Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 8314–8319

Surrogate grid model of an atmospheric pollutant spread

Juš Kocijan [*,**] Nadja Hvala [*] Boštjan Grašič [***]
Primož Mlakar [***]

[*] *Jožef Stefan Institute, Ljubljana, Slovenia*
(e-mail: jus.kocijan@ijs.si)
[**] *University of Nova Gorica, Nova Gorica, Slovenia*
[***] *MEIS d.o.o., Šmarje-Sap, Slovenia*

Abstract: This paper presents a method for developing computationally-efficient surrogate models for the spread of air pollution. Mitigating the pollution of Seveso-type accidents and designing evacuation scenarios require long-term prediction, which is obtained with numerical simulations of the spread of air pollution. Sophisticated simulation programs frequently possess high computational load and are not suitable for real-time computational studies and experiments. Data-driven surrogate models that are computationally fast are used for such investigations. We propose a grid of independent dynamical Gaussian-process models (GP-GIM) to simulate the spread of atmospheric pollution. This is demonstrated using a realistic example of limited complexity based on a thermal power plant in Šoštanj, Slovenia. The results show an acceptable behaviour match between the surrogate and original models, with a tenfold decrease in computational load. This confirms the feasibility of the proposed method and makes the resulting surrogate model suitable for further experiments.

Keywords: Machine learning for environmental applications, model reduction and dynamic emulation, modelling and identification of environmental systems

1. INTRODUCTION

This paper will investigate a method to develop a computationally efficient alternative model for the spread of air pollution.

Computational acceleration can be accomplished by building an approximation of the system's model known as a surrogate model, metamodel, emulator, or simulator. Developing surrogate models (Jiang et al., 2020) is an engineering method used when we cannot quickly compute the system's model response of interest. The method can reduce the computational cost required for various computationally intensive analyses. Data-driven approaches are frequently used for developing the surrogate model. The surrogate model, also the low-fidelity model, is identified from the input-output response of the original, also high-fidelity, simulation model. We capture the behaviour of the original mathematical model in the area of interest with appropriately selected excitation signals.

Surrogate models have been used in different fields of science for various tasks (Alizadeh et al., 2020), including in atmospheric sciences for air-pollution modelling. The goals of developing surrogate models in air-pollution modelling range from predicting spatial deposition, e.g., (Gunawardena et al., 2021; Mendil et al., 2022) to quantifying uncertainty, e.g., (Francom et al., 2019). The core of all these applications from the literature is to replace a computationally intensive model with a faster alternative. Still, each is slightly different regarding the purpose of the model or the methods used. However, the surrogate models listed use only information about the present and not the past and are, therefore, not dynamic models. One possible reason is that using data from earlier time instants would significantly increase the size of the input space.

A surrogate model tackling a similar problem as in this study, but based on decision trees, is described in (Kocijan et al., 2022a). The added value of using Gaussian-process models in this study is a measure of prediction confidence depending on the data used for the model training.

Problem statement In Europe, serious accidents involving dangerous chemicals are prevented and controlled by the Seveso Directive (European Commission, 2020). Seveso-type industrial installations and nuclear power plants have the potential for accidents with severe consequences, especially when releases occur in complex terrain. In such an accident, rapidly available forecasts would be required due to the immense environmental and population impact. The problem we describe is developing a data-driven surrogate model to simulate the spread of pollutants from a single source over complex terrain. The spread of air pollution in complex environments is typically modelled and simulated using the Lagrangian particle dispersion model (Girard et al., 2020), which is accurate enough but computationally expensive. Moreover, spread models are dynamic models, so they add complexity.

[⋆] 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 the Slovenian Research Agency financially supported.

10.1016/j.ifacol.2023.10.1020

We aim to develop dynamic surrogate models that can provide rapid predictions with indications of uncertainty for the spread of air pollution. If surrogate models of potentially important Seveso objects are available in advance, emergency services can use weather forecasts during disasters to quickly compute spread forecasts. The study is targeted at the ground layer as that is where most people are exposed and which is the most complex layer due to the effects of topography and land use. Investigations for higher layers using the same methods and corresponding data can also be performed.

Contribution We propose a dynamic input-output surrogate modelling method of continuous pollution at the origin of a complex terrain based on meteorological variables obtained from a digital weather forecasting system or other meteorological information sources.

The contribution of the study is twofold: A method for designing an alternative air-pollution spread model based on the input meteorological variables and the 2D representation of relative pollutant concentrations at the output as a grid of independent Gaussian-process dynamic models (GIMs) for each output cell, and quick and applicable case study demonstration of the method presented in the Seveso-type point release simulation of pollutants over complex terrain.

The focus is on developing a surrogate model that reduces the computational load of the prediction, approximates the overall simulation response with acceptable accuracy and provides information about the uncertainty of the prediction model. The accuracy of predictions in particular points is of secondary importance.

The following section describes the air-pollution Lagrangian particle dispersion model at the selected location with complex terrain. Section 3 describes the Gaussian-process model and its use for solving the fast-spread prediction problem. Results are discussed in Section 4, and conclusions are gathered in Section 5.

2. SIMULATION MODEL

The case study to demonstrate the development of a surrogate model for pollution spread is the thermal power plant Šoštanj. A regular and continuous source of sulfur dioxide (SO_2) pollutant emission with unit value has been assumed and can be sized, if necessary, to correspond to actual situations. The location of the Šoštanj power plant is at the edge of the Velenje basin in Slovenia. The Alps surround it to the north and northwest. The basin consists of narrow valleys crossed by rivers and thus presents a very complex topography. Winds are more assertive on the upper levels and weaker in the basin. Temperature inversions in winter and other circumstances further complicate the situation.

Pollution dispersion in such a complex terrain has been successfully modelled using the Lagrangian particle dispersion model (Mlakar et al., 2015), which represents an appropriate method to deal with the complexity of the landscape. The Lagrangian particle dispersion model SPRAY (Castelli et al., 2018) was combined with the MINERVE (Finardi et al., 1998) diagnostic mass consistent wind-field model, and the SURFPRO (Finardi et al., 1997) meteorological preprocessor. Inputs to the Lagrangian particle dispersion model are meteorological variables from different weather stations or weather forecasting programs, numerical data of terrain elevation, and land cover data for the area.

Instead of using all weather variables as described in (Mlakar et al., 2015), we simplified the weather situation as follows. The temperature, wind speed and direction signals are provided only at two altitudes of 10 m and 500 m, at the location of the thermal power plant Šoštanj, and the global solar radiation. The output variable is the relative concentration of the pollutant. Relative concentration (s/m^3) used is the ratio between the absolute concentration of a pollutant in $\mu g/m^3$ and the emission rate (kg/s) (Mlakar et al., 2019). This makes it possible to rescale the outcome to any other form of pollutant emission. We studied the region of 15×15 km described with $50\times 50 = 2500$ square cells of 300×300 m each.

The Lagrangian particle dispersion model simulation software is run on an i9 desktop with MS Windows operating system. The specified computer model calculated each response from half an hour prediction interval in about a few tens of seconds. Such computational performance is too slow to be used for numerical testing, which would be necessary for real-time accident prediction and especially not for long-term predictive studies because the computation time increases linearly with the length of the forecast horizon. This is why a surrogate model has to be developed.

3. SURROGATE MODEL

3.1 The grid of independent models - GIM

Air-pollution spread is a dynamic system, and it makes sense that the surrogate model is also dynamic. The static model will only be an approximation of the immediate dynamic model. The goal of the entire system's model is to forecast extended time periods, so long-term forecasting and the entire system's model have as many outputs as there are cells in the grid. The general schema of the GIM is shown in Figure 1. In our case study, each model in the

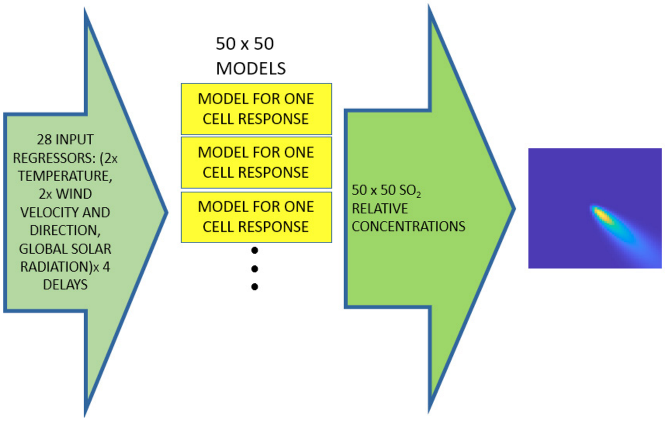
28 INPUT REGRESSORS: (2x TEMPERATURE, 2x WIND VELOCITY AND DIRECTION, GLOBAL SOLAR RADIATION)x 4 DELAYS
50 x 50 MODELS
MODEL FOR ONE CELL RESPONSE
MODEL FOR ONE CELL RESPONSE
MODEL FOR ONE CELL RESPONSE
50 x 50 SO2 RELATIVE CONCENTRATIONS

Fig. 1. The scheme of GIM with models' inputs and outputs indicated

GIM is a Gaussian-process (GP) model forming a GP-GIM model.

3.2 Gaussian-process model

Gaussian-process model and its development (Rasmussen and Williams, 2006) can be described as follows.

Let $\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon}$ be a model with the matrix of observed inputs $\mathbf{X} \in \mathcal{R}^{n \times D}$, latent function f, and observed output vector $\mathbf{y} \in \mathcal{R}^n$, where n is the number of observed data points. The observed output vector is presumed to be subject to white noise $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. Vector $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)]^T$ defines the vector of latent function values where $\mathbf{x}_n$ represents the n-th row of $\mathbf{X}$. The GP is then used to describe the prior over the vector of latent function values $\mathbf{f}$. It is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. It is specified by a mean function $m(\mathbf{x}_i)$ and a covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$

$$m(\mathbf{x}_i) = \mathbb{E}[f(\mathbf{x}_i)], \tag{1a}$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}[(f(\mathbf{x}_i) - m(\mathbf{x}_i))(f(\mathbf{x}_j) - m(\mathbf{x}_j))]. \tag{1b}$$

The distribution over the vector of latent function values $\mathbf{f}$ is defined with

$$\begin{aligned} p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) &= \mathcal{N}(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}) \\ &= \mathcal{N}(\boldsymbol{\mu}_f, \mathbf{K}_{ff}), \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}$ represents the hyperparameters, i.e., parameters of the mean and the covariance function with added likelihood noise. The likelihood is defined with $p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}, \mathbf{K}_{ff} + \sigma_n^2 \mathbf{I})$, where $\mathbf{y}$ is a noisy observation of $\mathbf{f}$ and σ_n^2 represents the variance of the likelihood noise. Without the loss of generality, the mean function is often selected as $m(\mathbf{X}) = \mathbf{0}$. The selection of the covariance function is more important as it incorporates our prior belief in the modelled function. A possible choice of the covariance function is the squared exponential covariance function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 e^{-\frac{1}{2l^2}||\mathbf{x}_i - \mathbf{x}_j||^2}, \tag{3}$$

where l represents a length scale parameter and σ_f is a scaling factor. Other covariance functions can be found in (Kocijan, 2016). In the case of the covariance function presented with equation (3), the hyperparameters are defined as $\boldsymbol{\theta} = Thel]$.

Learning Hyperparameters To find the hyperparameters $\boldsymbol{\theta}$, we first define a joint distribution of the observed and the unobserved vector of latent function values. Let $\mathbf{f}_*$ denote a vector of latent function values at the unobserved input $\mathbf{x}_*$. The joint distribution $p(\mathbf{f}, \mathbf{f}_*|\boldsymbol{\theta})$ is Gaussian and defined by

$$p(\mathbf{f}, \mathbf{f}_*|\mathbf{X}, \mathbf{x}_*, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{ff} & \mathbf{K}_{f*} \\ \mathbf{K}_{*f} & \mathbf{K}_{**} \end{bmatrix}), \tag{4}$$

where $\mathbf{K}_{ff}, \mathbf{K}_{f*}, \mathbf{K}_{**}$ represent the covariance matrices between the training inputs, training and the test inputs, and between the test inputs respectively. Joint distribution in equation (4) defines a prior, which is transformed to the posterior given the observed data

$$p(\mathbf{f}, \mathbf{f}_*|\mathbf{X}, \mathbf{x}_*, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}, \mathbf{f}_*|\mathbf{X}, \mathbf{x}_*, \boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})}. \tag{5}$$

The conditional dependency on $\mathbf{X}, \mathbf{x}_*$, and $\boldsymbol{\theta}$ are hereafter omitted for a more compact notation. Hyperparameters can be determined with the maximisation of the marginal log-likelihood in the denominator of equation (5) defined by $\log p(\mathbf{y}) = -\frac{1}{2}\log(|\mathbf{K}_{ff}|) - \frac{1}{2}\mathbf{y}^T \mathbf{K}_{ff}^{-1} \mathbf{y} - \frac{n}{2}\log(2\pi)$, with respect to $\boldsymbol{\theta}$. Learning the hyperparameters has a computational complexity of $\mathcal{O}(n^3)$, which limits the use of GPs for large datasets. The presented model is also limited to single-output and static problems.

Prediction The posterior distribution is obtained with the marginalisation over $p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}, \mathbf{f}_*|\mathbf{y}) d\mathbf{f}$. The mean and the variance of the predictive distribution can be evaluated in closed-form

$$\mu_*(p(\mathbf{f}_*|\mathbf{y})) = \mathbf{K}_{*f}[\mathbf{K}_{ff} + \sigma_n^2 \mathbf{I}]^{-1} \mathbf{y}, \tag{6a}$$

$$\sigma_*^2(p(\mathbf{f}_*|\mathbf{y})) = \mathbf{K}_{**} - \mathbf{K}_{*f}[\mathbf{K}_{ff} \sigma_n^2 \mathbf{I}]^{-1} \mathbf{K}_{f*}. \tag{6b}$$

Variational GP-NARX Model To reduce the computational complexity of GP models, sparse approximations consider m pseudo-input datapoints at location $\mathbf{x}_m$ with the corresponding vector of latent function values $\mathbf{u} = [u_1, \ldots, u_m]$. Assuming conditional independence of $\mathbf{f}$ and $\mathbf{f}_*$ given $\mathbf{u}$, their joint prior is approximated by $p(\mathbf{f}, \mathbf{f}_*) \cong \int p(\mathbf{f}|\mathbf{u}) p(\mathbf{f}_*|\mathbf{u}) p(\mathbf{u}) d\mathbf{u}$, where the distribution over the vector of latent function values conditioned on pseudo-input data points is given by

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \mathbf{u}, \mathbf{K}_{ff} - \mathbf{Q}_{ff}), \tag{7a}$$

$$p(\mathbf{f}_*|\mathbf{u}) = \mathcal{N}(\mathbf{K}_{*u} \mathbf{K}_{uu}^{-1} \mathbf{u}, \mathbf{K}_{**} - \mathbf{Q}_{**}), \tag{7b}$$

where $\mathbf{Q}_{ab} = \mathbf{K}_{au} \mathbf{K}_{uu}^{-1} \mathbf{K}_{ub}$. The model is completely specified and has a closed-form solution given the conditionals. The computational complexity of sparse approximations can be reduced to $\mathcal{O}(nm^2)$ with further assumptions (Rasmussen and Williams, 2006).

Variational Learning of Hyperparameters For a single-output model, variational learning lower bounds the marginal log-likelihood, approximating the joint distribution $p(\mathbf{f}, \mathbf{u}|\mathbf{y})$ by $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u}) q(\mathbf{u})$, where $q(\mathbf{u}) \sim \mathcal{N}(\mathbf{u}|\mathbf{m}, \boldsymbol{\Lambda})$ is chosen to be a free variational distribution. The parameters of the variational model are obtained by minimizing the Kullback-Leibler divergence between the variational distribution $q(\mathbf{f}, \mathbf{u})$ and the exact distribution $p(\mathbf{f}, \mathbf{u}|\mathbf{y})$. The lower bound is defined by $\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f})}\big[\log p(\mathbf{y}|\mathbf{f})\big] - \mathrm{KL}\Big[q(\mathbf{u})||p(\mathbf{u})\Big]$. To find the optimal parameters of the free variational distribution $q(\mathbf{u})$, the bound can be maximized with respect to the free variational distribution (Titsias, 2009) to obtain an optimal $q(\mathbf{u}) \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Lambda})$, where

$$\mathbf{m} = \boldsymbol{\Lambda}^{-1} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{y} \sigma_n^{-2}, \tag{8a}$$

$$\boldsymbol{\Lambda} = \mathbf{K}_{uu}^{-1} + \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \sigma_n^{-2}. \tag{8b}$$

The lower bound after finding the free variational distribution is defined by

$$\log p(\mathbf{y}) \geq \log\,[\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{Q}_{ff} + \sigma_n^2 \mathbf{I})] - \frac{1}{2\sigma_n^2} tr(\mathbf{K}_{ff} - \mathbf{Q}_{ff}), \tag{9}$$

where $\mathbf{Q}_{ff} = \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf}$.

The variational lower bound has a computational complexity of $\mathcal{O}(nm^2)$. The number of pseudo-input data points m

can be seen as a trade-off parameter between the model's accuracy and computational complexity.

Prediction The predictive distribution is obtained with marginalizing over the free variational distribution $p(\mathbf{f}_*^i) = \int p(\mathbf{f}_*^i|\mathbf{u})q(\mathbf{u})d\mathbf{u}$. The predictive distribution has a closed-form solution and is defined by

$$\mu_*(p(\mathbf{f}_*^i)) = \mathbf{K}_{*u}\mathbf{K}_{uu}^{-1}\mathbf{m}^i, \tag{10a}$$

$$\sigma_*^2(p(\mathbf{f}_*^i)) = \mathbf{K}_{**} - \mathbf{K}_{*u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u*} + \mathbf{K}_{*u}\mathbf{K}_{uu}^{-1}\mathbf{\Lambda}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u*}, \tag{10b}$$

where $\mathbf{m}^i = \mathbf{\Lambda}^{-1}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}\mathbf{y}^i\sigma_n^{-2}$ and $\mathbf{\Lambda}$ is shared between dimensions and was previously defined with equation (8b).

3.3 The modelling procedure

The procedure for developing the GIM model in our case is as follows (Kocijan et al., 2022a).

- The development of a Lagrangian particle dispersion model with accuracy that is suitable for the purpose of the model.
- Generate a dataset with the Lagrangian particle dispersion model for surrogate modelling.
- Select a data-driven modelling method to be used.
- Select a structure of a surrogate model (regressors, regression method, etc.).
- Data-driven modelling of a large number of independent models, one for each cell of interest.
- The validation of the surrogate model's prediction with data not used for modelling.

3.4 Constraints and assumptions

The problem of modelling a system with a large number of outputs can be solved by dimensionality reduction methods, e.g. (Girard et al., 2020). The alternative solution is to split the model into many submodels, as in the case in (Gunawardena et al., 2021; Carnevale et al., 2012), assuming that the outputs of individual cells can be well modelled without interactions with neighbouring cells. This also means the predicted variance is not continuous over the cells. The results described below suggest that this working hypothesis provides relevant results.

The surrogate model development usually starts with generating input samples intelligently distributed throughout the input space. The optimal experimental design or active learning is a standard method for it. In the case of modelling atmosphere, the input variables used for modelling are generally meteorological variables. In our case, the values at the input were not obtained with the designed experiment. Instead, we used weather forecasts from a digital weather simulator. This was because the actual combinations of input-variable values cannot fill all the subspaces of the input-variable space. After all, not all possible combinations are viable, and many combinations never occur. We, therefore, use data from available meteorological sources and not from optimal experimental design or active learning. Consequently, the amount of utilised data is not optimally distributed, and a large amount of data is needed to include relevant information.

3.5 Performance metrics

The modelling performance was evaluated with selected cost functions. The first one is chosen to assess time-dependent predictions of every submodel in the entire system compared to the original system. This cost function is the standardised mean-squared error – SMSE (Rasmussen and Williams, 2006): SMSE $= \frac{1}{N}\frac{\|\mathbf{y}-E(\hat{\mathbf{y}})\|^2}{\sigma_{\mathbf{y}}^2}$, where $\mathbf{y}$ is the vector of observations, $E(\hat{\mathbf{y}})$ is the mean value of estimations $\hat{\mathbf{y}}$, $\sigma_{\mathbf{y}}^2$ is the variance of observations and N is the number of observations. Pearson correlation coefficient R and the coefficient of determination R^2 are also given for comparison.

In our case, the presentation of pollution spread at ground level is a two-dimensional field - an image of the spread. Consequently, the second selected cost function, the statistical coefficient of the pollution space analysis, is the figure of merit in space – FMS (Mosca et al., 1998) also known as the Jaccard similarity coefficient FMS $= \frac{A_1 \bigcap A_2}{A_1 \bigcup A_2}$, where A_1 and A_2 represent the measured and predicted pollution areas, respectively. FMS is calculated at each time point, with a fixed threshold concentration level distinguishing two types of concentration values. Therefore, it does not confirm the concentration level but the pollution level. An FMS value close to 1 corresponds to a good performance of the model. Low FMS does not necessarily reflect poor model performance due to the moving pollution plumes. Therefore, the FMS value should be evaluated using a graphical representation of the regions measured A_1 and modelled A_2.

4. MODELLING AND RESULTS

The *data* for training, validation and testing are collected by simulation of the Lagrangian particle dispersion model. More details about the simulation model to be replaced are in (Mlakar et al., 2015). The data series contains three-year data (July 2018 - July 2021) with a sampling period of 30 minutes.

This set of more than 52,500 data points has been divided into training, validation and testing sets. To get as much data as possible for the training, we split the data into 51 subsets, one of which (June 2021 - July 2021) was immediately used as a test dataset. The rest of the data is used for training and validation (July 2018 - May 2021).

We chose Finite Impulse Response (FIR)*model structure*(Kocijan, 2016) for our submodels because too many outputs would make autoregressive models very inconvenient. The number of input lags corresponds to the time during which the impulse excitation response diminishes. Since air-pollution spread is a nonlinear process, we used a nonlinear Gaussian-process FIR (GP-NFIR) model. Gaussian process models are chosen because of the random variable at the output; the variance of this variable can be interpreted as a measure of confidence in the prediction. Other appropriate modelling methods can be used for the same purpose. The structure of the GP model for each cell includes a squared-exponential covariance function with automatic relevance determination, a zero mean function, and 257 pseudo-input data points. These parameters are arbitrarily determined based on a trade-off between model accuracy and GP-GIM training time. A cross-validation study on a single cell shows no significant difference in results when other covariance functions that provided top results are used.

The seven available input signals are used as inputs to the model, while the output is the relative concentration of SO_2 at each cell. The number of input lags was selected in the previous study (Kocijan et al., 2022a). We used the same lag on all inputs. The best SMSE and FMS results were obtained with a lag of 4 samples, corresponding to a lag of up to two hours. It is conceivable that the two-hour transient contains the most information about pollution spread.

The final structure was, therefore, as follows:

- nonlinear FIR model structure,
- 7 signals as inputs, each lagged up to 4 timesteps, i.e. 2 hours, which results in 28 regressors.

The models of GP-GIM have been trained each as a single-input single-output dynamic model and not altogether as a multiple-input multiple-output model. The complete data set without test data was used for training. Tensorflow in Python script was used to train and test *the resulting model.*

Two examples of images taken in two different weather conditions in the test data are shown in Figures 2 and 3. A complete set of test-data responses can be seen in a video (Kocijan et al., 2022b). The visual agreement between the predictions of the independent submodels and the original Lagrangian particle dispersion model is relatively good for our demonstration. The confidence of predictions is also clear from the Figures and indicates the reliability and trustworthiness of the results. Although there are no visual differences between the variances in Figures 2 and 3, a closer numerical inspection confirms that the figures are unequal.

The SMSE of predictions on the test set for each submodel, i.e. each of the 2500 cells, can be used to evaluate the accuracy of predictions. The average SMSE across the independent models is 0.5443, with $R^2 = 0.4557$ and R = 0.6750. These indicators are relatively low, but the inspection of the video (Kocijan et al., 2022b) shows that the primary dynamics of spread are well caught, while the computation of predictions is efficient.

The FMS values indicate the matching of the land coverage among the surrogate and original model at each time point in the test data sequence. The average FMS over time is 0.484.

The Lagrangian particle dispersion model was run on a dedicated computer (Intel Core i9 10900 @ 5.60 GHz, 32 GB RAM). The surrogate model was run on another computer (Intel Core i7 8700HQ CPU @ 3.70 GHz, 32 GB RAM) utilising the available GPU. The comparison due to circumstances can, therefore, only be made qualitatively. To predict about 1000 data samples, the original spread model took about 35,000 seconds on a dedicated computer, while the surrogate model took about 650 seconds (54 times less). The computational load increased proportionally with the number of predictions in both cases. This very crude comparison shows that predictions by the surrogate model are much faster than predictions by Lagrangian particle dispersion models.

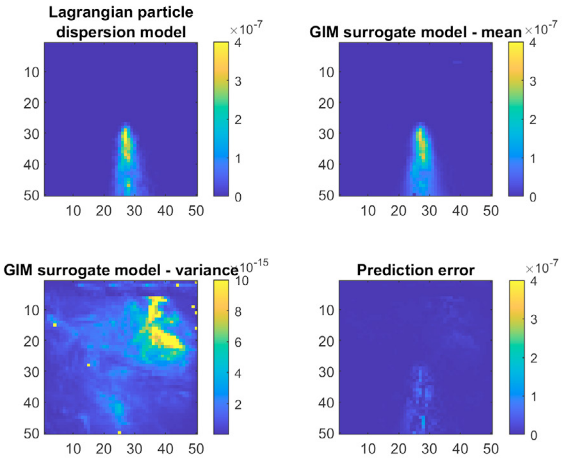

Fig. 2. An example of a weather situation with a strong wind from test data. The original-model response is in the left-top figure, and the GIM surrogate-model mean response is in the right-top figure. The scale is identical for both figures. GIM surrogate-model variance response is in the left-bottom figure, and the prediction error is in the right-bottom figure.

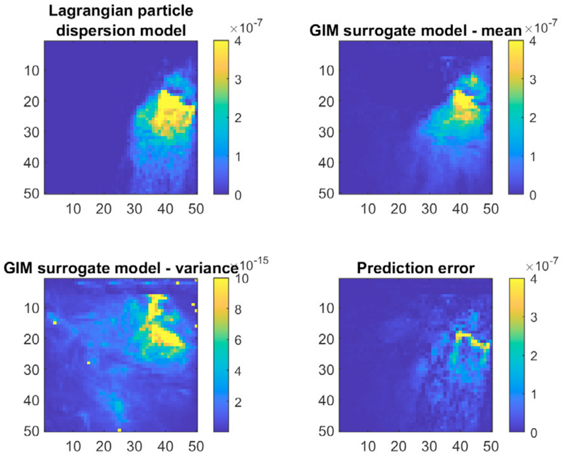

Fig. 3. An example of a weather situation with a weak wind from test data. The original-model response is in the left-top figure, and the GIM surrogate-model mean response is in the right-top figure. The scale is identical for both figures. GIM surrogate-model variance response is in the left-bottom figure, and the prediction error is in the right-bottom figure.

The surrogate model's training takes considerable, but still acceptable, time that is in the range of several hours. The computational cost of model training increases with the number of training data sets, as stated in Section 3.

The obtained results can be compared to those of (Kocijan et al., 2022a), where models based on ensembles of decision trees are used for modelling a similar problem with higher dimension (100×100 cells). While numerical results are similar, the advantage of GP modelling is additional information on prediction variances, and the disadvantage is a more considerable computational burden.

5. CONCLUSION

A method for developing a surrogate model to replace the air-pollution Lagrangian particle dispersion model for computationally-intensive applications is proposed in the paper. We demonstrated a grid of independent dynamical Gaussian-process models, which computationally facilitates numerical experiments. The resulting surrogate model can be used for long-term predictions or computationally intensive analyses instead of the Lagrangian particle dispersion model of air pollution.

The accuracy of the proposed surrogate model depends on the amount of training data used and its information content. While the computational load of surrogate-model training increases with the amount of training data nonlinearly, the computational load for prediction in the proposed model rises linearly. The computational burden of the surrogate model's prediction is much lower than that of the original Lagrangian particle dispersion model.

This study has used a different method than other studies using surrogate models for air-pollution spread, but the methods are difficult to compare on very different case studies. The contribution is the use of dynamic models and the utilisation of variances as confidence measures in 2D predictions of the surrogate model. These variances can be utilised for mining more data points where variances are relatively high. The idea of using a grid of models has been used before, but the proposed dynamic GP-based GIM for solving the modelling problem of interest is novel.

Future work will encompass different pollution situations and different model structures for better accuracy and improved computational efficiency.

ACKNOWLEDGEMENTS

We acknowledge the technical assistance of Rok Mokotar and Tadej Krivec with computer programming.

REFERENCES

Alizadeh, R., Allen, J.K., and Mistree, F. (2020). Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design*, 31(3), 275–298.

Carnevale, C., Finzi, G., Guariso, G., Pisoni, E., and Volta, M. (2012). Surrogate models to compute optimal air quality planning policies at a regional scale. *Environmental Modelling & Software*, 34, 44–50.

Castelli, S.T., Armand, P., Tinarelli, G., Duchenne, C., and Nibart, M. (2018). Validation of a Lagrangian particle dispersion model with wind tunnel and field experiments in urban environment. *Atmospheric environment*, 193, 273–289.

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, M.G., Brusasca, G., and Tinarelli, G. (1997). A 2–D meteorological pre–processor for real–time 3–D ATD models. *International Journal of Environment and Pollution*, 8(3-6), 478–488.

Finardi, S., Tinarelli, G., Faggian, P., and Brusasca, G. (1998). Evaluation of different wind field modeling techniques for wind energy applications over complex topography. *Journal of Wind Engineering and Industrial Aerodynamics*, 74, 283–294.

Francom, D., Sansó, B., Bulaevskaya, V., Lucas, D., and Simpson, M. (2019). Inferring atmospheric release characteristics in a large computer experiment using Bayesian adaptive splines. *Journal of the American Statistical Association*.

Girard, S., Armand, P., Duchenne, C., and Yalamas, T. (2020). Stochastic perturbations and dimension reduction for modelling uncertainty of atmospheric dispersion simulations. *Atmospheric Environment*, 224, 117313.

Gunawardena, N., Pallotta, G., Simpson, M., and Lucas, D.D. (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., and Shao, X. (2020). *Surrogate model-based engineering design and optimization.* Springer.

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

Kocijan, J., Hvala, N., Perne, M., Mlakar, P., Grašič, B., and Božnar, M.Z. (2022a). Surrogate modelling for the forecast of Seveso–type atmospheric pollutant dispersion. *Stochastic Environmental Research and Risk Assessment*, 1–16.

Kocijan, J., Hvala, N., Grašič, B., and Mlakar, P. (2022b). Surrogate grid model of an atmospheric pollutant propagation. URL `https://repo.ijs.si/e2pub/dispersion.git`.

Mendil, M., Leirens, S., Armand, P., and Duchenne, C. (2022). Hazardous atmospheric dispersion in urban areas: A deep learning approach for emergency pollution forecast. *Environmental Modelling & Software*, 152, 105387.

Mlakar, P., Božnar, M.Z., and Grašič, B. (2019). Relative doses instead of relative concentrations for the determination of the consequences of the radiological atmospheric releases. *Journal of environmental radioactivity*, 196, 1–8.

Mlakar, P., Božnar, M.Z., Grašič, B., Brusasca, G., Tinarelli, G., Morselli, M.G., and Finardi, S. (2015). Air pollution dispersion models validation dataset from complex terrain in Šoštanj. *International Journal of Environment and Pollution*, 57(3-4), 227–237.

Mosca, S., Graziani, G., Klug, W., Bellasio, R., and Bianconi, R. (1998). A statistical methodology for the evaluation of long-range dispersion models: an application to the ETEX exercise. *Atmospheric Environment*, 32(24), 4307–4324.

Rasmussen, C.E. and Williams, C.K. (2006). *Gaussian processes for machine learning.* MIT press Cambridge, MA.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, 567–574.