HIGH-ORDER GAUSSIAN PROCESS MODELS FOR PREDICTION OF OZONE CONCENTRATION IN THE AIR[1]

A. Grancharova[1], J. Kocijan[2, 3], A. Krastev[1], H. Hristova[1]

[1] Institute of Control and System Research, Bulgarian Academy of Sciences
Acad G. Bonchev str., Bl.2, P.O.Box 79, Sofia 1113, Bulgaria
[2] Department of Systems and Control, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
[3] University of Nova Gorica, Centre for Systems and Information Technologies
Vipavska 13, 5000 Nova Gorica, Slovenia

jus.kocijan@ijs.si (Juš Kocijan)

Abstract

Ozone is one of the main air pollutants with harmful influence over human health. Therefore, predicting the ozone concentration and informing the population when the air quality standards have been exceeded is an important task. In this paper, first- and high-order Gaussian process models for 1-hour ahead prediction of ozone concentration in the air of Bourgas, Bulgaria are identified and verified. For this purpose, the hourly measurements of the concentrations of ozone, sulfur dioxide, nitrogen dioxide, phenol and benzene in the air and the meteorological parameters, collected at the automatic measurement stations in Bourgas, are used.

Keywords: System identification, Ozone concentration prediction, Gaussian process models.

Presenting Author's biography

Jus Kocijan received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana. He is currently a senior researcher at the Department of Systems and Control, Jozef Stefan Institute and Professor of Electrical Engineering at the School of Engineering and Management, University of Nova Gorica. His main research interests are: applied nonlinear control and multiple model and probabilistic approaches to modelling and control. He is a member of SLOSIM - Slovenian Society for Simulation and Modelling, Automatic control society of Slovenia, IEEE.

[1] This work was financed by the National Science Fund of the Ministry of Education and Science of Republic of Bulgaria, contract №DO02-94/14.12.2008 and the Slovenian Research Agency, contract №BI-BG/09-10-005 ("Application of Gaussian processes to the modeling and control of complex stochastic systems")

1 Introduction

Ozone is one of the main air pollutants with harmful influence over human health. Standards which guarantee the human health protection are as follows [1]: *health protection level* 120 μg/m^3 eight hours mean concentration; *informing the public level* 180 μg/m^3 one hour mean concentration; *warning the public level* 240 μg/m^3 one hour mean concentration. Therefore, predicting the ozone concentration and informing the population when the air quality standards have been exceeded is an important task.

It has been shown in [2,3], that the ozone concentration has a strong daily cycle. Thus, the formation and collection of ozone in the air starts after 7h and it reaches its maximum between 13h and 16h. In [4], the relation between the ozone concentration and three meteorological parameters have been investigated using data about the region of Hessen in Germany. Based on these data, a linear regression model to predict the maximal daily ozone concentration in the air has been obtained. In [5,6], neural network models and Gaussian process models for ozone concentration forecasting in some regions of Slovenia have been developed and evaluated.

The region of Bourgas city is among the regions in Bulgaria with the highest ozone pollution of the air and thus it is of primary interest to obtain a prediction model for this region. In [7], *first-order* Gaussian process models for 1-hour ahead prediction of ozone concentration in the air of Bourgas are identified and verified based only on measurements of the air pollutants concentrations. The purpose of this paper is to develop high-order Gaussian process models for ozone concentration prediction by using measurements of both the air pollutants and the meteorological parameters. To achieve this, measurement data provided by the Executive Environmental Agency of Bulgaria are used.

The following notation will be used in the paper. For a random variable y with Gaussian distribution, $\mathcal{N}(\mu(y),\sigma^2(y))$ denotes its probability distribution, and $\mu(y)$ and $\sigma^2(y)$ are respectively its mean and variance.

2 Modelling of dynamic systems with Gaussian process models

The Gaussian process model is an example of a *non-parametric* probabilistic black-box model which, beside model predictions, inherently provides also the uncertainty of predictions. Its use and properties for modelling are reviewed in [8]. The use of Gaussian processes in the modelling of dynamic systems is a relatively recent development [9,10,11,12,13] and a retrospective review of dynamic systems modeling with Gaussian process models can be found in [14].

A Gaussian process is a collection of random variables which have a joint multivariate Gaussian distribution. Assuming a relationship of the form $y = f(z)$ between an input $z \in \mathbb{R}^D$ and output $y \in \mathbb{R}$, we have $y(1), y(2), \ldots, y(M) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\Sigma_{pq} = \mathrm{Cov}(y(p), y(q)) = C(z(p), z(q))$ gives the covariance between the output points $y(p)$ and $y(q)$ corresponding to the input points $z(p)$ and $z(q)$. Thus, the mean $\mu(z)$ (usually assumed to be zero) and the covariance function $C(z(p), z(q))$ fully specify the Gaussian process. Note that the covariance function $C(z(p), z(q))$ can be any function with the property that it generates a positive definite covariance matrix. A common choice is:

$$C(z(p), z(q)) = v_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{D} w_i (z_i(p) - z_i(q))^2\right] + v_0 \alpha_{pq} \quad (1)$$

where $\Theta = [w_1, \ldots, w_D, v_0, v_1]$ are the 'hyperparameters' of the covariance function, z_i denotes the i-th component of the D-dimensional input vector z, and α_{pq} is the Kronecker operator. The covariance function (1) is composed of two parts: the Gaussian covariance function for the modeling of system function and the covariance function for the modelling of noise. The noise, in our case, is presumed to be white. Other forms of covariance functions suitable for different applications can be found in [15]. For a given problem, the hyperparameters are learned (identified) using the data at hand. After the learning, one can use the w parameters as indicators of 'how important' the corresponding input components (dimensions) are: if w_i is zero or near zero it means that the inputs in dimension i contain little information and could possibly be removed.

Consider a set of M D-dimensional input vectors $\mathbf{Z} = [z(1), z(2), \ldots, z(M)]^T$ and a vector of output data $Y = [y(1), y(2), \ldots, y(M)]^T$. Based on the data $(\mathbf{Z}, Y)$, and given a new input vector z^*, we wish to estimate the probability distribution of the corresponding output y^*. Unlike other models, there is no model parameter determination as such, within a fixed model structure. With this model, most of the effort consists in *tuning* the parameters of the covariance function. This is done by maximizing the log-likelihood of the parameters, which is computationally relatively demanding since the inverse of the data covariance matrix ($M \times M$) has to be calculated at every iteration.

The described approach can be easily utilized for regression calculation. Based on a training set $\mathbf{Z}$, a covariance matrix $\mathbf{K}$ of size $M \times M$ is determined. As already mentioned before, the aim is to estimate the probability distribution of the corresponding output y^* at some new input vector z^*. For a new test input z^*, the predictive distribution of the corresponding output is $y^* \mid z^*, (\mathbf{Z}, Y)$ and is Gaussian, with mean and variance:

$$\begin{aligned} \mu(z^*) &= k(z^*)^T \mathbf{K}^{-1} Y \\ \sigma^2(z^*) &= k_0(z^*) - k(z^*)^T \mathbf{K}^{-1} k(z^*) \end{aligned} \quad (2)$$

where $k(z^*) = [C(z(1), z^*), ..., C(z(M), z^*)]^T$ is the $M \times 1$ vector of covariances between the test and training cases and $k_0(z^*) = C(z^*, z^*)$ is the covariance between the test input and itself.

Gaussian processes can be used to model static nonlinearities and can therefore be used for modelling of dynamic systems if delayed input and output signals are used as regressors [12]. In such cases an autoregressive model is considered, such that the current predicted output depends on previous estimated outputs, as well as on previous control inputs:

$$\begin{aligned} z(t) = [&\hat{y}(t-1), \hat{y}(t-2), ..., \hat{y}(t-L), u(t-1), \\ &u(t-2), ..., u(t-L)]^T \\ \hat{y}(t) = &f(z(t)) + \eta(t) \end{aligned} \quad (3)$$

where t denotes consecutive number of data sample, L is a given lag, and $\eta(t)$ is the prediction error. The quality of the mean values of predictions with a Gaussian process model can be assessed by computing the average squared error (ASE):

$$ASE = \frac{1}{M} \sum_{i=1}^{M} [\mu(\hat{y}(i)) - y(i)]^2 \quad (4)$$

and the log density error (LD) [8] is also a possible measure:

$$LD = \frac{\log(2\pi)}{2} \sum_{i=1}^{M} (\log[\sigma^2(\hat{y}(i))] + \frac{[\mu(\hat{y}(i)) - y(i)]^2}{\sigma^2(\hat{y}(i))}) \quad (5)$$

In (4), (5), $\mu(\hat{y}(i))$ and $\sigma^2(\hat{y}(i))$ are the prediction mean and variance, $y(i)$ is the system's output and M is the number of the training points.

The Gaussian process model now not only describes the dynamic characteristics of the non-linear system, but at the same time provides information about the confidence in the predictions. The Gaussian process can highlight areas of the input space where prediction quality is poor, due to the lack of data, by indicating the higher variance around the predicted mean.

3 Gaussian process models for prediction of ozone concentration in the air of Bourgas

3.1 Available data

Measurement data for the year 2008, collected at the automatic measurement station in the center of Bourgas, Bulgaria, are used. The data includes hourly measurements of the concentrations of ozone, sulfur dioxide, nitrogen dioxide, phenol and benzene. The meteorological parameters have not been measured at this station. However, in order to study how these parameters would influence the prediction of ozone concentration in the air of Bourgas, their measurements at two other stations in the regions of Bourgas city (in Dolno Ezerovo and Meden rudnik) are used. It is accepted that the values of the meteorological parameters in the center of Bourgas represent the average of the measurements collected in the stations in Dolno Ezerovo and Meden rudnik.

It should be noted that in the Gaussian process models, the mean hourly concentrations of ozone are used.

3.2 Daily cycle of hourly ozone concentrations

In the previous research [2,3], it has been shown that the ozone concentration has a strong daily cycle. Fig. 1 shows the daily cycle of the hourly ozone concentrations greater than 100 μg/m^3 [3]. The cycle is estimated by the count of data for every hour, normalized with the full data count. The count of data with concentration greater than 100 μg/m^3 has a minimum at 5–6h. After 7h it is growing fast, i.e. the formation and collection of ozone in the air starts after 7h. The maximum is reached at 13-16h, and between 11h and 17h envelop up to 60% of data. Therefore, if we are able to prognosticate correctly the ozone concentration at that hourly interval, we will be able to prognosticate the maximal hourly concentration for the day in all cases of high health risk.

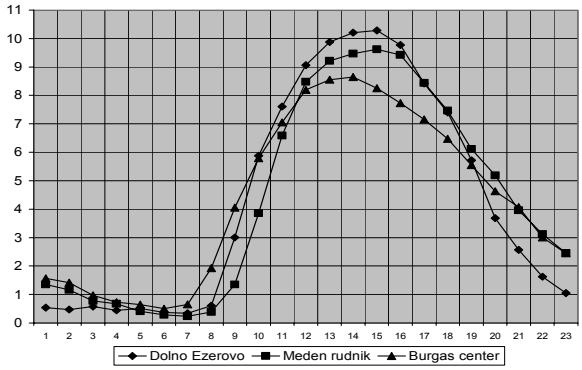

Fig. 1. Relative count of concentrations greater than 100 μg/m^3.

3.3 First-order Gaussian process models

Four *first-order* Gaussian process models for prediction of ozone concentration are identified and verified based on the available measurement data. The log-likelihoods (LL) of the obtained models, and the average squared errors (ASE) and the log density errors (LDE) computed for the training data and the validation data are given in Table 1.

The *training* data include the measurements from 9 till 16h at every 2-nd day of the year 2008. The reason to consider this time interval is related to the daily cycle of ozone concentration. Therefore, we are interested to obtain an accurate prediction of ozone concentration in the interval from 9 till 16h, where there is some risk to exceed the established air quality standards. Thus, the total number of the training data is 1032 corresponding to 129 days of the year. Note that the days, for which there is not a full collection of all measurements in the interval 9-16h, are excluded from the data set. The *validation* data include the measurements from 1 till 23h at every 9-th day of the year 2008 (by excluding the days which coincide with the 2-nd days). The days, for which there are measurements at some hours only, are excluded from the validation data set. The total number of the validation data is 299 corresponding to 13 days of the year.

The four identified *first-order* Gaussian process models have the following input parameters:

- The value of ozone concentration at the previous hour:

  Model A1: $c_{O_3}(t) = f_1(c_{O_3}(t-1))$ (6)

where t is the current hour of the day and c_{O_3} is the concentration of ozone in the air.

- The values of ozone concentration, the concentrations of the air pollutants, and the meteorological parameters at the previous hour:

  Model B1:

$$c_{O_3}(t) = f_2(c_{O_3}(t-1), c_{NO_2}(t-1), c_{SO_2}(t-1), c_{C_6H_5OH}(t-1), c_{C_6H_6}(t-1), h(t-1), p(t-1), sr(t-1), temp(t-1), ws(t-1)) \quad (7)$$

Here, c_{NO_2}, c_{SO_2}, $c_{C_6H_5OH}$, and $c_{C_6H_6}$ are the concentrations of nitrogen dioxide, sulfur dioxide, phenol and benzene in the air, h is the air humidity, p is the air pressure, sr is the sun radiation, $temp$ is the air temperature, ws is the wind speed.

- The values of ozone concentration and the concentrations of the air pollutants at the previous hour:

  Model C1:

$$c_{O_3}(t) = f_3(c_{O_3}(t-1), c_{NO_2}(t-1), c_{SO_2}(t-1), c_{C_6H_5OH}(t-1), c_{C_6H_6}(t-1)) \quad (8)$$

- The values of ozone concentration and the meteorological parameters at the previous hour:

  Model D1:

$$c_{O_3}(t) = f_4(c_{O_3}(t-1), h(t-1), p(t-1), sr(t-1), temp(t-1), ws(t-1)) \quad (9)$$

Table 1. The log-likelihoods (LL), average squared errors (ASE), and log density errors (LDE).

MODEL	LL	ASE_{TRAIN}	LDE_{TRAIN}	ASE_{VAL}	LDE_{VAL}
Model A1	626.6	0.0169	-0.6202	**0.0282**	**-0.2474**
Model B1	**726.3**	**0.0117**	**-0.8069**	0.0305	-0.2314
Model C1	682.5	0.0144	-0.7021	0.0386	-0.0816
Model D1	694.0	0.0137	-0.7253	0.0297	-0.1544

In Table 1, the best obtained values of LL, ASE_{TRAIN}, LDE_{TRAIN}, ASE_{VAL}, and LDE_{VAL} are given in bold. It can be seen that the largest log-likelihood of model parameters is obtained for model B1, as well as the smallest errors associated to the training data. However, the best validation results are obtained with model A1, which is also the simplest among the four obtained models. Therefore, model A1 is considered as the best model among the four *first-order* Gaussian process models.

The hyperparameters of model A1 have the following values:

$$\Theta = [w_1, v_1, v_0] = [2.6619,\ 1.4885,\ 0.1304] \quad (10)$$

The response of model A1 to validation data is shown in Fig. 2.

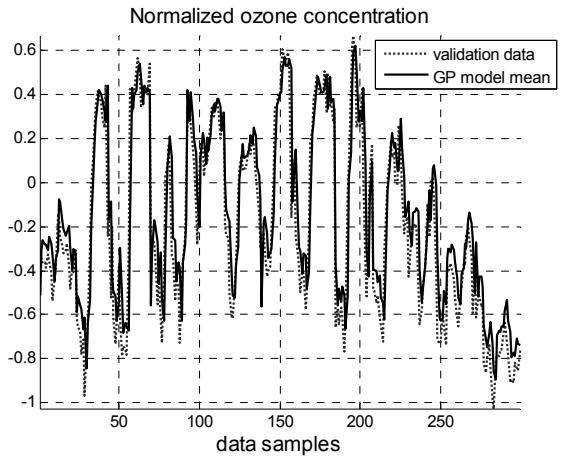

Fig. 2. The predicted mean value by model A1.

3.4 High-order Gaussian process models

Four *second-order* and one *third-order* Gaussian process models for prediction of ozone concentration are identified and verified based on the available measurement data. The log-likelihoods (LL) of the obtained models, and the average squared errors (ASE) and the log density errors (LDE) computed for the training data are given in Table 2. The *training* and the *validation* data are formed in the same way as for the *first-order* models.

The identified *high-order* Gaussian process models have the following input parameters:

- The values of ozone concentration at previous hours:

  Model A2:

  $$c_{O_3}(t) = f_5(c_{O_3}(t-1), c_{O_3}(t-2)) \quad (11)$$

  Model A3:

  $$c_{O_3}(t) = f_6(c_{O_3}(t-1), c_{O_3}(t-2), c_{O_3}(t-3)) \quad (12)$$

- The values of ozone concentration, the concentrations of the air pollutants, and the meteorological parameters at two previous hours:

  Model B2:

  $$\begin{aligned} c_{O_3}(t) = f_7(&c_{O_3}(t-1), c_{NO_2}(t-1), c_{SO_2}(t-1), \\ &c_{C_6H_5OH}(t-1), c_{C_6H_6}(t-1), h(t-1), p(t-1), \\ &sr(t-1), temp(t-1), ws(t-1), \\ &c_{O_3}(t-2), c_{NO_2}(t-2), c_{SO_2}(t-2), \\ &c_{C_6H_5OH}(t-2), c_{C_6H_6}(t-2), h(t-2), p(t-2), \\ &sr(t-2), temp(t-2), ws(t-2)) \end{aligned} \quad (13)$$

- The values of ozone concentration and the concentrations of the air pollutants at two previous hours:

  Model C2:

  $$\begin{aligned} c_{O_3}(t) = f_8(&c_{O_3}(t-1), c_{NO_2}(t-1), c_{SO_2}(t-1), \\ &c_{C_6H_5OH}(t-1), c_{C_6H_6}(t-1), \\ &c_{O_3}(t-2), c_{NO_2}(t-2), c_{SO_2}(t-2), \\ &c_{C_6H_5OH}(t-2), c_{C_6H_6}(t-2)) \end{aligned} \quad (14)$$

- The values of ozone concentration and the meteorological parameters at two previous hours:

  Model D2:

  $$\begin{aligned} c_{O_3}(t) = f_9(&c_{O_3}(t-1), h(t-1), p(t-1), sr(t-1), \\ &temp(t-1), ws(t-1), \\ &c_{O_3}(t-2), h(t-2), p(t-2), sr(t-2), \\ &temp(t-2), ws(t-2)) \end{aligned} \quad (15)$$

Table 2. The log-likelihoods (LL), average squared errors (ASE), and log density errors (LDE).

MODEL	LL	ASE_{TRAIN}	LDE_{TRAIN}	ASE_{VAL}	LDE_{VAL}
Model A2	622.3	0.0159	-0.6519	0.0265	-0.2802
Model A3	631.8	0.0149	-0.6849	0.0268	-0.2521
Model B2	**740.2**	**0.0100**	**-0.8864**	**0.0230**	-0.3447
Model C2	677.6	0.0128	-0.7629	0.0317	-0.1092
Model D2	711.7	0.0121	-0.7914	0.0231	**-0.3531**

It can be noticed from Tables 1 and 2 that for the same model type, the *high-order* models are more accurate than the *first-order* models (model A2 is more accurate than model A1, model B2 is more accurate than model B1 etc.).

It can be seen from Table 2 that the largest log-likelihood of model parameters is obtained for model B2, as well as the smallest errors associated to the training data and the smallest ASE_{VAL} error. However, the smallest LDE_{VAL} error is obtained with model D2, which is much simpler than model B2. Therefore, model D2 is considered as the best Gaussian process model for ozone concentration prediction.

The hyperparameters of model D2 have the following values:

$$\begin{aligned} \Theta &= [w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}, w_{12}, v_1, v_0] \\ &= [2.8933,\ 2162.1591,\ 30.9933,\ 8.0685,\ 2.6736, \\ &\quad 5.3512,\ 1.2902,\ 2249.5529,\ 22.0605,\ 7.0903, \\ &\quad 3.3699,\ 2065.0189,\ 1.4668,\ 0.1116] \end{aligned} \quad (16)$$

The response of model D2 to validation data is shown in Fig. 3.

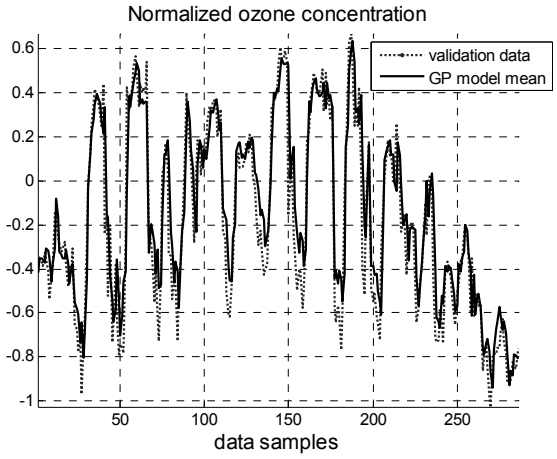

Fig. 3. The predicted mean value by model D2.

In Figures 4 to 10, the mean value and 95% confidence interval of ozone concentration predicted with model D2 are shown for some days of year 2008.

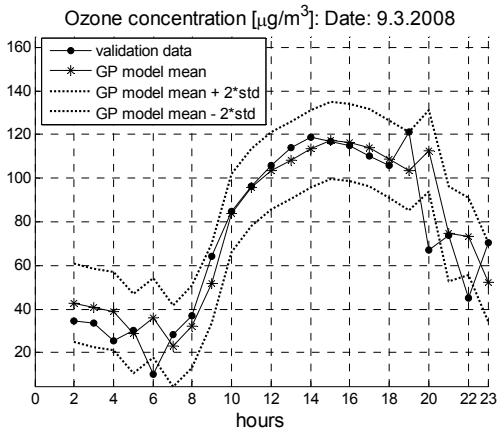

Fig. 4. The predicted mean value and 95% confidence interval of ozone concentration for 9-th March, 2008.

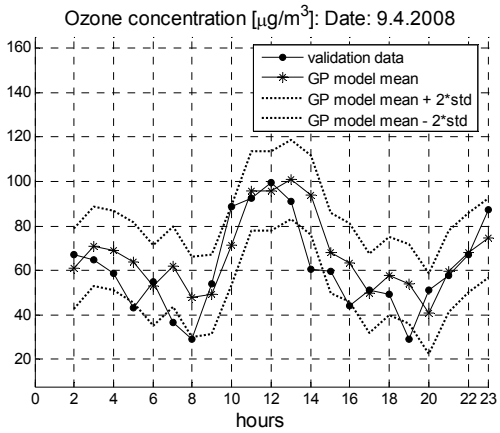

Fig. 5. The predicted mean value and 95% confidence interval of ozone concentration for 9-th April, 2008.

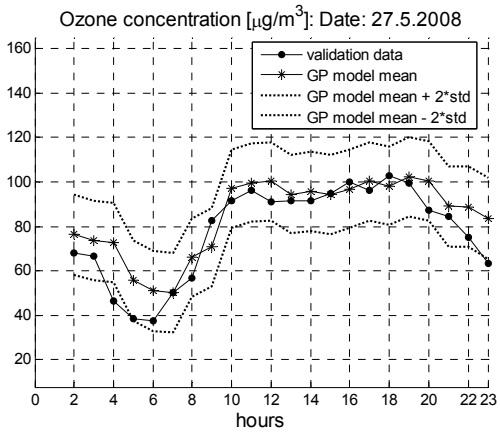

Fig. 6. The predicted mean value and 95% confidence interval of ozone concentration for 27-th May, 2008.

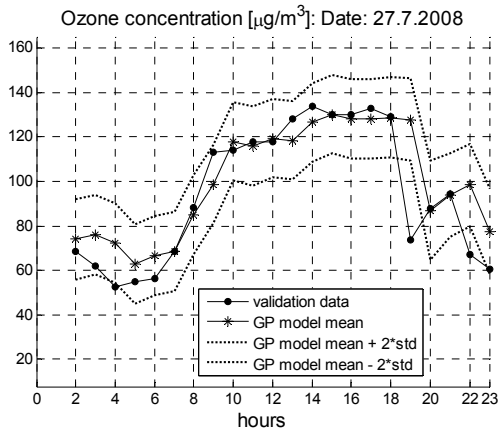

Fig. 7. The predicted mean value and 95% confidence interval of ozone concentration for 27-th July, 2008.

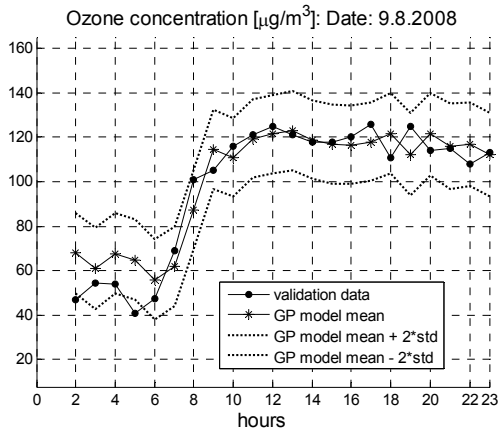

Fig. 8. The predicted mean value and 95% confidence interval of ozone concentration for 9-th August, 2008.

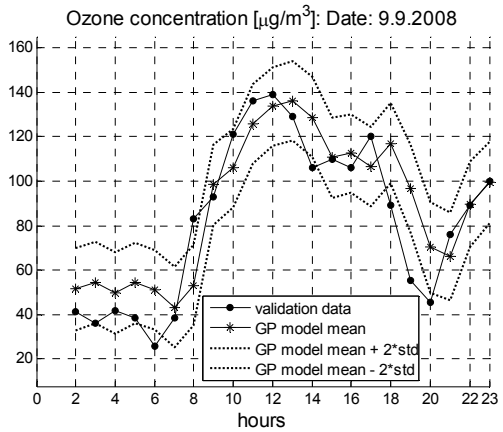

Fig. 9. The predicted mean value and 95% confidence interval of ozone concentration for 9-th September, 2008.

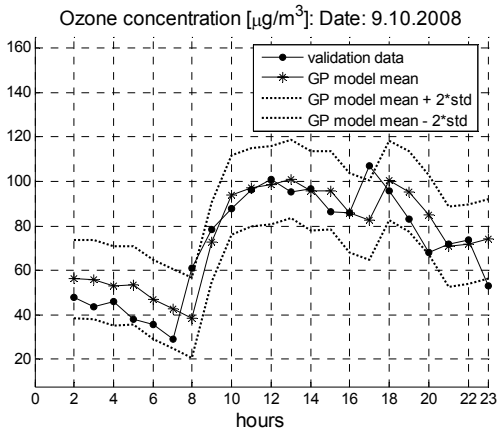

Fig. 10. The predicted mean value and 95% confidence interval of ozone concentration for 9-th October, 2008.

4 Conclusions

In this paper, *first-* and *high-order* Gaussian process models for 1-hour ahead prediction of ozone concentration in the air of Bourgas, Bulgaria are identified and compared. The models are of different types depending on the types of the input parameters. Results show that for the same model type, the *high-order* models are more accurate than the *first-order* models. The best model is the *second-order* Gaussian process model, whose input parameters are the values of the ozone concentration and the meteorological parameters at two previous hours.

5 References

[1] Directive 2002/3/EC of the European Parliament and of the Council of 12 February 2002 relating to ozone in ambient air.

[2] D. Nedialkov, M. Angelova, G. Baldjiev, A. Krastev, and H. Hristova. Ozone concentrations in the air – standards and daily cycle. *Proceedings of 19-th International Symposium on Bioprocess Systems*, Sofia, 2006.

[3] D. Nedialkov, A. Grancharova, H. Hristova, and A. Krastev. Linear regression models for prediction of ozone concentration in the air of Bourgas. *Proceedings of International Conference on Automatics and Informatics*, October, 2009, Sofia, Bulgaria, pp.IV-21-IV-24.

[4] D. Nedialkov, M. Angelova, A. Krastev, and H. Hristova. Prognostication of ozone concentration in the air. *Proceedings of 20-th International Symposium on Bioprocess Systems*, Sofia, November 6-7, 2007, pp. II.1-II.8.

[5] M. Božnar, P. Mlakar, and B. Grašič. Neural networks based ozone forecasting. In: *Modelling for Regulatory Purposes*, Garmisch-Partenkirchen, Germany, 1-4 June 2004, Proceedings, vol. 2. Karlsruhe: Forschungszentrum, 2004, pp. 356-360.

[6] B. Grašič, P. Mlakar, and M. Božnar. Ozone prediction based on neural networks and Gaussian processes. *Nuovo cimento Soc. ital. fis., C Geophys. space phys.*, vol. 29, No. 6, pp. 651-661, 2006.

[7] A. Grancharova, D. Nedialkov, J. Kocijan, H. Hristova, and A. Krastev. Application of Gaussian processes to the prediction of ozone concentration in the air of Bourgas. *Proceedings of International Conference on Automatics and Informatics*, October, 2009, Sofia, Bulgaria, pp.IV-17-IV-20.

[8] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*, MIT Press, Cambridge, MA, London, 2006.

[9] K. Ažman and J. Kocijan. Application of Gaussian processes for black-box modelling of biosystems. *ISA Transactions*, vol. 46, No. 4, pp. 443-457, 2007.

[10] A. Girard, C. E. Rasmussen, J. Quinonero Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs & application to multiple-step ahead time series forecasting. *Proceedings of NIPS 15*, Vancouver, Canada, MIT Press, 2003.

[11] A. Grancharova, J. Kocijan, and T. A. Johansen. Explicit stochastic predictive control of combustion plants based on Gaussian process models. *Automatica*, vol. 44, No. 6, pp. 1621-1631, 2008.

[12] J. Kocijan, A. Girard, B. Banko, and R. Murray-Smith. Dynamic systems identification with Gaussian processes. *Mathematical and Computer Modelling of Dynamic Systems*, vol. 11, No. 4, pp. 411-424, 2005.

[13] B. Likar and J. Kocijan. Predictive control of a gas-liquid separation plant based on a Gaussian process model. *Computers & Chemical Engineering*, vol. 31, pp. 142-152, 2007.

[14] J. Kocijan. Gaussian process models for systems identification. *Proceedings of the 9-th International PhD Workshop on Systems and Control: young generation viewpoint*, Izola, Simonov zaliv, 2008, 8 pages.

[15] C. E. Rasmussen. *Evaluation of Gaussian processes and other methods for non-linear regression*, Ph.D. Dissertation, Graduate Department of Computer Science, University of Toronto, Toronto, 1996.