Air Quality, Atmosphere & Health (2019) 12:743–754
https://doi.org/10.1007/s11869-019-00694-9

Spatial patterns of conditions leading to peak O_3 concentrations revealed by clustering analysis of modeled data

Andrea L. Pineda Rojas[1] · Julie A. Leloup[2] · Emilio Kropff[3]

Received: 11 January 2019 / Accepted: 5 April 2019 / Published online: 4 May 2019

Abstract

Air quality models are currently the best available tool to estimate ozone (O_3) concentrations in the Metropolitan Area of Buenos Aires (MABA). While the DAUMOD-GRS has been satisfactorily evaluated against observations in the urban area, a Monte Carlo (MC) analysis showed that it is the region around the MABA, where the lack of observations impedes model testing, that concentrates not only the greatest estimated O_3 peak levels but also the largest model uncertainty. In this work, we apply clustering analysis to these MC outcomes in order to study the spatial patterns of conditions leading to peak ozone hourly concentrations. Results show that families of conditions distribute, as emissions, radially around the city. A cluster exhibiting an O_3 morning peak dominates in low-emission areas, a behavior that can be explained both from theory and from the few monitoring campaigns carried out in the city. Its distinct dynamics compared with the typical O_3 diurnal profile occurring in the urban area suggests the need of new ozone measurements in the surroundings of the MABA which could contribute to improve our understanding of O_3 formation drivers in this region. The results illustrate the potential of applying clustering analysis on large ensembles of modeled data to better understand the variability in model solutions.

Keywords Air quality modeling · Buenos Aires · Clustering analysis · Monte Carlo simulations · Ozone

Introduction

The Metropolitan Area of Buenos Aires (MABA) is the third megacity in Latin America. In spite of regulations, the ozone (O_3) concentration in this region has not been measured regularly until 2015, and since then at only one air quality monitoring station. Therefore, air quality models are currently the only available tool to provide estimates of O_3 distribution across the MABA. The reliability of model results is assessed through model performance evaluations, including several steps among which the probabilistic evaluation plays an important role (Chang and Hanna 2005; Derwent et al. 2010). It assesses the uncertainty in model results that is caused by the uncertainties in model formulations, input variables or parameters, resolution, etc. Among all possible sources of error, uncertainties in the model input variables typically dominate the uncertainty in modeled pollutant concentrations (Russell and Dennis 2000). A widely used methodology to assess the uncertainty of modeled pollutant concentrations caused by possible errors in the input variables is the Monte Carlo (MC) analysis (e.g., Hanna et al. 1998; Bergin et al. 1999; Moore and Londergan 2001; Hanna et al. 2005, 2007; Rodriguez et al. 2007; Tang et al. 2010; Tan et al. 2014; Pineda Rojas et al. 2016). While this technique can be applied to any air quality model, high computational demand is among its main limitations. Typically, $N = 100$ is considered an acceptable number of MC runs to assess the (gridded) ensemble of modeled concentration solutions from which uncertainty is computed. Its analysis usually limits to the evaluation of the sensitivity of the output to uncertainties in the model input variables at a few receptors, providing quantitative

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11869-019-00694-9) contains supplementary material, which is available to authorized users.

✉ Andrea L. Pineda Rojas
pineda@cima.fcen.uba.ar

[1] Centro de Investigaciones del Mar y la Atmósfera, UMI-IFAECI/CNRS, Facultad de Ciencias Exactas y Naturales, CONICET, UBA, Universidad de Buenos Aires, Ciudad Universitaria, Pab. II, piso 2, 1428 Buenos Aires, Argentina

[2] LOCEAN/IPSL, UMR 7159, CNRS-IRD-MNHN, Sorbonne Universités, UPMC University Paris 6, Paris, France

[3] Fundación Instituto Leloir - IIBBA/CONICET, Buenos Aires, Argentina

measures (e.g., sensitivity coefficients) that show which are the variables that dominate the model uncertainty at those receptors. However, further analysis of gridded MC outcomes (i.e., the model output and associated input data) may provide some insight on the type of solutions that can be obtained with the model. The main limitation, the size and complexity of the dataset, can be tackled utilizing techniques from the field of big data.

Clustering analysis aims for an unbiased classification of big datasets into groups containing objects with similar characteristics. In air quality studies, it has been widely used to identify monitoring stations with similar pollutant concentrations (e.g., Flemming et al. 2005; Afif et al. 2009; Henne et al. 2010), classify monitoring sites based on their chemical composition (e.g., Austin et al. 2013; Wang et al. 2016; Park et al. 2018), study the impact of remote emission sources on urban levels of particulate matter (PM) concentrations (e.g., Borge et al. 2007; Karaca and Camci 2010; Dimitriou and Kassomenos 2014; Terrouche et al. 2016), and identify meteorological patterns associated to pollution episodes of O_3 (e.g., Beaver and Palazoglu 2006; Pakalapati et al. 2009; Khedairia and Khadir 2012; Awang et al. 2016) and PM (Rimetz-Planchon et al. 2008; Unal et al. 2011). Only a few works have applied clustering analysis to study gridded modeled pollutant concentrations (e.g., Jin et al. 2011). Despite of its wide application in air quality studies, it has not been used in combination with MC simulations to perform a systematic qualitative screening of the outcomes of air quality models.

DAUMOD-GRS (MODelo de Dispersión Atmosférica Urbano-Generic Reaction Set) is a simple urban-scale atmospheric dispersion model that allows the estimation of ground-level O_3 concentrations resulting from area source emissions of nitrogen oxides and volatile organic compounds, transport by the wind, atmospheric dispersion, and simplified photochemistry (Pineda Rojas and Venegas 2013). The statistical evaluation of the model has shown an acceptable performance to simulate O_3 hourly concentrations at 20 receptors within the MABA (Pineda Rojas 2014). In Pineda Rojas et al. (2016), the uncertainty of the summer maximum O_3 hourly concentration (C_{max}) was evaluated at each receptor of the MABA domain applying the MC analysis. Results from that work showed that the greatest uncertainties of C_{max} (up to 47 ppb) are obtained outside of the MABA, where the greatest values of C_{max} are estimated (up to 51 ppb) and the lack of observations impedes model testing. Given the amount of information obtained from the MC simulations, in this work, we apply clustering techniques to characterize the conditions leading to the occurrence of modeled C_{max} values. The objective is to further explore those gridded MC outcomes in order to better understand the type of model solutions that can be obtained with the DAUMOD-GRS throughout the whole MABA area.

Methodology

Description of the Monte Carlo outcomes used for clustering

The MC simulations are runs of a model fed with N different input datasets, obtained by perturbing the model variables randomly based on their error distributions and ranges. The base case input datasets (i.e., without perturbations) consist of surface hourly meteorological information registered at the domestic airport of Buenos Aires city (AEP: 34° 34′ S, 58° 30′ W) during a typical summer (2007), sounding meteorological data from the international airport (EZE: 34° 49′ S, 58° 30′ W), and area source emission rates of nitrogen oxides (NO_x) and volatile organic compounds (VOCs) from the emission inventory developed for the MABA by Venegas et al. (2011). A constant regional background concentration of 20 ppb is assumed for ozone based on a previous study (Mazzeo et al. 2005), and "clean air" concentration values are assumed for NO_x and VOCs given that the MABA is surrounded by non-urban areas.

The MC outcomes used in this work were obtained previously (Pineda Rojas et al. 2016) by perturbing nine input variables that feed the DAUMOD-GRS model: wind speed (WS) and direction (DIR), air temperature (T), sky cover (SC), total solar radiation (TSR), atmospheric stability class (KST), NO_x emission rate (QNO_x), VOC emission rate (QVOC), and regional background O_3 concentration ($[O_3]_r$). Due to the lack of information, the probability density functions and error ranges of these variables were taken from the literature (see Table S.1). Simple random sampling was used to obtain $N = 100$ sets of perturbations from these uncertainty distributions, with which the "base case" data described above were perturbed. In this way, 100 perturbed input datasets were generated to perform the MC runs. All simulations considered a temporal resolution of 1 h and a spatial resolution of 1 km × 1 km. At each hour, spatially constant meteorological conditions were assumed, and only the emissions were allowed to vary spatially. On the other hand, perturbations of all variables were considered constant both spatially and with time (see Pineda Rojas et al. (2016) for details).

The results obtained from these MC simulations include the value of C_{max} estimated during diurnal hours (7–19 h) at each square kilometer of the MABA domain (4647 receptors) and the values of the nine perturbed input variables at the moment of occurrence of C_{max}. It is worth noting that, at different receptors, C_{max} occurs at different times of the summer which results in a wide range of leading conditions in spite of the assumption of horizontally homogeneous meteorological variables (Pineda Rojas et al. 2016). This previous work also shows that the hour of occurrence of C_{max} (H) can vary considerably. For this reason, H was also considered a relevant variable for the clustering analysis. Hence, a total of 4,647,000 data (i.e., 10 variables ×

4647 receptors × 100 possible solutions) were obtained from that model uncertainty assessment. This size clearly limits the direct observation of the data.

Clustering analysis

Clustering analysis aims to find groups of "objects" within a large dataset based on their similarity. The *k*-means algorithm is a widely used clustering method for air quality studies (e.g., Davies et al. 1998; Beaver and Palazoglu 2006; Lu et al. 2006; Pakalapati et al. 2009; Jin et al. 2011; Khedairia and Khadir 2012; Austin et al. 2013; Gomez-Losada et al. 2018). It is a heuristic algorithm aiming to place k cluster centers (k, user defined) in a M-dimensional space (M, number of variables describing the objects) so as to minimize the mean distance from objects to their closest cluster center. The k centers are first distributed randomly, following which two steps are iterated until a steady solution is reached: (i) each object is assigned to the nearest cluster center and (ii) each cluster center is reset to the geometrical mean among all objects belonging to it.

Implementation of the *k*-means method

The definition of the objects and the form of standardization depend on the purpose of the clustering implementation. In this case, we aim to determine, for example, whether or not spatial patterns can be observed in the conditions of occurrence of C_{max} modeled with the DAUMOD-GRS. Hence, an object is considered the set of conditions (x_i, $i = 1, ..., M$) in which an individual C_{max}

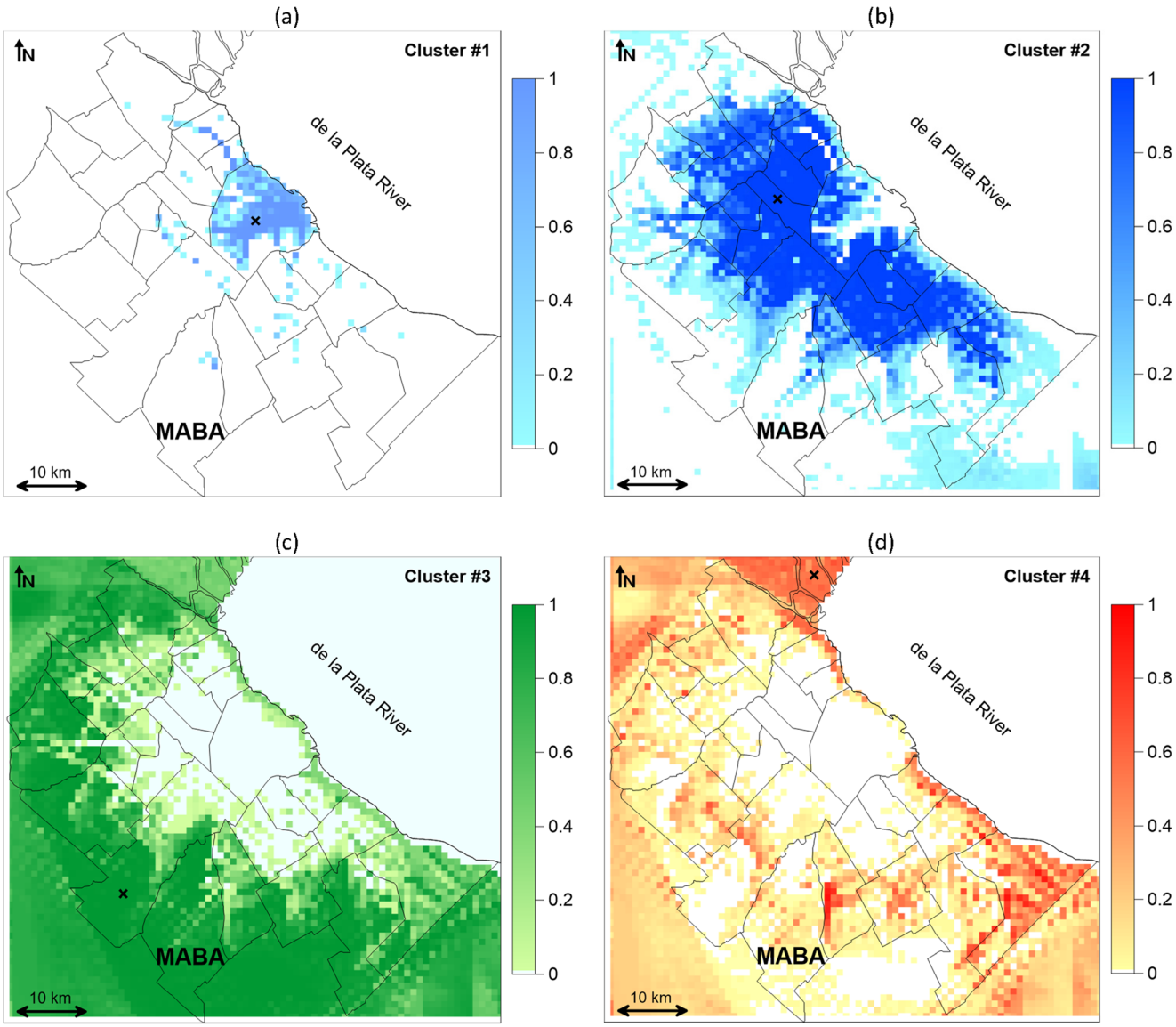

Fig. 1 Spatial distribution the frequency of occurrence of each cluster in the MC simulations (✖ indicates receptors selected for a more detailed analysis)

occurs, and the outcomes at all the receptors in the modeling domain are pooled together to define the object set (i.e., 100 MC repetitions at 4647 receptors = 464,700 objects). Given that the conditions of occurrence of C_{max} through the MC simulations are given by the nine perturbed input variables, all these variables together with the hour of occurrence of C_{max} are chosen to define the M-dimensional space (i.e., $M = 10$). Since variables are not comparable, each one is scaled subtracting its mean ($\overline{x_i}$) and dividing by its standard deviation (σ_{x_i}) across all objects in the dataset:

$$x_i^{'} = \left(x_i - \overline{x_i}\right)/\sigma_{x_i} \quad (1)$$

The wind variables (polar coordinates) are decomposed into their x and y components, which define a proper Euclidean space (otherwise, 0° and 360° would be far apart). The MATLAB function *kmeans* is used with $n = 100$ random initializations (to avoid suboptimal local solutions), and different values of k are considered. For a given value of k, among the n clustering solutions, the one with the lowest within-cluster sum of point-to-centroid distances is selected:

$$S = \sum_{j \in c} d\left(\overrightarrow{x_j},\ \overrightarrow{x_c}\right)^2 \quad (2)$$

where $\overrightarrow{x_j}$ is the position of point j in the normalized variables space $[\overrightarrow{x_j} = \left(x_1^{'}, x_2^{'}, \ldots, x_M^{'}\right)]$, $\overrightarrow{x_c}$ is the centroid position of cluster c, and the sum is performed over all elements j in cluster c and over all clusters $c = 1, \ldots, k$.

Choice of the best cluster distribution

There is no universal agreement regarding the optimal number of clusters for a given dataset. In this paper the silhouette criterion (Rouseeuw 1987) is applied. It compares, for different values of k, the average over all objects of:

$$S_j = \left(b_j - a_j\right)/\max\left(a_j, b_j\right) \quad (3)$$

where a_j is the mean distance from object j to all other objects in the same cluster and b_j is its mean distance to objects in other clusters. Silhouette maxima are typically used to determine k since they offer better cluster definition than its local neighbors.

In an initial exploration with downsampled data (100 times), k-means solutions were obtained for k ranging from 1 to 10. The MATLAB function *silhouette* was used to compare them, exhibiting two local maxima at $k = 4$ (0.51) and $k = 6$ (0.54). Both sets of solutions were analyzed for the complete dataset and qualitatively similar conclusions were extracted. In the remaining part of this paper, results for $k = 4$ are reported for the sake of simplicity in description and visualization.

Results

Spatial patterns

Our first observation is that when plotted on the map of the MABA, clusters present characteristic spatial patterns, both when the frequency of occurrence of each cluster (Fig. 1) and when the dominant cluster in each location (Fig. 2) are considered. Except for cluster 4 that dominates in emission transition areas, all other clusters exhibit a radial pattern resembling those of the NO_x and VOC emissions (see Pineda Rojas 2014). Clusters 1 and 2 appear mostly at receptors with emissions: in

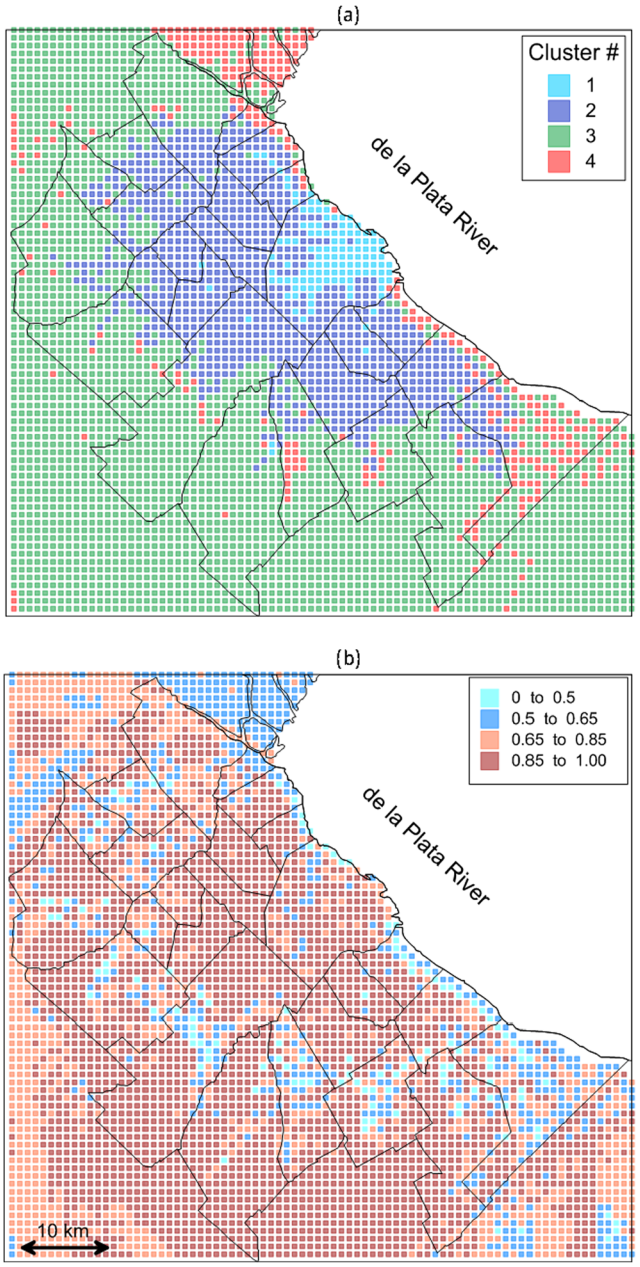

Fig. 2 Spatial distribution of **a** the dominant cluster at each receptor and **b** its frequency of occurrence in the MC simulations

the city of Buenos Aires (highest emission rates) and in the greater Buenos Aires (moderate emissions), respectively. Clusters 3 and 4 are mostly present in the suburbs and outside of the MABA where no emissions are considered and the highest C_{max} values are obtained (Pineda Rojas et al. 2016). In 57% of receptors, the frequency of occurrence of the dominant cluster at each receptor is ≥ 0.85; and in 81% of them, it is ≥ 0.65 (see Fig. 2b). (Note that only in 3% of the receptors, the frequency of the dominant cluster is < 0.5.) This means that in most of the analyzed domain, the family of leading conditions of C_{max} is well defined, while only in a small portion, two or more clusters can dominate depending on the specific MC run.

Cluster spatial patterns highlight the predominant role played by emissions in determining the conditions of occurrence of C_{max}.

Multivariate cluster structure

Are emissions the only variables determining the cluster structure (or separation) or do they interact with other variables? One way to look at the relative contributions of different variables to structure is to study the standard deviation of normalized variables x' (Fig. 3b). Its value for each cluster can be compared with the value for the whole dataset (which is equal

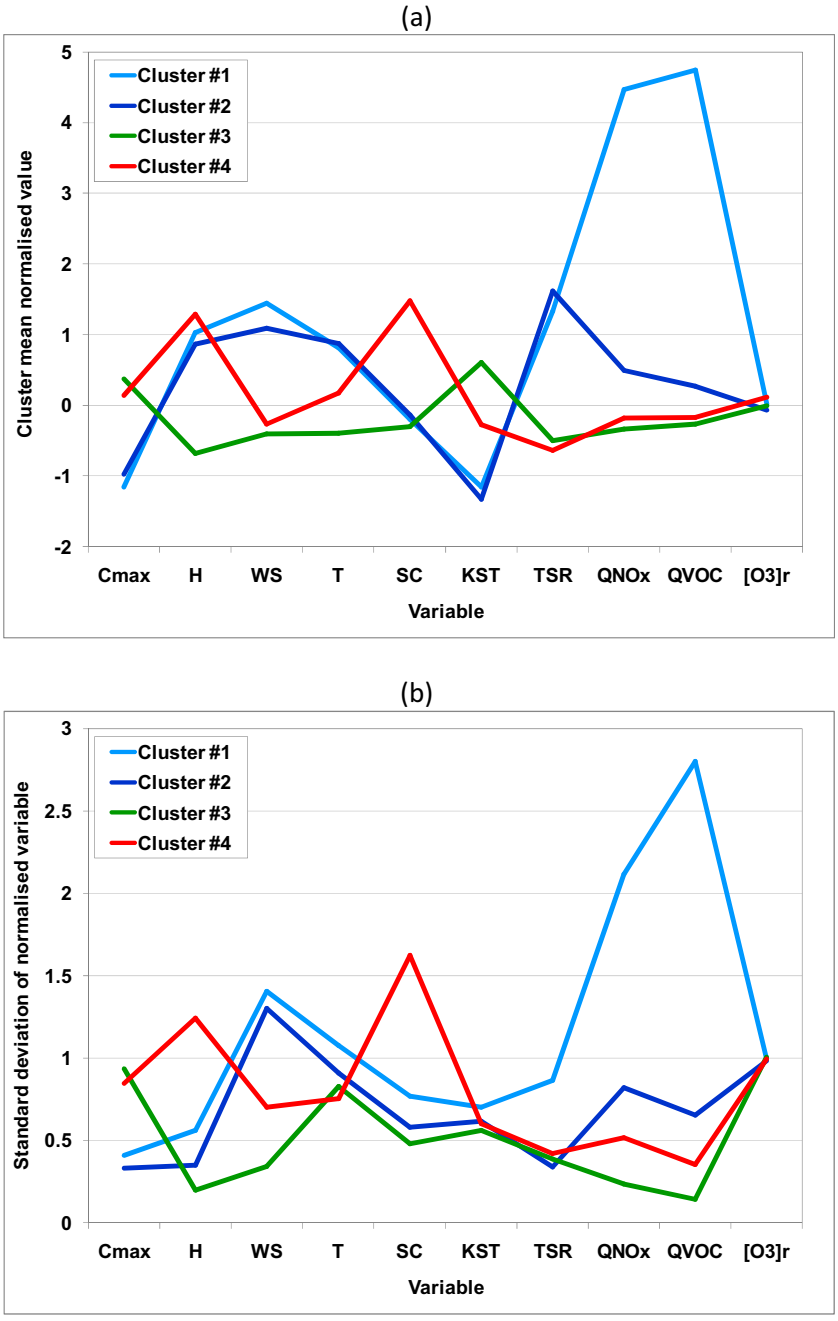

Fig. 3 **a** Mean values and **b** standard deviation of normalized variables (z-score) for each cluster (C_{max}, summer maximum O_3 concentration; H, hour of occurrence of C_{max}; WS, wind speed; T, air temperature; SC, sky cover; KST, atmospheric stability class; TSR, total solar radiation; QNO_x, NO_x emission rate; QVOC, VOC emission rate; $[O_3]_r$, regional background O_3 concentration)

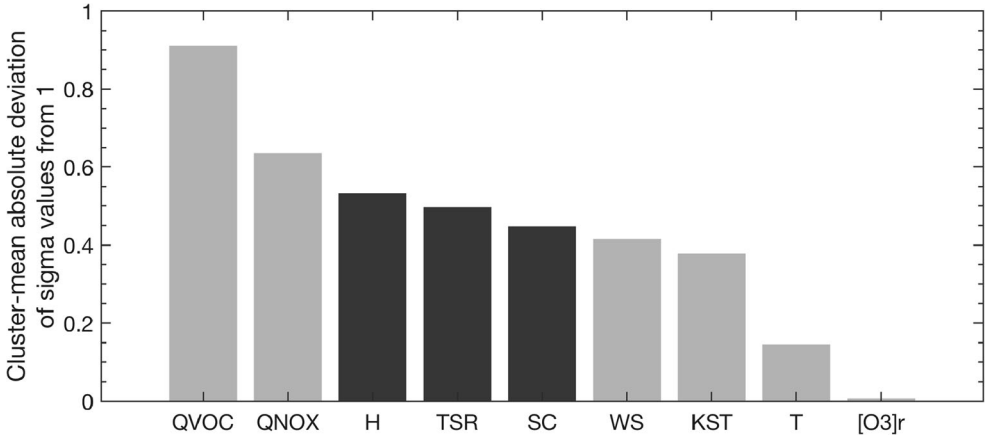

Fig. 4 Sorted absolute deviation of sigma values of normalized variables (Fig. 3b) from 1, averaged over the four clusters

to 1 due to the normalization given by Eq. (1)). A normalized variable that has a standard deviation close to 1 indicates that its spread within the cluster is similar to that of the population (for example, $[O_3]_r$ for all clusters). In contrast, a standard deviation close to 0 shows that the cluster specializes in a small range of values of the variable (for example, H in cluster 3), while a standard deviation greater than 1 is indicative of a complex structure (for example, SC in cluster 4). To understand which variables contribute to define a clustered data structure beyond the spatial distribution determined by emissions, their absolute deviations from 1 are averaged across all clusters (Fig. 4). The first three variables in the rank (putting aside emissions) are considered for visualization: the hour of occurrence of C_{max} (H), the total solar radiation (TSR), and the sky cover (SC). Contour curves surrounding the regions containing 99.5% of data within each cluster are plotted in the planes SC-H (Fig. 5a) and TSR-H (Fig. 5b). Under low SC conditions, cluster 3 appears in the morning hours with low TSR values while clusters 1 and 2 appear at midday with high TSR. Cluster 4 instead seems to have a more complex distribution, appearing mostly in early hours with high SC values and in late hours at low SC values. In midday hours, cluster 4 also appears with high SC and low TSR values. These projections help to understand the separation between clusters and the more complex scatter plot spanning all three variables (Fig. 5c, d). In this three-dimensional space, clusters distribute forming arcs that extend across the H dimension in successive TSR-H planes. When SC is low, the arc is highest in TSR, containing clusters 3 in the morning, 1 and 2 at noon, and part of 4 in late hours. As SC increases, these arcs become lower in TSR and are mostly formed by C_{max} belonging to cluster 4.

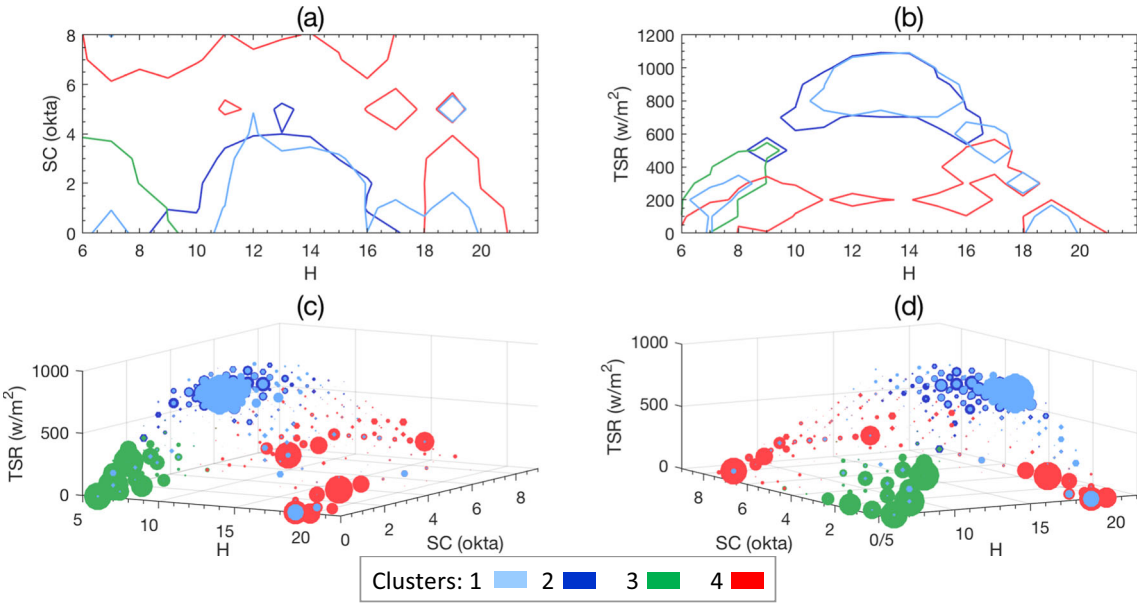

Fig. 5 Contour curves surrounding the regions containing 99.5% of data within each cluster in the SC-H (**a**) and TSR-H (**b**) planes and distribution of objects of each cluster in the H-SC-TSR space from two perspectives (**c** and **d**)

In summary, these results show that the structure of clusters spans multiple variables and that at different spatial locations, distinctive and complex sets of conditions lead to C_{max} modeled with the DAUMOD-GRS model.

Characterization of the clusters

In order to provide a full description of the clusters, the mean values and ranges of all variables are analyzed (except for the wind direction that is discussed apart in "Wind direction"). Table 1 presents the mean variables of each cluster, and Fig. 6 shows the 95% confidence range of each variable vs. the cluster number. Comparing clusters 1 and 2 (which dominate at urban and suburban receptors, respectively), these occur on average at 14 h and 13 h, respectively, under conditions of clear sky (mean SC = 1), moderate wind speeds (WS = 6.0 m/s and 5.1 m/s, respectively), relatively high values of temperature (T ~ 27 °C) and total solar radiation (TSR = 762 and 855 W/m^2, respectively), and atmospheric instability (i.e., lower mean values of KST) (see Table 1). Cluster 1 presents a wider range of variation of H, SC, KST, and TSR than cluster 2, presumably associated to a wider range of QNO_x and QVOC (which are produced by both their variations within the city and their uncertainty ranges in the Monte Carlo simulations).

On the other hand, comparing clusters 3 and 4 (that dominate at receptors with no emissions, where the largest C_{max} values are estimated), these occur on average at 7 h and 15 h, respectively under low wind conditions (WS ≤ 1.7 m/s) (see Table 1). The difference in the mean time of occurrence of C_{max} between these two clusters is also reflected in the mean values of the meteorological variables that present marked diurnal cycles, as the atmospheric stability class. However, this is not observed in the mean total solar radiation that is greater for cluster 3 (at H = 7 h) than that for cluster 4 (at H = 15 h), due to the fact that in cluster 4, C_{max} occurs with a mean sky cover value of 4 (partly cloudy sky). Regarding their within-cluster variations, while cluster 4 presents a wide range of H (as well as in other variables like WS, SC, and TSR), cluster 3 occurs only during early-morning hours (see Fig. 6b). The possible reasons for the occurrence of such an ozone morning peak under conditions of clusters 3 are discussed in "Discussion."

Wind direction

The role of wind direction (DIR) is more difficult to analyze because, when considering the information from all receptors, it is not possible to distinguish situations of DIR that bring more or less polluted air to the receptors (a same DIR may have different effects on the pollutant concentration at different receptors depending on the emission sources and how they distribute around them). However, it is worth inspecting whether differences exist among the most frequent wind directions of the clusters. Figure 7 presents the wind rose of each cluster. The four clusters show variable and different dominant wind directions. In cluster 1 (associated to the lowest C_{max} values), winds leading to the occurrence of C_{max} are mainly of moderate intensities (4 m/s) from the ENE (22%) or relatively intense (~ 8 m/s) from the SE-SSE (27%). In cluster 2, C_{max} occurs with lower mean wind speeds (3 m/s) from the ENE sector (9%), moderate (5 m/s) from the W (11%), and intense (8 m/s) from the S (9%). In cluster 3, the dominant wind directions are also variable: E (14%), WSW (13%), and NNW (14%), but wind intensities are low (< 2 m/s), as previously noted. The same is observed for cluster 4 but with different dominant wind directions: N-NNE (29%) and E-ESE (29%). (As shown in Table 1, each cluster presents a different number of objects and then the same frequency of DIR for two different clusters gives different number of wind situations with that DIR.)

Given the spatial distribution of clusters 3 and 4 (see Fig. 1), their corresponding wind roses (Fig. 7c, d, respectively) suggest that C_{max} could be occurring with winds that come from outside the MABA (i.e., bringing no emissions to those receptors). This can be easily confirmed with a histogram of wind directions at the time of occurrence of C_{max} at any of the receptors where these clusters dominate. For example, the histograms of DIR (not shown) at four selected receptors indicated in Fig. 1 verify this. At the suburban receptor selected in Fig. 1c, where cluster 3 dominates, C_{max} occurs with winds from the W-NW (in 59% of the MC simulations), while at the one chosen for cluster 4 (Fig. 1d), the maximum ozone

Table 1 Variables from Fig. 3, averaged for each cluster (*N* number of objects of each cluster (%))

Cluster no.	*N* (%)	C_{max} (ppb)	H	WS (m/s)	T (°C)	SC (okta)	KST	TSR (W/m^2)	QNO_x (g/km^2 s)	QVOC (g/km^2 s)	$[O_3]_r$ (ppb)
1	12,396 (3)	18.5	14	6.0	26.8	1	2	762	6.3	5.1	20.4
2	103,036 (22)	20.2	13	5.1	27.0	1	2	855	1.2	0.6	20.1
3	280,765 (60)	32.9	7	1.3	22.4	1	5	167	0.1	0.0	20.3
4	68,503 (15)	30.8	15	1.7	24.5	4	4	120	0.3	0.1	20.6

concentration is mostly (91%) associated with winds from the E-SE. The occurrence of C_{max} with winds that come from outside the MABA is intriguing yet compatible with theory and observations as analyzed in the following section.

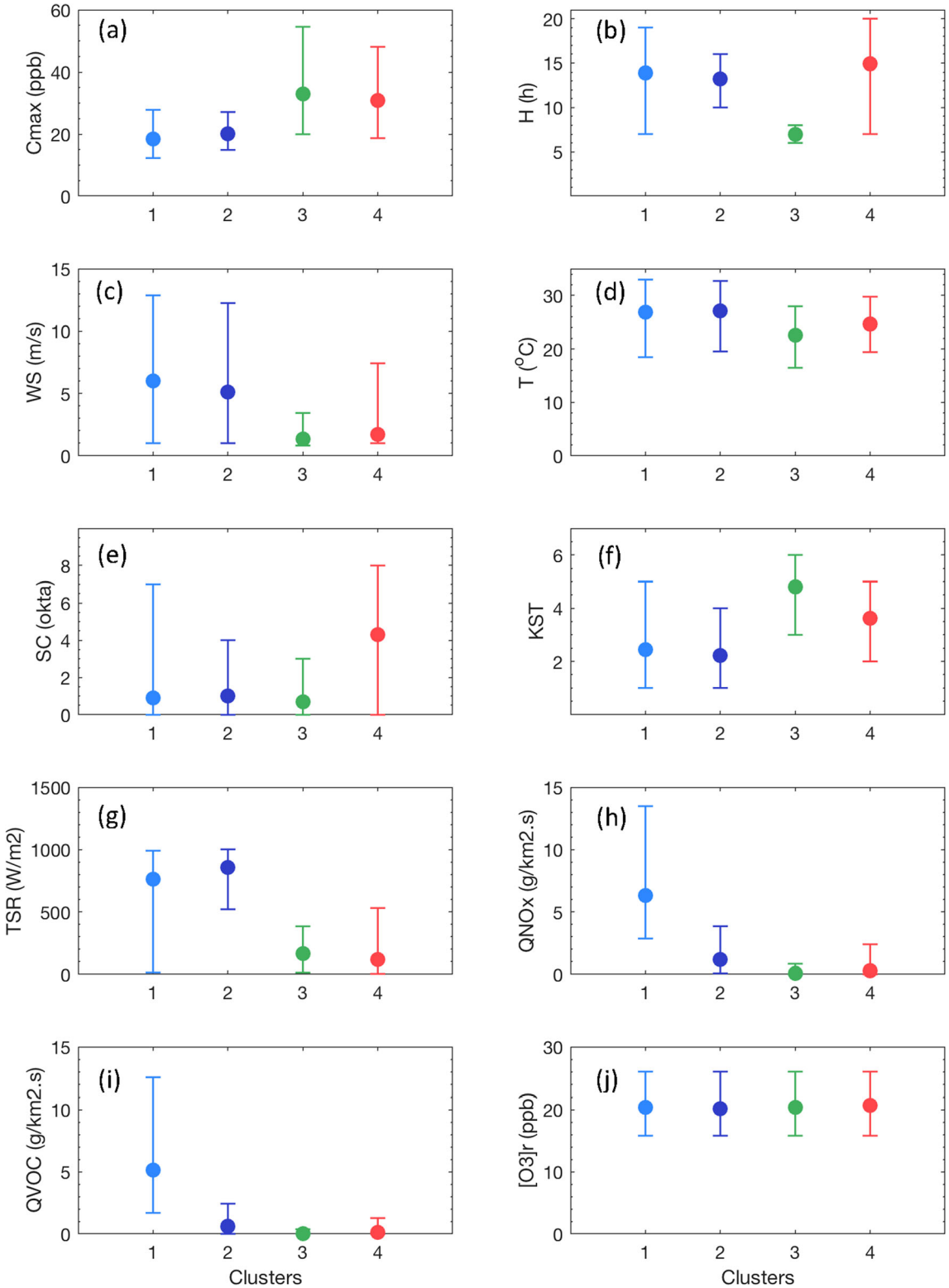

Fig. 6 Mean variables and their 95% confidence interval vs. cluster number: **a** summer maximum O_3 concentration (C_{max}), **b** hour of occurrence of C_{max} (H), **c** wind speed (WS), **d** air temperature (T), **e** sky cover (SC), **f** atmospheric stability class (KST), **g** total solar radiation (TSR), **h** NO_x emission rate (QNO_x), **i** VOC emission rate (QVOC), and **l** regional background O_3 concentration ($[O_3]_r$)

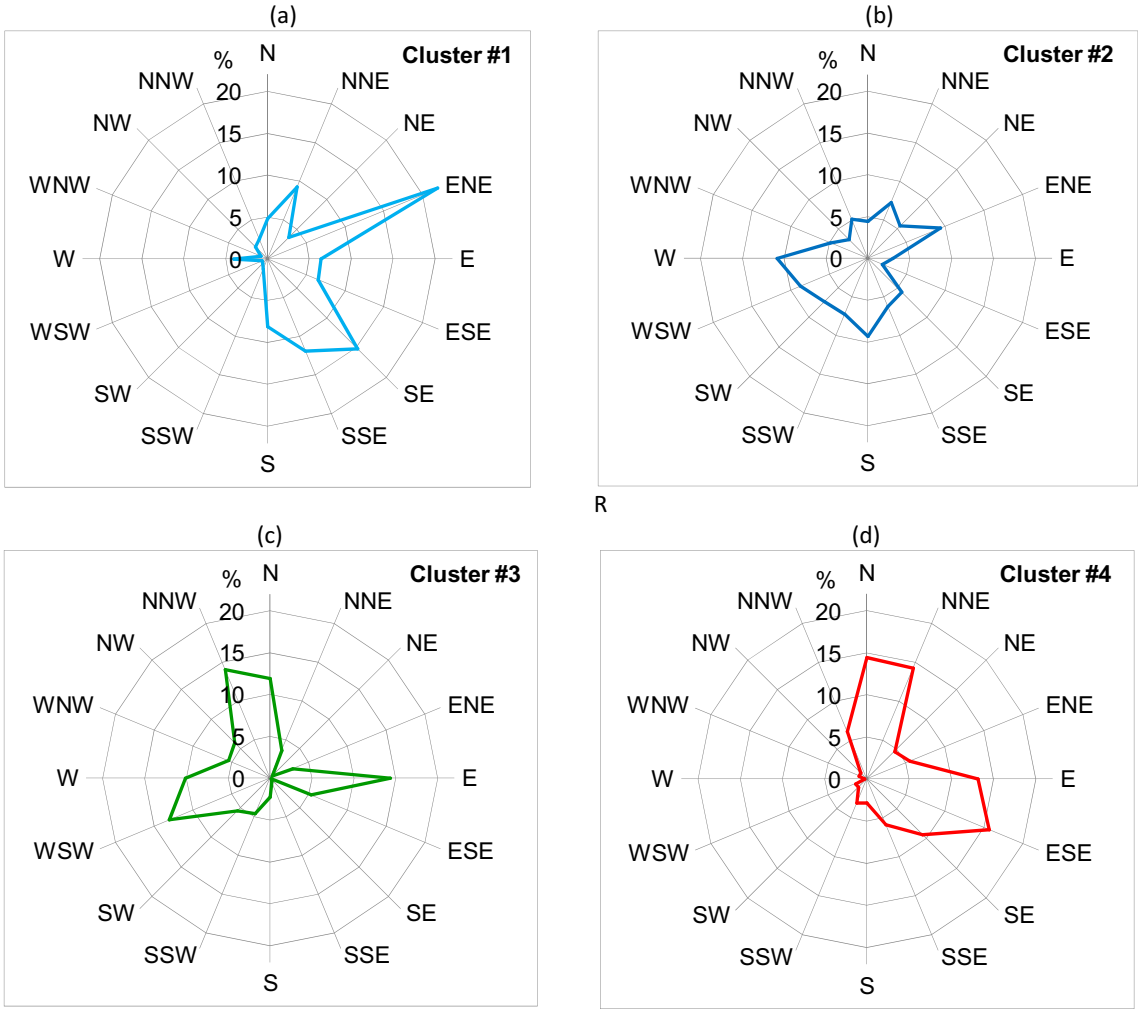

Fig. 7 Wind rose of each cluster

Discussion

The results obtained in this work show that at receptors of relatively high and moderate emission rates, the DAUMOD-GRS model gives relatively conventional results based on our knowledge of the typical O_3 diurnal profile (peaks occurring around midday hours due to photochemical conversion of NO_2 into O_3, enhanced by the presence of VOCs). However, at receptors with no emissions, cluster 3 shows that C_{max} can occur at early morning hours and/or with winds that come from outside the MABA, which is less expected. The clear spatial pattern of the obtained cluster distribution ("Spatial patterns") allows us to identify the region of the modeling domain where these unconventional results need to be explored. By choosing two representative examples of receptors with contrasting C_{max} leading conditions, an analysis of the potential causes of the O_3 morning peak is performed. Figure 8 shows the diurnal variations of the O_3 concentrations ($[O_3]$) and the initial concentrations of nitrogen monoxide ($[NO]_i$) and ozone ($[O_3]_i$), at an urban receptor (UR, Fig. 1a) where cluster 1 dominates and at a suburban receptor (SU, Fig. 1c) where cluster 3 is dominant. The reason to plot $[NO]_i$ and $[O_3]_i$ is that in the model, reaction $NO + O_3 \rightarrow NO_2$ (NO titration) is the only one occurring in the absence of solar radiation. For simplicity, the DAUMOD-GRS memory component was initially excluded from the analysis. A conventional hourly profile of O_3 concentration occurs at receptor UR (Fig. 8a). In this receptor, $[NO]_i$ is always greater than or equal to $[O_3]_i$, with a higher difference at early-morning and late-evening hours. At 6 h, $[O_3] \approx 0$ because the initial 20 ppb of ozone reacts with NO to generate NO_2 through the above reaction, and there is no solar radiation to form it photochemically. At the following hours, when solar radiation becomes important, the generated NO_2 is photolysed to form O_3, and $[O_3]$ increases reaching its maximum value (16.7 ppb) at 14 h. In turn, at receptor SU (Fig. 8b), at 6 h, $[NO]_i$ is close to zero due to a NW wind coming from outside the MABA. Consequently, the initial O_3 concentration cannot be consumed chemically and hence $[O_3]$ remains at around 20 ppb. At 7 h, the wind direction changes to NNW and some

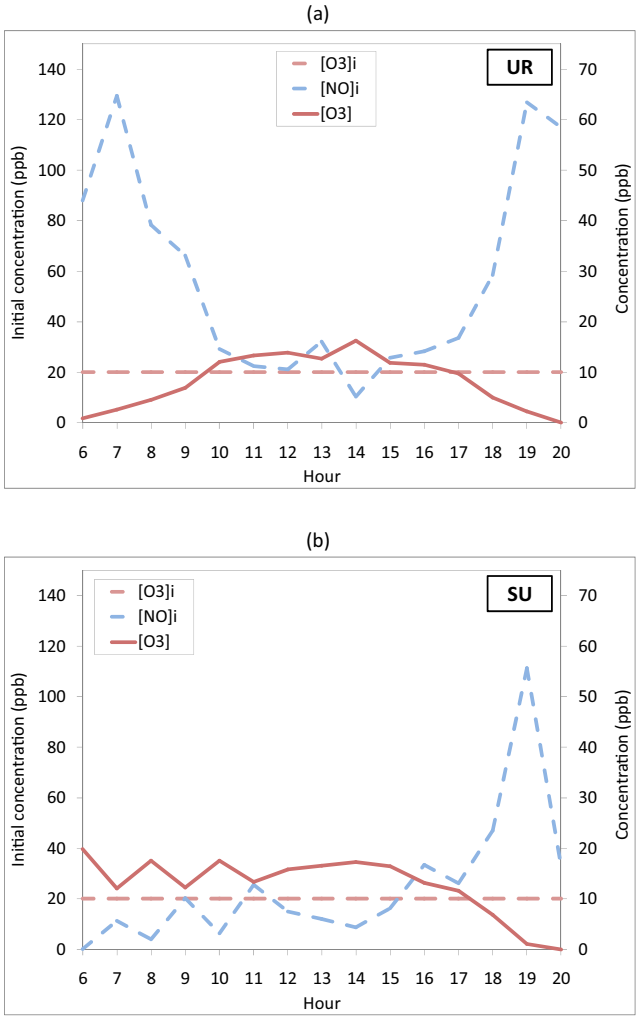

Fig. 8 Diurnal variations of the concentration of O_3 ($[O_3]$) and the initial concentrations of NO ($[NO]_i$) and O_3 ($[O_3]_i$), at two receptors of the MABA: **a** an urban receptor (UR) where cluster 1 dominates and **b** a suburban receptor (SU) where cluster 3 dominates, during their days of occurrence of C_{max}

NO (coming from the MABA) starts removing O_3 via the NO titration reaction. In the following hours, the wind keeps rotating hourly bringing pollutants from the MABA. As shown in Fig. 8b, from 7 to 10 h, $[O_3]$ still depends strongly on $[NO]_i$ (which is supported by a strong correlation between $[NO]_i$ and $[O_3]$ ($R^2 = 0.71$)). After that, photochemistry starts to dominate. In this case, the O_3 morning peak (17.6 ppb) occurs at 8 h and is slightly higher than the one occurring at 14 h (17.3 ppb). This means that when the solar radiation is low, an ozone morning maximum (higher than the midday peak) can occur if $[NO]_i \ll [O_3]_i$ and the diurnal amplitude is relatively small. When these simulations are repeated including the standard memory component of the model, the above analysis is still valid with the only addition of an increased morning peak at the suburban receptor (not shown).

This explains why the second type of modeled ozone profile only occurs in the MABA surroundings with winds from outside the urban area and not at receptors with high and moderate emissions (where $[NO]_i$ is always relatively large). These results are consistent with those obtained by Bogo et al. (1999) who measured O_3 hourly concentrations at a coastal site of the city in spring 1995 and found that the highest O_3 peak concentration occurred with very low NO concentration and wind coming from the vast de la Plata River estuary. The authors also show examples where the maximum O_3 concentration occurs during morning hours. Our results suggest that, while such a morning peak may not be responsible for C_{max} in the urban area of the MABA, where most of the measurements have been made, it is in the surroundings where morning peaks become more relevant. The observation of this distinct behavior suggests that it would be very interesting to monitor O_3, NO, and NO_2 concentrations at these less explored areas.

Summary and conclusions

Our main result is a qualitative characterization of the type of solutions that can be obtained with the DAUMOD-GRS to estimate the summer maximum O_3 concentration (C_{max}) in the MABA. In a previous work (Pineda Rojas et al. 2016), the gridded uncertainty of C_{max} due to the uncertainties in the DAUMOD-GRS input variables was assessed applying the Monte Carlo (MC) analysis. A sensitivity analysis performed at eight selected receptors showed that the relative contributions from nine input variables to total C_{max} uncertainty vary spatially, with the regional background O_3 concentration being the dominant input. The present work, in contrast, focuses on the identification and characterization of atmospheric and emission conditions leading to C_{max} in the gridded MC outcomes. To describe such conditions, we apply clustering analysis aiming to understand the dynamics of the DAUMOD-GRS model in different parts of the MABA, especially in its surroundings where the largest C_{max} and uncertainty values are estimated and the lack of observations impedes its statistical evaluation.

Applying the *k*-means algorithm, four families of conditions that lead to the occurrence of C_{max} are identified. The spatial variation of the dominant cluster (i.e., the most present in MC simulations) appears to be associated to that of the NO_x and VOC emissions in the MABA. At urban and suburban receptors (i.e., receptors with emissions), two clusters are mostly present: in both of them, C_{max} occurs on average at 13–14 h, under conditions of clear sky, moderate to intense winds, and relatively high air temperature and solar radiation. At the most urbanized area, a wider range of the emission rates appears to lead to a greater variation in the conditions under which C_{max} can occur, compared with the suburban zone. At the surroundings of the MABA (where no emissions are considered and the highest C_{max} values are simulated), other two clusters are obtained: one in which C_{max} occurs only during early-morning

hours, under clear sky conditions, low wind speed, and variable wind direction (in occasions coming from outside the MABA) and another cluster (mainly in emission transition areas) in which O_3 peak concentrations occur on average at 15 h under partly cloudy sky conditions. Less conventional model results revealed by one of the clusters (i.e., an ozone morning peak or values of C_{max} occurring with winds that come from outside the urban area) are consistent with the few measurements carried out in the city of Buenos Aires. This suggests that further monitoring efforts at suburban or rural areas could be particularly useful to enhance our current knowledge of O_3 dynamics in the MABA surroundings (and hence determine if further model adjustments or better parameter estimations are needed for this region). Our results exemplify the way in which clustering can be helpful in the analysis of Monte Carlo simulations, to unveil stereotypical spatial patterns in large collections of modeled concentration peaks and discriminate between the families of conditions generating them.

Funding information This study has been supported by the ANPCyT Project PICT2015-1676.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Afif C, Dutot AL, Jambert C, Abboud M, Adjizian-Gérard J, Farah W, Perros PE, Rizk T (2009) Statistical approach for the characterization of NO_2 concentrations in Beirut. Air Qual Atmos Health 2:57–67

Austin E, Coull BA, Zanobetti A, Koutrakis P (2013) A framework to spatially cluster air pollution monitoring sites in US based on the $PM_{2.5}$ composition. Environ Int 59:244–254

Awang NR, Elbayoumi M, Ramli NA, Yahaya AS (2016) Diurnal variations of ground-level ozone in three port cities in Malaysia. Air Qual Atmos Health 9(1):25–39

Beaver S, Palazoglu A (2006) A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. Atmos Environ 40:713–725

Bergin MS, Noblet GS, Petrini K, Dhieux JR, Milford JB, Harley RA (1999) Formal uncertainty analysis of a Lagrangian photochemical air pollution model. Environ Sci Technol 33:1116–1126

Bogo H, Negri RM, San Roman E (1999) Continuous measurement of gaseous pollutants in Buenos Aires City. Atmos Environ 33:2587–2598

Borge R, Lumbreras J, Vardoulakis S, Kassomenos P, Rodriguez E (2007) Analysis of long-range transport influences on urban PM_{10} using two-stage atmospheric trajectory clusters. Atmos Environ 41: 4434–4450

Chang JC, Hanna SR (2005) Technical descriptions and user's guide for the BOOT statistical model evaluation software package, version 2.0, p. 64. Available on. http://www.harmo.org/Kit/Download/BOOT_UG.pdf

Davies JM, Eder BK, Nychka D, Yang Q (1998) Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. Atmos Environ 32(14/15):2505–2520

Derwent D, Fraser A, Abbott J, Jenkin M, Willis P, Murrells T (2010) Evaluating the performance of air quality models. DEFRA report. Issue 3/June 2010. Available on. http://www.airquality.co.uk/reports/cat05/1006241607_100608_MIP_Final_Version.pdf

Dimitriou K, Kassomenos P (2014) Decomposing the profile of PM in two low polluted German cities - mapping of air mass residence time, focusing on potential long range transport impacts. Environ Pollut 190:91–100

Flemming J, Stern R, Yamartino RJ (2005) A new air quality regime classification scheme for O_3, NO_2, SO_2 and PM_{10} observations sites. Atmos Environ 39:6121–6129

Gomez-Losada A, Pires JCM, Pino-Mejías R (2018) Modelling background air pollution exposure in urban environments: implications for epidemiological research. Environ Model Soft 105:13–21

Hanna SR, Chang JC, Fernau ME (1998) Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. Atmos Environ 32(21):3619–3628

Hanna SR, Russell AG, Wilkinson JG, Vukovich J, Hansen DA (2005) Monte Carlo estimation of uncertainties in BEIS3 emission outputs and their effects on uncertainties in chemical transport model predictions. J Geophys Res 110:D01302. https://doi.org/10.1029/2004JD004986

Hanna SR, Paine R, Heinold D, Kintigh E, Baker D (2007) Uncertainties in air toxics calculated by the dispersion models AERMOD and ISCST3 in the Houston ship channel area. J Appl Meteorol Climatol 46:1372–1382

Henne S, Brunner D, Folini D, Solberg S, Klausen J, Buchmann B (2010) Assessment of parameters describing representativeness of air quality in-situ measurement sites. Atmos Chem Phys 10:3561–3581

Jin L, Harley RA, Brown NJ (2011) Ozone pollution regimes modelled for a summer season in California's San Joaquin Valley: a cluster analysis. Atmos Environ 45:4707–4718

Karaca F, Camci F (2010) Distant source contributions to PM_{10} profile evaluated by SOM based cluster analysis of air mass trajectory sets. Atmos Environ 44:892–899

Khedairia S, Khadir MT (2012) Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria. Atmos Res 113:89–101

Lu HC, Chang CL, Hsieh JC (2006) Classification of PM_{10} distributions in Taiwan. Atmos Environ 40:1452–1463

Mazzeo NA, Venegas LE, Choren H (2005) Analysis of NO, NO_2, O_3 and NO_x concentrations measured at a green area of Buenos Aires City during wintertime. Atmos Environ 39:3055–3068

Moore GE, Londergan RJ (2001) Sampled Monte Carlo uncertainty analysis for photochemical grid models. Atmos Environ 35:4863–4876

Pakalapati S, Beaver S, Romagnoli JA, Palazoglu A (2009) Sequencing diurnal air flow patterns for ozone exposure assessment around Houston, Texas. Atmos Environ 43:715–723

Park EH, Heo J, Hirakura S, Hashizume M, Deng F, Kim H, Yi S (2018) Characteristics of $PM_{2.5}$ and its chemical constituents in Beijing, Seoul, and Nagasaki. Air Qual Atmos Health 11(10):1167–1178

Pineda Rojas AL (2014) Simple atmospheric dispersion model to estimate hourly ground-level nitrogen dioxide and ozone concentrations at urban scale. Environ Model Softw 59:127–134

Pineda Rojas AL, Venegas LE (2013) Upgrade of the DAUMOD atmospheric dispersion model to estimate urban background NO_2 concentrations. Atmos Res 120-121:147–154

Pineda Rojas AL, Venegas LE, Mazzeo NA (2016) Uncertainty of modelled urban peak O_3 concentrations and its sensitivity to input data perturbations based on the Monte Carlo analysis. Atmos Environ 141:422–429

Rimetz-Planchon J, Perdrix E, Sobanska S, Bremard C (2008) PM_{10} air quality variations in an urbanized and industrialized harbour. Atmos Environ 47:7274–7283

Rodriguez MA, Bouwer J, Samuelsen GS, Dabdub D (2007) Air quality impacts of distributed power generation in the South Coast Air

Basin of California 2: model uncertainty and sensitivity analysis. Atmos Environ 41:5618–5635

Rouseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20(1):53–65

Russell A, Dennis R (2000) NARSTO critical review of photochemical models and modelling. Atmos Environ 34(12–14):2283–2324

Tan Y, Robinson AL, Presto AA (2014) Quantifying uncertainties in pollutant mapping studies using the Monte Carlo method. Atmos Environ 99:333–340

Tang X, Wang Z, Zhu J, Gbaguidi AE, Wu Q, Li J, Zhu T (2010) Sensitivity of ozone to precursor emissions in urban Beijing with a Monte Carlo scheme. Atmos Environ 44:3833–3842

Terrouche A, Ali-Khodja H, Kemmouche A, Bouziane M, Derradji A, Charron A (2016) Identification of sources of atmospheric particulate matter and trace metals in Constantine, Algeria. Air Qual Atmos Health 9(1):69–82

Unal YS, Toros H, Deniz A, Incecik S (2011) Influence of meteorological factors and emission sources on spatial and temporal variations of PM_{10} concentrations in Istanbul metropolitan area. Atmos Environ 45:5504–5513

Venegas LE, Mazzeo NA, Pineda Rojas AL (2011) Chapter 14: evaluation of an emission inventory and air pollution in the metropolitan area of Buenos Aires. In: Popovic (ed) Air Quality-models and Applications, Editorial In-tech, pp 261–288

Wang HL, Qiao LP, Lou SR, Zhou M, Ding AJ, Huang HY, Chen JM, Wang Q, Tao SK, Chen CH, Li L, Huang C (2016) Chemical composition of $PM_{2.5}$ and meteorological impact among three years in urban Shanghai, China. J Clean Prod 112:1302–1311