided by the
oxidation products formed within the chamber, we can see from Figure 2 that
amongst others, m/z 71 (methyl vinyl ketone and methacrolein), 75 (hydroxy
acetone), 83 (methyl furan) and 87 (C₄-hydroxycarbonyls/methacrylic acid) all
contribute to separation of the isoprene group, and m/z 237 (β-caryophyllene
aldehyde) and 235 and 253 (β-caryophyllene secondary ozonide and isomers
thereof) to that of the β-caryophyllene group. The monoterpene groupings are
influenced by the presence of m/z 107, 151 and 169 (primary aldehydes-
piononaldehyde and limononaldehyde) and 139 (primary ketone- limonaketone)
ions in their mass spectra. Helping to separate the straight chain from cyclic
monoterpenes are m/z 95 and 93, relatively dominant features in both the myrcene
and linalool spectra (relative abundance 10 – 24 % for m/z 93). m/z 93 has
previously been identified as a major fragment ion of first generation myrcene and linalool products 4-vinyl-4-pentenal and 4-hydroxy-4-methyl-5-hexen-1-al, respectively (Shu et al., 1997; Lee et al., 2006). A list of major ions contributing to the separation of spectra in statistical space is given in Table 3 along with potential identities and precursors. It is worthy of note here that these ions and the overall fragmentation patterns observed in this study are largely in-line with those reported by Lee et al. (2006), in their comprehensive PTR-MS analysis of a wide range of BVOC precursors and their associated oxidation products.

4.3 Implementation of the model to classify mesocosm data

Having employed the terpene data as a training set to construct a PCA model, a test set of mesocosm data was introduced in order to investigate the ability of the model to map the classification of more complex biogenic mixtures. In this instance the mesocosm test set comprised two birch tree and two fig tree photooxidation experiments, containing a more complex and “realistic” mixture of various different VOCs (Wyche et al., 2014). The resultant scores plot is shown in Figure 3.

Figure 3 demonstrates that the model can successfully distinguish between the two different types of mesocosm systems. Moreover, the model correctly classifies the mesocosm systems within the PCA space, with the birch trees (which primarily emit monoterpenes and only small quantities of isoprene; (Wyche et al., 2014)) grouped with the single precursor monoterpene cluster, and the fig trees (which primarily emit isoprene and camphor and only a small amount of monoterpenes; (Wyche et al., 2014)) grouped between the monoterpene and isoprene clusters. Investigation of the mesocosm mass spectra and PCA loadings shows that mass channels 137, 139, 107, 95, 93, 81 and 71 are amongst features important in classifying the birch tree systems, with the relatively strong presence of m/z 93 suggesting the emission of noncyclic as well as cyclic monoterpenes from the birch trees. This was confirmed by cross-reference with GC-MS analysis, which showed that the acyclic monoterpene, ocimene, was the third most abundant monoterpene present in the birch tree emissions (Wyche et al., 2014). For the fig tree systems, mass channels 153, 81, 73, 71 and 69 are key for classification, with the presence of small quantities of camphor
(m/z 153) and monoterpenes (m/z 81) causing the group to undergo a lateral shift in the PCA space, along PC1 away from the single precursor isoprene cluster.

As a further test of the technique to distinguish between and to classify VOCs and the matrix of oxidized organic compounds that may derive from their atmospheric chemistry, test data from an anthropogenic system was introduced into the model. In this instance, the toluene photooxidation system was employed. Toluene is an important pollutant in urban environments, originating from vehicle exhausts and fuel evaporation; furthermore it represents a model mono-aromatic, SOA precursor system (e.g. (Bloss et al., 2005)). As can be seen from the resultant scores plot in Figure 4, the model is also able to discriminate the anthropogenic system from those of biogenic origin. Besides the protonated toluene parent ion, those ions contributing to the positioning of the toluene cluster within the PCA space, include the protonated parent ions m/z 109 and 107, i.e. the ring retaining primary products benzaldehyde and phenol, respectively; m/z 123, i.e. the ring retaining secondary product, methyl benzoquinone and m/z 99 and 85, i.e. higher generation ring opening products (e.g. 4-oxo-2-pentenal and butenedial, respectively). A brief discussion regarding aromatic BVOCs is provided separately in the supplementary material.

4.4 Cluster analysis and classification

The relationships between the various terpene and mesocosm systems and their groupings with respect to one another can be explored further via the implementation of HCA; Figure 5 gives the dendrogram produced. Inspection of Figure 5 provides further evidence that the various systems in the four classes of terpenes investigated distinctly group together, with overall relatedness < 1 on the (centroid) distance between clusters scale using the Mahalanobis distance measure (Mahalanobis, 1936). Figure 5 shows that the sesquiterpene oxidation system has the most distinct spectral fingerprint (containing distinctive, higher mass oxidation products, e.g. m/z 253) and that the cyclic and straight chain monoterpenes systems appear the most similar (with some common features alongside key, unique precursor/mechanism specific product patterns, e.g. m/z 93 for myrcene and
linalool), grouping together with subclusters of cyclic and noncyclic precursors. The
monoterpane dominated birch tree mesocosm experiments are grouped with the
cyclic monoterpenes and show a close relationship with noncyclic monoterpane
systems. Being dominated by isoprene emissions, yet with some monoterpenes and
camphor present, the fig tree mesocosm experiments group separately but with a
close degree of relation to the single precursor isoprene experiments.

In order to advance our chemometric mapping of biogenic systems beyond PCA and
HCA (which do not consider user supplied a priori observation “class” information)
and to provide a degree of quantification to our analysis, a PLS-DA using six latent
variables (LVs) was conducted on the terpene and mesocosm data. For the PLS-DA,
the experiments were grouped into their respective “classes”, i.e. hemiterpene =
isoprene; cyclic monoterpane = α-pinene and limonene; sesquiterpene = β-
caryophyllene; noncyclic monoterpane = myrcene and linalool; birch trees; fig trees.

Figure 6 shows a plot of the resultant scores on the first three LVs (accounting for ~
85 % of the variance), from which it is clear that the PLS-DA is able to successfully
discriminate between the four terpene classes, and places the monoterpane
dominant birch experiments within the single precursor monoterpane cluster, and
the isoprene dominant fig experiments close to the single precursor isoprene cluster
within the PLS-DA model. **The greater spread in confidence of the noncyclic**
monoterpane group is once again likely to result to some extend from the low
number of repeat experiments employed (i.e. only two each for myrcene and
linalool).

As can be seen from inspection of Table 4, model classification sensitivity and
specificity was high in each instance. Each of the biogenic systems studied were
predicted with 100 % sensitivity (with the exception of birch mesocosm), meaning
that each set of experiments (again, except birch mesocosm) was predicted to fit
perfectly within its class. The relatively low sensitivity obtained for birch mesocosm
(50 %), is most likely a result of the use of only two repeat experiments in the model,
coupled with experiment limitations and ageing trees producing slightly lower
emissions during the final birch mesocosm experiment. All of the systems were
predicted with > 90 % specificity (four of the six with 100 % specificity), indicating that all experiments are highly unlikely to be incorrectly classified.

4.5 Mapping particle-phase composition

In order to explore similar classifications and linkages in the concomitant particle-phase, the PCA, HCA and PLS-DA techniques were also applied to the off-line LC-MS/MS spectra obtained from analysis of filter samples and on-line AMS spectra.

As can be seen from inspection of Figure 7, the detailed LC-MS/MS aerosol spectra produce PCA results somewhat similar to those of the gas-phase CIR-TOF-MS spectra, with distinct clusters of cyclic monoterpenes, straight chain monoterpenes and sesquiterpenes. From inspection of the loadings components of the bi-plot (Figure 7a), we can see that m/z 237 (3-[2,2-dimethyl-4-(1-methylene-4-oxo-butyl)]-cyclobutyl)-propanoic acid), 251 (β-caryophyllonic acid), 255 (4-(2-carboxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid), 267 (β-14-hydroxycaryophyllonic acid and β-10-hydroxycaryophyllonic acid) and 271 (4-(2-[3-hydroperoxy-3-oxopropyl]-3,3-dimethylcyclobutyl)-4-oxobutanoic acid or 4-(2-carboxy-1-hydroxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid), are amongst those ions dominant in classifying the sesquiterpenes. For further details regarding β-caryophyllene oxidation products, see for example Lee et al. (2006), Winterhalter et al. (2009), Hamilton et al. (2011), Chan et al. (2011), Li et al. (2011) and Jenkin et al. (2012) and references therein, and Sect. 5. Of this set of oxidation products, β-caryophyllonic acid is common between the gas- (i.e. m/z 253) and particle- (i.e. m/z 251) phases.

Similarly, those ions (compounds) significant in isolating the cyclic monoterpenes include, m/z 169 (pinalic-3-acid, ketolimononaldehyde and limonic acid), 183 (pinonic acid, limonic acid and 7-hydroxylimononaldehyde) and 185 (pinic acid, limonic acid), of which only those compounds of m/z 169 were observed to be of significant contribution to the gas-phase composition (observed as m/z 171; relative contribution as high as 1 - 5 % during α-pinene experiments). For further details regarding α-pinene and limonene oxidation products, see for example Larsen et al. (2001), Jaoui et al. (2003), Capouet et al. (2004), Jenkin (2004), Jaoui et al. (2006),
Lee et al. (2006), Ng et al. (2006), Camredon et al. (2010) and Hamilton et al. (2011) and references therein. Comparatively little information is available on the speciated composition of myrcene and linalool SOA, however, from Figure 7a it is clear that somewhat larger mass compounds are important in classifying straight chain monoterpenes, e.g. m/z 321 (adduct ion [M-H₂+FA+Na] M = 254 Da; potential formulae - C₁₂H₁₄O₆, six double bond equivalents or C₁₃H₁₈O₅, five double bond equivalents; indicative of oligomer formation), 325, 322 (the C13 peak for the m/z 321 ion), 227 (C₁₀H₁₁O₆), 215 (C₁₀H₁₅O₅) and 199 (C₈H₁₃O₃). Compounds of such high molecular weight were not observed in the concomitant gas-phase spectra. A list of major ions contributing to the separation of spectra in statistical space is given in Table 3 along with potential identities and precursors.

As with the PCA, the dendrogram produced via cluster analysis of the LC-MS/MS particle-phase data gave three distinct clusters (Figure 7b), i.e. cyclic monoterpane, straight chain monoterpane and sesquiterpene. The corresponding PLS-DA analysis reported 100% sensitivity in each case and 100% specificity for all systems except sesquiterpenes (i.e. β-caryophyllene = 83%), suggesting a good level of model classification for the three types of terpene systems studied.

Despite utilising the somewhat destructive electron impact (EI) ionisation technique, the cTOF-AMS produces spectra of sufficient chemical detail such that the PCA and HCA are able to successfully differentiate between the groups of terpenes tested (Figure 8a and b). However, unlike the outputs from the CIR-TOF-MS and LC-MS/MS PCA’s, the cyclic and straight chain monoterpenes in the AMS PCA do not group into two distinct classes, instead they tend to group in their species-specific sub-classes within the upper half of the PCA space. Indeed, the PLS-DA gave 100% sensitivity and specificity for the cyclic monoterpenes and sesquiterpenes, but only 75% sensitivity for the straight chain monoterpenes, suggesting that the model does less well at assigning myrcene and linalool cTOF-AMS spectra to their defined class.

As can be seen from inspection of Figure 8a, α-pinene, limonene and linalool tend in general to cluster towards the upper and right regions of the PCA space, primarily...
owing to the significant presence of m/z 43 and to a lesser extent m/z 44, in their spectra; both ions constituting common fragments observed in AMS of SOA (Alfarra et al., 2006). During such chamber experiments, the m/z 43 peak tends to comprise the CH$_3$CO$^+$ ion, originating from oxidised compounds containing carbonyl functionalities; it is usually representative of freshly oxidised material and semi-volatile oxygenated organic aerosol (SV-OOA; Alfarra et al., 2006)).

From further inspection of the loadings bi-plot (Figure 8a) we see that the four sesquiterpene (β-caryophyllene) experiments cluster towards the lower left hand quadrant, their clustering heavily influenced by the presence of m/z 41 in their spectra as well as m/z 55, 79 and 95. In EI-AMS, m/z 41 comprises the unsaturated C$_3$H$_5$ fragment (Alfarra et al., 2006). As well as being influenced by the m/z 41 ion, the myrcene cluster (situated in the region of both the α-pinene and β-caryophyllene clusters in the PCA space) is also influenced by m/z 44, i.e. most likely the CO$_2$ ion. In this instance m/z 44 would tend to result from low volatility oxygenated organic aerosol (LV-OOA), derived from highly oxidised compounds, including oxo- and dicarboxylic acids (Alfarra et al., 2004; Alfarra et al., 2006). For full details regarding the particle-phase specific experiments conducted at the MAC, see Hamilton et al. (2011), Jenkin et al., (2012) and Alfarra et al. (2012 and 2013).

5. Discussion

5.1 Mapping chemistry

Figure 9 provides a highly simplified overview of the current state of knowledge regarding the atmospheric oxidation of hemi-, sesqui-, cyclic and straight chain mono-terpenes, showing selected key steps and intermediates on route to SOA formation. The mechanisms outlined in Figure 9 underpin the findings reported here and explain how the atmospheric chemistry of the various terpene oxidation systems and their SOA can be chemometrically mapped with respect to one another.

From a review of recent literature and from the summary presented in Figure 9, it can be seen that isoprene can react to form condensable second and higher generation nitrates in the presence of NO$_x$, e.g. C$_4$-hydroxy nitrate peroxy acetyl
nitrate (C4-HN-PAN in Figure 9) (Surratt et al., 2010), as well as condensable OVOCs, e.g. hydroxymethyl-methyl-α-lactone (HMML) (Kjaergaard et al., 2012) and methacrylic acid epoxide (MAE) (Lin et al., 2013), via metharcolein (MACR) and methacryloyl-peroxy nitrate (MPAN). Alternatively, under “low NOx” conditions (e.g. < 1 ppbV) isoprene can react to form condensable second-generation epoxides, e.g. isoprene epoxides (IEPOX), via primary peroxides (ISOPOOH) (Paulot et al., 2009a; Surratt et al., 2006). Such C4 and C5 saturated, low volatility species constitute the monomer building blocks that proceed to form relatively high O:C ratio (nitrated in the presence of NOx and sulphated in the presence of H2SO4) isoprene SOA oligomers (e.g. 2-methyl tetrol dimer O:C = 7:9) (Claeys et al., 2004; Surratt et al., 2006; Surratt et al., 2010; Worton et al., 2013). Consequently, the gas-phase composition under conditions forming isoprene SOA will therefore be dominated by relatively low MW monomer precursors, e.g. MACR (MH+ = m/z 71), isoprene nitrates (ISOPN in Figure 9; MH+ - HNO3 = m/z 85) and MPAN (MH+·H2O - HNO3 = m/z 103) under “high NOx,” conditions (e.g. ~ 10’s – 100’s ppbV; (Paulot et al., 2009b; Surratt et al., 2010; Surratt et al., 2006)), and ISOPOOH and IEPOX (MH+·H2O = m/z 101) under “low NOx,” conditions. For the “high NOx,” isoprene experiments conducted here, besides m/z 71, i.e. MACR (measured together with methyl vinyl ketone), m/z 87, 85, 83 and 75 i.e. (tentatively assigned to be) C4- hydroxycarboxyls/methacrylic acid, ISOPN, C3-hydroxy carbonyls (CSHC in Figure 9)/3-methyl furan (3-MF) and hydroxy acetone, respectively, were significant in classifying the isoprene group; MPAN at the m/z 103 ion was only a minor contributor. It should be noted that in theory, both HMML and MAE (MH+ = m/z 103) may produce fragment ions of m/z 85 (i.e. MH+·H2O) following PTR ionisation, however without further detailed characterisation we are unable at this stage to postulate their fractional contribution to the measured m/z 85 signal.

Depending on the chemistry involved (Figure 9), potential SOA forming monoterpenes products will either be (six-member-) ring retaining (e.g. from reaction with OH) or (six-member-) ring cleaved (e.g. from reaction with OH or O3), producing gas-phase spectra with mid MW C9 and C10 oxygenated (and nitrated in the presence of NOx) products (e.g. (Kamens and Jaoui, 2001; Larsen et al., 2001; Capouet et al.,...
Both (six-member-) ring retaining and (six-member-) ring-opening products have been observed in monoterpene SOA (e.g. (Yu et al., 1999; Larsen et al., 2001; Camredon et al., 2010)), with the latter generally being dominant in terms of abundance (Camredon et al., 2010). Furthermore, (six-member-) ring-opening products are believed to undergo chemistry within the aerosol to form relatively low O:C ratio oligomers (e.g. 10-hydroxy-pinonic acid-pinonic acid dimer, O:C = 7:19) (Gao et al., 2004; Tolocka et al., 2004; Camredon et al., 2010).

OH will react with straight chain monoterpenes, such as myrcene, primarily by addition to either the isolated or the conjugated double bond system. Reaction at the isolated C=C bond can proceed via fragmentation of the carbon backbone, producing acetone and mid MW, unsaturated C7 OVOCs (and/or NVOCs, depending on NOx levels). Reaction at the conjugated double bond system in myrcene would be expected to form formaldehyde in conjunction with either a C9 aldehyde or C9 ketone. Structure activity relationships (SARs) predict that the conjugated double bond system accounts for almost half of the OH reactivity. The conjugated double bond would therefore be expected to have a partial rate coefficient of the order 1 x 10^{-10} (i.e. similar to OH + isoprene) (Atkinson and Arey, 2003b). Consistent with this, the reported yields of acetone and formaldehyde from OH + myrcene are similar (Atkinson and Arey, 2003b), suggesting that the isolated double bond and the conjugated double bond system have comparable OH reactivity, as such we would expect C9 and C7 co-products to be formed in comparable yields. However, with a significant fraction of reactions with OH leading to the loss of three carbon atoms from the parent structure, the straight chain monoterpene gas-phase spectra tend to contain fewer features of MW greater than that of the precursor and more mid MW features. It tends to be these mid MW features, such as m/z 111 and 93 (e.g. 4-vinyl-4-pentenal, MYR 1.2 in Figure 9, MH+ and MH+-H2O, respectively) and 113 and 95 (e.g. 2-methylenepentanedial MH+ and MH+-H2O, respectively) that assist in the classification of the straight chain monoterpene experiments within the statistical space. Besides these ions, m/z 139 (primary myrcene C9 aldehyde and/or C9 ketone product) also assists in separating the myrcene spectra from those of α-pinene.
By comparing both the gas- and particle-phase cyclic monoterpenes in Figures 2 and 7a, it is evident that the dominant loadings represent compounds of similar MW, i.e. 169, 151 and 107 (primary aldehyde product, e.g. pinonaldehyde- PINAL in Figure 9, parent ion and fragments thereof) and 139 (primary ketone product parent ion) for the gas-phase and 187, 185, 183 and 169 for the particle-phase. Conversely, for the straight chain monoterpenes experiments the major gas-phase loadings represent compounds of significantly smaller MW than their particle-phase counterparts, i.e. 113 and 95 and 111 and 93, compared to 325, 322, 321, 227 and 215. Indeed, the straight chain monoterpene LC-MS/MS spectra contained on average ~ 10 % more signal > 250 Da than the cyclic monoterpene spectra. Also, the composition of the ions observed in the straight chain monoterpene LC-MS/MS spectra suggests that the SOA particles contained both oligomers and highly oxidized species, with the C_{10} backbone intact (i.e. O:C = 0.6), similar in structure to (but a little less oxidised than) extremely low volatility organic vapours (ELV-VOC), which have been observed previously in significant yield from α-pinene and limonene (as well as 6-nonenal ozonolysis chamber experiments in the absence of an OH scavenger, as well as boreal forests in Finland (Ehn et al., 2014). Further evidence to elucidate the type of SOA formed from the oxidation of straight chain monoterpenes can be obtained from investigation of the grouping of myrcene spectra in the cTOF-AMS PCA (Figure 8a). In the hour-4 cTOF-AMS PCA loadings bi-plot, we see that the grouping of the myrcene spectra is influenced somewhat by both m/z 41 and 44, indicating the presence of LV-OOA in the SOA, potentially a result of oligomerisation or further oxidative heterogeneous chemistry involving reaction at remaining C=C double bond sites.

β-caryophyllene readily forms particulate matter on oxidation (e.g. Jaoui et al., 2003; Lee et al., 2006; Winterhalter et al., 2009; Alfarra et al., 2012; Chen et al., 2012), with reaction predominantly at one of the two C=C sites (e.g. with OH or O_3, although O_3 attack occurs almost exclusively at the endocyclic double bond (Jenkin et al., 2012)), yielding relatively low vapour pressure, unsaturated and oxygenated primary products (Figure 9), which have significant affinity for the particle-phase (Jenkin et al., 2012).
A further oxidation step involving the second C=C site can result in increased oxygen (and/or nitrogen, depending on NOx conditions) content, yet with little, if any reduction in the original C number. As with the cyclic monoterpene PCAs, the CIR-TOF-MS and LC-MS/MS PCA bi-plots demonstrate similarities in terms of classifying β-caryophyllene oxidation and SOA formation with comparable MW species, e.g. primary products β-caryophyllon aldehyde (MW 236, BCAL in Figure 9) and β-caryophyllene secondary ozonide in the gas-phase (MW 252, BCSOZ in Figure 9), β-caryophyllonic acid (MW 252, C141CO2H in Figure 9) in both phases and secondary product β-nocaryophyllinic acid (MW 254, C131CO2H in Figure 9) in the particle-phase. In the hour-4 cTOF-AMS PCA scores plot, the myrcene and β-caryophyllene clusters are located adjacent to one another, with β-caryophyllene classification also influenced by the m/z 41 peak, which similar to myrcene SOA for example, is indicative of higher oxidized content (Alfarra et al., 2012), a result of either the partitioning of higher generation gas-phase products or heterogeneous oxidation of condensed first or second generation products.

5.2 Mapping within a class

Within the monoterpene group there is a small degree of separation between the limonene and α-pinene experiments, with three out of the four α-pinene experiments located to the upper and right region of the monoterpene cluster. This distribution/separation within the group may be a consequence of precursor-specific reaction pathways; for instance, although structurally similar, α-pinene and limonene react at somewhat different rates with respect to both OH and O3 (Atkinson and Arey, 2003). Over a fixed time period, such system reactivity will govern the degree of oxygenated content present within a closed analyte matrix and may facilitate the isolation of specific reaction pathways. Furthermore, the separation of such similar gas-phase precursors within a class cluster may help us to elucidate differences in resultant SOA yield and composition (e.g. limonene tends to have a larger SOA yield than α-pinene; Lee et al., 2006; Fry et al., 2014). It therefore may be possible with the use of larger and more detailed data sets to employ loading information to determine the importance of certain products to SOA composition.
However, additional data to those reported here would be required to fully test this hypothesis.

5.3 Mapping reactivity

In order to explore how the PCA technique can be used to investigate product distributions driven by certain starting conditions, a separate analysis was conducted on the five toluene experiments. In this instance we investigate the product distribution dependency on initial VOC/NO\textsubscript{x} ratios. The VOC/NO\textsubscript{x} ratios employed nominally represent “low”, “medium” and “high” NO\textsubscript{x} conditions, with values of roughly 11 (i.e. low NO\textsubscript{x}, “NO\textsubscript{x}-limited” ozone formation conditions – as determined from simulation chamber ozone isopleth plots, see Wagner et al., 2003; two experiments), 4 (i.e. moderate NO\textsubscript{x}, two experiments) and 1 (i.e. high NO\textsubscript{x}, “VOC-limited”; one experiment), respectively. The resultant PCA loadings bi-plot (produced using the methodology described in Section 3) is given in Figure 10.

From inspection of the PCA loadings bi-plot in Figure 10, it is clear that the toluene photooxidation spectra distribute in statistical space according to their respective initial VOC/NO\textsubscript{x} ratios. Figure 10 shows the low NO\textsubscript{x}, high VOC/NO\textsubscript{x} ratio experiments grouped in the lower right-hand quadrant of the PCA space, principally influenced by loadings representing toluene (m/z 93 and 77, parent and fragment ions, respectively; note m/z 93 off-scale in Figure 10) and cresol (m/z 109). Summed spectra containing larger quantities of precursor would suggest the presence of a less reactive environment, which is the case here, where low NO\textsubscript{x} levels in the NO\textsubscript{x} limited regime, result in low [OH] (reduced radical cycling) and low [O\textsubscript{3}] (less NO to NO\textsubscript{2} conversions) (see also Bloss et al., 2005). Similarly, the relatively large contribution from cresol to the low NO\textsubscript{x} summed spectra, originates from a larger net cresol concentration across the experiment on account of low system reactivity (i.e. loss via reaction with OH).

The moderate NO\textsubscript{x}, medium VOC/NO\textsubscript{x} experiments group uniquely in the lower left-hand quadrant of the PCA space in Figure 10, principally on account of loadings representing benzaldehyde (m/z 107) and the ring-opening products, citraconic...
anhydride (m/z 113), 4-oxo-2-pentenal, maleic anhydride and/or angelicalactone (m/z 99) and methyl glyoxal (m/z 73). The greater abundance of higher generation, ring-opening products implies a more reactive environment (i.e. increased chemical processing) than that formed under low NOx conditions. Larger net benzaldehyde concentrations originate from greater system reactivity and greater abundance of NO to fuel the RO2 + NO reaction.

The high NOx, low VOC/NOx ratio experiment is sited in the left-hand half of Figure 10, on account of it possessing higher system reactivity (with respect to the low NOx experiments) and the resultant greater proportion of ring-opening product ions (as the case for the moderate NOx experiments). However, the low VOC/NOx ratio experiment is uniquely displaced into the upper region of the PCA space owing to a large contribution from 2-butenedial and/or 2(5H)-furanone (m/z 85, off scale in Figure 10) to the summed spectra (the yields of both of which are likely to be important under high NOx conditions, owing to reaction through the RO2 + NO channel).

6. Atmospheric relevance and future directions

Having successfully used the mechanistic fingerprints in the chamber data to construct descriptive statistical models of the gas- and particle-phases, and having applied the methodology to map mesocosm environments, a next logical step would be to use this detailed chemical knowledge to investigate ambient VOC and SOA composition data in an attempt to help elucidate and deconvolve the important chemistry controlling the gas- and particle-phase composition of inherently more complex real world environments.

If ambient biogenic gas/particle composition spectra of unknown origin, uncertain speciated composition and/or a high level of detail and complexity were to be mapped onto the relevant statistical model (i.e. introduced as a separate test set), their resultant vector description in the statistical space would provide information regarding the type of precursors present and the underlying chemical mechanisms at play, as exemplified by the classifying of the mesocosm experiments by the fraction...
of isoprene, monoterpane and sesquiterpene chemistry in the experimental fingerprints. Furthermore, as shown by the mapping of toluene photooxidation experiments into a separate and distinct cluster, the methodology is potentially able to be robust with respect to other chemical compositions expected for a “real world” environment that is significantly impacted by both anthropogenic and biogenic emissions (e.g. Houston, USA and the Black Forest – Munich, DE). This capability is important when attempting to understand the complex interactions that exist between urban and rural atmospheres and when attempting to understand VOC and SOA source identification.

One potential problem in moving from simulation chamber data to “real world” systems, would be the applicability of using “static” experimental spectra (i.e. time averaged) to build a model to accept “dynamic” data, in which there would be potentially overlapping reaction coordinates and multiple precursor and radical sources.

In order to investigate the impact of a more dynamic system on the composition of the gas-phase matrix and hence on the composition of the spectra employed to build the model, a zero-dimensional chamber box model was constructed for the α-pinene system and operated under three different scenarios:

1. **Basic chamber simulation**: α-pinene concentration constrained to measurements (initial concentration 124 ppbV); NO and NO₂ initialised according to measurements (31 and 41 ppbV, respectively).

2. **Spiked chamber simulation**: α-pinene constrained as in (1), but profile duplicated to represent a fresh injection of the precursor (at the midpoint of the experiment) on top of the already evolving matrix; constant 10 ppbV HONO employed as NO and radical source.

3. **Constant injection chamber simulation**: α-pinene and HONO constrained to constant values of 5 and 10 ppbV, respectively.
It should be noted here that the model runs are not idealised. The aim of these simulations is to provide systematically more complex chemical systems with which to compare and contrast a simulation representing the measured dataset. For work regarding the evaluation of the MCM with respect to single VOC precursor chamber experiments (including model-measurement intercomparison), see, for example, Bloss et al., 2005 (toluene), Metzger et al., 2008 and Rickard et al., 2010 (1,3,5-TMB), Camredon et al., 2010 (α-pinene) and Jenkin et al., 2012 (β-caryophyllene).

The results of the three different model scenarios are given in Figure 11, mapped through to (i.e. integrated across the experiment) the resultant simulated mass spectra.

Figures 11a and b show the results from scenario (1). Figure 11a gives the evolution of the system over the molecular weight region of interest with time and Figure 11b gives the scenario summed “model mass spectra”, i.e. the relative abundance of all simulated compounds within the gas-phase molecular weight region of interest (with relative contributions from isobaric species summed into a single “peak”). Scenario (1) and Figures 11a and b approximate the experimental data employed within this work and constitute the model base-case.

Figures 11c and d show the results from scenario (2). Figure 11c clearly shows the second α-pinene injection on top of the evolving matrix and the resultant system evolution. Figure 11d show the “difference model mass spectra” between scenarios (1) and (2), from which it can clearly be seen that there is very little difference between the spectra of the basic model and the “spiked” system. The difference in “mass channel” relative abundance (ΔMC) is generally ≤ 2 %, with the exceptions of MWs 168 and 186. MW 168 primarily comprises pinonaldehyde, with a ΔMC of around - 6 %; pinonaldehyde is a primary product and is slightly lower in relative abundance in scenario (2) owing to the longer reaction time employed and the greater proportion of pinonaldehyde reacted. MW 186 comprises a number of primary and secondary products and has a ΔMC of roughly + 3 %.
The results from model scenario (3) are given in Figures 11e and f. As with scenario (2), there is no dramatic difference between the simulated mass spectra of scenario (3) and the base-case scenario (1). In this instance $\Delta MC$ is generally $\leq 5\%$, with the exceptions of MWs 136 and 168 and MWs 121 and 245. The relative abundance of the precursor is lower in this case on account of the constraining method employed and once again the relative abundance of pinonaldehyde is slightly lower due to the longer reaction time. MW 121 solely comprises PAN and MW 245 primarily comprises a C$_{10}$ tertiary nitrate (C$_{10}$H$_{15}$NO$_6$, MCM designation: C106NO3). Both species are slightly elevated with respect to the base-case in scenario (3) owing to the longer reaction time and the continual input of OH and NO into the model in the form HONO.

Scenarios (2) and (3) represent complex mixtures with overlapping reaction coordinates, each one step closer to a “real world” case than scenario (1) and the chamber data employed within this work. However, despite the increase in complexity of the scenarios, both exhibit very little compositional difference to the base-case scenario and hence the chamber data employed in this work. These results give some confidence that despite being constructed from summed simulation chamber data, the statistical models employed here represents a solid framework onto which real atmosphere spectra could be mapped and interpreted.

A further step in increasing complexity and hence a further step towards the “real world” system, would be the addition of other (potentially unidentified) precursors to the simulation, which may be at different stages of oxidation or have passed through different reactive environments. Further increases in complexity, beyond the analysis discussed here, will form the focus of future work.

7. Conclusions

A chemometric dimension reduction methodology, comprising PCA, HCA and PLS-DA has been successfully applied for the first time to complex gas- and particle-phase composition spectra of a wide range of BVOC and mesocosm environmental simulation chamber photooxidation experiments. The results show that the oxidized
gas-phase atmosphere (i.e. the integrated reaction coordinate) of each different structural type of BVOC can be classified into a distinct group according to the controlling chemistry and the products formed. Indeed, a major potential strength of the data analysis methodology described here, could lie in the decoding of mechanisms into pathways (i.e. separation within a group on account of different underlying chemistry) and consequently linking chemical pathways to precursor compounds. Furthermore, the methodology was similarly able to differentiate between the types of SOA particles formed by each different class of terpene, both in the detailed and broad chemical composition spectra. In concert, these results show the different SOA formation chemistry, starting in the gas-phase, proceeding to govern the differences between the various terpene particle compositions.

The ability of the methodology employed here to efficiently and effectively “data mine” large and complex datasets becomes particularly pertinent when considering that modern instrumentation/techniques produce large quantities of high-resolution temporal and speciated data over potentially long observation periods. Such statistical mapping of organic reactivity offers the ability to simplify complex chemical datasets and provide rapid and meaningful insight into detailed reaction systems comprising hundreds of reactive species. Moreover, the demonstrated methodology has the potential to assist in the evaluation of (chamber and real world) modelling results, providing easy to use, comprehensive observational metrics with which to test and evaluate model mechanisms and outputs and thus help advance our understanding of complex organic oxidation chemistry and SOA formation.

8. Acknowledgements

The authors gratefully acknowledge the UK Natural Environment Research Council (NERC) for funding the APPRAISE ACES consortium (NE/E011217/1) and the TRAPOZ project (NE/E016081/1); the EU FP7 EUROCHAMP-2 programme for funding the TOXIC project (E2-2009-06-24-0001); the EU ACCENT Access to Infrastructures program for funding work at the PSI and the EU PEGASOS project (FP7-ENV-2010-265148) for funding used to support this work. Drs A. R. Rickard and M. R. Alfarra.
were supported by the NERC National Centre for Atmospheric Sciences (NCAS). The authors would like to thank the University of Leicester Atmospheric Chemistry group for assistance throughout all experiments, including Drs Alex Parker, Chris Whyte, Iain White and Timo Carr; co-workers at the University of Manchester for assistance with MAC experiments; co-workers from Fundacion CEAM, Drs Marie Camredon and Salim Alam for assistance with EUPHORE experiments and co-workers from the Laboratory of Atmospheric Chemistry smog chamber facility at the Paul Scherrer Institute (PSI) for assistance with PSISC experiments.

9. References


http://mcm.leeds.ac.uk/MCM:


Sindelarova, K., Granier, C., Bouarar, I., Guenther, A., Tilmes, S., Stavrakou, T., Muller, J. F., Kuhn, U., Stefani, P., and Knorr, W.: Global data set of biogenic VOC emissions calculated by the MEGAN model over the last 30 years, Atmospheric Chemistry and Physics, 14, 9317-9341, 10.5194/acp-14-9317-2014, 2014.


10. Tables

<table>
<thead>
<tr>
<th>Experiment ID</th>
<th>Precursor</th>
<th>Structure</th>
<th>$k(\text{OH}) / k(\text{O}_3)$ / cm$^3$ molec$^{-1}$ s$^{-1}$</th>
<th>Experiment Type (no.)</th>
<th>VOC/NO$_x$ Range</th>
<th>RH / % Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISOP1 – $10^{1.4 M.P}$</td>
<td>isoprene</td>
<td></td>
<td>9.9 x $10^{11}$ / 1.2 x $10^{17}$</td>
<td>Photooxidation (10)</td>
<td>1.3 – 20.0</td>
<td>49 – 72</td>
</tr>
<tr>
<td>APIN1 – $4^{2.3 M.P}$</td>
<td>α-pinene</td>
<td></td>
<td>5.3 x $10^{11}$ / 8.4 x $10^{17}$</td>
<td>Photooxidation (4)</td>
<td>1.3 – 2.0$^1$</td>
<td>49 – 73</td>
</tr>
<tr>
<td>LIM1 – $6^{2.3 M.E.P}$</td>
<td>limonene</td>
<td></td>
<td>1.7 x $10^{10}$ / 2.1 x $10^{16}$</td>
<td>Photooxidation (6)</td>
<td>1.4 – 2.0$^1$</td>
<td>50$^1$ – 82</td>
</tr>
<tr>
<td>BCARY1 – $10^{2.3 M}$</td>
<td>β-caryophyllene</td>
<td></td>
<td>2.0 x $10^{10}$ / 1.2 x $10^{14}$</td>
<td>Photooxidation (10)</td>
<td>0.6 – 2.0$^1$</td>
<td>50$^1$ – 72</td>
</tr>
<tr>
<td>MYRC1 – $2^{2.3 M}$</td>
<td>myrcene</td>
<td></td>
<td>2.1 x $10^{10}$ / 4.7 x $10^{16}$</td>
<td>Photooxidation (2)</td>
<td>1.4 – 1.9</td>
<td>52 – 54</td>
</tr>
<tr>
<td>Experiment ID</td>
<td>Precursor</td>
<td>Structure</td>
<td>$k(\text{OH}) / k(\text{O}_3)$ $/ \text{cm}^3 \text{molec}^{-1} \text{s}^{-1}$</td>
<td>Experiment Type (no.)</td>
<td>VOC/NOx Range</td>
<td>RH /% Range</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>-----------</td>
<td>---------------------------------</td>
<td>----------------------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LINA1 – 2.8 M</td>
<td>linalool</td>
<td><img src="image" alt="Structure of linalool" /></td>
<td>$1.6 \times 10^{10} / 4.5 \times 10^{16}$</td>
<td>Photooxidation (2)</td>
<td>1.4 – 2.6</td>
<td>42 – 47</td>
</tr>
<tr>
<td>BIR1 – 2 M</td>
<td>birch trees</td>
<td>Multiple emissions$^5$</td>
<td>Multiple emissions</td>
<td>Mesocosm</td>
<td>Photooxidation (2)</td>
<td></td>
</tr>
<tr>
<td>FIG1 – 2 M</td>
<td>fig trees</td>
<td>Multiple emissions$^5$</td>
<td>Multiple emissions</td>
<td>Mesocosm</td>
<td>Photooxidation (2)</td>
<td></td>
</tr>
<tr>
<td>TOL1 – 5 E</td>
<td>toluene</td>
<td><img src="image" alt="Structure of toluene" /></td>
<td>$3.7 \times 10^{12} /$</td>
<td>Photooxidation (5)</td>
<td>1.3 – 11.6</td>
<td>2 – 6</td>
</tr>
</tbody>
</table>

1 = Estimated using known volume of reactants injected  
2 = LC-MS/MS filter data available for at least one of these experiments (MAC)  
3 = c-TOF-AMS data available for at least one of these experiments (MAC)  
4 = From (Atkinson and Arey, 2003b;Sun et al., 2012;Khamaganov and Hites, 2001) and references therein  
5 = See Wyche et al., 2014
M = experiments conducted in the MAC
E = experiments conducted in the EUPHORE
P = experiments conducted in the PSISC
Table 2: Key technical features of MAC, EUPHORE and PSISC (Alfarra et al., 2012; Becker, 1996; Bloss et al., 2005; Camredon et al., 2010; Paulsen et al., 2005; Zador et al., 2006).

<table>
<thead>
<tr>
<th>Chamber</th>
<th>Material</th>
<th>Environment</th>
<th>Size</th>
<th>Light Source</th>
<th>Spectrum</th>
</tr>
</thead>
</table>
| MAC     | FEP Teflon| Indoor      | 18 m³, 3(H) x 3(L) x 2(W) m | 1 x 6 kW Xe arc lamp       | λ range = 290 – 800 nm  
|         |           |             |                             | Bank of halogen lamps       | $j_{NO2} = 6 \times 10^{-4} \text{ s}^{-1} (290 – 422 \text{ nm})$       |
| EUPHORE | FEP Teflon| Outdoor     | 200 m³, (hemispherical)     | Solar                      | Solar; 75 % transmission at 290 nm, 85 % transmission > 320 nm  
|         |           |             |                             |                            | $j_{NO2} = 5 – 9 \times 10^{-3} \text{ s}^{-1}$                       |
| PSISC   | FEP DuPontTedlar | Indoor | 27 m³, 3(H) x 3(L) x 3(W) m | 4 x 4 kW Xenon arc lamps   | λ range = 290 – 800 nm  
|         |           |             |                             |                            | $j_{NO2} = 0.12 \text{ min}^{-1}$                                      |
**Table 3:** List of certain major product ions integral to the separation of spectra in statistical space, their corresponding tentative assignments and their precursor. See main text, Section 4.2 for further information.

<table>
<thead>
<tr>
<th>Ion / m/z</th>
<th>CIR-TOF-MS</th>
<th>LC-MS/MS</th>
<th>cTOF-AMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assignment</td>
<td>Precursor</td>
<td>Assignment</td>
</tr>
<tr>
<td>41</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>43</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>44</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>71</td>
<td>methyl vinyl ketone + methacrolein</td>
<td>isoprene</td>
<td>-</td>
</tr>
<tr>
<td>75</td>
<td>hydroxyl acetone</td>
<td>isoprene</td>
<td>-</td>
</tr>
<tr>
<td>83</td>
<td>3-methyl furan</td>
<td>isoprene</td>
<td>-</td>
</tr>
<tr>
<td>87</td>
<td>C_{4}hydroxy carbonyls / methacrylic acid</td>
<td>isoprene</td>
<td>-</td>
</tr>
<tr>
<td>93</td>
<td>4-vinyl-4-pentenal / 4-hydroxy-4-methyl-5-hexen-1-al</td>
<td>myrcene / linalool</td>
<td>-</td>
</tr>
<tr>
<td>95</td>
<td>4-vinyl-4-pentenal / 4-hydroxy-4-methyl-5-hexen-1-al</td>
<td>myrcene / linalool</td>
<td>-</td>
</tr>
<tr>
<td>107</td>
<td>piononaldehyde / α-pinene</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>139</td>
<td>limononaldehyde</td>
<td>limonene</td>
<td>limonene</td>
</tr>
<tr>
<td>151</td>
<td>pinoonaldehyde / limononaldehyde</td>
<td>α-pinene / limonene</td>
<td>α-pinene / limonene</td>
</tr>
<tr>
<td>169</td>
<td>pinoonaldehyde / limononaldehyde</td>
<td>α-pinene / limonene</td>
<td>α-pinene / limonene</td>
</tr>
<tr>
<td>183</td>
<td>pinonic acid / limononic acid</td>
<td>α-pinene / limonene</td>
<td>α-pinene / limonene</td>
</tr>
<tr>
<td>185</td>
<td>pinic acid / limonic acid</td>
<td>α-pinene / limonene</td>
<td>α-pinene / limonene</td>
</tr>
<tr>
<td>199</td>
<td>C_{9}H_{11}O_{5}</td>
<td>myrcene</td>
<td>myrcene</td>
</tr>
<tr>
<td>215</td>
<td>C_{10}H_{15}O_{5}</td>
<td>myrcene</td>
<td>myrcene</td>
</tr>
<tr>
<td>235</td>
<td>β-caryophyllene secondary ozonide + isomers</td>
<td>β-caryophyllene</td>
<td>β-caryophyllene</td>
</tr>
<tr>
<td>227</td>
<td>C_{10}H_{11}O_{5}</td>
<td>myrcene</td>
<td>myrcene</td>
</tr>
<tr>
<td>237</td>
<td>β-caryophyllene aldehyde</td>
<td>β-caryophyllene</td>
<td>β-caryophyllene</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td>Mass</td>
<td>Formular</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------------------------------------------------</td>
<td>--------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>251</td>
<td>propanoic acid</td>
<td>85</td>
<td>C₃H₆O₂</td>
</tr>
<tr>
<td>253</td>
<td>β-caryophyllene</td>
<td>253</td>
<td>C₁₀H₁₆O₂</td>
</tr>
<tr>
<td></td>
<td>β-caryophyllene</td>
<td>253</td>
<td>C₁₀H₁₆O₂</td>
</tr>
<tr>
<td></td>
<td>acid</td>
<td></td>
<td>C₁₀H₁₆O₂</td>
</tr>
<tr>
<td>255</td>
<td>4-(2-(2-carboxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid</td>
<td>255</td>
<td>C₁₄H₁₈O₅</td>
</tr>
<tr>
<td>267</td>
<td>4-[(3-hydroperoxy-3-hydroxyethyl)-3,3-dimethylcyclobutyl]-4-oxobutanoic acid / 4-(2-(2-carboxy-1-hydroxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid</td>
<td>267</td>
<td>C₁₄H₁₈O₅</td>
</tr>
<tr>
<td>321</td>
<td>[M-H₂O+FA+Na]</td>
<td>321</td>
<td>C₁₂H₁₄O₅ / C₁₃H₁₆O₅</td>
</tr>
<tr>
<td></td>
<td>myrcene</td>
<td>321</td>
<td>C₁₃H₂₀O₅</td>
</tr>
</tbody>
</table>
Table 4: PLS-DA model classification sensitivity and specificity for the gas-phase biogenic air matrices

<table>
<thead>
<tr>
<th>Cross Validation</th>
<th>isoprene</th>
<th>cyclic-monoterpenes</th>
<th>sesquiterpene</th>
<th>straight-chain-monoterpenes</th>
<th>Fig tree</th>
<th>Birch tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>100.0</td>
<td>92.9</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>91.7</td>
</tr>
</tbody>
</table>
11. Figures

**Figure 1:** (a) NOx, O3, myrcene and 4-vinyl-4-pentenal (primary aldehyde product) and (b) particle mass (not wall loss corrected and assuming $\rho = 1.3$) and size evolution within the MAC during a typical photooxidation experiment.

**Figure 2:** PCA loadings bi-plot of the second vs. first principal components derived from the PCA analysis of the isoprene, cyclic monoterpane ("c-m-terpene" in the legend; $\alpha$-pinene and limonene), sesquiterpene ($\beta$-caryophyllene) and straight chain biogenic ("s-m-terpene" in the legend; myrcene and linalool) chamber data. Classification confidence limits = 95 %. Tentative assignments of major ions include $m/z$ 71 = methyl vinyl ketone and methacrolein, 75 = hydroxy acetone, 83 = e.g. 3-methyl furan, 87 = C4-hydroxycarboxyls/methacrylic acid, $m/z$ 237 = $\beta$-caryophyllene aldehyde, 235 and 253 = $\beta$-caryophyllene secondary ozonide (and isomers thereof), $m/z$ 107, 151 and 169 = piononaldehyde and limononaldehyde, 139 = limonketone, $m/z$ 95 and 93 = 4-vinyl-4-pentenal and 4-hydroxy-4-methyl-5-hexen-1-al. See main text, Section 4.2 and Table 3 for further information. For clarity, the scale has been set to show the bulk of the data, hence precursor parent ions and $m/z$ 71 are not shown.

**Figure 3:** PCA scores plot of the first vs. second principal components derived from the PCA analysis of the mesocosm test set using the PCA model developed from the isoprene, cyclic monoterpane ($\alpha$-pinene and limonene), sesquiterpene ($\beta$-caryophyllene) and straight chain monoterpane (myrcene and linalool) chamber data. Classification confidence limits = 95 %.

**Figure 4:** PCA scores plot of the first vs. second principal components derived from the PCA analysis of the toluene test set using the PCA model developed from the isoprene, cyclic monoterpane ($\alpha$-pinene and limonene), sesquiterpene ($\beta$-
caryophyllene) and straight chain monoterpenes (myrcene and linalool) chamber data. Classification confidence limits = 95%.

**Figure 5:** Dendrogram showing the grouping relationship between the various gas-phase matrices of systems examined. Red = isoprene, pink = fig, green = cyclic monoterpenes (α-pinene and limonene), yellow = birch, light blue = straight chain monoterpenes (myrcene and linalool) and dark blue = sesquiterpene (β-caryophyllene).

**Figure 6:** scores plot of the first three latent variables derived from the PLS-DA model analysis of the isoprene, cyclic monoterpenes (α-pinene and limonene), sesquiterpene (β-caryophyllene), straight chain monoterpenes (myrcene and linalool), fig and birch chamber data. Classification confidence limits = 95%.

**Figure 7:** (a) Loadings bi-plot of the second vs. first principal components obtained from the PCA of LC-MS aerosol spectra from a subset of terpene experiments and (b) the corresponding HCA dendrogram. See main text, Section 4.5 and Table 3 for further information, including ion assignments.

**Figure 8:** (a) Loadings bi-plot of the second vs. first principal components obtained from the PCA of AMS aerosol spectra from of a subset of terpene experiments and (b) the corresponding HCA dendrogram. See main text, Section 4.5 and Table 3 for further information, including ion assignments.

**Figure 9:** Simplified schematic illustrating some of the important mechanistic pathways in the gas-phase oxidation of isoprene, α-pinene, β-caryophyllene and
myrcene, and the associated mass transfer to the particle-phase. Red arrows and 
text = “high” NOx pathways, green arrows and text = “low NOx,” pathways, blue 
arrows and text = ozonolysis reactions, grey arrow and text = speculative, dashed 
arrows = multiple steps. * = multiple photooxidative routes initiated by reaction with 
OH (i.e. involving the reactants – OH, O2, NOx, HO2 and/or RO2), leading to structurally 
similar products containing different functional groups. α-pinene mechanism – X = 
OH, =O, OOH or ONO2; Y = CHO or C(O)OH; Z = OH, OOH or ONO2. β-caryophyllene 
mechanism – X = CH2OH(OH), CH2OH(OOH), CH2OH(ONO2) or =O. Myrcene 
mechanism – Y = OOH or ONO2; Z = CHO or C(O)OH. See text, section 5 for 
references.

**Figure 10:** PCA loadings bi-plot of the second vs. first principal components derived 
from the PCA analysis of the toluene experiments. Experiments were conducted 
under low NOx, high VOC/NOx ratio (red diamonds), moderate NOx, medium 
VOC/NOx ratio (green squares) and high NOx, low VOC/NOx ratio (blue triangle) 
conditions. For clarity, the scale has been set to show the bulk of the data, hence 
m/z 93 and 85 are not shown.

**Figure 11:** Results from MCM α-pinene photooxidation simulations. (a) and (b) = 
basic α-pinene photooxidation; (c) and (d) = spiked injection of α-pinene, 
continuous HONO input; (e) and (f) = continuous α-pinene and HONO input. Left 
hand image plots show the evolution of the respective systems over the molecular 
weight region of interest with time; colour scale = relative abundance (%). Right 
hand plots = (b) relative abundance of simulated molecular weights during straight 
α-pinene photooxidation; (d) difference in relative abundance of simulated 
molecular weights between double injection of α-pinene continuous HONO input 
and straight α-pinene photooxidation; (f) difference in relative abundance of 
simulated molecular weights between continuous α-pinene and HONO input and 
straight α-pinene photooxidation. See text for details.