Atmospheric Environment 81 (2013) 1–10

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Statistical downscaling of an air quality model using Fitted Empirical Orthogonal Functions

Farha A. Alkuwari [a,*], Serge Guillas [a], Yuhang Wang [b]

[a] Department of Statistical Science, University College London, UK
[b] School of Earth and Atmospheric Sciences, Georgia Institute of Technology, USA

HIGHLIGHTS

- We present a downscaling method using Fitted Empirical Orthogonal Functions (F-EOFs).
- We illustrate our downscaling method, for ozone levels over the US.
- We compare our method to linear and Principal Components (PC) regression methods.
- F-EOFs regression shows the best predictive ability compared to the other methods.
- F-EOFs outperform PC as a dimension reduction technique.

ARTICLE INFO

Article history:
Received 24 April 2013
Received in revised form
16 August 2013
Accepted 19 August 2013

Keywords:
Principal Fitted Components
Principal Components
Dimension reduction
Empirical Orthogonal Functions
Downscaling

ABSTRACT

Downscaling is a technique that is used to extract high-resolution information from regional scale variables produced by coarse resolution models such as Chemical Transport Models (CTMs). Statistical downscaling methods in geophysics often rely on Empirical Orthogonal Functions (EOFs). EOFs are spatial Principal Components (PCs) that display space-time modes of variability of a quantity over a region. Here we present a novel statistical downscaling method that employs Fitted Empirical Orthogonal Functions (F-EOFs) to provide local forecasts. F-EOFs differ from EOFs in that they represent space-time variations associated with a particular location through the use of inverse regression. We illustrate our downscaling method, for ozone levels over the US, with the Regional chEmical trAnsport Model (REAM) whose outputs are over 70 by 70 km grid cells. We use ground level ozone observations from monitoring stations within the south-eastern US region to downscale REAM. We select the first leading F-EOFs and regress our observations on the corresponding F-EOF loadings. We also compare our results to linear regression and PC regression. The regression on F-EOFs shows the best predictive ability. To examine the consistency of our results we repeat the analysis for different fitting and validation periods. Furthermore, in our application, PFC regression also outperforms PC regression as a dimension reduction technique.

1. Introduction

Ground-level ozone (O_3) is formed by a chemical reaction between volatile organic compounds (VOC) – such as automobile exhaust –, carbon monoxide (CO), and oxides of nitrogen (NO_x) with the existence of sunlight. Studies have shown that the exposure to elevated levels of ozone has major effects on human health (e.g. respiratory system problems) and vegetation (EPA, 2006). Ground-level ozone is a pollutant and is regulated under the National Ambient Air Quality Standards not to exceed 0.075 parts per million (ppm) as a measure over an 8-h average period in the USA. To manage the levels of tropospheric ozone, accurate spatial and temporal forecasts of its levels are needed.

Various statistical techniques have been used in the literature to model and forecast ground-level ozone, e.g. Sahu et al. (2007). Alternatively, numerical models were developed to produce forecasts of ozone and other atmospheric variables using meteorological information. Those models are known as Chemical Transport Models (CTMs). Although statistical methods were found to be satisfactory in predicting ozone, they do not capture the chemical and physical processes as well as CTMs. In general, CTMs produce simulations of climatological variables at large resolution, but as

* Corresponding author.
E-mail addresses: farha.alkuwari@qu.edu.qa, farha@stats.ucl.ac.uk (F.A. Alkuwari), serge@stats.ucl.ac.uk (S. Guillas), yuhang.wang@eas.gatech.edu (Y. Wang).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.08.031

computing power increases CTMs became more able to produce forecasts on a relatively small resolution (Eder et al. (2009), Lee et al. (2009), and Ngan et al. (2012)). CTMs produce their forecasts as an average over a grid cell. Although some CTMs provide forecasts at a fine grid cell resolution, in practice, point-specific information are needed. One approach that can improve the forecasts provided by CTMs is to combine the actual observations (usually produced by monitoring stations) with the CTM outputs. However, the model outputs and the actual ozone observations have different scales. Model outputs are produced as averages over a grid cell and ozone observations are recorded as points. To overcome the difference in scale issue we can downscale the model output to a point resolution. Downscaling is a technique that is used to extract high-resolution information from large/regional scale variables produced by numerical models. Many downscaling methods are available but pertain mostly to the climate modelling community. Statistical downscaling is based on developing a statistical relationship between observed small-scale variables (predictands) and large-scale variables (predictors) from a numerical model. The main advantage of statistical downscaling methods is that they are computationally inexpensive and appropriate when computational resources are limited (Wilby et al., 2004). Regression-based downscaling is a widely applied method in practice. It formalizes mathematically the relationship between large-scale predictors and the small-scale predictand.

There is an advantage in using statistical downscaling on coarse resolution air quality models instead of using a high resolution CTMs: downscaling combines the benefits of both statistical methods and CTMs. Statistical downscaling provide point-specific forecasts taking into consideration the physical and chemical process at a scale that is not available to even a fine scale CTM due to the local characteristics of the station, whereas CTMs show skills in the highly non-linear modelling of chemistry and transport at a more regional scale. Downscaling methods can be applied in air quality management. Indeed, on can downscale CTM forecasts, by using the established relationship between the model and observations that were estimated using historical data, to predict local air quality.

Few studies about downscaling air quality model have been published in the literature. Guillas et al. (2008) applied a two-step regression technique to downscale an air quality model in order to improve local forecasts at Use U.S. Environmental Protection Agency (EPA) monitoring stations from the Atlanta area. This improved ozone forecasts by up to 25% compared to the direct use of the numerical model. Using Bayesian approaches to downscale air quality has gained momentum recently (Berrocal et al., 2009, 2010, 2012).

Statistical downscaling often involves the use of multiple regression to downscale the data. However, this may produce unstable results when the predictors are correlated amongst each other (multicollinearity) or when the number of predictors is large. To overcome these problems, Principal Components (PCs) can be used to reduce the dimensionality of predictors and to eliminate multicollinearity. A number of studies in the climate literature make use of PC regression to downscale, e.g. Kim et al. (1984), Kidson and Thompson (1998), Hessami et al. (2008). In some downscaling scenarios the predictand does not necessarily have to be the same variable as the predictors, but they are strongly related to each other. For example, Schubert and Henderson-Sellers (1997) employed GCM-simulated pressure fields to estimate local temperature. Furthermore, predictors can be combined: Benestad (2002) downscaled temperatures over Northern Europe using so-called common EOFs that jointly reduce dimension of several spatial predictors, which is an improvement upon simple EOFs.

In this paper we present a new statistical downscaling approach that relies on Principal Fitted Components (PFC) regression (Cook, 2007). This technique reduces the dimension of predictors in a regression model with reference to the values of the predictand. We downscale an air quality model using PFC regression and compare the predictive ability of this technique to other downscaling techniques that are widely used in literature: multiple linear regression and PC regression. In the geophysical sciences, PCs are also called Empirical Orthogonal Functions EOFs (Lorenz, 1956). Indeed, EOFs are spatial PCs corresponding to the space-time variations of a quantity over a specific region. We introduce here a new kind of EOFs: Fitted Empirical Orthogonal Functions (F-EOFs). F-EOFs are spatial Principal Fitted Components (PFCs) that represent space-time variations over a region but are associated with a particular location through the use of inverse regression. The general method was developed by Cook (2007); it consists of obtaining Principal Components with reference to the response variable. PFCs are computed by performing PCA using the covariance matrix of fitted values that results from the inverse regression of the predictors on a vector valued function of the response. The simulation studies in Cook (2007) indicate that PFCs outperformed PCs as a regression dimension reduction technique and that PFC regression models exhibit better predictive ability than the Ordinary Least Squares (OLS) and PC regression models. In a follow up study Cook and Forzani (2008) presented a comprehensive theory of PFCs and further explained the advantages PFCs has (as a regression dimension reduction method) over PCs. The authors also identified the relationship between the PFC regression model and other methods (i.e. sliced inverse regression, partial and ordinary least squares, and seeded reductions). Johnson (2008) analysed the properties of PFCs and derived some theoretical properties. He studied Cook's simulation results and argued that PFCs outperformed PCs under Cook's model assumptions. Finally, Cook and Li (2009) extended the PFC methodology to regressions with categorical predictors or a mixture of categorical and continuous predictors.

One might question the plausibility of the use of the response to develop the predictors. Cook (2007) recounts the arguments between prominent statisticians: R.A. Fisher, F. Mosteller and J.W. Tukey being sceptical of nature's malice to relate the response to the least important PCs, and D.R. Cox, H. Hotelling, D.M. Hawkins and L.P. Fatti being open to the idea that in nature it is helpful to use the response to choose the predictors. Here, we refrain from entering the philosophical debate and argue that empirical evidence, in simulations (Cook, 2007) and in our case study, is inviting us to use the response in the choice of predictors. To our knowledge, there is no application of PFCs as a downscaling method in the literature. We present here the first of such application, where we downscale an air quality model.

In the next section we introduce the REAM modelling system. Section 3 describes the data that were used to perform the analysis, as well as the methodology that was followed. Section 4 presents the results of the analysis, and finally Section 5 is devoted to a discussion of the results and concluding remarks.

2. The REAM modelling system

REAM is the Regional chEmical trAnsport Model (e.g. Choi et al., 2008a,b; Wang et al., 2006, 2009; Zeng et al., 2006, Zhao et al., 2009, 2010). It adopts the photochemical, dry deposition, and biogenic emission modules from the GEOS-CHEM model, see (Bey et al., 2001) and references therein. We use the same model setup by Choi et al. (2008b) over North America. Anthropogenic and biogenic emission algorithms and inventories are adapted from the GEOS-CHEM model (Choi et al., 2005, 2008a). One exception is that the emissions of NO_x, CO, and ($\geq$C4 alkanes) over the US are prepared by Sparse Matrix Operator Kernel Emissions (SMOKE)

model (Houyoux et al., 2000) for 2005 projected from the VISTAS 2002 emission inventory.

We use the National Center for Atmospheric Research/Penn State MM5 dynamical model to provide the meteorological fields using four-dimensional data assimilation based on the National Center for Environmental Prediction (NCEP) reanalysis, rawinsonde, and surface observations (Georg et al., 1994). The REAM model used in this study has 70 km horizontal resolution with 21 vertical layers in the troposphere. The five extra grids on each side of the REAM domain are for minimizing potential transport anomalies near the boundary. The 2005 summertime GEOS-CHEM global chemical transport model (version 7.2) simulations are used to specify initial and boundary conditions for trace gases for June–August 2005 time period. The regional simulations are carried out in the last two weeks of May for spin up, and used to determine the initial chemical condition in the troposphere for the June–August 2005 simulation. Ozone 24 h ahead predictions are obtained from REAM. REAM produces forecasts within 104 grid cells (which cover the south-eastern region in the US) with 70 km spatial resolution. Out of the 104 and grid cells 5 overlap with the sea, as shown in Fig. 1. We do not consider these grid cells in our study.

3. Data and methods

3.1. Observations

The measurements are hourly ozone observations in south-eastern U.S.A. in the summer (June–August) of 2005 at stations maintained by the US Environmental Protection Agency (EPA). Data are available for 109 monitoring stations but some stations (in particular 15 stations) were outside or at the border of the grid cells range of REAM. As we want to relate regional patterns of ozone to local observations, we discarded these 15 stations and carried out the analysis for the remaining 94 stations, shown in Fig. 1. Since some of the stations had missing values (4.5% of the data were missing) we use linear interpolation to estimate the missing values. Prior to carrying out the analysis the data need to be centred to avoid the non stationary features that would prevent us from applying our statistical approach: one needs to keep stationary distributions in the errors to carry out regression. Due to the diurnal nature of the data, we centred the data by removing the diurnal cycle. The data were highly skewed for some of the stations, and this could distort the relationships in the models since our methodology in build under the assumption that errors have Gaussian distributions. To overcome this, we converted the ozone data to the square root scale to remove the skewness in the data. In our main analysis, for each of the 94 stations, we use the period from 6 June to 25 June as historical period and from 26 June to 30 June as forecasting period (i.e. validation period).

3.2. PCs and PFCs

3.2.1. Principal Component Analysis

Principal Component Analysis (PCA) is a technique for creating new variables which are a linear combination of the original variables (Anderson, 2003). Let $\mathbf{X(t)} = \mathbf{X}_1(t), \mathbf{X}_2(t), \ldots, \mathbf{X}_p(t)$ be an $n \times p$ centred matrix of predictors, where p is the number of predictors, and $t = 1, 2, \ldots, n$ (n is the sample size). The linear combinations:

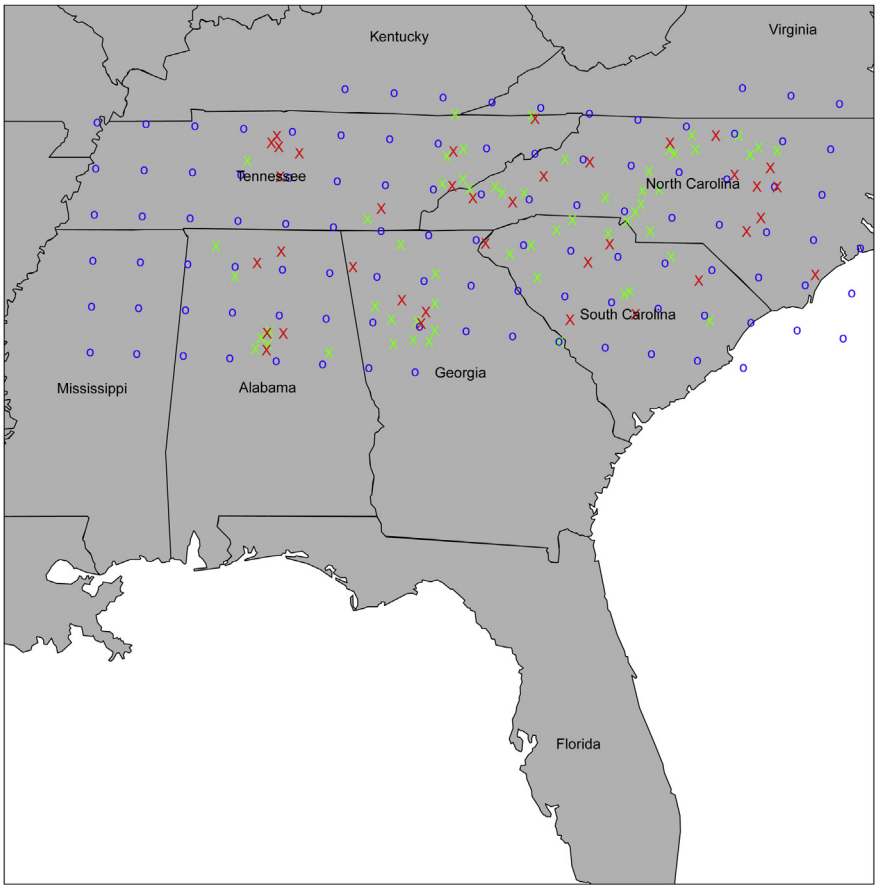

Fig. 1. Ozone monitoring stations (×) and REAM grid cells (circles).

$Y_1 = \mathbf{a}_1^T\mathbf{X}$, $Y_2 = \mathbf{a}_2^T\mathbf{X}$, …, $Y_p = \mathbf{a}_p^T\mathbf{X}$ are the Principal Components, when $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ are the eigenvectors that correspond to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of the covariance matrix of predictors $\mathbf{\Sigma} = \mathrm{Var}(\mathbf{X})$. The $\mathbf{a}$'s are known as *PC directions* or *PC loadings*. In PCA, the new variables are formed such that the first variable accounts for the maximum normalized variance in the original data. The second variable accounts for the maximum normalized variance in the original data uncorrelated to the first variable and the pth variable accounts for the maximum variation uncorrelated to the first $p-1$ variables. Thus, the first few PCs should contain a significant amount of information from the original data. PCA is a dimension reduction technique since one could use the first few PCs to represent the original data if the number of predictors is significantly large. As a result, PCs are often employed in regression to replace the original predictors. The use of PCs in regression overcomes the problem of predictors being correlated amongst each other (assuming normality).

3.2.2. Principal Fitted Component analysis

There are concerns regarding the use of PC as predictors in a regression model. First, PCs are obtained from the predictors without reference to the response variable. It may not be the case that the response variable depends upon the first few PCs but upon a smaller mode of variability. Second, PCs are not invariant (remains unchanged under some transformation) or equivariant (changes in a convenient way under some transformations) under full rank linear transformations of the predictors. PFCs as a dimension reduction approach was first presented by Cook (2007). PFCs have two major advantages over PCs when used as a dimension reduction in regression. They can be tailored to the value of the response and they are equivariant under full rank transformation of the predictors.

PFCs are obtained by gaining sufficient information about the response Y from the predictors X. One way to do this is by using inverse regression. In the inverse regression we obtain $E[X|Y = y]$, which consists of p one dimensional regressions instead of the forward regression $E[Y|X = x]$.

Assume that $\boldsymbol{\nu}_y = \boldsymbol{\beta}\,\mathbf{f}_y$ where $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, $d \leq r$, $\mathbf{f}_y \in \mathbb{R}^r$ is a known vector-valued function of the response with $\Sigma_y\,\mathbf{f}_y = 0$ (i.e. $\mathbf{f}_y$ is centred around its mean) and y is used to index observations instead of the usual "i" (Cook, 2007). The predictors $\mathbf{X}_y$ are regressed on $\mathbf{f}_y$, which is a function of the response Y. The function $\mathbf{f}_y$ can be formed using a particular basis function $\mathbf{g}_y$ and $\mathbf{f}_y$ is centred: $\mathbf{f}_y = \mathbf{g}_y - \overline{\mathbf{g}}$. An example of commonly used basis is the polynomial basis:

$$\mathbf{g}_y = \left(y,\, y^2,\, \cdots,\, y^r\right)^T, \tag{1}$$

Let $\widehat{\mathbf{\Sigma}}_{fit}$ be the sample covariance matrix of $\widehat{\mathbf{X}}$ (which is the $n \times p$ matrix of fitted values resulting from the regression of $\mathbf{X}_y$ on $\mathbf{f}_y$):

$$\widehat{\mathbf{\Sigma}}_{fit} = \frac{\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}}{n} \tag{2}$$

Then $\widehat{\Phi}_1^T\mathbf{X}$, $\widehat{\Phi}_2^T\mathbf{X}$, …, $\widehat{\Phi}_p^T\mathbf{X}$ are called *Principal Fitted Components*, where $\widehat{\Phi}_1$, …, $\widehat{\Phi}_p$ are the eigenvectors that corresponds to the eigenvalues $\widehat{\lambda}_1^{fit}$, $\widehat{\lambda}_2^{fit}$, …, $\widehat{\lambda}_p^{fit}$ of the covariance matrix $\widehat{\mathbf{\Sigma}}_{fit}$. Therefore, PFCs are obtained by performing PCA on the fitted sample covariance matrix.

Few PFCs (which account for high variation in the original data with reference to the response) can be used in the regression model instead of the original high dimensional predictors. Since PFCs are obtained using the response, they outperform PCs as regressors in many situations (Cook, 2007). However, the use of PFCs as predictors does not eliminate multicollinearity (Cook, 2007). Then PFC scores are obtained by multiplying the eigenvectors of $\widehat{\mathbf{\Sigma}}_{fit}$ by $\mathbf{X}$. Similarly to PCA, the choice of the number of PFCs to be considered in a regression model is rather subjective.

3.3. Methodology

Simple linear regression, PC regression, and PFC regression are employed to downscale the REAM model output and thus forecast local ozone levels. Models are fitted to the historical data, and then used to predict hourly ozone observations over the validation period. The Root Mean Square Error (RMSE) measures the forecasting accuracy. The overall results will be discussed in details in Section 4.

3.3.1. Downscaling REAM outputs by linear regression

For each station, we fit a linear regression model by regressing hourly ozone observations on the grid cell outputs that includes the station:

$$O_t = \beta_0 + \beta_1 M_t + \varepsilon_t \tag{3}$$

where O_t is hourly ozone observations, M_t is the REAM model output of the grid cell that include the station, β_0 and β_1 are regression model parameters and are estimated by the method of least squares, and ε_t is a normal error vector with mean 0 and a constant variance σ^2.

3.3.2. Downscaling REAM outputs by PC regression

The approach presented in Section 3.3.1, where we regress ozone observations on the grid cell that contains the station may not be very efficient. Indeed, the CTM may be misaligned and local ozone may be more closely related to regional conditions rather than the average over the grid cell. REAM forecasts ozone levels over $p = 99$ grid cells, and considering all 99 cells in the regression model raises the problem of "over fitting" – the model may have good fitting performance but poor predictive performance. To tackle this problem, we can reduce the number of predictors by applying PCA to the grid cells model output. Then, we can select a few PCs that capture the highest variation in the original grid cells data, and use them as predictors in the regression model. For each station we regress hourly ozone observations on a selected number of PCs:

$$O_t = \alpha_0 + \sum_{m=1}^{M} \alpha_m Z_m(t) + \varepsilon_t \tag{4}$$

where M is the number of PCs in the model, $M < p = 99$, $\alpha_0, \alpha_1, \ldots, \alpha_m$ are model parameters, $Z_1(t), Z_2(t), \ldots, Z_m(t)$ are PC scores, $t = 1, 2, \ldots, n$, and ε_t is a normal error vector with mean 0 and a constant variance σ^2.

There are numerous methods that can be employed to determine the number of PCs to be used in regression. A widely used PC selection technique chooses PCs that account for a large amount of variation of the predictors. However, PCs with low variance are not necessarily redundant. A PC might explain a small amount of the variation in the predictors but it could be a significant predictor for the dependent variable. Examples of such cases can be found in Kung and Sharif (1980) and Jolliffe (1982). For this purpose, in this paper we use a leave-one-out cross validation method (Mertens et al., 1995) where the PRESS (PREdicted Sum of Square) value helps determine the number of PCs to maintain in a regression model.

3.3.3. Downscaling REAM outputs by PFC regression

We downscale the REAM output by first applying a PFC analysis to reduce the dimension of the model outputs. We used a

polynomial basis function to compute the F-EOFs (we have explored other basis functions, e.g. slice, when computing the F-EOFs and polynomial basis function seemed to give better results than the other methods in our application). Then, we select a few F-EOFs as predictors in the regression model. For each station we regress hourly ozone observations on these F-EOFs:

$$O_t = \delta_0 + \sum_{d=1}^{D} \delta_d P_d(t) + \varepsilon_t \quad (5)$$

where D is the number of F-EOFs in the model, $D < p = 99$, δ_0, δ_1,…, δ_d are model parameters, $P_1(t)$, $P_2(t)$,…, $P_d(t)$ are PFC scores, $t = 1, 2,\ldots, n$, and ε_t is a normal error vector with mean 0 and a constant variance σ^2.

4. Results

Spatial plots of EOFs and F-EOFs can be very informative. An EOF plot shows the loadings of the corresponding PC or PFC geographically. It displays the locations at which PCs and PFCs contribute more strongly or weakly. The EOFs and the F-EOFs are associated with eigenspaces of dimension one, so the sign of the contribution does not matter.

Fig. 2 shows the leading EOFs. The EOF distributions reflect the general variation of ozone, the value of which are higher over emission regions and over land than over ocean. EOF1 illustrates mainly the ozone gradient decreasing towards the eastern coastline, and EOF4 shows the ozone gradient decreasing towards the southern coastline, reflecting generally much lower ozone concentrations over the ocean than land. EOF2 and EOF3 are associated with regional ozone distribution patterns, in which the variations are lower in Alabama and northern Georgia, and Mississippi and Tennessee, respectively. These spatial patterns are likely driven by meteorological systems that transport low ozone air masses to these regions.

Figs. 3 and 4 show the first F-EOFs of REAM outputs corresponding to eight selected stations over 6–25 June 2005. Stations 35 and 60 are generally associated with the west to east gradient of low ozone in the eastern coastline, which is somewhat similar to the EOF1 distribution (Fig. 3). The F-EOFs of Stations 75 and 5 have a mixture of EOF1 and EOF4 distributions, showing lower ozone variation in eastern and southern coastlines. The F-EOF of Station 51 has some resemblance to EOF2, showing lower ozone variation in Mississippi and Tennessee. Station 66 has an F-EOF similar to EOF3, showing low variation in Alabama and northern Georgia, although it extends further to eastern South Carolina. The F-EOF of Stations 96 and 83 are more complicated, none of which is a clear extension of the 4 EOFs. The uniqueness of the first F-EOFs implies that the PFC method is able to more efficiently reduce the dimension of the problem and capture the regional distribution pattern specifically relevant to the site of interest.

We first performed a simple simulation to gain further understanding of the features of PCs and PFCs in our context. The simulation is similar to the simulation in Cook (2007). The only difference is that the random variables generated here are seasonal, which mimics our situation where a diurnal cycle is present (and removed). First, we generate Y as a normal random variable with mean 0 and variance σ_Y^2 and has size n We added a seasonal pattern to the generated random variable. For simplicity the seasonality is a repetitive cycle of 10 values: 1, 2,…, 10. Second, we generate $\mathbf{X}_y$, which is an $n \times p$ matrix, according to the inverse model:

$$\mathbf{X}_y = \Gamma y + \sigma\varepsilon \quad (6)$$

where $\boldsymbol{\Gamma} = (1, 0,\ldots, 0)^{\boldsymbol{T}}$ and $\boldsymbol{\sigma} > 0$, and $\boldsymbol{\varepsilon}$ is a standard normal random variable. The number of predictors is $\boldsymbol{p} = 100$. We performed the simulation with sample sizes $\boldsymbol{n} = 200$, 500, and 1000. The forward regression model is:

$$\boldsymbol{Y} = \alpha_0 + \boldsymbol{\alpha}^{\boldsymbol{T}}\mathbf{X} + \boldsymbol{\sigma}_{\boldsymbol{Y}|\mathbf{X}}\boldsymbol{\varepsilon}$$

where $\mathbf{x}$ is the observed value of $\mathbf{X}$, $\boldsymbol{\sigma}_{\boldsymbol{Y}|\mathbf{X}}$ is constant, and $\boldsymbol{\varepsilon}$ is a standard normal random variable. Finally, we apply the three approaches in Section 3.3 to model the simulated data. For a straightforward comparison amongst all methods, as in Cook (2007), we restrict ourselves to $\boldsymbol{d} = 1$ ($\boldsymbol{d}$ is the dimension of the reduced space, i.e. the number of PCs and PFCs to be included in the regression model). We fit the PC model with one PC and the PFC

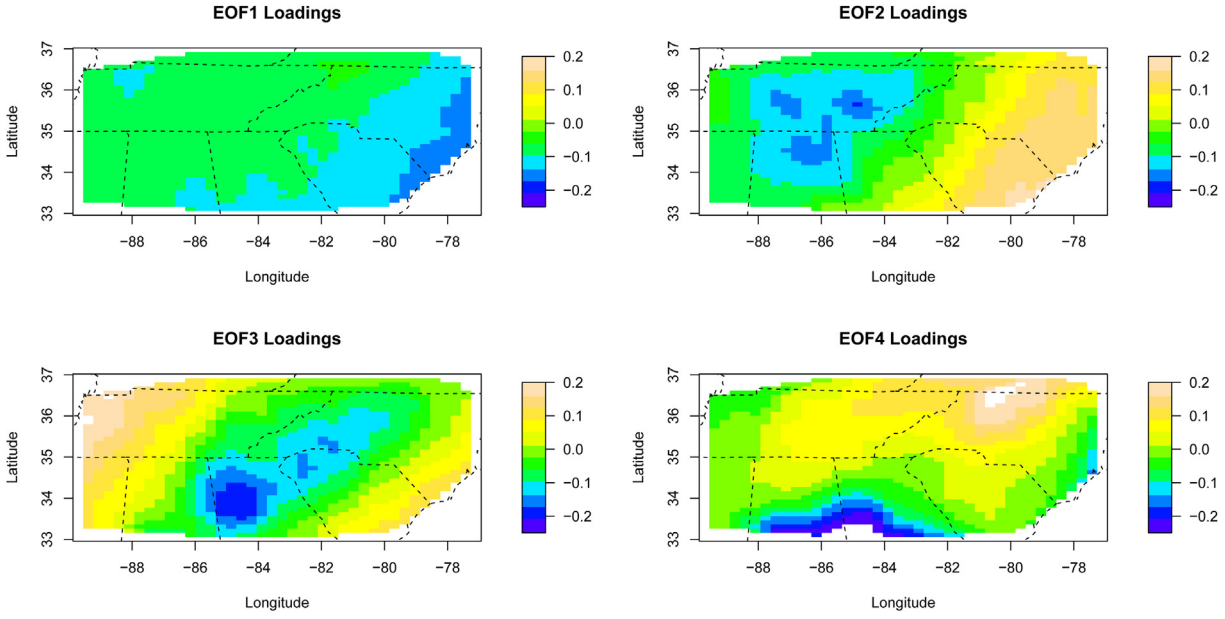

Fig. 2. The first four leading EOFs of the gridded REAM output from 6 June to 25 June 2005.

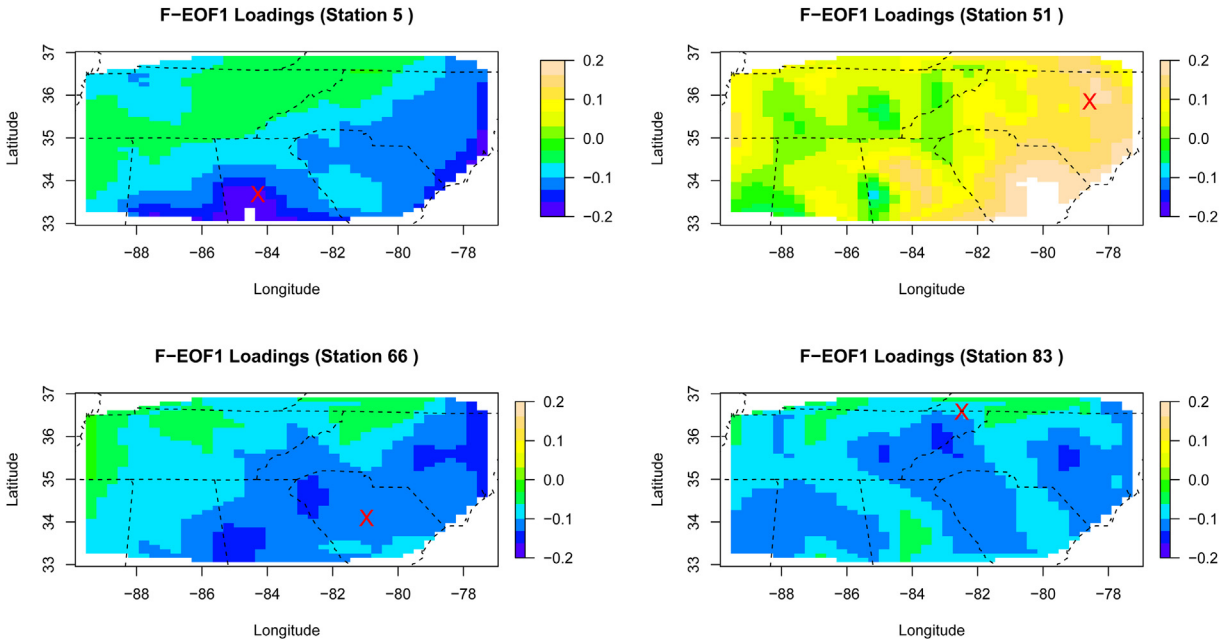

Fig. 3. The first F-EOFs (polynomial basis function with degree 1) of the REAM outputs for four stations, estimated over 6–25 June 2005. The location of the station is marked by '×'.

model with one PFC ($\mathbf{f}_{\boldsymbol{y}} = \boldsymbol{y} - \overline{\boldsymbol{y}_{\boldsymbol{s}}}$). For each simulated dataset we use the first 80% as a fitting period and the remaining as a validation period. Table 1 summarizes the results based on 100 replications. On average the PFC model seems to show a better predictive ability than the other models at all sample sizes.

According to the simulation results, PFCs performs comparatively better when the sample size is small compared to the dimension of predictors. This is the situation for the ozone data in hand, as the sample size (fitting period) is relatively small compared to the number of grid cells. Using the three approaches presented in Section 3.3, we use the period from 6 to 25 June as a training and the period from 26 to 30 June as validation period. Although the data were converted to the square root scale to adjust the skewness, the plots and tables presented in this section show the results after converting them back to the original scale. The number of PCs selected in the regression model was determined after performing the leave one out cross validation presented by Mertens et al. (1995). We performed the cross validation for each stations individually and limited the number of PCs in the regression model to a maximum of 20 to avoid over fitting. The cross validation results show that each station should be fitted with a different number of PCs. For example, for some stations using only one PC in the regression model seems to be enough, while for other stations all 20 PCs should be used as predictors to obtain significant results. Hence, according to the cross validation results a PC regression model with different number of PCs have been fitted for

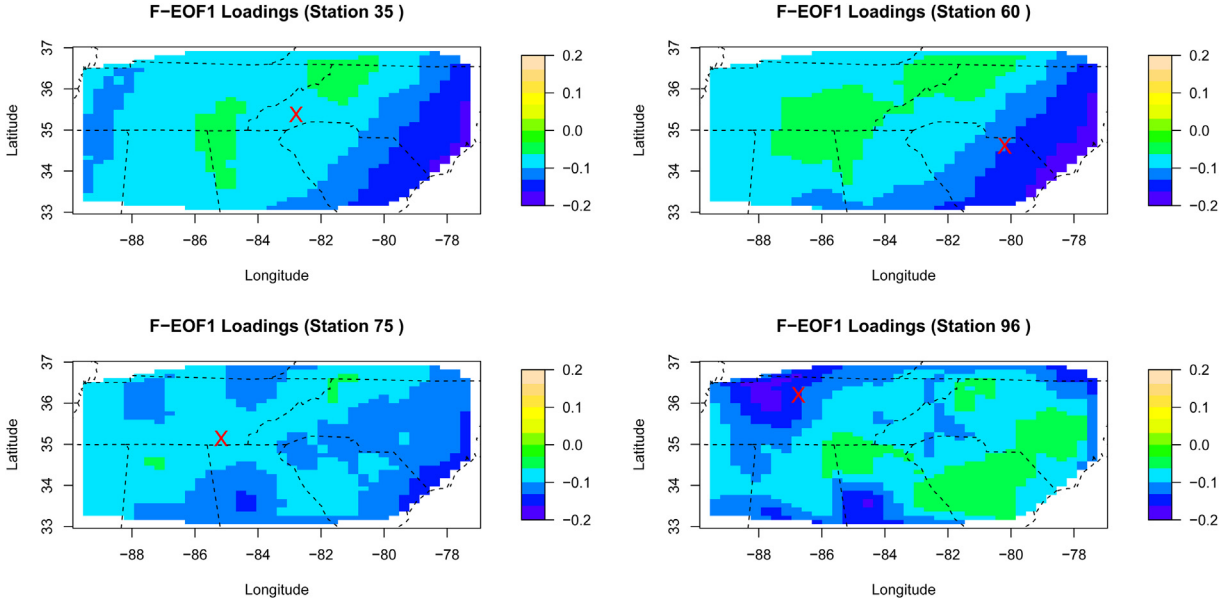

Fig. 4. The first F-EOFs (polynomial basis function with degree 1) of the REAM outputs for four stations, estimated over 6–25 June 2005. The location of the station is marked by '×'.

Table 1
Simulation results: RMSEs averaged over 100 replications for linear, PC, and PFC regression. The PC model was fitted with one PC and the PFC model was fitted with one polynomial PFC.

Sample size	Linear regression	PC regression	PFC regression
200	1.02	0.79	0.67
500	0.78	0.76	0.71
1000	0.74	0.73	0.71

Table 2
RMSEs: training period is 6–25 June, validation period is 26–30 June. The PC models were fitted for: station 29 with 18 PCs, station 84 with 5 PCs, and station 107 with 20 PCs. The PFC models were fitted with 1 PFC (polynomial basis function with degree one).

Regression model	All stations	Station 29	Station 84	Station 107
REAM	18.98	15.10	26.95	27.40
Linear	10.01	11.57	13.33	10.08
PC	9.35	11.31	10.27	9.89
PFC	9.13	9.86	10.36	8.91
Linear with AR(2) errors	10.66	12.77	12.33	9.17
PC with AR(2) errors	9.61	11.17	10.27	9.53
PFC with AR(2) errors	9.39	10.14	10.36	8.76

each station. Fig. 5 shows a summary of the number stations (i.e. regression models) versus the number of PCs needed to fit the regression model. The plot shows that for most of the stations in the study area using only one PC in the regression model seems to be significant. We fit the PFC model using one PFC (PFCs were obtained using a polynomial basis function of degree one). Table 2 shows the RMSEs for some stations within the study region and the average RMSE. The average RMSE indicate that overall, PFC regression outperformed both PC and simple regression methods. The PFC model has significantly improved the predictive ability relative to the REAM model (the ozone prediction error has been reduced by 52% relative to the REAM model predictions). Using PFCs improved ozone predictions by approximately 10% relative to the simple regression model, but (on average) it showed a 3% improvement relative to the PC model. Although RMSE values indicate that the PFC model has better predictive ability than the other approaches, they show that PFCs did not perform best in 36 stations compared to the linear and PC regression, which approximately accounts for 38% of the stations in the study region (these stations are marked in red 'x' in Fig. 1). In 18 out of these 36 stations the PC model seems to outperform the other methods. Simple regression appears to perform worst on average, which reinforces our view that regional variations ought to be taken into account.

Fig. 6 displays ozone observations and the corresponding REAM outputs, simple regression forecasts, PC regression forecasts, and PFC regression forecasts over the period from 26 to 30 June for a selected stations. The plots indicate that for the selected prediction period, the PFC model produced forecasts that are very close to the actual ozone concentrations at most times of the day.

The nature of our data suggests that there might be a great possibility that the model errors could be correlated. One way of verifying this, is to plot the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the model residuals. The ACF and the PACF plots (not shown here) for the residuals of the simple regression, PC regression, and PFC regression indicate that the models errors seems to be autocorrelated. The models errors are not white noise; hence they should be modelled as an autoregressive model (AR). Note that this structure in the error is actually due to the different short-term behaviours in time of observations and REAM outputs, and was already modelled in Guillas et al. (2008). The ACF and PACF plots suggest that modelling the residuals using a second order autoregressive model AR(2) seems to be a sensible choice overall. We refitted the regression models presented in this paper with an AR(2) model for the error. Table 2 shows the RMSEs averaged over the stations within the study region and for three randomly selected stations. We would expect that the RMSE values to be smaller for the models with and AR errors. However, the results show that modelling the errors with an AR(2) models did not improve the models predictions in general.

To investigate the change in prediction errors when using a longer fitting period, we repeated the analysis using the period from 6 June to 10 July as a fitting period and used the proceeding five days as validation period. Table 3 shows the RMSE values averaged over all stations in the study region and for selected stations. The results indicate that the PFC model improved the predictive performance by 45% compared to the REAM model. Moreover, using PFCs improved ozone predictions by 2% relative to the simple regression model and by 4% compared to the PC model. On average, the PC model seems to perform the worst in this case. Although predictions errors seem to be smaller when using a longer fitting period, the PFC model does not seem to have a significant prediction improvement compared to the other downscaling methods. This coincides with the simulation results when we used a relatively large sample size.

To further verify the consistency of our findings, we repeated the analysis and model fitting for different fitting and validation periods. We selected a fitting period of size $\boldsymbol{n}$ days and predict for the next $\boldsymbol{k}$ days. Then, we move the $\boldsymbol{n}$ day fitting period one day

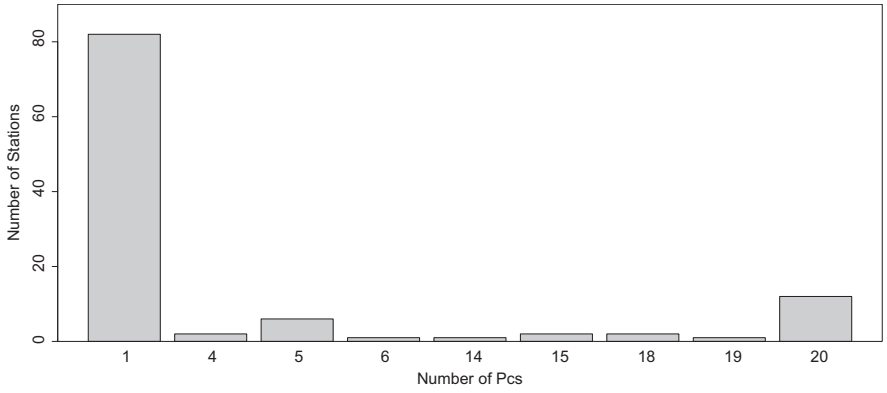

Fig. 5. The bar chart shows a summary of the number of stations versus the number of PCs needed to fit the regression model for the station. The number of PCs were determined using the PRESS cross validation method by Mertens et al. (1995).

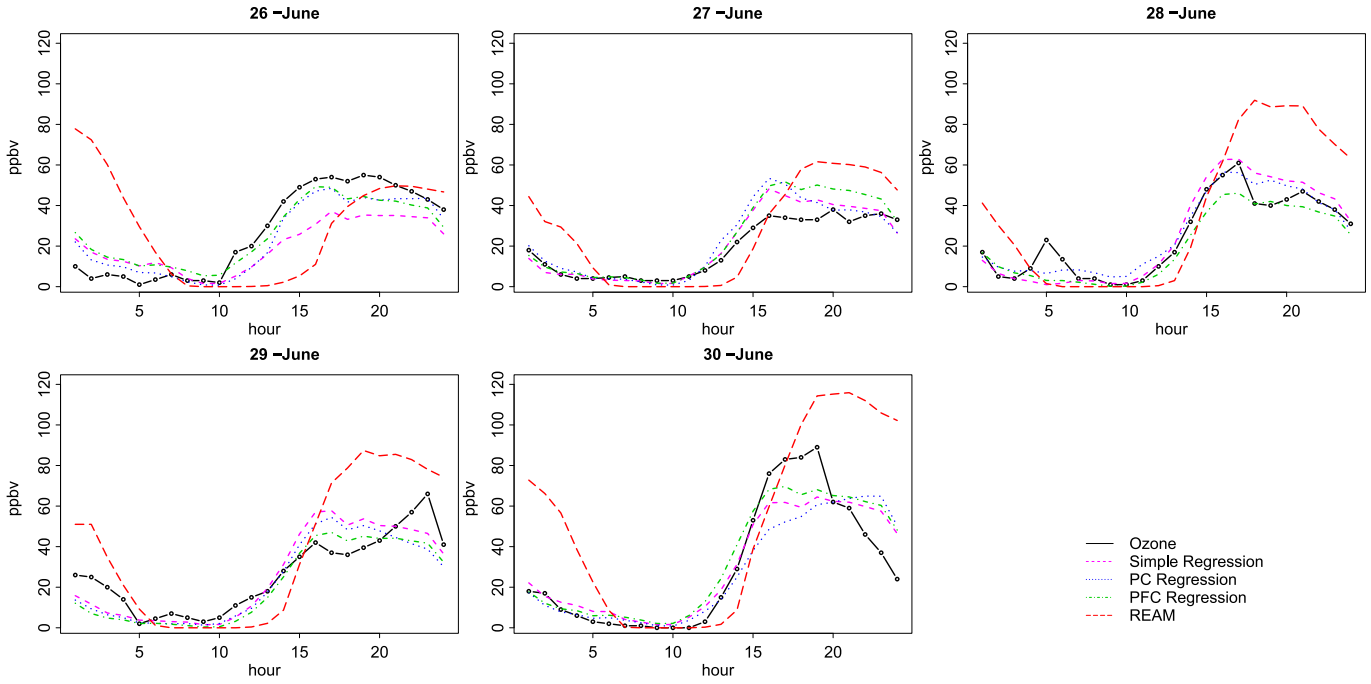

Fig. 6. Prediction plots for station 107 (26–30 June). Observations (black line), REAM outputs (dashed red line), linear regression predictions (dashed pink line), PC predictions (20 PCs, dotted blue line), and PFC predictions (polynomial basis function with degree one, dashed green line).

ahead and predict for the next $\boldsymbol{k}$ days and so on. The data are available from 2 June 2005 to 31 August 2005. Within 2 June to 26 August, we select a fitting period of $\boldsymbol{n} = 20$ consecutive days (i.e. 480 data points) and we allocate the following $\boldsymbol{k} = 5$ days (i.e. 120 data points) to the validation period. Then we compute the RMSE for each set of predictions. This computation is repeated 60 times, as we move the fitting period by one day ahead and proceed with the prediction for the corresponding validation period. Table 4 shows the RMSEs (averaged over the 60 runs) for stations 29, 80, and 107. For station 29 the PC model were fitted with 18 PCs and for stations 80 and 107 the PC model were fitted with 20 PCs. The PFC models were fitted with 1 PFC, which was computed using a polynomial basis function with degree one. The table also shows the average RMSE over all stations and over all 60 runs. PFCs outperform other methods in terms of predictive ability. However, PFCs did not perform well for 38 stations, which is approximately 40% of the stations in the study region. These results coincides with the results we obtained from Table 2. We conclude that overall, PFCs show a significant improvement over PCs as our analysis rely on the 60 chosen fitting and validation periods.

We now use the Jackknife to assess the intrinsic uncertainties in the PC and PFC regression approaches (Efron and Tibshirani, 1994). We remove one day (i.e. 24 data points) out of the fitting period (6–25 June) and carry out the estimation and the prediction of the validation period (26–30 June). This procedure is repeated removing one day at a time over the fitting period. Moreover, we add each model predictor (PC or PFC score) progressively. We cap the number of predictors in the models to a maximum of 10 (this is to avoid over-fitting and for computational convenience). To compute the PFCs we used a polynomial basis function with degree 10 as the number of PFCs in a model should not exceed the size of the basis function. We then compute the RMSEs for each set of predictions. Table 5 show the jackknife RMSEs (averaged over all 94 stations) for the PC and the PFC model. It shows that PFCs have better predictive ability than PCs when considering one PC and one PFC as predictors in the regression model. Furthermore, increasing the number of predictors in the PC and PFC models does not improve the predictive ability of the model. It seems that adding more predictors adds more noise for both types of models. The results indicates that having only one PFC in the regression does not only shows better predictive performance, but also indicates that PFCs outperforms PCs as a dimension reduction technique.

5. Discussion and conclusion

We used Fitted Empirical Orthogonal Functions (F-EOFs) to downscale an air quality model for ozone over the southeastern U.S. The results were compared to two downscaling approaches: simple linear regression and Principal Components (PC) regression. The analysis has been done for each site separately. The PFC regression outperformed the other methods in terms of predictive ability in most stations within the study region. However, the PFC method did not work better in roughly one third of the stations. This might

Table 3
RMSEs: training period is 6 June to 10 July, validation period is 11–15 July. The PC models were fitted for: station 29 with 18 PCs, station 84 with 5 PCs, and station 107 with 20 PCs. The PFC models were fitted with 1 PFC (polynomial basis function with degree one).

Regression model	All stations	Station 29	Station 84	Station 107
REAM	15.84	11.46	17.61	18.07
Linear	8.87	9.66	7.18	8.06
PC	9.04	11.56	6.45	7.82
PFC	8.68	9.27	6.54	7.82

Table 4
The RMSE value for some selected stations in the study region. We selected a fitting period of 20 days (i.e. 480 data points) and we used the following 5 days (i.e. 120 data points) as a validation period. We chose 60 different fitting and validation periods for each station. The RMSE values are averaged based on 60 runs.

Regression model	All stations	Station 29	Station 80	Station 107
Linear	9.33	9.50	8.79	10.31
PC	9.32	9.96	8.84	10.94
PFC	9.03	9.23	7.44	9.52

Table 5
Jackknife RMSE for the PC and PFC regressions. The values are averaged over all 94 stations of the study region. The training period was 6 June to 25 June and the validation period was from 26 June to 30 June. The PFC model was computed based on a polynomial basis with degree 10.

No. of predictors	PC model	PFC model
1	9.20	9.14
2	9.24	9.34
3	9.34	9.45
4	9.48	9.50
5	9.45	9.56
6	9.49	9.63
7	9.39	9.70
8	9.43	9.73
9	9.44	9.76
10	9.42	9.82

be because there are limited grid cells covering the locations of those stations, which might be resolved by enlarging the domain of the model, as some of stations in the study area are located at the border of the grid cells domain. We considered the autocorrelated nature of the data by fitting an AR(2) model for errors. The results did not show any improvement in the models predictive ability. This may be because the autoregressive structure of the errors is not strong, hence the models without AR(2) errors might already captures the essential features of the relationship between REAM and the observations; the AR modelling step adds noise instead of reducing uncertainties. We repeated the analysis for different fitting and validation periods and the results coincide with the illustrative period we initially chose. We examined the uncertainty in the PC and the PFC models by applying the jackknife method, and PFCs outperform PCs as a dimension reduction technique. These results are consistent with the simulation results by Cook (2007).

One might argue that our method lacks the information that can be obtained by using a spatially and temporally varying coefficient in the regression model such as the downscaler used in Berrocal et al. (2012). This type of weighted downscalers have the advantage of borrowing strength for coefficients in neighbouring locations and across space, while in our case we obtain a different coefficient for each location (station) individually. Although our technique does not borrow strength from neighbouring grid cell, we believe that our method could be more adaptive to spatial properties that might exist in the region, such as teleconnections and anisotropy.

Another important avenue to improve the forecasting ability of the PFC model would be to consider additional factors that affect ozone levels (e.g. temperature). One may think of the common EOFs (Benestad, 2002) as a potential tool for that goal. Initial investigation was done using Temperature as a predictor in the downscaling process in addition to REAM. Although we would expect the predictive ability of the model to improve, the results showed that adding temperature as a predictor in the downscaling process did not improve the models predictive performance. This might be because REAM already seems to capture well the impact of temperature on ozone. Finally, the uncertainties in the model itself, due to numerical errors or unknown parametrizations of chemistry and transport of ozone and its precursors, ought to also trickle down to the location of interest through downscaling. This is a challenging task that most probably requires a Bayesian framework to reflect prior scientific knowledge.

PFCs are estimated using the sample covariance matrix. However, it has been shown that it is not a good estimator of the population covariance matrix, see Dempster (1969). One approach that could be used to obtain a better estimate of the covariance matrix is thresholding (Bickel and Levina, 2004). One of the main advantages of thresholding is that it is computationally inexpensive. A useful extension to our work would be to threshold the covariance matrix used in the estimation of PFCs. It would be interesting to examine the effect of thresholding on the predictive ability of the PFC regression model. We are currently investigating the impact of thresholding the covariance matrix on the forecasting performance of the PC and PFC models. We have reasons to believe that it could enhance the overall forecasting ability.

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis. Wiley.
Benestad, R., 2002. Empirically downscaled temperature scenarios for northern Europe based on a multi-model ensemble. Clim. Res. 21 (21), 105–125.
Berrocal, V., Gelfand, A., Holland, D., 2009. A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. 15 (2), 176–197.
Berrocal, V., Gelfand, A., Holland, D., 2010. A bivariate space-time downscaler under space and time misalignment. Ann. Appl. Stat. 4 (4), 1942–1975.
Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2012. Space-time data fusion under error in computer model output: an application to modeling air quality. Biometrics 68 (3), 837–848. ISSN 837-0420.
Bey, I., Jacob, D.J., Yantosca, R., Logan, J., Field, B., Fiore, A., Li, Q., Liu, H., Mickley, L., Schultz, M., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: model description and evaluation. J. Geophys. Res. 106.
Bickel, P., Levina, E., 2004. Covariance regularization by thresholding. Ann. Stat. 36 (6), 2577–2604.
Choi, Y., Wang, Y., Cunnold, D., Zeng, T., Shim, C., Luo, M., Eldering, A., Bucsela, E., Gleason, J., 2008a. Spring to summer northward migration of high O_3 over the western North Atlantic. Geophys. Res. Lett. 35.
Choi, Y., Wang, Y., Zeng, T., Cunnold, D., Yang, E., Martin, R., Chance, K., Thouret, V., Edgerton, E., 2008b. Springtime transitions of NO_2, CO, and O_3 over North America: model evaluation and analysis. J. Geophys. Res. 113.
Choi, Y., Wang, Y., Zeng, T., Martin, R., Kurosu, T., Chance, K., 2005. Evidence of lightning NO_x and convective transport of pollutants in satellite observations over North America. Geophys. Res. Lett. 32.
Cook, R., Forzani, L., 2008. Principal fitted components for dimension reduction in regression. Stat. Sci. 23 (4), 485–501.
Cook, R., Li, L., 2009. Dimension reduction in regressions with exponential family predictors. J. Comput. Graph. Stat. 18 (3), 774–791.
Cook, R.D., 2007. Fisher lecture: dimension reduction in regression. Stat. Sci. 22 (1), 1–26.
Dempster, A., 1969. Elements of continuous multivariate analysis. Addison-Wesley series in behavioral sciences. Quant. Methods.
Eder, B., Kang, D., Mathur, R., Pleim, J., Yu, S., Otte, T., Pouliot, G., 2009. A performance evaluation of the national air quality forecast capability for the summer of 2007. Atmos. Environ. 43 (14), 2312–2320.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman and Hall.
EPA, U., 2006. Air Quality Criteria for Ozone and Related Photochemical Oxidants (2006 Final) (Tech. rep., Washington, DC).
Georg, A., Dudhia, J., Stauffer, D., 1994. A Description of the Fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR (Technical Note: NCAR/TN-398pSTR).
Guillas, S., Bao, J., Choi, Y., Wang, Y., 2008. Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. Atmos. Environ. 42 (6), 1338–1348.
Hessami, M., Gachon, P., Ouarda, T., St.-Hilaire, A., 2008. Automated regression-based statistical downscaling tool. Environ. Modell. Softw. 23 (6), 813–834.
Houyoux, M., Vukovich, J., Brandmeyer, J., 2000. Sparse Matrix Kernal Emission Modeling System: SMOKE User Manual. MCNC-North Carolina Supercomputing Center. URL http://www.smoke-model.org.
Johnson, O., 2008. Theoretical properties of Cook's PFCs dimension reduction algorithm for linear regression. Electron. J. Stat. 2, 807–828.
Jolliffe, I., 1982. A note on the use of principal components in regression. J. R. Stat. Soc. Ser. C Appl. Stat. 31, 300–303.
Kidson, J., Thompson, C., 1998. A comparison of statistical and model-based downscaling techniques for estimating local climate variations. J. Clim. 11 (4), 735–753.
Kim, J.W., Chang, J.T., Baker, N.L., Wilks, D.S., Gates, W.L., 1984. The statistical problem of climate inversion: determination of the relationship between local and large-scale climate. Am. Meteor. Soc. 112 (10), 2069–2077.
Kung, E., Sharif, T., 1980. Multi-regression forecasting of the Indian summer monsoon with antecedent patterns of the large-scale circulation. In: WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting, pp. 295–302.
Lee, S.-M., Princevac, M., Mitsutomi, S., Cassmassi, J., 2009. MM5 simulations for air quality modeling: an application to a coastal area with complex terrain. Atmos. Environ. 43 (2), 447–457.
Lorenz, E.N., 1956. Empirical Orthogonal Functions and Statistical Weather Prediction (Tech. Rep. 1). M.I.T., Statistical Forecasting Project.
Mertens, B., Fearn, T., Thompson, M., 1995. The efficient cross-validation of principal components applied to principal component regression. Stat. Comput. 5, 227–235. http://dx.doi.org/10.1007/BF00142664.
Ngan, F., Byun, D., Kim, H., Lee, D., Rappenglck, B., Pour-Biazar, A., 2012. Performance assessment of retrospective meteorological inputs for use in air quality modeling during texaqs 2006. Atmos. Environ. 54 (0), 86–96.

Sahu, S., Gelfand, A., Holland, D., 2007. High resolution space-time ozone modeling for assessing trends. J. Am. Stat. Assoc. 102 (480), 1221–1234.
Schubert, S., Henderson-Sellers, A., 1997. A statistical model to downscale local daily temperature extremes from synoptic-scale atmospheric circulation patterns in the Australian region. Clim. Dyn. 13 (3), 223–234.
Wang, Y., Choi, Y., Zeng, T., Ridley, B., Blake, N., Blake, D., Flocke, F., 2006. Late-spring increase of trans-Pacific pollution transport in the upper troposphere. Geophys. Res. Lett. 33 (1), L01811.
Wang, Y., Hao, J., McElroy, M.B., Munger, J.W., Ma, H., Chen, D., Nielsen, C.P., 2009. Ozone air quality during the 2008 Beijing Olympics: effectiveness of emission restrictions. Atmos. Chem. Phys. 9, 5237–5251.
Wilby, R., Charles, S., Zorita, E., Timbal, B., Whetton, P., Mearns, L., 2004. Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods (Tech. rep).
Zeng, T., Wang, Y., Chance, K., Blake, N., Blake, D., Ridley, B., 2006. Halogen-driven low altitude O_3 and hydrocarbon losses in spring at northern high latitudes. J. Geophys. Res. 111 (D17313), 55–557.
Zhao, C., Wang, Y., Yang, Q., Fu, R., Cunnold, D., Choi, Y., 2010. Impact of east asian summer monsoon on air quality over China: the view from space. J. Geophys. Res. 115 (D09301).
Zhao, C., Wang, Y., Zeng, T., 2009. East China plains: a basin of ozone pollution. Environ. Sci. Technol. 43, 1911–1915.