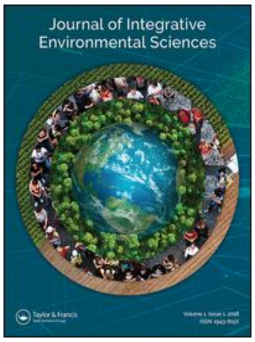

Journal of Integrative Environmental Sciences

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nens20

Statistical analysis of factors driving surface ozone variability over continental South Africa

Tracey Leah Laban , Pieter Gideon Van Zyl , Johan Paul Beukes , Santtu Mikkonen , Leonard Santana , Miroslav Josipovic , Ville Vakkari , Anne M. Thompson , Markku Kulmala & Lauri Laakso

To cite this article: Tracey Leah Laban , Pieter Gideon Van Zyl , Johan Paul Beukes , Santtu Mikkonen , Leonard Santana , Miroslav Josipovic , Ville Vakkari , Anne M. Thompson , Markku Kulmala & Lauri Laakso (2020): Statistical analysis of factors driving surface ozone variability over continental South Africa, Journal of Integrative Environmental Sciences, DOI: 10.1080/1943815X.2020.1768550

To link to this article: https://doi.org/10.1080/1943815X.2020.1768550

Published online: 03 Jun 2020.

JOURNAL OF INTEGRATIVE ENVIRONMENTAL SCIENCES
https://doi.org/10.1080/1943815X.2020.1768550

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Statistical analysis of factors driving surface ozone variability over continental South Africa

Tracey Leah Laban[a], Pieter Gideon Van Zyl[a], Johan Paul Beukes[a], Santtu Mikkonen[b], Leonard Santana[c], Miroslav Josipovic[a], Ville Vakkari[d], Anne M. Thompson[e], Markku Kulmala[f] and Lauri Laakso[d]

[a]Chemical Resource Beneficiation, North-West University, Potchefstroom, South Africa; [b]Department of Applied Physics, University of Eastern Finland, Kuopio, Finland; [c]School of Mathematical and Statistical Sciences, North-West University, Potchefstroom, South Africa; [d]Finnish Meteorological Institute, Helsinki, Finland; [e]Earth Sciences Division, NASA/Goddard Space Flight Center, Greenbelt, MD, USA; [f]Department of Physics, University of Helsinki, Helsinki, Finland

ABSTRACT

Statistical relationships between surface ozone (O_3) concentration, precursor species and meteorological conditions in continental South Africa were examined from data obtained from measurement stations in north-eastern South Africa. Three multivariate statistical methods were applied in the investigation, i.e. multiple linear regression (MLR), principal component analysis (PCA) and –regression (PCR), and generalised additive model (GAM) analysis. The daily maximum 8-h moving average O_3 concentrations were considered in these statistical models (dependent variable). MLR models indicated that meteorology and precursor species concentrations are able to explain ~50% of the variability in daily maximum O_3 levels. MLR analysis revealed that atmospheric carbon monoxide (CO), temperature and relative humidity were the strongest factors affecting the daily O_3 variability. In summer, daily O_3 variances were mostly associated with relative humidity, while winter O_3 levels were mostly linked to temperature and CO. PCA indicated that CO, temperature and relative humidity were not strongly collinear. GAM also identified CO, temperature and relative humidity as the strongest factors affecting the daily variation of O_3. Partial residual plots found that temperature, radiation and nitrogen oxides most likely have a non-linear relationship with O_3,while the relationship with relative humidity and CO is probably linear. An inter-comparison between O_3 levels modelled with the three statistical models compared to measured O_3 concentrations showed that the GAM model offered a slight improvement over the MLR model. These findings emphasise the critical role of regional-scale O_3 precursors coupled with meteorological conditions in daily variances of O_3 levels in continental South Africa.

ARTICLE HISTORY
Received 23 August 2019
Accepted 21 April 2020

KEYWORDS
Tropospheric ozone (O_3); multiple linear regression; principal component analysis; generalized additive models; Welgegund

CONTACT Pieter Gideon van Zyl pieter.vanzyl@nwu.ac.za Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa

1. Introduction

Surface O_3 is a secondary pollutant, which is considered a relatively short-lived (lifetime ranging between days to weeks) greenhouse gas (Ordonez et al. 2005). In general, high surface O_3 concentrations are a concern because of its detrimental impacts on human health and ecosystem functioning (NRC 2008). The potential for O_3 damage to plants is, especially, a concern when agricultural yields are reduced, which threatens the food security and economies of countries that rely strongly on agricultural production. However, an important consequence of plant damage caused by increased O_3 levels relate to the reduced removal of CO_2 in the atmosphere and thereby O_3 also indirectly contributes to climate change. In addition, tropospheric O_3 can also affect new particle formation in the atmosphere (e.g. Mikkonen et al. 2011), which also impacts climate change directly (e.g. scattering) and indirectly (e.g. cloud formation).

O_3 in the troposphere is produced by the photochemical oxidation of nitrogen dioxide (NO_2):

$$NO_2 + h\nu \rightarrow NO + O \tag{1.1}$$

$$O + O_2 + M \rightarrow O_3 + M \tag{1.2}$$

The photolytically formed O_3 reacts with NO to regenerate NO_2:

$$O_3 + NO \rightarrow NO_2 + O_2 \tag{1.3}$$

This is a continuous process termed the NO_x-dependent photo-stationary state (PSS), which results in no net O_3 production (Seinfeld and Pandis 2006; Awang et al. 2018). However, when this PSS is altered in the presence of carbon monoxide (CO) and volatile organic compounds (VOCs), net O_3 production occurs. High O_3 levels are not only a result of chemistry associated with precursor emissions but are also related to meteorological conditions conducive to the formation, transport and removal of air pollutants (Melkonyan and Kuttler 2012). Local meteorological parameters, such as temperature, relative humidity, sunlight, and wind speed and -direction play a significant role in O_3 variability (Ooka et al. 2011; Tsakiri and Zurbenko 2011). These multiple factors influencing surface O_3 levels have confounded the effect of individual parameters on ground-level O_3, thereby making it challenging to separate the impacts of local emissions, meteorology and transport on surface O_3 concentrations (Gorai et al. 2015).

Statistical models relating ambient O_3 concentrations to meteorological variables have been developed for the purpose of the prediction of O_3 concentrations, the estimation of long-term O_3 trends, as well as explaining the underlying chemical and meteorological processes affecting O_3 concentrations (Thompson et al. 2001). Some of these statistical methods were critically reviewed by Thompson et al. (2001), which included regression-based methods (Fiore et al. 1998; Abdul-Wahab et al. 2005; Ooka et al. 2011), time-series filtering (Rao and Zurbenko 1994; Milanchus et al. 1998; Tsakiri and Zurbenko 2011), multivariate statistical techniques such as cluster analysis and principal component analysis (PCA) (Abdul-Wahab et al. 2005; Melkonyan and Kuttler 2012; Dominick et al. 2012; Awang et al. 2015), as well as neural networks (Comrie 1997; Gardner and Dorling 1998, 2000; Guardani et al. 2003). The most widely used statistical technique to relate O_3 concentrations to influencing factors is linear regression, because of its user-friendliness and straightforward interpretability (Comrie 1997; Cardelino et al. 2001). However, the relationship between O_3 levels and certain meteorological effects is typically non-linear, while some explanatory variables are collinear (Neter et al. 1996). Although non-linear regression models for O_3 forecasting have been developed (Bloomfield et al. 1996; Thompson et al. 2001; Lin and Cobourn 2007), these models are difficult to interpret and explain in summarized form to the public (Thompson et al. 2001; Pearce et al. 2011). However, generalized additive models (GAM), which are an extension of linear regression, are able to handle non-linear associations between atmospheric parameters and are simpler to interpret or justify (Hastie and Tibshirani 1990). Melkonyan and Kuttler (2012) suggested that PCA is the most appropriate method to identify multivariate relationships between pollutants and meteorological factors.

Southern Africa is the largest industrialized region in Africa, where high O_3 levels may be expected due to the high rate of precursor emissions from anthropogenic sources, coupled with the abundance of sunlight throughout the year (Zunckel et al., 2006). In addition, this region is also influenced by large-scale open biomass burning, which is considered to be a significant source of O_3 precursor species. Laban et al. (2018) indicated that CO emissions associated with biomass burning (household combustion and open biomass burning) contributed significantly to high O_3 levels, while it was also indicated that large parts of the regional background in South Africa can be considered VOC-limited. Although the temporal and spatial variability is generally attributed to meteorological conditions and/or precursor emissions, the response of O_3 with respect to changing emission levels and meteorological fluctuations is not well understood for this region (Laban et al. 2018). Therefore, the aim of this study was to utilize statistical models to distinguish the complex effects of meteorological parameters and precursor emissions influencing O_3 chemistry and concentrations in continental South Africa, as well as to quantify the strength of association of O_3 with these factors in order to better understand the underlying mechanisms responsible for the changes in surface O_3 levels in this region.

2. Material and methods

2.1. Description of the study area

Data from continuous *in-situ* measurements conducted at four measurement sites (indicated in Table 1) in the north-eastern interior of South Africa were obtained for statistical analysis. This region is the largest industrial area in South Africa, with substantial

Table 1. Measurement stations from which meteorological- and air pollutant data utilized for statistical analysis were obtained.

Measurement site	Latitude Longitude (decimal degrees)	Elevation (m) a.s.l.	Measurement period	Site description
Welgegund	26.57° S 26.94° E	1480	May 2010-Dec 2015	Rural, background
Botsalano	25.54° S 25.75° E	1420	Jul 2006-Jan 2008	Rural, background
Marikana	25.70° S 27.48° E	1170	Feb 2008-Apr 2010	Rural, residential, industrial
Elandsfontein	26.25° S 29.42° E	1750	Feb 2009-Jan 2011	Rural, industrial

emissions of atmospheric pollutants from anthropogenic activities, e.g. industries, domestic fuel burning and vehicles (Lourens et al. 2011, 2012). A combination of meteorology and anthropogenic activities has amplified pollution levels within the region. Detailed descriptions of the locations of these four measurement stations and their surroundings are provided in Laban et al. (2018).

Measurements were conducted from 20 July 2006 until 5 February 2008 at Botsalano, 8 February 2008 to 16 May 2010 at Marikana, 20 May 2010 to 31 December 2015 at Welgegund and 11 February 2009 to 31 December 2010 at Elandsfontein. These four measurement stations represent high quality, high resolution data, which include comprehensive continuous measurements of aerosols, trace gases and meteorological parameters. Data quality was ensured through regular site visits, while data collected from these four sites were subjected to meticulous cleaning (e.g. excluding measurements recorded during calibrations and maintenance). The data were available as 15-min averages.

2.2. Data treatment

Respiratory symptoms have been found to be associated with the daily maximum of the eight-hour average O_3 concentration (Schlink et al. 2006). Therefore, the South African National Ambient Air Quality Standards and other international standards, designed to protect human health, are based on this metric. Consequently, the daily maximum 8-h moving average O_3 concentrations (daily max 8-h O_3) were utilized in the statistical analysis (dependent variable). The choice of input (independent) variables for the models was based on literature (Dueñas et al. 2002; Ordonez et al. 2005; Abdul-Wahab et al. 2005; Camalier et al. 2007; Awang et al. 2015), as well as exploratory analysis and a general understanding of O_3-related processes (Equation 1.1–1.3). Daytime (11:00–17:00 local time) daily average concentrations were calculated for NO_2, NO and CO, while daily mean values for zonal (u) wind component, meridional (v) wind component, relative humidity and solar radiation were determined. Daily maximum temperatures were included in models. Only daytime measurements were used in the statistical models, since the boundary layer is deep and well mixed during this period, as well as to exclude night-time chemistry (Cooper et al. 2012). Other variables such as soil moisture and precipitation, as well as SO_2- and H_2S levels were also explored, but were found to have only a minor influence on daily max 8-h O_3. Since the O_3 data utilized in this study were normally distributed, it was not necessary to log-transform the original data to satisfy parametric test assumptions.

Exploratory descriptive statistics (calculation of mean, median, minimum, maximum and standard deviation) were employed prior to the statistical analyses in order to gain a general understanding of meteorological, O_3, NO_x and CO variations at the measurement locations. Correlation coefficients were also calculated as a measure of the linear relationship between O_3 and each variable.

2.3. Statistical methods

Three different statistical methods, namely MLR, PCA and GAM were used to statistically evaluate the datasets. A separate model was built for each measurement site and used to investigate the influence of meteorological and precursor species (indicated in section 2.2.) variability on daily max 8-h O_3 at each site. The statistical calculations were performed using MATLAB version R2013a or R software environment (R Development Core Team 2009).

2.3.1. Multiple linear regression (MLR)

Multiple linear regression modelling was used to relate O_3 concentrations (daily max 8-h O_3) to meteorological and pollutant factors, as well as the relative contribution of each of these factors. The general equation for an MLR model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \tag{1}$$

where Y is the response variable, $X_1, X_2, \ldots, X_p$ are the exploratory variables, $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients, and ε is an error term or residual value associated with deviation between the observed value of Y and the predicted Y value from the regression equation. The ordinary least squares procedure is the standard method to estimate the coefficients in the MLR equation. With this method, the regression procedure is based on finding coefficient values that minimize the sum of the squares of the residuals. A forward stepwise regression procedure was used in which each variable was added individually to the starting model according to their statistical significance and overall increase in the explanation capability of the model. This was done to remove the least important predictor variables and to obtain the optimal combination of variables depending on the statistical indices.

The strength of relationship between each independent variable and O_3 was evaluated in terms of the magnitude of the t-statistic and associated p-value for statistical significance. The performance of the model was evaluated with R^2, adjusted R^2 and root mean square error (RMSE). The adjusted-R^2 is an R^2 measure that does not increase unless the new variables have additional predictive capability (unlike R^2 that increases when variables are added to the equation even when the new variables have no real predictive capability). The optimum MLR models considered had the largest R^2 and adjusted R^2, and smallest RMSE from a minimum number of independent variables. The main assumptions of the model are true underlying linearity, residuals are mutually independent with constant variance (homoscedasticity), and residuals are normally distributed (Ordonez et al. 2005). Multicollinearity in the regression model was verified by examining the variance inflation factor (VIF) for each of the predictor variables (Abdul-Wahab et al. 2005; Otero et al. 2016).

2.3.2. Principal component analysis (PCA) and -regression (PCR)

Parameters such as solar radiation, temperature and relative humidity are related properties, which could be inessential in MLR. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of interrelated variables into a set of uncorrelated variables, i.e. principal components. Therefore, PCA is able to separate interrelationships (collinearity) into statistically independent basic components (Abdul-Wahab et al. 2005) and determine the most important uncorrelated variables. Each principal component is a linear combination of the original predictor variables that account for the variance in the data. All the principal components are orthogonal to each other, which implies that they are uncorrelated to each other. The first principal component is calculated such that it accounts for the highest possible variance in the dataset, followed by the concurrent components. Since the variables are measured in different units, it is necessary to standardize data before a principal component analysis is carried out, which involves scaling every variable to have a mean equal to 0 and a standard deviation equal to 1. The principal component model presents the i^{th} principal component as a linear function of the p measured variables as expressed in Eq. (2) below:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + a_{i2}X_2 + \ldots + a_{ip}X_p \tag{2}$$

where *"Z"* is the principal component, *"a"* is the component loading, and *"X"* is the measured variable. The full set of principal components is as large as the original set of variables, but it is common for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data. By examining plots of these few new variables, researchers often develop a deeper understanding of the driving forces that generated the original data.

PCA was first applied to the original independent variables to transform these variables into an equal number of principal components. Only those principal components with an eigenvector greater than 1 were retained (according to the Kaiser criterion), which were then subjected to Varimax rotation to maximize the loading of a predictor variable on one component (Abdul-Wahab et al. 2005). Since the eigenvectors are the correlation of the component variables with the original variables, they comprise coefficients (loadings) that indicate the relative weight of each variable in the component, which is important, since they represent the extent of the correlation between the measured variable and the principal components. Variables that load highly on a specific principal component form a related group.

PCR is a combination of PCA and MLR (Awang et al. 2015), where the outputs from the PCA are used as potential predictors in order to improve the original MLR model (Abdul-Wahab et al. 2005; Awang et al. 2015). Either the original independent variables associated with each of the principal components with high loadings (Abdul-Wahab et al. 2005) or the principal components with high loadings (Awang et al. 2015) are selected to be included in the regression equation.

2.3.3. Generalized additive models (GAMs)

GAMs extend traditional linear models by allowing for an alternative distribution for the modelling of response variables that have a non-normal error distribution. In addition, GAMs do not force dependent variables to be linearly related to independent variables as in MLR, and recognize that the relationship of some explanatory variables (e.g. daily temperature) and the response variable (i.e. ozone in this study) may not be linear (Gardner and Dorling 2000). In GAMs, the response variable depends additively on unknown smoothing functions of the individual predictors that can be (linear) parametric or non-parametric (Hastie and Tibshirani 1990). The GAM model equation developed by Hastie and Tibshirani (1990) is given by

$$g(E(Y_i)) = \beta_0 + s_1(X_{i1}) + s_2(X_{i2}) + \ldots + s_p(X_{ip}) + \varepsilon_i \tag{3}$$

where Y_i is the response variable, $E(Y_i)$ denotes the expected value and $g(\cdot)$ denotes the link function that links the expected value to the predictor variables $X_{i1}, \ldots, X_{ip}$, β_0 is an intercept and ε_i is an i.i.d. random error. For the purposes of the analysis performed in this study, the link function chosen was the identify transformation $g(E(Y_i)) = E(Y_i)$.The terms $s_1(\cdot), s_2(\cdot), \ldots, s_p(\cdot)$ are smooth functions that are estimated in a nonparametric fashion (Hastie and Tibshirani 1990). We can estimate these smooth relationships simultaneously from the data and then predict $g(E(Y_i))$ by simply adding up these functions. The estimated smooth functions s_k are the analogues of the coefficients $\beta_k$in linear regression. In contrast to MLR, an additive regression is done by using a back-fitting procedure and thereby controlling the effects of the other predictors. GAM is able to identify covariates, X_k relevant to Y for a large set of potential factors (Hayn et al. 2009), while it does not require any prior knowledge on the underlying relationship between Y and its covariates. The latter can be obtained through separate partial residual plots, which allow visualization of the relationships between each variable X_k and the response variable, Y, after accounting for the effects of the other explanatory variables in the model.

Smooth parameters were automatically selected in the "mgcv" package (Wood 2017) in the R software environment used in this study, which is based on maximum probability methods that minimize the Akaike information criterion (AIC) score. The AIC measures the goodness-of-fit of the model in such a manner that the final model selected has the smallest AIC. The models were also evaluated with R^2 values and generalized cross-validation (GCV) scores (estimate of the prediction error).

3. Results and discussion

3.1. Exploratory analysis

3.1.1. Descriptive statistics

As indicated in Section 2.2, descriptive statistics were performed prior to the statistical analyses in order to gain a general understanding of meteorological, O_3, NO_x and CO variations at the measurement locations, which are presented in Table 2. It is evident that Elandsfontein and Marikana are the more polluted sites, as indicated by higher NO_2, NO and CO median values, whereas Botsalano had the lowest median values for NO_2, NO and CO. Note that O_3 concentrations are similar at all sites, even though Botsalano and Welgegund are considered regional background sites. The regional problem associated with O_3 in southern Africa was indicated by Laban et al. (2018). The large standard deviations of NO_2 and NO concentrations can be attributed to occasional high pollution events.

3.1.2. Calculation of correlation coefficients

In Table 3, Pearson correlation coefficients (r) relating O_3 concentration with individual atmospheric parameters at the four measurement locations are presented. It is evident that O_3 has a positive correlation with temperature and global radiation, while it is negatively correlated with relative humidity. A relatively strong positive correlation with CO was observed at Welgegund, Botsalano and Marikana, with NO_2 and NO correlations with O_3 almost negligible at these sites due to the time scale. The correlations with u and v wind components are also weak, as given by their low correlation coefficients. Exploratory Pearson correlations indicate that variability in O_3 levels is in general associated (positively or negatively) with CO (r(O_3,CO) = 0.3 to 0.6), relative humidity (r(O_3, RH) = –0.2 to –0.5) and temperature (r(O_3, T) = 0.2 to 0.5). The significance of CO on O_3 levels in this north-eastern

Table 2. Descriptive statistics of the daily summaries of the key variables used in the study.

	Time scale	Statistics	Welgegund	Botsalano	Marikana	Elandsfontein
[O_3] ppb	Daily 8-h max	Mean	47	47	50	48
		Median	46	48	48	47
		Min	8	21	14	11
		Max	114	73	113	102
		Std Dev	11	9	16	16
[NO_2] ppb	Daily average	Mean	2.0	1.5	5.7	13.2
		Median	1.4	1.3	4.8	10.8
		Min	−0.4	0.2	0.0	0.2
		Max	21.2	11.4	20.9	68.3
		Std Dev	1.9	1.0	3.3	9.7
[NO] ppb	Daily average	Mean	0.4	0.3	2.8	4.5
		Median	0.2	0.2	1.6	2.6
		Min	−0.4	−0.1	−0.3	0.1
		Max	6.9	5.3	52.8	42.5
		Std Dev	0.7	0.4	3.8	5.4
[CO] ppb	Daily average	Mean	126	118	197	
		Median	116	109	181	
		Min	23	57	85	
		Max	412	308	591	
		Std Dev	45	35	68	
Solar Radiation W/m^2	Daily average	Mean	508	508	462	522
		Median	490	504	458	541
		Min	14	31	24	3
		Max	871	835	884	1005
		Std Dev	154	137	146	156
Temperature °C	Daily maximum	Mean	24	25	26	21
		Median	25	26	27	21
		Min	5	8	10	6
		Max	38	36	37	30
		Std Dev	5	5	5	4
Relative Humidity %	Daily average	Mean	42	40	49	52
		Median	40	38	48	53
		Min	6	7	10	9
		Max	100	95	100	96
		Std Dev	18	19	18	18
Zonal (u) wind component (m/s)	Daily average	Mean	0.7	−2.8	0.5	0.4
		Median	1.1	−3.3	0.5	0.9
		Min	−13.1	−13.4	−6.9	−9.1
		Max	12.9	10.0	8.0	8.7
		Std Dev	3.6	3.9	2.4	3.2
Meridional (v) wind component (m/s)	Daily average	Mean	−0.8	−0.6	−0.3	−0.8
		Median	−0.8	−0.6	−0.2	−0.7
		Min	−10.4	−7.4	−5.7	−10.0
		Max	10.9	6.3	5.9	5.2
		Std Dev	2.7	1.9	1.4	2.4

interior of South Africa was indicated by Laban et al. (2018). The relative significance of CO, relative humidity and temperature highlighted with these correlations is further explored in subsequent sections through more advanced statistical methods, as indicated in section 2.3.

3.2. Multiple linear regression (MLR) analysis

A summary of the contributions of independent variables to variation of the dependent variable (daily max 8-h O_3) included in the optimum MLR models obtained for each of the measurement sites is presented in Table 4. VIF values ranging between 1.00 and 2.00 for all the independent variables indicated moderate collinearity, which did not contribute to unstable parameter estimates or the necessity to remove any independent variables from

Table 3. Pearson correlation coefficient (r) for the different variables with their associated p-values (P) for data from the four sites.

		Daily 8-h max O_3 (ppb)			
		Welgegund	Botsalano	Marikana	Elandsfontein
Daily average NO_2 (ppb)	r	0.113	0.061	0.128	−0.096
	P	0.000	0.197	0.001	0.018
Daily average NO (ppb)	r	−0.077	−0.141	−0.026	−0.211
	P	0.001	0.003	0.508	0.000
Daily average CO (ppb)	r	0.554	0.543	0.330	
	P	0.000	0.000	0.000	
Daily average radiation (W/m^2)	r	0.204	0.324	0.290	0.237
	P	0.000	0.000	0.000	0.000
Daily maximum temp (°C)	r	0.374	0.518	0.434	0.207
	P	0.000	0.000	0.000	0.000
Daily average relative humidity (%)	r	−0.428	−0.242	−0.486	−0.451
	P	0.000	0.000	0.000	0.000
Zonal (u) wind component (m/s)	r	−0.002	−0.094	0.074	0.079
	P	0.921	0.033	0.042	0.052
Meridional (v) wind component (m/s)	r	−0.167	−0.253	−0.083	−0.070
	P	0.000	0.000	0.023	0.085

the models. Regression analysis explained approximately 50% of the variability ($R^2 \approx 0.5$) of daily max 8-h O_3 concentrations at Welgegund, Botsalano and Marikana, with lower R^2 (0.261) at Elandsfontein attributed to CO not measured at this site and not included in the MLR.

From Table 4, it is evident that CO, T and RH make the most significant contributions to the variance in daily max 8-h O_3 at Welgegund, Botsalano and Marikana as indicated by the magnitude of the t-statistics. In the absence of CO measurements at Elandsfontein, RH and NO predominantly contributed to variances in daily max 8-h O_3, while notable contributions are also made by NO levels at Welgegund. A positive regression coefficient associated with temperature is expected due to the photochemical production of O_3 (Equations 1.1–1.3). In addition, evaporative emissions of anthropogenic VOCs increase at high temperatures (Ordonez et al. 2005; Jaars et al. 2014), which could favour O_3 formation as previously mentioned. Relative humidity had a negative regression coefficient and a significant t-statistic at three of the sites, which indicate that low relative humidity is associated with high daily max 8-h O_3. This influence of relative humidity on O_3 variances suggests that atmospheric wet conditions can affect O_3 production and loss, which will be explored later in this paper. Surprisingly, the contribution of relative humidity to O_3 variation was similar to that of temperature at Welgegund, while it had the most significant contribution at Elandsfontein (in the absence of any CO measurements). CO levels have the highest contribution to variations in daily max 8-h O_3 at Welgegund and Botsalano, i.e. the two regional background sites, while it had the second highest contribution at the industrialized Marikana site. Laban et al. (2018) indicated that CO emissions associated with regional open biomass burning, as well as household combustion for space heating and cooking, contributed significantly to O_3 levels in the interior of southern Africa. Negative regression coefficients associated with NO at Welgegund and Elandsfontein can be attributed to O_3 titration in the presence of high NO levels (Equation 1.3).

Since O_3 has strong seasonal variation, MLR analysis was also performed for each season: winter (JJA), spring (SON), summer (DJF) and autumn (MAM) in order to evaluate

Table 4. Summary of the optimum MLR models for each site showing the individual variable contributions to daily max 8-h O_3.

WELGEGUND	Constant	T (°C)	RH (%)	u (m/s)	v (m/s)	NO (ppb)	CO (ppb)
Regression coefficient (β)	29.31	0.41	−0.17	−0.28	0.11	−2.99	0.12
Standard error	1.17	0.03	0.01	0.06	0.06	0.27	0.00
t-statistic	25.10	11.71	−16.44	−5.00	1.74	−11.23	29.87
P-value	4.61E-119	1.45E-30	1.41E-56	6.20E-07	0.082156854	2.60E-28	5.85E-159
$R^2 = 0.529$	Adjusted $R^2 = 0.528$		RMSE = 6.75		F-statistic = 330		
BOTSALANO	Constant	T (°C)	Rad (W/m^2)	CO (ppb)			
Regression coefficient (β)	8.03	0.69	0.01	0.14			
Standard error	1.72	0.08	0.00	0.01			
t-statistic	4.67	8.65	2.91	16.17			
P-value	3.86E-06	7.34E-17	0.003734848	2.03E-47			
$R^2 = 0.531$	Adjusted $R^2 = 0.528$		RMSE = 6.41		F-statistic = 184		
MARIKANA	Constant	T (°C)	RH (%)	u (m/s)	NO (ppb)	CO (ppb)	
Regression coefficient (β)	8.92	1.45	−0.25	−0.83	−0.57	0.09	
Standard error	4.98	0.12	0.03	0.23	0.15	0.01	
t-statistic	1.79	12.58	−7.53	−3.68	−3.87	10.46	
P-value	7.35E-02	1.66E-32	1.73E-13	2.58E-04	0.000121169	1.07E-23	
$R^2 = 0.454$	Adjusted $R^2 = 0.449$		RMSE = 12.46		F-statistic = 104		
ELANDSFONTEIN	Constant	RH (%)	v (m/s)	NO (ppb)			
Regression coefficient (β)	71.25	−0.39	−0.64	−0.67			
Standard error	1.73	0.03	0.23	0.10			
t-statistic	41.16	−12.90	−2.79	−6.54			
P-value	4.20E-176	9.66E-34	5.47E-03	1.29E-10			
$R^2 = 0.261$	Adjusted $R^2 = 0.257$		RMSE = 13.56		F-statistic = 70		

where T is daily maximum temperature, Rad is daily average global radiation, RH is daily average relative humidity, u is the zonal (east-west) wind component, v is the meridional (north-south) wind component, NO_2 is the daily average NO_2 concentration, NO is the daily average NO concentration and CO is the daily average CO concentration.

the major factors driving O_3 variability during different seasons. Maximum O_3 concentrations generally occur in late winter and spring (August–November) for continental southern Africa (Zunckel et al., 2004; Combrink et al. 1995; Diab et al. 2004). In Table 5, the independent variables with the most significant contributions (i.e. highest t-statistic values in the optimum model) to O_3 variability for different seasons are presented for each site.

CO makes the highest contribution to the variance in daily max 8-h O_3 during all the seasons at Botsalano, during autumn, winter and spring at Welgegund, as well as during spring (second highest in winter) at Marikana, which signifies the influence of CO levels on O_3 concentrations in continental South Africa. The seasonal pattern of CO is also reflected in the seasonal variations of contributing factors to O_3 variability as indicated by a less important influence of CO levels on the variance in O_3 during summer at Welgegund and Marikana. Increased CO emissions in this region are associated with increased household combustion and open biomass burning during winter and spring (Laban et al. 2018). This is also indicated by increased contributions of NO and NO_2 to O_3 variances at Welgegund and Marikana during summer, i.e. increased O_3 titration/formation mainly associated with NO and NO_2 levels (Equation 1.1–1.3). CO has the highest influence on variation O_3 throughout the year at Botsalano, which can be ascribed to the site being more removed from source regions compared to Welgegund. The important influence of relative humidity on O_3 levels is also apparent, as indicated by increases in its contribution to O_3 variances during months coinciding with the wet season, i.e. mid-October to mid-May (mostly summer and autumn). The wet season is also characterized by lower concentrations of air pollutants (and O_3 precursors) due to wet deposition. Daily maximum temperature remains an important contributor to variance in daily max 8-h O_3, except during summer at Welgegund, Botsalano and Marikana. This can be attributed to relatively constant higher temperatures occurring during summer, with O_3 variability associated with other influencing factors, e.g. relative humidity. In the absence of CO measurements at Elandsfontein, daily maximum temperature contributes most significantly to O_3

Table 5. Most important explanatory variables for daily max 8-h O_3 for each season (ranked in decreasing order of importance as given by the magnitude of their t-statistic).

	Summer	Autumn	Winter	Spring
WELGEGUND	NO	CO	CO	CO
	NO_2	RH	NO	NO
	RH	NO	T	T
	CO	T	u	v
	Rad			RH
BOTSALANO	CO	CO	CO	CO
	RH	T	T	NO
			RH	Rad
				T
MARIKANA	RH	NO_2	T	CO
	NO_2	NO	CO	T
	NO	RH		
	v	T		
		u		
		v		
ELANDSFONTEIN	T	T	NO	T
		NO	NO_2	NO_2
		RH	T	Rad

variability at Elandsfontein on a seasonal scale, which can be attributed to the influence of temperature on the vertical mixing of tall stack emissions of power plants (Ordonez et al. 2005). The highest contribution of NO on O_3 variance at Elandsfontein in winter can be attributed to more pronounced inversion layers, as well as increased household combustion for space heating and cooking.

3.3. Principal component analysis (PCA)

PCA revealed four principal components (factors) with eigenvalues greater than 1 at each of the sites, which explained approximately 80% of the variation in the data. Only these four factors (labelled Factor 1, Factor 2, Factor 3 and Factor 4) were subjected to Varimax rotation, which are presented with their respective loadings, eigenvalues and variances in Table 6. Factor loadings ≥0.5 (or close to 0.5) were considered significant, i.e. strongly correlated within each principal component.

Similar factor loadings were determined for each of the four principal components identified for each site, i.e. a factor with high loadings of T and Rad, a factor with high loadings of NO and NO_2 and a factor with a high loading of RH. A factor with a high loading of CO was determined at Welgegund, Botsalano and Marikana, while one factor was highly loaded with the wind direction vectors at Elandsfontein where CO was not measured. Therefore, PCA indicated that the predominant factors identified by MLR driving variances in daily max 8-h O_3, i.e. CO, T and RH (as well as NO levels in certain instances) are not inter-correlated. Collinearity is expected between T and radiation, as well as NO and NO_2 as revealed by PCA. In addition, Factor 1 at Marikana with high loadings of CO and NO_2 (and NO) is indicative of the influence of household combustion at this site, as indicated by Venter et al. (2012). Furthermore, the correlation between NO_2, NO and CO at Welgegund in Factor 1 also reflects the influence of similar sources of these species at Welgegund and signifies that Welgegund lies in a region between a NO_x- and VOC-limited O_3 production regime, as indicated by Laban et al. (2018). CO is also strongly correlated to meridional wind vector in Factor 4 at the regional background site Welgegund, which can be attributed the regional transport of CO emissions. Welgegund is influenced by the major source regions in the interior of South Africa and a relatively clean background sector to the west (Tiitta et al. 2014; Jaars et al. 2014). In addition, Welgegund is also impacted on by regional biomass burning, contributing to increased CO emissions (Vakkari et al. 2013). In contrast to Welgegund, CO at Botsalano is not correlated to NO and NO_2 and is the only major loading in Factor 4 at this site.

3.4. Generalized additive model (GAM) analysis

Given the complex and non-linear chemistry of O_3 (NRC 1991), the datasets were also statistically analysed with GAM. A summary of the optimum (highest R^2 and lowest AIC) GAM models is shown in Table 7. According to the F-statistics of the optimum models obtained with GAM, RH and CO make the highest contributions to variances in O_3 concentrations Welgegund, Botsalano and Marikana, with T and NO also contributing to O_3 variances at these sites. NO, RH and T contributed to O_3 variability at Elandsfontein where no CO measurements were conducted. These results correspond to the most significant independent variables contributing to variance in O_3 levels indicated by MLR.

Table 6. Factor loadings after PCA followed by Varimax rotation at the four measurement sites. Loadings ≥ 0.5 (or close to 0.5) are indicated in bold.

	Rotated principal component loadings			
Welgegund	Factor 1	Factor 2	Factor 3	Factor 4
T (°C)	−0.060	0.640	−0.012	−0.215
Rad (W/m^2)	0.060	0.728	−0.031	0.166
RH (%)	−0.132	−0.076	0.755	−0.035
u (m/s)	−0.274	−0.028	−0.549	0.020
v (m/s)	0.145	−0.089	−0.157	0.802
NO_2 (ppb)	0.637	−0.093	−0.020	−0.052
NO (ppb)	0.545	0.167	0.179	0.188
CO (ppb)	0.420	−0.101	−0.266	−0.493
Eigenvalue (variance)	2.260	1.620	1.379	1.230
Variance (%)	28.658	20.545	17.484	15.603
Cumulative variance (%)	28.658	49.203	66.687	82.290

	Rotated principal component loadings			
Botsalano	Factor 1	Factor 2	Factor 3	Factor 4
T (°C)	−0.001	−0.049	−0.646	0.068
Rad (W/m^2)	0.039	−0.003	−0.673	−0.136
RH (%)	0.066	−0.540	0.270	−0.404
u (m/s)	−0.070	0.667	0.049	−0.012
v (m/s)	0.185	0.506	0.177	−0.300
NO_2 (ppb)	0.668	−0.042	0.045	0.211
NO (ppb)	0.710	0.029	−0.084	−0.157
CO (ppb)	0.070	−0.052	0.115	0.809
Eigenvalue (variance)	1.802	1.746	1.701	1.328
Variance (%)	22.709	22.003	21.430	16.726
Cumulative variance (%)	22.709	44.712	66.142	82.868

	Rotated principal component loadings			
Marikana	Factor 1	Factor 2	Factor 3	Factor 4
T (°C)	−0.110	−0.602	0.016	−0.108
Rad (W/m^2)	−0.089	−0.625	−0.105	0.021
RH (%)	−0.376	0.484	−0.181	−0.177
u (m/s)	−0.039	0.037	0.973	−0.020
v (m/s)	−0.034	0.032	−0.020	0.967
NO_2 (ppb)	0.563	0.096	−0.063	−0.001
NO (ppb)	0.439	0.034	0.057	0.070
CO (ppb)	0.571	−0.011	−0.048	−0.130
Eigenvalue (variance)	2.510	2.194	1.031	0.996
Variance (%)	30.728	26.864	12.626	12.189
Cumulative variance (%)	30.728	57.592	70.218	82.407

	Rotated principal component loadings			
Elandsfontein	Factor 1	Factor 2	Factor 3	Factor 4
T (°C)	0.006	0.762	−0.108	0.154
Rad (W/m^2)	0.004	0.616	0.103	−0.236
RH (%)	0.034	−0.083	−0.067	0.802
u (m/s)	0.109	−0.137	−0.580	−0.439
v (m/s)	0.073	−0.084	0.798	−0.203
NO_2 (ppb)	0.651	−0.049	−0.001	−0.130
NO (ppb)	0.747	0.066	0.011	0.161
Eigenvalue (variance)	1.676	1.556	1.313	1.287
Variance (%)	24.328	22.581	19.062	18.685
Cumulative variance (%)	24.328	46.910	65.972	84.657

Table 7. Summary of the optimum GAM for each site showing the individual variable contributions to daily max 8-h O_3. This was done with the function gamm in R, which takes into account autocorrelation in the O_3 data.

	GAMM (Welgegund)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	daily max 8-h O_3 ~ s(T) + s(RH) + s(u) + s(NO_2) + s(NO) + s(CO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	45.59	0.40	114.00	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(T)	2.78	2.78	4.29	4.10E-03
2	s(RH)	1.00	1.00	100.93	< 2e-16
3	s(u)	2.07	2.07	3.18	3.47E-02
4	s(NO_2)	4.86	4.86	9.83	5.26E-09
5	s(NO)	3.87	3.87	24.84	< 2e-16
6	s(CO)	5.53	5.53	36.18	< 2e-16
	R-sq. (adj) = 0.487	AIC = 10,756		n = 1767	
	GAMM (Botsalano)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	daily max 8-h O_3 ~ s(T) + s(RH) + s(CO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	46.6921	0.60	77.38	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(T)	2.57	2.57	11.24	1.96E-06
2	s(RH)	1.00	1.00	22.14	3.28E-06
3	s(CO)	4.09	4.09	46.74	< 2e-16
	R-sq. (adj) = 0.522	AIC = 3013		n = 492	
	GAMM (Marikana)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	daily max 8-h O_3 ~ s(T) + s(RH) + s(NO_2) + s(NO) + s(CO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	51.36	1.91	26.89	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(T)	1	1.00	9.47	2.18E-03
2	s(RH)	1	1.00	19.228	1.36E-05
3	s(NO_2)	3.194	3.19	3.16	2.23E-02
4	s(NO)	6.452	6.45	12.06	1.64E-13
5	s(CO)	1	1.00	52.93	9.85E-13
	R-sq. (adj) = 0.352	AIC = 4327		n = 630	
	GAMM (Elandsfontein)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	daily max 8-h O_3 ~ s(T) + s(RH) + s(u) + s(NO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	48.444	1.47	32.94	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(T)	2.10	2.10	8.686	1.68E-04
2	s(RH)	1.00	1.00	10.033	1.62E-03
3	s(u)	4.15	4.15	3.323	1.60E-02
4	s(NO)	1.00	1.00	28.852	1.11E-07
	R-sq. (adj) = 0.180	AIC = 4449		n = 598	

To diagnose the nature of the relationships between O_3 and each of the independent variables, partial residual plots were examined (Figure 1). The partial residual plot of each independent variable, X_k, versus the smooth function, $s(X_k)$, shows the relationship between X_k and Y, given that the other independent variables are also included in the model. These residual plots indicate that, in the temperature range 20°C to 35°C, the relationship between daily max 8-h O_3 and T is positive and linear at Welgegund, Botsalano and Elandsfontein, while a change in slope is evident at lower temperatures. At Marikana, however, T is linearly and positively correlated for the entire T range. At all four sites, the change in O_3 with a change in relative humidity is linear and negatively correlated over the entire humidity range. For CO, the partial residual plot identified a positive linear relationship (although there is a small change in slope around 150–200 ppb for Welgegund and Botsalano) across the concentration range for Marikana. For NO and NO_2, there is sometimes a more complex (non-linear) fit in their partial residual response, suggesting other effects confounding with NO and NO_2.

3.5. Comparison of statistical models

In order to relate the statistical models utilized in this study, the differences between O_3 concentrations calculated with each model and measured O_3 levels (expressed as R^2 and RMSE) were compared and presented in Table 8. The factors obtained with PCA were also included in an MLR model to perform PCR, as indicated in section 2.3.2, which are presented in Table 8. Previous-day daily max 8-h O_3 was also included as an independent variable in the evaluation of these models in order to deal with the autocorrelation (persistence) in the data and to increase model performance (Comrie 1997), since it could also contribute to daily max 8-h O_3 (Otero et al. 2016). Previous-day daily max 8-h O_3 was not included in sections 3.2 to 3.4 where the influence of different independent variables on variances of O_3 was evaluated, since it could suppress the influence of other independent variables (Achen 2001). The complete statistics from each of the models are presented in Tables A1–A3 of the appendix. It is evident from Table 8 that inclusion of the previous-day daily max 8-h O_3 increases the performance of the MLR and GAM models, as reflected by the relative contribution to total explained variance (i.e. R^2 significantly increases). The results show that the O_3 concentrations calculated with non-parametric GAM compared slightly better to measured O_3 concentrations than O_3 levels calculated with MLR and PCR, as indicated by the highest R^2- and smallest RMSE values for GAM. However, less complicated MLR models are also suitable to evaluate contributions of factors to variances in O_3 levels. In addition, the inclusion of only previous-day daily max 8-h O_3, T, RH and CO in these statistical models explained approximately 70% of the variance in daily max 8-h O_3, which implies that these are the main factors influencing variations in O_3 concentrations in continental South Africa.

3.6. Insights into major factors driving O_3 variances

As indicated above, CO, RH and T were identified by all three statistical models as the major factors driving variances in O_3 levels in southern Africa. In many empirical and modelling studies, temperature is generally considered the most strongly correlated with O_3 concentrations (Jacob et al. 1993; Ryan 1995; Hubbard and Cobourn 1998; Baertsch-

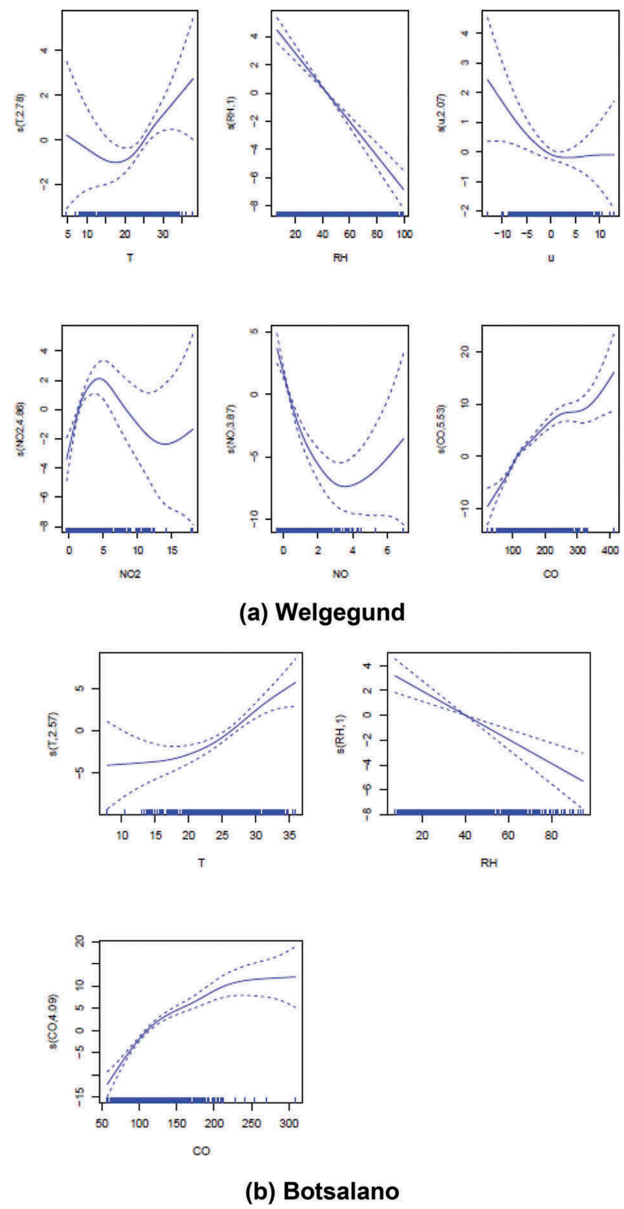

Figure 1. Partial residual plots of independent variables contained in the optimum solution from the GAM for O_3. The solid line in each plot is the estimate of the spline smooth function bounded by 95% confidence limits (i.e. ±2 standard errors of the estimate). The tick marks along the horizontal axis represent the density of data points of each explanatory variable (rug plot).

Ritter et al. 2004; Camalier et al. 2007; Dawson et al. 2007; Lin and Cobourn 2007; Cobourn 2007), which therefore has been used as a reasonable proxy to account for the combined influence of meteorological and chemical factors on O_3 concentrations (Jacob et al. 1993; Tsakiri and Zurbenko 2011; Rasmussen et al. 2012). High temperatures are usually associated with high solar radiation that contributes to increased photochemical reaction

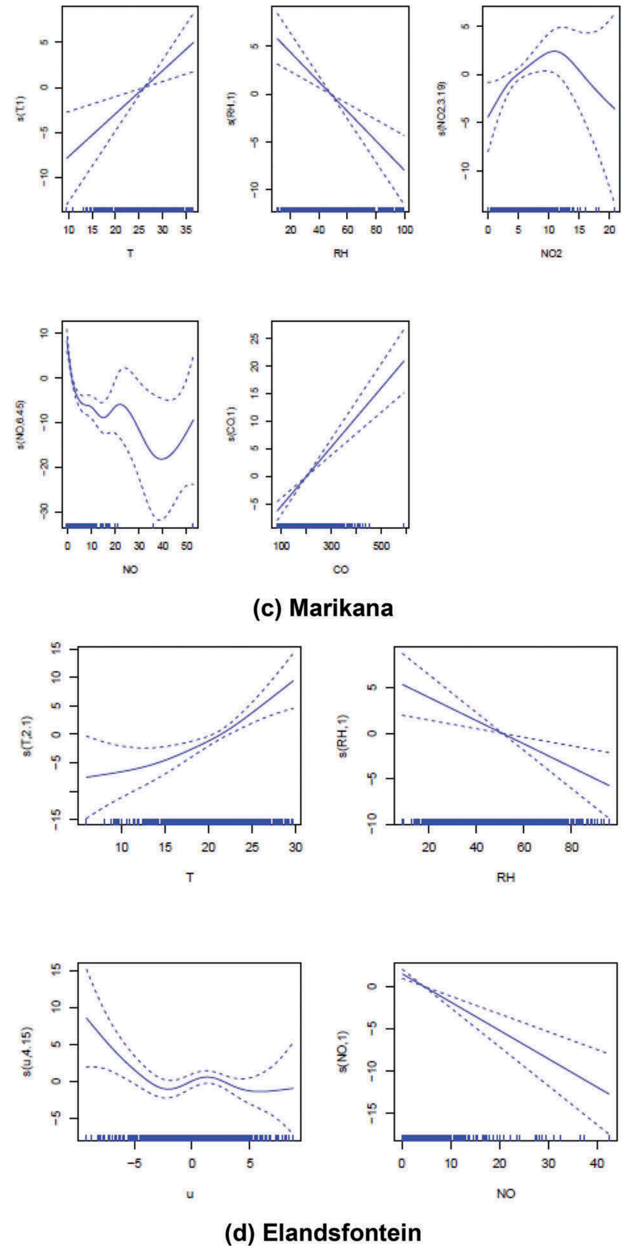

Figure 1. (Continued).

rates (Equation 1.1 and 1.2), as well as other meteorological conditions favouring O_3 production, such as high pressure, stagnation of air masses and reduced cloud cover (NRC 1991; Jacob et al. 1993). Jaars et al. (2014) also indicated that increased ambient VOC concentrations at Welgegund were associated with higher temperatures resulting from higher evaporation rates, which could also contribute to the increased O_3 formation potential of VOCs. The positive correlation between O_3 and temperature is also largely

Table 8. Comparison of statistical models in predicting daily max 8-h O_3 at the four measurement sites.

Measurement site	Method	Model	R^2	RMSE
WELGEGUND	MLR	daily max 8-h O_3 = 9.10 + 0.59*O_3-1 + 0.28*T − 0.10*RH − 0.21*u + 0.08*v − 1.44*NO + 0.07*CO	0.77	4.75
	PCR	daily max 8-h O_3 = −0.13 − 0.42*PC1 + 5.96*PC2 + 0.86*PC3 − 0.71*PC4	0.62	6.00
	GAM	daily max 8-h O_3 = 45.59 + s(O_3-1) + s(T) + s(RH) + s(u) + s(v) + s(NO_2) + s(NO) + s(CO)	0.79	4.47
BOTSALANO	MLR	daily max 8-h O_3 = −0.31 + 0.48*O_3-1 + 0.45*T + 0.005*Rad + 0.09*CO	0.70	5.14
	PCR	daily max 8-h O_3 = −0.25 − 4.88*PC1 − 0.09*PC2 + 0.07*PC3 − 2.18*PC4	0.64	5.63
	GAM	daily max 8-h O_3 = 46.75 + s(O_3-1) + s(T) + s(Rad) + s(u) + s(v) + s(CO)	0.73	4.69
MARIKANA	MLR	daily max 8-h O_3 = −18.19 + 0.73*O_3-1 + 0.48*T + 0.01*Rad + 0.70*v − 0.24*NO + 0.07*CO	0.83	6.93
	PCR	daily max 8-h O_3 = 0.01 + 4.15*PC1 − 3.73*PC2 + 1.01*PC3 − 9.93*PC4	0.77	8.04
	GAM	daily max 8-h O_3 = 51.09 + s(O_3-1) + s(T) + s(Rad) + s(u) + s(NO_2) + s(NO) + s(CO)	0.85	6.40
ELANDSFONTEIN	MLR	daily max 8-h O_3 = 18.45 + 0.68*O_3-1 + 0.32*T − 0.19*RH − 0.29*v + 0.15*NO_2 − 0.49*NO	0.67	9.03
	PCR	daily max 8-h O_3 = −0.31 − 0.38*PC1 − 2.31*PC2 − 1.44*PC3 + 10.36*PC4	0.61	9.88
	GAM	daily max 8-h O_3 = 48.81 + s(O_3-1) + s(T) + s(RH) + s(u) + s(NO_2) + s(NO)	0.69	8.64

driven by the chemical equilibrium between NO_x and peroxyacetylnitrate (PAN), which serves as a reservoir for NO_x (Jacob et al. 1993). The enhanced decomposition of PAN at high temperatures to regenerate stored NO_x results in local O_3 production being maximized (Jacob et al. 1993; Sillman and Samson 1995; Sillman 1999).

Some studies have indicated the significance of relative humidity to surface O_3 concentrations (Camalier et al. 2007; Davis et al. 2011; Awang et al. 2018). In the eastern United States, for instance, a north-south divide in terms of meteorological parameters controlling O_3 levels has been discussed in various studies (Camalier et al. 2007; Zheng et al. 2007; Davis et al. 2011; Rasmussen et al. 2012; Tawfik and Steiner 2013), with temperature most strongly correlated with O_3 at high latitude and strongly negatively correlated with relative humidity at lower latitude. This strong negative relationship between O_3 and relative humidity is not widely understood, with several authors presenting possible explanations:

- The O_3-relative humidity correlation is closely related to the O_3-temperature correlation, where temperature is the actual cause of O_3 variability, simultaneously affecting relative humidity and O_3 concentration (Camalier et al. 2007; Bloomer et al. 2009);
- High relative humidity can be associated with increased cloud cover and reduced UV radiation, which limits the photochemical production of O_3 to occur (Camalier et al. 2007; Davis et al. 2011; Porter et al. 2015);
- High relative humidity is associated with wet deposition (precipitation), which does not affect O_3 directly, but leads to the removal of soluble species such as HNO_3 and H_2O_2 and consequently the availability of NO_x and OH (Wild 2007). Furthermore, increased relative humidity increases the stomatal conductance of plants (Kavassalis and Murphy (2017)) and therefore also the dry deposition of surface O_3;
- Increased concentrations of atmospheric water vapour provide a chemical sink for O_3 through the reaction with water after photolysis, instead of the quenching reaction where O_3 is regenerated;

- Higher relative humidity can lead to more liquid water on aerosol particles, causing increased loss of gas phase NO_x via the heterogeneous reaction of dinitrogen pentoxide (N_2O_5) on particulates (Bertram and Thornton 2009). Jia and Xu (2014) also showed that increased relative humidity can greatly reduce O_3 through the transfer of NO_2- and ONO_2-containing species (reactive nitrogen species) to the particulate phase;
- Increased surface O_3 concentrations associated with stratospheric intrusions are associated with low water vapour (Thompson et al. 2014, 2015; Stauffer et al. 2017);
- O_3-relative humidity correlation can also result from a shift in the soil-moisture atmosphere coupling regime (evapotranspiration-limiting regimes), reflecting the simultaneous impact of soil moisture deficit on near-surface humidity, temperature and radiation (Tawfik and Steiner 2013).

All these afore-mentioned explanations could contribute to the significant (negative) correlation between O_3 variances and relative humidity observed for southern Africa. However, the relative role of temperature and relative humidity in driving O_3 variability is not yet fully disentangled due to their interdependency with the order of their significance possibly related to short-term dependencies, i.e. weather- and precursor emissions fluctuations. The significance of the influence of temperature and relative humidity on surface O_3 is also indicated by substantial higher O_3 concentrations measured during spring in 2015 at Welgegund. Dry and warm conditions were associated with the El Niño weather cycle, which persisted into the first half of 2016 with the 2015/2016 rain season being one of the warmest and driest in approximately 35 years.

The influence of CO on tropospheric O_3 formations is well known. CO and VOCs are the main sources of peroxy radicals that alter the PSS of O_3 production. Laban et al. (2018) indicated the important influence of CO on surface O_3 levels in southern Africa. CO emissions were attributed to household combustion for space heating and regional open biomass burning. Source maps indicated that O_3 and CO had similar regional sources with the highest concentrations of these species corresponding with the regions where a large number of wild fire events occurred. Furthermore, it was also indicated by Laban et al. (2018) that increased surface O_3 levels correlated with higher CO concentrations at Welgegund, Botsalano and Marikana, while it was implied that regional background regions in southern Africa could be considered VOC limited.

4. Conclusions

Three multivariate statistical models were utilized in order to provide some insights into major factors driving surface O_3 variability in continental southern Africa. Concentrations of precursors species and meteorological parameters measured at four sites located in the north-eastern interior of South Africa were included as input parameters. MLR indicated that CO, temperature and relative humidity made the largest contribution in explaining variances in daily max 8-h O_3. PCA indicated that parameters calculated with MLR are not strongly collinear and contributed independently to variances. Nonlinear GAM also revealed that CO, temperature and relative humidity were the most important parameters influencing variances in O_3 levels. Partial residual plots indicated that NO_x most likely have a non-linear relationship with O_3, while the relationship with temperature, relative humidity and CO is probably linear. Comparison of the measured O_3 concentrations with O_3 levels calculated with MLR and GAM indicated that O_3 levels calculated with both these models compared well to measured O_3 values, with GAM performing slightly better.

The influence of temperature on O_3 variability is expected, while Laban et al. (2018) indicated the significance of CO emissions associated with biomass burning on surface O_3 levels in southern Africa. The significant effect of relative humidity on O_3 variability, i.e. lower O_3 associated with increased relative humidity, was unexpected. Therefore, the influence of relative humidity should not be underestimated in atmospheric O_3 formation and prediction models.

In conjunction with variables utilized in this study, other synoptic-scale meteorological contributions to surface O_3 should also be investigated, e.g. large-scale atmospheric circulation over this region. It is also important that VOCs are included in statistical models. No continuous long-term VOC measurements were conducted at any of the sites. Although Jaars et al. (2014) and Jaars et al. (2016) did report on VOCs collected with grab samples during a two-year sampling campaign at Welgegund, this data was not from a statistical perspective considered sufficient to be included in the statistical models. Photochemical box models can also be used to investigate the main reactions that participate in O_3 formation. A greater scientific understanding of the factors influencing surface O_3 concentrations in South Africa will allow regional air quality models to be improved for the prediction of surface O_3 concentrations. It could be a step towards developing operational O_3 forecast models for cities and towns in South Africa.

Acknowledgments

V Vakkari is a beneficiary of an AXA Research Fund postdoctoral grant. The authors are also grateful to Eskom for supplying the Elandsfontein data., North-West University, Private Bag x6001, Potchefstroom 2520, South Africa.

Disclosure statement

The authors declare that they have no conflict of interest.

Funding

This work was partly funded by the Academy of Finland Centre of Excellence program [272041 and 307331] and the National Research Foundation of South Africa (grant numbers 97006 and 111287). Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

ORCID

Markku Kulmala http://orcid.org/0000-0003-3464-7825

Data availability

The data of this paper are available upon request to Pieter van Zyl (pieter.vanzyl@nwu.ac.za) or Johan Paul Beukes (paul.beukes@nwu.ac.za).

References

Abdul-Wahab SA, Bakheit CS, Al-Alawi SM. 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ Model Softw. 20(10):1263–1271. doi:10.1016/j.envsoft.2004.09.001.

Achen CH. 2001. Why lagged dependent variables can suppress the explanatory power of other independent variables. Ann Arbor. 1001:41248–48106.

Awang NR, Ramli NA, Shith S, Zainordin NS, Manogaran H. 2018. Transformational characteristics of ground-level ozone during high particulate events in urban area of Malaysia. Air Quality, Atmosphere & Health. 11(6):715–727. doi:10.1007/s11869-018-0578-0.

Awang NR, Ramli NA, Yahaya AS, Elbayoumi M. 2015. Multivariate methods to predict ground level ozone during daytime, nighttime, and critical conversion time in urban areas. Atmos Pollut Res. 6 (5):726–734. doi:10.5094/APR.2015.081.

Baertsch-Ritter N, Keller J, Dommen J, Prevot A. 2004. Effects of various meteorological conditions and spatial emission resolutions on the ozone concentration and ROG/NO_x limitation in the Milan area (I). Atmos Chem Phys. 4(2):423–438. doi:10.5194/acp-4-423-2004.

Bertram T, Thornton J. 2009. Toward a general parameterization of N_2O_5 reactivity on aqueous particles: the competing effects of particle liquid water, nitrate and chloride. Atmos Chem Phys. 9 (21):8351–8363. doi:10.5194/acp-9-8351-2009.

Bloomer BJ, Stehr JW, Piety CA, Salawitch RJ, Dickerson R. 2009. R.: observed relationships of ozone air pollution with temperature and emissions. Geophys Res Lett. 36(9). doi:10.1029/2009GL037308.

Bloomfield P, Royle JA, Steinberg LJ, Yang Q. 1996. Accounting for meteorological effects in measuring urban ozone levels and trends. Atmos Environ. 30(17):3067–3077. doi:10.1016/1352-2310(95)00347-9.

Camalier L, Cox W, Dolwick P. 2007. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. Atmos Environ. 41(33):7127–7137. doi:10.1016/j.atmosenv.2007.04.061.

Cardelino C, Chang M, John JS, Murphey B, Cordle J, Ballagas R, Patterson L, Powell K, Stogner J, Zimmer-Dauphinee S. 2001. Ozone predictions in Atlanta, Georgia: analysis of the 1999 ozone season. J Air Waste Manage Assoc. 51(8):1227–1236. doi:10.1080/10473289.2001.10464342.

Cobourn WG. 2007. Accuracy and reliability of an automated air quality forecast system for ozone in seven Kentucky metropolitan areas. Atmos Environ. 41(28):5863–5875. doi:10.1016/j.atmosenv.2007.03.024.

Combrink J, Diab R, Sokolic F, Brunke E. 1995. Relationship between surface, free tropospheric and total column ozone in two contrasting areas in South Africa. Atmos Environ. 29(6):685–691. doi:10.1016/1352-2310(94)00313-A.

Comrie AC. 1997. Comparing neural networks and regression models for ozone forecasting. Air Waste Manage. Assoc. 47(6):653–663. doi:10.1080/10473289.1997.10463925.

Cooper OR, Gao RS, Tarasick D, Leblanc T, Sweeney C. 2012. Long-term ozone trends at rural ozone monitoring sites across the United States, 1990–2010. J Geophys Res, 117: D22307. doi:10.1029/2012JD018261.

Davis J, Cox W, Reff A, Dolwick P. 2011. A comparison of CMAQ-based and observation-based statistical models relating ozone to meteorological parameters. Atmos Environ. 45 (20):3481–3487. doi:10.1016/j.atmosenv.2010.12.060.

Dawson JP, Adams PJ, Pandis SN. 2007. Sensitivity of ozone to summertime climate in the eastern USA: A modeling case study. Atmos Environ. 41(7):1494–1511. doi:10.1016/j.atmosenv.2006.10.033.

Diab R, Thompson A, Mari K, Ramsay L, Coetzee G. 2004. Tropospheric ozone climatology over Irene, South Africa, from 1990 to 1994 and 1998 to 2002. J Geophys Res. 109(D20). doi:10.1029/2004JD004793.

Dominick D, Juahir H, Latif MT, Zain SM, Aris AZ. 2012. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmos Environ. 60:172–181. doi:10.1016/j.atmosenv.2012.06.021.

Dueñas C, Fernández MC, Cañete S, Carretero J, Liger E. 2002. Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. Sci Total Environ. 299 (1–3):97–113. doi:10.1016/S0048-9697(02)00251-6.

Fiore AM, Jacob DJ, Logan JA, Yin JH. 1998. Long-term trends in ground level ozone over the contiguous United States, 1980–1995. J Geophys Res. 103(D1):1471–1480. doi:10.1029/97JD03036.

Gardner M, Dorling S. 2000. Meteorologically adjusted trends in UK daily maximum surface ozone concentrations. Atmos Environ. 34(2):171–176. doi:10.1016/S1352-2310(99)00315-5.

Gardner MW, Dorling S. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ. 32(14–15):2627–2636. doi:10.1016/S1352-2310(97)00447-0.

Gorai A, Tuluri F, Tchounwou P, Ambinakudige S. 2015. Influence of local meteorology and NO_2 conditions on ground-level ozone concentrations in the eastern part of Texas, USA, Air Quality. Atmos Health. 8(1):81–96. doi:10.1007/s11869-014-0276-5.

Guardani R, Aguiar JL, Nascimento CA, Lacava CI, Yanagi Y. 2003. Ground-level ozone mapping in large urban areas using multivariate statistical analysis: application to the Sao Paulo Metropolitan area. J Air Waste Manage Assoc. 53(5):553–559. doi:10.1080/10473289.2003.10466188.

Hastie T, Tibshirani R. 1990. Generalized additive models. Wiley Online Library.

Hayn M, Beirle S, Hamprecht FA, Platt U, Menze BH, Wagner T. 2009. Analysing spatio-temporal patterns of the global NO_2-distribution retrieved from GOME satellite observations using a generalized additive model. Atmos Chem Phys. 9(17):6459–6477. doi:10.5194/acp-9-6459-2009.

Hubbard MC, Cobourn WG. 1998. Development of a regression model to forecast ground-level ozone concentration in Louisville, KY. Atmos Environ. 32(14–15):2637–2647. doi:10.1016/S1352-2310(97)00444-5.

Jaars K, Beukes JP, van Zyl PG, Venter AD, Josipovic M, Pienaar JJ, Vakkari V, Aaltonen H, Laakso H, Kulmala M, et al. 2014. Ambient aromatic hydrocarbon measurements at Welgegund, South Africa. Atmos Chem Phys. 14(13):7075–7089. doi:10.5194/acp-14-7075-2014.

Jaars K, van Zyl PG, Beukes JP, Hellén H, Vakkari V, Josipovic M, Venter AD, Räsänen M, Knoetze L, Cilliers DP, et al. 2016. Measurements of biogenic volatile organic compounds at a grazed savannah grassland agricultural landscape in South Africa. Atmos Chem Phys. 16 (24):15665–15688. doi:10.5194/acp-16-15665-2016.

Jacob DJ, Logan JA, Gardner GM, Yevich RM, Spivakovsky CM, Wofsy SC, Sillman S, Prather MJ. 1993. Factors regulating ozone over the United States and its export to the global atmosphere. J Geophys Res. 98(D8):14817–14826. doi:10.1029/98JD01224.

Jia L, Xu Y. 2014. Effects of relative humidity on ozone and secondary organic aerosol formation from the photooxidation of benzene and ethylbenzene. Aerosol Sci and Tech. 48(1):1–12. doi:10.1080/02786826.2013.847269.

Kavassalis SC, Murphy JG. 2017. Understanding ozone-meteorology correlations: A role for dry deposition. Geophys Res Lett. 44(6):2922–2931. doi:10.1002/2016GL071791.

Laban TL, van Zyl PG, Beukes JP, Vakkari V, Jaars K, Borduas-Dedekind N, Josipovic M, Thompson AM, Kulmala M, Laakso L. 2018. Seasonal influences on surface ozone variability in continental South Africa and implications for air quality. Atmos Chem Phys Discuss. 18(20):15491–15514. doi:10.5194/acp-2017-1115.

Lin Y, Cobourn WG. 2007. Fuzzy system models combined with nonlinear regression for daily ground-level ozone predictions. Atmos Environ. 41(16):3502–3513. doi:10.1016/j.atmosenv.2006.11.060.

Lourens AS, Beukes JP, Van Zyl PG, Fourie GD, Burger JW, Pienaar JJ, Read CE, Jordaan JH. 2011. Spatial and temporal assessment of gaseous pollutants in the Highveld of South Africa. S Afr J Sci. 107(1/2):1–8. doi:10.4102/sajs.v107i1/2.269.

Lourens ASM, Butler TM, Beukes JP, Van Zyl PG, Beirle S, Wagner TK, Heue K-P, Pienaar JJ, Fourie GD, Lawrence MG. 2012. Re-evaluating the NO_2 hotspot over the South African Highveld. South African J Sci. doi:10.4102/sajs.v108i11/12.1146.

Melkonyan A, Kuttler W. 2012. Long-term analysis of NO, NO2 and O3 concentrations in North Rhine-Westphalia, Germany. Atmos Environ. 60:316–326. doi:10.1016/j.atmosenv.2012.06.048.

Mikkonen S, Korhonen H, Romakkaniemi S, Smith JN, Joutsensaari J, Lehtinen KEJ, Hamed A, Breider TJ, Birmili W, Spindler G, et al. 2011. Meteorological and trace gas factors affecting the number concentration of atmospheric Aitken (D_p= 50 nm) particles in the continental boundary layer: $_p$parameterization using a multivariate mixed effects model. Geosci Model Dev. 4(1):1–13. doi:10.5194/gmd-4-1-2011.

Milanchus ML, Rao ST, Zurbenko IG. 1998. Evaluating the effectiveness of ozone management efforts in the presence of meteorological variability. J Air Waste Manage Assoc. 48(3):201–215. doi:10.1080/10473289.1998.10463673.

Neter J, Kutner M, Nachtsheim C, Wasserman W. 1996. Applied linear statistical models. 4th ed. New York: McGraw-Hill; p. 283.

NRC. 1991. Rethinking the ozone problem in urban and regional air pollution. Washington (DC): The National Academies Press; p. 524.

NRC. 2008. Estimating mortality risk reduction and economic benefits from controlling ozone air pollution. Washington, DC: National Academies Press. https://doi.org/10.17226/12198.

Ooka R, Khiem M, Hayami H, Yoshikado H, Huang H, Kawamoto Y. 2011. Influence of meteorological conditions on summer ozone levels in the central Kanto area of Japan. Procedia Environ. Sci. 4:138–150. doi:10.1016/j.proenv.2011.03.017.

Ordonez C, Mathis H, Furger M, Henne S, Hüglin C, Staehelin J, Prévôt A. 2005. Changes of daily surface ozone maxima in Switzerland in all seasons from 1992 to 2002 and discussion of summer 2003. Atmos Chem Phys. 5(5):1187–1203. doi:10.5194/acp-5-1187-2005.

Otero N, Sillmann J, Schnell JL, Rust HW, Butler T. 2016. Synoptic and meteorological drivers of extreme ozone concentrations over Europe. Environ Res Lett. 11(2):024005. doi:10.1088/1748-9326/11/2/024005.

Pearce JL, Beringer J, Nicholls N, Hyndman RJ, Tapper NJ. 2011. Quantifying the influence of local meteorology on air quality using generalized additive models. Atmos Environ. 45(6):1328–1336. doi:10.1016/j.atmosenv.2010.11.051.

Porter WC, Heald CL, Cooley D, Russell B. 2015. Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression. Atmos Chem Phys. 15 (18):10349–10366. doi:10.5194/acp-15-10349-2015.

R Development Core Team. 2009. R: A language and environment for statistical computing. [accessed 2018 Mar 12]. http://www.R-project.org.

Rao ST, Zurbenko IG. 1994. Detecting and tracking changes in ozone air quality. Air Waste. 44 (9):1089–1092. doi:10.1080/10473289.1994.10467303.

Rasmussen D, Fiore A, Naik V, Horowitz L, McGinnis S, Schultz M. 2012. Surface ozone-temperature relationships in the eastern US: A monthly climatology for evaluating chemistry-climate models. Atmos Environ. 47:142–153. doi:10.1016/j.atmosenv.2011.11.021.

Ryan WF. 1995. Forecasting severe ozone episodes in the Baltimore metropolitan area. Atmos Environ. 29(17):2387–2398. doi:10.1016/1352-2310(94)00302-2.

Schlink U, Herbarth O, Richter M, Dorling S, Nunnari G, Cawley G, Pelikan E. 2006. Statistical models to assess the health effects and to forecast ground-level ozone. Environ Model Softw. 21 (4):547–558. doi:10.1016/j.envsoft.2004.12.002.

Seinfeld JH, Pandis SN. 2006. Atmospheric chemistry and physics: from air pollution to climate change. Vol. xxviii, 2nd ed. New York: Wiley. p. 1202

Sillman S. 1999. The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. Atmos Environ. 33(12):1821–1845. doi:10.1016/S1352-2310(98)00345-8.

Sillman S, Samson PJ. 1995. Impact of temperature on oxidant photochemistry in urban, polluted rural and remote environments. J Geophys Res Atmospheres. 100(D6):11497–11508. doi:10.1029/94JD02146.
Stauffer RM, Thompson AM, Oltmans SJ, Johnson BJ. 2017. Tropospheric ozonesonde profiles at long-term US monitoring sites: 2. Links between Trinidad Head, CA, profile clusters and inland surface ozone measurements. J Geophys Res. 122:1261–1280.
Tawfik AB, Steiner AL. 2013. A proposed physical mechanism for ozone-meteorology correlations using land–atmosphere coupling regimes. Atmos Environ. 72:50–59. doi:10.1016/j.atmosenv.2013.03.002.
Thompson AM, Balashov NV, Witte JC, Coetzee JGR, Thouret V, Posny F. 2014. Tropospheric ozone increases over the southern Africa region: bellwether for rapid growth in Southern Hemisphere pollution? Atmos Chem Phys. 14(18):9855–9869. doi:10.5194/acp-14-9855-2014.
Thompson AM, Stauffer RM, Miller SK, Martins DK, Joseph E, Weinheimer AJ, Diskin GS. 2015. Ozone profiles in the Baltimore-Washington region (2006–2011): satellite comparisons and DISCOVER-AQ observations. J Atmos Chem. 72(3–4):393–422. doi:10.1007/s10874-014-9283-z.
Thompson ML, Reynolds J, Cox LH, Guttorp P, Sampson PD. 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmos Environ. 35(3):617–630. doi:10.1016/S1352-2310(00)00261-2.
Tiitta P, Vakkari V, Croteau P, Beukes JP, van Zyl PG, Josipovic M, Venter AD, Jaars K, Pienaar JJ, Ng NL, et al. 2014. Chemical composition, main sources and temporal variability of PM_1 aerosols in southern African grassland. Atmos Chem Phys. 14(4):1909–1927. doi:10.5194/acp-14-1909-2014.
Tsakiri KG, Zurbenko IG. 2011. Prediction of ozone concentrations using atmospheric variables. Air Qual Atmos Health. 4(2):111–120. doi:10.1007/s11869-010-0084-5.
Vakkari V, Beukes JP, Laakso H, Mabaso D, Pienaar JJ, Kulmala M, Laakso L. 2013. Long-term observations of aerosol size distributions in semi-clean and polluted savannah in South Africa. Atmos Chem Phys. 13(4):1751–1770. doi:10.5194/acp-13-1751-2013.
Venter AD, Vakkari V, Beukes JP, Van Zyl PG, Laakso H, Mabaso D, Tiitta P, Josipovic M, Kulmala M, Pienaar JJ, et al. 2012. An air quality assessment in the industrialised western Bushveld Igneous Complex, South Africa. S Afr J Sci. 108(9/10). doi:10.4102/sajs.v108i9/10.1059.
Wild O. 2007. Modelling the global tropospheric ozone budget: exploring the variability in current models. Atmos Chem Phys. 7(10):2643–2660. doi:10.5194/acp-7-2643-2007.
Wood SN. 2017. Generalized additive models: an introduction with R. Boca Raton, FL: CRC press.
Zheng J, Swall JL, Cox WM, Davis JM. 2007. Interannual variation in meteorologically adjusted ozone levels in the eastern United States: A comparison of two approaches. Atmos Environ. 41(4):705–716. doi:10.1016/j.atmosenv.2006.09.010.
Zunckel M, Venjonoka K, Pienaar JJ, Brunke EG, Pretorius O, Koosialee A, Raghunandan A, van Tienhoven AM. 2004. Surface ozone over southern Africa: synthesis of monitoring results during the cross border air pollution impact assessment project. Atmos Environ. 38:6139–6147. doi:10.1016/j.atmosenv.2004.07.029

Appendix

Table A1. MLR models for prediction of daily max 8-h O_3 for each measurement site.

WELGEGUND	Constant	O_3-1 (ppb)	T (°C)	RH (%)	u (m/s)	v (m/s)	NO (ppb)	CO (ppb)
Regression coefficient (β)	9.10	0.59	0.28	−0.10	−0.21	0.08	−1.44	0.07
Standard error	0.95	0.01	0.02	0.01	0.04	0.05	0.19	0.00
t-statistic	9.57	42.36	11.23	−13.26	−5.39	1.83	−7.54	20.36
P-value	3.56E-21	1.16E-270	2.65E-28	2.57E-38	7.99E-08	0.06747224	7.77E-14	6.94E-83
R^2 = 0.768	**Adjusted R^2 = 0.767**			**RMSE = 4.75**		**F-statistic = 828**		
BOTSALANO	**Constant**	**O_3-1 (ppb)**	**T (°C)**	**Rad (W/m^2)**	**CO (ppb)**			
Regression coefficient (β)	−0.31	0.48	0.45	0.01	0.09			
Standard error	1.51	0.03	0.07	0.00	0.01			
t-statistic	−0.20	16.23	6.78	2.09	11.94			
P-value	0.83958521	1.94E-47	3.65E-11	0.03712663	6.49E-29			
R^2 = 0.695	**Adjusted R^2 = 0.693**			**RMSE = 5.14**		**F-statistic = 271**		
MARIKANA	**Constant**	**O_3-1 (ppb)**	**T (°C)**	**Rad (W/m^2)**	**v (m/s)**	**NO (ppb)**	**CO (ppb)**	
Regression coefficient (β)	−18.19	0.73	0.48	0.01	0.70	−0.24	0.07	
Standard error	1.91	0.02	0.09	0.00	0.21	0.08	0.00	
t-statistic	−9.55	39.12	5.08	4.07	3.40	−2.91	14.46	
P-value	3.16E-20	8.56E-169	4.96E-07	5.36E-05	0.00070864	0.00369032	5.48E-41	
R^2 = 0.830	**Adjusted R^2 = 0.828**			**RMSE = 6.93**		**F-statistic = 498**		
ELANDSFONTEIN	**Constant**	**O_3-1 (ppb)**	**T (°C)**	**RH (%)**	**v (m/s)**	**NO_2 (ppb)**	**NO (ppb)**	
Regression coefficient (β)	18.45	0.69	0.32	−0.19	−0.30	0.15	−0.50	
Standard error	3.16	0.03	0.09	0.02	0.16	0.05	0.09	
t-statistic	5.84	26.31	3.47	−8.32	−1.83	2.76	−5.28	
P-value	8.63E-09	2.92E-100	0.00055166	6.48E-16	0.06774829	0.00593864	1.87E-07	
R^2 = 0.672	**Adjusted R^2 = 0.669**			**RMSE = 9.04**		**F-statistic = 193**		

Table A2. PCR models for prediction of daily max 8-h O_3 for each measurement site.

Welgegund	Constant	PC1	PC2	PC3	PC4
Regression coefficient	−0.13	−0.42	5.96	0.86	−0.71
R^2	0.62				
F-statistic	444				
P-value	1.63E-226				
Estimate of error variance (MSE)	35.98				
RMSE	6.00				
Botsalano	**Constant**	**PC1**	**PC2**	**PC3**	**PC4**
Regression coefficient	−0.25	−4.88	−0.09	0.07	−2.18
R^2	0.64				
F-statistic	191				
P-value	2.75E-94				
Estimate of error variance (MSE)	31.67				
RMSE	5.63				
Marikana	**Constant**	**PC1**	**PC2**	**PC3**	**PC4**
Regression coefficient	0.01	4.15	−3.73	1.01	−9.93
R^2	0.77				
F-statistic	516				
P-value	8.94E-195				
Estimate of error variance (MSE)	64.69				
RMSE	8.04				
Elandsfontein	**Constant**	**PC1**	**PC2**	**PC3**	**PC4**
Regression coefficient	−0.31	−0.38	−2.31	−1.44	10.36
R^2	0.61				
F-statistic	217				
P-value	1.53E-112				
Estimate of error variance (MSE)	97.66				
RMSE	9.88				

Table A3. GAMs for prediction of daily max 8-h O_3 for each measurement site: includes tests for each smooth, the degrees of freedom for each smooth, adjusted R-squared for the model and deviance for the model.

	GAM (Welgegund)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	O_3 ~ s(O_3-1) + s(T) + s(RH) + s(u) + s(v) + s(NO_2) + s(NO) + s(CO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	45.5907	0.1075	423.9	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(O_3-1)	6.782	7.943	220.022	< 2e-16
2	s(T)	6.277	7.493	16.296	< 2e-16
3	s(RH)	4.422	5.485	40.596	< 2e-16
4	s(u)	2.933	3.77	3.683	7.76E-03
5	s(v)	2.204	2.86	3.395	2.68E-02
6	s(NO_2)	4.006	4.993	9.407	7.52E-09
7	s(NO)	2.532	3.248	20.323	1.76E-13
8	s(CO)	8.323	8.877	36.994	< 2e-16
	R-sq. (adj) = 0.79	Deviance explained = 79.3%			
	GCV score = 20.85	Scale est. = 20.39		n = 1763	
	AIC score = 10,349	RMSE = 4.47			
	GAM (Botsalano)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	O_3 ~ s(O_3-1) + s(T) + s(Rad) + s(u) + s(v) + s(CO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	46.7474	0.2187	213.7	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(O_3-1)	3.088	3.912	66.936	< 2e-16
2	s(T)	1	1	26.666	3.56E-07
3	s(Rad)	2.07	2.641	5.668	1.89E-03
4	s(u)	4.829	5.965	2.677	1.49E-02
5	s(v)	3.733	4.743	2.443	3.60E-02
6	s(CO)	4.332	5.396	35.054	< 2e-16
	R-sq. (adj) = 0.73	Deviance explained = 74.3%			
	GCV score = 23.96	Scale est. = 22.96		n = 480	
	AIC score = 28,888	RMSE = 4.69			
	GAM (Marikana)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	O_3 ~ s(O_3-1) + s(T) + s(Rad) + s(u) + s(NO_2) + s(NO) + s(CO)				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	51.0886	0.2629	194.3	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	s(O_3-1)	4.12	5.137	305.399	< 2e-16
2	s(T)	1	1	18.271	2.22E-05
3	s(Rad)	1	1	29.151	9.58E-08
4	s(u)	4.441	5.541	3.117	6.80E-03
5	s(NO_2)	5.213	6.323	3.46	2.25E-03
6	s(NO)	7.72	8.536	4.481	2.10E-05
7	s(CO)	3.619	4.574	23.334	< 2e-16
	R-sq. (adj) = 0.85	Deviance explained = 85.4%			
	GCV score = 44.90	Scale est. = 42.86		n = 620	
	AIC score = 4119	RMSE = 6.39			

(*Continued*)

Table A3. (Continued).

	GAM (Welgegund)				
	GAM (Elandsfontein)				
	Family: Gaussian				
	Link function: identity				
	Formula:				
	$O_3 \sim s(O_3\text{-}1) + s(T) + s(RH) + s(u) + s(NO_2) + s(NO)$				
	Parametric coefficients:				
	term	Estimate	Std. Error	t value	Pr(> I t I)
	(Intercept)	48.8052	0.3657	133.5	<2e-16
	Approximate significance of smooth terms:				
	term	edf	Ref.df	F	p-value
1	$s(O_3\text{-}1)$	1.664	2.1	341.565	< 2e-16
2	s(T)	2.298	2.923	4.403	5.54E-03
3	s(RH)	1	1	61.999	1.58E-14
4	s(u)	4.759	5.871	5.371	3.04E-05
5	$s(NO_2)$	1	1	4.94	2.66E-02
6	s(NO)	2.41	3.026	10.294	1.20E-06
	R-sq. (adj) = 0.69	Deviance explained = 69.6%			
	GCV score = 78.56	Scale est. = 76.62		n = 573	
	AIC score = 4128	RMSE = 8.64			