https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

Atmospheric Chemistry and Physics
Discussions
Open Access

Technical Note: Unsupervised classification of ozone profiles in UKESM1

Fouzia Fahrin[1,2], Daniel C. Jones[3], Yan Wu[2], James Keeble[4,5], and Alexander T. Archibald[4,5]

[1]Department of Geological and Atmospheric Sciences, Iowa State University, USA
[2]Department of Mathematical Sciences, Georgia Southern University, USA
[3]British Antarctic Survey, NERC, UKRI, Cambridge, UK
[4]Department of Chemistry, University of Cambridge, Cambridge, UK
[5]National Centre for Atmospheric Science (NCAS), University of Cambridge, Cambridge, UK

Correspondence: Fouzia Fahrin (ffahrin@iastate.edu)

Abstract. The vertical distribution of ozone in the atmosphere, which features complex spatial and temporal variability set by a balance of production, loss, and advection, is relevant for both surface air pollution and for climate via its role in radiative forcing. At present, the way in which regions of coherent ozone structure are defined relies on somewhat arbitrarily drawn boundaries. Here we consider a more general, data-driven method for defining coherent regimes of ozone structure; we apply an unsupervised classification technique called Gaussian Mixture Modelling (GMM), which represents the underlying distribution of ozone profiles as a linear combination of multi-dimensional Gaussian functions. In doing so, GMM identifies coherent groups or sub-populations of the ozone profile distribution. As a proof-of-concept study, we apply GMM to ozone profiles from three subsets of the UKESM1 coupled climate model runs carried out for CMIP6: specifically, a historical decade and two decades from two different future climate projections (i.e. SSP1-2.6, SSP5-8.5). Despite not being given any spatiotemporal information, GMM identifies several spatially coherent regions of ozone structure. Using a combination of statistical guidance and post-hoc judgment, we select a six-class representation of global ozone, consisting of two tropical classes and four mid-to-high latitude classes. The tropical classes feature a relatively high-altitude tropopause, while the higher-latitude classes feature a lower-altitude tropopause and low values of tropospheric ozone, as expected based on broad patterns observed in the atmosphere. Both of the future projections feature lower tropospheric ozone concentrations than the historical benchmark, with signatures of ozone hole recovery. We find that the area occupied by the tropical classes is expanded in both future projections, in consistency with the tropical broadening hypothesis. Our results suggest that GMM may be a useful method for identifying coherent ozone regimes, particularly in the context of model analysis.

1 Introduction

Earth's atmospheric ozone distribution is a topic of interest because of its effect on climate and its role in protecting surface-dwelling organisms from harmful ultraviolet radiation (Newman and Todara, 2003; Monks et al., 2015). The distribution of ozone varies both vertically and horizontally. Nearly 90% of ozone is found in the stratosphere, the layer of the atmosphere between 10-50 km, while 10% is found in the troposphere, the atmospheric layer extending from the surface to 10 km. Strato-

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

spheric ozone protects surface-dwelling life by reducing the number of high energy photons reaching the surface, which would otherwise lead to high occurrences of skin cancer, cataracts, and impaired immune systems (Newman and Todara, 2003; Monks et al., 2009). In contrast, near-surface tropospheric ozone poses a threat to human health as it is a pollutant (Monks et al., 2015).

The spatial variation in ozone is driven by complex atmospheric processes. Unlike many of the important trace gas species studied in the atmosphere, ozone is not directly emitted from natural or anthropogenic sources. Instead, atmospheric ozone concentrations are controlled by chemical, radiative, and dynamical processes that operate on a range of timescales. Adding further complication is the fact that these processes vary significantly with altitude. In the stratosphere, gas phase photochemical reactions involving oxygen produce ozone (Chapman, 1930), while it is destroyed through reactions involving chlorine, nitrogen, hydrogen, and bromine radical species (Bates and Nicolet, 1950; Crutzen, 1970; Johnston, 1971; Molina and Rowland, 1974; Cicerone et al., 1974). In contrast, tropospheric ozone is produced through photochemical oxidation of ozone precursors such as carbon monoxide (CO), methane (CH_4) and non-methane volatile organic compounds (NMVOCs) in the presence of nitrogen oxides (NO and NO_2). In a similar way, transport processes differ between the stratosphere and troposphere. Because of these different processes, understanding patterns in the vertical distribution of ozone remains a challenge (Monks et al., 2015). These ozone precursors can be transported far downwind from their source locations (Chameides et al., 1992; Monks et al., 2009).

Not only are there significant differences in the processes controlling local ozone mixing ratios at different altitudes, but these processes respond differently to changes in atmospheric composition and global climate. Past changes in anthropogenic emissions, biomass burning, and lightning have all contributed to increased emissions of ozone precursors and increased tropospheric ozone (Griffiths et al., 2021; Jaffe and Wigder, 2012; Monks et al., 2015; Laban et al., 2018). In contrast, emissions of halogenated ozone-depleting substances (ODS) at the end of the 20th century led to significant decreases in stratospheric ozone concentrations and the formation of the ozone hole (Keeble et al., 2021). Future projections of ozone concentrations are dependent on assumptions made about greenhouse gas, ozone precursor, and halogenated ODS emissions, and these changes may work against each other. For example, stratospheric ozone mixing ratios are expected to increase in the coming decades as ODS levels decline. However, an acceleration of the Brewer-Dobson circulation (BDC) associated with increasing greenhouse gas concentrations may lead to reductions in lower tropical stratospheric ozone mixing ratios, while increasing transport of ozone into the mid-latitudes troposphere. Because of these complex interactions, understanding future changes to the vertical distribution of ozone requires simulations performed by complex models.

Because of this complexity, chemistry-climate and Earth system models are often used to explore changes in atmospheric ozone. A key component in this evaluation is the comparison of ozone derived from different models and/or from different scenarios in the same model (Griffiths et al., 2021; Keeble et al., 2021). Often this is done at the global scale, but if regional comparisons are made, this is often done by averaging ozone profiles over set latitude ranges. However, owing to the complex, spatially heterogeneous processes controlling the distribution of ozone described above, this is a poor method for identifying regions with similar profiles. As climate and ozone mixing ratios change in the future, the boundaries between ozone profiles with similar characteristics might be expected to move. This feature would not be captured by averaging profiles over fixed

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

latitude ranges. In this work, in order to address this limitation in latitude-based averaging methods, we describe the vertical ozone structure with an unsupervised classification method that groups profiles into classes based on their similarity.

Clustering techniques have been already been used in ozone concentration studies for understanding long-term variability. Boleti et al. (2020) have applied a multidimensional clustering technique to understand the long-term trend of ozone. Chang et al. (2017) used a classification technique that is latitude dependent for regional ozone trend analysis. In our study, we adopt a Gaussian Mixture Modelling (GMM) approach, an automated, robust, and standardized unsupervised classification technique that has previously been applied to ocean structure and dynamics (Bishop, 2006; Maze et al., 2017; Jones et al., 2019; Sonnewald et al., 2019; Rosso et al., 2020). GMM does not use any latitude or longitude information to identify similar profiles and cluster them together, which makes it more general than a latitude-based averaging method. In section 2, we describe the method adopted in the study and the data set used in the study. In section 3, we present the results of the GMM-based clustering analysis. Finally, we end with a brief discussion 4 and conclusions 5.

2 Methods and data

Our approach is based on Gaussian Mixture Modelling (GMM), which is a type of unsupervised classification method. We want to model the vertical ozone structure, i.e to understand how we can identify different ozone profile types in a dataset. To do so, we analyze the diversity of vertical ozone profiles by way of identification of recurrent patterns throughout the collection of profiles using unsupervised learning.

2.1 UKESM1 Experiment Selection

The UK Earth System Model 1 (UKESM1, https://ukesm.ac.uk/) is a coupled climate model with a well-resolved stratosphere, tropospheric-stratospheric chemistry, ocean-atmosphere carbon and aerosol coupling, and terrestrial biogeochemistry (Sellar et al., 2019). The model has a horizontal resolution of 1.25^0 latitude by 1.875^0 longitude, with 85 vertical levels on a terrain-following hybrid height coordinate and a model top at 85 km (0.004 hPa). UKESM1's complex physical-biogeochemical coupling and its realistic representation of historical ozone structure and trends make it a suitable choice for our study (Keeble et al., 2021). Using the Pangeo platform, we selected annual mean ozone profile data from three different UKESM1 experiments (Abernathey et al., 2021). We chose annual mean profiles in order to focus on longer-term variability, but in principle one could use monthly means to include seasonal variations in ozone structure. Changes in ozone precursor emissions have an effect on future tropospheric ozone concentrations; reductions in precursor emissions drive ozone decreases in shared socioeconomic pathways (SSPs) (Griffiths et al., 2021). To explore the effect of emissions on the class properties, we used ozone data from three different experiments:

- **Historical**: Annual means covering the years 2004-2014.
- **SSP1-2.6**: Annual means covering the years 2090-2100 (strong emission reductions).
- **SSP5-8.5**: Annual means covering the years 2090-2100 (no emission reductions).

Here each simulation year contains 304128 annual mean profiles.

In order to create a training dataset for the GMM algorithm, we combined data from all three of the above experiments. Essentially, we trained the GMM in such a way that it “sees" structures from all three experiments and is thereby better able to represent the full range of possible structures, i.e. the training process is not biased towards one particular experiment. Using the trained GMM, we labeled the full dataset of ozone profiles from all three experiments. We then used the fully labeled dataset to look for differences in structure among the historical, SSP1-2.6, and SSP5-8.5 experiments.

At present, standard implementations of GMM cannot handle missing values. So in this context, one has to select a subset of the ozone profiles that feature values on every selected standard pressure level. We discarded any profiles with NaN values. As such, we only worked with profiles with values over the entire pressure range, from 1-1000 hPa. This means that much of our analysis takes place over the ocean and only partially covers land-based areas, i.e. out of necessity we omit grid cells with surface pressures lower than 1000 hPa due to topography.

2.2 Gaussian mixture modelling

Gaussian Mixture Modelling (GMM), a machine learning method, uses a probabilistic approach for describing and classifying data by representing the underlying data distribution using a linear combination of multi-dimensional Gaussian functions (McLachlan and Basford, 1988). By using a sufficient number of Gaussians, any continuous density field can be approximated to arbitrary accuracy. This allows us to identify and model the typical vertical structure represented in the collection of profiles.

Although GMM has been used in several oceanographic studies to date (Maze et al., 2017; Jones et al., 2019; Sonnewald et al., 2019; Houghton and Wilson, 2020; Sonnewald et al., 2020; Rosso et al., 2020; Desbruyères et al., 2021; Boehme and Rosso, 2021), to our knowledge, our application is novel in the field of atmospheric chemistry. One unique aspect of this approach is that we do not use any geographical information about the profiles to identify groups of similar profiles. Specifically, we withhold latitude, longitude, and time information from the unsupervised classification algorithm; it only sees the values of the ozone concentration on each standard pressure level. The motivation behind withholding the geographical information is that there is no reason for the vertical ozone structure of the profile to be unique to a given region (Maze et al., 2017).

The core foundation of a GMM, as described in Bishop (2006), is that any Probability Density Function (PDF) can be described as closely as desired with a model of weighted sums of Gaussian PDFs:

$$p(x) = \sum_{k=1}^{K} \lambda_k N(x|\mu_k, \Sigma_k) \tag{1}$$

which is called a mixture of Gaussians. Each Gaussian density $N(x|\mu_k, \Sigma_k)$, a multidimensional normal probability density function (PDF), is called a component of the mixture and has its own mean μ_k and covariance Σ_k. Where x is a single profile taken from the complete array X.

We use an expectation-maximization algorithm (Appendix B) to find the maximum likelihood solution for the model, which is effectively “training" the GMM to represent the underlying structure of the ozone data as represented in abstract principal

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

component space (section 2.3).

2.3 Dimension reduction

The abstract “feature space" in which we perform the clustering is relatively high-dimensional; ozone is defined on 19 standard pressure levels in our dataset. Because GMM becomes less efficient for high-dimensional problems, we apply a dimension reduction scheme to reduce the computational expense of the training step. A large number of dimensions in the problem fundamentally translates into a large number of parameters to be determined in the Gaussian covariance matrices. Here we used Principal Component Analysis (PCA), a dimension reduction method that is often used to reduce the dimension of large data sets by transforming a large set of variables into a smaller set that still retains an acceptable percentage of the variability.

As a first prepossessing step, we standardize the ozone values on each pressure level. Since the ozone values on each pressure level are standardized independently, “small" variations in ozone on levels with low variability can have roughly the same effect as “large" variations in ozone on levels with high variability. This ensures that the structure seen by GMM is not just dominated by the pressure levels on which the variability is high. This prepossessing step also helps to speed up the algorithm (Jaadi, 2019).

In last step of PCA, we express each ozone profile as a linear combination of eigenfunctions, using the following equation for $x(z)$:

$$x(z) = \sum_{j=1}^{d} P(z,j)y(j) \tag{2}$$

where z is the pressure level, d is the total number of PCs (index j), and $P(z,j)$ is the transformation matrix between pressure space and PC space. $P \in \mathbb{R}^{D\times d}$ and $y \in\ \mathbb{R}^{d\times N}$ with $d \leq D$. The first row of P contains profiles maximizing the structural variance throughout the collection of profiles. Thus, if we choose $d \leq D$, we can reduce the number of dimensions of the data set x while preserving most of its structure. This creates a new space where the N profiles are not defined with D vertical level values (the x array) but with only d values (y array). The transition between one space to the other is done through the matrix P containing the definition of the new dimensions in the original ones (d vertical profiles of D levels, the eigenvectors of the covariance matrix $x^T x$) (Figure A1).

We find that with 10 PCs, this transformation captures 99% of the variance in the vertical structure, which appropriately reduces the number of dimensions we need to describe the profile structure from UKESM1, that is, from 19 pressure levels to 10 PCs. A reduction to an even smaller number of PCs is possible at the expense of losing more of the variability in the original dataset.

2.4 Selection of the number of classes

We used a random sampling technique to select a subset to perform BIC to find the appropriate K for classes. The reason for random sampling is to test for the sensitivity of our results to the sample selection process. Under random sampling, each observation of the data set subset has an equal opportunity to be chosen as a part of the sampling process. Note that this sampling is not related to unbiased spatial sampling.

In our application, for each potential value of K we chose 20 different sets of 1000 random samples from the full dataset of 678,810 profiles. This sampling approach allowed us to estimate the mean and standard deviation of BIC at each K. We used the same random seed each time, so there is no variability associated with the random initial guesses for the cluster centers. The mean BIC curve appears to flatten after $K = 8$, indicating a point of diminishing returns for increasing K (Figure 1). The overfitting penalty term starts to dominate for $K > 12$, indicating an upper bound for the number of classes.

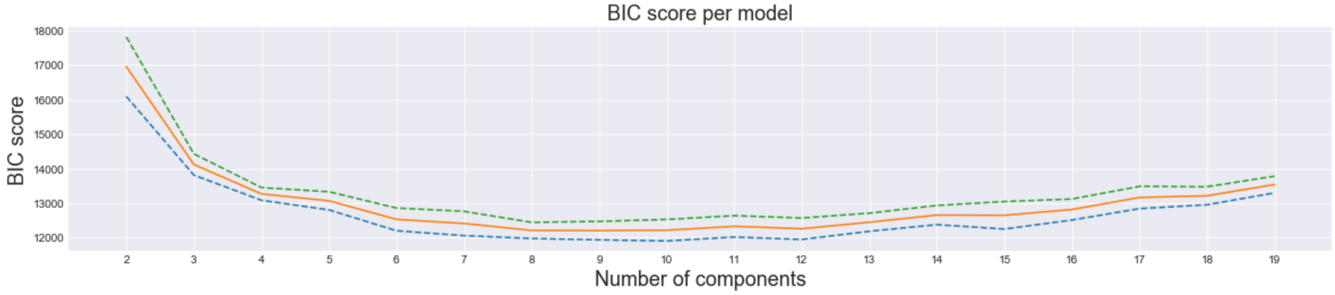


Figure 1. BIC Score versus the specified number of classes K for UKESM1 data. The solid line is the mean BIC value and the dashed lines represent one standard deviation on either side of the mean.

After examining the class structures produced by the $K = 8$ model and observing the very close similarity in structure and variability between some of the classes, we decided to manually merge two of the classes together. Specifically, we found a class in the high-latitude Northern Hemisphere that occupied a very small surface area and had a very similar structure to the much larger Northern Hemispheric class, so we merged them. This sort of post-hoc grouping has been used in oceanographic applications for similar reasons, illustrating the continued importance of domain expertise in this particular machine learning application (Rosso et al., 2020). After manually grouping these classes, we converged on a $K = 6$ model that features two Southern Hemispheric classes, two tropical classes, and two Northern Hemispheric classes. The BIC curve indicates that the loss of two classes from $K = 8$ to $K = 6$ comes at the cost of a relatively small decrease in likelihood, which is acceptable given the increase in ease of interpretability. Generally speaking, as one enhances the statistical complexity of the GMM by increasing K, the statistical model may become harder to interpret in terms of simpler conceptual models (Sonnewald et al., 2021).

3 Classification of UKESM ozone profiles

3.1 Ozone profile description

Our purpose is to identify coherent patterns within the collection of profiles using unsupervised machine learning. Overall, the profiles reveal relatively high ozone levels in the lower and middle stratosphere which peak and then decrease gradually in the upper stratosphere. The tropopause is located at around 300 hPa, with a high concentration of ozone just above it. Ozone concentrations peak at around 70 hPa then decrease at higher altitudes (Figure 2). In the troposphere, the ozone concentration is fairly constant and then increases towards the surface, in part due to pollution and biomass burning (Jaffe and Wigder, 2012).

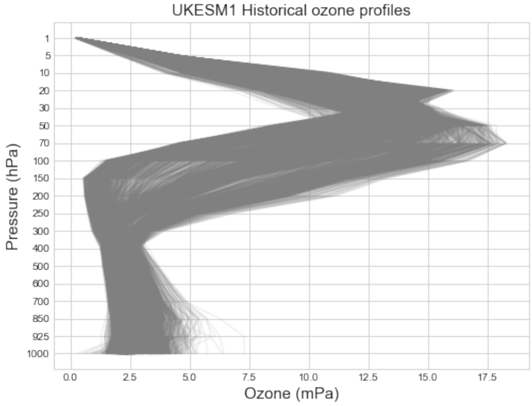


Figure 2. Ozone profile of UK Earth System model data (2004-2014 year period), as an example of how ozone concentration is changing with height.

3.2 Classification of ozone profiles in the historical experiment

In this section, we analyze the vertical structure of ozone data from the UKESM simulations that represent a chosen decade of the historical period (2004-2014). Our results are not especially sensitive to the choice of decade, since we train the GMM using multiple decades from a variety of atmospheric ozone states. The classes are sorted by mean latitude for ease of interpretation.

Proceeding from south to north: classes 1 and 2 are high-latitude Southern Ocean classes with similar mean profiles but different variability structures as measured by the standard deviation curves (Figure 3). They both feature relatively low-altitude and gentle tropopauses as indicated by the slope of the ozone curves. Class 1 has lowest surface ozone value with a mean 2.02 ± 0.25 mPa (Table 1); it has high variability in the middle stratosphere (at 50 hPa, the standard deviation is 1.40 mPa) and lower stratospheric ozone than any other class (only 14 mPa at 50 hPa) (Table 2). The relatively low value of stratospheric ozone is associated with the ozone hole and has the largest effect on class 1, based on its position at high southern latitudes (Wargan et al., 2020). The mean posterior probability, which in the context of a given statistical model is a measure of the algorithm's confidence in its assignment, is somewhat lower for class 2 than for class 1, indicating that there is some ambiguity associated with the assignments into class 2, which may be somewhat of a boundary or transition class between the high southern latitudes and the tropics. Note that high posterior probabilities do not necessarily indicate that the particular GMM is the best fit to the data, only that the selected GMM is confident in its assignment as measured by the uncertainty. Class 2 is highly variable throughout the upper troposphere and tropopause. Notably, all of the high-latitude Southern Hemispheric classes feature relatively low near-surface ozone values with small variability - they are relatively “clean" in terms of surface ozone pollution (Table 1).

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

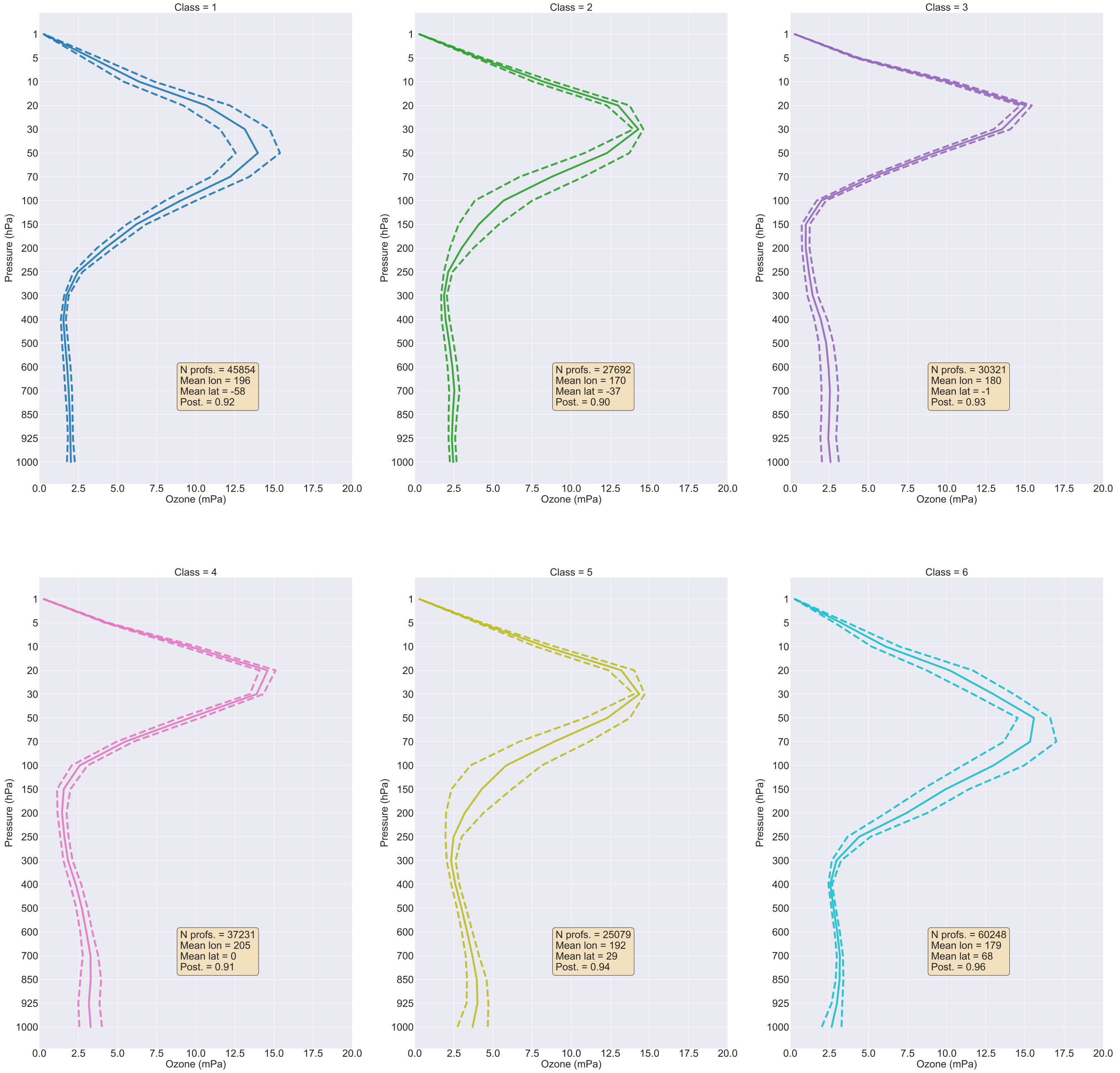


Figure 3. Ozone concentration statistics of UK Earth System model data (2004-2014 year period), separated by class, as a function of pressure, sorted by latitude. Shown are the mean (solid lines) and the mean plus or minus one standard deviation (dashed line) for all profiles in the indicated class. Also shown are the number of profiles in each class and the class mean values for longitude, latitude, and posterior probability.

Classes 3 and 4 are tropical classes, with higher surface ozone concentrations and a higher-altitude tropopause compared with the Southern Hemispheric classes (Figure 3). Class 3 and class 4 share similar kind of structure from surface to the upper stratosphere. Class 4 features more surface ozone and higher variability than class 3. In the stratosphere (around 20 hPa), class 4 exhibits the highest mean ozone concentration than any other class (14.60 ± 0.5 mPa) (Table 2). Finally, classes 5 and 6 are northern hemispheric classes with high near-surface ozone concentrations and large variability from tropopause to stratosphere(Table 1 & 2). These higher surface values result from greater surface pollutants, including the associated ozone precursor emissions, which tend to be concentrated in the Northern Hemisphere due to anthropogenic emissions (Monks et al., 2009, 2015).

Progressing from south to north, we see that the altitude of the maximum ozone concentration generally increases in height from the high-latitude southern hemisphere to the tropics and then decreases in height from the tropics to the high-latitude northern hemisphere (Figure 3). This structure is consistent with observations and is enforced by the meridional Brewer-Dobson circulation (Butchart, 2014), which is associated with upwelling in the tropics and downwelling in the extratropics, somewhat favoring the southern hemisphere (Butchart, 2014; Li and Thompson, 2013; Newman and Todara, 2003; Weber et al., 2011). The imprint of this circulation pattern is a low-altitude tropopause at the poles and a higher-altitude tropopause at the equator.

The label map indicates the geographic distribution of the classes during the historical decade (Figure 4). Notably, although the GMM algorithm was not given any latitude or longitude information, it was nevertheless able to identify spatially coherent groups. Over the ocean, the classes are largely organized in roughly zonal bands, with some exceptions (e.g. the tropical Atlantic). This zonal structure reflects the spatial structure of the zonal and meridional circulation patterns and ozone chemistry that leave their imprint on the atmospheric ozone structure.

From the tropopause to stratosphere to the stratosphere, classes 5 and 6 feature a relatively large standard deviation, especially in the lower and mid stratosphere, suggesting that these classes consist of a wide variety of profiles. The high surface ozone concentration in class 5 (3.69 mPa) highlights the bulk of biomass burning and wildfire, which occurs primarily near the Arctic Circle, Africa, and also in some parts of North America (Laban et al., 2018; Jaffe and Wigder, 2012). In the last few decades, wildfires/biomass burning have gained much attention as they have been recognized as the second-largest source of ozone precursor emissions (Monks et al., 2015). Boreal forest fires are reasonable source of high-surface ozone over North America (Jaffe and Wigder, 2012). Africa produces a significant amount of ozone precursor by biomass burning. Arctic boreal fire and biomass burning are sources of high ozone precursors over the Northern extratropical and temperate zone (Laban et al., 2018; Monks et al., 2015; Jaffe and Wigder, 2012).

Classes at the northern high latitudes (i.e. class 6) have more stratospheric ozone than those at southern high latitudes. This is a consequence of the fact that the northern hemisphere ozone hole is not especially present or dominant. Larger amplitudes of upward propagating planetary waves like Rossby waves can propagate from troposphere to stratosphere with eastward wind, where these waves can perturb stratospheric circulation and reduce the speed of polar night jet (Lee, 2021; Oehrlein et al., 2020; Waugh et al., 2017). In the Northern Hemisphere, the layout of the continents and mountain ranges accelerate this wave activity than in the Southern Hemisphere (Lee, 2021; Waugh et al., 2017). Consequently, the Arctic stratospheric vortex is much weaker

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

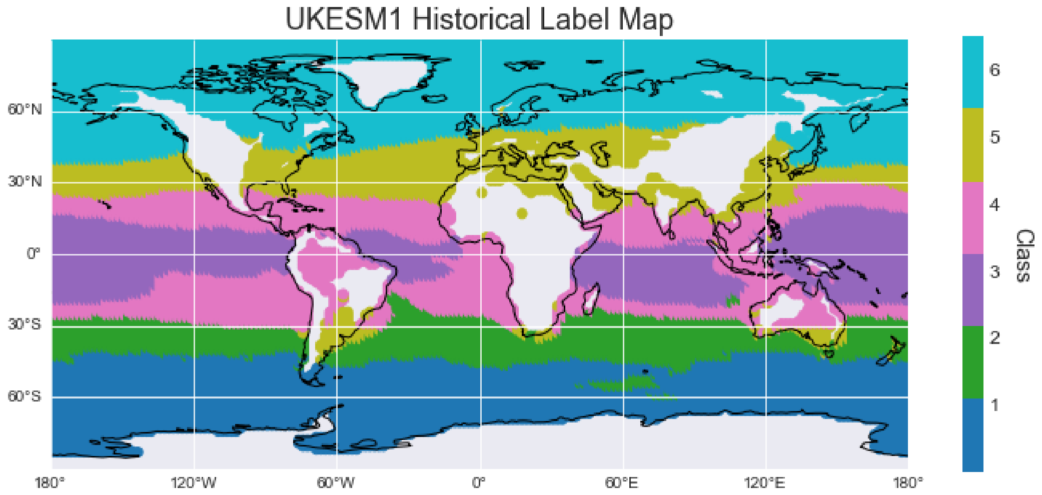


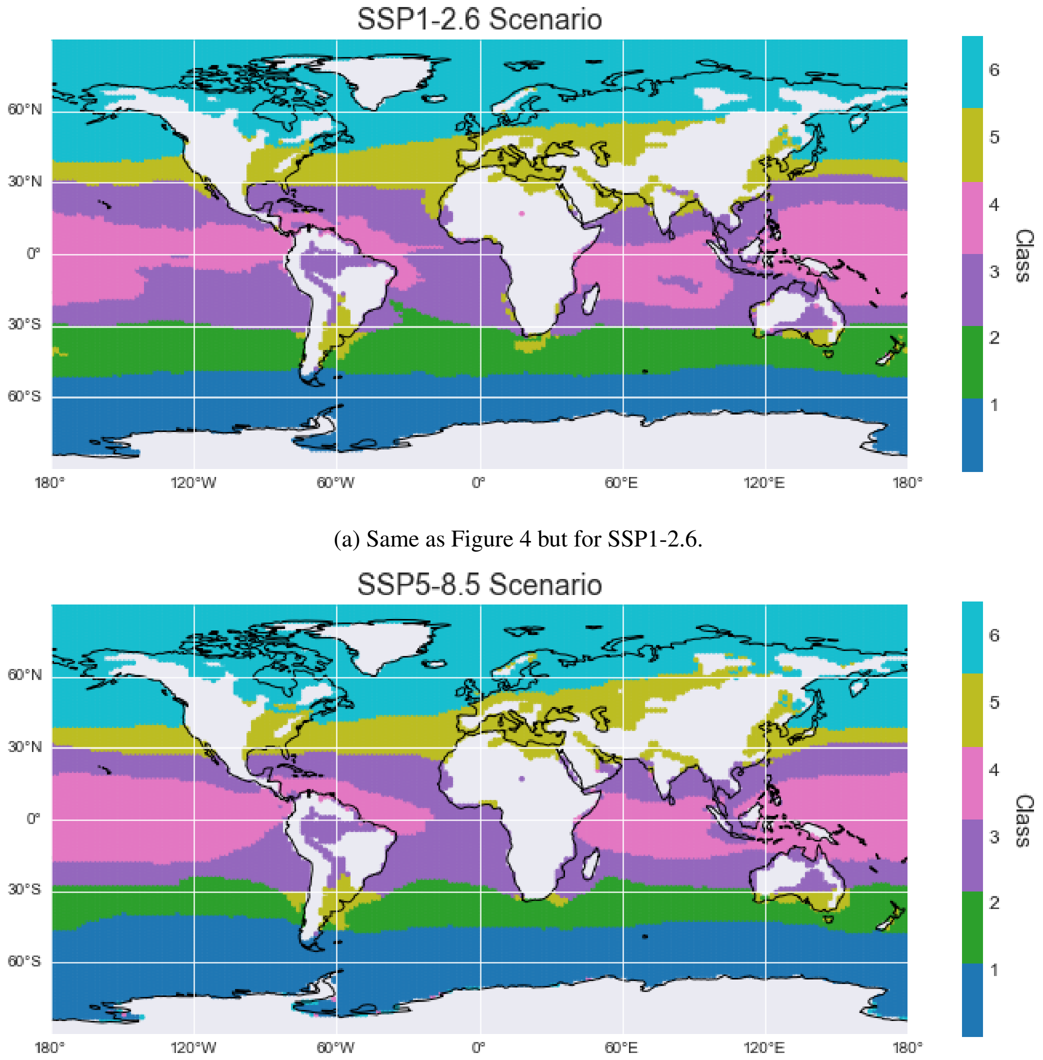
Figure 4. Map of profiles color-coded with the class they have been attributed to for model historical data (annual means covering 2004-2014 at each model grid cell). We discarded profiles that have missing values, such as those which do not start at 1000 hPa. The labels are assigned to every annual mean profile; the above plot indicates the median label assigned to the profiles at each model grid cell.

and more variable than its Antarctic counterpart which features larger meanders in the meridional extent. It is for this reason that, unlike the Antarctic, a large ozone hole does not form in the Arctic stratosphere each winter. As the Arctic temperature is higher than the Antarctic, a strong Antarctic vortex allows for the formation of polar stratospheric clouds that catalyze ozone depletion (Waugh et al., 2017; Lee, 2021; Newman and Todara, 2003). This allows redistribution of stratospheric ozone and pulls ozone from the tropics in Northern Hemisphere (Lee, 2021; Newman and Todara, 2003). The strong polar vortex in the south pole prevents the region from having a high stratospheric ozone (Newman and Todara, 2003).

3.3 Classification of ozone profiles in the future climate projections SSP1-2.6 and SSP5-8.5

We examine the distribution and structure of ozone in two chosen future climate projections, namely SSP1-2.6 and SSP5-8.5. SSP1-2.6 is a scenario with strong emission reductions and SSP5-8.5 is with increased emissions. We chose these two experiments as end-members representing two drastically different future projections. In the SSP1-2.6 case, with reduced emissions of ozone precursors, the total surface ozone concentration gets smaller, as expected (Table 1). In the SSP5-8.5 case, with increased emissions of ozone precursors, the total surface ozone concentration is slightly increased or approximately steady, also as expected. Classes 1 and 6 in particular, which are affected by the ozone hole, display a large increase in stratospheric ozone between 2004-2014 and 2090-2100 in both cases, which is a signature of the closing of the ozone hole (Keeble et al., 2021). The maximum concentration is located around 50 hPa in the historical case, which is just above the region of maximum ozone depletion. The recovery of the ozone hole also shifts the level of maximum ozone concentration to lower altitudes (higher pressures, i.e. from 50 hPa to 70 hPa) for both hemispheres in future projections. In the next subsections, we investigate differences in the spatial structure of the two future emissions experiments.

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022



(a) Same as Figure 4 but for SSP1-2.6.

(b) Same as Figure 4 but for SSP5-8.5

Figure 5. SSP Label Map covering year 2090-2100.

3.3.1 Ozone profile structure in SSP1-2.6

Here we examine the ozone structure in SSP1-2.6 over the decade 2090-2100. As with the historical experiment, class 1 has the lowest surface ozone (Table 1), in consistency with the reduction in surface ozone precursors. The maximum value of stratospheric ozone increases under this scenario, from 14.0 mPa to 17.4 mPa in the mean (Table 2), which is a signature of the recovery of the ozone hole (Keeble et al., 2021).

Moving northwards, class 2 appears to have a similar structure to its historical counterpart, with higher stratospheric ozone and considerable variability in the upper troposphere to middle atmosphere (Figure 5a and 6). It is a mid-latitude southern hemispheric class occupying roughly the same total surface area as it did in the historical experiment (Table 3). As with the historical case, classes 3 and 4 are tropical classes with relatively high-altitude, sharp tropopause, and larger variability near the surface. Notably, the relative positions of classes 3 and 4 are swapped between the historical case and the 2090-2100 decade of SSP1-2.6, indicating that these two classes may be difficult to unambiguously differentiate over this period. Classes 5 and 6 are similar to their historical counterpart in structure, except with reduced surface ozone concentrations and increased stratospheric ozone for high latitude classes, consistent with continued ozone precursor emissions reductions (Tables 1 and 2). The tropospheric ozone decrease is significant in the NH than in other scenarios, helping to mitigate climate change and air quality impacts (Table 1) (Keeble et al., 2021).

3.3.2 Ozone profile structure in SSP5-8.5

Here we examine the structure of atmospheric ozone in the 2090-2100 decade of the SSP5-8.5 experiment, where ozone mixing ratios are projected to increase throughout much of the troposphere and upper stratosphere (Keeble et al., 2021). Proceeding from south to north, we see that classes 1 and 2 are similar to their historical counterparts, covering a similar proportion of area, albeit with increased stratospheric ozone at the pressure level with maximum concentration (Table 2). The mean posterior probability of class 2 is higher than before, indicating a better fit for class 2 during this last decade of SSP5-8.5 (Figures 6 & 7). The vertical structure of class 2 features more variability in the upper troposphere compared with its historical and SSP1-2.6 counterparts. Moving northwards, class 3 is similar to its historical and SSP1-2.6 counterparts, albeit with larger variability throughout the lower and upper troposphere (Figure 7). Future ozone depletion decrease will lead to ozone concentration increase throughout the atmosphere, and both hemispheric high-latitude upper stratosphere will have the largest changes (Table 2 class 1 in SSP5-8.5) (Griffiths et al., 2021). However, an increasing amount of greenhouse gas emission will yield a more complex pattern of ozone changes, which will lead to a possible strengthening of the Brewer-Dobson circulation to an increase in net stratospheric influx, and high tropospheric ozone in the Southern Hemisphere class is the result of circulation changes (Class 1 in SSP5-8.5 in Table 1) (Young et al., 2013; Monks et al., 2015; Butchart, 2014; Griffiths et al., 2021; Lu et al., 2019).

The tropical classes (i.e. 3 and 4) are similar to those seen in SSP1-2.6 and have switched places relative to the historical case (Figure 5b). Notably, the vertical structures of these two classes in SSP5-8.5 have much higher mean posterior probabilities, indicating that these classes are a more suitable fit for this time period (Figure 7). Again we see that training the GMM for all three decades has produced classes that are influenced by structures in all three experiments. Finally, classes 5 and 6 remain large-scale Northern Hemispheric classes, although class 6 has increased surface ozone concentrations relative to SSP1-2.6, in part due to continued precursor emissions. In response to tropospheric warming driven by greenhouse gas in SSP5-8.5, the subtropical tropospheric jets intensify, while the contribution of gravity waves increases in the middle stratosphere (Palmeiro et al., 2014). As a result, stratospheric ozone increases in high latitude classes (Table 2).

The modelled surface ozone results suggest that levels of surface ozone over the ocean are lower than those over any of the land-based regions (Figures 4, 5a and 5b). The oceans are major sinks of tropospheric ozone at the surface and there are few

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

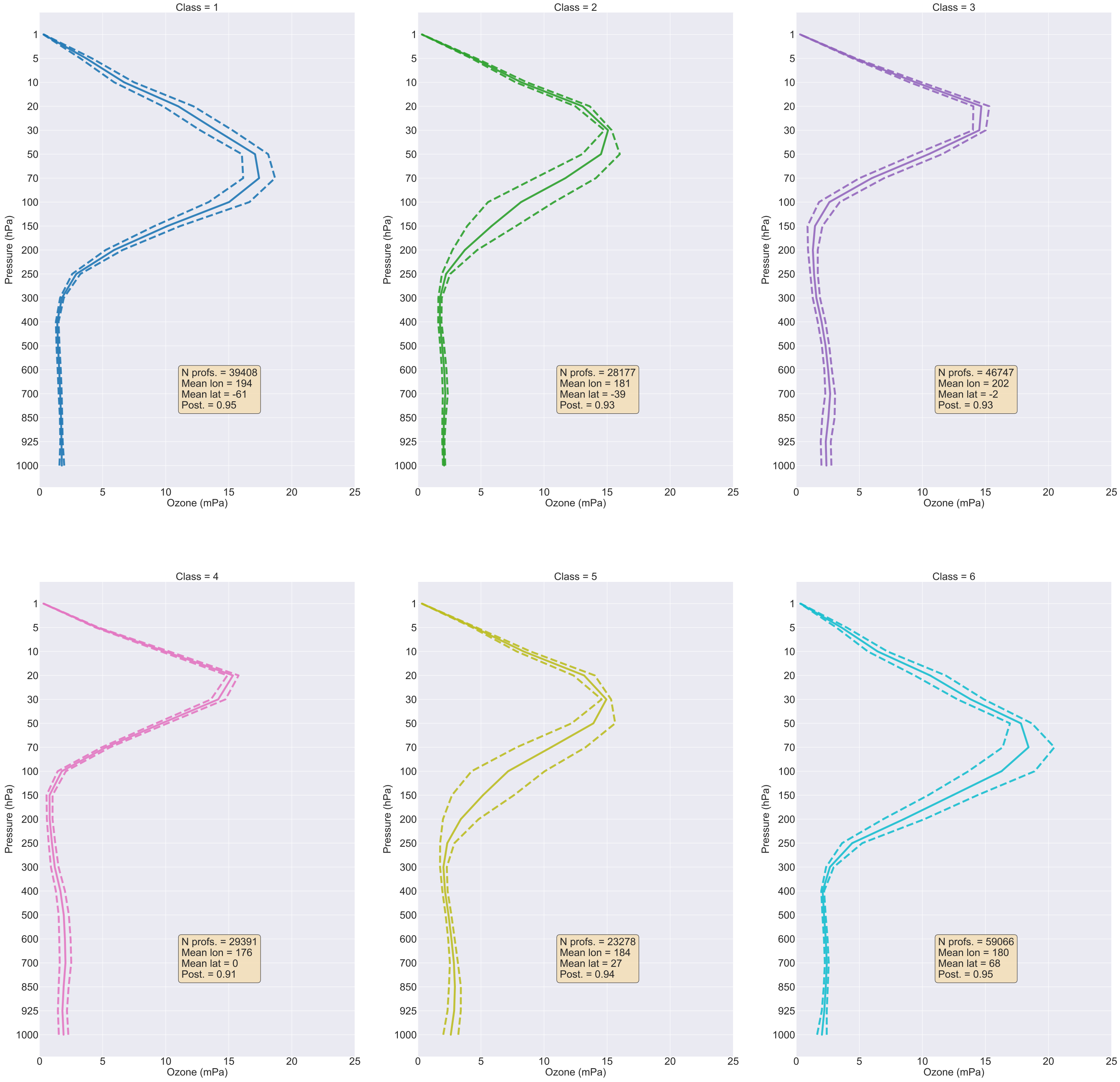


Figure 6. Same as Figure 3 but for SSP1-2.6 covering 2090-2100.

direct sources of ozone precursors present over the ocean (Archibald et al., 2020a, b). Advection of emission-driven ozone production over the land or an increase of ozone transport from the stratosphere are responsible for ozone increase over the ocean (Archibald et al., 2020a, b).

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

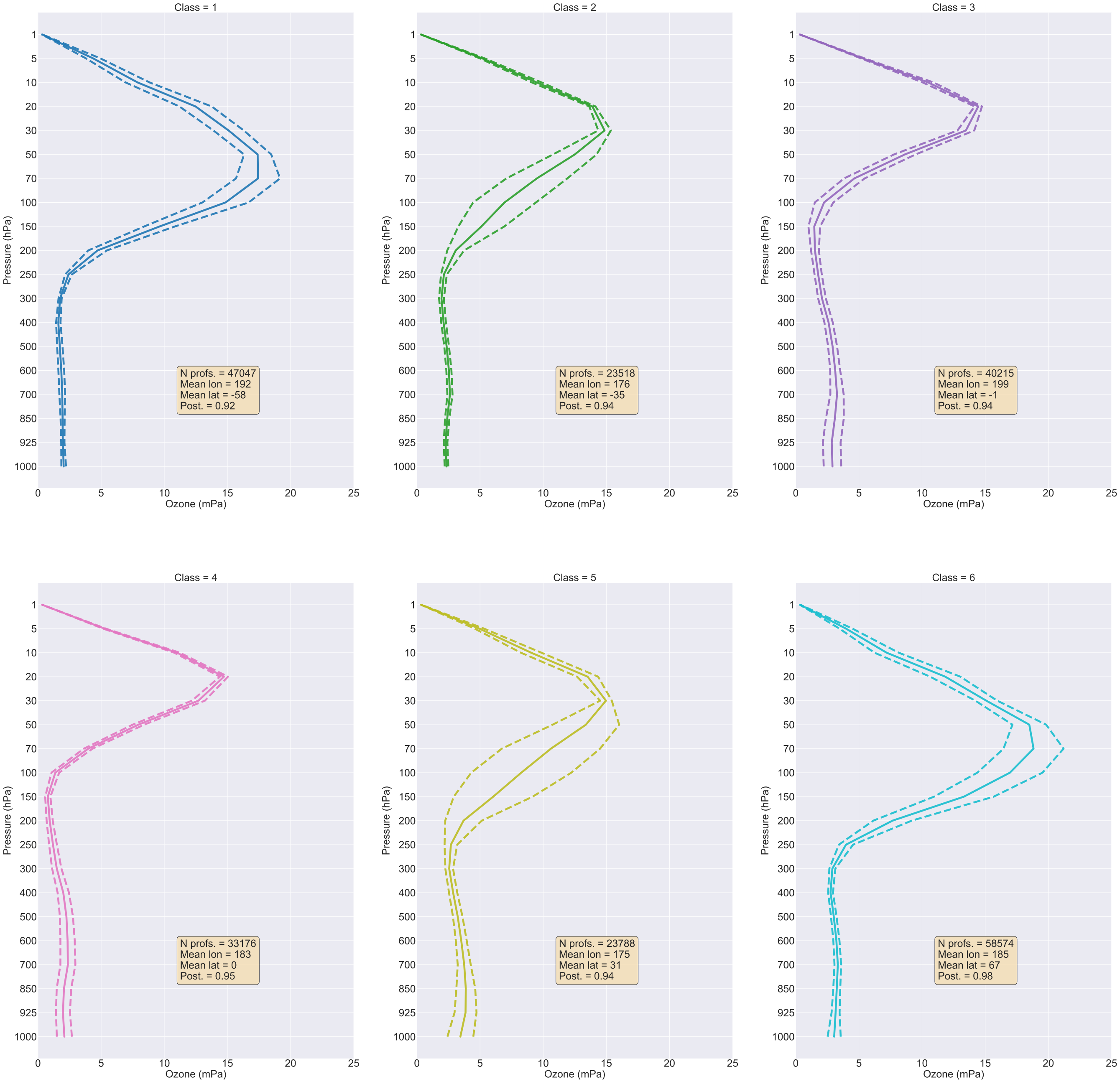


Figure 7. Same as Figure 3 but for SSP5-8.5 covering 2090-2100.

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

4 Discussion

The GMM-based classification of ozone profiles highlights some key features in the UKESM1 ozone data. One of the main points of our study is that even though the GMM algorithm was not supplied with the latitudes of the profiles, the classes nevertheless vary with latitude, and we can provide plausible physical-chemical explanations for the changes in the vertical structure of each class. The tropics, for example, feature profiles from two distinct classes broadly occupying the area between 30°S-30°N with the highest tropopause height, while, for the polar classes (classes 1 and 6) the tropopause height is the lowest.

Our results indicate that under both future climate states explored in this study, the area covered by the tropical classes (3 and 4) is projected to increase poleward at the expense of the polar and midlatitude classes, and Figure 5 suggests that expansion is expected to mostly take place in the Northern Hemisphere (NH).

In the projections of future climate considered here, both Hemispheric high latitudes show large variations in stratospheric ozone. These changes in the ozone concentration for high-latitude classes (i.e. classes 1 and 6) in future projections show the potential changes due to changes in precursor emissions and changes in ozone advection. Southern Hemispheric tropospheric ozone levels are generally low for all three cases considered here. There are larger fluctuations at the surface at high latitudes of the Northern Hemisphere (class 6), which could be related to differences in precursor transport and chemistry from lower latitudes.

This study focuses on model comparison. When working with model data, we typically have access to fairly uniform spatial and temporal ozone coverage, at least in parts of the atmosphere with a full range of pressures from 1000-1 hPa. This coverage allows us to train our mixture model in a way that is relatively unbiased with respect to location and time. The trained mixture model is thus able to identify coherent regimes with similar patterns of vertical variability in a way that is more general than drawing somewhat ad-hoc latitude-longitude boxes. Because we can train the mixture model using data from a variety of times and experiments, it is possible to train a GMM that can in principle represent the full range of data structures found within a selected ensemble. Although we did not attempt to do so here, it should be possible to use GMM for inter-model comparison, allowing for the structures and differences in structures to be derived directly from model data.

It is possible to apply GMM to observed ozone profiles as well. At present, ozone observations are biased towards a few specific locations where long-term monitoring has taken place (locations of World Ozone and Ultraviolet Data Center (WOUDC) ozonesonde stations can be found on the following link: http://dx.doi.org/10.14287/10000008); training a GMM on this data would necessarily bias the classes towards particular locations and times, making direct comparisons between models and observations difficult. One possible solution would be to train a GMM on model data and then apply it to observations, although any systematic biases would have to be treated carefully during the data cleaning and prepossessing steps. In any case, to the extent that the classes derived in this work, which uses a state-of-the-art and thoroughly verified Earth system model with coupled chemistry and climate, are representative of the structures present in the atmosphere, our results suggest that we may be over-sampling some regions of the global ozone distribution and under-sampling others.

Class	Hist (mean)	(std)	SSP1-2.6 (mean)	(std)	SSP5-8.5 (mean)	(std)
1	2.020	0.250	1.760	0.190	2.040	0.180
2	2.460	0.210	2.080	0.090	2.340	0.150
3	2.570	0.540	2.390	0.390	2.910	0.690
4	3.280	0.720	1.900	0.380	2.090	0.600
5	3.690	0.970	2.610	0.600	3.440	1.020
6	2.640	0.630	2.030	0.380	3.040	0.520

Table 1. Ozone concentration statistics at 1000 hPa for the historical, SSP1-2.6, and SSP5-8.5 epxeriments, shown in mPa

Class	Hist (lev) [hPa]	(mean)	(std)	SSP1-2.6 (lev) [hPa]	(mean)	(std)	SSP5-8.5 (lev) [hPa]	(mean)	(std)
1	50	14.00	1.40	70	17.4	1.26	70	17.4	1.70
2	30	14.30	0.32	30	15.1	0.29	30	14.9	0.50
3	20	15.10	0.40	20	14.7	0.63	20	14.4	0.29
4	20	14.60	0.50	20	15.3	0.44	20	14.7	0.32
5	30	14.35	0.40	30	14.9	0.37	30	15.0	0.46
6	50	15.60	1.04	70	18.4	2.05	70	18.8	2.40

Table 2. Pressure level (lev) of the maximum value of class mean ozone concentration. The mean and standard deviation values of the class statistics are given in mPa.

In terms of working towards a more optimized ozone observing system, it may be useful to use GMM and similar classification methods to identify which regions feature coherent variability.

Class	Historical	SSP1-2.6	SSP5-8.5
1	15.5	12.4	16.0
2	14.4	14.2	12.4
3	19.7	28.8	24.8
4	23.2	19.1	21.5
5	13.4	12.1	11.8
6	13.9	13.5	13.5

Table 3. Relative area occupied by each class, shown in percentages.

Region	Historical	SSP1-2.6	SSP5-8.5
Southern hemispheric (classes 1+2)	29.9	26.4	28.4
Tropical (classes 3+4)	42.9	47.9	46.3
Northern hemispheric (classes 5+6)	27.2	25.6	25.3

Table 4. Relative area coverage of three combined regions, shown in percentages.

5 Conclusions

In this study, we applied Gaussian Mixture Modelling (GMM), an unsupervised classification method, to ozone profiles from the UKESM1 coupled climate model in order to robustly and objectively identify coherent sets of ozone profile types. Our motive was to investigate the ozone structure using a limited number of classes. We used Principal Component Analysis (PCA) to reduce the computational complexity of the problem, increasing the computational efficiency at the expense of only 1% of the variability in the dataset. We used a statistical approach (i.e. BIC) and post-hoc expert judgment to inform our choice of the number of classes, settling on a six-class representation of the ozone profiles. This six-class system included two tropical classes and four mid-to-high latitude classes. We found that, although the GMM algorithm was not given any spatiotemporal information, it was able to identify a set of spatially coherent regions of ozone structure. We trained the GMM using data from all three model cases in order to expose it to the full range of profile types in our classification problem. We compared surface and maximum ozone concentrations for three model cases and their spatial extents. Surface ozone in the SSP1-2.6 case is projected to decrease than both historical and SSP5-8.5 case (Table 1). High stratospheric ozone in classes 1 and 6 in both of the future projection cases indicates a decrease in ozone depletion and possible ozone hole recovery, which results in a decrease in tropopause height and a transition of maximum ozone from 50 hPa to 70 hPa because of the thicker stratospheric ozone layer (Table 2). The modelled surface ozone shows high variability in the Northern Hemisphere (NH) and low variability in the Southern Hemisphere (SH). Notably, the spatial area occupied by the tropical classes increased in both future projections relative to the historical benchmark, in consistency with the tropical broadening hypothesis, i.e. the expected expansion of tropical upwelling. GMM can be applied to identify data-derived regions of coherent ozone structure and may therefore be useful for model-model comparison or model-data comparison.

Appendix A: Principal Component Analysis (PCA)

The principal component analysis shown in figure A1 is adopted for dimensionality reduction in this work. The figure shows the eigenfunctions. These eigenfunctions came from the eigenvalues and corresponding eigenvectors of the covariance matrix to find the directions along which the variability is the largest.

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

Atmospheric Chemistry and Physics Discussions

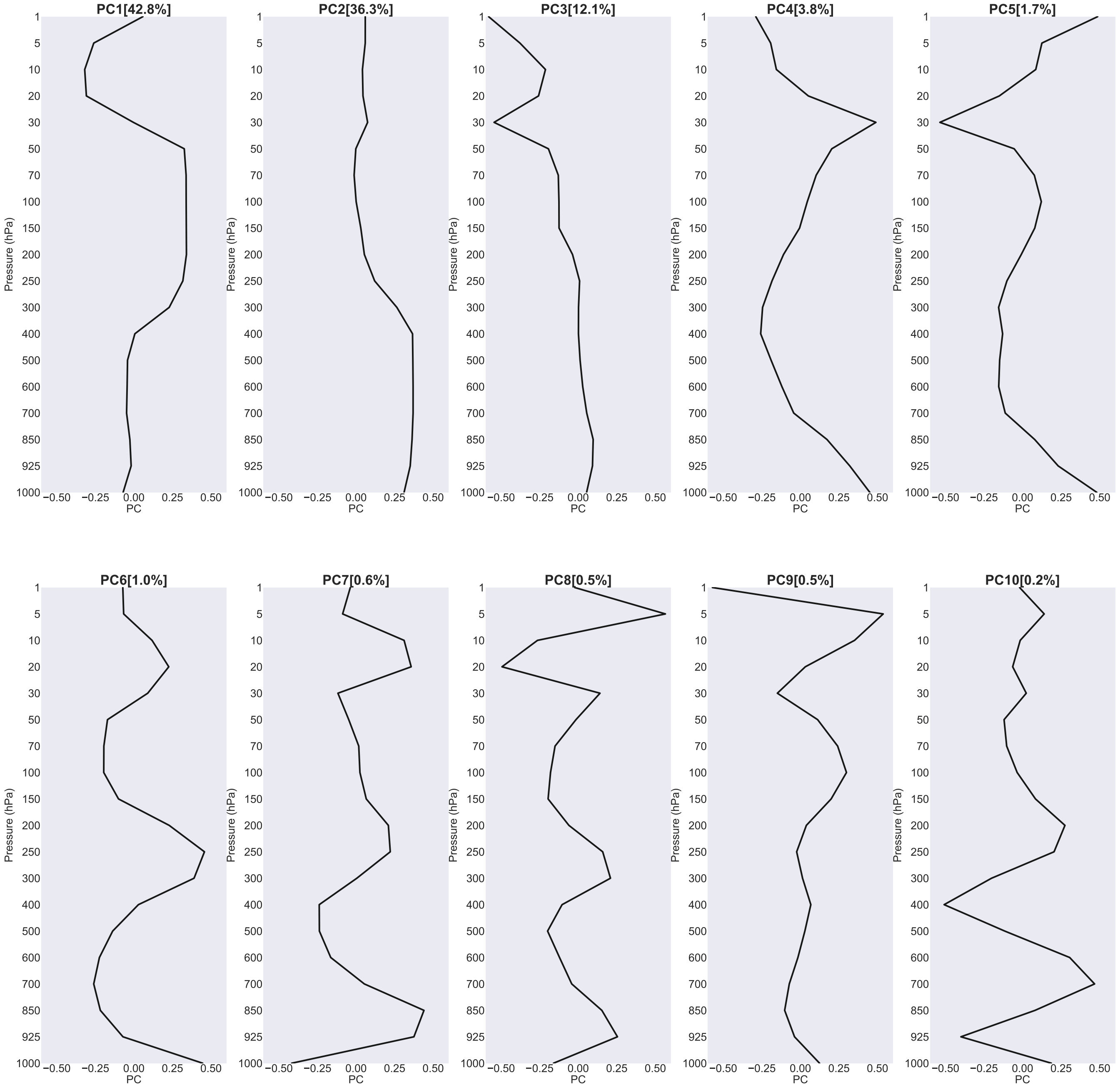


Figure A1. Principal components (PCs) with percent variance statistically explained by each PC is shown (in parenthesis).

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

Appendix B: GMM details

For details of the GMM classification algorithm we refer the readers to Bishop (2006). The classification algorithm is adopted from Bishop (2006); Maze et al. (2017)

B1 Probability density function of profiles

The key ingredient of GMM: a multidimensional normal probability density function (PDF) with mean μ and covariance Σ:

$$N(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^D}|\Sigma|^{1/2}} \exp\big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\big) \tag{B1}$$

In this study, $x \in \mathbb{R}^{D\times 1}$ is a profile of the $\mathbf{X} \in \mathbb{R}^{D\times N}$ collection, μ_k is a D-dimensional mean vector and $\mu_k \in \mathbb{R}^{D\times 1}$, $\Sigma \in \mathbb{R}^{D\times D}$ a covariance matrix and $|\Sigma|$ is the determinant.
In other words, the array $\mathbf{X}$ is the data set we want to analyze; it is made of N vertical profiles (as columns) of D pressure levels (as rows). The functional dependence of the Gaussian on the x is through the quadratic form, $\Delta^2 = (x-\mu)^T\Sigma^{-1}(x-\mu)$, which appears in the exponent in Eq. (B1). We consider a superposition of K Gaussian densities of the form, where the quantity Δ is called the *Mahalanobis distance* from μ to x and it reduces to the Euclidean distance when Σ is the identity matrix (Bishop, 2006).

The joint distribution will be $p(z)p(x|z)$ and the marginal distribution of x is,

$$p(x) = \sum_z p(x,z) = \sum_z p(z)p(x|z) \tag{B2}$$

Here, $\sum_z p(x,z)$ is the probability distribution for the observations $x_1, ..., x_N$. So for every observed data point x_n, there is a corresponding latent variable z_n.

GMM represents the PDF as weighted sum of K Gaussian classes as in Eq. (1). If we integrate Eq. (B1) with respect to x, and note that both p(x) and Gaussian components are normalized we obtain,

$$\sum_{k=1}^{K} \lambda_k = 1 \tag{B3}$$

We call the parameters λ_k mixing coefficients. The requirement $p(x) \geq 0$ together with $N(x|\mu_k,\Sigma_k) \geq 0$ implies $\lambda_k \geq 0$ for all k.
Combining these conditions, we can write, $0 \leq \lambda_k \leq 1$. The latent variable z is a K- dimensional binary random variable, having a 1-of-K representation in which a particular element $z_k = 1$ and the rest are equal to 0. Therefore, $z_k \in \{0,1\}$ and $\sum_k z_k = 1$ and there are K possible states for the vector z according to which element is nonzero. The joint distribution $p(x,z)$

https://doi.org/10.5194/acp-2022-423
Preprint. Discussion started: 26 July 2022

in terms of a marginal distribution $p(z)$ and a conditional distribution $p(x|z)$. The marginal distribution over z is specified in terms of the mixing coefficients λ_k, such that

$$p(z_k = 1) = \lambda_k$$

Because, z uses a 1-of-K representation, Eq. (1) can be written in the from

$$p(z) = \prod_{k=1}^{K} \lambda_k^{z_k} \tag{B4}$$

The conditional distribution of x given a particular value for z is a Gaussian

$$p(x|z_k = 1) = N(x|\mu_k, \Sigma_k) \tag{B5}$$

which can be written in the form,

$$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k} \tag{B6}$$

The joint distribution will be $p(z)p(x|z)$ and the marginal distribution of x is,

$$\begin{aligned} p(x) &= \sum_z p(x,z) = \sum_z p(z)p(x|z) \\ &= \sum_{k=1}^{K} \lambda_k N(x|\mu_k, \Sigma_k) \\ &= \sum_{k=1}^{K} p(z_k = 1)p(x|z_k = 1) \\ &= \sum_{k=1}^{K} \lambda_k p_k(x) \end{aligned} \tag{B7}$$

This equation is also called *Mixture distribution*.

Here, $p(x)$ stands for the observed PDF, and $\sum_z p(x,z)$ is the probability distribution for the observations $x_1, ..., x_N$. So for every observed data point x_n, there is a corresponding latent variable z_n.

Gaussian mixture modelling nails down to an optimization problem that can be tackled by maximizing the likelihood of observed profiles. This optimization is referred to as a *model training*. It is solved with the Expectation- Maximization method. The conditional probability of z given x plays an important role in the Expectation-Maximization algorithm. $\gamma(z_k)$ represents $p(z_k = 1|x)$ whose value can be found using the Bayes theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

So,

$$\gamma(z_k) \equiv p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{k=1}^{K} p(z_k = 1)p(x|z_k = 1)}$$
$$= \frac{\lambda_k N(x|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \lambda_k N(x|\mu_k, \Sigma_k)} \tag{B8}$$

Here, λ_k is the prior probability of $z_k = 1$ and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed x. The posterior probability for each component in GMM from which the data set was generated is called the *responsibilities*. Responsibilities sum to 1. This helps us predict which Gaussian is responsible for which data point.

Since the latent variables are never observed, and the correct values are not known in advance Expectation Maximization is useful to figure out what z represents, without someone to specify it beforehand.

EM method aims to iteratively improve the results based on some initial assumptions on the mean, standard deviation, and latent values. Every single iteration is of two steps - the so E step and the M step.

In the *expectation* step, it uses current values for the parameters to evaluate the posterior probabilities or responsibilities, given by Eq. (B8). We then use these probabilities in the *maximization* step to re-estimate the means, covariances, and mixing coefficients

EM for Gaussian Mixture

1. Initialization of the parameters and evaluate the initial values for log likelihood. Parameters are: Means μ_k, covariances Σ_k and mixing coefficients λ_k

2. **E step :** Evaluation of the responsibilities using the current parameter values.

$$\gamma(z_{ik}) == \frac{\lambda_k N(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \lambda_k N(x_i|\mu_k, \Sigma_k)}$$

3. **M step:** Re-estimate the parameters using the current responsibilities

   . $\mu_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) x_i$

   . $\Sigma_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik})(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T$

   . $\lambda_k^{new} = \frac{N_k}{N}$

   where, $N_k = \sum_{k}^{N} \gamma(z_{ik})$

4. Evaluate the log likelihood

$$\ln p(X|\lambda, \mu, \Sigma) = \sum_{i=1}^{N} \ln\{\sum_{k=1}^{K} \lambda_k N(x_i|\mu_k, \Sigma_k)\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

Appendix C: Selecting the number of classes

The main free input parameter to the model training procedure is the number of mixture components K. Determining the most appropriate number of components automatically is a difficult problem that often contains a degree of subjectivity, requiring domain expertise. Here we use a combination of statistical guidance and expert judgment to select the number of classes.

For statistical guidance, we use BIC, which stands for the *Bayesian Information Criterion*. The BIC is an empirical approach of the model probability computed as:

$$BIC(K) = -2\ell(K) + N_f(K)log(n) \tag{C1}$$

where $\ell(K)$ is the log likelihood of the trained model with K classes, n is the number of profiles used in the BIC test. The log-likelihood function as below,

$$\begin{aligned}\ell = \ln p(X|\lambda,\mu,\Sigma) &= \sum_{i=1}^{N} \ln\{\sum_{k=1}^{K} \lambda_k N(x_i|\mu_k,\Sigma_k)\} \\ &= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} (\lambda_k p(x_i))\end{aligned} \tag{C2}$$

The log-likelihood of the data set, assuming independent observations, is:

$$\ell(\theta) = \sum_{i=1}^{N} logp(x_i;\theta), \tag{C3}$$

where it is explicit that the log-likelihood is a function of the set of parameters θ, and where $p(x_i;\theta)$ is the probability given in equation C2 for the data set instance x_i using the parameters θ. N_f is the number of the independent parameters to be estimated (the sum of the component weights, Gaussian means and covariance matrix elements in the d-dimensional data space, after PCA our new dimension is d):

$$N_f(k) = (K-1) + Kd + \frac{Kd(d-1)}{2} \tag{C4}$$

The BIC is empirical; the first r.h.s term in equation C1 decreases as the likelihood of the statistical model increases, while the second r.h.s term is a penalty term that increases with K and thus discourages over-fitting (Maze et al., 2017). The “ideal" value for K, in terms of this statistical metric, would be one that minimizes BIC, i.e. the likelihood of the model has been maximized without overfitting. One may also find that the BIC curve “plateaus", indicating that the model has reached maximum likelihood, i.e. further increases in the statistical complexity of the model no longer noticeably improve the likelihood. Empirical approaches like BIC are often used in statistics, especially when constraining the parameters is difficult or subjective. They can give us a rough estimate of what data collection might look like if we were able to survey the entire population (Maze et al., 2017).

Here, $\theta = \{\lambda, \mu, \Sigma\}$ is the set of parameters that minimize the misfit between the PDF of the data set that is going to be used for calculation and the PDF of the original data set. To train a GMM, i.e., to maximize $\ell(\theta)$ with regard to θ so that our BIC can be lowest, we need a data set x and a given number of components K (Maze et al., 2017).

Code and data availability. Data from UKESM1 is part of the CMIP6 data suite, which is freely available from a number of sources. For this study, we used Pangeo (https://pangeo.io/) for rapid data access and averaging. We used "preprocessing" script from Julian Busecke (https://github.com/jbusecke/cmip6_preprocessing). All scripts used to data process and produce figures for this paper are online via Zenodo (Fahrin and Jones, 2022)

Author contributions. DCJ designed the initial project and developed much of the software. FF performed the analysis, worked with the software, and created the figures. JK and ATA provided expert guidance on analysing the results and placing them in the wider context of atmospheric chemistry. FF and DCJ wrote the initial manuscript, JK edited the introduction and all authors assisted with edits.

Competing interests. The contact author and the co-authors do not have any competing interest.

Acknowledgements. This work originated as a master's project in the Department of Mathematical Sciences at Georgia Southern University. The authors wish to thank Guillaume Maze for suggesting the particular training dataset method used here. We acknowledge the use of the Pangeo platform in obtaining our data (https://pangeo.io/). DJ acknowledges funding from a UKRI Future Leaders Fellowship (reference MR/T020822/1) and the North Atlantic Climate System Integrated Study (ACSIS) (grant NE/N018028/1). JK and ATA thank the Met Office CSSP-China programme for funding the POzSUM project.

References

Abernathey, R. P., Augspurger, T., Banihirwe, A., Blackmon-Luca, C. C., Crone, T. J., Gentemann, C. L., Hamman, J. J., Henderson, N., Lepore, C., McCaie, T. A., Robinson, N. H., and Signell, R. P.: Cloud-Native Repositories for Big Scientific Data, Computing in Science Engineering, 23, 26–35, https://doi.org/10.1109/MCSE.2021.3059437, 2021.

Archibald, A., Neu, J., Elshorbany, Y., Cooper, O., Young, P., Akiyoshi, H., Cox, R., Coyle, M., Derwent, R., Deushi, M., et al.: Tropospheric Ozone Assessment ReportA critical review of changes in the tropospheric ozone burden and budget from 1850 to 2100, Elementa: Science of the Anthropocene, 8, 2020a.

Archibald, A. T., Turnock, S. T., Griffiths, P. T., Cox, T., Derwent, R. G., Knote, C., and Shin, M.: On the changes in surface ozone over the twenty-first century: sensitivity to changes in surface temperature and chemical mechanisms, Philosophical Transactions of the Royal Society A, 378, 20190 329, 2020b.

Bates, D. R. and Nicolet, M.: The photochemistry of atmospheric water vapor, Journal of Geophysical Research, 55, 301–327, 1950.

Bishop, C. M.: Pattern recognition, Machine learning, 128, 2006.

Boehme, L. and Rosso, I.: Classifying Oceanographic Structures in the Amundsen Sea, Antarctica, Geophysical Research Letters, 48, e2020GL089 412, https://doi.org/https://doi.org/10.1029/2020GL089412, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL089412, e2020GL089412 2020GL089412, 2021.

Boleti, E., Hueglin, C., Grange, S. K., Prévôt, A. S., and Takahama, S.: Temporal and spatial analysis of ozone concentrations in Europe based on timescale decomposition and a multi-clustering approach, Atmospheric Chemistry and Physics, 20, 9051–9066, 2020.

Butchart, N.: The Brewer-Dobson circulation, Reviews of geophysics, 52, 157–184, 2014.

Chameides, W., Fehsenfeld, F., Rodgers, M., Cardelino, C., Martinez, J., Parrish, D., Lonneman, W., Lawson, D., Rasmussen, R., Zimmerman, P., et al.: Ozone precursor relationships in the ambient atmosphere, Journal of Geophysical Research: Atmospheres, 97, 6037–6055, 1992.

Chang, K.-L., Petropavlovskikh, I., Cooper, O. R., Schultz, M. G., Wang, T., Helmig, D., and Lewis, A.: Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia, Elementa: Science of the Anthropocene, 5, 2017.

Chapman, S.: XXXV. On ozone and atomic oxygen in the upper atmosphere, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 10, 369–383, 1930.

Cicerone, R. J., Stolarski, R. S., and Walters, S.: Stratospheric ozone destruction by man-made chlorofluoromethanes, Science, 185, 1165–1167, 1974.

Crutzen, P. J.: The influence of nitrogen oxides on the atmospheric ozone content, Quarterly Journal of the Royal Meteorological Society, 96, 320–325, 1970.

Desbruyères, D., Chafik, L., and Maze, G.: A shift in the ocean circulation has warmed the subpolar North Atlantic Ocean since 2016, Communications Earth & Environment, 2, https://doi.org/10.1038/s43247-021-00120-y, 2021.

Fahrin, F. and Jones, D.: UKESM1_Ozone_clustering, https://doi.org/10.5281/zenodo.6837484, https://doi.org/10.5281/zenodo.6837484, 2022.

Griffiths, P. T., Murray, L. T., Zeng, G., Shin, Y. M., Abraham, N. L., Archibald, A. T., Deushi, M., Emmons, L. K., Galbally, I. E., Hassler, B., et al.: Tropospheric ozone in CMIP6 simulations, Atmospheric Chemistry and Physics, 21, 4187–4218, 2021.

Houghton, I. A. and Wilson, J. D.: El Niño Detection Via Unsupervised Clustering of Argo Temperature Profiles, Journal of Geophysical Research: Oceans, 125, e2019JC015 947, https://doi.org/https://doi.org/10.1029/2019JC015947, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015947, e2019JC015947 10.1029/2019JC015947, 2020.

Jaadi, Z.: A step by step explanation of Principal Component Analysis, Towards Data Science, pp. 1–9, 2019.

Jaffe, D. A. and Wigder, N. L.: Ozone production from wildfires: A critical review, Atmospheric Environment, 51, 1–10, 2012.

Johnston, H.: Reduction of stratospheric ozone by nitrogen oxide catalysts from supersonic transport exhaust, Science, 173, 517–522, 1971.

Jones, D. C., Holt, H. J., Meijers, A. J., and Shuckburgh, E.: Unsupervised clustering of Southern Ocean Argo float temperature profiles, Journal of Geophysical Research: Oceans, 124, 390–402, 2019.

Keeble, J., Hassler, B., Banerjee, A., Checa-Garcia, R., Chiodo, G., Davis, S., Eyring, V., Griffiths, P. T., Morgenstern, O., Nowack, P., Zeng, G., Zhang, J., Bodeker, G., Burrows, S., Cameron-Smith, P., Cugnet, D., Danek, C., Deushi, M., Horowitz, L. W., Kubin, A., Li, L., Lohmann, G., Michou, M., Mills, M. J., Nabat, P., Olivie, D., Park, S., Seland, O., Stoll, J., Wieners, K.-H., and Wu, T.: Evaluating stratospheric ozone and water vapour changes in CMIP6 models from 1850 to 2100, Atmospheric Chemistry and Physics, 21, 5015–5061, https://doi.org/10.5194/acp-21-5015-2021, https://acp.copernicus.org/articles/21/5015/2021/, 2021.

Laban, T. L., Zyl, P. G. v., Beukes, J. P., Vakkari, V., Jaars, K., Borduas-Dedekind, N., Josipovic, M., Thompson, A. M., Kulmala, M., and Laakso, L.: Seasonal influences on surface ozone variability in continental South Africa and implications for air quality, Atmospheric chemistry and physics, 18, 15 491–15 514, 2018.

Lee, S. H.: The stratospheric polar vortex and sudden stratospheric warmings, Weather, 76, 12–13, 2021.

Li, Y. and Thompson, D. W.: The signature of the stratospheric Brewer–Dobson circulation in tropospheric clouds, Journal of Geophysical Research: Atmospheres, 118, 3486–3494, 2013.

Lu, X., Zhang, L., Zhao, Y., Jacob, D. J., Hu, Y., Hu, L., Gao, M., Liu, X., Petropavlovskikh, I., McClure-Begley, A., et al.: Surface and tropospheric ozone trends in the Southern Hemisphere since 1990: possible linkages to poleward expansion of the Hadley circulation, Science Bulletin, 64, 400–409, 2019.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Radcenco, M. L., Lenca, P., Feucher, C., and Le Goff, C.: Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean, Progress in Oceanography, 151, 275–292, 2017.

McLachlan, G. J. and Basford, K. E.: Mixture models: Inference and applications to clustering, vol. 38, M. Dekker New York, 1988.

Molina, M. J. and Rowland, F. S.: Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone, Nature, 249, 810–812, 1974.

Monks, P. S., Granier, C., Fuzzi, S., Stohl, A., Williams, M. L., Akimoto, H., Amann, M., Baklanov, A., Baltensperger, U., Bey, I., et al.: Atmospheric composition change–global and regional air quality, Atmospheric environment, 43, 5268–5350, 2009.

Monks, P. S., Archibald, A., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G., et al.: Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer, Atmospheric Chemistry and Physics, 15, 8889–8973, 2015.

Newman, P. and Todara, R.: Stratospheric Ozone; An Electronic Textbook, Studying Earths Environment From Space. NASA, 480, 2003.

Oehrlein, J., Chiodo, G., and Polvani, L. M.: The effect of interactive ozone chemistry on weak and strong stratospheric polar vortex events, Atmospheric Chemistry and Physics, 20, 10 531–10 544, 2020.

Palmeiro, F. M., Calvo, N., and Garcia, R. R.: Future changes in the Brewer–Dobson circulation under different greenhouse gas concentrations in WACCM4, Journal of the atmospheric sciences, 71, 2962–2975, 2014.

Rosso, I., Mazloff, M. R., Talley, L. D., Purkey, S. G., Freeman, N. M., and Maze, G.: Water Mass and Biogeochemical Variability in the Kerguelen Sector of the Southern Ocean: A Machine Learning Approach for a Mixing Hot Spot, Journal of Geophysical Research: Oceans, 125, e2019JC015 877, https://doi.org/https://doi.org/10.1029/2019JC015877, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015877, e2019JC015877 10.1029/2019JC015877, 2020.

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., Mora, L. d., Kuhlbrodt, T., Rumbold, S. T., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G., and Zerroukat, M.: UKESM1: Description and Evaluation of the U.K. Earth System Model, Journal of Advances in Modeling Earth Systems, 11, 4513–4558, https://doi.org/10.1029/2019MS001739, 2019.

Sonnewald, M., Wunsch, C., and Heimbach, P.: Unsupervised learning reveals geography of global ocean dynamical regions, Earth and Space Science, 6, 784–794, 2019.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G.: Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces, Science Advances, 6, 1–12, https://doi.org/10.1126/sciadv.aay4740, 2020.

Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., and Balaji, V.: Bridging observations, theory and numerical simulation of the ocean using machine learning, Environmental Research Letters, 16, 073 008, https://doi.org/10.1088/1748-9326/ac0eb0, https://iopscience.iop.org/article/10.1088/1748-9326/ac0eb0, 2021.

Wargan, K., Weir, B., Manney, G. L., Cohn, S. E., and Livesey, N. J.: The anomalous 2019 Antarctic ozone hole in the GEOS Constituent Data Assimilation System with MLS observations, Journal of Geophysical Research: Atmospheres, 125, e2020JD033 335, 2020.

Waugh, D. W., Sobel, A. H., and Polvani, L. M.: What is the polar vortex and how does it influence weather?, Bulletin of the American Meteorological Society, 98, 37–44, 2017.

Weber, M., Dikty, S., Burrows, J. P., Garny, H., Dameris, M., Kubin, A., Abalichin, J., and Langematz, U.: The Brewer-Dobson circulation and total ozone from seasonal to decadal time scales, Atmospheric Chemistry and Physics, 11, 11 221–11 235, 2011.

Young, P., Archibald, A., Bowman, K., Lamarque, J.-F., Naik, V., Stevenson, D., Tilmes, S., Voulgarakis, A., Wild, O., Bergmann, D., et al.: Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmospheric Chemistry and Physics, 13, 2063–2090, 2013.