United States
Environmental Protection
Agency

Atmospheric Research and Exposure
Assessment Laboratory
Research Triangle Park, NC 27711

Research and Development

EPA/600/SR-94/081 June 1994

EPA

Project Summary

On the Feasibility of Using Satellite Derived Data to Infer Surface-Layer Ozone Concentration Patterns

Brian K. Eder

Principal Component Analysis (PCA) was applied to six years (1985-1990) of surface and satellite ozone (O_3) data collected over the eastern United States to determine whether O_3 measurements derived from satellites could be used to infer surface-layer concentrations. Examination of the spatial and temporal characteristics associated with the first nonrotated principal components (which are the dominant components, explaining 37.95% and 41.25% of the total variance of the surface and satellite data sets, respectively) revealed considerable coherence between the data sets, suggesting that on continental-scales, seasonal 0_3 patterns derived from the satellite data replicate quite well those of the surface. This coherence diminishes, however, when daily patterns are compared. Upon orthogonal rotation, the PCA delineated four contiguous and statistically unique subregions with each data set (the *Northwest, Northeast, Southwest,* and *Southeast)* that were very similar, suggesting that the satellite data may be able to discern 0_3 patterns on spatial scales as small as 1000 km. The temporal characteristics associated with the *Southwest* and *Southeast* subregions exhibited cross-data set similarities; however, those associated with the *Northwest* and *Northeast* subregions were somewhat dissimilar.

This Project Summary was developed by EPA's Atmospheric Research and Exposure Assessment Laboratory, Research Triangle Park, NC, to announce key findings of the research project that is fully documented in a separate report of the same title (see Project Report ordering information at back).

Introduction

Traditionally, ozone (0_3) has been characterized as an urban-scale pollutant. Increasingly, however, it has been recognized by scientists as a regional and even global-scale phenomenon, as high concentrations are routinely observed over vast, non-urban areas of most industrialized countries, where forest retardation and crop injury are becoming growing environmental concerns. Daily maximum 0_3 concentrations in these areas are often comparable to those found in urban areas, and daily average concentrations can even exceed urban concentrations due to a lack of nitric oxide (NO) scavenging. Coinciding with these realizations has been the advent of satellite-derived 0_3 measurements. Using data derived from the Total 0zone Mapping Spectrometer (TOMS), which measures total column 0_3 concentrations, and the Stratospheric Aerosol and Gas Experiment (SAGE), which measures the stratospheric 0_3 concentration, scientists have been able to estimate the tropospheric (residual) concentration and, under certain meteorological conditions, estimate the surface-layer 0_3 concentration.

The appropriateness of this remote sensing approach will be evaluated using Principal Component Analysis (PCA) as applied to surface data obtained from EPA's Aerometric Information and Retrieval System (AIRS) and residual 0_3 data from the National Satellite Service Data Center (NSSDC) at NASA's Goddard Space Flight Center. This analysis, which employs data from the six-year period

985-1990, will enable us to determine to hat extent, if any, the major modes of atial and temporal surface O_3 variability e being captured by the satellite data. he advantages of employing a technique ch as PCA are numerous. First, we are aling with widely varying spatial scales hich prohibit point-by-point comparisons. he surface data obtained from the AIRS twork is representative of meso-scales - 100 km), while the satellite data is presentative of macro-scales (≥ 100 km). CA circumvents this impediment by proding similar spatial scale results that will ow for pattern recognition and compari-n of O_3 concentrations. Second, because the copious amount of data resulting om such a large-scale study (nearly ,000 surface observations and over ,000 satellite observations) and because e individual data tend to be erratic or isy, it is advantageous to employ an alysis technique that identifies, through reduction of data, the recurring and inpendent modes of variation within the rger data sets. And finally, the analysis O_3 characteristics and trends between e data sets is based on an aggregation data from many stations (grid cells), as posed to individual stations (grid cells), inimizing the effects of anomalous or en erroneous data often associated with single observation.

ata

The surface O_3 concentration (ppb) data nployed in this analysis were obtained om AIRS, which operates under strict onitoring criteria including multipoint calitions, independent audits, and data valiation based upon frequent zero, span, d precision checks. The O_3 measureents were made during the "ozone seaon" (June 15th through October 31st for is study) using either chemiluminescence alyzers, which are sensitive to light emitd by the reaction between O_3 and ethyle, or ultraviolet photometers, which meare the absorption of light by O_3.

A major goal of this study was to estabh a complete, regionally-representative rface O_3 data base, unencumbered by issing data or local-scale variability. Atinment of this goal was achieved by sing many selection criteria. First, the ily 1-hour maximum concentration was sed to help minimize local-scale variabil- because at the time of maximum surce concentration (typically between 1 d 3 pm LST), the boundary layer is nerally uniformly mixed and the surface ncentration is therefore most represenive of the boundary layer concentration. dditionally, both the primary standard, degned to protect human health, and the secondary standard, designed to protect human welfare, established by EPA as part of the NAAQS, are based on the daily 1-hour maximum concentration. Second, to avoid NO scavenging effects found in close proximity to urban areas, only those stations classified as either rural or suburban and reporting a land use of either forest, agriculture, or residential were employed in the analysis. Rural stations received highest priority; however, to meet our third criteria, spatial completeness, several suburban stations were also included in the data base. Finally, only those stations reporting a capture rate of 90.0% or better for the study period were considered. These criteria resulted in the inclusion of 77 stations across the eastern half of the United States, the majority of which (55) are classified as rural. Several combinations of "station-seasons" were examined before the optimum period of 1985-1990 was selected. The total capture rate for the period was > 95.0%. All missing data were replaced using a linear interpolation scheme, across time.

The TOMS and SAGE O_3 data, measured in Dobson Units (DU) (1 DU = 2.69 x 10^{16} molecules of O_3 cm^{-2}), were obtained from the archived data sets available at the NSSDC located at the Goddard Space Flight Center. The column of O_3 in the troposphere, or the "residual O_3," is determined by subtracting the integrated amount of O_3 above the tropopause derived from the SAGE profiles from the concurrent amount of total O_3 observed from the TOMS measurements. These data, which cover the eastern half of the United States for this study, are gridded with a resolution of 5° longitude by 2.5° latitude, resulting in a total of 54 grid cells.

Methodology

One of the main objectives of PCA is to identify, through a reduction of data, the recurring and independent modes of variation (signals) within large, noisy data sets, thereby summarizing the essential information of the data sets so that meaningful and descriptive conclusions can be made. The analysis sorts initially correlated data into a hierarchy of statistically independent modes of variations which explain successively less and less of the total variance.

The PCA of both the AIRS and TOMS-SAGE residuals data sets began with the extraction of square, symmetrical correlation matrices (**R**), having dimensions ${}_{77}\mathbf{R}_{77}$ (77 rows and 77 columns) for the AIRS data and ${}_{54}\mathbf{R}_{54}$ (54 rows and 54 columns) for the satellite data, from their original data matrices having dimensions of 77 [stations] x 834 [days] (or 64,218 observations) and 54 [grid cells] x 834 [days] (or 45,036 observations), respectively. By using **R** and the identity matrix (**I**), of the same dimensions, n = 77 (54) eigenvectors can be derived that represent the mutually orthogonal linear combinations (modes of variation) of the matrix. Their associated eigenvalues represent the amount of total variance that is explained by each of the eigenvectors. By retaining only the first few eigenvector-eigenvalue pairs, or principal components, a substantial amount of the total variance can be explained while ignoring the higher order principal components that explain minimal amounts of the total variance and can therefore be viewed as noise. The exact number of components that should be retained was determined through the use of the Scree Test and suggested four component solutions for both the AIRS and Satellite data sets; subsequently the first four principal components were retained for analysis and comparison.

When the elements of each eigenvector are multiplied by the square root of the associated eigenvalue, one obtains the *principal component loading (L)*, which represents the correlation between the component and the station or grid cell. When these principal component loadings are spatially mapped onto their respective stations (grid cells) for each component, isopleths of component loadings can be drawn, which identify the major modes of spatial variability.

Initially, the principal component analysis replaced 77 stations (54 grid cells), measured over 834 days, with four principal components having no temporal measure. By introducing the *principal component score* (PC_s), a derivation of similar temporal measurement for the principal components over the same 834 days can be achieved. The principal components are identified in terms of the original stations (grid cells), the larger the loading, the more important the station is in the interpretation of the component. Therefore, if a day has high values for the stations with large loadings, then it should have a large value on that component. The scores have been standardized; therefore, they have a mean of zero and a standard deviation of one.

Results

Examination of the spatial characteristics associated with the first nonrotated principal components (which are the dominant components, explaining 37.95% and 41.25% of the total variance of the surface and satellite data sets, respectively) revealed considerable coherence between the data sets. With only one minor excep-

tion (that being southern Florida in the surface data set), each spatial pattern revealed an in-phase oscillation with the area of greatest variance centered around the Ohio River Valley (the centers of maximum variance are within 200 km of each other). This would suggest that on continental scales, spatial patterns derived from the satellite mirror those of the surface.

Inspection of the seasonal time series (as defined by the smoothed median of the six years of daily principal component scores) associated with these dominant components also revealed considerable coherence. With both data sets, the highest 0_3 concentrations consistently occur during the period from June 15 through the middle of August. The transition to low concentrations occurs slightly later (by roughly a week) in the satellite time series and is not quite as sharp as the transition of the surface data set. The correlation coefficient between these seasonal data sets is high (0.75), indicating that on a seasonal scale at least, the satellite data could be used to infer surface-layer concentrations.

Examination of the daily principal component scores associated with these two dominant components was less encouraging, however, as marked differences were revealed. Surface concentrations during 1988 were by far the highest of any year during the study; however, this year does not stand out in the satellite data as being unusually high. In fact, the satellite data indicates that 1990, not 1988, was the year experiencing the highest concentrations. This and other discrepancies are reflected in the correlation coefficient between the daily principal component scores of the two data sets, which is quite low (0.33). Although statistically significant ($\alpha = 0.0001$), this value does not strongly support the use of the satellite data as a surrogate for surface-layer concentrations on a daily basis.

Another goal of this study was to identify, within each data set, areas of homogeneous 0_3 concentrations. This delineation, which was achieved through an orthogonal rotation of the original principal components, revealed four contiguous and statistically unique subregions within each data set. These subregions were quite similar, suggesting that the satellite-derived measurements could be useful on spatial scales as small as 1000 km. The first rotated component of the satellite data set encompasses the *Southeast* part of the study domain. This area corresponds well with the third rotated component of the surface data (note the ordering of the components is insignificant). The only major difference between these subregions is the extension of the satellite's *Southeast* subregion into southern Texas. The second rotated component of the satellite data set encompasses the *Northeast* states from Northern Virginia and Ohio east and northward. This region corresponds quite well with the second rotated component of the surface data (with the exception of its inclusion of northern Ohio and eastern Michigan). The third rotated component of the satellite data corresponds fairly well with the fourth rotated surface component, although there are differences. The satellite's *Southwest* subregion extends further into the Tennessee Valley, and, as mentioned above, does not include southern Texas. And finally, the satellite's fourth rotated component encompasses the *Northwest* portion of the domain, which corresponds well with the surface *Northwest* subregion, although it is somewhat smaller.

The seasonal time series of the data sets associated with two (the Southwest and the Southeast) of the subregions were quite similar. The surface and the satellite-derived time series associated with both Southwest Subregions exhibited minimal seasonal variation, while the time series associated with the Southeast Subregions exhibited a strong seasonality (though somewhat less pronounced in the surface data set). The transition from high to low 0_3 concentrations occurs at roughly the same time (end of August) for each of these Southeast Subregions.

Less similar are the time series associated with the Northeast and Northwest Subregions. The seasonal trend in the satellite's Northeast Subregion is weaker and its transition from high to low concentrations is delayed when compared to the surface's Northeast Subregion. More noticeable is the difference between the satellite and the surface Northwest Subregions. While the surface data indicate a more typical seasonal trend, the data derived from the satellite indicate an abnormal seasonality in which low concentrations occur during the first month of the season and again during the last month, with the remainder of the period experiencing higher concentrations. These discrepancies may be attributable to problems inherent in the satellite 0_3 data, where interference associated with the lowering of the tropopause that occurs in proximity of the jet stream.

Recommendations

The accuracy of the 0_3 data derived from the TOMS-SAGE residuals may be limited due to the inherently poor temporal and spatial resolution of the SAGE system. These limitations, which are still being explored, may be circumvented through use of data obtained from the Solar Backscattered Ultraviolet (SBUV) instrument. These data, which should become available in the near future, are collected on spatial and temporal scales more comparable to data collected by the TOMS instrument. Accordingly, any additional research should use this TOMS-SBUV residual database when it becomes available.

☆U.S. GOVERNMENT PRINTING OFFICE: 1994 - 550-067/80263

The EPA author, ***Brian K. Eder*** *(also the EPA Project Officer, see below),* is on assignment to the Atmospheric Research and Exposure Assessment Laboratory, Research Triangle Park, NC 27711, from the National Oceanic and Atmospheric Administration.

The complete report, entitled "On the Feasibility of Using Satellite Derived Data to Infer Surface-Layer Ozone Concentration Patterns," (Order No. PB94-170263; Cost: $17.50, subject to change) will be available only from:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: 703-487-4650

The EPA Project Officer can be contacted at:

Atmospheric Research and Exposure Assessment Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

nited States
nvironmental Protection Agency
enter for Environmental Research Information
incinnati, OH 45268

ficial Business
enalty for Private Use $300

PA/600/SR-94/081