Atmospheric Environment 199 (2019) 415–422

Contents lists available at ScienceDirect

Atmospheric Environment

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

A statistical emulator for multivariate model outputs with missing values

Francesco Finazzi[a], Yoana Napier[b,*], Marian Scott[b], Alan Hills[c], Michela Cameletti[d]

[a] Dept. of Management, Information and Production Engineering, University of Bergamo,Viale Marconi, 5 24044 Dalmine, BG, Italy
[b] School of Mathematics and Statistics, University of Glasgow, University Place, Glasgow, G12 8TA, UK
[c] Scottish Environment Protection Agency, Angus Smith Building, 6 Parklands Avenue, Eurocentral, Holytown, North Lanarkshire, ML1 4WQ, UK
[d] Dept. of Management, Economics and Quantitative Methods, University of Bergamo, via dei Caniana 2, Bergamo, 24127, Italy

ARTICLE INFO

Keywords:
EM algorithm
Sensitivity analysis
D-STEM software
Nitrogen dioxide

ABSTRACT

Statistical emulators are used to approximate the output of complex physical models when their computational burden limits any sensitivity and uncertainty analysis of model output to variation in the model inputs.

In this paper, we introduce a flexible emulator which is able to handle multivariate model outputs and missing values. The emulator is based on a spatial model and the D-STEM software, which is extended to include emulator fitting using the EM algorithm. The missing values handling capabilities of the emulator are exploited to keep the number of model output realisations as low as possible when the computing burden of each realisation is high.

As a case study, we emulate the output of the Atmospheric Dispersion Modelling System (ADMS) used by the Scottish Environment Protection Agency (SEPA) to model the air quality of the city of Aberdeen (UK). With the emulator, we study the city air quality under a discrete set of realisations and identify conditions under which, with a given probability, the 40 $\mu g\ m^{-3}$ yearly average concentration limit for NO_2 of EU legislation is not exceeded at the locations of the city monitoring stations.

The effect of missing values on the emulator estimation and probability of exceedances are studied by means of simulations.

1. Introduction

Air pollution is one of the most important environmental problems because of its impact on people's health. According to the Royal College of Physicians (RCP) report from 2016, an estimated 40,000 deaths a year in the UK are attributed to ambient (outdoor) air pollution (Royal College of Physicians, 2016). The United Kingdom (UK) has an obligation to provide clean air to its citizens as defined by Directive 2008/50/EC (EC, 2008) of the European Union (EU). However, the UK government is struggling to tackle air pollution, which caused ClientEarth to sue them for a third time for not implementing changes as fast as possible (Carrington and Taylor, 2017). The Scottish government works closely with the Department for Environment, Food and Rural Affairs (DEFRA) to develop "domestic policies and initiatives to improve air quality and reduce risks to human health" (The Scottish Government, 2016). In Scotland, the air pollution objectives are presented in the Cleaner Air for Scotland Strategy (CAFS). It is crucial to note that the goal is not only to achieve compliance with the air pollution guidelines but also to be able to ensure that the limits will not be broken under different meteorological conditions.

Outdoor air pollution is measured typically by a network of reference method automatic monitors such as the AURN (Automatic Urban and Rural Network) or diffusion tubes. Reference method monitors (stations) are expensive and they require secure boxes and power, so the number of installed stations is quite small. Diffusion tubes are low-cost and do not need power, so they can be installed in more locations. However, they only report monthly concentrations and uncertainties are larger than automatic monitors (Local Air Quality Management (LAQM, 2008). In general, monitoring air quality using automatic monitors or diffusion tubes is not enough to fully understand the complex space-time dynamics of the pollutants, so may not be sufficient to implement actions aimed at reducing the pollutant concentrations.

The simplest and cheapest way to understand the pollutants' movement and concentrations is to use a computer model. However, producing many different realisations (from different combinations of input factors) using a model is time-consuming and often expensive. A common approach to reduce the running time for simulations is to use an emulator. The emulator can "predict the output at inputs not previously run on the simulator" (Bastos and OHagan, 2009). It is crucial

[*] Corresponding author.
E-mail address: y.napier.1@research.gla.ac.uk (Y. Napier).

https://doi.org/10.1016/j.atmosenv.2018.11.025
Received 13 July 2018; Received in revised form 7 November 2018; Accepted 12 November 2018
Available online 19 November 2018

to note that using an emulator does not eliminate the need for simulations but simply reduces the number of model simulations to be run. Blanc (2017) uses a simple emulator, which is applied to multiple responses, for gridded crop models. The performance of the emulator is assessed by using both in- and out-of-sample validation. Bayarri et al. (2009) have built an emulator for quantifying volcanic hazards, which uses a mixture of frequentist and Bayesian estimation techniques and the emulator can be used to calculate the probability of catastrophic events. The one disadvantage of the Bayarri et al. (2009) emulator is that it does not produce bias and uncertainty estimates for the outputs. The emulator presented in Fricker et al. (2013) propose a multivariate emulator, which models a small number of responses at the same time. The multivariate emulator is applied to a simple climate model and information, which would be lost implementing multiple univariate emulators, was captured. However, the multivariate emulator opens the question of what combination of inputs should be used. Bastos and O'Hagan, Park (1994) and Fricker and others suggest using a Latin hypercube. In this work, a Latin Hypercube design is applied.

In this paper, a multivariate emulator is developed to jointly model a set of variables which are output from the physical model, provided that the number of variables is not large ($\geq$ 20). Additionally, the emulator calculates the uncertainty of its parameter estimates, which is key for the application. Depending on the application, variables can be either heterogeneous (variables with different unit of measure) or homogeneous (the same variable observed at different points in space or time). The emulator is implemented by extending the D-STEM software (Finazzi and Fassò, 2014), which allows multivariate data sets to be handled and to estimate multivariate statistical models from data sets affected by missing values by extending the work of Fricker et al. (2013). In this work, the capability of D-STEM in handling missing values is exploited. In general, missing values are unobserved realisations but in this work missingness is designed in order to reduce the computational burden. Instead of reducing the number of realisations, the fact that variables are jointly modelled is exploited and, for each given realisation, missing values are forced on some variables. The more the variables are cross-correlated the more it is convenient to reduce the number of observed variables, for a given realisation, rather than to reduce the number of realisations. In the special case of perfect correlation, only one variable needs to be observed for each realisation. This approach is used when the model output takes hours for each variable for each realisation.

A case study on air quality management for the city of Aberdeen (UK) is considered. Aberdeen is an oil industry centre and until 2012, all the traffic passed through the city, making the city intriguing to investigate. Air quality has improved in Aberdeen a great deal since 2012, mainly due to fewer old heavy goods vehicles (HGVs) and generally less traffic. The air quality of Aberdeen is modelled by the Scottish Environment Protection Agency (SEPA) using an atmospheric dispersion model (ADMS) (Pasquill, 1979; Turner, 1994), which is used to assess the pollutant concentration on the basis of anthropogenic and environmental factors. In particular, ADMS allows the pollutant concentration at any desired location in space and time to be estimated using a set of inputs. The measurements from ADMS are validated to AURN monitors measurements as seen in (Hood et al., 2018). In this paper, the ADMS-Urban extension is used as this model is specifically developed to use for modelling air pollution in cities. The ADMS-Urban predictions for Aberdeen are compared to the actual data in (SEPA and Natural Scotland, 2017). In Scotland, the ADMS-Urban model is often used to make recommendations regarding improving Air Quality but these studies have been limited to small areas or road sections as seen in (Stratton, 2014) and (Callaghan, 2014). The computational burden related to ADMS-Urban, however, is often high and the number of realisations that can be addressed is low, limiting any sensitivity analysis of model outputs to variation in the model inputs. Additionally, a small number of realisations may not be enough to solve the so-called *inverse problem*, where the goal is to identify all the combinations of model inputs that guarantee model outputs satisfying a given criterion. A similar study to this one is presented by Mallet et al. (2018), where the ADMS-Urban is emulated on an hourly basis with thousands of runs of the ADMS-Urban model. This paper aims to develop an emulator, which uses a smaller number of runs of the ADMS-Urban model but only predicts the annual average concentration of NO_2.

In the case study, an airborne pollutant, NO_2, is taken into account and probabilities of compliance with the limits of the 2008 ambient air quality directive (2008/50/EC) (EC, 2008) are given at different locations across Aberdeen. There are three main questions to answer:

(i) what are the annual yearly average pollutant concentrations at multiple spatial locations under different realisations?
(ii) given the emulated yearly average pollutant concentrations at multiple spatial locations, what is the risk of non-compliance with the EU directive due to factors which cannot be controlled (meteorological factors such as wind speed, wind direction, temperature)?
(iii) what are the effects of missing values on the output from the emulator?

In this paper, the concentration at each location is considered as a variable as the ADMS-Urban produces a different concentration level depending on the features of the spatial location. Although the pollutant concentration is independently computed at each location, variables are presumed to be cross-correlated as nearby locations are expected to share similar features. This justifies the adoption of the multivariate emulator in order to answer (i) and using inverse regression to answer (ii). The rest of the paper is organised as follows: Section 2 introduces the multivariate emulator and the inverse problem. The case study is discussed in Section 3 while the effect of missing values on model estimation is detailed in Section 4. Conclusions are given in Section 5.

2. Methodology

In this section of the paper the developed methodology is presented. Subsection 2.1 gives the theoretical background for the emulator and Subsection 2.2 presents the theory for the inverse problem.

2.1. Emulator

The multivariate emulator, built to answer question (i) from the introduction, belongs to the class of Gaussian emulators (Sacks et al., 1989). The input values are formed for different input variable combinations called realisations. In this paper, homogeneous variables are used as only one pollutant is considered but concentrations are computed at many locations. Let $\boldsymbol{y}_i = (y_{1,i},\dots,y_{M,i})'$ be the annual mean pollutant concentration estimated by the ADMS-Urban at M monitoring stations under the i-th realisation.

Each realisation is defined by the vector of independent variables (inputs) $\boldsymbol{x}_i = (x_{1,i},\dots,x_{B,i})^T$, where each $x_{b,i}$ takes values in the range $[x_b^l, x_b^u]$, $b = 1, \dots, B$. The emulator equation is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{w}(\boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i \quad (1)$$

- $\boldsymbol{X}_i = blockdiag(\tilde{\boldsymbol{x}}_{1,i},\dots,\tilde{\boldsymbol{x}}_{M,i})$ is a block-diagonal matrix with $\tilde{\boldsymbol{x}}_{m,i}$, $m = 1, \dots, M$, is a $1 \times \tilde{B}_m$ vector including the constant, all or a subset of the elements of $\boldsymbol{x}_i$ and, possibly, interactions of any order;
- $\boldsymbol{\beta} = (\boldsymbol{\beta}_1{}',\dots,\boldsymbol{\beta}'_M)'$ is the vector of parameters;
- $\boldsymbol{A}$ is a diagonal matrix of parameters;
- $\boldsymbol{w} = (w_1,\dots,w_M)'$ is a zero-mean multivariate Gaussian process (GP) with marginal unit variance;
- $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma}_\varepsilon)$ is an unstructured Gaussian random error independent across the M components of $\boldsymbol{y}_i$.

The multivariate Gaussian process $\boldsymbol{w}$ has the following cross-correlation matrix function

$$\boldsymbol{\Gamma_w} = \boldsymbol{V} \otimes \ \rho(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) \tag{2}$$

where $\boldsymbol{V}$ is a valid correlation matrix and ρ is the product exponential correlation function (Bayarri et al., 2009)

$$\rho(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \exp\left(-\frac{|x_{1,i} - x_{1,j}|}{\theta_1}\right)\cdot...\cdot\exp\left(-\frac{|x_{B,i} - x_{B,j}|}{\theta_B}\right),$$

for each $\boldsymbol{x}_i, \boldsymbol{x}_j \in [x_1^l, x_1^u] \times ...\times[x_B^l, x_B^u] = \mathscr{X} \subset \mathbb{R}^B$, $\boldsymbol{\theta} = (\theta_1,...,\theta_B)$. Note that the geographic distance between the monitoring stations is not explicitly modelled and the cross-correlations across stations are handled by $\boldsymbol{V}$. The model parameter set Ψ is estimated by means of the expectation-maximization algorithm (EM) (Fassò and Finazzi, 2011) and is defined to be

$$\Psi = \{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_\varepsilon^2, \boldsymbol{V}, \boldsymbol{\theta}\}$$

with $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_M)' = diag(\boldsymbol{A})$ and $\boldsymbol{\sigma}_\varepsilon^2 = (\sigma^2_{\varepsilon,1},...,\sigma^2_{\varepsilon,M})' = diag(\boldsymbol{\Sigma}_\varepsilon)$.

Let $\tilde{\boldsymbol{y}}'_i$ be the non-missing values subvector of $\boldsymbol{y}_i'$. Considering the joint vector $\tilde{\boldsymbol{y}} = (\tilde{\boldsymbol{y}}_1', .., \tilde{\boldsymbol{y}}'_N)$ of N realisations. For any new realisation $\boldsymbol{x} \in \mathscr{X}$, the pollutant concentration at the M monitoring stations is given by

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{A}}\hat{\boldsymbol{w}}(\boldsymbol{x}) \tag{3}$$

where $\hat{\boldsymbol{w}}(\boldsymbol{x}) = E_{\hat{\Psi}}[\boldsymbol{w}(\boldsymbol{x})|\tilde{\boldsymbol{y}}]$ (see Fassò and Finazzi, 2011) and $\hat{\Psi}$ is the estimated parameter set.

The D-STEM software estimates multivariate space-time models from data collected over space and time. In this work, D-STEM has been extended to accommodate multivariate emulators. From a software point of view, spatial models and emulators are very similar objects and they only differ in the way the GP is adopted. In a spatial model, the GP is defined over the geographic space while in an emulator the GP is defined over the explored region $\mathscr{X}$.

2.2. Inverse problem

The emulator estimated in the previous section is used here to solve the inverse problem and hence, address question (ii) stipulated in the Introduction. Given a value L, the aim is to find, for the M spatial locations, the region $\mathscr{X}_m \subset \mathscr{X}$, $m = 1, ..., M$, where the probability of exceedance L is below δ.

Let $\boldsymbol{y}^0(\boldsymbol{x}) = (y_1^0(\boldsymbol{x}),...,y_M^0(\boldsymbol{x}))'$ be the yearly average pollutant concentration vector in any future year under a given realisation $\boldsymbol{x}_i$. The vector $\boldsymbol{y}^0(\boldsymbol{x})$ is essentially equal to $\hat{\boldsymbol{y}}(\boldsymbol{x}) + \boldsymbol{\varepsilon}_i$ as we are interested in the individual realisation of the pollutant concentrations.

In order to evaluate $P(\boldsymbol{y}^0(\boldsymbol{x}) > L)$ it is necessary to assess the individual prediction variance $Var(\boldsymbol{y}^0(\boldsymbol{x}))$. The prediction variance for $y_m^0(\boldsymbol{x})$, $m = 1, ..., M$, is given by

$$Var_{\hat{\Psi}}[y_m^0(\boldsymbol{x})] = \hat{\sigma}^2_{\varepsilon,m}[1 + \tilde{\boldsymbol{x}}'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{x}}] + \hat{\alpha}_m^2\hat{\sigma}^2_{w_m}(\boldsymbol{x}) \tag{4}$$

where $\tilde{\boldsymbol{x}}$ is constructed similarly to $\tilde{\boldsymbol{x}}_{m,i}$ and $\tilde{\boldsymbol{X}}$ is the $N \times \tilde{B}_m$ matrix of the stacked $\tilde{\boldsymbol{x}}_{m,i}$ vectors. Moreover, $\hat{\sigma}^2_{w_m}(\boldsymbol{x}) = \mathrm{Var}_{\hat{\Psi}}(w_m(\boldsymbol{x})|\boldsymbol{Y})$ is the prediction variance of $\hat{w}_i(\boldsymbol{x})$ and it is computed as detailed in Fassò and Finazzi (2011).

The distribution of $y_m^0(\boldsymbol{x})$ is

$$y_m^0(\boldsymbol{x}) \sim N\left(\tilde{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_m + \hat{\alpha}_m\hat{w}_m(\boldsymbol{x}), Var_{\hat{\Psi}}[y_m^0(\boldsymbol{x})]\right) \tag{5}$$

and $P(y_m^0(\boldsymbol{x}) > L)$ is easily evaluated for each $\boldsymbol{x} \in \mathscr{X}$.

The region $\mathscr{X}_m$ is given by

$$\mathscr{X}_m = \{\boldsymbol{x} \in \mathscr{X} | P(y_m^0(\boldsymbol{x}) > L) < \delta\}. \tag{6}$$

and it is assessed under the estimated parameter set $\hat{\Psi}$.

Following the bootstrap approach detailed in Finazzi et al. (2013), the uncertainty of $\mathscr{X}_m$ is evaluated considering the asymptotic distribution of $\hat{\Psi}$ which is given by $N(\hat{\Psi}, \hat{\mathfrak{I}}^{-1})$, where $\hat{\mathfrak{I}}$ is the approximated Fisher information matrix and, with abuse of notation, $\hat{\Psi}$ are the model parameters stacked in a vector. The asymptotic distribution is sampled R times to obtain the collection $\boldsymbol{\Psi} = (\Psi^{(1)},...,\Psi^{(R)})$. For each vector in $\boldsymbol{\Psi}$, the distribution of $y_m^0(\boldsymbol{x})$ is recomputed and a new region $\mathscr{X}_m$ is evaluated. The collection of R regions is eventually used to compute bootstrap confidence intervals as shown in Section 4. Hence, the emulator provides uncertainty measures.

3. Aberdeen case study

3.1. Introduction

In this section, the multivariate emulator introduced in Section 2 is used to emulate the output of the Atmospheric Dispersion Modelling System (ADMS), (Stocker et al., 2012), developed by Cambridge Environmental Research Consultants of the UK and adopted by SEPA. There are $M = 6$ monitoring stations and $i = 100$ realisations. SEPA uses the ADMS to predict pollutant concentrations at the city level, at hourly temporal resolution and at 75 m spatial resolution. This is done to understand under which conditions (of emissions, traffic, meteorology, etc.) the annual ambient pollutant concentration complies with the limits of the air quality directive (2008/50/EC). Since compliance with the directive is based on the concentrations measured at the monitoring stations, the ADMS output analysis is restricted to the spatial locations of the stations, installed in Aberdeen (Fig. 1). A main advantage of the emulator is that it directly provides an annual average for the pollutants, whereas when using the ADMS annual average of the pollutant has to be estimated from hourly simulations.

NO_2 is considered because it is a pollutant for which at some locations the annual limit is exceeded. The regulations state that the yearly average NO_2 concentration must be below 40 $\mu g\, m^{-3}$. However, in Aberdeen, there are three (Market Street 2, Union Street and Wellington Road) monitoring stations which regularly fail to comply with the regulations as their annual averages are above 40 $\mu g\, m^{-3}$ as seen in Table 1. The data were accessed from http://www.scottishairquality.co.uk/data/data-selector on 22/02/2018.

3.2. Realisations

Each realisation of the ADMS NO_2 output is identified by three independent variables, namely pollutant emission (x_1), wind speed (x_2) and wind direction (x_3). Year 2012 is taken as baseline and all the realisations are defined as variations from the baseline. Variations in the emission are presumed to be uniform across all the emission sources while wind speed and wind direction are presumed to vary across the city as the ADMS model takes into account the buildings and street canyons around each of the stations. The range of variation is [−50%, +20%] for the pollutant emission, [−20%, +20%] for the wind

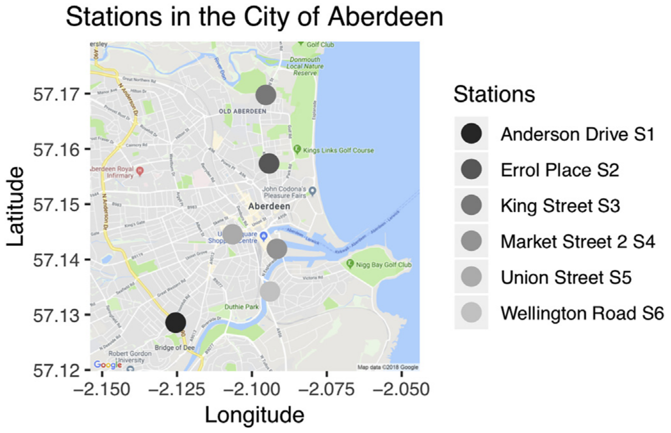

Fig. 1. Aberdeen automatic air quality monitoring network stations.

Table 1
The annual means for NO_2 concentration for each of the six stations across Aberdeen for period 2010–2017. The stations which have broken the annual average limit of 40 $\mu g\,m^{-3}$ are in red. Additionally, the bottom line has the annual wind speed variations for each of the years discussed.

Station	2010	2011	2012	2013	2014	2015	2016	2017
Anderson Drive (S1)	27.00	23.34	30.38	22.32	25.97	21.97	21.44	16.98
Errol Place (S2)	22.19	22.75	21.00	20.30	21.70	22.65	20.83	21.19
King Street (S3)	29.48	32.02	29.17	28.37	27.40	27.60	27.64	23.25
Market Street 2 (S4)	43.90	40.27	44.04	43.46	40.33	35.78	35.40	30.63
Union Street (S5)	57.99	43.53	52.84	48.26	46.63	46.10	42.96	39.66
Wellington Road (S6)	52.45	51.28	59.02	51.94	47.70	40.30	45.62	39.51
Wind Speed Variation	3.4%	10.8%	0.0%	21.6%	21.4%	24.3%	8.0%	22.0%

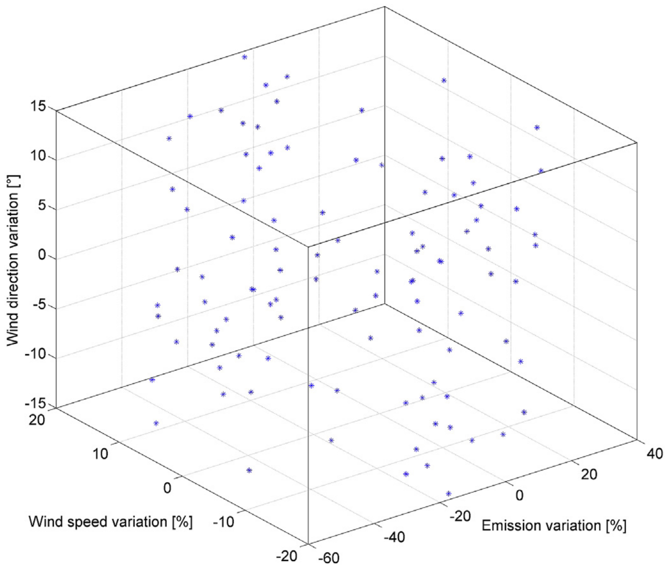

Fig. 2. Latin Hypercube Sampling design in the space $\mathscr{X}$ composed on 100 realisations.

speed and [−15°, +15°] for the wind direction. The ranges of the variations of each input were chosen by experts as they believe that in Aberdeen values outside these ranges are highly unlikely to occur. The annual wind speed variations from 2010 to 2017 relative to 2012 are presented in the bottom row of Table 1. The percentage of variation was calculated by taking the difference between the annual average wind speed for any year (2010–2017) and the year 2012, diving by the annual average for 2012 and multiplying by 100%. All the percentage variations in Table 1 are positive because 2012 was a year with very low annual average wind speed. From these results, a 20% variation around zero in average wind speed was considered appropriate for the simulations. For each combination $\boldsymbol{x} = (x_1, x_2, x_3)$, variations are applied to the hourly figures of the baseline, producing a new realisation, and the annual average is then calculated.

In a realisation, decision makers may only change pollutant emission while wind speed and wind direction are not controllable. Nonetheless, it is important that the study includes variables which are known to have an impact on the pollutant concentration such as temperature and wind. This allows emulating the risk that the actions taken to achieve a given target are nullified by random variations of the non-controllable variables.

3.3. Design of experiment

In order to estimate the emulator (1), the ADMS output must be produced under N realisations. In this work N is fixed to 100 which is a good compromise between computing time and estimation accuracy. Since the independent variables are continuous in their range of variation, a Latin Hypercube Sampling design (Park, 1994) is adopted to define the realisations in the space $\mathscr{X}$ (see Fig. 2). For each realisation $\boldsymbol{x}_i \in X$, $i = 1, \ldots, N$, the ADMS is used to compute the NO_2 concentration at the locations of the six monitoring station with hourly temporal resolution. The vector $\boldsymbol{y}_i$ is then obtained by computing annual averages from the hourly data.

3.4. Emulator estimation

The emulator estimation is performed using the D-STEM software in MATLAB. The vectors $\tilde{\boldsymbol{x}}_{m,i}$ which are used to build the matrix $\boldsymbol{X}_i$, are equal to $(1, x_{1,i}, x_{2,i}, x_{3,i}, x_{1,i}{\cdot}x_{2,i}, x_{1,i}{\cdot}x_{3,i}, x_{2,i}{\cdot}x_{3,i})$, namely the first order interactions are included. Of the 100 realisations generated by the Latin Hypercube Sampling design, two of them were discarded as the respective $\boldsymbol{y}_i$ vectors include anomalous values. The emulator is thus estimated using the remaining 98 realisations.

The estimation of Ψ is based on the EM algorithm, which is stable even when the number of parameters is high and typically converges in a small number of iterations (less than 100). For this example, the EM algorithm reaches convergence in 23 iterations and in around 13 min on a standard laptop. On the same laptop, the emulator is able to provide the output $\hat{\boldsymbol{y}}(\boldsymbol{x})$ with a rate of about 1400 realisations per seconds. D-STEM also provides the standard deviation of each estimated parameter, easing model selection and inference.

Model estimation results are reported in Table 2 (fixed effect coefficients), Table 3 (random effect coefficients) and Table 4 (range parameters). The coefficients $\hat{\beta}_0$ in Table 2 directly give the yearly average pollutant concentration under the baseline realisation at the monitoring station locations. The fixed effects of emission, wind speed and wind direction are significant, as well as the interaction between emission and wind speed. The remaining interactions are not significantly different from zero. The elements of $\hat{\boldsymbol{\alpha}}$ are close to zero, suggesting that the random component $\boldsymbol{w}$ explains a low percentage of the pollutant concentration variability across the realisations. It is found that $\boldsymbol{w}$ accounts for less than 0.1% of the variability at all spatial locations. The matrix $\boldsymbol{V}$ shows high cross-correlation between the spatial locations, while the estimated range parameter vector $\hat{\boldsymbol{\theta}}$ suggests that $\boldsymbol{w}$ is highly correlated across $\mathscr{X}$. However, the small values in $\boldsymbol{\alpha}$ implies that $\boldsymbol{V}$ and $\boldsymbol{\theta}$ are poorly identifiable, as confirmed by the high standard deviations on some elements of $\boldsymbol{V}$. Finally, the residual variance $\boldsymbol{\sigma}_{\varepsilon}^2$ is close to zero which implies R^2 statistics close to one for all the six spatial locations.

In order to keep the emulator as simple as possible, the random effect $\boldsymbol{w}$ is removed and the emulator is re-estimated considering only the fixed effect and the significant interaction between emission and wind speed. The estimated $\hat{\boldsymbol{\beta}}$ coefficients are not reported here as they differ by less than 0.03% from the values given in Table 2. Essentially, the estimated emulator is close to a deterministic linear model. This is a consequence of the fact that the ADMS output is here linear with respect to variations in the variables defining the realisations. From a modelling point of view, this implies that the emulator accurately provides the yearly average pollutant concentration at the six spatial locations, without the need to produce the ADMS output at hourly temporal resolution.

3.5. Cross validation

To check the performance of the emulation model, cross validation was performed by using a training (70%) and a validation (30%) set. The validation annual average concentration were removed at random for all six stations. The emulator was refitted with the training data set and the model was used to predict the annual average concentrations for the validation data set. The bias, the root mean squared error (RMSE) and the correlation were estimated for the predictions compared to the true values as described in Table 1 in (Mallet et al., 2018). The statistics are summarised in Table 5 below:

From Table 5, it is clear that the bias for all estimates is very low

Table 2
Emulator fixed effect coefficients with standard deviations in brackets.

Coef.	Anderson Dr	Errol Pl	King St	Market St 2	Union St	Wellington Rd
	S1	S2	S3	S4	S5	S6
$\hat{\beta}_0$	31.337	28.102	36.003	47.429	49.452	43.895
	$(1.3\cdot10^{-2})$	$(9.5\cdot10^{-3})$	$(1.9\cdot10^{-3})$	$(3.4\cdot10^{-2})$	$(3.3\cdot10^{-2})$	$(3.4\cdot10^{-2})$
$\hat{\beta}_{x1}$	0.078	0.046	0.120	0.217	0.235	0.190
	$(5.2\cdot10^{-4})$	$(3.7\cdot10^{-4})$	$(7.2\cdot10^{-4})$	$(1.3\cdot10^{-3})$	$(1.3\cdot10^{-3})$	$(1.3\cdot10^{-3})$
$\hat{\beta}_{x2}$	-0.066	-0.051	-0.114	-0.183	-0.206	-0.177
	$(1.2\cdot10^{-3})$	$(8.3\cdot10^{-4})$	$(1.7\cdot10^{-3})$	$(3.0\cdot10^{-3})$	$(2.9\cdot10^{-3})$	$(2.9\cdot10^{-3})$
$\hat{\beta}_{x3}$	0.024	0.013	0.050	0.065	0.022	0.076
	$(1.5\cdot10^{-3})$	$(1.1\cdot10^{-3})$	$(2.2\cdot10^{-3})$	$(4.0\cdot10^{-3})$	$(3.8\cdot10^{-3})$	$(3.8\cdot10^{-3})$
$\hat{\beta}_{x1x2}$	$-5.4\cdot10^{-4}$	$-3.9\cdot10^{-4}$	$-9.2\cdot10^{-4}$	$-1.3\cdot10^{-3}$	$-1.4\cdot10^{-3}$	$-1.3\cdot10^{-3}$
	$(5.6\cdot10^{-5})$	$(3.9\cdot10^{-5})$	$(8.0\cdot10^{-5})$	$(1.4\cdot10^{-4})$	$(1.3\cdot10^{-4})$	$(1.4\cdot10^{-4})$
$\hat{\beta}_{x1x3}$	$9.4\cdot10^{-5}$	$-2.8\cdot10^{-6}$	$2.9\cdot10^{-4}$	$2.1\cdot10^{-4}$	$1.2\cdot10^{-4}$	$3.7\cdot10^{-4}$
	$(6.2\cdot10^{-5})$	$(4.4\cdot10^{-5})$	$(9.0\cdot10^{-5})$	$(1.6\cdot10^{-4})$	$(1.5\cdot10^{-4})$	$(1.6\cdot10^{-4})$
$\hat{\beta}_{x2x3}$	$-1.0\cdot10^{-4}$	$2.9\cdot10^{-5}$	$-8.1\cdot10^{-5}$	$-1.1\cdot10^{-4}$	$1.6\cdot10^{-4}$	$-1.6\cdot10^{-4}$
	$(1.2\cdot10^{-4})$	$(8.2\cdot10^{-5})$	$(1.6\cdot10^{-4})$	$(2.9\cdot10^{-4})$	$(2.8\cdot10^{-4})$	$(2.8\cdot10^{-4})$

Table 3
Emulator random effect parameters with standard deviations in brackets. S1-S6 are the station code as reported in Table 2.

	$\hat{V}$						$\hat{\alpha}$	σ_ε^2
	S1	S2	S3	S4	S5	S6		
S1	1	0.97	0.95	0.95	0.83	0.97	0.071	$1.9\cdot10^{-4}$
		(7.13)	(1.57)	(0.03)	(0.03)	(0.05)	(0.006)	$(1.1\cdot10^{-4})$
S2		1	0.97	0.94	0.88	0.96	0.053	$9.2\cdot10^{-5}$
			(0.11)	(0.02)	(0.04)	(0.05)	(0.005)	$(5.7\cdot10^{-5})$
S3			1	0.95	0.93	0.97	0.112	$3.1\cdot10^{-4}$
				(0.09)	(0.03)	(0.03)	(0.010)	$(1.8\cdot10^{-4})$
S4				1	0.92	0.98	0.186	$6.6\cdot10^{-4}$
					(0.02)	(0.01)	(0.017)	$(4.1\cdot10^{-4})$
S5					1	0.89	0.243	$1.4\cdot10^{-2}$
						(0.05)	(0.024)	$(3.0\cdot10^{-3})$
S6						1	0.173	$5.9\cdot10^{-4}$
							(0.015)	$(3.6\cdot10^{-4})$

Table 4
Emulator spatial correlation function parameters.

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
66.27 (12.88)	36.63 (7.13)	9.42 (1.57)

Table 5
The cross validation bias (in $\mu g\ m^{-3}$), RMSE (in $\mu g\ m^{-3}$) and correlation for each station.

Station	Bias	RMSE	Correlation
Anderson Drive (S1)	-0.003	0.117	0.999
Errol Place (S2)	-0.002	0.092	0.998
King Street (S3)	0.007	0.204	0.999
Market Street 2 (S4)	-0.040	0.335	0.999
Union Street (S5)	-0.012	0.329	0.999
Wellington Road (S6)	-0.011	0.336	0.999

with the largest bias being -0.040 $\mu g\ m^{-3}$ for the Market Street 2 station. This indicates that the estimates from the model have negligible bias. The cross validation RMSE is found to be very small for all the stations which shows that the emulator is performing well in predicting annual average concentrations. The correlations for all stations are 0.99 indicating a very strong positive relationship between the actual values and the predicted ones. Overall, these statistics show that the emulator is performing with high accuracy.

3.6. *Inverse problem*

The yearly average pollutant concentration vector $\hat{\boldsymbol{y}}(\boldsymbol{x})$ is obtained over a discretised version $\mathscr{X}^d$ of $\mathscr{X}$. Each element of the discrete space is called a voxel. In this paper, a 3D voxel is used for the combinations of the three inputs (emission, temperature and wind direction). In this particular example, voxels of length 0.5 along all the dimensions of $\mathscr{X}$ are used. For each voxel in $\mathscr{X}^d$, $P(y_m^0(\boldsymbol{x}^d) > L)$ is computed considering the distribution (5), where $\boldsymbol{x}^d$ is the centre of the voxel. Since $\boldsymbol{w} \equiv 0$, the second term of the rhs of (4) is equal to zero and the region $\mathscr{X}_m^d$, which is defined similarly to (6), is a compact 3D solid embedded in $\mathscr{X}^d$. In particular, the region $\mathscr{X}_m^d$ is given by the set of voxels that satisfy $P(y_m^0(\boldsymbol{x}^d) > L) < \delta$ with $\delta = 0.05$.

For all spatial locations, Fig. 3 shows $P(y_m^0(\boldsymbol{x}^d) > L)$, $\boldsymbol{x}^d \in \mathscr{X}_m^d$, when $x_3 = 0°$ (no change in wind direction). The black circle in each plot represents the 2012 baseline. Under the baseline realisation, three of the six spatial locations have a probability equal to one to exceed the 40 $\mu g\ m^{-3}$ yearly average concentration, which is in agreement with the observed exceedances in Table 1. In each plot, the boundary of $\mathscr{X}_m^d$ is depicted by a red curve. Realisations $\boldsymbol{x}^d$ on the left of the curve ensure that $P(y_m^0(\boldsymbol{x}^d) > L)$ is lower than 0.05. Since the emulator's R^2 is close to one, the transition in $\mathscr{X}_m^d$ from probability zero to probability one is quite rapid. However, this only reflects the capabilities of the emulator in emulating the ADMS, and not the ADMS capabilities in predicting the true pollutant concentrations.

From the analysis of the plots in Fig. 3 it is possible to conclude that, assuming no changes in wind speed and direction, the goal is achieved at all the spatial locations if emissions are reduced by at least 42%, with Union St. the most critical location. The 95% bootstrap confidence interval is (−42.2%, −41.9%). The interval length is small since $\boldsymbol{w} \equiv 0$ and the uncertainty on the estimated emulator parameters is small. Since not very informative, bootstrap confidence intervals are not provided in the rest of this section.

Fig. 4 shows the probability map for Union St. and Wellington Rd. when the wind direction changes from $-15°$ to $15°$ with respect to the baseline. From Table 2, it is observed that the effect of wind direction on the yearly average pollutant concentration at Wellington Rd. is around 3.5 higher than the effect at Union St. This is reflected in the

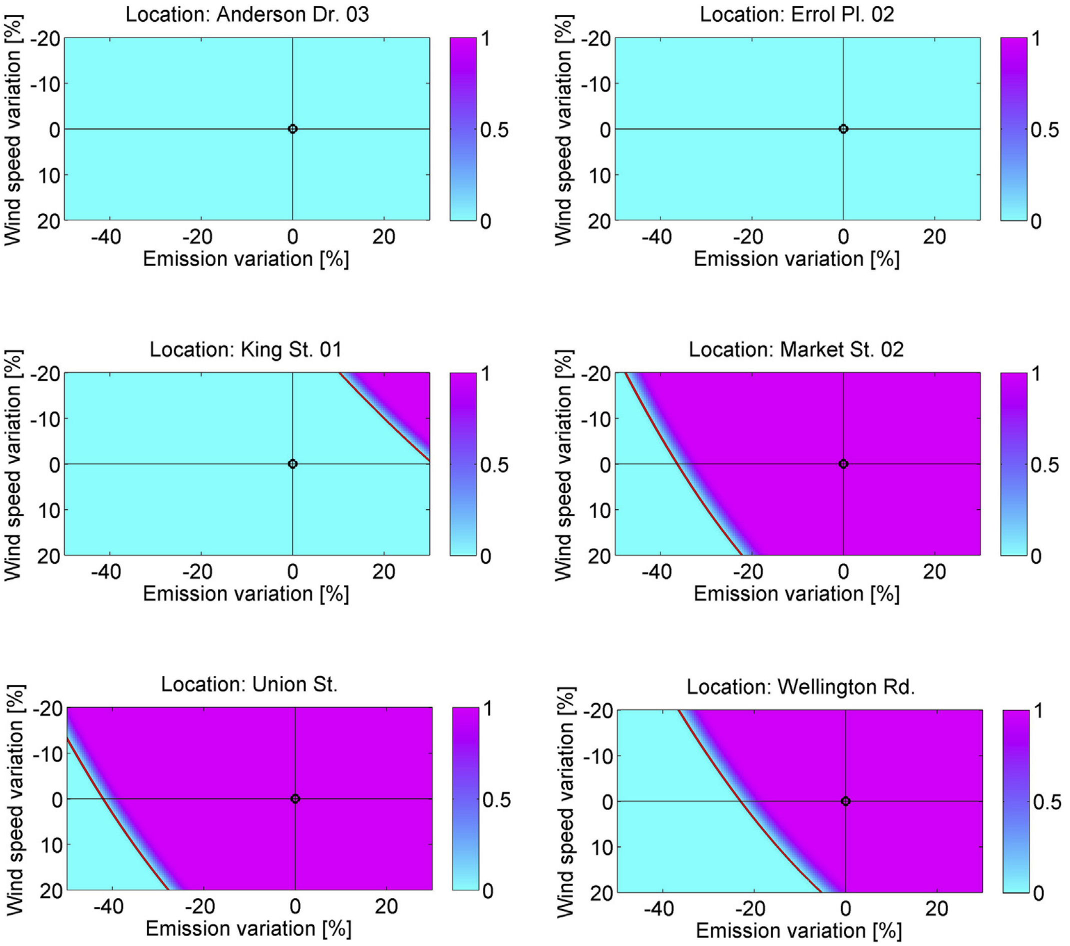

Fig. 3. Exceedance probabilities $P(y_m^0(\boldsymbol{x}^d) > 40\ \mu g\ m^{-3})$ at the six monitoring station locations with respect to variations in emission and wind speed assuming no variation in the wind direction. In each plot, the black circle depicts the baseline realisation. The red curve delimits the sub-region of $\mathscr{X}^d$ where $P(y_m^0(\boldsymbol{x}^d) > 40\mu g m^{-3}) < 0.05$. The (0,0) coordinate on each plot is the baseline point. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

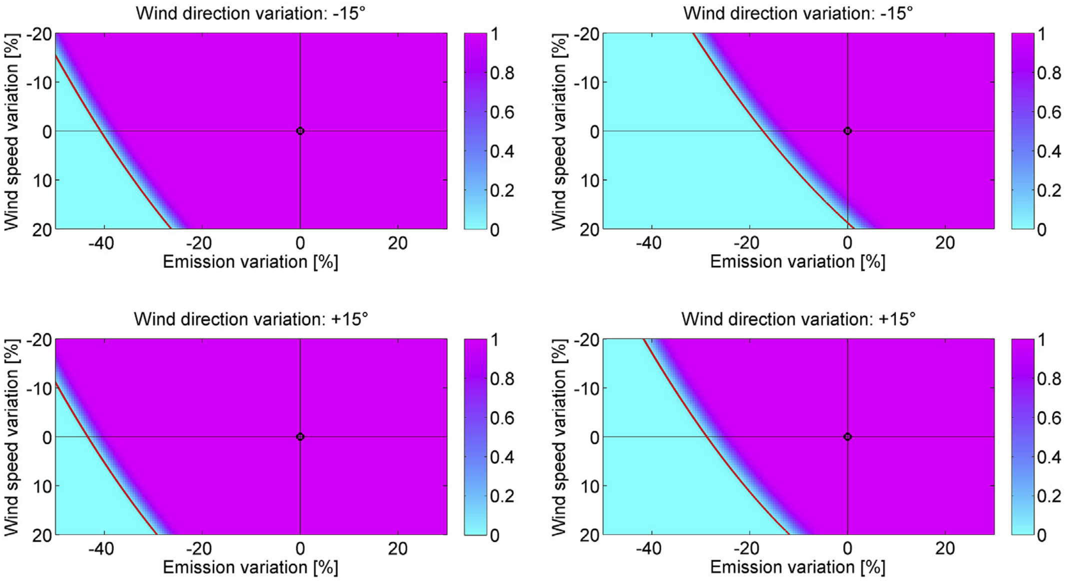

Fig. 4. Exceedance probabilities $P(y_m^0(\boldsymbol{x}^d) > 40\ \mu g\ m^{-3})$ at Union St. (left panels) and Wellington Rd. (right panels) assuming a wind direction variation of − 15° (top panels) and + 15° (bottom panels) with respect to the baseline. In each plot, the black circle depicts the baseline realisation. The red curve delimits the sub-region of $\mathscr{X}^d$ where $P(y_m^0(\boldsymbol{x}^d) > 40\ \mu g\ m^{-3}) < 0.05$. The (0,0) coordinate on each plot is the baseline point. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

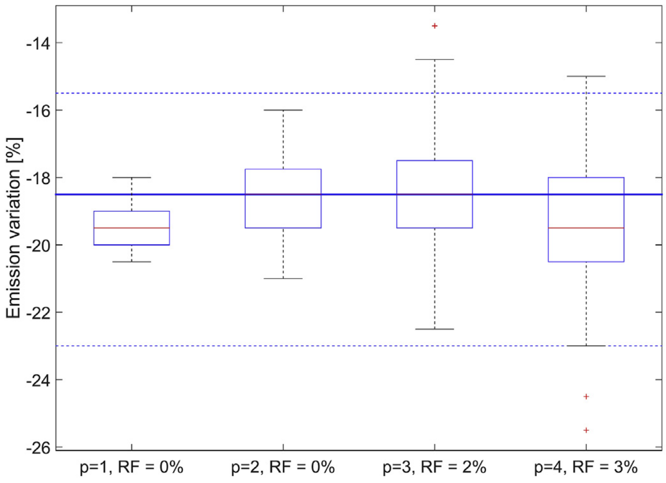

Fig. 5. Risk factor assessment when the number of missing values in $\boldsymbol{y}_i$ increases from $p = 1$ to $p = 4$. The horizontal lines depicts $\hat{x}_1^0$ and its 95% bootstrap confidence interval. Each box-plot represents the estimates $\hat{x}_{1,1}, ...,\hat{x}_{1,T}$ for the different p values.

decrease of emissions required to achieve the compliance goal. At Union St., the decrease in emission indeed ranges from 41% to 43.5% when the wind direction changes from − 15° to 15°. At Wellington Rd., the same change in wind direction implies a reduction of emissions that go between 17.5% and 28.8%.

The information from Figs. 3 and 4 suggests that wind speed and wind direction should both be considered as risk factors in failing to comply with the regulations when the emission reduction is implemented. If their natural variability from year to year is known, then this information can be used to choose an emission reduction which is robust against their variability. If the information is not available, then it is wise to identify an emission reduction which is not too close to the boundary of $\mathscr{X}_m^d$ to avoid non-compliance due to meteorological conditions. Wind direction is crucial for modelling the concentrations from monitors in complex locations near buildings.

4. Emulator estimation with missing values

In this section, a simulation study is carried out to understand the impact of missing values (by which we mean missing monitoring stations) on the emulator estimation and, in turn, on the action to be performed to reduce the pollutant concentration to a desired level. Although the case study presents only complete cases, there are many other circumstances which could results in missing values. Alternatively, allowing for missing values, would speed up the emulator. The method presented by Fricker et al. (2013) is extended allowing the vector $\boldsymbol{y}_i$ to include missing values. Indeed, when the computing time is high and/or M is large it may be convenient to produce $\boldsymbol{y}_i$ only partially, allowing some elements (at random) to be missing.

4.1. Simulation design

The case study of the previous section is used as a basis for the simulation. In particular, the same realisations and the same spatial distribution of the monitoring stations are considered. For each realisation i, the observation vector $\boldsymbol{y}_i$ is directly simulated from the emulator equation (1). To obtain realistic pollutant concentrations, the emulator parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}_\varepsilon^2$ are initialised using the values in Tables (2) and (3). On the other hand, in order to deal with a more general emulator with a significant random component $\boldsymbol{w}(\boldsymbol{x}_i)$, all the elements of $\boldsymbol{\alpha}$ are set to 2. This value implies that the random component accounts for around 12% of the variance of $\boldsymbol{y}_i$. Finally, the cross-correlation matrix $\boldsymbol{V}$ is initialised with all extra-diagonal elements equal to 0.8, while $\theta = (8,5,3)$. In this study, the simulation values are chosen arbitrary except for the correlation which was chosen to be a high value to emphasise on the benefits of high-cross correlation across stations.

Data simulated through the emulator are used to estimate the emulator parameters. However, missing values are forced on the observation vectors $\boldsymbol{y}_i$. This is done by removing all the outputs for a specific monitoring station. To better understand the impact of missing values, four cases are considered by changing the number of missing stations ($p = 1,2,3,4$) in each 6×1 vector $\boldsymbol{y}_i$. For each realisation i, the elements of $\boldsymbol{y}_i$, which are replaced with missing values (NaN values), are randomly chosen.

Since the estimation result is affected by the position of the missing values in $\boldsymbol{y}_i$, the emulator is estimated $T = 100$ times for each p, randomly changing the position of the missing values. Note that each emulator estimate gives rise to a set of regions $\mathscr{X}_{m,t}$, $m = 1, ...,M$, $t = 1, ...,T$. In order to simplify the discussion of the results, the focus is limited to the monitoring station S5 and thus to $\mathscr{X}_{5,t}$. In particular, for each $\mathscr{X}_{5,t}$, the probability $P(y_m^0(\boldsymbol{x}^d) > L)$ is assessed considering $L = 40\,\mu gm^{-3}$, $x_2 = 0\%$ (no change in wind speed) and $x_3 = 0°$ (no change in wind direction). Under these constraints, the emission variation, x_1, is optimised in such a way that the above probability is lower than 0.05. This gives rise to the set of estimates $\hat{x}_{1,1}, ...,\hat{x}_{1,T}$, where each estimate is related to a permutation of the missing values.

Now, suppose that the generic emission variation $\hat{x}_{1,t}$ is the action taken by decision makers in order to reduce the pollutant concentration below L. Since the emulator was estimated with missing values in the observation vector, there is a risk that the action is wrong, namely that it is too mild to attain the goal or that it is overly demanding. In the latter case, the goal is attained but at a higher cost.

4.2. Evaluation of the results

In order to assess this risk, actions $\hat{x}_{1,t}$ are compared with the same action evaluated in the non-missing values case, say $\hat{x}_1^0$. For each $p = 1, ...,4$, the risk factor RF is defined as the percentage of the actions $\hat{x}_{1,1}, ...,\hat{x}_{1,T}$ that fall outside the 95% bootstrap confidence interval of $\hat{x}_1^0$, which is computed as detailed in Section 2.2. In Fig. 5, the box-plots represent the estimates $\hat{x}_{1,1}, ...,\hat{x}_{1,T}$ with respect to $p = 1, ...,4$. The horizontal lines depict $\hat{x}_1^0$ and its 95% confidence interval. It can be seen that, though the box-plot increases in height when moving from $p = 1$ to $p = 4$, most of the estimates $\hat{x}_{1,t}$ are within the confidence interval of $\hat{x}_1^0$ and RF only reaches 3% when $p = 4$. This suggests that missing values do not compromise the emulator and, in turn, the estimate of the action to be taken to reduce the pollutant concentration at the desired level. Therefore, further additional computations savings can be made.

5. Conclusions

This paper provides an empirical proof of the capabilities of a multivariate emulator, which is suitable for emulating physical model outputs where the output is vector-valued. The emulator, therefore, allows multiple variables and, due to its implementation within the D-STEM software, it is able to deal with missing values in the output vector.

The data missingness can be exploited when the physical model is very computationally demanding and/or the number of variables is high, in which case the model output is produced for only a subset of the variables for each realisation. This approach works if the variables are cross-correlated and allows the number of observed variables in each realisation rather than the number of realisations itself to be reduced, keeping the computation feasible.

The emulator was applied to an air quality management case study related to the city of Aberdeen (UK). Considering a continuous set of realisations defined by emission, wind speed and wind direction, the emulator was used to directly produce the yearly average NO_2 concentration at the spatial locations where the city monitoring stations are installed. This allows us to identify in which realisations the yearly average concentration is expected, with a given probability, to be below the 40 $\mu g\ m^{-3}$ limit at all the stations.

Finally, based on the case study, a simulation study was conducted to understand the impact of missing values on the emulator estimation and the resulting risk of non-compliance. Using a bootstrap technique, in this case study, it was shown that missing values only mildly affect the emulator estimation and that the risk of failing to comply by reducing emissions is very low even when more than half of the observation vector includes missing values.

In this paper, the emulator was built as a quick computational tool to help SEPA visualise changes in concentrations under different emission variations as well as meteorological data (wind speed and wind direction). However, this has resulted in loss of information, for instance, temporal information was lost by using just the annual average concentrations rather than the hourly data. In future work, the emulator will be extended to handle hourly concentrations. Furthermore, geographical space information was not used to build the emulator. Instead, the emulator was built on the distances between the sets of inputs in the Latin Hypercube space. Therefore, the model could be further extended to include the geographical space in addition to the Latin Hypercube space. This will allow for predicting continuously in the geographical area of interest.

Acknowledgements

This work is developed under the PRIN2015 supported-project "Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)" funded by MIUR (Italian Ministry of Education, University and Scientific Research). Francesco Finazzi and Michela Cameletti have been supported by the PRIN EphaStat Project funding (Project No. 20154X8K23, https://sites.google.com/site/ephastat/) provided by the Italian Ministry for Education, University and Research. Yoana Napier acknowledges financial support under a doctoral training grant from NERC NE/N008367/1.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2018.11.025.

References

Bastos, L.S., OHagan, A., 2009. Diagnostics for Gaussian process emulators. Technometrics 51 (4), 425–438.

Bayarri, M.J., Berger, J.O., Calder, E.S., Dalbey, K., Lunagomez, S., Patra, A.K., Pitman, E.B., Spiller, E.T., Wolpert, R.L., 2009. Using statistical and computer models to quantify volcanic hazards. Technometrics 51 (4), 402–413.

Callaghan, D., 2014. Glasgow City Council Detailed Assessment. https://www.glasgow.gov.uk/CHttpHandler.ashx?id=32459&p=0.

Carrington, F., Taylor, M., 2017. UK Government Sued for Third Time over Deadly Pollution. The Guardian. https://www.theguardian.com/environment/2017/nov/07/uk-government-sued-for-third-time-over-deadly-air-pollution.

EC, 2008. Directive 2008/50/EC of the European parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe. Offic. J. Eur. Union L 152/1.

Fassò, A., Finazzi, F., 2011. Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Environmetrics 22 (6), 735–748.

Finazzi, F., Fassò, A., 12 2014. D-STEM: a software for the analysis and mapping of environmental space-time variables. J. Stat. Software 62 (6), 1–29.

Finazzi, F., Scott, E.M., Fassò, A., 2013. A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of scottish air quality data. J. Roy. Stat. Soc. C 62 (2), 287–308.

Fricker, T.E., Oakley, J.E., Urban, N.M., 2013. Multivariate Gaussian process emulators with nonseparable covariance structures. Technometrics 55 (1), 47–56.

Hood, C., et al., 2018. Air quality simulations for London using a coupled regional-to-local modelling system. Atmos. Chem. Phys. 18, 11221–11245. https://doi.org/10.5194/acp-18-11221-2018.

Local Air Quality Management (LAQM), 2008. Diffusion Tubes for Ambient NO_2 Monitoring: Practical Guidance for Laboratories and Users.

Mallet, V., et al., 2018. Meta-modelling of ADMS-Urban by dimension reduction and emulation. Atmos. Environ. 184, 37–46.

Park, J.-S., 1994. Optimal Latin-hypercube designs for computer experiments. J. Stat. Plann. Inference 39 (1), 95–111.

Pasquill, F., 1979. Atmospheric dispersion modeling. J. Air Pollut. Contr. Assoc. 29 (2), 117–119.

Royal College of Physicians, 2016. Every Breath We Take: the Lifelong Impact of Air Pollution. https://www.rcplondon.ac.uk/file/2916/download?token=RzylFzis.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Stat. Sci. 409–423.

SEPA and Natural Scotland, 2017. Aberdeen Air Quality Modelling Pilot Project Technical Report. (Not published).

Stocker, J., Hood, C., Carruthers, D., McHugh, C., 2012. Adms-urban: developments in modelling dispersion from the city scale to the local scale. Int. J. Environ. Pollut. 50 (1–4), 308–316.

Stratton, S., 2014. Air Quality Further Assessment for Musselburgh. East Lothian Council. https://www.eastlothian.gov.uk/download/downloads/id/23471/air_quality_further_assessment_2014.pdf.

The Scottish Government, 2016. Air Quality in Scotland. http://www.gov.scot/Topics/Environment/waste-and-pollution/Pollution-1/16215.

Turner, D.B., 1994. Workbook of Atmospheric Dispersion Estimates: an Introduction to Dispersion Modeling. CRC press.