




A statistical downscaling framework for environmental mapping

Kassim S. Mwitondi¹  · Farha A. Al-Kuwari² · Raed A. Saeed³ · Shahrzad Zargari¹

Published online: 12 October 2018
© The Author(s) 2018

Abstract

In recent years, knowledge extraction from data has become increasingly popular, with many numerical forecasting models, mainly falling into two major categories—chemical transport models (CTMs) and conventional statistical methods. However, due to data and model variability, data-driven knowledge extraction from high-dimensional, multifaceted data in such applications require generalisations of global to regional or local conditions. Typically, generalisation is achieved via mapping global conditions to local ecosystems and human habitats which amounts to tracking and monitoring environmental dynamics in various geographical areas and their regional and global implications on human livelihood. Statistical downscaling techniques have been widely used to extract high-resolution information from regional-scale variables produced by CTMs in climate model. Conventional applications of these methods are predominantly dimensional reduction in nature, designed to reduce spatial dimension of gridded model outputs without loss of essential spatial information. Their downside is twofold—complete dependence on unlabelled design matrix and reliance on underlying distributional assumptions. We propose a novel statistical downscaling framework for dealing with data and model variability. Its power derives from training and testing multiple models on multiple samples, narrowing down global environmental phenomena to regional discordance through dimensional reduction and visualisation. Hourly ground-level ozone observations were obtained from various environmental stations maintained by the US Environmental Protection Agency, covering the summer period (June–August 2005). Regional patterns of ozone are related to local observations via repeated runs and performance assessment of multiple versions of empirical orthogonal functions or principal components and principal fitted com-

✉ Kassim S. Mwitondi
k.mwitondi@shu.ac.uk

¹ Faculty of Science, Technology and Arts, Sheffield Hallam University, Sheffield, UK

² Statistics and Research Division, Qatar Development Bank, Doha, Qatar

³ Canadian University Dubai, Dubai, UAE

ponents via an algorithm with fully adaptable parameters. We demonstrate how the algorithm can be extended to weather-dependent and other applications with inherent data randomness and model variability via its built-in interdisciplinary computational power that connects data sources with end-users.

Keywords Chemical transport models · Downscaling · Empirical orthogonal functions · Ensemble modelling · Interdisciplinary computation · Principal component analysis · Principal fitted components · Unsupervised modelling

1 Introduction and general background

Changing climatic conditions due to space and terrestrial activities and how they affect our livelihood are well documented. The relationship between solar activities and terrestrial climatic change has been hotly debated in recent years. Two schools of thought have emerged, one believing that the climate on earth is directly impacted by activities on the sun and we therefore cannot predict the climate on earth until we understand solar activities [1], and the other not seeing a direct solar impact on terrestrial phenomena. Many studies suggest that climatic conditions affect the levels of air pollutants [2], with human activity particularly blamed for the increase in greenhouse gases. While greenhouse gases reabsorb heat reflected from the Earth's surface is essential in regulating the Earth's temperature, disproportional increase in greenhouse gases tends to hinder additional thermal radiation from escaping from the Earth; hence, many researchers have attributed this to rising sea levels and temperature. The recent announcement that concentrations of carbon dioxide CO₂ in the atmosphere had surged to a record high [3] shows that we have serious issues to address. We are therefore called upon to enhance our understanding not only of the conditions on the ground—like ground-level ozone levels and extreme weather events such as flooding, but also of the heat-trapping CO₂ in the atmosphere and of what happens out in space. Despite ubiquitous enhancements in data acquisition, sharing and modelling technologies, many questions remain unanswered and, hence, new developments call for equally sophisticated real-life applications, including those focusing on tackling environmental challenges for global sustainability. We therefore need robust ways for generalising the impact of various phenomena. Downscaling can be envisioned from different perspectives—the circulation pattern over a specific geographical region could be viewed as a *large-scale variable* while precipitation in a local area within the region can be described as a *small-scale variable* [4].

In recent years, knowledge extraction from data has become increasingly popular, with many numerical forecasting models, typically falling into two major categories—chemical transport models (CTMs) [5,6] and conventional statistical methods [7]. However, due to data and model variability, data-driven knowledge extraction from high-dimensional, multifaceted data in such applications requires generalisations of global to regional or local conditions. This paper proposes a general framework for statistical downscaling based on a data-driven procedure for making use of *prior* information at *large scales* for generalisation at *local scales*. The idea is to develop statistical relationships between local and global attributes that map the latter to the former.

Mapping global conditions to local ecosystems and human habitats amounts to tracking and monitoring environmental dynamics in various geographical areas and their regional and global implications on human livelihood. Conventional applications of these methods are predominantly dimensional reduction in nature, designed to reduce spatial dimension of gridded model outputs without loss of essential spatial information. They heavily depend on unlabelled design matrix and distributional assumptions.

We propose a novel statistical downscaling framework for dealing with data and model variability. Its power derives from training and testing multiple models on multiple samples, narrowing down global environmental phenomena to regional discordance through dimensional reduction and visualisation. Repeated samples of ground-level ozone data are used to illustrate its power for narrowing down global environmental phenomena to regional discordance through an iterative process—train, validate, assess, repeat executed by an algorithm with built-in interdisciplinary computational power. The proposed algorithm is embedded with the applications agility for addressing the foregoing issues in different contexts, and we demonstrate how, by connecting data sources with end-users, it can be extended to weather-dependent and other applications with inherent data randomness and model variability. The paper is organised as follows. Study motivation, aims and objectives are given in Sect. 1, followed by methods in Sect. 2—describing data sources in Sect. 2.1 and implementation strategy in Sect. 2.2. Implementation, results and related discussions and computational environment are given in Sect. 3 and concluding remarks and future directions in Sect. 4.

1.1 Motivation and rationale

The key motivation for this paper is twofold. Firstly, the complexity of the dynamics of our ecosystem—particularly how it is affected by spatio-temporal activities, entails rigorous and robust strategies to understand and sustain. Narrowing down global environmental phenomena to regional discordance may provide insights into the phenomena and therefore help in planning for mitigating strategies. Real-life phenomena such as space–terrestrial, climatic conditions–human activities and pollutions fit into this context. Secondly, the notional functional relationship between the *large* and *small-scale* variables makes it possible for one to be described in terms of the other. Modelling this relationship via downscaling will correct the spatial mismatch between the variables without loss of useful information which conforms to the mechanics of dimensional reduction methods, such as principal component analysis (PCA) [8]. Many natural and physical phenomena render themselves readily for such relationships, which may not always be so obvious. This paper seeks to derive knowledge on local from regional variables via the objectives outlined below.

1.2 General and specific objectives

To account for variability, the paper adopts an ensemble downscaling approach and enhances its robustness capacity. Outlined below are the general and specific objectives of the paper.

1. *General: To develop a general framework for training and testing data models*
This is fundamental because quite often we adopt simplification due to our lack of adequate understanding of processes and/or phenomena. We need robust tools, techniques and analytical skills to scale down real-world data phenomena for a better understanding. The specific objectives for addressing data complexity are
 - 1.1. To narrow down global environmental phenomena to regional discordance.
 - 1.2. To extract knowledge on simple/local from complex/regional variables.
2. *General: Model assessment and optimisation* Data randomness and spatio-temporal parametrisation stipulate that intrinsic characteristics of the environment being modelled will be a function of specified attributes location, time and samples, say. In other words, air quality model parameters are not geophysical constants, and so they cannot be measured with 100% accuracy. More specifically, the paper seeks
 - 2.1. To carry out multiple sampling and testing for robustness.
 - 2.2. To provide an adaptive statistical downscaling tool.
 - 2.3. To demonstrate how the framework can be adapted to other applications.

Air quality modelling is an area of strong research interest for many reasons. Within the European Union (EU), for instance, member states are required to design appropriate air quality plans for zones where the air quality does not comply with specified limit values which has led to a wide range of air quality modelling tools and techniques [9]. A thorough assessment of the effects of local and regional emission on air quality and human health and to identify methodologies and their limitations is proposed in [9]. The evaluation, based on 59 one-off appraisal contributions from 13 EU member states, relies on indicators collected from questionnaires, which does not exhibit robustness of the models used. Recent spatio-temporal variations as in [10] provide good insights into spatio-temporal variations through simulations. The foregoing research objectives allude to an interdisciplinary approach to air quality modelling and to the general research philosophy that interdisciplinary formalisation of multifaceted environmental-related data, analytical methods and procedures potentially yields consistent, comprehensive, robust and veracious results as they help minimise the effect of data randomness [11]. The paper presents a comparative statistical downscaling framework based on the core ideas of dimensional reduction as described in the following exposition.

2 Methods

This section describes the data sources in Sect. 2.1 and the implementation strategy in Sect. 2.2. An **unsupervised** algorithm for **learning** rules from **training** spatio-temporal data for future replication elsewhere is developed and implemented via a data-adaptive algorithm. The section outlines the path to fulfilling the sequence of objectives in Sect. 1.2, results presentation, computational environment, discussions and making recommendations for extensions.

2.1 Data sources

We have 24-hour ground-level ozone predictions by the **R**egional **C**hEμical t**R**ansport **M**odel (**REAM**) adopted from [12–16] and [17,18]. The forecasts span across 104 grid cells over the south-eastern region of the USA with a resolution of 70 km × 21 vertical layers in the troposphere. The data, representing hourly ozone observations from various environmental stations in Fig. 1 maintained by the US Environmental Protection Agency (EPA), cover the summer period of June–August 2005, with regional simulations carried out during the last two weeks of May 2005.

The data are sampled from an $n \times p$ data matrix in which the rows, $n = 2203$, represent hours of day from 05:00hrs to 23:00hrs on 01 June, while the **columns 1 to 109 are the actual** ground-level ozone concentrations for 109 monitoring stations scattered across south-eastern USA and **columns 110 to 213 are the REAM forecasts** for the 104 grid cells covering the same south-eastern region. Comparability is on model performance deriving from multiple versions of unsupervised models for identifying underlying structures in sampled data. The goal is to extract improved knowledge on a *smaller subset* $\mathbf{x} \subset \mathbf{X}$ from the *super-set* $\mathbf{X} \supset \mathbf{x}$ through higher spatial resolution. Let

$$\begin{aligned} \mathbf{X} &= X_{i,j} \text{ where } i = 1, 2, 3, \dots, T; \quad j = 1, \dots, K; \quad K \ll T \\ \mathbf{x} &= x_{i,j} \text{ where } i = 1, 2, 3, \dots, t \leq T; \quad j = 1, \dots, k; \quad k \ll t \end{aligned} \quad (1)$$

represent the data in Fig. 1. The most obvious approach would be, for each monitoring station to regress hourly ozone observations on the grid cell that includes the station. Equation 2, where Z_t represent hourly ozone observations and $M_t(X)$ is the data-dependent REAM model output of the grid cell, exhibits how this can be implemented.

$$Z_t = \beta_0 + \beta_1 M_t + \epsilon_t. \quad (2)$$

The strategy is to repeatedly sample $\mathbf{x} \subset \mathbf{X}$ over multiple combinations of t and k , recording performance parameters. As noted earlier, modelling Eq. 2—i.e. the notional functional relationship between the *large* and *small-scale* variables, conforms to mechanics of dimensional reduction techniques such as PCA as expounded below.

2.2 Implementation strategy

The main idea is to use ensemble *unsupervised* models to map regional to local air quality patterns based on repeated sampling and performance assessment of multiple versions for different techniques. We adopt PCA—a technique that creates new uncorrelated variables by linearly combining the original $x_{i,j}$. The components are extracted in succession, with the first accounting for as much of the variability in the data as possible and each succeeding component accounting for as much of the remaining variability as possible. Ordering extracted components in descending order of the variability each accounts for, allows dimension reduction, and hence, only the first few components can be used to describe the original data. Each component is estimated as a weighted sum of the variables as in Eq. 3.

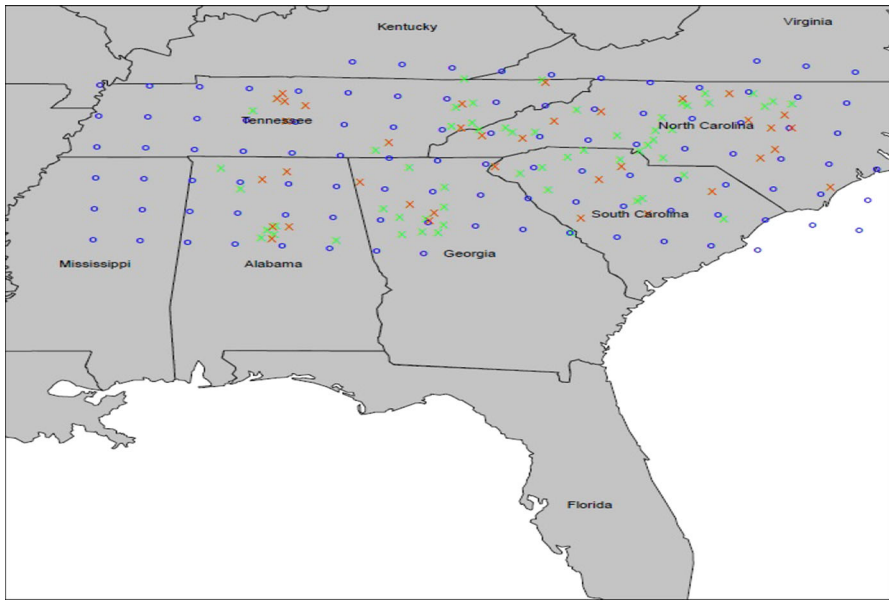


Fig. 1 Selected ozone monitoring stations in the south-eastern USA shown as \times s and REAM grid cells as circles [16]

$$C_{i,j} = \alpha_{ij}^T \mathbf{X} = \sum_{j=1}^K \sum_{i=1}^K \alpha_{ij}^T X_i, \tag{3}$$

where $i, j = 1, 2, \dots, K$ and $\alpha_1, \alpha_2, \dots, \alpha_K$ are the eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \dots \geq \lambda_K$ of the covariance matrix $x \in \mathbf{X}$. The coefficients in Eq. 3 are the directions or loadings of the components, and as such, they represent the weights associated with each of the i variables. For each sample of $k \subset K$ variables, we can extract k components and the decision to retain components is based on the amount of variance each component accounts for—typically, retaining components for which $\lambda \geq 1$. To put this into perspective, the grid cells of our REAM model hourly data represent an $n \times p$ source which will be linearly combined to form a fewer than p grid cells information from which can be generalised to any of the p cells. Further, if we treat $C_{i,j}^s$ as notional predictors of ozone, we obtain the multi-variate version of Eq. 2 as follows:

$$\mathcal{Z}_t = \beta_0 + \sum_{i=1}^K \beta_i M_K(t) + \epsilon_t, \tag{4}$$

where \mathcal{Z}_t is the ozone output, $M \leq p$ is the number of C_{ij}^s , β_i are model parameter and $M_1(t), M_2(t), \dots, M_K(t)$ are $C_{i,j}^s$. There are a number of issues of concern up to this point. Firstly, PCA extracts p $C_{i,j}^s$ and even though retention is determined by, say, the $\lambda \geq 1$ criterion, randomness due to sampling, produces variations in loadings. Secondly, this approach to downscaling fully relies on unlabelled data [19] and therefore proper interpretations of each cluster, i.e. each $C_{i,j}^s$, are required in order to incorporate the target variable in downscaling.

Algorithm 1 Unsupervised

```

1: procedure OPTIMISING NATURALLY ARISING STRUCTURES( $\mathcal{C}_{ij}^s$ )
2:  $\mathbf{X} := X_{i,j}; i = 1, 2, 3, \dots, T; j = 1, \dots, K; K \ll T$ 
3:  $\mathbf{x} := x_{i,j}; i = 1, 2, 3, \dots, t \leq T; j = 1, \dots, k; k \ll t$ 
4: Initialise:  $\Phi(\cdot) \leftarrow \text{NULL}; M := \text{Large constant}; \nu := 0; m := 0$ 
5: while  $m \leq M$  do
6:    $\mathcal{S}_m \leftarrow m^{\text{th}}$  sample  $x_{m,i,j} \subset X_{i,j}$ 
7:    $\mathcal{C}_{i,j}^s \leftarrow x_{m,i,j}$ 
8:   Update  $\Phi_m(\cdot) \leftarrow \mathcal{S}_m \{\alpha_{m,1}, \alpha_{m,2}, \dots, \alpha_{m,k}; \lambda_{m,1} \geq \lambda_{m,2} \dots \geq \lambda_{m,k}\}$ 
9:    $m := m + 1$ 
10:  for  $\nu = 1 \rightarrow \dim \Phi_m(\cdot)$  do
11:    Update  $\Phi_{\nu,m}(\cdot) \leftarrow \lambda_{m,1} \geq \lambda_{m,2} \dots \geq \lambda_{m,k} \leftarrow \mathcal{S}_m \{\mathbf{Av} [\alpha_{m-1,k}, \alpha_{m,k}]\}$ 
12:    Test optimality of the parameters in  $\mathcal{S}_{\nu,m}(\cdot)$ 
13:    if  $\Phi_{\nu,m}(\cdot) \cong \Phi_{\nu+1,m}(\cdot)$  then
14:       $\mathcal{C}_{\nu,i,j}^s \leftarrow \Phi_{\nu,m}(\cdot)$ 
15:    else
16:      if  $\Phi_{\nu,m}(\cdot) \not\cong \Phi_{\nu+1,m}(\cdot)$  then
17:         $\mathcal{C}_{\nu,i,j}^s \leftarrow \Phi_{\nu+1,m}(\cdot)$ 
18:      end if
19:    end if
20:  end for
21: end while
22: Output and interpret  $\mathcal{C}_{M,i,j}^s$ 
23: end procedure

```

Notice that **Algorithm 1** is generic, designed to cater for various types of unsupervised modelling and, hence, Φ is initialised as an empty set of extracted components—in case of PCA; map dimensionality, in case of SOMs or the number of clusters, in case of **K**-Means, say. The large constant M is basically the number of iterations or the number of different sized samples extracted from $X_{i,j}$, and it is determined by the investigator, typically depending on $X_{i,j}$ and the visualisations emerging from Φ . Note that ν is incremented and used to update and optimise Φ , a typical example being that of dimensionality as stated in the algorithm, using eigenvalues and vectors. The conditional checks in lines 13 through 18 of the algorithm underline the need for proper interpretation of the resulting structures—i.e. not fully conditioning retention to the mechanics of the methods and this helps minimise randomness due to sampling.

As each component retains elements of randomness, for each $\mathcal{C}_{i,j}^s$ we record loadings, eigenvalues and crucial model parameters which are subsequently used to determine the optimal number of naturally arising groupings in the data. This process is repeated many times over, storing and comparing relevant parameters as illustrated in **Algorithm 1**. The loadings determine the formation of $\mathcal{C}_{i,j}^s$ and their *meaningfulness* is determined by their magnitude, direction and domain knowledge at each step of the algorithm using numerical methods, graphical visualisation or both. Each $\mathcal{C}_{i,j}^s \propto Y$ and so the final $\mathcal{C}_{M,i,j}^s$ can be interpreted as class labels for supervised applications. Further illustrations of the computational power of the algorithm are provided in Sect. 3.3.

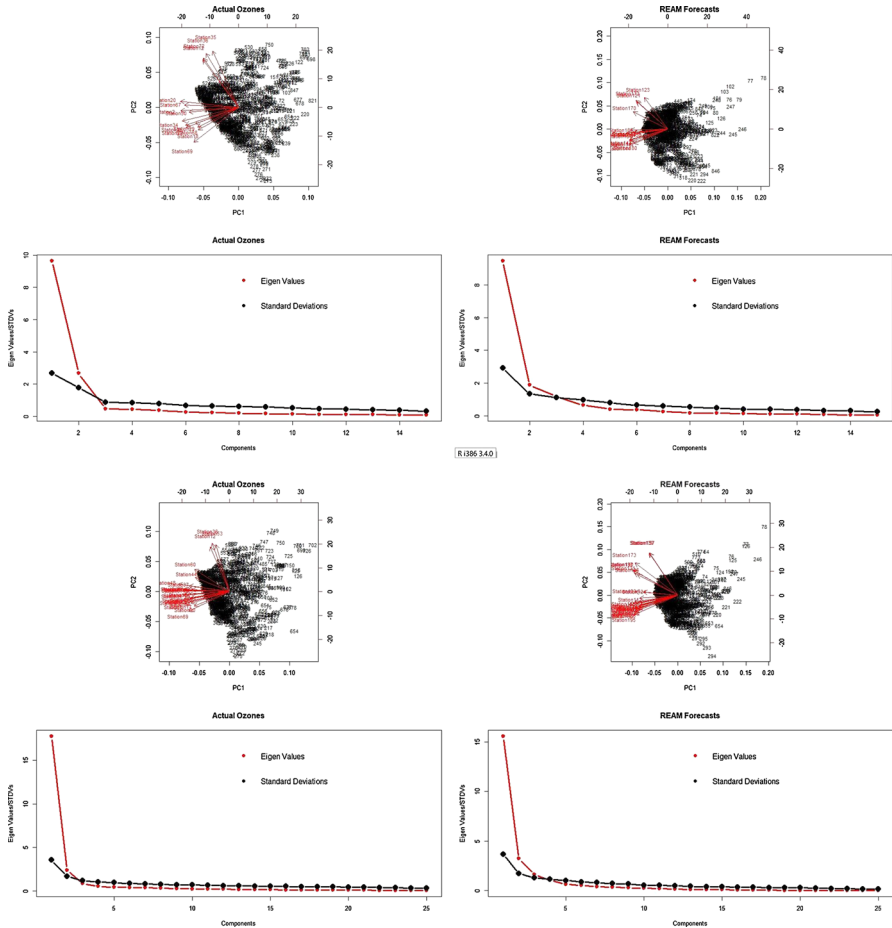


Fig. 2 PC patterns from sampled 15 observation stations in the south-eastern USA

3 Implementation, analyses and discussions

The typically low spatial resolution of air quality model forecasts and the relatively large spatial dimension of gridded model outputs entail the application of the foregoing methods. The underlying idea is that dimension reduction would retain most of the influential spatial and regional information provided by the air quality model. We present results from PCA and principal fitted components (PFC) downscaling. Thus, we combine the power of dimensional reduction with that of predictive modelling for the purpose of attaining robustness as outlined below.

3.1 Principal components and principal fitted components

PC patterns from sampled 15 observation station for the actual Ozone levels and REAM forecasts are presented in Fig. 2. Both cases exhibit two clear

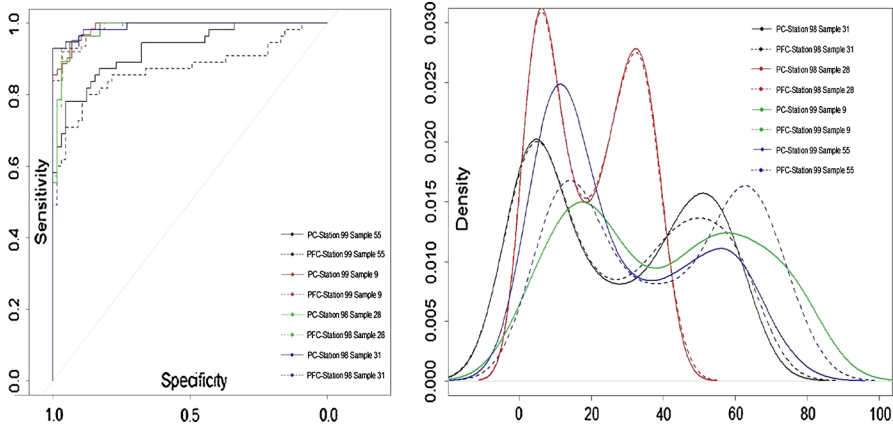


Fig. 3 Performance assessment for selected samples and stations

components, a detailed discussion of which is provided in Sect. 3.2 and via Figs 3, 4.

The two panels are selected plots from multiple samples of size 15 (top) and 20 (bottom), and in both cases, the bimodality of the vectors is evident. The bi-plots at the top of each panel represent the first 2 C_{ij}^s , and they are based on the loadings–eigenvectors relationship as described in Eq. 3. The high variation accounted by the first and second C_{ij}^s is evident from the panels. Of great values to the first C_{ij} are several stations—such as 35, 36, 72 and 12 for the actual plot and 121, 123 and 170 for the forecasts—having high absolute influences in the construction of both C_{1j} and C_{2j} . The pattern is repeated in the bottom panel, for stations 12, 36 and 53 and 36, 121 and 173, respectively. Although in most cases eigenvalues greater than 1 corresponded to 3 PCs, plots involving the third and higher PCs revealed near-random relationships, suggesting that very little variability remains after extracting C_{1j} and C_{2j} .

3.2 Performance assessment and key results

We assess the performance of both PCA and PFC different techniques using the receiver operating characteristic (ROC) curves [20] and kernel density estimations for the same parameters as shown in Fig. 3. The panel on the left-hand side of the plot shows the ROC curves from selected stations and $n \times p$ samples representing ozone readings in time. While a classifier is usually said to be optimal if it yields results in the north-western corner of the plot, classifier superiority must always be decided by taking variation into consideration which is precisely what Algorithm 1 seeks to achieve. By setting κ large, averaging of ROC curves can yield good, reliable results through cross-validation or bagging techniques. Multiple runs on κ provide scope for measuring the margins by which curves vary. The final part of the algorithm seeks to obtain optimal models based on the information obtained from repeated searches using transformed plots which may be achieved by evaluating multiple patterns of the type shown in the two panels of Fig. 3. For the ROC curves, rules may derive

from multiple tangent lines, *iso-performance* [21], optimising or otherwise. Extracted parameters from the plots can be used as inputs in repeated training. The relationships between different slopes corresponding to the different model versions may guide us choose a slope or slopes in different points which we can use to adapt the model architecture and so on. It is also possible to explore the impact of covariates on the ROC curves by examining the way they inter-cross. For the kernel density estimates of forecasts, the bi-modal patterns on the right-hand side suggest that ground-level ozone may be correlated with location. Hence, identifying the grid cell locations for the stations that constrain the system the most will make it possible to generalise local conditions from information gathered over the 70 km × 21 vertical layers resolution as discussed below.

Figure 4 displays the Empirical orthogonal function (EOF) plots—i.e. spatio-temporal patterns from the data. Like in PCA, orthogonality is with respect to the bases functions, with the *i*th basis function being orthogonal to, and capturing more variation than, the (*i* − 1)th. These spatio-temporal plots provide great insights into the strength of the extracted C_{ij}^s via the absolute values of loadings which display the locations in which the C_{ij}^s contribute more strongly or weakly.

The distribution of the leading EOFs in Fig. 4 reflects general variations of ozone levels—higher over emission regions and over land than over the open ocean. The top left-hand side panel (EOF1) loadings exhibit an ozone gradient decreasing towards the eastern coastline, while the top right-hand side panel (EOF4) gradient slides down towards the southern coastline, reflecting generally much lower ozone concentrations over the ocean than over land. Both the top right and bottom left panels (EOF2 and EOF3, respectively) reflect regional ozone distribution patterns, with low variations in Alabama, Northern Georgia, Mississippi and Tennessee likely due to meteorological conditions.

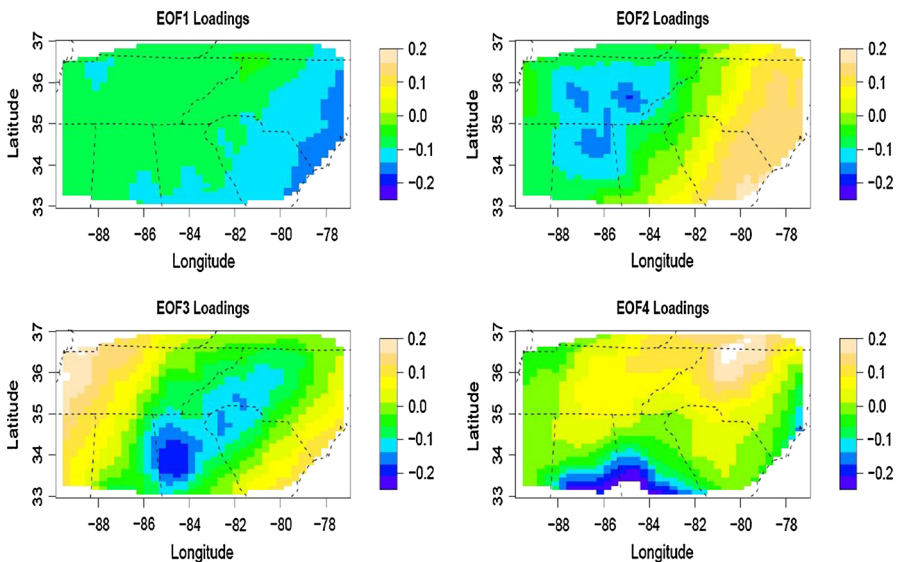


Fig. 4 Plots of the first four empirical orthogonal functions of the REAM from 6–25 June 2005

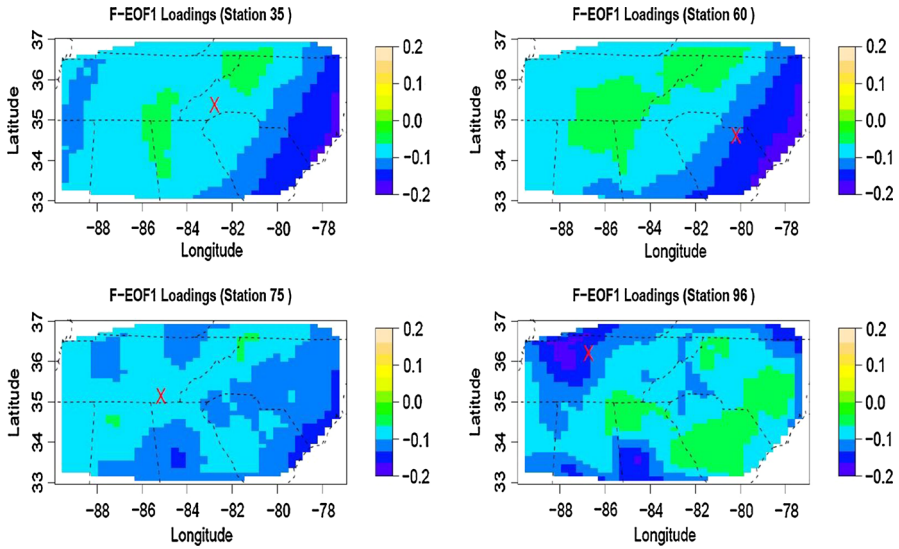


Fig. 5 Plots of the first four fitted empirical orthogonal functions of the REAM from 6 to 25 June 2005

Table 1 Root-mean-squared errors: training period is 6–25 June 2005 and validation period is 26–30 June 2005

Model	All stations (%)	Station 29 (%)	Station 84 (%)	Station 107 (%)
PC	9.35	11.31	10.27	9.89
PFC	9.13	9.86	10.36	8.91

Like the EOFs, F-EOFs display locations at which PFCs contribute more strongly or weakly. Figure 5 shows the first F-EOFs of REAM outputs corresponding to selected stations over the period 6–30 June 2005. A west to east sliding gradient in stations 35, 60 and 75 exhibits low ozone levels on the eastern coastline, while station 96 exhibits a mixture of low ozone levels in the north-western and southern parts. The PFC appears to be quite capable of reducing the dimension of the problem and capturing spatio-temporal patterns more efficiently than just the C_{ij}^s .

Numerical comparisons of the two models are presented in Table 1, based on their respective root-mean-squared error (RMSE) at selected stations and overall. Station 29 was fitted with 18 C_{ij}^s 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) which shows an overall PFC superiority although PFC was actually outperformed in 36 stations—i.e. 38.29% out of the total. One possible explanation is the gridding of cells covering the locations of those stations as some stations may be located at the grid cells borders—analogueous to near-overlapping clusters which reinforces the assertion of regional variations.

As the locations of these stations are known, any set can be adopted as a regional set from which one or more local stations may be isolated for testing by simply masking actual (observed) values. We illustrate the computational set-up of this adoption, isolation and integration, based on the mechanics of Algorithm 1, in the next exposition.

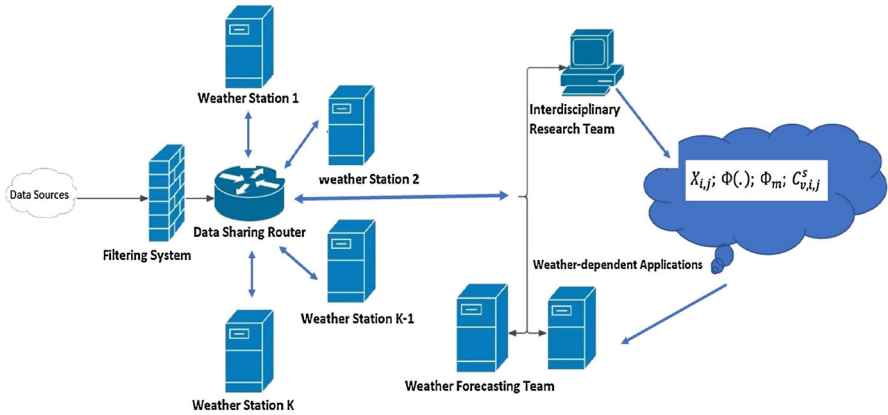


Fig. 6 A graphical illustration of the interdisciplinary computational environment

3.3 Computational environment

Algorithm 1 seeks to balance model accuracy and reliability, and in this particular application, one of the focal characteristics is resolution. Attaining high-resolution seasonal and climate change forecasts has always been at the heart of climate-related research aimed at enhancing efficiency in many applications, including agriculture and energy. The algorithm emphasises the interdisciplinary nature of statistical downscaling and its inherent requirement to access, process and share large volumes of high-dimensional heterogeneous data. Successful application of the algorithm will therefore depend on how these computational environment conditions are fulfilled. Figure 6 graphically illustrates how weather-dependent applications benefit from an interdisciplinary computational power that connects data sources with end-users. This work was accomplished by combining the data acquisition power of the weather data sources. An integrated interdisciplinary computation is the approach our work seeks to open research paths to.

Implementation of the interdisciplinary layout in Fig. 6 relies on distributed computing systems providing real-time data sharing and allowing model training and testing across the computational nodes. Its ultimate usefulness is in being accessible to weather-dependent applications. There are several tools in use today to meet such requirements—Hadoop and Apache Spark are currently two of the most popular open-source distributed computing frameworks. Algorithm- 1 was implemented using the statistical open source, **R**, which is readily adaptable to the foregoing distributed system.

4 Concluding remarks

The paper proposes a framework for deriving knowledge on local from regional data attributes. An algorithm is presented in a specific framework that provides scope for model selection. Results show that given air quality data of similar structure, the algo-

rithm can be used to order the models in terms of optimality. Its mechanics derive from notional functional relationship between the *large*- and *small-scale* variables that makes it possible for one to be described in terms of the other, and it seeks to minimise variability through repeated samples and multiple learning models. The generic graphical illustrations in Fig. 3 exhibit how rules can be learnt from training data and applied to new, previously unseen, data rules which are driven by domain knowledge and therefore make the approach readily adaptable to other applications. However, while many natural and physical phenomena render themselves readily adaptable to notional functional relationship between the *large* and *small-scale* variables, some useful information may remain hidden in the data attributes. For instance, while solar energy may not produce environmental pollution, it may indirectly impact the environment in that its use may induce some potentially hazardous and toxic materials and chemicals, despite strict environmental laws and regulations. Large-scale farming may produce more food but it may also leave lasting effects on the ecosystems—both issues can be addressed via downscaling techniques.

The algorithm has the potential for deployment of different data mining models, and it has scope for extension to other applications, focusing on how to carry out comparisons, address inconsistencies, draw conclusions in cases of partial agreement and account for the effect of data and model variability. Applications of the algorithm may be extended to interpolations of mean fields for oceanographic variables at various ocean levels to provide statistical downscaling of average readings at various ocean depth levels for each variable of interest—i.e. a variable fulfilling specific attributes. Finally, its interdisciplinary implementation layout, amenable to distributed computing, not only enhances its computational power and real-time data sharing, but it is also a perfect environment for model training, validation and testing across the nodes. We expect that this work will lead to new research avenues in various areas.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Lockwood M, Harrison R, Woollings T, Solanki S (2010) Are cold winters in Europe associated with low solar activity? *Environ Res Lett* 5(2):1–7
2. Weaver CP, Cooter E, Gilliam R, Gilliland A, Grambsch A, Grano D, Hemming B, Hunt SW, Nolte C, Winner DA, Liang X-Z, Zhu J, Caughey M, Kunkel K, Lin J-T, Tao Z, Williams A, Wuebbles DJ, Adams PJ, Dawson JP, Amar P, He S, Avise J, Chen J, Cohen RC, Goldstein AH, Harley RA, Steiner AL, Tonse S, Guenther A, Lamarque J-F, Wiedinmyer C, Gustafson WI, Leung LR, Hogrefe C, Huang H-C, Jacob DJ, Mickley LJ, Wu S, Kinney PL, Lamb B, Larkin NK, McKenzie D, Liao K-J, Manomaiphiboon K, Russell AG, Tagaris E, Lynn BH, Mass C, Salathé E, O’neill SM, Pandis SN, Racherla PN, Rosenzweig C, Woo J-H (2009) A preliminary synthesis of modeled climate change impacts on U.S. regional ozone concentrations. *Bull Am Meteorol Soc* 90(12):1843–1863 2015/01/20
3. WMO (2016) The state of greenhouse gases in the atmosphere based on global observations through 2015. *WMO Greenhouse Gas Bulletin*, WMO, No. 12, 24 October 2016. ISSN 2078-0796
4. Benestad E, Hanssen-Bauer I, Chen D (2008) *Empirical—statistical downscaling*. World Scientific Publishing, Singapore

5. Rotman DA, Atherton CS, Bergmann DJ, Cameron-Smith PJ, Chuang CC, Connell PS, Dignon JE, Franz A, Grant KE, Kinnison DE, Molenkamp CR, Proctor DD, Tannahill JR (2004) Impact, the LLNL 3-d global atmospheric chemical transport model for the combined troposphere and stratosphere: model description and analysis of ozone and other trace gases. *J Geophys Res Atmos* 109(D4):D04303
6. Jacob DJ (2004) *Introduction to atmospheric chemistry*. Princeton University Press, Princeton
7. Sahu S, Gelfand AE, Holland DM (2007) High resolution space-time ozone modeling for assessing trends. *J Am Stat Assoc* 102(480):1221–1234
8. Jolliffe I (2013) *Principal component analysis*. Springer, Berlin
9. Thunis P, Miranda A, Baldasano J, Blond N, Douros J, Graff A, Janssen S, Juda-Rezler K, Karvosenoja N, Maffei G, Martilli A, Rasoloharimahefa M, Real E, Viaene P, Volta M, White L (2016) Overview of current regional and local scale air quality modelling practices: assessment and planning tools in the EU. *Environ Sci Policy* 65(Supplement C):13–21 (**Multidisciplinary research findings in support to the EU air quality policy: experiences from the APPRAISAL, SEFIRA and ACCENT-Plus EU FP7 projects**)
10. Tao X, Yixuan Z, Guannan G, Bo Z, Xujia J, Qiang Z, Kebin H (2017) Fusing observational, satellite remote sensing and air quality model simulated data to estimate spatiotemporal variations of PM_{2.5} exposure in china. *Remote Sens* 9(3):221
11. Mwitondi K, Said R (2013) A data-based method for harmonising heterogeneous data modelling techniques across data mining applications. *Stat Appl Probab* 2(3):293–305
12. Choi Y, Wang Y, Cunnold D, Zeng T, Shim C, Luo M, Eldering A, Bucsela E, Gleason J (2008) Spring to summer northward migration of high O₃ over the western North Atlantic. *Geophys Res Lett* 35:L04818
13. Choi Y, Kim J, Eldering A, Osterman G, Yung YL, Gu Y, Liou KN (2009) Lightning and anthropogenic NO_x sources over the United States and the western North Atlantic Ocean: impact on OLR and radiative effects. *Geophys Res Lett* 36:L17806
14. Choi Y, Osterman G, Eldering A, Wang Y, Edgerton E (2010) Understanding the contributions of anthropogenic and biogenic sources to CO enhancements and outflow observed over North America and the western Atlantic Ocean by TES and MOPITT. *Atmos Environ* 44:2033–2042
15. 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
16. Wang Y, Hao J, McElroy MB, Munger JW, Ma H, Chen D, Nielsen CP (2009) Ozone air quality during the 2008 Beijing olympics: effectiveness of emission restrictions. *Atmos Chem Phys* 9:5237–5251
17. Zhao C, Wang Y (2009) Assimilated inversion of NO_x emissions over east Asia using OMI NO₂ column measurements. *Geophys Res Lett* 36(6):1–5
18. 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* . <https://doi.org/10.1029/2009JD012745>
19. Cook RD, Forzani L (2008) Principal fitted components for dimension reduction in regression. *Stat Sci* 23(4):485–501
20. Robin X, Turck N, Hainard A, Tiberti N, Lisacek Frdrique, Sanchez Jean-Charles, Miller Markus (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform* 12:77
21. Provost F, Fawcett T (2001) Robust classification for imprecise environments. In: *Proceedings of Machine Learning: 25th Anniversary of Machine Learning*, vol 42(3), pp 203–231