Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 51-5 (2018) 43–48

Identification of Atmospheric Variable Using Deep Gaussian Processes ⋆

Mitja Jančič * **Juš Kocijan** ** **Boštjan Grašič** ***

* *Jozef Stefan Institute, Ljubljana, Slovenia (e-mail: mitjajancic@gmail.com).*
** *Jozef Stefan Institute, Ljubljana, Slovenia and University of Nova Gorica, Nova Gorica, Slovenia (e-mail: jus.kocijan@ijs.si).*
*** *MEIS d.o.o., Šmarje-Sap, Slovenia (e-mail: bostjan.grasic@meis.si).*

Abstract: Mathematical and physical modelling only provide an approximate description of the true nature of a dynamic system. The higher the accuracy of the model, the more likely it becomes analytically intractable; therefore, empirical models or black box models are used. When dynamic systems are considered as black box models, almost no prior knowledge about the system is considered. Deep Gaussian Processes, which use hierarchical structure to provide adequate identification of very complex systems, can be used to identify the mapping between the system input and output values. With the given mapping function, we can provide one-step ahead prediction of the system output values together with its uncertainty, which can be used advantageously. In this paper, we use deep Gaussian Processes to identify a dynamic system and evaluate the method empirically. In the illustrative case, we study one-step-ahead prediction of air temperature in the atmospheric surface layer.

Keywords: System identification, deep Gaussian Processes, atmospheric temperature, big data.

1. INTRODUCTION

A crucial issue in the case of an accident or the failure of a nuclear power plant's safety systems is that authorities are able to act efficiently and promptly. The action of the authorities includes following the prescribed security protocols for such a situation and the potential evacuation of local inhabitants in the vicinity of the nuclear power plant. Authorities can take advantage of an Integrated Assessment Modelling System that provides expert services to assess and predict the consequences of radioactive-material release to the atmosphere (Chang and Weng, 2013). This paper describes an evaluation that is a part of a modelling study of atmospheric variables necessary for the modelling of air-pollution with radionuclides. This modelling is intended to be a part of an Integrated Assessment Modelling System intended for the case of an accident at a nuclear plant situated in complex terrain. The Integrated Assessment Modeling System will be composed of meteorological, dispersion and exposure assessment models. The prediction of air-pollution dispersion in the vicinity, i.e., a radius of at least ten kilometers, is necessary for the action of the mentioned authorities. Nevertheless, the prediction of the highly focused climatic condition is a necessary input in the pollution-dispersion modelling system. The goal is, therefore, to get the best possible model of the 3D condition of the atmosphere above the area under observation and consequently the dynamics of the cloud of radionuclides.

The atmosphere, as a very complex system, highly depends on the terrain below it. Physical or deterministic models (Zhang et al., 2012a,b) contain relations among the physical and chemical variables and as such provide an insight into the atmospheric and air-quality dynamics (Zhang et al., 2012a). These kinds of air-quality models provide prognostic time- and spatially-resolved concentrations for various typical and atypical scenarios. Unfortunately, the forecasts of weather over complex terrain, which comprises a large part of Europe, are not yet sufficiently localised, and this is exactly the problem we address in our study. The input data for the correct function of air-pollution dispersion models are those signals which describe the condition of the atmosphere at the location of interest during the passage of the radioactive pollution cloud.

The alternative to physical models are empirical or statistical models. The main goal of empirical modelling or system identification is to match the modelled variable as close as possible to the measured variable based on available observations of other variables. New approaches to modelling based on numerical data have emerged in recent times. Research in the area of methodology and the application of Gaussian Processes (GPs) for numerical modelling has raised a great deal of interest. The pioneering work in this field is described in various works, e.g. (Rasmussen and Williams, 2006; Shi and Choi, 2011; Kocijan, 2016).

⋆ The authors acknowledge the project "Method for the forecasting of local radiological pollution of atmosphere using Gaussian process models", ID L2-8174, and research core funding No. P2-0001, which were financially supported by the Slovenian Research Agency. We are grateful to the NPP Krško for the measurement data from their automatic measuring system.

2405-8963 © 2018, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.06.197

The authors have shown that the approach is very useful for experimental modelling or identification and is often superior to other similar methods. However, it is a new approach in the forecasting of meteorological variables in association with the dispersion of radiological pollution.

The aim of this paper is to describe results of deep Gaussian process (GP) method (Damianou and Lawrence, 2013) evaluation for the modelling of an atmospheric variable, namely temperature. Deep GP modelling is a particular type of GP modelling. The purpose of this evaluation is to test the utility of the method for the air-quality and the big data modelling at the same time. Consequently, a temperature is selected to be modelled because it is a measured variable dependent from other variables that are also measured at the site of a nuclear plant, which facilitates the modelling procedure considerably.

In particular, the study focuses on the empirical modelling and the short-term prediction of air temperature 2 meters above the ground. The air temperature is meant to be only one of the inputs into a physical model for radiological dispersion. The air temperature is a complex dynamic system (Holton and Hakim, 2012). Besides daily variations, it also contains seasonal variations, temperature-drop after rain, heat waves, polar waves and other events impacting the air temperature. The temperature dynamics is relatively slow and related to the change of other meteorological variables, e.g., air pressure, winds, solar radiation. Trends of temperature variations may be dependent on a large number of variables.

Nevertheless, if a deep GP method proves itself as appropriate for the described kind of big data modelling problem, it can be used for the modelling of some other, indirectly measured, variables. There are also other methods that can be used for large amounts of data, e.g., (Kocijan et al., 2016), but are beyond the scope of this paper.

The rest of the paper is organised as follows. In section 2, the details of GP modelling with Gaussian process latent variable model and deep GPs are briefly described. In section 3, the results are reviewed and discussed. Finally, conclusions are drawn and lessons learnt are described.

2. GAUSSIAN PROCESS LATENT VARIABLE MODEL AND DEEP GAUSSIAN PROCESSES

2.1 Modelling with Gaussian processes

GP models are probabilistic, non-parametric models based on the principles of Bayesian probability. GPs can be seen as the kernel methods with a Bayesian interpretation (Rasmussen and Williams, 2006). A GP model does not approximate the modelled system by fitting the parameters of the selected basis functions, but implies a relationship among the measured data. The use of GP models and the properties for modelling are thoroughly described in (Rasmussen and Williams, 2006; Shi and Choi, 2011; Kocijan, 2016).

GP models can be used for regression, where the task is to infer a mapping from a set of N D-dimensional regression vectors represented by the regression matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ to a vector of output data $\mathbf{y} = [y_1, y_2, \ldots, y_N]$. The outputs are usually assumed to be noisy realisations of the underlying function $f(\mathbf{x}_i)$. A GP model assumes that the output is a realisation of a GP with a joint probability density function:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{m}, \mathbf{K}), \tag{1}$$

with the mean $\mathbf{m}$ and the covariance $\mathbf{K} = [k_{ij}]$ being functions of the inputs $\mathbf{x}$. Usually, the mean function is defined as $\mathbf{0}$, while the covariance function or kernel

$$k_{ij} = C(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

defines the characteristics of the process to be modelled, i.e., the statistical stationarity, smoothness, etc. The value of the covariance function $C(\mathbf{x}_i, \mathbf{x}_j)$ expresses the correlation between the individual outputs $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ with respect to the inputs $\mathbf{x}_i$ and $\mathbf{x}_j$. Assuming the statistically stationary data is contaminated with white noise, the most commonly used covariance function is the composition of the square exponential (SE) covariance function with 'automatic relevance determination' (ARD) hyperparameters (Kocijan, 2016) and a constant covariance function assuming white noise:

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \Lambda^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right] + \delta_{ij}\sigma_n^2, \tag{3}$$

where Λ^{-1} is a diagonal matrix $\Lambda^{-1} = \mathrm{diag}([l_1^{-2}, \ldots, l_D^{-2}])$ of the ARD hyperparameters, σ_f^2 and σ_n^2 are hyperparameters of the covariance function, and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The hyperparameters can be written as a vector $\theta = [l_1^{-2}, \ldots, l_D^{-2}, \sigma_f^2, \sigma_n^2]^T$. The ARD property means that $l_i^{-2}; i = 1, \ldots, D$ indicates the importance of the individual inputs. If l_i^{-2} is zero or near zero, it means that the inputs in dimension i contain only a little information and could possibly be discarded. Further covariance functions suitable for various applications can be found in, e.g., (Kocijan, 2016).

The common aim of regression is to predict the output y^* in an unobserved test location $\mathbf{x}^*$ given the training data, a known mean function and a known covariance function C. The output predictive distribution can be obtained by using the Bayes' rule. The effect of unknown hyperparameters θ has to be taken into account. This leads to a computationally very demanding, sometimes intractable, task. A frequently used approximate solution to the problem of computation is to estimate the hyperparameters by maximising the marginal likelihood from the Bayes' rule. The details of inferring hyperparameters can be found in (Rasmussen and Williams, 2006; Kocijan, 2016).

Once the hyperparameter values are obtained, the predictive normal distribution of the output for a new test input can be calculated using

$$\mu(y^*) = \mathbf{k}(\mathbf{x}^*)^T \mathbf{K}^{-1} \mathbf{y}, \tag{4}$$

$$\sigma^2(y^*) = \kappa(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}^*), \tag{5}$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}_1, \mathbf{x}^*), \ldots, C(\mathbf{x}_N, \mathbf{x}^*)]^{\mathrm{T}}$ is the $N \times 1$ vector of covariances between the test and the training cases, and $\kappa(x^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test input itself.

A prediction of the GP model, in addition to the mean value (4), also provides information about the confidence of the prediction using the prediction variance (5). Usually, the confidence in the prediction is interpreted with a 2σ interval. This confidence interval highlights the areas of

the input space where the prediction quality is poor, due to the lack of data or noisy data, by indicating a wider confidence interval around the predicted mean.

2.2 Gaussian Process Latent Variable Model

One type of GP model is the Gaussian Process Latent Variable Model (GP-LVM). GP-LVM is originally known as a dimensionality-reduction method proven to be very robust in big data applications (Damianou et al., 2016; Lawrence, 2006). Historically, GP-LVM was introduced for the needs of unsupervised learning (Damianou, 2015); however, transition to supervised learning requires a minimum effort of applying prior belief to input as it is illustrated in Fig. 1, where the vector of inputs $\mathbf{z}$ is added to variational GP-LVM.

The primary goal of a GP-LVM model in both cases of learning, besides dimensionality reduction, is to find a mathematical relation between high dimensional input vector $\mathbf{X} \in \mathbb{R}^{N\times D}$ and lower dimensional space of latent variables $\mathbf{Y} \in \mathbb{R}^{N\times P}$. The main challenge here is the unobserved vector $\mathbf{X}$ elements of input values in the case of unsupervised learning. The GP-LVM model provides an elegant solution to the challenge by treating the input vector as latent variables and at the same time deploying P independent GPs as prior belief (Lawrence, 2006): $\mathbf{f} = f(\mathbf{X}) = (f_1(\mathbf{X}), \cdots, f_p(\mathbf{X}))$ in a way that

$$f_j(\mathbf{X}) \sim \mathcal{GP}(0, C(\mathbf{X}, \mathbf{X}')), \quad j = 1, \cdots, P. \tag{6}$$

Choosing nonlinear covariance functions in equation (6) enables nonlinear dimensionality reduction of a given regression problem. More details on GP-LVM can be found in (Lawrence, 2006).

The problem of big data training can be solved by introducing auxiliary *inducing points* $\mathbf{U}$, which expand the probability space. Inducing points are interpreted as additional variables yet to be optimised by an optimisation algorithm (Snelson, 2006). They are used for low-rank approximations for covariance matrix, leading to highly reduced computational cost (Damianou, 2015).

Another challenging part of Bayesian methodology is propagation of prior probability $p(\mathbf{X})$ belief through nonlinear mapping function f. The optimisation algorithm requires the calculation of the joint probability

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})\Big(p(\mathbf{f}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}\Big)\mathrm{d}\mathbf{f}, \tag{7}$$

where $\mathbf{y}$ is a vector of targets, which corresponds to, e.g., regression model with single output. As it turns out, the inputs $\mathbf{X}$ of the kernel matrix $\mathbf{K}$ are contained in the joint probability (7) in a very complex nonlinear manner, leaving the integration over domain $\mathbf{X}$ in most cases intractable.

To avoid this intractability a standard variational Bayesian methodology is used to approximate the marginal likelihood of $p(\mathbf{y})$ with a variational lower bound (Damianou, 2015).

Effectively applying standard variational Bayesian methodology makes the lower bound of marginal likelihood $p(\mathbf{y})$ tractable, meaning we are able to propagate all the uncertainties related to the input data, even in the case of nonlinear mapping.

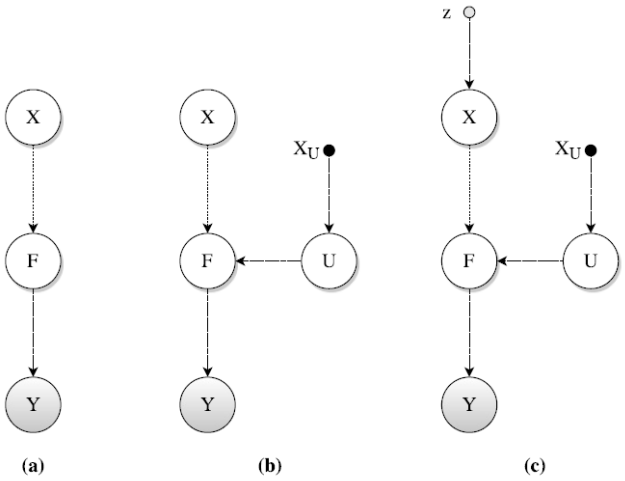

Fig. 1. Graphical representation of a) standard GP-LVM model, b) variational GP-LVM model and c) variational GP-LVM model for supervised learning.

GP-LVMs can be used to assemble deep GPs.

2.3 Deep Gaussian processes

Deep GPs were introduced as a flexible non-parametric approach to deep learning (Damianou, 2015; Damianou and Lawrence, 2013). The deep GP consists of L hidden layers of latent variables $\mathbf{h}_l$. Gaussian processes govern the mappings between the layers. Simply put, a deep GP model are nested GPs, in our case GP-LVMs, where outputs of a GP are treated as inputs to another GP (see Fig. 2):

$$\mathbf{y} = \mathbf{f}_{1:L} + \epsilon = f_L(f_{L-1}(\cdots f_1(\mathbf{X}))) + \epsilon, \tag{8}$$

where each f_i is an independent GP model and L number of hidden layers, also called the depth of a deep learning model. In this paper, white noise with normal distribution is added to the outputs of each layer (Damianou, 2015). Joint distribution of a deep GP model with L hidden layers can be written as

$$p(\mathbf{y}, \{\mathbf{h}_l\}_{l=1}^{L}) = p(\mathbf{y}|\mathbf{h}_l)p(\mathbf{h}_L, \mathbf{h}_{L-1}) \cdots p(\mathbf{h}_2|\mathbf{h}_1)p(\mathbf{h}_1) \tag{9}$$

for every set of latent variables $\mathbf{h}_l$.

The whole process itself is no longer interpreted as a GP model or GP-LVM. By recursing the procedure, we form multiple layers to an arbitrary depth of a deep learning model. Inputs to each layer in the hierarchical structure are considered to be latent, leading to an analytically tractable non-parametric model as explained in the previous subsection. This again leads to a numerical model appropriate for unsupervised learning where the transition to supervised learning effectively means only additionally bounding the input data with a prior belief, similar to the representation of supervised learning in Fig. 1(c).

Most of the deep learning algorithms need big data to identify mapping functions $f_i; i = 1 \ldots L$ (Goodfellow et al., 2016). Using complex and deep models on a low number of input data seems to be redundant, particularly when computational complexity and the amount of all necessary approximations to make the model analytically tractable are considered (Damianou, 2015).

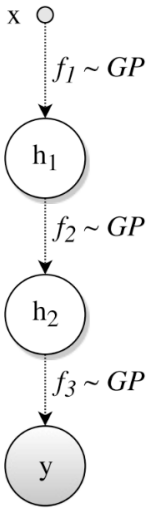

Fig. 2. Deep GP with two hidden layers

Using hierarchy provokes the question as to whether considering more hidden layers always results in better identification of a dynamic system, i.e., ignoring overfitting and the increase of the number of parameters. In practice, very deep models are rarely used as they normally result in identifying special cases instead of general behaviour of a dynamic system (Damianou, 2015). Otherwise said, it turns out that deep models focus their computational power on a very small sample of input data. Luckily, Bayesian methodology prevents the algorithm from looking for too complex structures in the input data (Hensman and Lawrence, 2014) in the first place.

The computational cost of NLM^3 of deep GP models, where M is number of inducing points, grows linearly with the number of layers L and remains almost unaffected by the number of output dimensions P (Damianou, 2015), which makes the model very usable for high dimensional data (Bui et al., 2016; Hensman and Lawrence, 2014). For comparison, the computational cost of GP-LVM is N^3 for supervised and NM^2 for unsupervised learning (Damianou, 2015).

3. RESULTS AND DISCUSSION

The wider domain of interest in our case is a size of 25 km × 25 km around the nuclear power plant in Krško, Slovenia. The model of interest, however, deals with variables at the site of the nuclear plant, which is situated in a complex terrain (Fig. 3).

The set of meteorological variables that we would like to use for the modelling of temperature should be as rich as possible. Nevertheless, we are constrained by available observations and we are trying to gain maximal information about modelled-variable dynamics from what is at hand.

A comparison of deep GP as a multilayer GP-LVM and a single GP-LVM is done in this study to highlight the utility of deep GP.

MEIS company has been pursuing measurement activities and analysis for the nuclear plant for years. Available

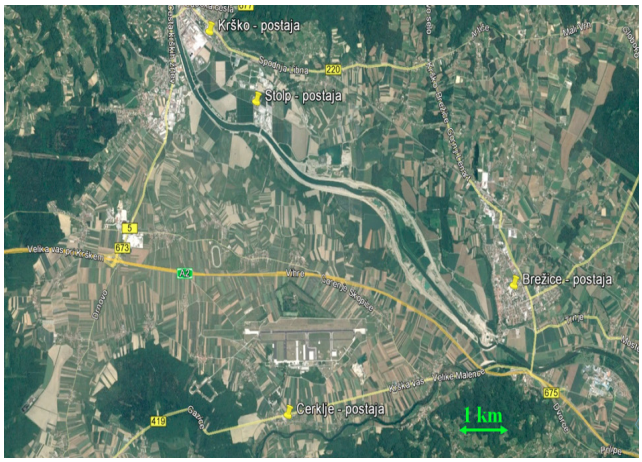

Fig. 3. The geographical features of the surrounding terrain and the measurement station. The plant and its measurement station (marked as 'Stolp – postaja') is situated in the basin surrounded by hills and valleys, which influence micro-climate conditions.

measurements for the period of four years (2013–2016), which can be used for modelling, are:

- temperature (T) 2 m above ground,
- relative humidity (φ) 2 m above ground,
- atmosphere stability (PG),
- air pressure (pr),
- global solar radiation (R), and
- wind speed (v) 10 m above ground.

The atmosphere stability is defined with Pasquill-Girard stability classes A–G, where A means extremely unstable and G extremely stable atmosphere (Air Resource Laboratory, 2009). Stability classes are defined qualitatively, e.g., class A means cloudless sunny day with less than 2 m/s general wind. All measurements are acquired at 30 min interval, which means approximately 17,500 samples per year per variable or approximately 70,000 samples per variable for four years. Gaussian process modelling is, in our case, pursued with software (Hensman et al., 2012) and (Dai et al., 2017).

The data are normalised with mean value 0 and variance 1 and further divided into identification and validation sets. The measurements from years 2013, 2014 and 2015 are used as training data, and 2016 as validation data. A training-data period of three years should be long enough to encompass for most of the seasonal and other temperature variations.

The regressor for system identification is selected using the backward-elimination method (May et al., 2011) starting with dynamic systems order or lag of 4. The evaluation criterion was normalised root-mean-square error (NMRSE), but the 95 % confidence interval was checked as well.

$$\text{NMRSE} = 1 - \frac{\|\mathbf{y} - \mu\|^2}{\|\mathbf{y} - E(\mathbf{y})\|^2}, \tag{10}$$

where $\mathbf{y}$ is the vector of validation values, μ is the vector of mean predicted values and $E(\mathbf{y})$ is the mean value of $\mathbf{y}$. NRMSE has value 1 for a perfect match and $-\infty$ for extremely bad match of validation and mean predicted values.

The final regression vector is obtained after an exhaustive search and contains 12 regressors as follows:

$$\begin{aligned}\mathbf{z} = [&T(k-2), T(k-1),\\ &\varphi(k-2), \varphi(k-1),\\ &PG(k-2), PG(k-1),\\ &R(k-4), R(k-3), R(k-2), R(k-1),\\ &pr(k-2), pr(k-1)]^{\mathrm{T}},\end{aligned} \tag{11}$$

where k represents consecutive time instant. The regressor selection procedure eliminated wind speed of 10 m above ground from regression vector. Values of identified hyper-parameters with their relative importance are shown in Fig. 4.

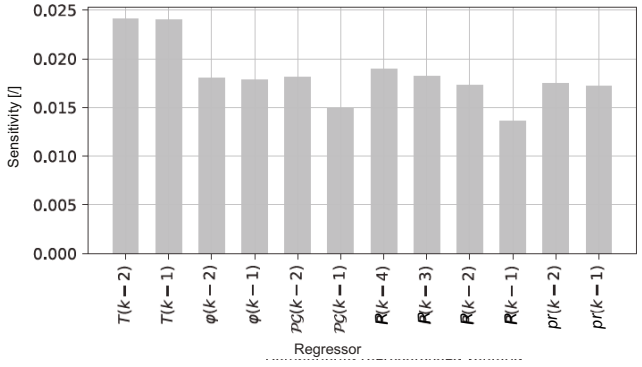

Fig. 4. Components of regression vector and their relative importance

Another structural issue that has to be solved using deep GP is the number of the deep model's layers. The number of layers is, in our case, determined with a trial-and-error method. One hidden layer proved to be enough for our modelling and provided the best NMRSE results.

The inducing points are, in our case, selected randomly with uniform distribution. 10 inducing points are utilised in every layer of deep GP, while GP-LVM contains 100 inducing points. Again these parameters are determined with a trial-and-error method. The best obtained results according to NRMSE and corresponding computational time after multiple runs are shown in Table 1.

Table 1. Comparison of NRMSE and computational time for both kinds of models on a desktop computer with 64-bit operation system, i5-6400 2,7 GHz quad-core processor and 24 GB RAM.

Model	No. induced varb.	Computn. time [s]	NRMSE
GP-LVM	100	400	0.923
deep GP	10	172	0.947

Segments of obtained prediction results for normalised validation data of the GP-LVM model and the deep GP model to illustrate prediction ability of the model are shown in Figs. 5 and 6. Absolute errors and 95 % confidence interval for both models are shown in Fig. 7.

Both identified models, GP-LVM and deep GP, predict temperature relatively well 2 m above ground as can be seen from evaluation criterion in Table 1 and from Fig. 7.

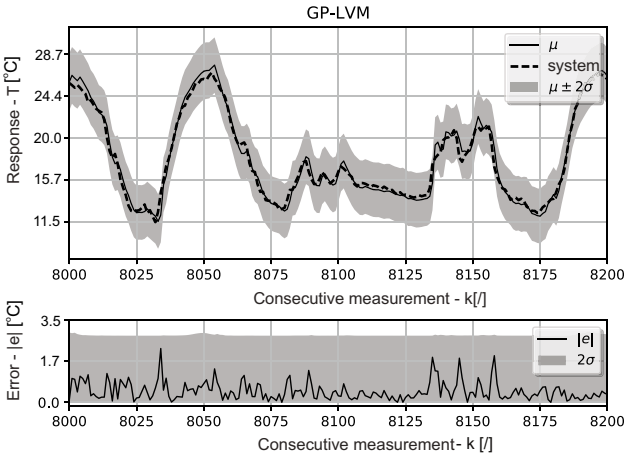

Fig. 5. One-step-ahead prediction with the GP-LVM model (mean value – full line, 95 % confidence interval – grey band) on normalised validation data (measured data – dashed line) with corresponding absolute error and 95 % confidence interval (bottom figure) – a segment of 200 samples to illustrate prediction ability of the model.

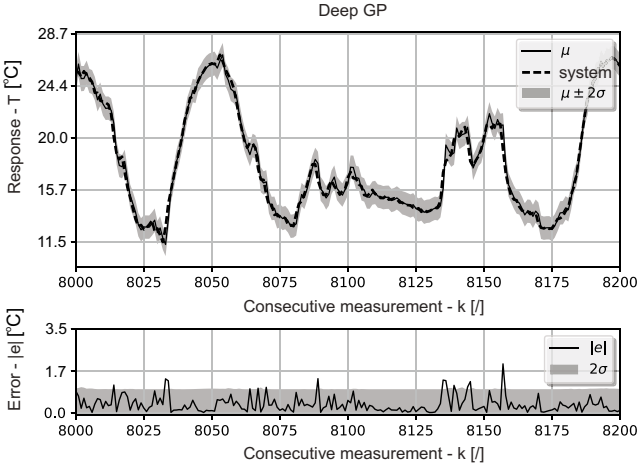

Fig. 6. One-step-ahead prediction with the deep GP model (mean value – full line, 95 % confidence interval – grey band) on validation data (measured data – dashed line) with corresponding absolute error and 95 % confidence interval (bottom figure) – a segment of 200 samples to illustrate prediction ability of the model.

Illustrative segments of results on Figs. 5 and 6 confirm this statement. The difference in NRMSE is not significant. The main difference is computational consumption. The deep GP model of the selected structure is identified faster and provides more realistic confidence bands. Nevertheless, it is important to keep in mind that deep GP becomes reasonable to utilise when the number of data is high, much higher than a few thousand data samples per variable, which is the number that can be handled with most other GP models.

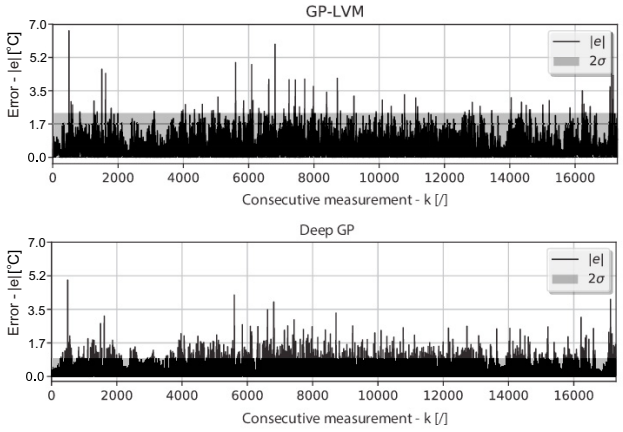

Fig. 7. Absolute error and 95 % confidence interval for model predictions on the entire set of validation data.

4. CONCLUSION

The paper describes an attempt to use deep GP modelling and prediction of air temperature 2 m above the ground. Accurate prediction of this variable is necessary as it is one of the variables that are used as the input into the physical model of pollution dispersion at the selected microlocation.

The obtained results show that deep GP models may be successfully used for the modelling when there are large amounts of measurement data. For the problem at stake, deep GP provided slightly better results more efficiently.

The experimentation with deep GP model structure shows no improvement in the case of multiple hidden layers, and random selection of induced variables provided satisfactory results. This does not mean that a more directed search of induced variables would not provide better results, but this investigation remains the topic of future research.

The temperature 2 m above ground is just one of many variables to be modelled and predicted. Other meteorological variables of interest such as a complete temperature profile containing heights up to a hundred meters, air pressure, global solar radiation, wind speed and direction, and others are yet to be modelled with this or other modelling methods. Moreover, other modelling methods will also be evaluated on the same and similar problems. The obtained models shall be accurate enough to enable long-range prediction.

REFERENCES

Air Resource Laboratory (2009). Pasquill stability classes. `www.ready.noaa.gov/READYpgclass.php`. Last access January 8, 2018.

Bui, T.D., Hernández-Lobato, J.M., Hernández-Lobato, D., Li, Y., and Turner, R.E. (2016). Deep Gaussian processes for regression using approximate expectation propagation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 1472–1481. JMLR.org.

Chang, N.B. and Weng, Y.C. (2013). Short-term emergency response planning and risk assessment via an integrated modeling system for nuclear power plants in complex terrain. *Frontiers of Earth Science*, 7(1), 1–27. doi:10.1007/s11707-012-0342-y.

Dai, Z., Damianou, A., and Gonzalez, J. (2017). PyDeepGP. `github.com/SheffieldML/PyDeepGP`. Last access January 8, 2018.

Damianou, A. (2015). *Deep Gaussian processes and variational propagation of uncertainty*. Ph.D. thesis, University of Sheffield.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In C. Carvalho and P. Ravikumar (eds.), *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, AISTATS '13, 207–215. JMLR W&CP 31.

Damianou, A.C., Titsias, M.K., and Lawrence, N.D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17(42), 1–62.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Hensman, J., Fusi, N., Andrade, R., Durrande, N., Saul, A., and Lawrence, N.D. (2012). GPy. `github.com/SheffieldML/GPy`. Last access January 8, 2018.

Hensman, J. and Lawrence, N.D. (2014). Nested variational compression in deep Gaussian processes. *arXiv preprint arXiv:1412.1370*.

Holton, J.R. and Hakim, G.J. (2012). *An introduction to dynamic meteorology*, volume 88. Academic press.

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

Kocijan, J., Gradišar, D., Božnar, M.Z., Grašič, B., and Mlakar, P. (2016). On-line algorithm for ground-level ozone prediction with a mobile station. *Atmospheric Environment*, 131, 326–333.

Lawrence, N.D. (2006). The Gaussian Process Latent Variable Model. `ftp.dcs.shef.ac.uk/home/neil/gplvmTutorial.pdf`. Last access January 8, 2018.

May, R., Dandy, G., and Maier, H. (2011). Review of input variable selection methods for artificial neural networks. In *Artificial neural networks-methodological advances and biomedical applications*. InTech.

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

Shi, J.Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press.

Snelson, E. (2006). Variable noise and dimensionality reduction for sparse Gaussian processes. In *Proc. of UAI-06*. AUAI Press.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A. (2012a). Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60, 632 – 655. doi:/10.1016/j.atmosenv.2012.06.031.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A. (2012b). Real-time air quality forecasting, part ii: State of the science, current research needs, and future prospects. *Atmospheric Environment*, 60, 656 – 676. doi:/10.1016/j.atmosenv.2012.02.041.