AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE

10.1002/2015JD023641

Key Points:
- O_3 profile clusters correspond to large-scale meteorological conditions
- Three clusters contain O_3 >100 ppbv above climatology near the tropopause
- Clustering captures O_3 profile variability better than climatological means

Supporting Information:
- Figures S1 and S2 and Table S1
- Figure S1
- Figure S2

Correspondence to:
R. M. Stauffer,
rms5539@psu.edu

Citation:
Stauffer, R. M., A. M. Thompson, and G. S. Young (2016), Tropospheric ozonesonde profiles at long-term U.S. monitoring sites: 1. A climatology based on self-organizing maps, *J. Geophys. Res. Atmos.*, *121*, 1320–1339, doi:10.1002/2015JD023641.

Received 6 MAY 2015
Accepted 4 JAN 2016
Accepted article online 10 JAN 2016
Published online 6 FEB 2016

Tropospheric ozonesonde profiles at long-term U.S. monitoring sites: 1. A climatology based on self-organizing maps

Ryan M. Stauffer[1,2], Anne M. Thompson[2,3], and George S. Young[2]

[1]Earth System Science Interdisciplinary Center, University of Maryland, College Park, Maryland, USA, [2]Department of Meteorology, Pennsylvania State University, University Park, Pennsylvania, USA, [3]Earth Sciences Division, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

Abstract Sonde-based climatologies of tropospheric ozone (O_3) are vital for developing satellite retrieval algorithms and evaluating chemical transport model output. Typical O_3 climatologies average measurements by latitude or region, and season. A recent analysis using self-organizing maps (SOM) to cluster ozonesondes from two tropical sites found that clusters of O_3 mixing ratio profiles are an excellent way to capture O_3 variability and link meteorological influences to O_3 profiles. Clusters correspond to distinct meteorological conditions, e.g., convection, subsidence, cloud cover, and transported pollution. Here the SOM technique is extended to four long-term U.S. sites (Boulder, CO; Huntsville, AL; Trinidad Head, CA; and Wallops Island, VA) with 4530 total profiles. Sensitivity tests on *k*-means algorithm and SOM justify use of 3×3 SOM (nine clusters). At each site, SOM clusters together O_3 profiles with similar tropopause height, 500 hPa height/temperature, and amount of tropospheric and total column O_3. Cluster means are compared to monthly O_3 climatologies. For all four sites, near-tropopause O_3 is double (over +100 parts per billion by volume; ppbv) the monthly climatological O_3 mixing ratio in three clusters that contain 13–16% of profiles, mostly in winter and spring. Large midtropospheric deviations from monthly means (−6 ppbv, +7–10 ppbv O_3 at 6 km) are found in two of the most populated clusters (combined 36–39% of profiles). These two clusters contain distinctly polluted (summer) and clean O_3 (fall-winter, high tropopause) profiles, respectively. As for tropical profiles previously analyzed with SOM, O_3 averages are often poor representations of U.S. O_3 profile statistics.

1. Introduction

1.1. Ozone Climatologies

Since the 1960s, the global ozonesonde network has provided a comprehensive O_3 data set of increasing spatial coverage and density, as well as recent quantification of short time scale processes such as pollution transport [e.g., *Cooper et al.*, 2011] through campaign-based networks [*Thompson et al.*, 2011]. Campaign sonde networks also capture the evolution of stratosphere-to-troposphere exchange (STE) [*Holton et al.*, 1995; *Lin et al.*, 2012] events. STE greatly affects the O_3 profile shape on short time scales with pronounced O_3 and potential vorticity correlations in the upper troposphere [*Danielsen*, 1968; *Rao et al.*, 2003]. Ozonesondes provide the highest vertical resolution measurements of O_3 available from the surface to above 30 km, at accuracies as high as ±5% [*Komhyr et al.*, 1995]. For these reasons, ozonesondes are the preferred reference measurements with which to compare chemical model output and satellite O_3 profile and column retrievals.

There have been efforts to establish global O_3 climatologies for comparison with model output and satellite measurements. These studies have relied heavily on the global ozonesonde network, using climatology as a baseline for trends [*Logan*, 1985, 1994; *Logan et al.*, 1999, 2012; *Oltmans et al.*, 2006, 2013], O_3 distribution in latitudinal bands [*Stevenson et al.*, 2006], and climatology for specific regions [*Newchurch et al.*, 2003; *Tilmes et al.*, 2012] including the tropics [*Thompson et al.*, 2003a, 2003b, 2012]. Climatologies from ozonesondes and satellite retrievals in the stratosphere have also been developed to increase accuracy of total column O_3 integration from ozonesonde profiles [*McPeters et al.*, 1997; *McPeters and Labow*, 2012]. Model and satellite performance is judged primarily on replication of seasonal variability at one location or region, particularly in the upper troposphere/lower stratosphere [e.g., *Considine et al.*, 2008].

Recently, *Tilmes et al.* [2012] assembled a global O_3 climatology from 42 ozonesonde sites from 1980–1994 and 1995–2011. Their analysis separated the stations into 12 regions that exhibited similar O_3 probability

density functions. They demonstrated an application of the climatology via the improvements in Community Atmosphere Model with Chemistry [*Lamarque et al*., 2012] model simulations that used derived stratospheric O_3, as opposed to a monthly and latitudinally invariant stratospheric O_3 climatology. There are many processes, however, such as synoptic-scale wave dynamics, that can cause significant deviations in the profile from a typical O_3 climatology. Thus, an investigation into these processes is performed using a technique that tends to classify O_3 profiles according to meteorological conditions and other influences on tropospheric O_3 profile shape.

1.2. Ozone Profile Clustering

Two studies in particular set a precedent for clustering ozonesonde profiles. *Diab et al*. [2004] classified over 100 ozonesonde profiles launched from late 1998 to 2002 from a subtropical Southern Hemisphere Additional Ozonesondes (SHADOZ) [*Thompson et al*., 2003a] site, Irene, South Africa. Their analysis yielded six clusters including distinct "background" and "polluted" clusters, containing well below and well above average tropospheric O_3 mixing ratios. *Diab et al*. [2004] also found a cluster containing 48% of all profiles, which could not be ascribed to a particular meteorological regime, season, or source region. They labeled this cluster as representative of "typical" Irene O_3, arguing that the representative cluster is more informative and descriptive of Irene O_3 than a mean profile because it is not influenced by extreme values and is not necessarily confined to a particular season. Their clustering analysis also allowed identification of STE/low tropopause height O_3 profiles, the majority of which occurred during the Southern Hemisphere winter, when influences more characteristic of the midlatitudes, namely, the subtropical jet, frequently affect Irene.

Jensen et al. [2012] performed a cluster analysis on over 900 tropical ozonesonde profiles. They employed self-organizing maps (SOM) [*Kohonen*, 1995], which have been used as a clustering algorithm across many disciplines, including several recent meteorology and climate studies [*Hewitson and Crane*, 2002; *Hong et al*., 2004; *Liu et al*., 2006; *Nowotarski and Jensen*, 2013]. *Jensen et al*. [2012] used SOM to describe influences dictating O_3 variability at two SHADOZ stations, Natal, Brazil, and Ascension Island, from 1998 to 2009. Their four-cluster results were similar to those found in *Diab et al*. [2004] and were dominated by the seasonal influences of biomass burning and convection. Clusters representing a background state, a polluted state, and a mean state cluster with a plurality of profiles, were found at both locations. The polluted clusters corresponded to the African biomass burning season in the Southern Hemisphere spring, leading to sharp O_3 gradients above the boundary layer and large midtropospheric O_3 amounts. The clean clusters contained launches primarily in the convective season, during which near-surface, low-O_3 tropical air is lifted into the free troposphere.

Because of such source and synoptic effects that govern O_3 profile evolution throughout the year, clusters of O_3 profiles may be a better way to present a site's O_3 profile variability than monthly or seasonal averages. Thus, we are motivated to extend these techniques to data from several long-term midlatitude ozonesonde sites. Following the approach of *Jensen et al*. [2012], SOM is applied to Contiguous United States (CONUS) tropospheric ozonesonde profiles. CONUS represents a somewhat confined, but varied geographic area, with thousands of high-quality O_3 profiles from decades-long records available. The extremes of short-term vertical variability of O_3 in the midlatitudes are much greater than in the tropics, presenting a new challenge in interpreting the O_3 clustering statistics.

Our goals for this study are the following:

1. We aim to cluster tropospheric O_3 mixing ratio profiles at four CONUS sites using SOM. SOM is also evaluated against the similar *k*-means clustering algorithm to determine which method to apply in this paper. Sensitivity tests comparing the two methods and supporting the decision to use SOM are presented in Appendix A, as is as a technical discussion of both methods.
2. We aim to provide meteorological and geophysical interpretations of SOM clusters and organization. Although chemical processes probably play a role in SOM classification, a scarcity of colocated trace gas data precludes characterization of chemistry's added influence.
3. We aim to evaluate O_3 variability at each site by assessing the representativeness of a monthly O_3 profile climatology, focusing on deviations from the monthly climatology, midtropospheric O_3, and the near-tropopause region.

Table 1. CONUS Ozonesonde Sites Used in This Study[a]

Location	Latitude/Longitude (deg)	Altitude (m)	Length of Record	# of Profiles
Boulder, CO	40.0/−105.3	1734	1979–2013	1376
Huntsville, AL	34.7/−86.6	196	1999–2012	686
Trinidad Head, CA	40.8/−124.2	20	1997–2012	868
Wallops Island, VA	37.9/−75.5	13	1970–2013	1600

[a]Latitude and longitude, altitude amsl, record length, and number of profiles used are shown. Note that the Wallops Island site was moved slightly (from 4 m elevation) to its current location listed in the table in October 1982.

2. Sonde Measurements and Analysis Techniques

There are four ozonesonde sites with records of more than 15 years in CONUS: Boulder, CO; Huntsville, AL; Trinidad Head, CA; and Wallops Island, VA (Table 1). *Newchurch et al.* [2003] examined data from these four sites and compiled the first CONUS ozonesonde climatology using data from April 1995 to March 2002. Our analysis extends their data set back in time to include the beginning of the Boulder and Wallops Island records and adds a decade of observations beyond 2002 to each of the four sites. In this study, variability of O_3 is described without the constraints of monthly averages; rather they are used as context in this case. This method filters background or polluted O_3 cases and variations in tropopause height that are otherwise diluted by averaging.

The CONUS locations in this study span about 6° of latitude, and each is surrounded by unique terrain and experiences different regional influences. Boulder is just downwind of the Rocky Mountains and part of the Denver metro area. Huntsville, the southernmost site, is located in the southeast U.S. and exhibits more subtropical characteristics compared to the other sites [*Newchurch et al.*, 2003; *Tilmes et al.*, 2012]. Trinidad Head is located on the coast of northern CA, is influenced by marine air masses from the Pacific, and of the four CONUS sites is impacted most frequently by STE [*Newchurch et al.*, 2003]. Wallops Island is located on the Atlantic coast of the Delmarva Peninsula in southeast Virginia, often downwind of large emissions sources in the Ohio River Valley and the Baltimore/Washington D.C. region. For reference, *Tilmes et al.* [2012] combined Huntsville and Wallops Island into an "Eastern U.S." region, while Boulder and Trinidad Head each remained isolated and unassigned to a regional grouping of sonde stations. Much like *Tilmes et al.* [2012], O_3 profile shapes and distributions are used to characterize ozonesonde sites. However, instead of developing a new O_3 climatology from SOM, the tendency of an O_3 climatology to describe clusters of O_3 profiles is evaluated. This is accomplished through use of a stricter, monthly O_3 climatology as opposed to a seasonal one as in *Tilmes et al.* [2012].

2.1. Data

Ozonesonde data from the four CONUS locations in Table 1 were accessed through either the World Ozone and Ultraviolet Radiation Data Centre (WOUDC; ftp://ftp.tor.ec.gc.ca/pub/woudc/; Wallops Island, VA; portions of Boulder, CO) or NOAA Earth System Research Laboratory Global Monitoring Division (ESRL GMD; ftp://ftp.cmdl.noaa.gov/data/ozwv/Ozonesonde/; portions of Boulder, CO; Huntsville, AL; and Trinidad Head, CA). Ozonesondes are launched approximately weekly with each month of the year well represented at all sites (Figure 1). There are occasional increases in frequency for measurement campaigns, resulting in a total sample size of 4530 profiles.

The ozonesondes in this study use the electrochemical cell instrument and processing technique described in *Komhyr* [1969]. Typical uncertainties of ozonesonde measurements range from −7 to +17% in the troposphere, to ±5% in the stratosphere [*Komhyr et al.*, 1995]. All profiles include measurements of pressure, temperature, and O_3 partial pressure. If not included with the standard measurements provided for each profile, variables such as O_3 mixing ratio, geopotential altitude, and potential temperature are calculated from existing data. Vertical resolution in the data varies over two of the long-term records. For example, vertical resolutions of 250 m are available for Boulder, CO, launches from 1979 to 1989. Vertical resolutions of approximately 250–350 m (derived by recording one data point per minute) are available for Wallops Island, VA, from 1970 to 1995. The remainder of the data have resolutions at or better than 100 m. Accounting for response time of the ozonesonde and the ascent rate of the balloon, the true vertical resolution of the O_3 measurements is approximately 100–150 m. For uniformity, data from all sondes were interpolated linearly to 100 m.

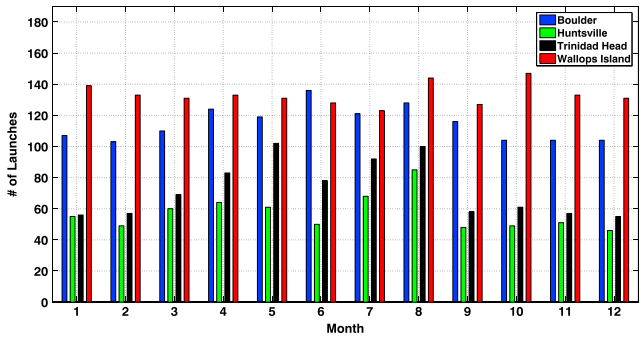

Figure 1. Histogram of number of launches contained in each month for every site.

Ancillary meteorological data were added to assist the geophysical interpretation of the ozonesonde profile clusters. HYSPLIT (Hybrid Single Particle Lagrangian Integrated Trajectory) [*Draxler and Hess*, 1997] back trajectories were computed starting at the time and location of each ozonesonde profile. The HYSPLIT trajectories were forced with National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis [*Kalnay et al.*, 1996], which is available globally from 1948 to present with 17 pressure levels and 2.5° × 2.5° horizontal resolution. Meteorological variables of temperature, potential vorticity (PV), and geopotential height were extracted at four levels (850, 700, 500, and 250 hPa) from the European Centre for Medium-Range Weather Forecasts Interim Reanalysis (ERA-Interim) [*Dee et al.*, 2011]. Mean sea level pressure (MSLP), total cloud cover, and 2 m temperature were also analyzed. ERA-Interim data are available globally from 1979 to present with 37 pressure levels and a horizontal resolution of ~0.7° × 0.7°.

2.2. Self-Organizing Maps

Presented here is a brief introduction to SOM clustering. Additional discussion of user-selectable SOM parameters appears in Appendix A. Appendix A also contains sensitivity tests and arguments that explain the selection of SOM over *k*-means for clustering, as well as the choice of SOM map size.

SOM, developed and described by *Kohonen* [1995], is an artificial neural network, that is, a network of "nodes" or "neurons" that learn from and are used to represent an input data set. SOM is often used for data visualization or dimensional reduction of the original data set [*Liu et al.*, 2006, and references therein]. In the application to CONUS O_3 mixing ratio profiles, SOM is employed primarily as a clustering algorithm.

The SOM algorithm is configured via a user-specified map (network) size and shape that dictates the number and relationship of the nodes that will represent the data. The map can be of any dimension, with 2-D maps (e.g., a 4 × 4 map of 16 nodes) preferred in recent related meteorological applications [e.g., *Jensen et al.*, 2012; *Nowotarski and Jensen*, 2013]. The initial values of the nodes, analogous to cluster centroids in *k*-means, can be obtained in a number of ways. Here a principal component analysis (PCA) decomposition of the input data set yields a subspace across which the initial nodes are distributed over a rectangular grid. This linear initialization approach is taken so as to cover as much of the input data set variability as possible in the array of initial nodes. SOM nodes can also be initialized randomly as in *k*-means, although randomly initialized SOM maps converge more slowly and have larger error [*Liu et al.*, 2006]. Linear initialization also guarantees that the same map is consistently produced for a given set of inputs and is thus preferred here.

Once the SOM nodes are initialized, the SOM algorithm is executed on the input data set in either the batch or sequential mode. These two iterative update modes result in similar SOMs, and the batch algorithm is much more computationally efficient [*Vesanto et al.*, 2000], and so batch is used in this study. The SOM algorithm clusters the input data (i.e., O_3 profiles) in such a way that each cluster is most similar to those holding adjacent positions in the map. This feature is utilized in section 3. Our study uses code from the Matlab SOM Toolbox described in *Vesanto et al.* [2000], available for download from the Helsinki University of Technology, Finland (http://www.cis.hut.fi/projects/somtoolbox/). Further discussion of the SOM algorithm, comparisons and sensitivity tests with *k*-means clustering, and optimization of SOM geometry and topology for the CONUS sondes appear in Appendix A.

3. Results and Discussion

Average monthly O_3 mixing ratio profiles (surface to 12 km above mean sea level (amsl)) at the four CONUS locations are first presented (Figure 2). The least seasonal variability throughout the lower to middle troposphere is observed at Trinidad Head, with O_3 mixing ratios generally averaging between 50 and 65 parts per billion by volume (ppbv) from 2 to 6 km throughout the year. Larger seasonal variability in low to middle

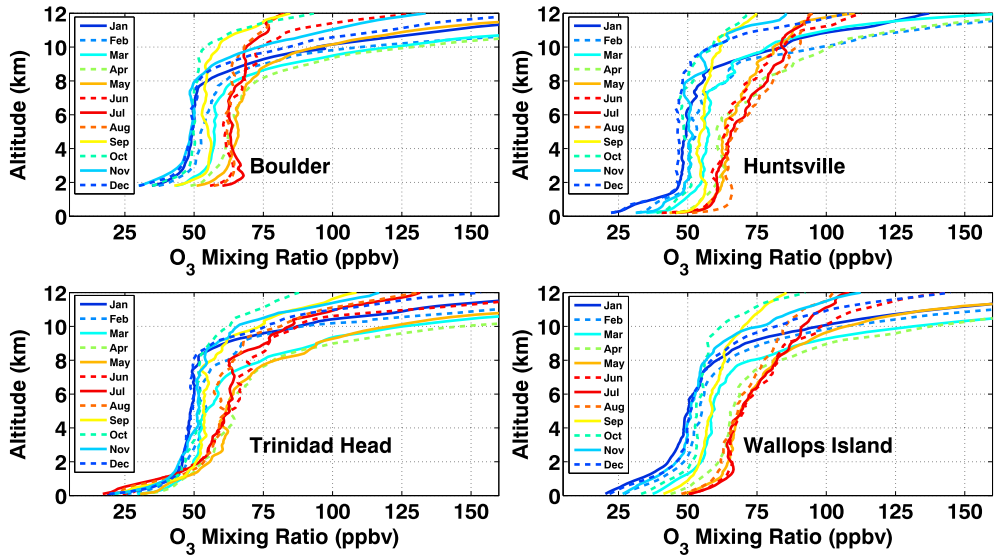

Figure 2. Monthly averaged O_3 mixing ratio profiles for each site from the surface to 12 km amsl.

troposphere O_3 is observed at Huntsville and Wallops Island (50–75 ppbv). Higher summer O_3 mixing ratios at the two latter sites reflect intermittent transport of photochemical pollution from upwind regional sources. Boulder exhibits less extreme seasonal variability. The small yearly range in low to middle tropospheric O_3 at Trinidad Head was first observed in the shorter (4.5 years) record examined by *Newchurch et al*. [2003]. Trinidad Head also exhibits the lowest concentrations of near-surface and boundary layer O_3 ($<$40 ppbv). Although Trinidad Head is often influenced by clean, marine air masses [*Newchurch et al*., 2003], the measurements may also be affected by local launch times that are one to three solar hours earlier than those of the other sites. The majority of launches occur from 17 to 19 UTC at all locations, equating to 9–11 local standard time at the west coast Trinidad Head site.

The seasonal cycle in tropopause height, seen as sharp O_3 increases near 10–12 km in Figure 2, displays a minimum in late winter/spring (March-April-May), and maximum in the fall (September-October-November). In this study, tropopause height is calculated using the thermal lapse rate tropopause as defined by the World Meteorological Organization (WMO): the lowest altitude at which the temperature lapse rate increases to $-2\,\mathrm{K\,km^{-1}}$ or less and persists for a depth of at least 2 km [*World Meteorological Organization*, 1957].

3.1. The 3 × 3 SOM Cluster Results

In order to avoid the complexity of O_3 gradients in the tropical tropopause layer, *Jensen et al*. [2012] ran SOM on profiles to 15 km amsl altitude. In contrast, we wish to capture variability in the tropopause altitude; therefore our SOM uses O_3 mixing ratio data from the surface to 12 km amsl. At our CONUS locations, 12 km altitude is sufficient to include seasonal tropopause altitude variability in the SOM clusters and encompasses most or all of the troposphere. This altitude ceiling also prevents stratospheric O_3 mixing ratios from dominating the SOM clusters. The 3 × 3 SOM output (nine nodes/clusters) for Wallops Island is shown in Figure 3. Clusters of individual profiles (dark blue) corresponding to each SOM node (red) are plotted along with the entire data set mean and twentieth and eightieth percentile O_3 (cyan) for comparison. SOM nodes are labeled 1–9 and will be referred to by number when discussing the characteristics of each O_3 profile cluster. Each SOM node is identical to the mean of its respective cluster. A major advantage of SOM over other clustering algorithms is adjacency of like nodes (e.g., 2 and 3, Figure 3) and the separation of contrasting nodes (e.g., 3 and 7). This allows us to visualize subtle differences between the neighboring clusters of O_3 profiles and distinguishes unique characteristics of nodes and groups of nodes through variation of specific features across the SOM map. For example, traversing nodes 1–3 shows a lowering of the altitude of tropopause O_3 gradients; O_3 is well above the mean and eightieth percentile O_3 near 8–12 km in these nodes. Likewise, a distinct rise in the amount of lower tropospheric O_3 from nodes 7 to 9 is observed. Node 7 contains profiles with O_3 below the twentieth percentile and $<$50 ppbv through nearly the entire surface to 12 km profile.

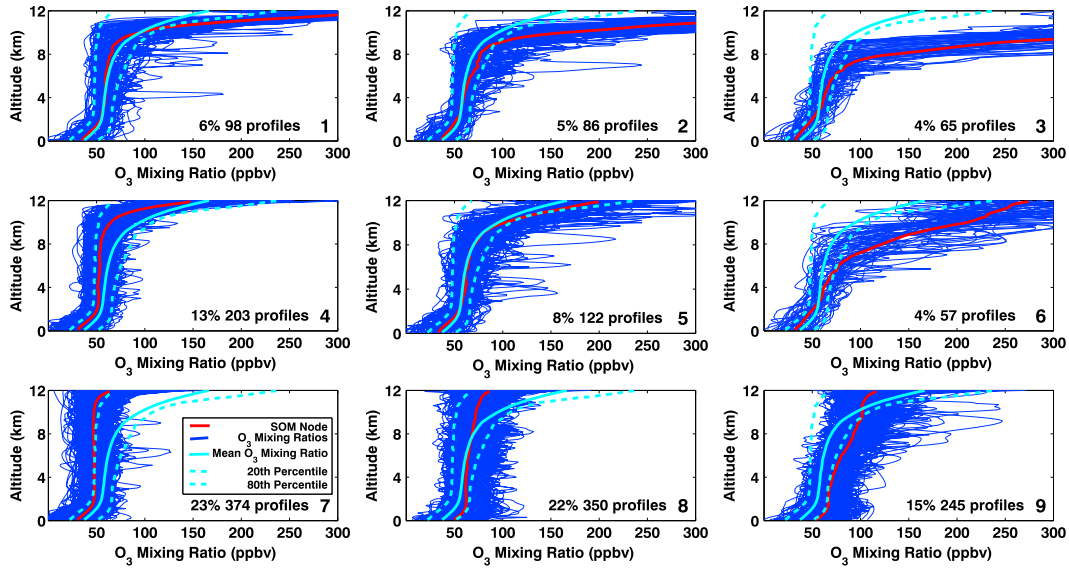

Figure 3. The 3 × 3 SOM output for Wallops Island, VA. SOM nodes are shown in red, with the corresponding individual O_3 mixing ratio profiles in dark blue. For reference, the overall site average O_3 mixing ratio profile and twentieth and eightieth percentile O_3 are shown in cyan.

Node 8 contains near-average O_3 in the low to middle troposphere, and node 9 exceeds 70 ppbv O_3 and remains above the eightieth percentile through nearly the entire troposphere.

The percentage and total number of profiles corresponding to each node quantifies the frequency with which each O_3 profile type is observed, and which cluster(s) may be most representative of a site's typical O_3 profile. At Wallops Island, the top row of nodes 1–3 contains just 15% of all profiles, whereas the bottom row 7–9 contains 60% and generally contains data below the steep tropopause O_3 gradients.

All nine SOM nodes from surface to 12 km O_3 mixing ratios for each of the four CONUS sites appear in Figure 4. The topological ordering and shapes of the SOM nodes are nearly identical. This similar topological ordering from SOM clustering results from the linear initialization of SOM nodes. Differences in nodes 5 and 6 between Huntsville and the other sites is evident in Figure 4, presumably stemming from Huntsville's higher average tropopause and the subtropical-like characteristics noted in *Newchurch et al.* [2003] and *Tilmes et al.* [2012]. Much like the Wallops Island SOM in Figure 3, there is a lowering of the altitude of O_3 gradients indicating the tropopause across nodes 1–3. Nodes 1–3 represent 13–16% of each site's profiles. The increasing amount of lower tropospheric O_3 across nodes 7–9, corresponding to 57–61% of profiles, is also prominent. Node 7 (Figure 4) represents a background state at every site, with O_3 averaging <50 ppbv throughout most of the troposphere, whereas node 9 contains polluted profiles with well above average low to middle tropospheric O_3.

3.2. Seasonal and Meteorological Influences on SOM Nodes

Given the meteorological and seasonal influences on O_3 profiles in the midlatitudes, the SOM nodes at the CONUS ozonesonde locations are expected to correspond to seasonality, midlatitude ridge and trough Rossby wave patterns, and column O_3 amounts. The number of profiles from each month in each SOM node is expressed as a percentage and is shown in Figure 5. Many of the clusters contain launches from only a few months. However, profiles from several nodes (e.g., 4–6) exhibit no distinct seasonality and are found throughout the year. The altitude of the tropopause O_3 gradient and the photochemical O_3 season both contribute to the node/seasonality relationship. The lowest tropopause altitudes, predominantly found in nodes 1–3, occur mainly in late winter and spring. During these seasons, increased latitudinal temperature gradients and synoptic-scale Rossby wave dynamics cause large meanders in the polar jet, with associated lower tropopause altitude. At every site, >90% of all profiles in node 3 occurred between January and May. Conversely, tropopause O_3 gradients are generally not found below 10 km in profiles corresponding to the bottom SOM row. Rather, the low to middle tropospheric O_3 increase across nodes 7–9 (~50 to 60 to 70 ppbv in Figure 4) represents a sequence of increasing photochemical O_3 pollution. The majority of profiles in node

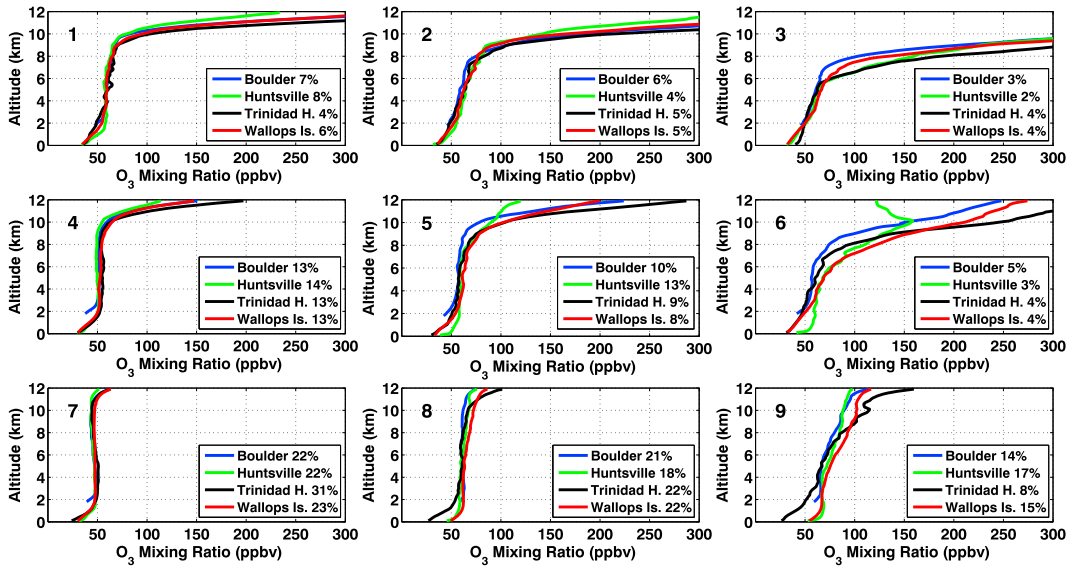

Figure 4. The 3 × 3 SOM nodes for each of the four CONUS sites shown as O_3 mixing ratio profiles. SOM nodes are labeled from 1 to 9 with the percentage of each site's profiles corresponding to each node shown in the legend.

7 occur in fall and winter, whereas >80% of profiles in node 9 occur between May and August at all sites. Figures 4 and 5 show that there is reduced tropospheric O_3 pollution year round at Trinidad Head compared to the other stations. Nearly 30% of node 7 profiles from Trinidad Head were launched in June-July-August (JJA). This is a far greater portion of summer launches in node 7 than for any other site. Fewer than 5% of launches in node 7 are from JJA at Huntsville and Wallops Island. Trinidad Head also includes the greatest percentage of its total launches in node 7 (31%) compared to other sites and the fewest in the polluted node 9 (8%).

Many of the remaining SOM node clusters are difficult to explain through seasonality and tropopause height alone and will require analysis of additional sources of data. Meteorological information examined from the radiosondes attached to the corresponding ozonesonde is used to infer influences on the SOM node O_3 clusters. Radiosonde and ozonesonde data from SOM clusters are compared against monthly climatological means for each site. So that all CONUS locations can be considered together, an anomaly approach is taken.

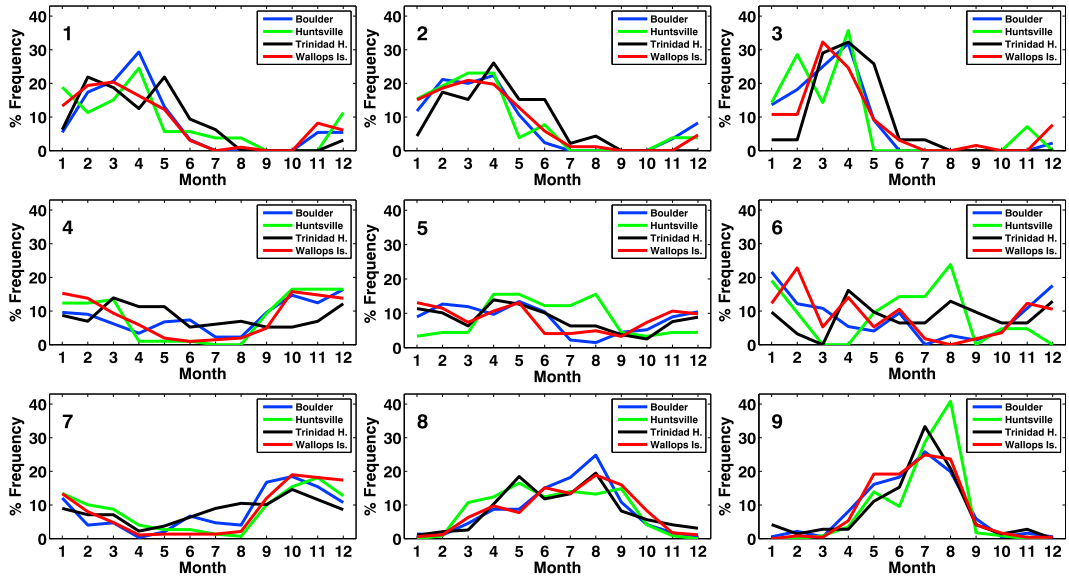

Figure 5. Seasonality of SOM nodes 1–9 shown as the relative frequency of month within each SOM node. Each of the nine histograms totals 100% at every site.

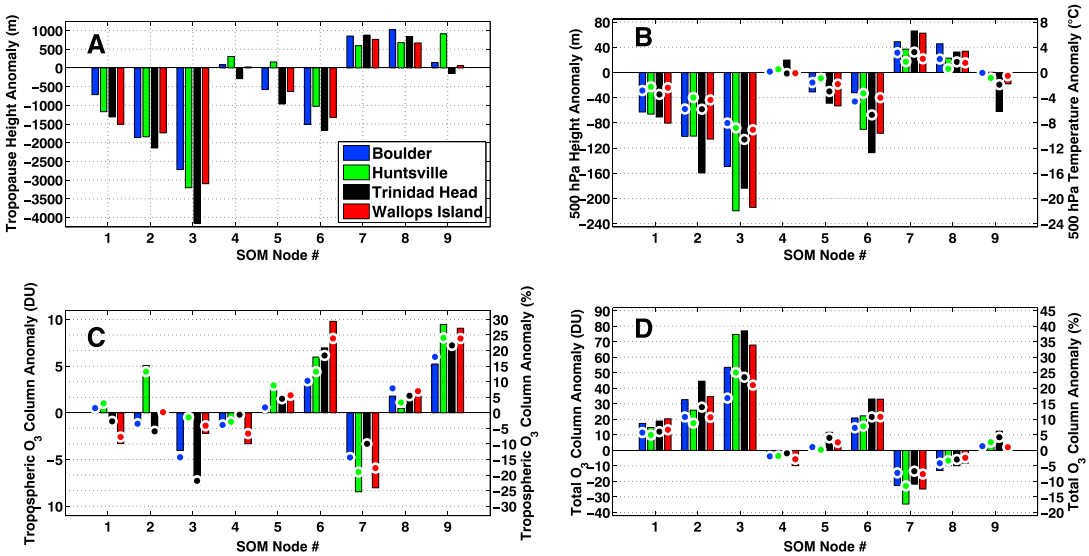

Figure 6. (a) Average tropopause height anomaly (in meters) from monthly climatology for each SOM node 1–9. (b) Average 500 hPa geopotential height anomaly (meters, bars, left axis) and 500 hPa temperature anomaly (°C, dots, right axis) from monthly climatology for each SOM node 1–9. (c) Average tropospheric column O_3 anomaly amount (DU, bars, left axis) and percentage (%, dots, right axis) from monthly climatology for each SOM node 1–9. (d) Average total column O_3 anomaly amount (DU, bars, left axis) and percentage (%, dots, right axis) from monthly climatology for each SOM node 1–9. Methodology for calculating anomalies is given in section 3.2.

Our approach to calculating anomalies from monthly climatology is as follows: (1) Calculate the climatology with 12 monthly means using a site's entire profile data set for the variable of interest, (2) calculate the variable of interest for every profile in each SOM node and compare to its corresponding monthly mean climatology (measurement—climatology), and (3) average the results of all profiles within each SOM node. Average anomalies for each node at each site are calculated. The result of applying this technique to the WMO lapse rate tropopause is shown in Figure 6a. Nodes 1–3 represent an incremental lowering of the tropopause altitude, with the lowest relative tropopause averaging over 4000 m lower than climatology in node 3 at Trinidad Head. Nodes 4 and 5 contain about average or slightly lower than average tropopause heights, whereas node 6 tropopause heights lie >1000 m below climatology at each site. Nodes 7 and 8 appear to be related in that they contain similar tropopause height anomalies, with a slight increase in low to middle tropospheric O_3 from nodes 7 and 8 as seasonality shifts from mainly fall-winter to spring-fall. The polluted, summertime node 9 tropopause altitudes are close to climatology, except at Huntsville, which averages 900 m higher. Huntsville's uniqueness is also displayed in node 5, a consequence of its more subtropical characteristics and distinct O_3 profiles and seasonality (Figures 4 and 5).

Anomalies of 500 hPa geopotential height and temperature (Figure 6b) are similar to the tropopause height anomalies (Figure 6a). Nodes 1–3 are associated with lower and progressively colder 500 hPa surfaces, suggesting that these profiles are influenced by Rossby wave troughs through most of the troposphere. Notably, Huntsville contains the largest 500 hPa height anomalies in node 3. Many of the 500 hPa surfaces in node 3 at Huntsville lie below 5.5 km. These are some of the lowest 500 hPa heights in the entire CONUS ozonesonde record. However, node 3 O_3 profiles represent only 2% of Huntsville's data set. As in Figure 6a, nodes 4, 5, and 9 lie close to climatological mean 500 hPa heights. The large-scale ridge pattern implied by the positive 500 hPa heights in node 7 is an indication of subtropical influence and, as a result, lower O_3 amounts. The 500 hPa heights and temperatures in node 6 are well below average and in line with its low tropopause height. With the evidence presented thus far, node 6 seems to be a miscellaneous cluster with highly variable O_3 profiles and unclear seasonality. However, node 6 contains <5% of data at each site. This cluster appears to be equivalent to the wastebin taxon in biology [e.g., *Friedman and Brazeau*, 2011].

3.3. Column Ozone Anomalies From Monthly Climatology

Using the same anomaly approach as in Figures 6a and 6b, integrated tropospheric column O_3 anomalies, calculated from the surface to the tropopause, for each SOM node (Figure 6c) are presented in Dobson units (1 DU = 2.69×10^{16} molecules cm^{-2}). The tropopause is defined using the same WMO lapse rate definition.

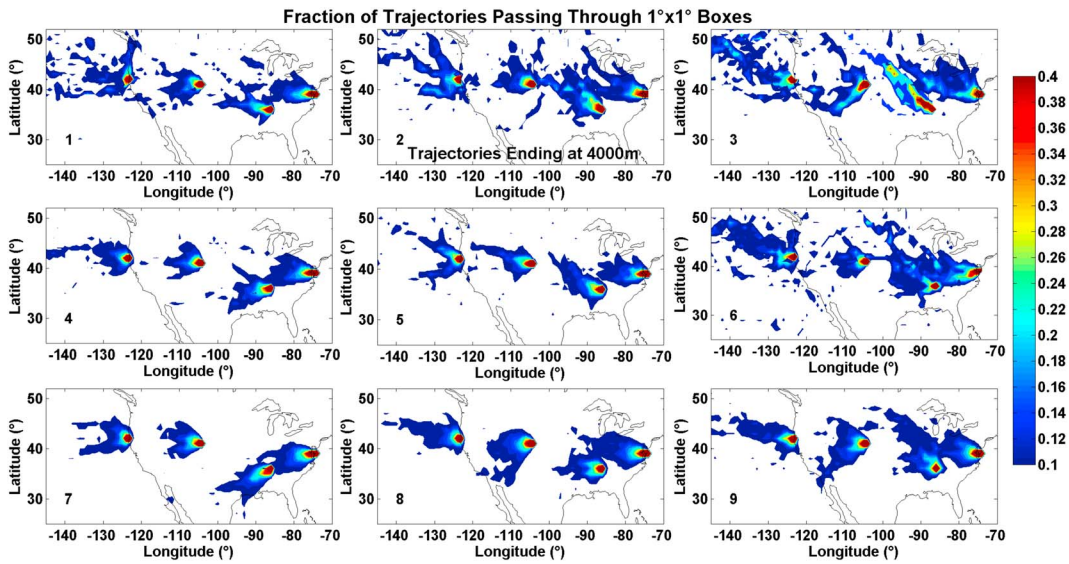

Figure 7. Contoured heat map of HYSPLIT back trajectories terminating at 4000 m at time and location of every O_3 profile. Data are contoured based on fraction of trajectories passing through 1° × 1° grid boxes. Contours are drawn every 0.02 from 0.10 to 0.40.

Tropospheric column O_3 anomalies for nodes 1–3 display no pattern, unlike the meteorological anomalies in Figures 6a and 6b. There is no clear relationship between the low tropopause/500 hPa heights and tropospheric column amount. Average tropospheric column O_3 anomalies generally lie within a few DU or ±10% of the climatology for most sites in nodes 1–3. However, distinct patterns in tropospheric O_3 amount are observed for nodes 7–9. The tropospheric O_3 increase from Figure 4 is prominent in tropospheric column anomalies in Figure 6c across node 7 (−4 to −8 DU; −10 to −20%), node 8 (+2 DU; +5%), and node 9 (+5 to +9 DU; +17 to +25%). Nodes 4 and 5 (Figure 6c) display near-climatological values, with tropospheric O_3 anomalies averaging ±2 DU (±5%). Stratospheric O_3 intrusions may contribute to the node 6 profiles, given the +4–10 DU (+10–25%) anomalies occurring in conjunction with low tropopause heights. Node 6 also contains the largest O_3 DU km^{-1} values in the troposphere at all sites.

Total column O_3 is calculated from each sonde to investigate the synoptic meteorological influences on the integrated profile. Total column O_3 from the ozonesondes is derived by first integrating the O_3 profile from surface to balloon burst or 10 hPa, whichever is higher in pressure (lower in altitude). The 10 hPa cut off has been shown to reduce errors in the total column O_3 calculation resulting from increasing measurement uncertainties in the midstratosphere [e.g., *Stauffer et al.*, 2014]. The *McPeters and Labow* [2012] above-balloon burst O_3 column climatology is then added to the ozonesonde column O_3 amount, yielding a total column O_3 amount. The *McPeters and Labow* [2012] O_3 climatology is based on a combination of ozonesonde and Aura Microwave Limb Sounder climatology. Profiles that did not reach 30 hPa were discarded.

The resulting total column O_3 anomalies in Figure 6d reflect the tropopause height anomalies in Figure 6a. Nodes 1–3 contain increasing total column O_3 corresponding to the lowering tropopause, yielding a deeper stratosphere. Node 3 profile columns frequently exceed 400 DU, representing a ~55–75 DU (~15–25%) increase in total column O_3 over climatology. An increase of about 20–30 DU (10%) above average O_3 appears in the highly variable O_3 profiles in node 6. Nodes 4, 5, 8, and 9 contain near-average total column O_3 within ±5% of climatology. Node 7 is the only cluster with notably low total column O_3, 20–35 DU below the climatological average.

3.4. Meteorological Interpretations

The meteorological and seasonal influences on nodes 1–3 (winter/spring, low tropopause O_3 profiles) and 7 (fall/winter, high tropopause/500 hPa heights, subtropical influence), are obvious. However, ancillary data are required to interpret the remaining SOM nodes. A contoured heat map of HYSPLIT back trajectories ending at 4 km provides a summary of source regions for all SOM nodes and sites (Figure 7). The 4 km altitude was chosen because this altitude is typically located in the free troposphere, well above effects from boundary layer processes yet low enough to avoid the largely zonal winds at higher altitude that result from thermal wind

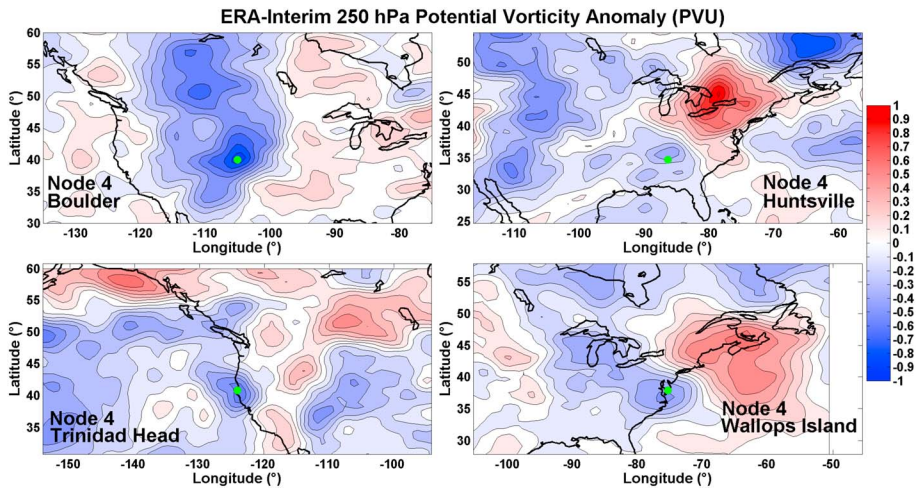

Figure 8. Contoured map of average ERA-Interim 250 hPa PV (PVU) anomalies from monthly climatology for node 4 at each site. Data are contoured every 0.1 PVU from −1 to 1 PVU. Blue colors represent negative anomalies, and red colors represent positive anomalies. The green dot represents the site location.

balance. Distinct cyclonic curvature in the trajectories is evident in nodes 2, 3, and 6 in Figure 7, confirming that large-scale troughs (e.g., 500 hPa heights, Figure 6b) are the driving force behind the profiles in those clusters. In node 7 at Huntsville and Wallops Island, anticyclonic curvature and a source region to the southwest illustrate the previously noted subtropical influences; node 4 contains similar trajectories. Trajectories from other nodes are mostly zonal.

Meteorological variables from ERA-Interim reanalysis were analyzed to further explore the origins of nodes 4–6, 8, and 9. Each node is evaluated individually using meteorological anomalies calculated with the same methodology used for Figure 6. We note that the beginning of the Wallops Island record is not covered by the ERA-Interim data set (1979 to present), but this fact is not expected to influence the following results as a large sample size remains. While some of the variables examined could simply be extracted from the sonde data, it is more useful to examine 3-D meteorological fields to aid geophysical interpretation of the remaining nodes.

Node 4 exhibits negative PV anomalies at 250 hPa (Figure 8), indicating reduced stratospheric influence at upper levels. The anomalies, on the order of −0.3 to −1.0 potential vorticity units (PVU), are observed at all sites. Positive anomalies of MSLP and geopotential height at 850 and 700 hPa at all sites (see supporting information Figures S1 and S2), and the southwesterly trajectory tendency at Huntsville and Wallops Island (Figure 7), support the hypothesis that node 4 profiles are indeed influenced by subtropical air, leading to slightly below average tropospheric O_3 amounts. Quantitative values of the meteorological anomalies (Figures 6a and 6b), however, make it apparent that the degree to which subtropical air affects node 4 profiles is much less than for node 7.

Node 5 profiles exhibit slightly above average total and tropospheric column O_3 and slightly below average tropopause and 500 hPa geopotential heights (Figure 6). ERA-Interim 250 hPa geopotential height anomalies (Figure 9) exhibit an upper level trough influence on these profiles with negative anomalies (except Huntsville) of −35 to −90 m. The disparity between Huntsville and the other sites is evident in the 250 hPa height anomalies, consistent with the node 5 differences in SOM profile shape in Figure 4 at 10–12 km. The lower 250 hPa heights reflect a correspondingly lower tropopause height which increases O_3 at this level, an effect not as evident in node 5 Huntsville profiles as at other sites. The remainder of the ERA-Interim pressure level and surface data are void of anomalies in node 5; only altitudes near the tropopause show appreciable meteorological influence on the O_3 profiles.

Node 6 profiles are hypothesized to be influenced by STE, given the low 500 hPa and tropopause heights (Figures 6a and 6b), well above average tropospheric column O_3 (Figure 6c), and cyclonic-curving back trajectories (Figure 7). Positive maxima of 500 hPa PV anomalies (Figure 10; +0.1 to +0.4 PVU) centered near each site indicate stratospheric influence in the midlevels of node 6 profiles. Though node 6 members represent an assortment of profiles, many contain layers of high O_3 mixing ratios from stratospheric origins in the midtroposphere.

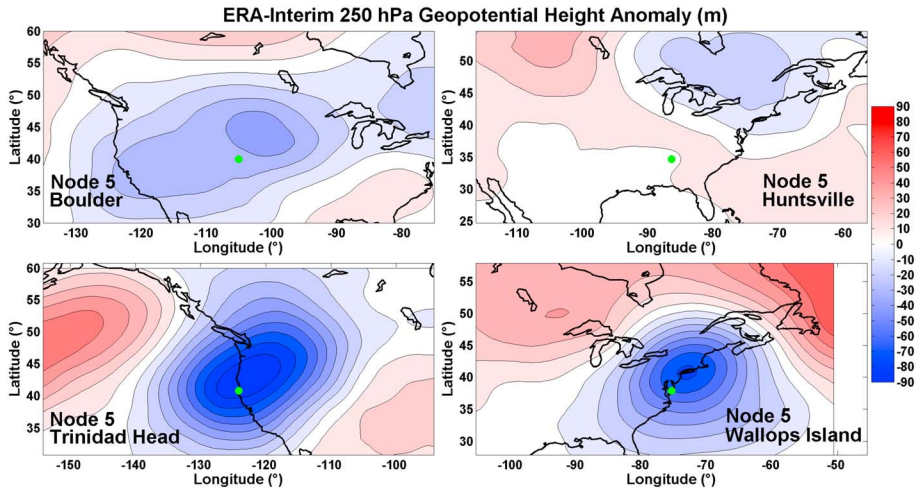

Figure 9. Contoured map of average ERA-Interim 250 hPa geopotential height (m) anomalies from monthly climatology for node 5 at each site. Data are contoured every 10 m from −90 to 90 m. Blue colors represent negative anomalies, and red colors represent positive anomalies. The green dot represents the site location.

Node 8 profiles contain positive tropopause and 500 hPa height anomalies similar to node 7 profiles (Figure 6) but contain near-average tropospheric column O_3 amounts. The 500 hPa geopotential height anomalies derived from ERA-Interim are shown in Figure 11. The 500 hPa geopotential height anomalies derived from ERA-Interim (see Figure 11) differ only slightly from those derived from sonde data. A clear trough-ridge structure is visible in the average 500 hPa geopotential height anomalies of node 8 throughout all CONUS sites. Node 8 profiles have positive geopotential height and temperature anomalies through all four (850, 700, 500, and 250 hPa) extracted pressure levels. This is also true for node 7 profiles. In terms of meteorological anomalies, node 8 is nearly identical to node 7. Thus, the dichotomy in seasonality (Figure 5) of nodes 7 and 8 is likely the major driver behind the O_3 profile differences in these two clusters.

Node 9 profiles have fewer defining meteorological characteristics. The most distinct features are observed in the ERA-Interim temperature anomalies, especially at 2 m (Figure 12). All sites but Trinidad Head are anomalously warm (+0.7 to +1.5°C) near the surface in the largely summertime node 9 profiles. Trinidad Head, which is rarely polluted near the surface and contains relatively few profiles in node 9, averages 1°C cooler than

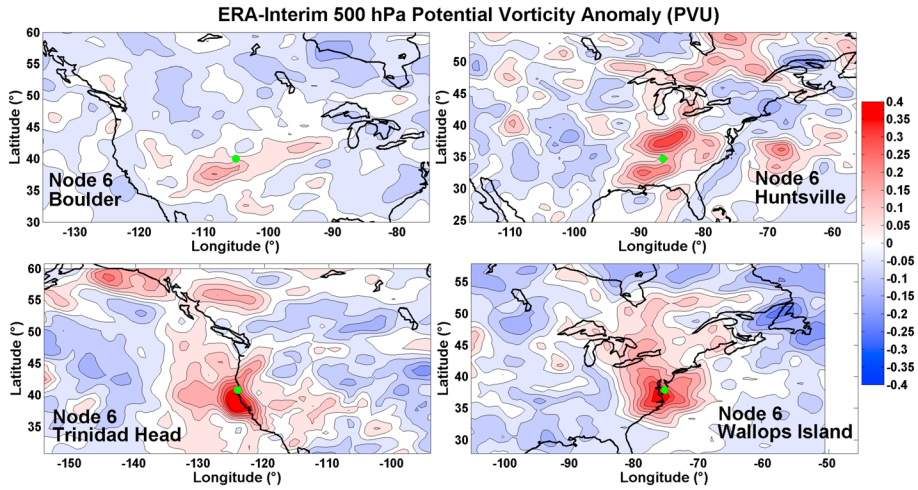

Figure 10. Contoured map of average ERA-Interim 500 hPa PV (PVU) anomalies from monthly climatology for node 6 at each site. Data are contoured every 0.05 PVU from −0.4 to 0.4 PVU. Blue colors represent negative anomalies, and red colors represent positive anomalies. The green dot represents the site location.

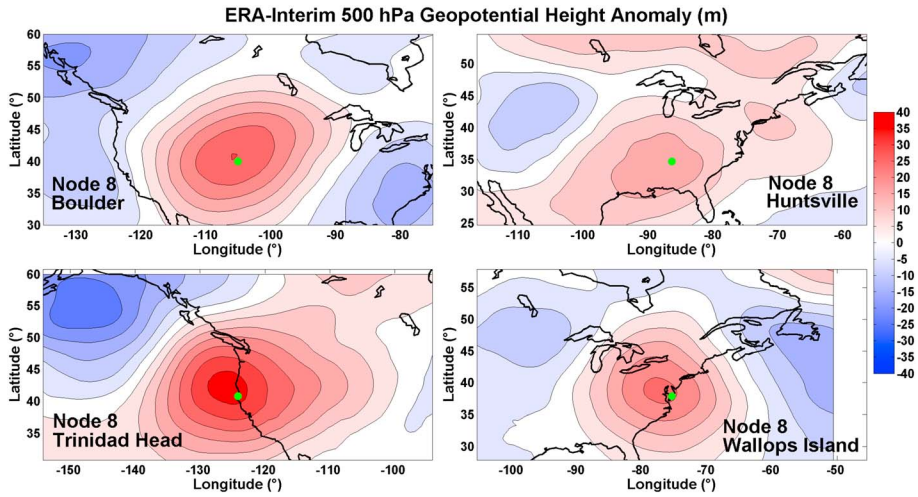

Figure 11. Contoured map of average ERA-Interim 500 hPa geopotential height (m) anomalies from monthly climatology for node 8 at each site. Data are contoured every 5 m from −40 to 40 m. Blue colors represent negative anomalies, and red colors represent positive anomalies. The green dot represents the site location.

climatology at 2 m. This temperature behavior holds true through most of the troposphere, with Boulder, Huntsville, and Wallops Island being warmer than normal at 2 m, 850, 700, and 500 hPa, and Trinidad Head being cooler than normal at these levels. Additionally, Trinidad Head exhibits positive PV anomalies at 500 hPa in node 9 profiles. Processes leading to enhanced tropospheric O_3 amounts at Trinidad Head, such as STE and transport of pollution across the Pacific Ocean, are different from the other three sites. Other than anomalous temperatures at all sites and PV in the midlevels at Trinidad Head, there is a lack of significant dynamic meteorological forcing in node 9. Therefore, node 9 is hypothesized to result from transported pollution during the high-sun-angle summer months, facilitating photochemical production in the troposphere and the resulting O_3 profiles.

3.5. Ozone Profile Anomalies From Monthly Climatology

SOM clusters show characteristics based on meteorological measurements and reanalysis, seasonality, and O_3 column amounts. The next step is to evaluate how closely the monthly O_3 climatology describes the vertical O_3 profiles in the SOM node clusters. The average difference between profiles from each node and their respective monthly O_3 mixing ratio climatology is calculated. Results for all nodes at each site are presented in terms of O_3 mixing ratio anomalies in ppbv (Figure 13). Not surprisingly, given their low tropopause

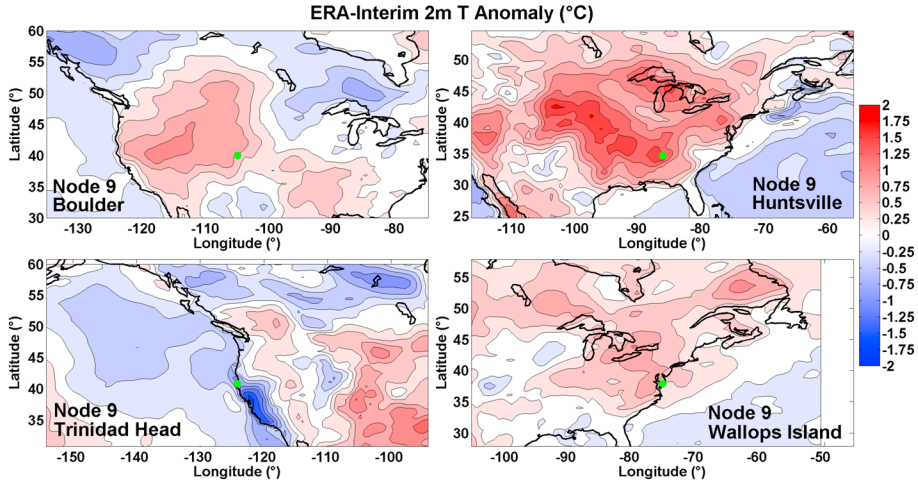

Figure 12. Contoured map of average ERA-Interim 2 m temperature (°C) anomalies from monthly climatology for node 9 at each site. Data are contoured every 0.25°C from −2 to 2°C. Blue colors represent negative anomalies, and red colors represent positive anomalies. The green dot represents the site location.

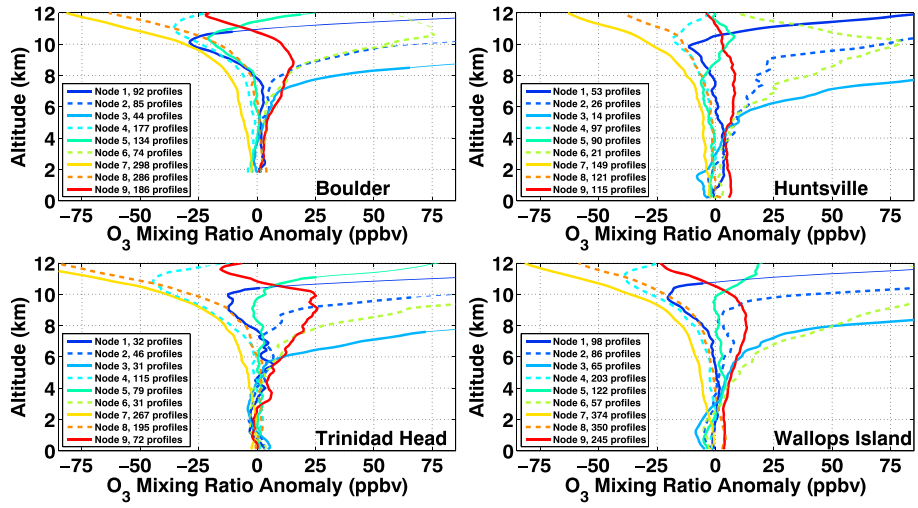

Figure 13. Average O_3 mixing ratio anomaly (ppbv) from monthly climatology with altitude for each SOM node 1–9. For this figure, each O_3 mixing ratio profile was compared with its corresponding monthly averaged O_3 profile. A mean anomaly was then calculated for each node and is shown. For reference, values above the average tropopause for each node are shown as thin lines. The number of profiles in each node is given in the legend.

heights, nodes 1–3 exceed monthly climatological O_3 by over 100 ppbv and in many cases more than double the climatological O_3 from 8 to 12 km. Extreme O_3 increases above climatology of over +75 ppbv also appear at all sites in node 6. Ozone anomalies in node 6 retreat above 10 km at Boulder and Huntsville, accounting for the complex O_3 profile shapes and stratospheric intrusion layers in that cluster. Nodes 4 and 5 differ by only a few ppbv from climatology, typically within ±3 ppbv in the low to middle troposphere. The largest deviations from climatological O_3 in nodes 4 and 5 occur above 10 km, coincident with the PV and geopotential height anomalies evident in Figures 8 and 9. Even discounting variations in the tropopause region, climatological O_3 averages may fail to describe a large percentage of the O_3 profiles at the CONUS sites. At 6 km at all sites, node 7 profiles on average fall more than 6 ppbv below monthly climatology, in conjunction with tropospheric column O_3 amounts that are 10–20% below normal (Figure 6c). Conversely, node 9 profiles lie well above climatology, exceeding it by over +7 ppbv at all sites, and up to +10 ppbv at Wallops Island and Trinidad Head. Nodes 7 and 9 correspond to 36–39% of all profiles at each of the ozonesonde sites. Clearly, monthly O_3 climatologies do not adequately describe the variability of CONUS O_3 observations, either in terms of profile shape or column O_3 amount. Use of nine SOM clusters has more accurately captured the distribution of these O_3 profile data sets than monthly averages, particularly near the highly variable tropopause altitude.

Based on findings from Figures 6a, 6c, and 6d, one expects a relationship between the tropopause height and tropospheric column O_3. Figure 14 shows how that relationship may change given the O_3 anomalies observed in Figures 6c and 13. Figure 14 presents correlation coefficients and least squares fits of tropopause height and tropospheric column O_3. Except for node 7 profiles, each SOM node results in a similar correlation or contains few enough profiles so as to not affect the overall data set correlation. In fact, excluding node 7 profiles greatly increases the total data set correlation coefficient between tropopause height and tropospheric column O_3 at all sites (Boulder, $r = 0.59$ to 0.73; Huntsville, $r = 0.57$ to 0.69; Trinidad Head, $r = 0.44$ to 0.65; Wallops Island, $r = 0.53$ to 0.71). The low tropospheric O_3 amount (lowest average of all nodes) and high tropopause heights found in node 7 represent a separate regime (red line in Figure 14) signaled by a displacement in the relationship between the two variables compared to the rest of the data set. SOM node 7 profiles contain the lowest O_3 DU km^{-1} value of all nodes in the troposphere. The typical tropopause height/tropospheric column O_3 relationship observed in 70–80% of CONUS profiles is adjusted when fall and winter profiles contain positive tropopause and 500 hPa height anomalies.

A summarizing table (Table 2) is provided to outline the meteorological and O_3 characteristics of the SOM nodes at each site.

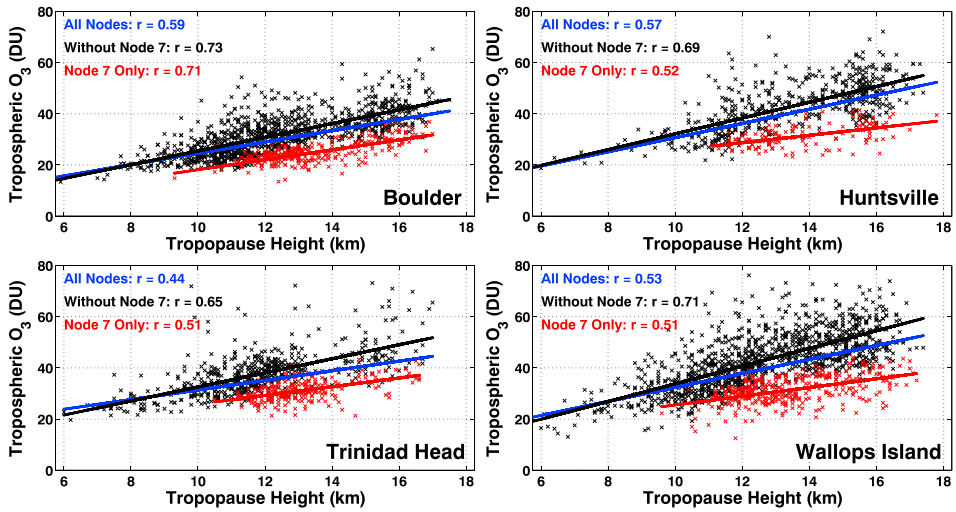

Figure 14. Scatterplots of tropospheric column O_3 (DU) and WMO lapse rate tropopause height (km). Correlation coefficients and least squares best fit lines are shown for three cases: (1) all profiles/SOM nodes (blue), (2) excluding node 7 (black), and (3) only node 7 (red). Individual launches are shown as black crosses with node 7 launches in red.

4. Conclusions

The application of SOM to 4530 CONUS O_3 profiles led to clustering that is primarily based on two main factors: (1) the altitude of the tropopause O_3 gradient and (2) the amount of O_3 in the low to middle troposphere. Though profiles in some SOM clusters were mostly confined to a few months in the year, several exhibited no distinct seasonality, indicating more than just temporal effects on O_3 profile variability over CONUS. The top row of nodes 1–3 (13–16% of CONUS profiles) at each site represented an incremental lowering of the tropopause O_3 gradient. The nodes were associated with synoptic-scale troughs and contained double the climatological O_3 amount from 8 to 12 km. Thus, capturing the variability in tropopause height is vital for reproducing the significant day-to-day changes in O_3 profile shape at these CONUS sites.

Challenges in characterizing CONUS O_3 variability go beyond knowledge of the tropopause height. Nodes 7 and 9 displayed the largest deviations from climatological O_3 in the low to middle troposphere. Ozone in nodes 7 and 9 was generally beyond ±6 ppbv from 6 km climatological O_3 but up to +10 ppbv (+25% tropospheric column O_3) in node 9 at Trinidad Head and Wallops Island. Inclusion of node 7 profiles, which represent a different regime in the tropopause height/tropospheric column O_3 relationship, greatly reduced correlation coefficients between tropopause height and tropospheric column O_3 for the entire data set. Nodes 7 and 9 contained nearly 40% of CONUS O_3 profiles. Understanding the large-scale conditions outlined here that lead to clean tropospheric O_3 profiles of subtropical origins (node 7) and summertime tropospheric pollution events (node 9) is key to better design of chemical models and satellite algorithms.

Although SOM nodes are explained by large-scale meteorological conditions, we will explore SOM connections to chemical processes for locations and periods with more chemical measurements.

Table 2. Summaries of SOM Node Seasonal, Meteorological, and O_3 Characteristics[a]

Node	Percentage of O_3 Profiles	Seasonality	Tropopause Height Anomaly (m)	Tropospheric Column O_3 Anomaly (DU)	6 km O_3 Anomaly (ppbv)
1	4–8%	Winter-Spring	−1170 (−1500, −710)	−0.6 (−3.3, 1.1)	1.1 (−0.8, 2.2)
2	4–6%	Winter-Spring	−1900 (−2140, −1740)	0.5 (−2.1, 5.1)	7.3 (3.8, 13.2)
3	2–4%	Winter-Spring	−3290 (−4150, −2720)	−3.5 (−7.7, −0.1)	11.9 (6.0, 19.5)
4	13–14%	Most/All Months	30 (−290, 310)	−1.5 (−3.3, −0.3)	−2.2 (−3.0, −0.9)
5	8–13%	All Months	−500 (−960, 160)	1.8 (0.8, 2.7)	−0.2 (−3.8, 3.2)
6	3–5%	Varies by Site	−1380 (−1670, −1020)	6.4 (3.1, 9.8)	10.5 (3.6, 16.2)
7	22–31%	Fall-Winter	770 (600, 880)	−6.2 (−8.5, −3.7)	−7.2 (−8.3, −6.6)
8	18–22%	Spring-Fall	800 (670, 1030)	1.5 (0.5, 1.9)	−0.5 (−2.2, 0.5)
9	8–17%	Summer	240 (−150, 910)	7.8 (5.2, 9.5)	8.8 (7.3, 11.0)

[a]Because all sites displayed similar characteristics organized by SOM node, all sites are summarized together. Averages and the range of values from the four sites are presented for tropopause height (m), tropospheric column O_3 (DU), and 6 km O_3 mixing ratio anomaly (ppbv).

The finding that simple time means are inadequate for describing the complexity of individual O_3 profiles is particularly true near the tropopause at the CONUS sites addressed here. The diversity of CONUS O_3 profile shapes is more appropriately expressed using nine SOM clusters than by using 12 monthly averages. In the midlatitudes, O_3 profile evolution is highly dependent upon trough and ridge systems associated with large-scale Rossby waves, pollution transport, and the influence of subtropical air, all of which cause appreciable changes over short time scales. SOM graphically depicts the contributions to this variability in the tropospheric O_3 profile. SOM provides exceptional insights into the seasonality, or lack thereof, of observed profile shapes and the frequency of extreme, dynamic-induced changes at the tropopause level observed in SOM nodes 1–3 in this paper. SOM also distinguishes O_3 profiles with very low and high (nodes 7 and 9) tropospheric O_3 amounts, useful for quantifying typical baseline and polluted O_3 levels.

Appendix A

This appendix provides explanations of the *k*-means clustering algorithm and details of the SOM clustering user-selectable settings. Sensitivity tests on the user-selectable SOM settings are then compared to the output with *k*-means, using both randomly generated data and real ozonesonde O_3 mixing ratio profiles. Results from these tests are used to justify our choice of clustering algorithm ultimately used in this study, the 3×3 SOM map with nine nodes/clusters.

A1. *k*-Means

The *k*-means algorithm partitions an input data set into a user-defined number (*k*) of clusters. Cluster centroids are initialized through random selection of *k* vectors from the input data set. Each remaining input vector is then assigned to the closest (in Euclidean distance) centroid. Finally, the centroid of each cluster is updated:

$$\boldsymbol{m}_i(t+1) = \frac{1}{n_i}\sum\nolimits_{j \in n_i} x_j \tag{A1}$$

where $\boldsymbol{m}$ is the *i*th cluster centroid, *x* is the *j*th data vector belonging to the *i*th cluster, *n* is the number of data vectors belonging to the *i*th cluster, and *t* is the iteration. The new centroid is thus the average of the previous centroid's assigned input vectors. All of the data vectors are then reassigned to clusters based on these new centroids. This process is repeated until there are no new vector assignments—the algorithm has converged, with the input data separated into *k* exclusive clusters. Defining the centroids as the mean of the corresponding data vectors guarantees that the average Euclidean distance between a cluster's centroid and its member vectors is minimized.

A2. The Batch SOM Algorithm

In the batch SOM algorithm, each vector in the data set is grouped with its closest (in Euclidean distance) initial SOM node, called the best matching unit (BMU). The batch SOM equation is then applied to the data set and can be repeated thousands of times (each repeat is called an epoch), if necessary, to converge to a final map. The batch equation is as follows:

$$\boldsymbol{m}_i(t+1) = \sum\nolimits_{j=1}^{M} n_j h_{ij}(t)\overline{\boldsymbol{x}}_j / \sum\nolimits_{j-1}^{M} n_j h_{ij} \tag{A2}$$

where $\boldsymbol{m}$ is the *i*th of *M* total nodes, *n* is the number of vectors for which node *j* is the BMU, $\overline{\boldsymbol{x}}$ is the mean of the vectors for which node *j* is the BMU, *h* is the value of the neighborhood function (dependent on factors discussed below), and *t* is the epoch. This equation is repeatedly calculated for the user-defined number of epochs. Essentially, each $\boldsymbol{m}_i$ node is updated with the mean of its member vectors $\overline{\boldsymbol{x}}_j$ when $i = j$, and $\overline{\boldsymbol{x}}_j$ multiplied by the neighborhood function, a value from 0 to 1, when $i \neq j$.

A3. The SOM Neighborhood Function

The neighborhood function distinguishes the batch trained SOM from the *k*-means algorithm. This function allows updating node $\boldsymbol{m}_i$ to learn from nearby nodes' member vectors in addition to its own. Typically in SOM, the neighborhood function value decreases gradually with increasing distance between nodes *i* and *j*. This causes node $\boldsymbol{m}_i$ to learn less from neighboring nodes' member vectors than its own but more than the case with *k*-means where such intercluster learning does not occur at all. The addition of intercluster learning leads to a topographical ordering of SOM nodes that is absent in *k*-means clusters. However, in equation (A2), if the neighborhood function is 1 only when $i = j$, and 0 otherwise, the SOM learning becomes identical to *k*-means.

In that limiting case, each node $\boldsymbol{m}_i$ is updated only with its own member vectors, there is no neighborhood node learning, and the nodes are independent of each other.

The neighborhood function depends on the Euclidean distance between the updating node $\boldsymbol{m}_i$, the current node iteration j, and the user-defined neighborhood radius. The distance between nodes is 0 when $i=j$, ranging to $\sqrt{32}$ in a 4×4 map, for example. The neighborhood radius is reduced linearly with epoch so the neighborhood function value decays, diminishing the neighborhood learning and allowing the map solution to converge. We explore the effects of four neighborhood functions, each available as an option in the Matlab SOM Toolbox [*Vesanto et al.*, 2000; *Liu et al.*, 2006], as follows:

$$h_{ij}(t)=\begin{cases}\exp\left(-\dfrac{d_{ij}^2}{2r_t^2}\right) & \text{Gaussian}\\ \exp\left(-\dfrac{d_{ij}^2}{2r_t^2}\right)F\left(r_t^2-d_{ij}^2\right) & \text{Cut Gauss}\\ F\left(r_t^2-d_{ij}^2\right) & \text{Bubble}\\ \left[\dfrac{1-d_{ij}^2}{r_t^2}\right]F\left(r_t^2-d_{ij}^2\right) & \text{Epanechnikov (Ep)}\end{cases} \tag{A3}$$

Here r is the neighborhood radius at epoch t, d is the distance between nodes i and j, and $F(x)$ is a step function with a value of 1 if $x\geq0$, and 0 otherwise. The basic geometry of these functions is found in *Vesanto et al.* [2000]. The Gaussian function decays to a nonzero value as distance between nodes increases. The Cut Gauss, Bubble, and Ep functions are zero once the node distance d is greater than the neighborhood radius r.

A4. SOM Neighborhood Functions Test

The user's choice of map size/cluster number and SOM neighborhood function can have significant impacts on the amount of data assigned to each cluster and the quality of fit between a node/centroid and its member data. A sensitivity analysis is conducted to determine user-chosen parameter settings that produce the most effective O_3 profile clusters from SOM or k-means. Neighborhood functions will be evaluated first, followed by map size/cluster number.

To compare the performance of k-means and the four SOM neighborhood functions, each algorithm was applied to an identical, random, 1000 point, 2-D data set using a 3×3 SOM map and, equivalently, $k=9$ clusters (Figure A1). For this test, the SOM neighborhood radius decreased linearly from 3 to 1 over 100 epochs. The k-means algorithm was repeated to convergence and was initialized identically to the SOMs (via PCA) to avoid the stochastic outcomes that typically result from random initialization (a random k-means initialization will be considered in other tests). Figure A1 represents how neighborhood functions organize and cluster a randomly generated set of 2-D data. Given the use of neighborhood learning, the nodes in SOM depend upon others' member vectors, in contrast with the independent and arbitrarily organized clusters in k-means. Consistent ordering of SOM nodes, regardless of neighborhood function, is displayed by the SOM node number labels in Figure A1.

Clustering often seeks to maximize the distance between centroids/nodes to better distinguish signals in the data set. The Ep function SOM nodes (Figure A1; top left, large color dots) converge most similarly to the k-means algorithm (black diamonds) because of the Ep neighborhood functions' sharp decrease with increasing distance between nodes; each node is less dependent on nonmember vectors than other functions. This yields the greatest distances among nodes among the neighborhood functions. Gaussian, the slowest decaying function with node distance, yields nodes that cluster close together near the overall data set mean, greatly contrasting the independent and farthest separated k-means centroids. The varied clustering resulting from use of different neighborhood functions also causes disparity in performance as measured by error metrics, as explained in the next section.

A5. Error Metrics

There are two standard measures of SOM error: quantization error (QE) and topographical error (TE). Figure A2 shows QE, the average Euclidean distance between a node/centroid and its respective member

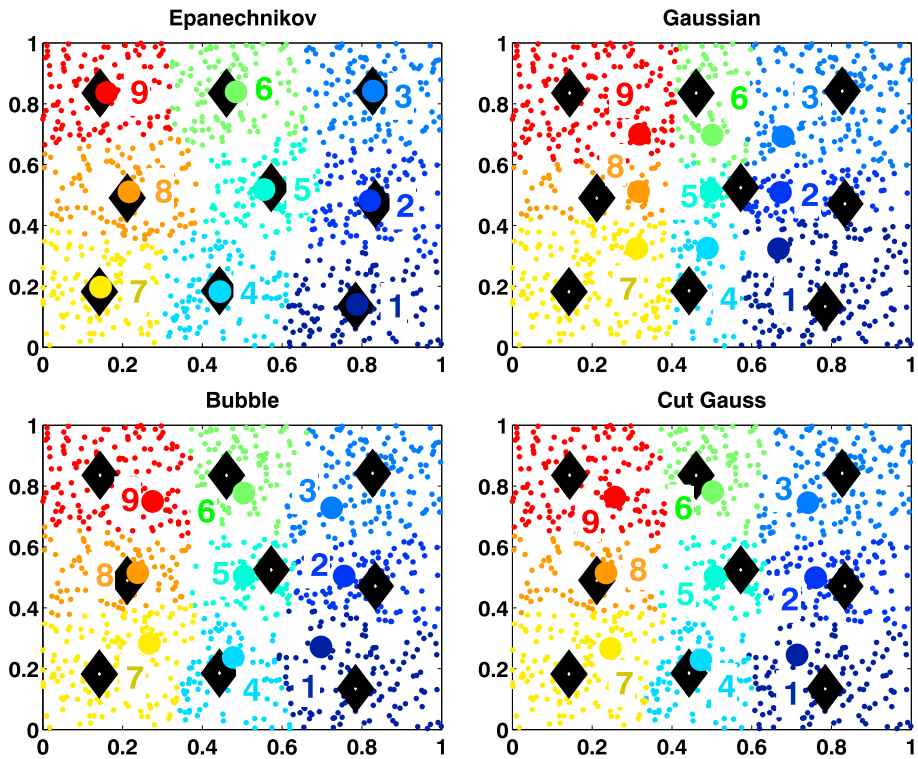

Figure A1. Example of 3 × 3 SOM nodes (numbered large colored dots) and $k = 9$ clusters (black diamonds, PCA initialized) of randomly generated data (small colored dots) for four neighborhood functions. Data are colored and labeled according to their respective SOM node in each plot. *k*-means runs were unchanged and run to convergence. SOM was run for 100 epochs with neighborhood radii decreasing linearly from 3 to 1 over the 100 epochs.

vectors, for both 3 × 3 SOM and *k*-means (both random and PCA-initialized are included). SOM and *k*-means are performed on the combined CONUS sites' O_3 mixing ratio profile data sets. The altitude range covers surface to 12 km amsl as throughout the paper.

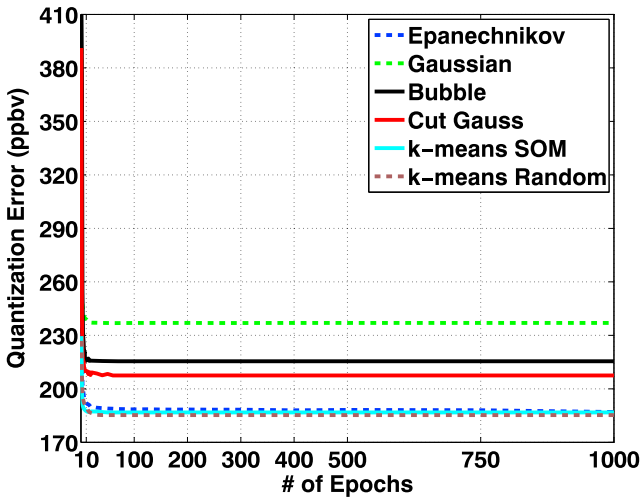

Figure A2. The four-site mean quantization error, defined as the average Euclidean distance between a node/centroid and its member profiles. Data are shown for 3 × 3 SOM/$k = 9$ clusters, with increasing epochs. Four SOM neighborhood functions are tested with SOM-initialized *k*-means and randomly initialized *k*-means. Note that *k*-means converges before 100 iterations and is constant thereafter. Lower error indicates a better fit between a node/centroid and its member data.

The four-site average QE value (small values are desired) is shown as a function of epoch in Figure A2. To remain consistent with SOM settings used in the body of the paper, QE values up to 1000 epochs are presented (note that *k*-means converges to its solution long before 1000 iterations). Given the distribution of nodes for each neighborhood function in Figure A1, the results are not surprising. The Ep function mimics both *k*-means runs, which by definition minimize the QE metric for clusters. Because the Gaussian function clusters tend toward the overall mean, the fits between its nodes and member data are the worst of the six options shown in Figure A2. The Bubble and Cut Gauss functions fall between these two extremes.

TE provides a measure of how well the SOM map fits the data manifold. TE is the fraction of input data vectors whose

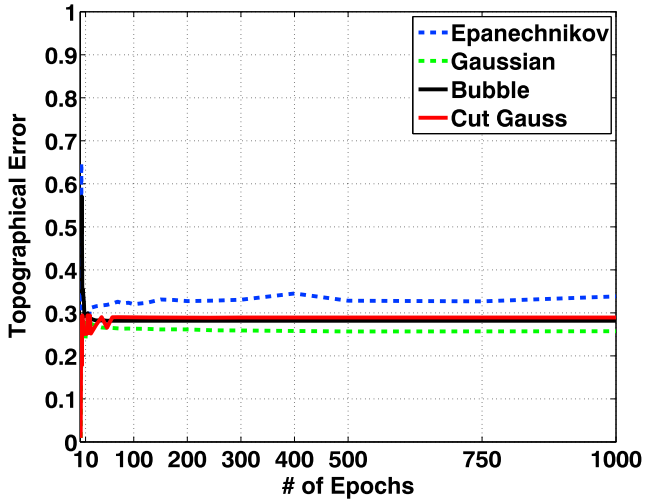

Figure A3. The four-site mean topographical error, defined as the fraction of profiles whose second closest node is *not* adjacent to its BMU in the map, for 3 × 3 SOM, $k = 9$ clusters, with number of epochs for four SOM neighborhood functions. Data are from the same output as in Figure A2. Lower percentages typically indicate better-ordered neighboring nodes.

BMU is *not* adjacent to its second closest node (Figure A3; same SOM settings as Figure A2), with smaller percentages generally indicating better organization. This metric quantifies a major advantage of SOM over *k*-means. The organization of clusters by SOM provides superior data visualization (Figures 3 and 4). Because *k*-means clusters are randomly ordered and unorganized, TE is not relevant to them. In Figure A3, the Gaussian neighborhood function yields the lowest TE, indicating that adjacent nodes are most alike with that function. TE is highest for the Ep function, which may be a result of its more independent, and thus more unique, nodes. Still, the TE metric varies by only a few percent between neighborhood functions, and the errors show well-ordered maps in any case. Given the small variation in TE between neighborhood functions, and the improved QE and uniqueness performance of Ep compared to other neighborhood functions, we choose the Ep function to further compare SOM against *k*-means.

A6. Cluster Number/SOM Map Size

The final sensitivity test evaluates randomly and PCA-initialized *k*-means and SOM using the Ep neighborhood function. A balance is sought between the number of profiles assigned to each cluster and the total number of clusters. The choice of number of clusters must be enough to capture the variability in the data set, and each cluster must contain enough cases to sufficiently describe geophysical meaning for each cluster. The percentage of profiles in both the most and least populous clusters for each case is chosen to provide a measure of this balance. For each ozonesonde location, varying numbers of clusters and SOM nodes are analyzed to evaluate cluster membership (Table A1). The SOMs were run to 1000 epochs with the same settings as prior SOM tests. Small *k* and SOM map sizes result in highly populated clusters. Single SOM and *k*-means clusters often contain over half the data in the 2 × 2 SOM and $k = 4$ *k*-means solutions. At all CONUS sites, the two most populated clusters in the 2 × 2 SOM/$k = 4$ solutions contain ~80% of all profiles, making characterization and interpretation of those

Table A1. Percentages of Total Profiles Corresponding to Most and Least Populous Clusters for Given SOM Map Sizes (*k* Clusters)[a]

		SOM		*k*-Means (Random)		*k*-Means (PCA)	
Site	Map Size (*k*)	Max%	Min%	Max%	Min%	Max%	Min%
Boulder	2 × 2 (4)	44.9	6.3	62.7	3.9	62.7	3.9
	3 × 3 (9)	21.7	3.2	25.5	2.3	24.6	2.3
	4 × 4 (16)	12.7	1.5	13.8	0.1	14.5	0.9
Huntsville	2 × 2 (4)	46.4	3.6	46.6	3.5	46.6	3.5
	3 × 3 (9)	21.7	2.0	24.9	1.0	24.5	0.7
	4 × 4 (16)	15.2	1.0	15.7	0.3	14.6	0 (1)
Trinidad Head	2 × 2 (4)	52.1	5.8	62.3	4.4	62.3	4.4
	3 × 3 (9)	30.8	3.6	31.3	3.1	32.7	3.2
	4 × 4 (16)	18.0	2.1	11.6	1.6	17.4	0 (1)
Wallops Island	2 × 2 (4)	43.5	5.8	46.6	4.7	46.6	4.7
	3 × 3 (9)	23.4	3.6	25.8	2.0	26.9	1.9
	4 × 4 (16)	12.4	1.3	14.1	1.4	15.9	1.1

[a]Tests were run using SOM with the Epanechnikov neighborhood function, randomly initialized *k*-means, and *k*-means initialized identically to SOM (via PCA) for 1000 epochs. Note that *k*-means converges long before 1000 iterations. Cells marked "0 (1)" indicate that the cluster contained only one profile.

clusters difficult. This statement is supported by closer inspection of differences between profile membership in the 2 × 2 and 3 × 3 SOM options (Table S1). The CONUS O_3 profile 2 × 2 SOM nodes are combinations of specific 3 × 3 SOM nodes, masking the unique meteorological conditions characteristic of many (particularly nodes 4–6, and 8) 3 × 3 SOM nodes found in the body of this paper.

As the map size and cluster number increases, membership of the least populous SOM and *k*-means clusters drops precipitously. The *k*-means centroids appear to be affected by outlier profiles, yielding several one-member clusters when $k = 16$. Presumably, this results from a lack of neighborhood learning. Even when $k = 9$, the least populous *k*-means cluster contains 1% or less of O_3 profiles in several cases. Considering the excessively large cluster membership in the 2 × 2 SOM/$k = 4$ *k*-means, and the lack of profiles associated with nodes and centroids in the 4 × 4 SOM/$k = 16$ or 9 *k*-means, the 3 × 3 SOM with the Ep neighborhood function appears to be optimum for examining O_3 profile clustering at the CONUS sites.

Acknowledgments
Funding for this project was provided by the following NASA grants: NNG05G062G, NNX10AR39G, NNX11AQ44G, and NNX12AF05G. The continued operation of CONUS ozonesonde stations are the combined efforts of many institutions and individuals: Boulder, CO, and Trinidad Head, CA: Samuel Oltmans and Bryan Johnson (NOAA ESRL GMD); Huntsville, AL: Michael Newchurch (University of Alabama in Huntsville); Wallops Island, VA: Frank Schmidlin and E. Thomas Northam (NASA/Wallops Flight Facility). Thanks to the World Ozone and Ultraviolet Radiation Data Centre (WOUDC) for continued availability of ozonesonde data sets. Thanks also to Bryan Johnson for providing high-resolution profile data from 1979 to 1989 for the Boulder, CO, station. Thanks to Anders Jensen (Penn State University) for initial assistance with SOM. WOUDC data were accessed at ftp://ftp.tor.ec.gc.ca/pub/woudc/. NOAA ESRL GMD data were accessed at ftp://ftp.cmdl.noaa.gov/data/ozwv/Ozonesonde/. ERA-Interim reanalysis data were accessed at http://rda.ucar.edu/datasets/ds627.0/. NCEP/NCAR reanalysis data were accessed at ftp://ftp.cdc.noaa.gov/. This paper is the basis for a chapter in the first author's PhD thesis. The authors also thank the Editor and three anonymous reviewers for suggestions that improved this manuscript.

References

Considine, D. B., J. A. Logan, and M. A. Olsen (2008), Evaluation of near-tropopause ozone distributions in the Global Modeling Initiative combined stratosphere/troposphere model with ozonesonde data, *Atmos. Chem. Phys.*, *8*, 2365–2385, doi:10.5194/acp-8-2365-2008.

Cooper, O. R., et al. (2011), Measurement of western U.S. baseline ozone from the surface to the tropopause and assessment of downwind impact regions, *J. Geophys. Res.*, *116*, D00V03, doi:10.1029/2011JD016095.

Danielsen, E. F. (1968), Stratospheric-tropospheric exchange based on radioactivity, ozone and potential vorticity, *J. Atmos. Sci.*, *25*, 502–518, doi:10.1175/1520-0469(1968)025<0502:STEBOR>2.0.CO;2.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*, 553–597, doi:10.1002/qj.828.

Diab, R. D., A. M. Thompson, K. Mari, L. Ramsay, and G. J. R. Coetzee (2004), Tropospheric ozone climatology over Irene, South Africa, from 1990 to 1994 and 1998 to 2002, *J. Geophys. Res.*, *109*, D20301, doi:10.1029/2004JD004793.

Draxler, R. R., and G. D. Hess (1997), Description of the HYSPLIT_4 modeling system, NOAA Tech. Memo, ERL ARL-224, NOAA Air Resour. Lab., Silver Spring, Md., 24 pp.

Friedman, M., and M. D. Brazeau (2011), Sequences, stratigraphy and scenarios: What can we say about the fossil record of the earliest tetrapods?, *Proc. R. Soc. Edinburgh*, *278*, 432–439, doi:10.1098/rspb.2010.1321.

Hewitson, B. C., and R. G. Crane (2002), Self-organizing maps: Applications to synoptic climatology, *Clim. Res.*, *22*, 13–26, doi:10.3354/cr022013.

Holton, J. R., P. H. Haynes, M. E. McIntyre, A. R. Douglass, R. B. Rood, and L. Pfister (1995), Stratosphere-troposphere exchange, *Rev. Geophys.*, *33*(4), 403–439, doi:10.1029/95RG02097.

Hong, Y., K. Hsu, S. Sorooshian, and X. Gao (2004), Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system, *J. Appl. Meteorol.*, *43*, 1834–1853, doi:10.1175/JAM2173.

Jensen, A. A., A. M. Thompson, and F. J. Schmidlin (2012), Classification of Ascension Island and Natal ozonesondes using self-organizing maps, *J. Geophys. Res.*, *117*, D04302, doi:10.1029/2011JD016573.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kohonen, T. (1995), The basic SOM, in *Self-Organizing Maps*, pp. 77–130, Springer, New York.

Komhyr, W. D. (1969), Electrochemical concentration cells for gas analysis, *Ann. Geophys.*, *25*, 203–210.

Komhyr, W. D., R. A. Barnes, G. B. Brothers, J. A. Lathrop, and D. P. Opperman (1995), Electrochemical concentration cell ozonesonde performance evaluation during STOIC 1989, *J. Geophys. Res.*, *100*(D5), 9231–9244, doi:10.1029/94JD02175.

Lamarque, J.-F., et al. (2012), CAM-chem: Description and evaluation of interactive atmospheric chemistry in the Community Earth System Model, *Geosci. Model Dev.*, *5*, 369–411, doi:10.5194/gmd-5-369-2012.

Lin, M., A. M. Fiore, O. R. Cooper, L. W. Horowitz, A. O. Langford, H. Levy II, B. J. Johnson, V. Naik, S. J. Oltmans, and C. J. Senff (2012), Springtime high surface ozone events over the western United States: Quantifying the role of stratospheric intrusions, *J. Geophys. Res.*, *117*, D00V22, doi:10.1029/2012JD018151.

Liu, Y., R. H. Weisberg, and C. N. K. Mooers (2006), Performance evaluation of the self-organizing map for feature extraction, *J. Geophys. Res.*, *111*, C05018, doi:10.1029/2005JC003117.

Logan, J. A. (1985), Tropospheric ozone: Seasonal behavior, trends, and anthropogenic influence, *J. Geophys. Res.*, *90*(D6), 10,463–10,482, doi:10.1029/JD090iD06p10463.

Logan, J. A. (1994), Trends in the vertical distribution of ozone: An analysis of ozonesonde data, *J. Geophys. Res.*, *99*(D12), 25,553–25,585, doi:10.1029/94JD02333.

Logan, J. A., et al. (1999), Trends in the vertical distribution of ozone: A comparison of two analyses of ozonesonde data, *J. Geophys. Res.*, *104*(D21), 26,373–26,399, doi:10.1029/1999JD900300.

Logan, J. A., et al. (2012), Changes in ozone over Europe: Analysis of ozone measurements from sondes, regular aircraft (MOZAIC) and alpine surface sites, *J. Geophys. Res.*, *117*, D09301, doi:10.1029/2011JD016952.

McPeters, R. D., and G. J. Labow (2012), Climatology 2011: An MLS and sonde derived ozone climatology for satellite retrieval algorithms, *J. Geophys. Res.*, *117*, D10303, doi:10.1029/2011JD017006.

McPeters, R. D., G. J. Labow, and B. J. Johnson (1997), A satellite-derived ozone climatology for balloonsonde estimation of total column ozone, *J. Geophys. Res.*, *102*(D7), 8875–8885, doi:10.1029/96JD02977.

Newchurch, M. J., M. A. Ayoub, S. Oltmans, B. Johnson, and F. J. Schmidlin (2003), Vertical distribution of ozone at four sites in the United States, *J. Geophys. Res.*, *108*(D1), 4031, doi:10.1029/2002JD002059.

Nowotarski, C. J., and A. A. Jensen (2013), Classifying proximity soundings with self-organizing maps toward improving supercell and tornado forecasting, *Weather Forecasting*, *28*, 783–801, doi:10.1175/WAF-D-12-00125.1.

Oltmans, S. J., et al. (2006), Long-term changes in tropospheric ozone, *Atmos. Environ.*, *40*(17), 3156–3173, doi:10.1016/j.atmosenv.2006.01.029.

Oltmans, S. J., et al. (2013), Recent tropospheric ozone changes—A pattern dominated by slow or no growth, *Atmos. Environ.*, *67*, 331–351, doi:10.1016/j.atmosenv.2012.10.057.

Rao, T. N., S. Kirkwood, J. Arvelius, P. von der Gathen, and R. Kivi (2003), Climatology of UTLS ozone and the ratio of ozone and potential vorticity over northern Europe, *J. Geophys. Res.*, *108*(D22), 4703, doi:10.1029/2003JD003860.

Stauffer, R. M., G. A. Morris, A. M. Thompson, E. Joseph, G. J. R. Coetzee, and N. R. Nalli (2014), Propagation of radiosonde pressure sensor errors to ozonesonde measurements, *Atmos. Meas. Tech.*, *7*, 65–79, doi:10.5194/amt-7-65-2014.

Stevenson, D. S., et al. (2006), Multimodel ensemble simulations of present-day and near-future tropospheric ozone, *J. Geophys. Res.*, *111*, D08301, doi:10.1029/2005JD006338.

Thompson, A. M., et al. (2003a), Southern Hemisphere Additional Ozonesondes (SHADOZ) 1998–2000 tropical ozone climatology: 1. Comparison with Total Ozone Mapping Spectrometer (TOMS), *J. Geophys. Res.*, *108*(D2), 8238, doi:10.1029/2001JD000967.

Thompson, A. M., et al. (2003b), Southern Hemisphere Additional Ozonesondes (SHADOZ) 1998–2000 tropical ozone climatology 2. Tropospheric variability and the zonal wave-one, *J. Geophys. Res.*, *108*(D2), 8241, doi:10.1029/2002JD002241.

Thompson, A. M., S. J. Oltmans, D. W. Tarasick, P. Von der Gathen, H. G. J. Smit, and J. C. Witte (2011), Strategic ozone sounding networks: Review of design and accomplishments, *Atmos. Environ.*, *45*(13), 2145–2163, doi:10.1016/j.atmosenv.2010.05.002.

Thompson, A. M., et al. (2012), Southern Hemisphere Additional Ozonesondes (SHADOZ) ozone climatology (2005–2009): Tropospheric and tropical tropopause layer (TTL) profiles with comparisons to OMI-based ozone products, *J. Geophys. Res.*, *117*, D23301, doi:10.1029/2011JD016911.

Tilmes, S., et al. (2012), Technical note: Ozonesonde climatology between 1995 and 2011: Description, evaluation and applications, *Atmos. Chem. Phys.*, *12*, 7475–7497, doi:10.5194/acp-12-7475-2012.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas (2000), SOM Toolbox for Matlab 5, report, Helsinki Univ. of Technol., Helsinki, Finland.

World Meteorological Organization (1957), Meteorology—A three-dimensional science: Second session of the Commission for Aerology, *WMO Bull.*, *4*(4), 134–138.