S1 FLAME-4 FL12 emission profiles similarity study

This section presents a comparison and analysis of emission profiles using data from the FLAME-4 FL12 campaign, and provides the rationale for the focus on monoterpenoids in the current application of the PR and classification algorithms. The FLAME-4 FL12 data set was chosen for the comparison and analysis because it currently contains a larger number and broader range of compounds for which calibrated mixing ratios are available and EFs have been calculated. In the near future, the FIREX FL16 data set will provide additional opportunities for analysis and application of the PR and classification algorithms. The published FLAME-4 FL 12 data set (Hatch et al. (2015)) includes 458 compounds across six samples (ponderosa pine, black spruce, Indonesian peat, rice straw, wiregrass and giant cutgrass). With only six samples, the statistical requirements of the PR algorithm could not be met. Therefore, a simplified approach was applied to evaluate the similarity in the emission profiles, using two distance metrics: 1) cosine distance (Eq. S1), and 2) Euclidean distance (Eq. S2):

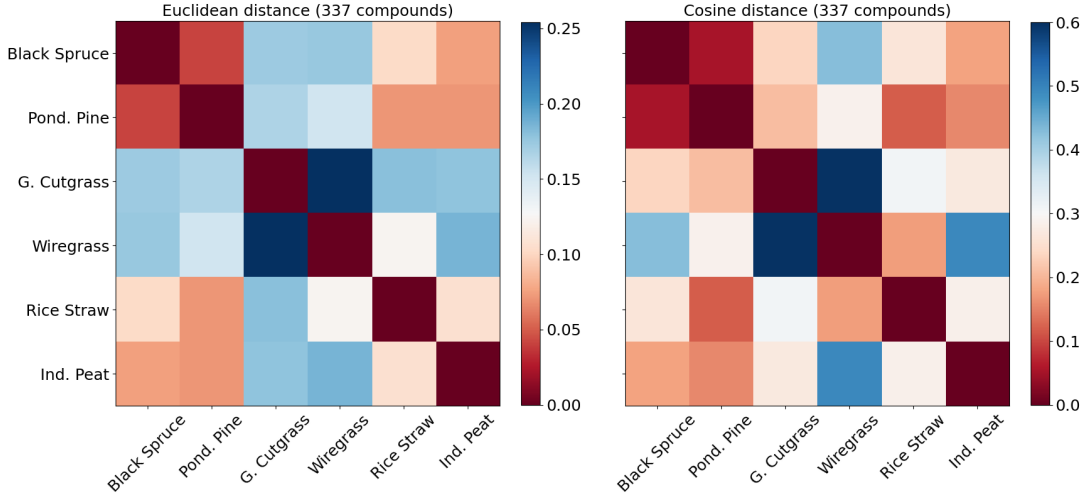
$$d_1 = 1 - \frac{u \cdot v}{||u||_2 ||v||_2} \quad (S1)$$

$$d_2 = ||u - v||_2 = (\sum_{j=1}^{n} |u_j - v_j|^2)^{1/2} \quad (S2)$$

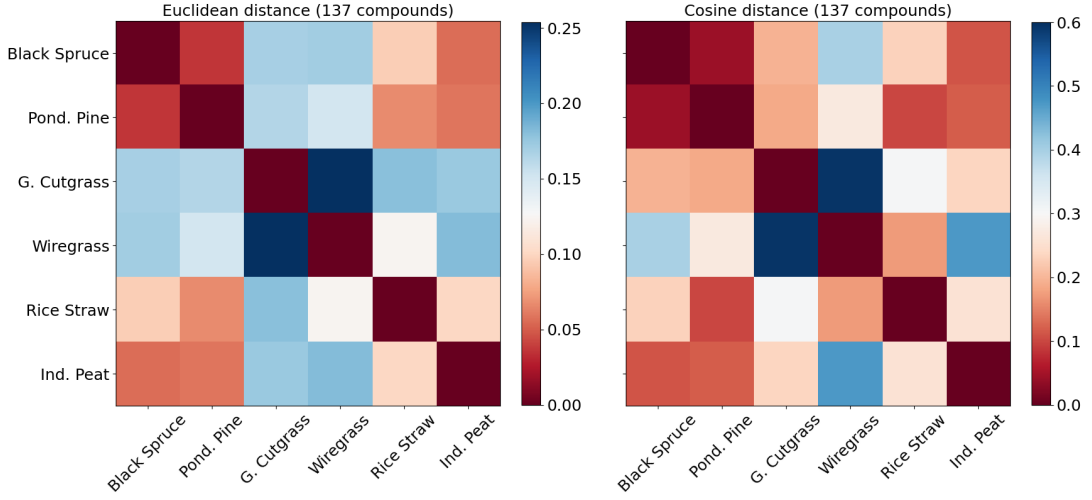
where d_1 and d_2 are the cosine and Euclidean distances, respectively, and u and v are vectors that represent fuel types using their emission profiles. The index j corresponds to a specific compound in the emissions profile of both u and v, and n is the total number of compounds in the emission profiles of u and v.

Before the calculation of either distance metric, the emission profiles of all samples were normalized using the total NMOC EF for each fuel. Normalization was necessary because Euclidean distance is sensitive to magnitude. Furthermore because Eqs. 1 and 2 cannot be applied with missing values (in this data set, values below background), the compounds were filtered and missing values were replaced. Compounds were filtered using two different thresholds, missing in three or more samples and missing in four or more samples. Compounds that exceeded the thresholds were removed from further analysis. For the remaining compounds any missing values were replaced with zeros. Using a threshold of three yielded 137 compounds out of 458, and using a threshold of four yielded 337 out of 458. After normalizing and filtering, the cosine and Euclidean distances were calculated pairwise for all fuel types. The results for both thresholds are shown in Fig. S1 as heatmaps.

From S1 it can be seen that the two conifers (black spruce and ponderosa pine) have the smallest pairwise distances (cosine and Euclidean) and show the greatest similarity. For the rest of the fuels the average distances were 0.2-0.3 (cosine) and 0.1 (Euclidean); and the maximum distances were 0.6 (cosine) and 0.25 (Euclidean). The implication of these results are that the distance methods, applied to the FLAME-4 FL 12 NMOC profiles, can be used to differentiate some of the fuel types (i.e., peat from grasses, grasses from conifers), but they can not be used to differentiate the two conifers. Given that conifers are a dominant fuel type in the western US and that Hatch et al. (2019) showed that monoterpenoids were strongly correlated with conifers,the PR algorithm was developed and applied to differentiate coniferous fuels using monoterpenoids measured during the FLAME-4 FL 12 and FIREX FL16 campaigns.



[a]



[b]

S 1. The Euclidean and cosine distances for the FLAME-4 FL12 samples, using a threshold of three or more compounds [a] and four or more compounds [b].

S2 Analysis of variance (ANOVA) technical details

In this work, ANOVA (King (2010)) was used for feature selection. ANOVA is a statistical test used to analyze the difference between the means of more than two groups for variable(s) of interest. The null hypothesis (H_o) of ANOVA is that there is no difference among group means. The alternative hypothesis (H_a) is that the mean of the dependent variable for at least one group differs significantly from the other mean(s). ANOVA uses the *F*-test for statistical significance. This allows for comparison of multiple groups at one time, because the error is calculated for the set of means rather than for each pair as is done with a *t*-test. The *F*-test compares the variance in each group mean from the overall group variance. If the variance within groups is smaller than the variance between groups, the *F*-test will find a higher *F*-ratio, defined as (Eq. S3):

$$F_{ratio} = \frac{V_b}{V_w} \quad (S3)$$

where V_b is the mean sum of squares between groups/classes and V_w is the sum of squares within the same group/class. A higher *F*-ratio indicates a higher likelihood that the observed difference is real and not random. Table 1 shows the formulas for the calculation of the *F*-ratio, in which *k* and *l* correspond to the number of classes and samples in each class, respectively, and $\bar{X}$ and $\bar{X}_j$ are the overall average of the samples for a specific compound and average of a compound for a specific class.

Table 1. Formulas for ANOVA

Variance Source	Sum of Squares	Degrees of Freedom	Mean Squares
Within	$SSW = \sum_{j=1}^{k}\sum_{i=1}^{l}(X - \bar{X}_j)^2$	$df_w = k - 1$	$V_b = \frac{SSW}{df_w}$
Between	$SSB = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$	$df_b = n - k$	$V_w = \frac{SSB}{df_b}$
Total	$SST = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$	$df_t = n - 1$	

S3 Principal component analysis (PCA) with singular value decomposition (SVD)

The first step of PCA with SVD involves the decomposition of the data matrix. The matrix notation for SVD is given by Eq. S4:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \quad (\mathrm{S4})$$

where $\boldsymbol{X}$ is the original (n x d) data matrix, $\boldsymbol{U}$ (n x r) is the matrix of left singular vectors, $\boldsymbol{S}$ (r x r) is a diagonal matrix of singular values for $\boldsymbol{X}$, and $\boldsymbol{V}$ (r x d) is the matrix of right singular vectors that correspond to the new principal directions, and r is the rank of matrix $\boldsymbol{X}$. The principal component scores which are the projection of $\boldsymbol{X}$ to a lower dimensional space are given by Eq. S5:

$$\mathrm{scores(PCs)} = \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{S} \quad (\mathrm{S5})$$

Equation S5 in a reduced format gives:

$$PC_i = X * u_i = ((x_i - \mu)u_i, \ldots, (x_n - \mu)^T u_i)^T \quad (S6)$$

where $x_i, \ldots\ x_n$ are the original features. Finally the reduced dimensionality is achieved by using an appropriate number of PCs ($m < d$) to represent the projected data. The selection of components is based on the minimization of the reconstruction error (Eq. S7), calculated by:

$$\mathrm{error} = \min|\boldsymbol{X}_{\mathrm{original}} - \boldsymbol{X}_{\mathrm{reconstructed}}| \quad (\mathrm{S7})$$

The error is a measure of how much information was lost during the dimensionality reduction using fewer components than the number of dimensions to reconstruct the original data matrix. Because the calculation of the error might be computationally expensive, especially for large matrices, there is another metric which is related to the reconstruction error and is much easier to calculate. That metric is called the explained variance ratio (Jolliffe (2002)) and is equal to (Eq. S8):

$$EV\% = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} * 100 \quad (S8)$$

where k is the number of retained components, p is the number of original variables, λ are the eigenvalues of the decomposition that are obtained by Eq. S9:

$$\Lambda = \frac{\boldsymbol{S}^2}{(n-1)} \quad (S9)$$

or in reduced format (Eq. S10):

$$\lambda = \frac{s_i^2}{(n-1)} \quad (10)$$

where s are the singular values of the SVD decomposition and n the number of samples in the data set.

S4 Transformations for linear discriminant analysis (LDA)

The LDA algorithm uses the PCA scores from the PR algorithm for training. The PC scores were calculated using the original data matrix after standardization (Lever et al. (2017)). Therefore samples that will be classified and were not part of the training set need to be standardized and then transformed to their representation in the PCA space. Standardization is the process of mean centering and division by the average and standard deviation of each selected compound across all samples, calculated by Eq. S11:

$$\text{compound}_i = \frac{\text{ER}_i - \mu_{\text{training}}}{\sigma_{\text{training}}} \quad (S11)$$

where ER_i is the emission ratio of the compound to be standardized, and μ_{training} and σ_{training} are the average and standard deviation for the same compound obtained from the training set. In this application, the averages and standard deviations were obtained using the training set samples so that the new samples were projected to the same PCA space as the training samples.

S5 Classification using LDA

The probability for class assignment, for more than two classes, in LDA is given by *P(Y = n | X = x)* (Eq. S12):

$$P(Y = n|X = x) = \frac{\pi_n f_n(x)}{\sum_{l=1}^{N} \pi_l f_l(x)} \quad (S12)$$

where *n* is the class number, *x* the sample to be classified and *l* denotes the *l*th class. π_k is the fraction of the training observations that belong to the *n*th class and $f_n(X)$ is the probability density function. In this case $f_n(X)$ is assumed to be a multivariate Gaussian distribution (Eq. S13):

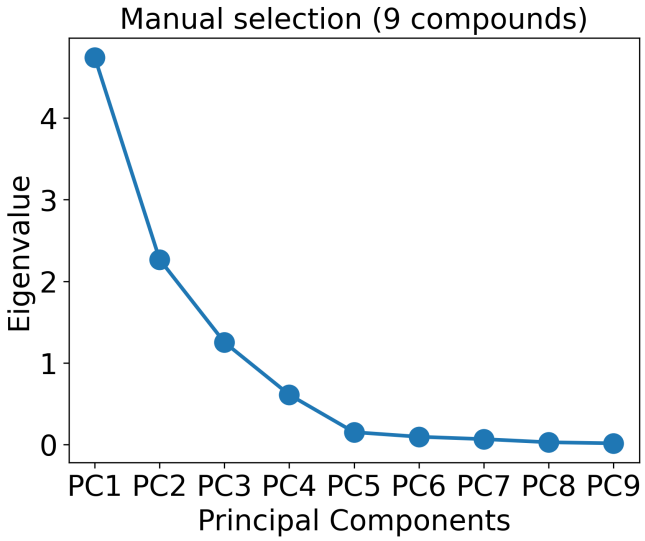
$$f_n(x) = -\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \quad (S13)$$

where μ is the vector of the means for each class based on the selected features, Σ is the common covariance matrix for the classes in the training set and *p* is the number of features. Substituting Eq. S13 into Eq. S12 and taking the natural logarithm of each side yields:

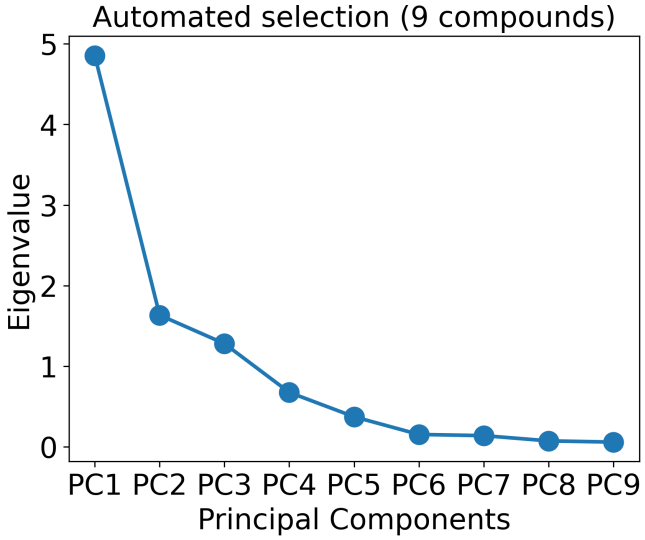
$$\log P(y = n|x) = -\frac{1}{2}(x-\mu_n)^t\Sigma^{-1}(x-\mu_n) + \log P(y = n) + C_{st} \quad (S14)$$

where C_{st} is a term that includes constants from the multivariate Gaussian distribution (Eq. S13).

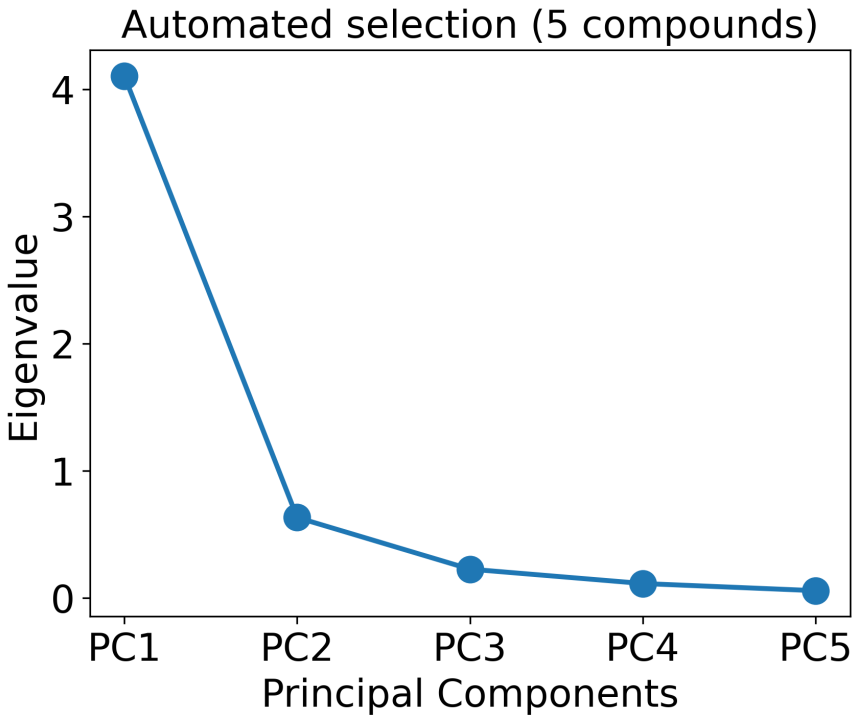
S6 Algorithm performance figures and overstory composition figures for Blodgett Forest Research Station



S 2. Eigenvalues from PCA runs for the manual selection of nine compounds versus the number of principal components.

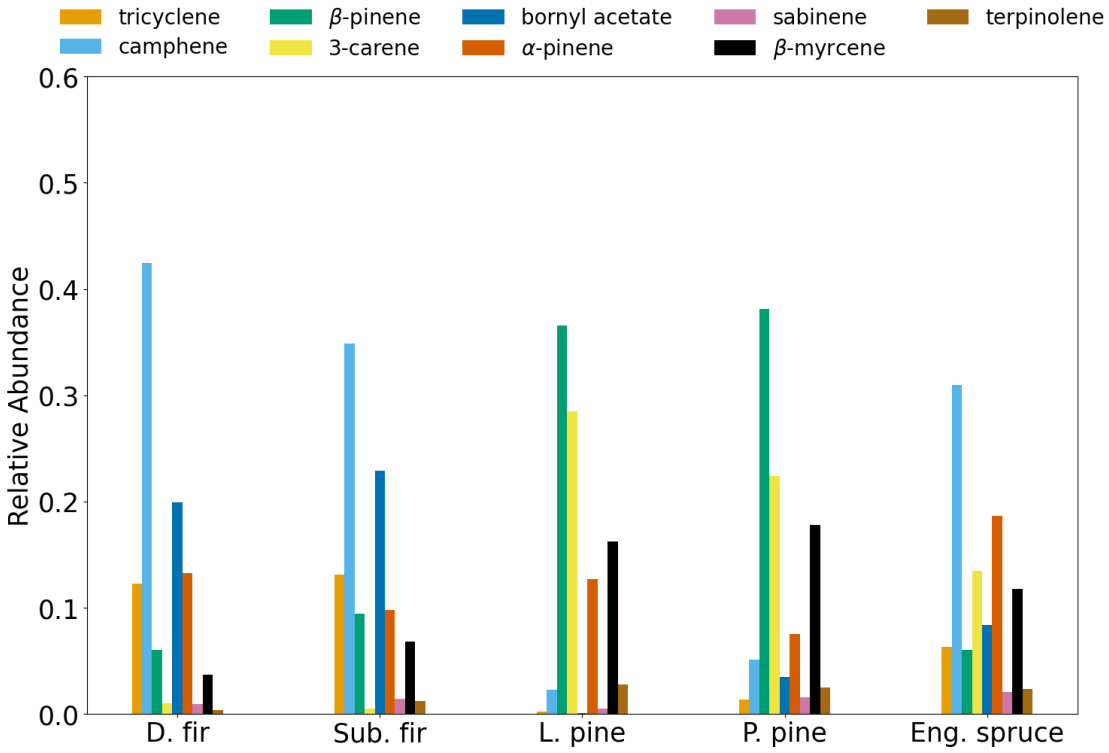


S 3. Eigenvalues from PCA runs for the automated selection of nine compounds versus the number of principal components.

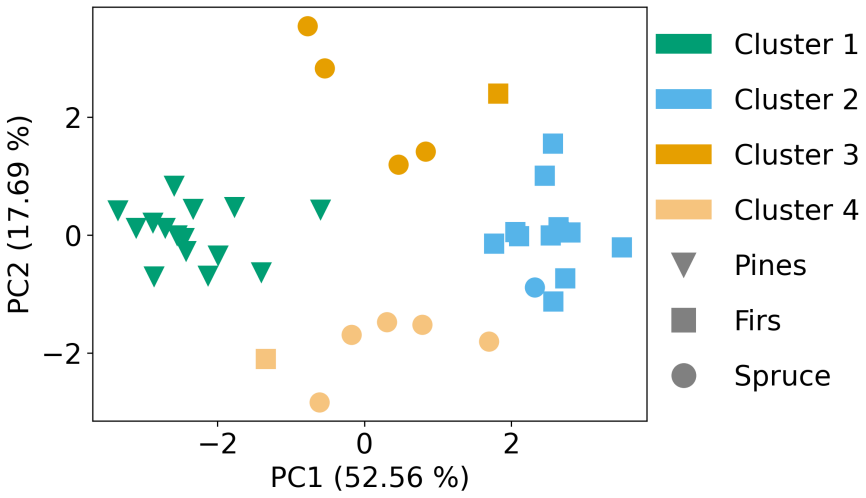


S 4. Eigenvalues from PCA runs for automated selection for five compounds versus the number of principal components.

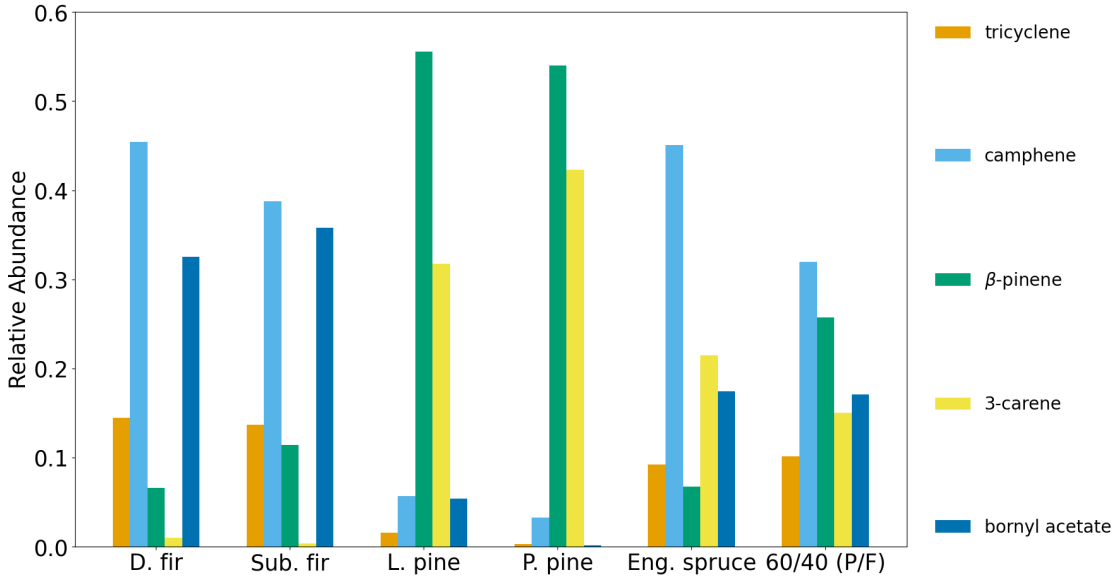
The PR algorithm was run twice; once for five compounds (optimal run case) and once for nine compounds (manuscript section 3.2.1). The effectiveness of dimensionality reduction was compared for the two runs and with manual selection of compounds. In the main manuscript, the quality of feature selection was compared for the three selection methods (automated 9 compounds, automated 5 compounds, and manual) using three different metrics to evaluate the separation. Here the effectiveness of the two automated selection methods is further evaluated by comparing the number of necessary components relative to the initial dimensions to achieve the 80% variance threshold. The selection of five compounds required about 20% - 40% of the number of original dimensions for an effective dimensionality reduction. The additional four selected compounds, for the nine compounds run, required about 44% - 55% of the number of original dimensions. This shows that the solutions for nine compounds require about half the number of original dimensions compared to the five compound solution.



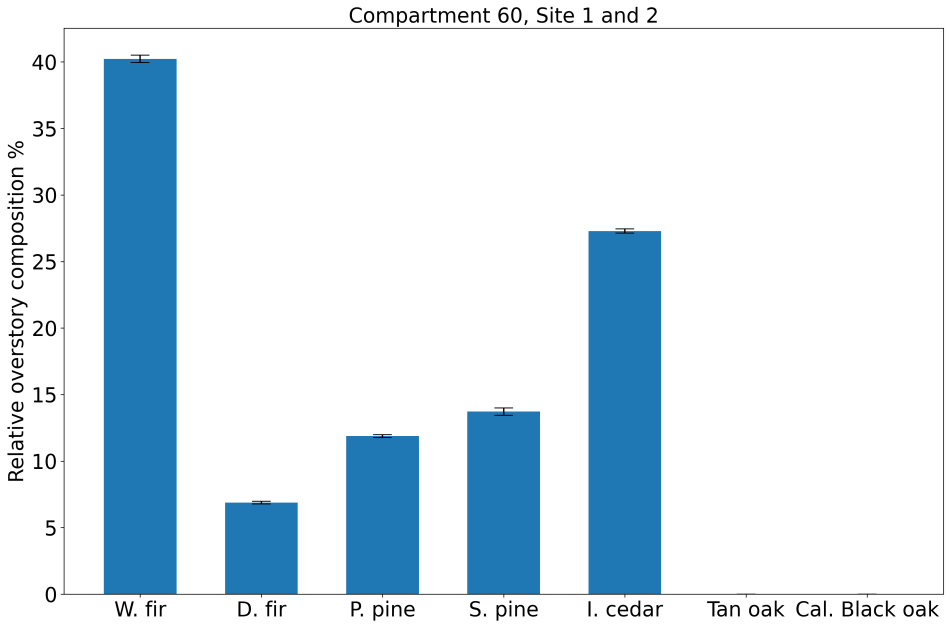
S 5. Normalized emission ratio profiles for nine selected compounds for pines, firs and spruce.



S 6. PCA coupled with k-means clustering results for nine selected compounds for the PC1 and PC2 pair.

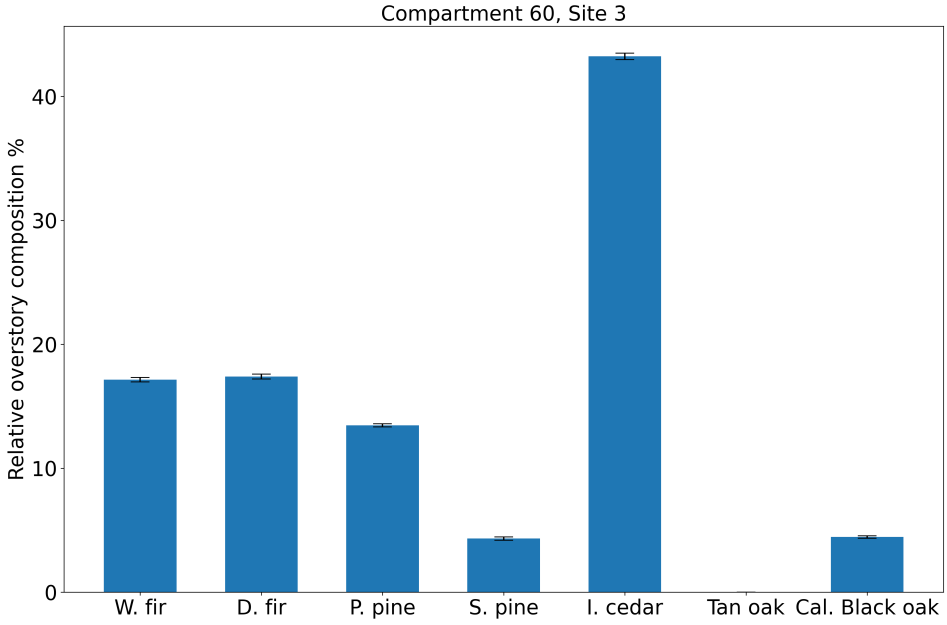


S 7. Normalized emission ratio profiles for five selected compounds for pines, firs, spruce and the synthetic 60/40 pine/fir sample.

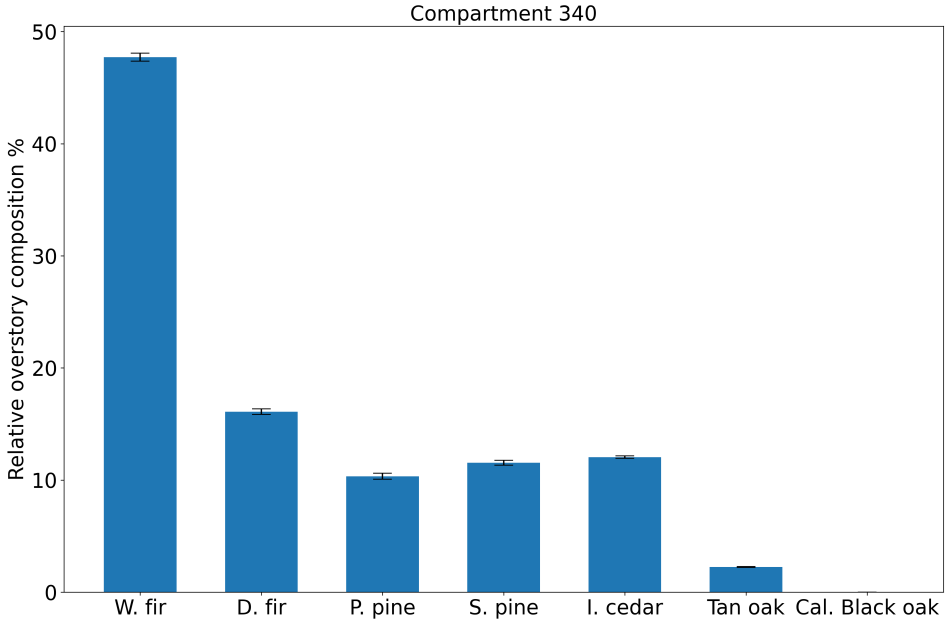


S 8. Average relative fractional composition for compartment 60 (burned on 31 October 2017) close to site 1 and 2. The fractional contribution of each species was determined by the proportion of basal area among all trees with diameter >11.4 cm at breast height. Data for the overstory composition were taken from the supplemental material in Hatch et al. (2019)

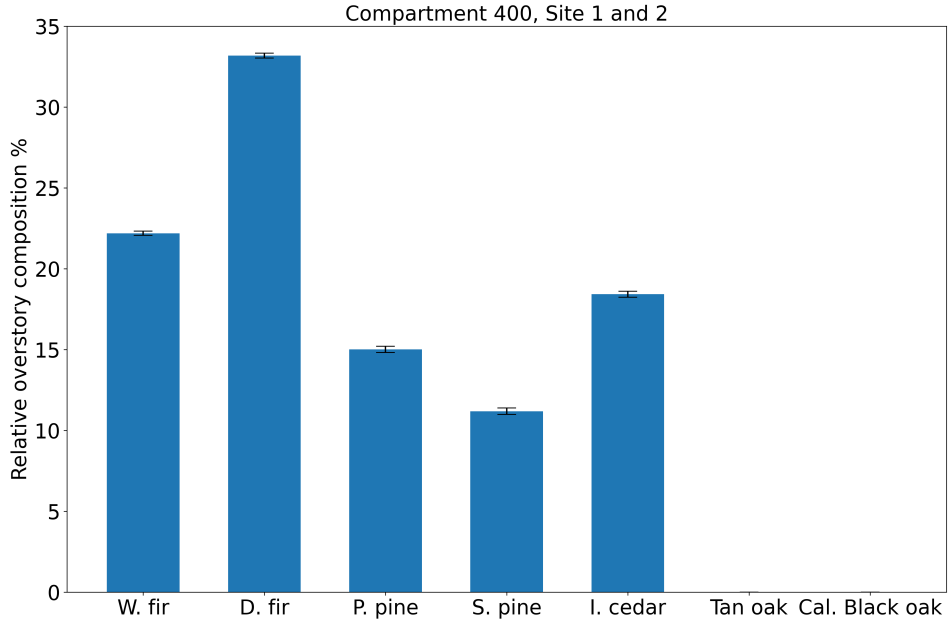
.



S 9. As in S8 but for compartment 60 (burned on 31 October 2017) site 3.



S 10. As in S8 but for compartment 340 (burned on 30 October 2017) site 2.



S 11. As in S8, for compartment 400 (burned on 01 November 2017) for sites 1 and 2.

Table 2. Sample number, fuel types and emission ratios (ppb/ppb) for retained compounds after preprocessing (*continues on the next page*)

Sample no.	Fuel type	tricyclene	a-pinene	camphene	sabinene	b-pinene	b-myrcene
Burn 10	Douglas fir	0.00992	0.01092	0.04101	0.00308	0.00195	0.00219
Burn 11	Douglas fir	0.02357	0.02191	0.06427	0.00021	0.00848	0.00718
Burn 14	Douglas fir	0.01275	0.01756	0.06219	0	0.00661	0.00463
Burn 18	Douglas fir	0.03135	0.02425	0.11072	0.0003	0.01249	0.00854
Burn 22	Douglas fir	0.00017	0.00309	0.00096	0	0.01097	0.00203
Burn 45	Douglas fir	0.04538	0.05038	0.13323	0.0046	0.02133	0.01323
Burn 57	Douglas fir	0.0097	0.0151	0.04669	0.00163	0.00351	0.00252
Burn 64	Douglas fir	0.02979	0.02479	0.04825	0.01182	0.01022	0.00932
Burn 08	Engelmann spruce	0.00051	0	0.00145	0	0	0.0013
Burn 09	Engelmann spruce	0.00057	0	0.00156	0	0.00205	0.00205
Burn 12	Engelmann spruce	0.00072	0.001	0.00136	0.00029	0.00058	0.00047
Burn 17	Engelmann spruce	0.0169	0.0465	0.09012	0.00625	0.00557	0.02764
Burn 25	Engelmann spruce	0	0.00345	0.003	0	0.00514	0.00258
Burn 26	Engelmann spruce	0.00036	0	0.00092	0	0	0.00016
Burn 36	Engelmann spruce	0.00069	0.00262	0.00125	0.0004	0.0021	0.00046
Burn 44	Engelmann spruce	0.00109	0.0034	0.00292	0	0.00147	0.00131
Burn 52	Engelmann spruce	0.00025	0.00454	0.00057	0.00011	0.0026	0.00188
Burn 54	Engelmann spruce	0.00044	0.00227	0.0029	0.00011	0.00104	0.00234
Burn 05	Lodgepole pine	0	0	0	0.00025	0.00842	0.00276
Burn 07	Lodgepole pine	0.00029	0.0068	0.00109	0.00049	0.01236	0.00785
Burn 20	Lodgepole pine	0	0.00041	0.00025	0.00015	0.01054	0.00363
Burn 40	Lodgepole pine	0.00078	0.00508	0.00173	0.00131	0.02675	0.01456
Burn 42	Lodgepole pine	0.00161	0.00357	0.00782	0.00088	0.01889	0.00692
Burn 58	Lodgepole pine	0.00014	0.00139	0	0.00037	0.00665	0.00299
Burn 63	Lodgepole pine	0.00063	0.00202	0.0021	0.0006	0.01356	0.00666
Burn 03	Ponderosa pine	0	0.00685	0.00072	0.00046	0.03033	0.00965
Burn 04	Ponderosa pine	0	0.00295	0.00058	0.00029	0.01187	0.00832
Burn 16	Ponderosa pine	0.00045	0.00754	0.00257	0.00015	0.01157	0.00601
Burn 19	Ponderosa pine	0	0.00211	0	0.00038	0.02297	0.01013
Burn 37	Ponderosa pine	0.00027	0.01244	0.00171	0.0006	0.02439	0.01592
Burn 39	Ponderosa pine	0.00051	0.02943	0.00569	0.00126	0.08736	0.03142
Burn 59	Ponderosa pine	0.00018	0.01183	0.00228	0.00057	0.02012	0.01994
Burn 72	Ponderosa pine	0.00033	0.01879	0.00311	VALUE!	0.05579	0.01587
Burn 15	Subalpine fir	0.03068	0.02024	0.07088	0.00053	0.02839	0.03026
Burn 23	Subalpine fir	0.00321	0.00085	0.01438	0	0.00369	0.00159
Burn 51	Subalpine fir	0.0203	0.01045	0.0715	0	0.00521	0.00066
Burn 56	Subalpine fir	0.00194	0.00523	0.0104	0	0.00226	0.00088
Burn 67	Subalpine fir	0.00968	0.00753	0.03566	0.00029	0.01159	0.00323
Burn 47	Subalpine fir	0.02292	0.01978	0.06436	0.00049	0.02384	0.01552

Sample no.	Fuel type	MT isomer	3-carene	p-cymene	D-Limonene	terpinolene	bornyl acetate
Burn 10	Douglas fir	0	0.00037	0.00199	0.01467	0.00082	0.01658
Burn 11	Douglas fir	0.00084	0.00075	0.00344	0.03394	0.00101	0.03628
Burn 14	Douglas fir	0.0005	0.00399	0.00194	0.02414	0.00074	0.03925
Burn 18	Douglas fir	0.00081	0	0.0024	0.03966	0.00103	0.043
Burn 22	Douglas fir	0.00024	0.00217	0.00169	0.0041	0	0
Burn 45	Douglas fir	0.00145	0.00129	0.00585	0.04843	0	0.05537
Burn 57	Douglas fir	0.00027	0.00244	0.00157	0.01411	0.00075	0.02472
Burn 64	Douglas fir	0.00114	0.00182	0.00519	0.0351	0.00611	0.04345
Burn 08	Engelmann spruce	0	0	0.00168	0.00122	0	0.00132
Burn 09	Engelmann spruce	0	0.00046	0.0023	0.00169	0	0.00101
Burn 12	Engelmann spruce	0.00388	0.00401	0.00654	0.0014	0.00056	0.00017
Burn 17	Engelmann spruce	0.00136	0.0377	0.00666	0.1423	0.00634	0.02395
Burn 25	Engelmann spruce	0	0	0.0007	0.00086	0	0
Burn 26	Engelmann spruce	0.00413	0.00365	0.00984	0.00322	0.00076	0
Burn 36	Engelmann spruce	0	0	0.0049	0.00494	0.00032	0.00049
Burn 44	Engelmann spruce	0	0	0.00096	0.00266	0	0.00079
Burn 52	Engelmann spruce	0.00016	0.00027	0.00145	0.00259	0	0.00058
Burn 54	Engelmann spruce	0.00018	0	0.00275	0.00304	0	0.00032
Burn 05	Lodgepole pine	0.00044	0.00435	0.00154	0.00202	0.00044	0
Burn 07	Lodgepole pine	0.00109	0.00571	0.00287	0.00908	0.00074	0.00032
Burn 20	Lodgepole pine	0.00063	0.00131	0.00114	0.00157	0.00036	0
Burn 40	Lodgepole pine	0.00252	0.02214	0.00634	0.01234	0.0021	0.00066
Burn 42	Lodgepole pine	0.0009	0.00838	0.00368	0.01118	0.00102	0.00628
Burn 58	Lodgepole pine	0.00055	0.00781	0.00258	0.00467	0.00067	0.00034
Burn 63	Lodgepole pine	0.0009	0.00743	0.00417	0.01474	0.00109	0.00136
Burn 03	Ponderosa pine	0.00099	0.03552	0.00251	0.01071	0.00187	0
Burn 04	Ponderosa pine	0.00132	0.0188	0.00218	0.00961	0.00125	0
Burn 16	Ponderosa pine	0.001	0.00862	0.0062	0.0164	0.00042	0.0005
Burn 19	Ponderosa pine	0.00139	0.02758	0.00389	0.00851	0.00141	0
Burn 37	Ponderosa pine	0.00189	0.02063	0.00397	0.01733	0.0018	0
Burn 39	Ponderosa pine	0.00558	0.04626	0.00725	0.04677	0.00753	0
Burn 59	Ponderosa pine	0.00214	0.02791	0.00333	0.02039	0.00266	0
Burn 72	Ponderosa pine	0.00199	0.0209	0.00441	0.02076	0.0031	0.00027
Burn 15	Subalpine fir	0.00576	0.00101	0.00533	0.02981	0.00253	0.05803
Burn 23	Subalpine fir	0	0	0.00057	0.00543	0	0.01054
Burn 51	Subalpine fir	0	0	0.00305	0.02098	0	0.0184
Burn 56	Subalpine fir	0.00119	0	0.00712	0.00658	0	0.00375
Burn 67	Subalpine fir	0.00052	0.00136	0.00262	0.01375	0.00029	0.01083
Burn 47	Subalpine fir	0.00198	0.00022	0.00568	0.03509	0.00194	0.06205

References

Hatch, L. E., Luo, W., Pankow, J. F., Yokelson, R. J., Stockwell, C. E., and Barsanti, K. C.: Identification and quantification of gaseous organic compounds emitted from biomass burning using two-dimensional gas chromatography–time-of-flight mass spectrometry, Atmos. Chem. Phys., 15, 1865–1899, https://doi.org/10.5194/acp-15-1865-2015, 2015.

Hatch, L. E., Jen, C. N., Kreisberg, N. M., Selimovic, V., Yokelson, R. J., Stamatis, C., York, R. A., Foster, D., Stephens, S. L., Goldstein, A. H., and Barsanti, K. C.: Highly Speciated Measurements of Terpenoids Emitted from Laboratory and Mixed-Conifer Forest Prescribed Fires, Environmental Science & Technology, 53, 9418–9428, https://doi.org/10.1021/acs.est.9b02612, pMID: 31318536, 2019.

Jolliffe, I.: Principal Component Analysis, Springer, 2002.

King, B. M.: Analysis of Variance, pp. 32–36, Elsevier, Oxford, https://www.sciencedirect.com/science/article/pii/B9780080448947013063, 2010.

Lever, J., Krzywinski, M., and Altman, N.: Principal component analysis, Nature Methods, 14, 641–642, https://doi.org/10.1038/nmeth.4346, 2017.