Stochastic Environmental Research and Risk Assessment (2023) 37:275–290
https://doi.org/10.1007/s00477-022-02288-x

ORIGINAL PAPER

Surrogate modelling for the forecast of Seveso-type atmospheric pollutant dispersion

Juš Kocijan[1,2] · Nadja Hvala[1] · Matija Perne[1,4] · Primož Mlakar[3] · Boštjan Grašič[3] · Marija Zlata Božnar[3]

Accepted: 18 July 2022 / Published online: 30 August 2022

Abstract
This paper presents a framework for the development of a computationally-efficient surrogate model for air pollution dispersion. Numerical simulation of air pollution dispersion is of fundamental importance for the mitigation of pollution in Seveso-type accidents, and, in extreme cases, for the design of evacuation scenarios for which long-range forecasting is necessary. Due to the high computational load, sophisticated simulation programs are not always useful for prompt computational studies and experimentation in real time. Surrogate models are data-driven models that mimic the behaviour of more accurate and more complex models in limited conditions. These models are computationally fast and enable efficient computer experimentation with them. We propose two methods. The first method develops a grid of independent dynamic models of the air pollution dispersion. The second method develops a reduced grid with interpolation of outputs. Both are demonstrated in an example of a realistic, controlled experiment with limited complexity based on an approximately 7 km radius around the thermal power plant in Šoštanj, Slovenia. The results show acceptable matching of behaviour between the surrogate and original model and noticeable improvement in the computational load. This makes the obtained surrogate models appropriate for further experimentation and confirms the feasibility of the proposed method.

Keywords Air pollution · Seveso-type accident · Pollution dispersion · Surrogate modelling · Dynamic systems · Data-driven modelling

1 Introduction

A framework for the development of a computationally-efficient surrogate model for air pollution dispersion is studied in this paper.

Surrogate modelling (Keane et al. 2008; Koziel and Leifsson 2013; Jiang et al. 2020) is an engineering method that is used when we cannot compute the response of the model of the system of interest easily enough. It is a method that helps to alleviate computationally demanding computations necessary for tasks like design optimisation, design-space exploration, and various computationally intensive analyses. Speeding up computations can be done by constructing an approximation model known as a surrogate model, meta model, emulation model, or emulator. The surrogate model is usually developed by selecting a suitable data-driven black-box model. Such a model is obtained from the input-output response of the original mathematical model. It is constructed based on appropriately selected input values that excite the original mathematical model in the region of interest as illustrated in Fig. 1.

Surrogate models have been used in different scientific fields for different tasks, see (Alizadeh et al. 2020) for a review. The methodology is used also in atmospheric sciences for dispersion modelling, e.g. (Carnevale et al. 2012; Bowman and Woods 2016; Gunawardena et al. 2021; Le et al. 2019; Francom et al. 2019; Girard et al. 2020). The relevance of dispersion modelling is non-disputable and

✉ Juš Kocijan
jus.kocijan@ijs.si

[1] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

[2] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

[3] MEIS d.o.o., Mali Vrh pri Šmarju 78, 1293 Šmarje-Sap, Slovenia

[4] QUANTECTUM, prognoza potresov, d.o.o., Dunajska cesta 156, 1000 Ljubljana, Slovenia

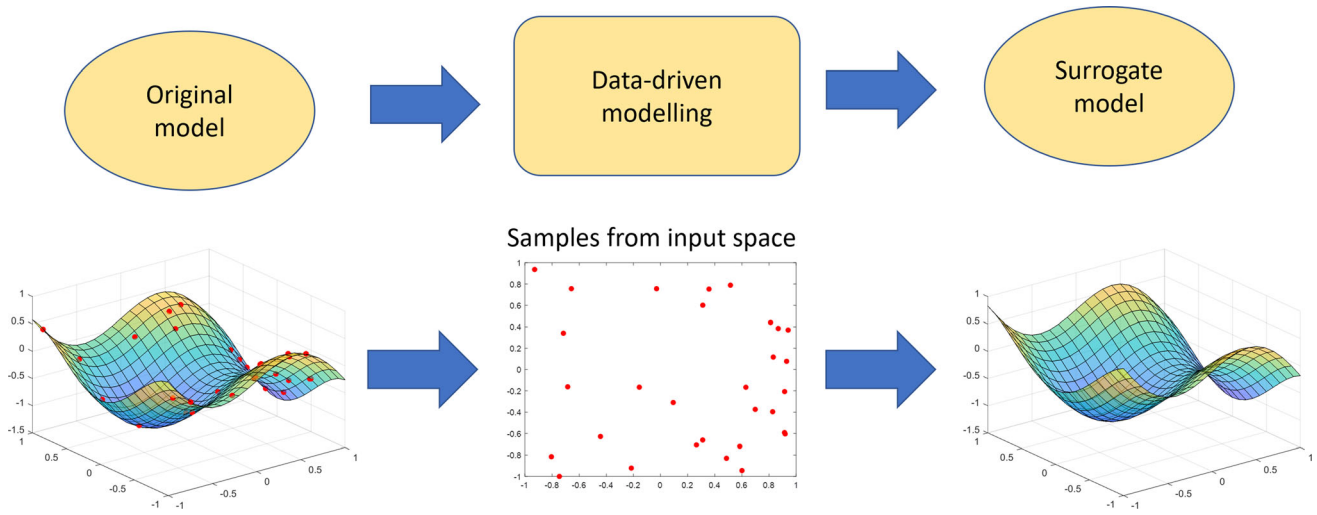

Fig. 1 The principle of surrogate modelling

addressed in numerous studies dealing with different environments, situations and methods, e.g. (Ravina et al. 2021).

The purposes of developing surrogate models range from uncertainty quantification (Francom et al. 2019) to spatial-deposition prediction (Gunawardena et al. 2021). The core of all applications is to replace a computationally demanding model with a faster surrogate one. Details of some recent investigations are as follows.

To investigate the relative impact of a collection of uncertain model inputs and their interactions on the outputs of the atmospheric dispersion for the Fukushima nuclear catastrophe, Girard et al. (2016) applied Sobol's global sensitivity analysis using Gaussian-process emulation. At 64 measurement locations, the emulators' ability to forecast time- and space-aggregated gamma dose rates, as well as time-integrated gamma radiation rates, were assessed.

Le et al. (2019) describe the use of surrogate models for the prediction of integrated statistical measures, e.g. Root Mean Square Error, based on inputs describing meteorological forecasts and a source term. The authors have shown an application of surrogate modelling for replacing an Eulerian model for the Fukushima event.

Pal et al. (2019) developed a surrogate model of physics-based radiation model using deep neural networks to reduce the computational cost.

For 2D dose prediction in a Brazilian nuclear power plant, Desterro et al. (2020) utilised a Deep Rectifier Neural Network. The method was developed for immediate prediction up-to 1 h after the accident and considered five model inputs (wind velocity, wind direction, position x, position y and time after the accident started). The data samples for the investigation were generated by the simulation model and, consequently, the obtained model can be considered as a surrogate.

Gunawardena et al. (2021) proposed a data-driven surrogate model to predict the spatial deposition of radioactive materials from a nuclear power plant for a single radiological release over a wide area for a particular period of 48 h. A grid of linear regression and logistic models is used for the surrogate model with categorical variables from an NWP system as inputs.

Carnevale et al. (2012) used a grid of neural networks for the static mapping between precursor emissions in each cell and its neighbouring cells and PM10 pollutant in a domain of interest.

Bowman and Woods (2016) proposed a surrogate model of atmospheric dispersion for a Gaussian puff model. The surrogate model is composed of basis functions whose coefficients are modelled with a Gaussian-process model. They compare different modelling methods with the uncertainty quantification of coefficients. The surrogate model is a static model and does not take dynamics into account. The static model means that the surrogate model output at every time instant does not encounter the past or derivations of input and/or output variables.

Uncertainty quantification of spline coefficients is also the purpose of (Francom et al. 2019), which efficiently uses Bayesian adaptive splines for surrogate modelling to which they use categorical inputs, i.e. discrete values from a finite set of choices from a numerical weather prediction (NWP) system. The purpose of the developed surrogate model is the characterisation of atmospheric release. In particular, the authors model the weights of empirical orthogonal functions in space and time for plume emulation using the adaptive splines.

Mallet et al. (2018) built a meta-model that replicates the key aspects of the air quality model ADMS-Urban to simulate NO_2 and PM10 emissions on an urban scale with street resolution and continuous emissions from emission

sources. The original model lacks a temporal dimension in inputs and outputs and is static with low-dimensional inputs and high-dimensional outputs. The principal component analysis is used to condense the model's outputs before multiple linear regression and Kriging interpolate them.

For anticipating atmospheric dispersion of methane (CH_4) in 2D over complicated terrain, Lauret et al. (2016) developed cellular automata paired with an artificial-neural-network model. A dynamic model based on wind field data processed with Computational-Fluid-Dynamics simulation is utilised to supply input data for the artificial neural network.

A surrogate model for the Computational-Fluid-Dynamics wind model is developed by Moonen and Allegrini (2015). The surrogate model was a Gaussian-process model applied in a case study representing an urban area.

A surrogate model of a Computational-Fluid-Dynamics model of pollution dispersion (Mendil et al. 2021) and recently (Mendil et al. 2022) is a deep-neural-network model that mimics the accidental release of a radioactive pollutant from multiple sources for up to 2 h ahead in an urban area.

There are other uses of surrogate modelling that prove the utility of the methodology. Each of the described cases is a bit different in the model's purpose or different in the methods used, but with a common goal of reducing computational burden with the use of some approximation. Nevertheless, the listed surrogate models only utilise the information about the present and not about the past. In this sense, they are not dynamic models. One possible reason is that using information from previous time instants considerably increases the dimension of input space.

1.1 Problem statement

The problem at stake, different from those published, deals with modelling a dispersion of a pollutant with a single source of pollution over complex terrain. In Europe, major accidents involving dangerous chemicals are prevented and controlled through the Seveso Directive (European Commission 2020). Seveso-type industrial facilities and nuclear power plants have the potential for accidents with serious consequences, especially when the release occurs over complex terrain. When such an accident occurs, the consequences for the environment and the human population must be quickly forecasted. Air pollution dispersion in a complex environment is commonly modelled and simulated with the Lagrangian particle dispersion model (Girard et al. 2020), where the cost for its forecast accuracy is a heavy computational burden. The dispersion model is a dynamic model, which adds to its complexity.

The problem we tackle is to develop a surrogate model that will provide reliable forecasts of air pollution dispersion relatively quickly so that authorities will have enough time and information to act upon it. The surrogate model also should take into account the dynamics of the air pollution dispersion. The primary goal is to improve the computational speed of prediction to use the model for potential experimentation and long-range forecasting, i.e. forecasting for an extensive period of time, as long as information on weather variables or average weather variables is available. If the surrogate model is available for potentially important Seveso objects in advance, then in the event of a disaster, emergency services could use meteorological forecasts and calculate the dispersion forecasts quickly. The investigation is intended for the ground layer because that is where most people are exposed. The investigation can be done also for higher layers with the same methods and corresponding data. The ground layer is usually also the most complex layer due to the influence of terrain and land use.

1.2 Contribution

We propose two methods for modelling an input-output dynamic surrogate model of continuous point-source originating pollution over complex terrain based on meteorological variables obtained from an NWP system or other sources of weather information.

The contribution of the investigation is as follows:

- A method for the modelling of a surrogate model of air pollution dispersion based on meteorological variables as inputs and 2D representation of relative pollution concentration at the output as a grid of independent dynamic models (GIM) for each output cell.
- A method for the modelling of a surrogate model of air pollution dispersion based on meteorological variables as inputs and 2D representation of relative pollution concentration at the output as a reduced grid of dynamic models with interpolation of outputs (RGI).
- A fast and applicable case-study demonstration of the listed methods on a simulated Seveso-type point release of a pollutant over complex terrain.

In addition to developing the surrogate model, the contributed methods also can be used for modelling a data-driven dispersion model if the necessary training data is available. Nevertheless, this is not the emphasis of the paper.

The emphasis of the paper is on developing a surrogate model that reduces the computational burden of model prediction and forecasting with acceptable accuracy. Note that this paper is not about the particular data-driven machine learning method nor is its focus exclusively on the

point-accuracy of dispersion predictions. Different methods can be used in the proposed framework and the speed of computation is the focus of the study. The proposed framework can be used effectively for developing surrogate models for computationally demanding experimentations.

The structure of the paper is as follows. The following section describes the air pollution Lagrangian particle dispersion model at the selected location with complex terrain. Section 3 describes decision-tree models and the Gaussian-process-grid model that are used for solving the fast-dispersion prediction problem. Results are discussed in Sect. 4, and conclusions are gathered in Sect. 5.

2 Mathematical model and numerical simulation

The case study for demonstrating the development of a surrogate model for pollution dispersion is the Šoštanj thermal power plant. The pollution dispersion of this power plant frequently attracts attention and was also used for early modelling with artificial neural networks (Božnar et al. 1993).

A constant and continuous sulphur-dioxide (SO_2) pollution source emission of unit value was presumed, and which can be, when necessary, proportionally sized for real-life situations. The location of the Šoštanj power plant is at the edge of the Velenje Basin in Slovenia (Fig. 2). It is surrounded by the Alps to the north and north-west. The basin consists of narrow valleys with rivers flowing along them and, as such, represents highly complex terrain. Winds are stronger on elevated levels and weaker in the basin. A temperature inversion in winter and other circumstances additionally complicate the situation.

Pollution dispersion in such a complex terrain was successfully modelled by an air pollution Lagrangian particle dispersion model (Mlakar et al. 2015), which represents a suitable method to deal with the complexity of the terrain. The Lagrangian particle dispersion model was combined with a corresponding meteorological preprocessor able to reconstruct a three-dimensional diagnostic nondivergent wind field. In particular, the SPRAY Lagrangian particle dispersion model (Castelli et al. 2018), the MINERVE diagnostic mass consistent wind field model (Finardi et al. 1998) and the SURFPRO meteorological preprocessor (Finardi et al. 1997) were used for the dispersion modelling (Mlakar et al. 2015). Inputs in the Lagrangian particle dispersion model were weather variables that can be collected at different weather stations or weather-forecast variables, the digital model of the terrain heights, and land cover data of the region. Realistic information regarding land use was used in the Lagrangian particle dispersion model. The weather-forecast variables are valuable, especially in the case of major accidents

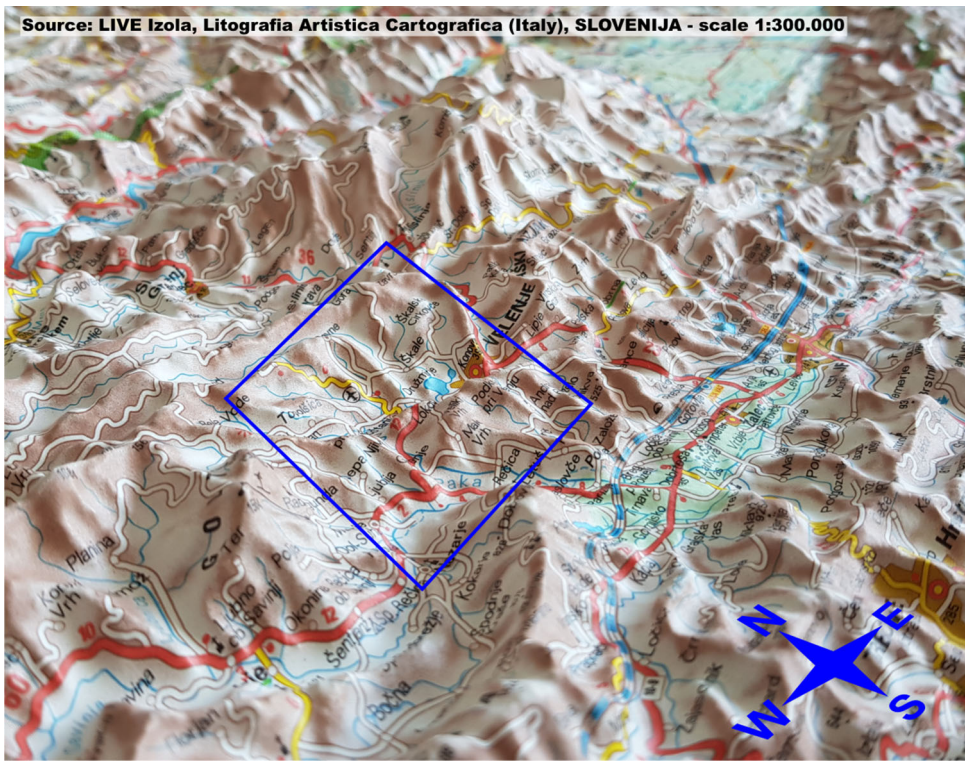

Fig. 2 Relief view of the position of Šoštanj power plant (46°22′26.56″ N, 15°3′5.15″ E)

involving dangerous chemicals, where we would like to forecast the direction and values of pollution dispersion.

For our demonstration, we did not use all the weather variables as described in (Mlakar et al. 2015), but have made a study for the simplified weather situation as follows. The temperature profile, wind velocity and direction were provided only at two altitudes, 10 m and 500 m, at the location of the thermal power plant in Šoštanj. The weather was presumed sunny with a clear sky. The output variable was the relative concentration of SO_2, but any other chemical of interest could be used in the study. The relative concentration (s/m^3) is a ratio between absolute pollutant concentration in $\mu g/m^3$ and the rate of emission (kg/s) (Mlakar et al. 2019). The relative concentration enables rescaling the results to any other form of pollutant emission. The region of interest was 15×15 km described with 100×100 square cells of 150×150 m each. Consequently, we dealt with the situation in 10,000 cells.

What we were striving for was the computational aspect of the described model. Nevertheless, the matching between the Lagrangian particle air pollution dispersion model output and the real situation in the field played a noticeable role in our study, but we were aware that it can be improved a bit further with additional training data and modelling effort. The software was run on an i9 desktop computer with MS Windows operating system. The model at the designated computer calculated every half-hour response at approximately a few tens of seconds. While this was an expected and acceptable computational performance, it was too slow to serve for numerical experimentation, which would be necessary for the real-time forecasting of accidents and especially not for long-range forecast studies because the time of calculation increases linearly with the length of the forecast horizon.

We took the developed Lagrangian particle dispersion model combined with the corresponding meteorological preprocessor as a benchmark and tried to improve the computational performance at an acceptable loss of accuracy with a surrogate model.

3 Methods

3.1 Surrogate modelling

3.1.1 The modelling procedure

The brief procedure for developing a surrogate model in our case is as follows (Fig. 3).

1. The development of a Lagrangian particle dispersion model with the accuracy suitable for the model's purpose.

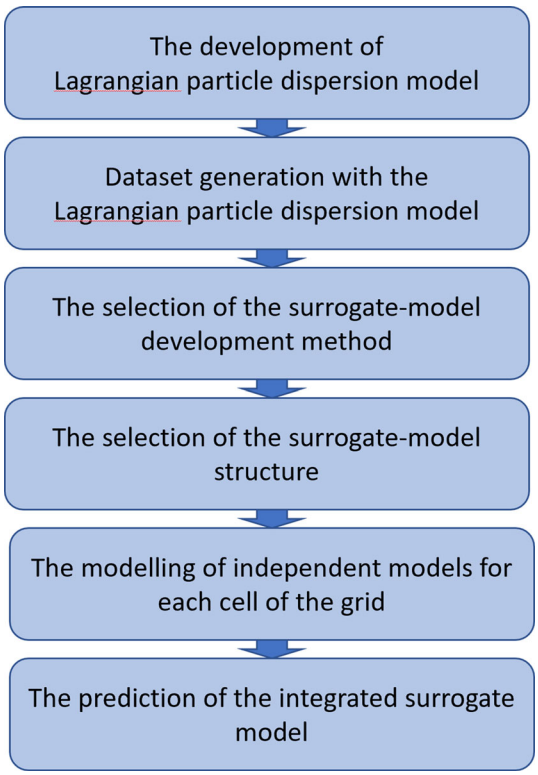

Fig. 3 Flow chart for the development of a surrogate model

2. Dataset generation with the Lagrangian particle dispersion model for surrogate modelling.
3. The selection of a model-development method for the surrogate model.
4. The selection of a surrogate-model structure (regressors, regression method, etc.).
5. Data-driven modelling of a large number of independent models, each for a cell of interest, or modelling of a smaller number of independent models and interpolation of their responses to the field of cells.
6. The prediction of the obtained surrogate model with data not used for modelling.

3.1.2 Constraints and assumptions

One of the most important steps in surrogate modelling is the generation of input samples that are intelligently distributed within the entire input space. This is usually done with optimal experimental design or with active learning (Breiman et al. 2017). In the case of atmospheric phenomena, the input variables utilised for modelling are commonly weather variables. This was also in our case. Values at the inputs were not obtained with a designed experiment and its implementation, but we used weather forecasts from a numerical weather simulator. The reason for this choice is that realistic combinations of values of

input variables cannot fill all the subspaces in the input-variables space because all possible combinations are not natural. Therefore, lots of combinations never occur. Consequently, we use data from available weather sources and not optimal experimental design or active learning. This particular selection of input data means that the number of data is not optimal and a large amount of data is necessary to encompass the relevant information.

The dispersion provided by the Lagrangian particle dispersion module has a resolution of 10,000 cells, with each cell measuring 150 × 150 m. This means that we deal with a system of 10,000 outputs, which introduces an identification problem with the excessive number of outputs. Such a problem can be tackled with dimension-reduction methods, e.g. (Girard et al. 2020). We decided to approach the problem differently and offer an alternative solution. The idea is to divide the model into a large number of submodels, as it was utilised in (Gunawardena et al. 2021; Carnevale et al. 2012), under the assumption that outputs do not influence each other. Nevertheless, the results described in the continuation showed that this working assumption provides some applicable results. Further research on alternative models and some output reduction methods applicable to this problem is envisaged for the future.

3.1.3 Performance metrics

Modelling performance was evaluated with two cost functions. The first is selected to evaluate time-dependent predictions of every submodel in the entire system in comparison with the original system. This evaluation is done with the standardised mean-squared error—SMSE (Le et al. 2019; Rasmussen and Williams 2006):

$$\mathrm{SMSE} = \frac{1}{N}\frac{\|\mathbf{y} - E(\hat{\mathbf{y}})\|^2}{\sigma_{\mathbf{y}}^2}, \tag{1}$$

where

$\mathbf{y}$–the vector of observations,
$E(\hat{\mathbf{y}})$–the mean value of estimations $\hat{\mathbf{y}}$,
$\sigma_{\mathbf{y}}^2$–the variance of observations,
N–the number of observations.

SMSE is a frequently used standardised measure for the accuracy of predictions' mean values with values between 0 and 1, where the value 0 is the result of a perfect model.

Pearson correlation coefficient R and the coefficient of determination R^2 are also given for comparison. Pearson correlation coefficient is defined as

$$\mathrm{R} = \frac{\mathrm{cov}(\mathbf{y}, E(\hat{\mathbf{y}}))}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}, \tag{2}$$

where cov is covariance, and $\sigma_{\hat{\mathbf{y}}}$ is the standard deviation of the mean values of estimations. The value is between − 1 and 1, and the more positive value is better.

The coefficient of determination is defined as

$$\mathrm{R}^2 = 1 - \frac{\|\mathbf{y} - E(\hat{\mathbf{y}})\|^2}{N\sigma_{\mathbf{y}}^2} = 1 - \mathrm{SMSE}. \tag{3}$$

The value is between 0 and 1, and a bigger value is better.

The presentation of pollution dispersion at ground level in our case is a two-dimensional field—an image of the dispersion. Consequently, the second selected cost function, the statistical coefficient of the space analysis, is the figure of merit in space—FMS (Mosca et al. 1998) also known as the Jaccard similarity coefficient

$$\mathrm{FMS} = \frac{A_1 \bigcap A_2}{A_1 \bigcup A_2}, \tag{4}$$

where A_1 and A_2 represent the measured and predicted areas respectively. The FMS is calculated at each time instant, with a fixed-threshold concentration level that distinguishes two categories of concentration values. Therefore it does not validate concentration levels, but the coverage of pollution. Values of the FMS close to 1 correspond to good model performance. Low FMS does not necessarily correspond to a bad model performance due to the shift of pollution plumes. Therefore, the FMS value should be evaluated together with a graphical representation of the measured A_1 and modelled areas A_2.

Ensembles of decision trees and a Gaussian-process grid were used for the development of surrogate models for the selected case study.

3.2 Decision trees

The literature provides many different algorithms for learning decision trees, which can be classification or regression trees (Breiman et al. 2017). While classification trees have a categorical output that indicates belonging to a finite set of outputs, regression trees provide numeric responses. A binary model tree, whether it is of a classification or regression type, consists of a split node with a threshold test of a particular variable $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a set of regressors. Given identification data $D = (\mathbf{x}, \mathbf{y}) \mid x \in \mathbb{R}^p, y \in \mathbb{R}^r$, where $\mathbf{x}$ is the vector of regressors (inputs) and y is the output, a model tree partitions the input space $\mathbb{R}^r$ into several partitions, called leaves. The split node creates a binary partition of the input space and has left and right offspring nodes. Such tree-like structures are frequently used for surrogate modelling (Alizadeh et al. 2020). They are robust, have the internal regressors' selection mechanism, are computationally efficient, are interpretable to a certain extent, etc. The

disadvantage is that piece-wise continuous estimates, as in the case of regression trees, may create certain inaccuracy, particularly for small trees.

The accuracy of regression trees can be improved with the use of tree ensembles (Mendes-Moreira et al. 2012; Aleksovski et al. 2016). Tree ensembles are created from several uncorrelated regression-tree models. A combination of the imperfect predictions obtained from each model tree should improve the prediction accuracy over a single tree and thus provide a more accurate model. Different principles exist to create ensembles. One of them is the bagging principle (Breiman 1996; Breiman et al. 2017). In bagging, bootstrap replicates are created, i.e. random samples with replacement of the training dataset D that have an equal number of data points as the training set. Each of the replicates D_i is used to build one model tree. The learning procedure starts by creating n bootstrap replicates of the training data D. Using each of the n data samples, a collection of model trees is built: $m_1, m_2, \ldots, m_n$. Denoting the predictions of the n single-output model trees of the ensemble with $\hat{f}_i(x)$, the overall prediction from the model tree ensemble is the average of the base model predictions for one output variable:

$$\hat{y}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_i(\mathbf{x}). \tag{5}$$

The rationale for selecting ensembles of bagged regression trees for modelling surrogate models was twofold: (1) learning of regression trees is faster than for other tested models and (2) ensembles improved the accuracy of predictions. Other selections might be viable as well.

3.3 Gaussian-process grid

Gaussian process modelling, also known as kriging, is a further method used in surrogate modelling (Alizadeh et al. 2020). Gaussian process models (Rasmussen and Williams 2006; Kocijan 2016) describe the input-output mapping of data $f(\mathbf{x})$ from regression vector $\mathbf{x}$ with a Gaussian process (GP). GP is a stochastic process containing random variables $f(\mathbf{x}_i)$ with a normal probability distribution,

$$p(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N) \mid \mathbf{x}_1, \ldots, \mathbf{x}_N)) = \mathcal{N}(\mathbf{m}, \mathbf{K}). \tag{6}$$

The vectors $\mathbf{x}_i$ are regressor vectors, f is the GP, $\mathbf{m}$ is the mean vector and $\mathbf{K}$ is the covariance matrix of the Gaussian distribution $\mathcal{N}$. In GP modelling, we describe the GP with a mean function and a covariance function,

$$\mathbf{m}_i = m(\mathbf{x}_i), \quad \mathbf{K}_{ij} = C(\mathbf{x}_i, \mathbf{x}_j), \tag{7}$$

where $m(\mathbf{x}_i)$ is the mean function and $C(\mathbf{x}_i, \mathbf{x}_j)$ is the covariance function. GP models are flexible function approximators, which can be used to represent complex structures with covariance kernels (Kocijan 2016).

When dealing with the grid structure of output data, grid-based covariance approximations can be utilised (Wilson and Nickisch 2015). This is the way to decrease computational costs in training and prediction. In the case of multidimensional inputs on a Cartesian grid, $\mathbf{x} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$, and a product kernel across the grid dimensions, $C(\mathbf{x}_i, \mathbf{x}_j) = \prod_{k=1}^{p} C(\mathbf{x}_{ik}, \mathbf{x}_{jk})$, then the $n \times n$ covariance matrix $\mathbf{K}$ can be expressed as a Kronecker product $\mathbf{K} = \mathbf{K}_1 \otimes \cdots \otimes \mathbf{K}_p$. The product of grid sizes in all the dimensions $N = \prod_{i=1}^{p} n_p$ is a product of the number of points n_p per grid dimension. The method takes advantage of the computational properties of a grid-based covariance matrix via the structured kernel interpolation. The reader is referred to (Wilson and Nickisch 2015) for more details.

4 Results and discussion

4.1 Grid of independent models—GIM

4.1.1 Data

Data for training, validation and testing were obtained with the simulation of air pollution Lagrangian particle dispersion model for the Šoštanj thermal power plant. The data sequence contained three years of data (July 2018–July 2021) with a 30-min sampling time. The data sequence contained SO_2 concentration as a system's output signal and 7 meteorological variables as a system's input signals, which were temperature, wind velocity and wind direction at heights 10 m and 500 m, the global solar IR radiation and the ground projection of the dispersion forecast for each time instant.

This set of more than 52,500 data pieces was divided into training, validation and test sets. Since we need as much data as possible for training to get as much information as possible, we ended dividing data into 51 subsets where one was immediately set aside as the test dataset (June 2021–July 2021), while the rest of the data was used for training and validation (July 2018–May 2021). The test data incorporated samples at the end of the completed dataset.

The data that was not used for the test was normalised, namely centred to have a mean of 0 and scaled to have a standard deviation of 1. The obtained mean values and variances were then also used for the normalisation of the test data.

4.1.2 Structure

Dispersion of air pollution is a dynamic system. It nonlinearly depends on meteorological and other environmental variables and not just on their present time values, but also the variable's history. Therefore the surrogate models have to be dynamic models as well. A static model would be just an approximation of a dynamic one in one time instant. The purpose of the entire system of models is forecasting for extended time horizons, therefore long-range forecasting and the entire system of models has a large number of outputs. The scheme of GIM is shown in Fig. 4.

Dynamics is incorporated in a model with regressors, which are the delayed values of outputs and inputs in the form of infinite-impulse-response or finite-impulse-response models (Nelles 2002).

We selected a finite-impulse-response (FIR) *model structure* for our submodels because an excessive number of outputs would make their feedback very inconvenient. These models have delayed values of inputs only (no delayed outputs) for regressors and the number of delays corresponds to the time in which impulse-excited response fades out. Since air pollution dispersion is a nonlinear process, we used a nonlinear-FIR (NFIR) model, in particular an ensemble model of bagged regression trees. Regression trees were selected due to the speed of training and their accuracy was improved using ensembles. Other data-driven modelling methods were tested, but all have performed worse than the selected one regarding the speed of training.

The next step in the structure selection is the selection of *regressors*. All seven available input signals were used as inputs, while outputs were values of SO_2 at each cell. The number of input delays was selected with 4-fold cross-validation on the dataset of about 21,000 data points. The reason for not taking a larger amount of data is that we wanted to keep a reasonable computational time. The training and validation sets were large enough to make the obtained results instrumental also for larger datasets even though the values of cost functions for cross-validation are worse than in the case when a larger dataset is used.

The results of 4-fold cross-validation for delay-selection are given in "Appendix A". Note that we have validated the same amount of delays on all inputs, without eliminating particular uninformative regressors due to simplicity. Therefore, if the maximal delay is, e.g. 3, this means that we have 21 regressors, 7 with delay 1, 7 with delay 2, and 7 with delay 3 for the prediction of the value at the moment of delay 0, i.e. at present moment. The cross-validation study was pursued for the delay interval between 2 and 6, where the minimum was located.

The best SMSE and FMS results were obtained with the delay of 4 samples, which corresponds to delays up to 2 h. This can be interpreted that a transient of 2 h encompasses the most information about the pollution dispersion.

The number of *observations per tree leaf*(or partition) was selected next. Using the already selected model parameters, a 4-fold cross-validation study was done to determine the optimal number of observations. The obtained results are given in "Appendix A". Averages of the SMSE measure are shown in Fig. 5. SMSE was used for the evaluation because we are evaluating the prediction ability of each independent submodel in the GIM model.

It is clear from Fig. 5 that the prediction-quality index is the lowest when 10 observations per tree leaf are used.

The next selection concerns the structure of ensembles, in particular the *number of regression models within the ensemble*. Using the already selected model parameters, a 4-fold cross-validation study was done to determine the

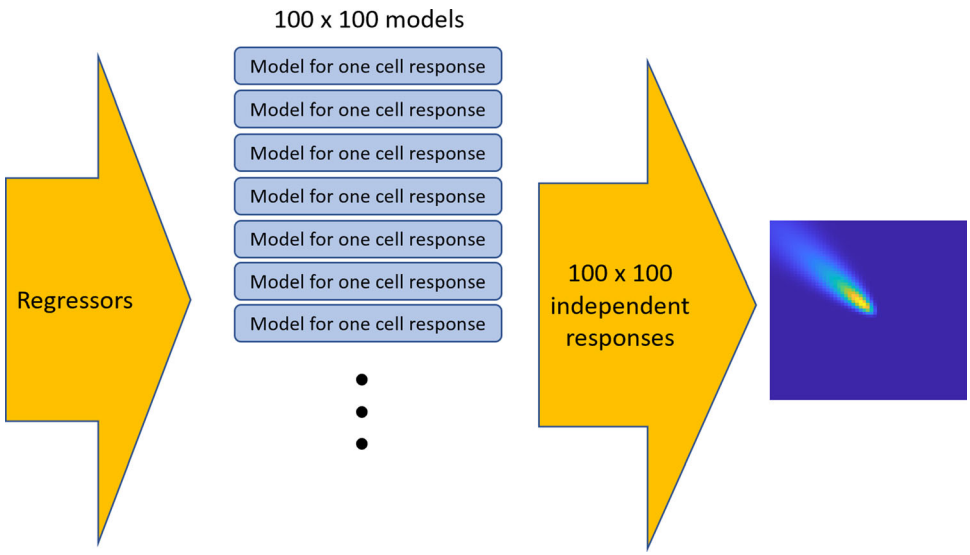

Fig. 4 The scheme of GIM

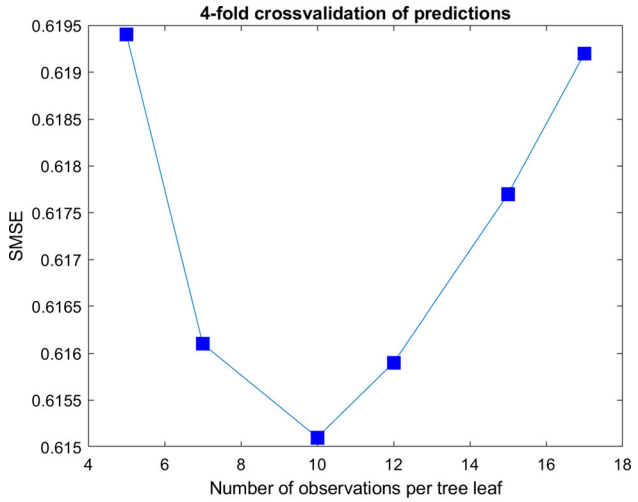

Fig. 5 The dependency of the predictions' SMSE from the number of observations per tree leaf

optimal number of models. The obtained results are given in "Appendix A" and the average SMSE measure is shown in Fig. 6. SMSE was used for the evaluation because we are again evaluating the prediction ability of each independent submodel in the GIM model.

It is clear from Fig. 6 that the quality of the model increases with the increasing number of models in ensembles. However, the computation burden also increases with the number of models. Consequently, we selected the corresponding number of models, where the change of SMSE is the biggest. This can be seen in Fig. 6 as a 'knee' of depicted function. Having in mind the computational burden, 60 models were selected as an acceptable number of models in ensembles.

The final structure was therefore as follows:

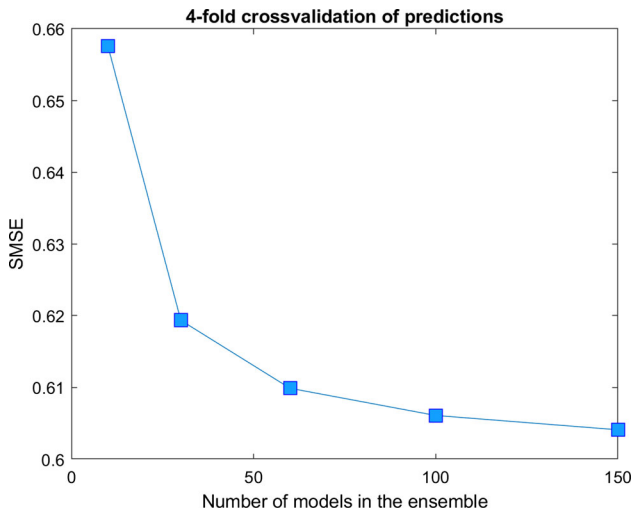

Fig. 6 The dependency of the predictions' SMSE from the number of models in the ensemble

- NFIR model structure,
- 7 signals as inputs, each delayed up to 4 time steps, i.e. 2 h, which results in 28 regressors,
- ensembles of regression trees composed of 60 models with 10 observations per leaf.

4.1.3 Test results

The complete set of data excluding the test data was used for training. The Statistics and Machine Learning Toolbox of Matlab was used for the training and testing of the obtained models.

Two examples of the obtained images for two different weather situations contained in the test data are given in Figs. 7 and 8. The complete set of test-data responses can be seen in the video (Online Resources 1 and 2—(Kocijan et al. 2022)). The visual matching of the predictions of the grid of independent submodels and the original Lagrangian particle dispersion model is relatively good. However, it depends on the purpose of the surrogate model which level of accuracy and details are important.

While the direct comparison of the computation time was not possible because the Lagrangian particle dispersion model was run on a designated computer (Intel Core i9 10900 @ 5.60 GHz, 32 GB RAM) and the surrogate model was run on another computer (Intel Core i7 8700HQ CPU @ 3.70GHz, 32 GB RAM), the comparison can be done only qualitatively. The purpose of the surrogate model was to make long-range predictions. For the prediction of approximately 1000 data samples, the original dispersion model at the designated computer took about 35,000 s, while the surrogate model took about 300 s. The computing load increased with the number of predictions linearly. This very rough comparison provides an indicator that the prediction with the surrogate model is much faster than with the Lagrangian particle dispersion model.

Regarding the accuracy of the model, we can compare the SMSE of the predictions on the test set for each of the submodels, that is for each of the 10,000 cells. These are graphically shown in Fig. 9 as an image. Differences in the quality of predictions can be observed in the figure, which is mainly an indicator of the different information content of inputs and outputs concerning cells of the two-dimensional representation. The average SMSE over the independent models is 0.5167, while R^2 = 0.4833 and R = 0.6952.

How the predictions of dispersion in each time instant cover the original model's predictions of dispersion is given with FMS values at each time instant of the test data sequence. The FMS values are graphically shown in Fig. 10. The average FMS over time is 0.595.

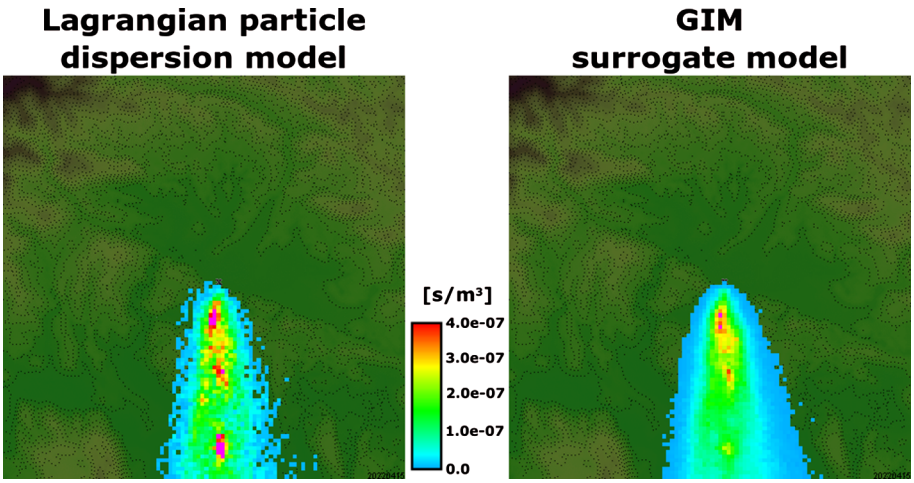

Fig. 7 An example of a weather situation with a strong wind from test data. Original-model response in the left figure, GIM surrogate-model response in the right figure. The scale is identical for both figures, values over the maximum value of scale are drawn in magenta colour, the maximum value in the left figure is $7.530 \cdot 10^{-7}$ s/m^3 and in the right figure is $4.190 \cdot 10^{-7}$ s/m^3

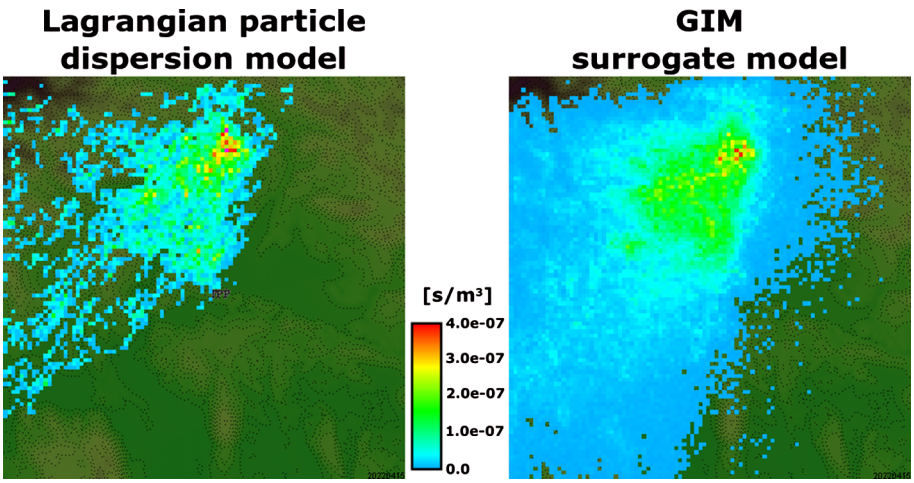

Fig. 8 An example of a weather situation with a weak wind from test data. Original-model response in the left figure, GIM surrogate-model response in the right figure. The scale is identical for both figures, values over the maximum value of scale are drawn in magenta colour, the maximum value in the left figure is $4.310 \cdot 10^{-7}$ s/m^3 and in the right figure is $3.830 \cdot 10^{-7}$ s/m^3

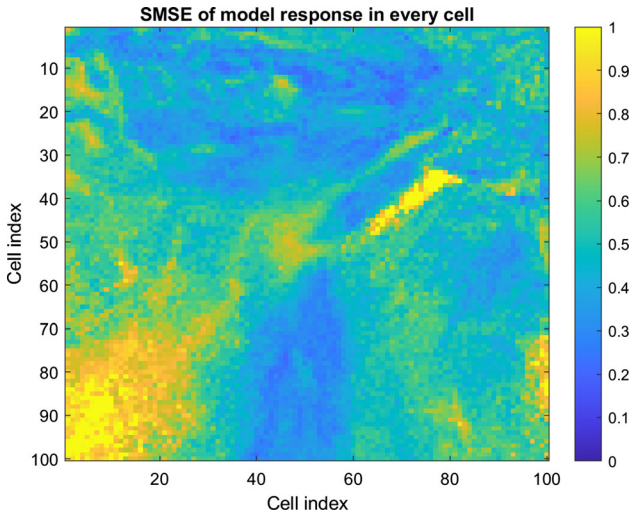

Fig. 9 The planar distribution of model-responses SMSE values on the test data, where the value 0 is the result of a perfect model

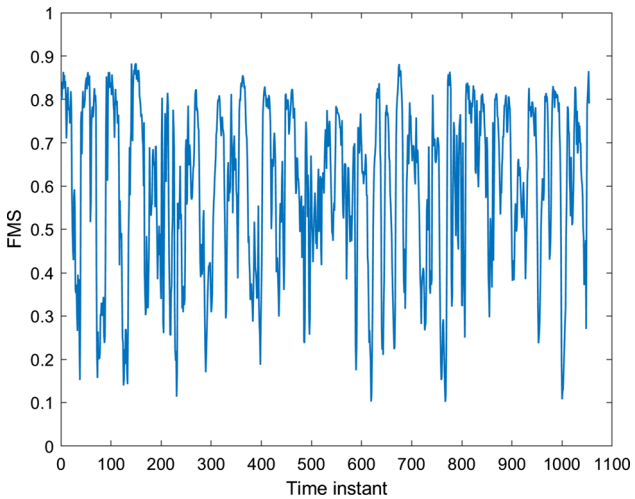

Fig. 10 FMS of the surrogate model for each time instant on the test data

Even though the GIM accelerates the prediction in comparison with the Lagrangian particle dispersion model, the following section describes an alternative in which a lower number of independent submodels is calculated and predictions in each cell are not independent anymore. This can be achieved with the RGI.

4.2 Reduced grid with interpolation of outputs—RGI

RGI is intended as the structure for the acceleration of computation in the comparison with GIM. The same *data* was used for modelling and testing and with the same division as in the case of GIM.

4.2.1 Structure

The entire model is composed of two parts: GIM, but not for the complete grid, and the Gaussian-process grid for filling up the gaps in the grid of predictions. The scheme of RGI is shown in Fig. 11.

The *structure of independent cell models* was kept as it was for the complete GIM. We selected every third cell to be modelled, which reduced the number of cells to be modelled on $34 \times 34 = 1156$ cells. This is a considerable reduction. In our case, the cells to be modelled with independent models are distributed uniformly, but this is not necessary.

Predictions of available independent cell models were inputs into the GP grid. The GP grid uses covariance functions for the calculation of the grid-based covariance matrix. The individual covariance functions in our case were two squared exponential covariance functions with isotropic distance measures, each of them as

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{r^2}{2l^2}\right). \quad (8)$$

The hyperparameter σ_f^2 represents the scaling factor of the possible variations of the function or the vertical scaling factor and the hyperparameter l is called the horizontal scaling factor and determines the relative weight on distance for the input variable $\mathbf{x}$. The variable r is the input-distance measure and is $r = |\mathbf{x}_i - \mathbf{x}_j|$, where $\mathbf{x}$ is a regression vector for the GP grid. In RGI's case, regressors of the GP grid are the cell's indices.

The Kronecker covariance matrix (Wilson and Nickisch 2015) is calculated from two covariance matrices, each for one input dimension.

Hyperparameters σ_f^2 and l were optimised for the best grid performance.

4.2.2 Test results

The predictions for two different weather situations are provided in Figs. 12 and 13.

SMSEs of the predictions on the test set for all models in the GIM model combined into an image is shown in Fig. 14 and FMS values at each time instant of the test data sequence are graphically shown in Fig. 15. The average SMSE over predictions on the test data and all cells is 0.5283, while $R^2 = 0.4717$ and $R = 0.6868$ and the average FMS over time is 0.599, which are both around 1 % difference in comparison with the GIM and consequently provides similar graphs.

It is clear from Figs. 12, 13, 14 and 15 that the accuracy of RGI predictions is close, but not equal to those of GIM. Especially, maximum values are not that well predicted. However, the computation of predictions for the test data, approximately 1000 data samples, takes about one-third of

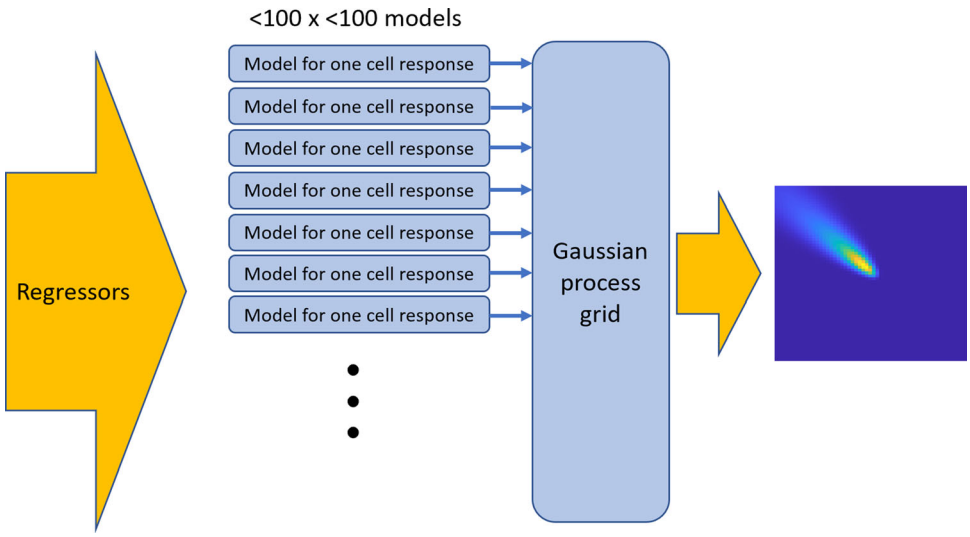

Fig. 11 The scheme of RGI

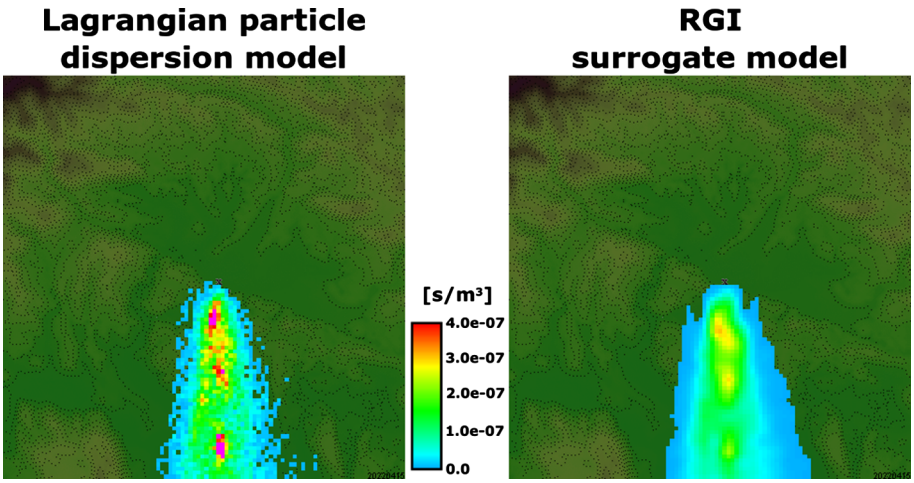

Fig. 12 An example of a weather situation with a strong wind from test data. Original-model response in the left figure, RGI surrogate-model response in the right figure. The scale is identical for both figures, values over the maximum value of scale are drawn in magenta colour, the maximum value in the left figure is $7.530 \cdot 10^{-7}$ s/m^3 and in the right figure is $3.070 \cdot 10^{-7}$ s/m^3

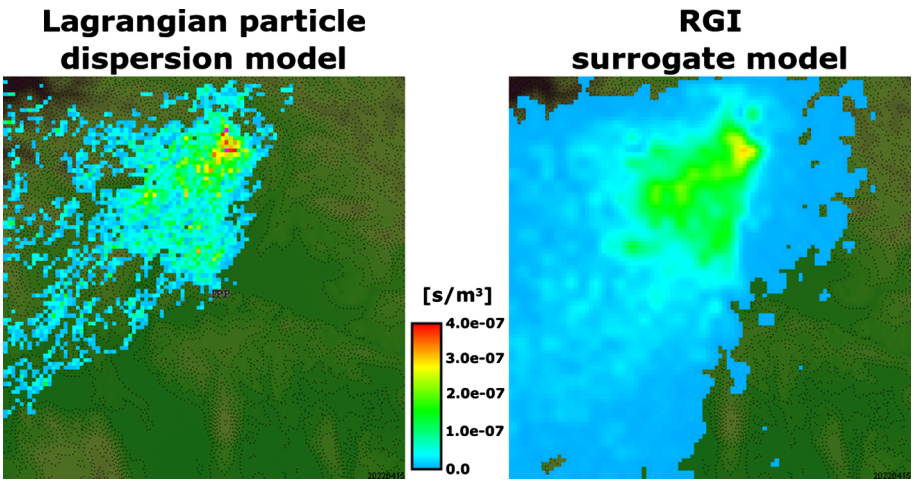

Fig. 13 An example of a weather situation with a weak wind from test data. Original-model response in the left figure, RGI surrogate-model response in the right figure. The scale is identical for both figures, values over the maximum value of scale are drawn in magenta colour, the maximum value in the left figure is $4.310 \cdot 10^{-7}$ s/m^3 and in the right figure is $2.980 \cdot 10^{-7}$ s/m^3

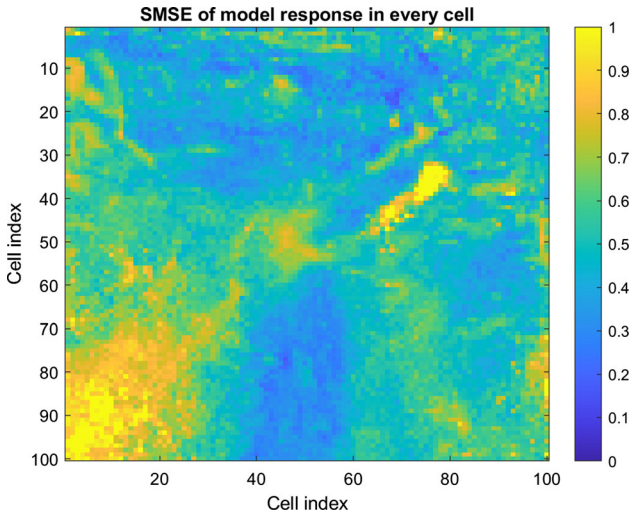

Fig. 14 The planar distribution of model-responses SMSE values on the test data, where the value 0 is the result of a perfect model

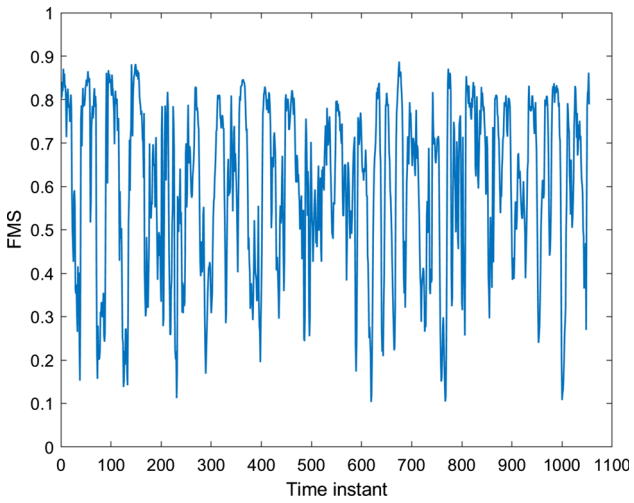

Fig. 15 FMS of the surrogate model for each time instant on the test data

the time with the combined model. The complete set of responses can be seen in the video (Online Resources 1 and 2—(Kocijan et al. 2022)).

4.3 Comparison of responses

A different aspect to the matching between the responses of the original model, GIM model and RGI model is provided with some additional processing. Figure 16 depicts the average relative concentrations over the domain calculated for the period of time of the test dataset. The comparison of responses of the original dispersion system and its surrogate model shows a good match in regards to both location and value. A very good match of maximum values of averages can also be observed. The response of the GIM surrogate model matches some of the details of the original system better than the response of the RGI surrogate model.

Figure 17 shows maximum relative concentrations over the domain calculated for the period of time of the test dataset. The comparison of responses of the original dispersion system and its surrogate model shows, like in Fig. 16, a good match location-wise as well as value-wise. The same is with the maximum values of maximum relative concentrations. The match of maximum values of surrogate models is a bit worse than the match of average relative concentrations. Nevertheless, these deviations are expected, because the prediction of maximum values is a difficult task also in the original system itself.

Figure 18 shows the 95th percentile of relative concentrations over the domain calculated for the time period of the test dataset. The match of responses and maximums of the 95th percentile of relative concentrations among models are good. The match is not as good as the other two comparisons, but still acceptable for our purpose. The response of the GIM surrogate model again matches some of the details of the original system better than the response of the RGI surrogate model.

The 95th percentile of relative concentrations represents the top 5 % response values. These are predicted well for the Seveso-type event in space and time. The comparison of maximum values in Fig. 17 confirms a good match based on previous evaluation investigations (Grašič et al. 2011) in the field of pollution dispersion. This all confirms that the developed surrogate models fulfilled the purpose of replacing the original dispersion system in the Seveso-type simulations at a lower computational cost.

The acceptable accuracy of a surrogate model depends on the purpose of the model. The modeller has to put enough effort into the modelling to achieve the acceptable accuracy. Moreover, if more data or data with better information content is used, a better model can be obtained.

One has to keep in mind that the training of models takes a considerably longer time than the prediction, but the time is still reasonable. The computational load of the model training increases nonlinearly with the number of training data, but the increase in computational load for surrogate-model predictions is linear with the number of test data and not considerable in comparison with the prediction time of the original Lagrangian particles model.

5 Conclusion

The objective of our investigation was to develop a surrogate model that will replace the air pollution Lagrangian particle dispersion model for computationally intensive applications like computer experimentations at a considerably lower computational time. We proposed two methods, i.e. GIM and RGI, that considerably raise the potential

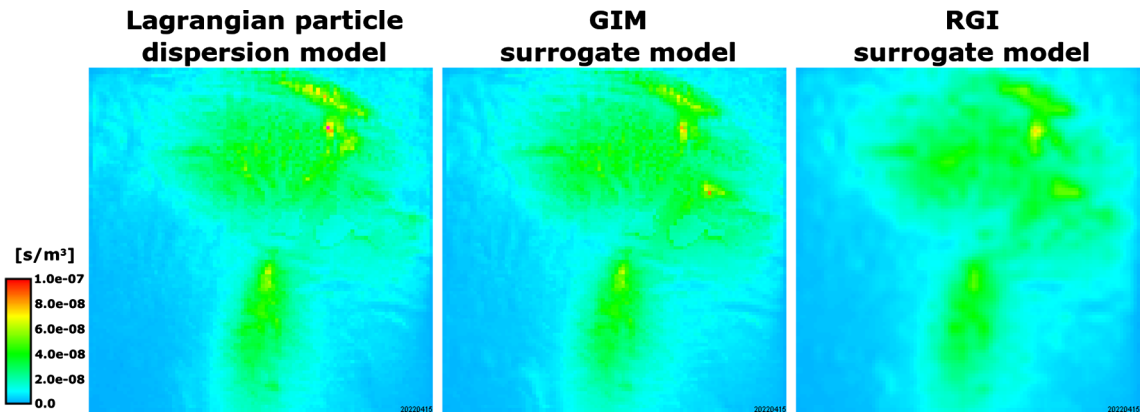

Fig. 16 Average relative concentrations over domain calculated for the validation dataset period June 9, 2021–June 30, 2021; the scale is identical for all figures; values over the maximum value of scale are drawn in magenta colour; the maximum value in the left figure is $1.004 \cdot 10^{-7}$ s/m^3, in the middle figure is $0.8470 \cdot 10^{-7}$ s/m^3 and in the right figure is $0.7273 \cdot 10^{-7}$ s/m^3

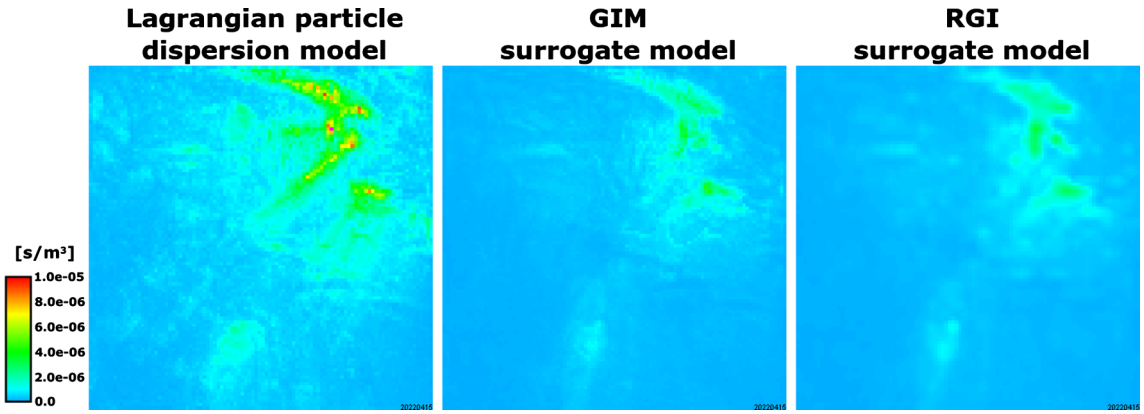

Fig. 17 Maximal relative concentrations over domain calculated for the validation dataset period June 9, 2021–June 30, 2021; scale is identical for all figures; values over the maximum value of scale are drawn in magenta colour; the maximum value in the left figure is $1.098 \cdot 10^{-5}$ s/m^3, in the middle figure is $0.4450 \cdot 10^{-5}$ s/m^3 and in the right figure is $0.3430 \cdot 10^{-5}$ s/m^3

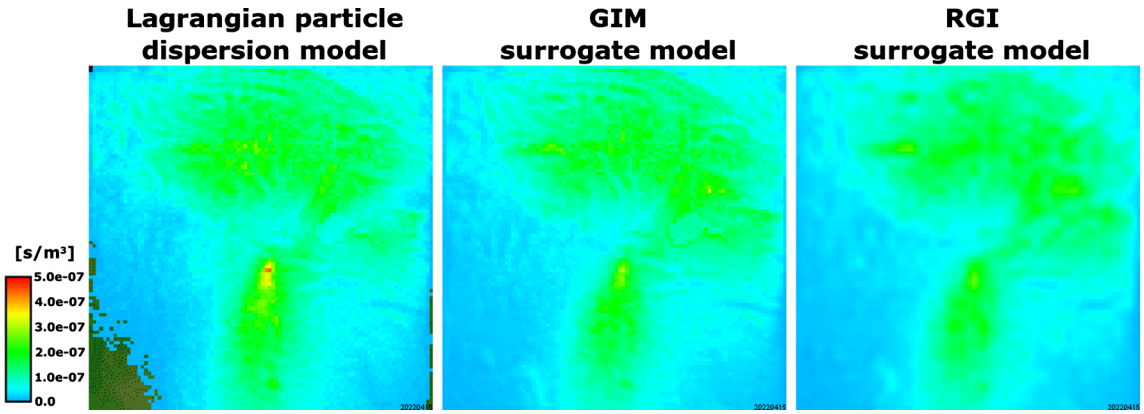

Fig. 18 The 95th percentile of relative concentrations over domain calculated for the validation dataset period June 9, 2021–June 30, 2021; scale is identical for all figures; values over the maximum value of scale are drawn in magenta colour; the maximum value in the left figure is $4.415 \cdot 10^{-7}$ s/m^3, in the middle figure is $3.200 \cdot 10^{-7}$ s/m^3 and in the right figure is $2.560 \cdot 10^{-7}$ s/m^3

for numerical experimentation. The obtained surrogate models can be used for computer experimentation like long-range predictions, simulations, parameter optimisation, etc., which would be very time-consuming with the air pollution Lagrangian particle dispersion model.

The accuracy of the surrogate model depends on the amount of training data used and its information content. The computational load for surrogate model development increases with the number of training data. However, the increase in computational load for predictions is linear for the proposed models and not considerable in comparison with the original model.

Other studies using surrogate models have not tackled the problem of air pollution dispersion in the same way as the present one. Two alternative methods that represent a solution of surrogate modelling for air-pollution dispersion based on dynamic models were demonstrated. While the idea of using a grid of models has been used before in somewhat different contexts, the use of dynamic GIM and RGI is novel for the dispersion-modelling problem of interest. Moreover, the investigation demonstrated the utility of surrogate modelling in the modelling of air pollution dispersion over complex terrain.

Further studies exploring more complex models, using different model structures, output-reduction methods and different kinds of dispersion problems are envisaged in future research.

Appendix A: Cross-validation results

Results of 4-fold cross-validation for delay-selection are given in Table 1 for the SMSE cost function that represents the average quality of GIM and in Table 2 for the FMS cost function that represents the quality of the cover of forecasted pollution plumes.

Table 1 Table of delays SMSE—4-fold cross-validation

Max delay	1st fold	2nd fold	3rd fold	4th fold	Average
− 2	0.6301	0.7234	0.6065	0.6450	0.6513
− 3	0.5973	0.6933	0.5811	0.6291	0.6252
− 4	0.5862	0.6847	0.5765	0.6301	**0.6194**
− 5	0.5865	0.6900	0.5817	0.5817	0.6242
− 6	0.5912	0.6946	0.5884	0.6473	0.6304

Table 2 Table of delays FMS—4-fold cross-validation

Max delay	1st fold	2nd fold	3rd fold	4th fold	Average
− 2	0.507	0.395	0.514	0.546	0.491
− 3	0.522	0.403	0.520	0.550	0.499
− 4	0.528	0.407	0.520	0.547	**0.501**
− 5	0.523	0.406	0.514	0.540	0.496
− 6	0.517	0.403	0.508	0.532	0.490

Table 3 Table of SMSE of model predictions depending on observations per tree leaf—4-fold cross-validation

No. of models	1st fold	2nd fold	3rd fold	4th fold	Average
5	0.5862	0.6847	0.5765	0.6301	0.6194
7	0.5827	0.6772	0.5750	0.6294	0.6161
10	0.5824	0.6719	0.5757	0.6303	**0.6151**
12	0.5831	0.6715	0.5774	0.6318	0.6159
15	0.5851	0.6715	0.5800	0.6341	0.6177
17	0.5866	0.6725	0.5822	0.6355	0.6192

Table 4 Table of number of models SMSE in ensembles—4-fold cross-validation

No. of models	1st fold	2nd fold	3rd fold	4th fold	Average
10	0.6252	0.7260	0.6192	0.6600	0.6576
30	0.5862	0.6847	0.5765	0.6301	0.6194
60	0.5768	0.6746	0.5657	0.6227	0.6099
100	0.5728	0.6706	0.5615	0.6195	0.6061
150	0.5708	0.6685	0.5593	0.6180	0.6041

Results of the 4-fold cross-validation study for the determination of the optimal number of observations per leaf are given in Table 3.

Results of the 4-fold cross-validation study for the determination of the optimal number of models are given in Table 4.

Author Contributions All authors contributed to the study conception, material preparation, analysis and design. Data collection was performed by BG and PM. The first draft of the manuscript was written by JK and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors acknowledge projects "Sources, transport and fate of persistent air pollutants in the environment of Slovenia", ID J1-1716, and "Modelling the dynamics of short-term exposure to radiation", ID L2-2615, and research core funding No. P2-0001, which were financially supported by the Slovenian Research Agency.

Declarations

Conflict of interest The authors have no relevant financial or nonfinancial interests to disclose.

References

Aleksovski D, Kocijan J, Džeroski S (2016) Ensembles of fuzzy linear model trees for the identification of multioutput systems. IEEE Trans Fuzzy Syst 24(4):916–929

Alizadeh R, Allen JK, Mistree F (2020) Managing computational complexity using surrogate models: a critical review. Res Eng Des 31(3):275–298

Božnar M, Lesjak M, Mlakar P (1993) A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. Atmos Environ Part B Urban Atmos 27(2):221–230

Bowman VE, Woods DC (2016) Emulation of multivariate simulators using thin-plate splines with application to atmospheric dispersion. SIAM/ASA J Uncertain Quantif 4(1):1323–1344

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L, Friedman JH, Olshen RA et al (2017) Classification and regression trees. Routledge

Carnevale C, Finzi G, Guariso G et al (2012) Surrogate models to compute optimal air quality planning policies at a regional scale. Environ Model Softw 34:44–50

Castelli ST, Armand P, Tinarelli G et al (2018) Validation of a Lagrangian particle dispersion model with wind tunnel and field experiments in urban environment. Atmos Environ 193:273–289

Desterro FS, Santos MC, Gomes KJ et al (2020) Development of a Deep Rectifier Neural Network for dose prediction in nuclear emergencies with radioactive material releases. Prog Nucl Energy 118(103):110

European Commission (2020) Major accident hazards: the Seveso directive—technological disaster risk reduction. https://ec.europa.eu/environment/seveso/. Accessed 21 April 2020

Finardi S, Morselli MG, Brusasca G et al (1997) A 2-D meteorological pre-processor for real-time 3-D ATD models. Int J Environ Pollut 8(3–6):478–488

Finardi S, Tinarelli G, Faggian P et al (1998) Evaluation of different wind field modeling techniques for wind energy applications over complex topography. J Wind Eng Ind Aerodyn 74:283–294

Francom D, Sansó B, Bulaevskaya V et al (2019) Inferring atmospheric release characteristics in a large computer experiment using Bayesian adaptive splines. J Am Stat Assoc 114:1450–1465

Girard S, Mallet V, Korsakissok I et al (2016) Emulation and Sobol'sensitivity analysis of an atmospheric dispersion model applied to the Fukushima nuclear accident. J Geophys Res Atmos 121(7):3484–3496

Girard S, Armand P, Duchenne C et al (2020) Stochastic perturbations and dimension reduction for modelling uncertainty of atmospheric dispersion simulations. Atmos Environ 224(117):313

Grašič B, Mlakar P, Božnar MZ (2011) Method for validation of Lagrangian particle air pollution dispersion model based on experimental field data set from complex terrain. In: Advanced air pollution. IntechOpen

Gunawardena N, Pallotta G, Simpson M et al (2021) Machine learning emulation of spatial deposition from a multi-physics ensemble of weather and atmospheric transport models. Atmosphere 12(8):953

Jiang P, Zhou Q, Shao X (2020) Surrogate model-based engineering design and optimization. Springer

Keane A, Forrester A, Sobester A (2008) Engineering design via surrogate modelling: a practical guide. American Institute of Aeronautics and Astronautics, Inc

Kocijan J (2016) Modelling and control of dynamic systems using Gaussian process models. Springer

Kocijan J, Hvala N, Perne M et al (2022) Surrogate modelling for the forecast of Seveso-type atmospheric pollutant dispersion. https://doi.org/10.5281/zenodo.6820934

Koziel S, Leifsson L (2013) Surrogate-based modeling and optimization. Springer

Lauret P, Heymes F, Aprin L et al (2016) Atmospheric dispersion modeling using artificial neural network based cellular automata. Environ Model Softw 85:56–69

Le NBT, Mallet V, Korsakissok I, et al (2019) Calibration of a surrogate dispersion model applied to the Fukushima nuclear disaster. In: 3rd international conference on uncertainty quantification in computational sciences and engineering, UNCECOMP 2019, pp 215–228

Mallet V, Tilloy A, Poulet D et al (2018) Meta-modeling of ADMS-Urban by dimension reduction and emulation. Atmos Environ 184:37–46

Mendes-Moreira J, Soares C, Jorge AM et al (2012) Ensemble approaches for regression: a survey. ACM Comput Surv 45(1):1–40

Mendil M, Leirens S, Armand P, et al (2021) Synthetic data and deep neural networks for atmospheric dispersion modelling in urban areas. In: 20th international conference on harmonisation within atmospheric dispersion modelling for regulatory purposes

Mendil M, Leirens S, Armand P et al (2022) Hazardous atmospheric dispersion in urban areas: a deep learning approach for emergency pollution forecast. Environ Model Softw 152(105):387

Mlakar P, Božnar MZ, Grašič B et al (2015) Air pollution dispersion models validation dataset from complex terrain in Šoštanj. Int J Environ Pollut 57(3–4):227–237

Mlakar P, Božnar MZ, Grašič B (2019) Relative doses instead of relative concentrations for the determination of the consequences of the radiological atmospheric releases. J Environ Radioact 196:1–8

Moonen P, Allegrini J (2015) Employing statistical model emulation as a surrogate for CFD. Environ Model Softw 72:77–91

Mosca S, Graziani G, Klug W et al (1998) A statistical methodology for the evaluation of long-range dispersion models: an application to the ETEX exercise. Atmos Environ 32(24):4307–4324

Nelles O (2002) Nonlinear system identification. IOP Publishing

Pal A, Mahajan S, Norman MR (2019) Using deep neural networks as cost-effective surrogate models for super-parameterized E3SM radiative transfer. Geophys Res Lett 46(11):6069–6079

Rasmussen CE, Williams CK (2006) Gaussian processes for machine learning, vol 2. MIT Press, MA

Ravina M, Esfandabadi ZS, Panepinto D et al (2021) Traffic-induced atmospheric pollution during the COVID-19 lockdown: dispersion modeling based on traffic flow monitoring in Turin, Italy. J Clean Prod 317(128):425

Wilson A, Nickisch H (2015) Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In: International conference on machine learning, PMLR, pp 1775–1784