Stochastic Environmental Research and Risk Assessment manuscript No.
(will be inserted by the editor)

Selection of the data time interval for the prediction of maximum ozone concentrations

Juš Kocijan · Dejan Gradišar · Martin Stepančič · Marija Zlata Božnar · Boštjan Grašič · Primož Mlakar

the date of receipt and acceptance should be inserted later

Abstract This paper highlights the problem of step-length selection for the one-step-ahead prediction of ozone called the data time interval. This is done using a case study-based comparison of two approaches for predicting the maximum daily values of tropospheric ozone. The first approach is the one-day-ahead prediction and the second is the prediction of the maximum values based on a multi-step-ahead iteration of one-hour predictions. Gaussian process modelling is utilised for this comparison. In particular, evolving Gaussian-process models are used that update on-line with the incoming measurement data. These sorts of models have been successfully used in the past for the prediction of ozone pollution. This paper contributes an assessment of the way that the maximum ozone values are predicted. A comparison of the daily maximum ozone values forecasted by a model based on one-day-ahead predictions with those obtained by iterated one-hour-ahead predictions of the ozone with predictions at predetermined hours of the day is given. The forecast results are in favour of the on-line model based on hourly predictions when approaching closer to the real maximum values of ozone, and in favour of the daily predictions when they are made on a daily basis.

J. Kocijan
Jožef Stefan Institute,
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
Tel.: +386-1-4773661
Fax.: +386-1-4773994
E-mail: jus.kocijan@ijs.si
`http://orcid.org/0000-0002-1221-946X`
and
University of Nova Gorica,
Vipavska 13, SI-5000 Nova Gorica, Slovenia

D. Gradišar, M. Stepančič
Jožef Stefan Institute,
Jamova cesta 39, SI-1000 Ljubljana, Slovenia

M. Z. Božnar, B. Grašič, P. Mlakar
MEIS d.o.o.,
Mali Vrh pri Šmarju 78, SI-1293 Šmarje-Sap, Slovenia

Keywords Air pollution · Ozone · Prediction of ozone concentration · Statistical modelling · Gaussian-process model · Evolving model

1 Introduction

Tropospheric ozone concentration forecasting is recognised by international regulations as one of the important factors in determining air quality. As such, considerable efforts are being employed to develop models that provide forecasts (hereafter predictions) of this pollutant that can be used to provide warnings in addition to other purposes.

Mathematical models that constitute the basis for computer models can be roughly divided into deterministic (i.e., physical) models and statistical (i.e., empirical) models, and can also be a combination of the two.

Deterministic or physical models (Zhang et al., 2012a,b) contain relations among the physical and chemical meteorological and air-pollution variables and as such provide an insight into pollutant-formation processes (Zhang et al., 2012a). These types of air-quality models provide prognostic time- and spatially-resolved concentrations for various typical and atypical scenarios. An overview of such models for the prediction of ozone concentration is given in (Im et al., 2015).

Statistical or empirical models, on the other hand, are developed based on historical measurements of meteorological and air-pollution variables and utilise mainly statistical, i.e., regression, methods to develop prediction models based on the historical record of the included variables. These models provide predictions of higher accuracy and with better computational efficiency in relation to acquired measurements than deterministic models, provided that the empirical models are developed correctly and sufficiently well (Zhang et al., 2012a). Unfortunately, the physical and chemical processes determining the meteorology and air quality are not transparent in empirical models.

Nevertheless, the dilemma of whether deterministic models are superior over empirical models is artificial. These two types of models should be complementary in use. Deterministic air-quality and meteorological models cover large areas of geographical, 3-dimensional space, including locations that are not monitored (Žabkar et al., 2015), and for long prediction horizons. Empirical air-quality and meteorological models, on the other hand, are advantageous in predicting locally, and as such are more suitable for modelling the concentrations of air pollutants over terrain with complex topography and consequently complex climatological conditions.

This paper deals with the assessment of empirical models for modelling maximum ozone concentrations. When empirical models are chosen for development, there are several modelling decisions to be made regarding the methodology. Besides the regression method and, consequently, the content of the regression vector, there are other design parameters to be selected including the interval of prediction, the interval of sampling data from databases, the interval of averaging measurements, and the sources of the data. Regulatory frameworks provide directions for some of these design parameters, e.g., the interval of the measurements' averaging, but most of them are decided by modelling experts.

There are various modelling methods that can be used for empirical models. An abundance of publications exist that describe the use of empirical models for the modelling and prediction of ozone concentrations, and air quality in general, using the different methods and applied in various regions, e.g., (Gong and Ordieres-Meré, 2016; Taylan, 2017; Ding et al., 2016), to name a few of the most current.

This paper continues the investigation of finding the optimal method for predicting the daily maximum ozone concentrations as described in (Kocijan et al., 2016), but with a different focus. While in (Kocijan et al., 2016) the selections of the modelling algorithm and the regressors were emphasised, here the selection of step length, which is the interval for the one-step-ahead prediction, is highlighted. We refer to this step length as the data time interval in the subsequent text.

Often the applied data time intervals coincide with daily predictions, but ozone cannot be predicted well on a daily basis because its dynamics is faster. If approached as a dynamic system, the ozone prediction makes it necessary to refine the data time interval for the prediction. This is also suggested when considering daily values. Various published investigations differ in the data time interval used for the one-step-ahead prediction due to the availability of the data sampled at a certain sampling rate. Besides one-day-ahead predictions, some investigations used 15-minutes-ahead predictions (Feng et al., 2011), others used one-hour-ahead predictions, e.g., (Petelin et al., 2013; Bruno and Paci, 2014; Duenas et al., 2005; Al-Alawi et al., 2008), and another used three-hour-ahead predictions (Faris et al., 2014).

The selection of the data time interval in prediction modelling is a known problem and has been investigated in various domains, e.g., economic sciences, e.g., (Andrawis et al., 2011; Casals et al., 2009; Kourentzes et al., 2014), medicine, e.g., (Sud et al., 2002), and hydrology, e.g., (Liu and Han, 2013). However, this issue has not yet been highlighted for air-quality modelling (Alyousifi et al., 2017) or ozone-concentration modelling and prediction (Conde-Amboage et al., 2017).

The problem considered in this paper is how one-hour-ahead predictions can be used for predicting the daily maximum value of the 1-hour-averaged ozone concentrations. We require that the obtained predictions contain information about their uncertainty. Because one-hour-ahead predictions provide results too late to be utilised for alarm purposes, they have to be iterated to obtain multi-hour-ahead predictions that can provide results early enough. The results based on one-hour predictions should be compared to the results of one-day-ahead predictions and the relations between them should be assessed.

The paper is structured as follows. The problem is described in the next section. The modelling method used is introduced in Section 2. Section 3 discusses experiments used to assess the described approaches and the results are presented in Section 4. The conclusions are gathered at the end of the paper.

2 Methods

2.1 Gaussian Process Models

Gaussian Process (GP) models are probabilistic, non-parametric models based on the principles of Bayesian probability. GPs can be considered as kernel methods with a Bayesian interpretation (Rasmussen and Williams, 2006; Gregorčič and Lightbody, 2008). 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; Kocijan, 2016; Shi and Choi, 2011).

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 $\mathbf{K}$ being functions of the inputs $\mathbf{x}$. Usually, the mean function is defined as $\mathbf{0}$, while the covariance function or kernel

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

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 (MacKay, 1998) and a constant covariance function assuming white noise:

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

where $\boldsymbol{\Lambda}^{-1}$ is a diagonal matrix $\boldsymbol{\Lambda}^{-1} = \mathrm{diag}([l_1^{-2}, \ldots, l_D^{-2}])$ 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 $\boldsymbol{\theta} = [l_1^{-2}, \ldots, l_D^{-2}, \sigma_f^2, \sigma_n^2]^T$. 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 $\boldsymbol{\theta}$ has to be taken into account. This leads to a computationally 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 a lack of data or noisy data, by indicating a wider confidence interval around the predicted mean.

A drawback of GP modelling is the computational load that comes with a large training dataset. This load increases with the third power of the amount of input data due to the calculation of the inverse of the covariance matrix. To overcome this drawback, various sparse-approximation methods have been suggested. A common property of all these sparse-approximation methods is that they try to retain the bulk of the information contained in the full training dataset, but reduce the size of the covariance matrix. They achieve this by using the subset of the training data that contains most of the information. This subset of the training data is called the active set. For more details see (Quinonero-Candela et al., 2007). The subsequently described method adapts the active set sequentially and in parallel adjusts the hyperparameter values.

The properties of GP models have made them attractive for modelling case studies in various domains like chemical engineering (Chan et al., 2013) and process control (Likar and Kocijan, 2007), biomedical engineering (Faul et al., 2007), biological systems and medicine (Bukhari and Hong, 2014), power systems (Leith et al., 2004) and engineering (Leithead et al., 2005), motion recognition (Kang and Park, 2015), etc., to list just a few, and have also found their utility in the field of environmental research (Grašič et al., 2006; Kocijan et al., 2016; Schliep et al., 2017).

2.2 Evolving Gaussian process modelling

The Evolving GP (EGP) model is a self-developing system that sequentially adapts elements of the GP model with incoming data, including the hyperparameter values.

This enables the fast and efficient adaptation of the GP model to changes. The EGP concept is described in, e.g., (Petelin et al., 2013) and (Kocijan, 2016). This concept considers the adaptation of four main elements of the GP model: the active set, the hyperparameter values, the covariance function, and the regressors. To simplify the concept we decided, like with (Petelin et al., 2013), to use the fixed covariance function SE with ARD as we assume the smoothness and stationarity of the stochastic process used for the modelling of the nonlinear system input/output mapping. The ARD functionality is able to find influential regressors. With the optimization of the hyperparameter values, the uninfluential regressors have a smaller weight and, as a consequence, have a smaller influence on the result. Therefore, all the available regressors can be used and, consequently, only the active set and the hyperparameter values are adapted sequentially.

The proposed method consists of roughly three main steps to adapt the GP model sequentially. In the first step, the new data is processed in the sense of including the incoming data in the active set $\mathcal{X}$. In the following step, the hyperparameter values $\boldsymbol{\theta}$ are optimized, while in the last step the covariance matrix $\mathbf{K}$ and its inversion $\mathbf{K}^{-1}$ are updated in accordance with the changes from the first two steps. More details can be found in (Petelin et al., 2013) and (Kocijan, 2016).

3 Description of the experiments

The investigation is made empirically as a comparison between one-day ahead predictions of the daily maximum ozone concentrations based on historical data from previous days and one-hour ahead predictions of the daily maximum 1-hour-averaged ozone concentrations based on historical data from previous hours with iterative predictions for several steps ahead.

The one-step-ahead prediction is

$$\hat{y}(k+1) = \hat{f}(\mathbf{x}(k)), \tag{6}$$

where k represents consecutive time instants and $\hat{f}$ is the nonlinear mapping from the regression vector $\mathbf{x}$ to the prediction $\hat{y}$. The iterative predictions are implemented as follows:

Step	Prediction	Regression vector
1	$\hat{y}(k+1)$	$\mathbf{x}(k) = [y(k), y(k-1), \ldots, \mathbf{u}(k), \mathbf{u}(k-1), \ldots]^{\mathrm{T}}$
2	$\hat{y}(k+2)$	$\mathbf{x}(k+1) = [\hat{y}(k+1), y(k), \ldots, \mathbf{u}(k+1), \mathbf{u}(k), \ldots]^{\mathrm{T}}$
3	$\hat{y}(k+3)$	$\mathbf{x}(k+2) = [\hat{y}(k+2), \hat{y}(k+1), \ldots, \mathbf{u}(k+2), \mathbf{u}(k+1), \ldots]^{\mathrm{T}}$
$\vdots$	$\vdots$	$\vdots$

where y represents measurements of the output, $\mathbf{u}$ is the vector of measurements of the input signals and $\hat{y}$ represents the predictions.

Remark In practice, the vector of the input signals $\mathbf{u}$ for the future steps cannot be measured and is usually predicted by some model predictions. The regression vector in step i should, in such cases, be written as follows:

$$\mathbf{x}(k+i) = [\hat{y}(k+i-1), \hat{y}(k+i-2), \ldots, \hat{\mathbf{u}}(k+i-1), \hat{\mathbf{u}}(k+i-2), \ldots]^{\mathrm{T}}.$$

However, to avoid the prediction uncertainty for other variables as well as to enable a fair comparison strictly following the assessment methodology as used in the investigation described in (Kocijan et al., 2016), where the focus is on ozone only, we used the measurements instead of the predictions for all those variables that are not ozone. The use of numerical predictions from other models as regressors is certainly a feasible option, but it is beyond the scope of this paper.

The one-hour-ahead predictions are made at 6:00, 12:00 and 15:00 hours with iterative predictions forward to find daily maximum-value predictions. The procedure and results of the modelling and predictions for the daily maximum ozone concentrations one-day ahead are taken from the investigation described in (Kocijan et al., 2016), where the details can be found.

The empirical investigation of the two prediction concepts is pursued on a set of data acquired from a measurement station in the city of Koper, Slovenia. Koper

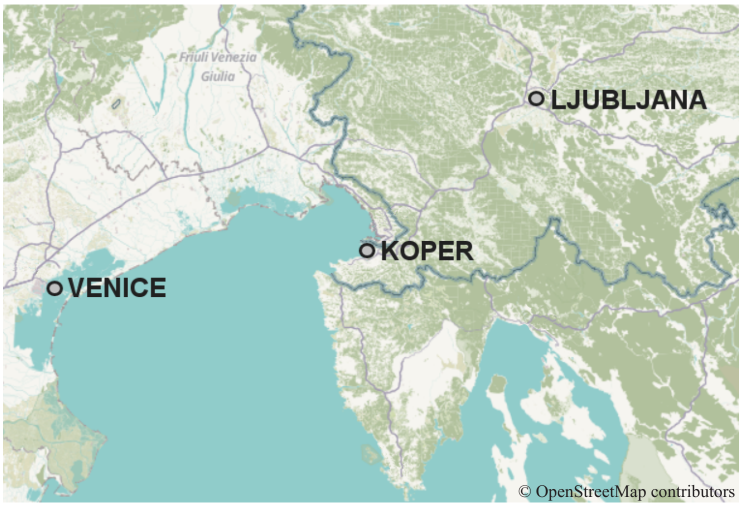

Fig. 1 Geographical location of the treated location in Slovenia.

is an industrial and port town on the Adriatic coast with a Mediterranean climate. Its geographical location is depicted in Figure 1.

The meteorological and air-quality variables at these locations are measured every half an hour and are stored in a database. The measured data were acquired for all the available variables, as listed in Table 1, for each location for a period of 3 years (from the beginning of 2012 to the end of 2014).

Table 1 Available variables' measurements at the city of Koper

Variable	Description
$O3$	Ozone concentration
$GlSolRad$	Global solar radiation
$AirTemp$	Air temperature
$RelHum$	Relative humidity
$WindSpd$	Wind speed
$WindDir$	Wind direction
NOx	Nitrogen oxides concentration
$NO2$	Nitrogen dioxide concentration
$Dust$	Solid particles
$Precip$	Precipitation
$DifSolRad$	Diffuse solar radiation
$Pressure$	Atmospheric pressure

Based on the collected half-hour measurements for each variable, their 1-h averages are calculated according to the regulatory framework (EU-Commission, 2008). Because the ozone concentration depends on the past and present values of the variables, the present values are also added. In practice, as already mentioned, this is done with model, e.g. physical- or empirical-model, predictions.

The regression vector for the data-based model is expected to contain the values of variables as well as their lagged values, because the dynamic model is developed. However, the number of all the regressors is very large, so it is only necessary to select the regressors that add the most information to the prediction.

The procedure for the selection of regressors was using a conventional off-line GP regression model for determining the best set of regressors with an exhaustive search. The GP model was based on a squared exponential covariance function with the ARD property and with a noise covariance function as described by Equation (3). The GP model was trained on a subset of data that was recognized as the most significant, i.e., periods of the year with high maximum values of ozone concentrations. The available data were divided into 11 subsets. A 10-fold cross-validation was used in the training procedure and the remaining set was used for testing the predictions. Performance measures described in the Appendix were used for the evaluation. The final set of regressors is shown in Table 2 together with the set of regressors for the daily maximum one-day ahead model from (Kocijan et al., 2016). It is clear from Table 2 that even though the regressors for the one-day-ahead and one-hour-ahead models are different, the variables, from which the regressors are sampled, are the same. This confirms the significance of particular variables for daily maximum predictions at the investigated location.

Table 2 Regressors for the final models. k denotes consecutive time instants, where $k+1$ means the present-time values, i.e., at the prediction time, and k the recent values, i.e., one day or one hour backwards. Instead of predictions, measurements are used for the present-time values, as explained earlier in the text.

	One-day-ahead	One-hour-ahead
1	$O3(k)$	$O3(k)$
2	$RelHum(k+1)$	$RelHum(k+1)$
3	$AirTemp(k+1)$	$AirTemp(k+1)$
4	$NOx(k+1)$	$NOx(k+1)$
5	$Pressure(k+1)$	$Pressure(k+1)$
6	$GlSolRad(k+1)$	$GlSolRad(k)$
7	$AirTemp(k)$	$NOx(k)$
8	$GlSolRad(k)$	$Pressure(k)$
9	$Pressure(k)$	/

When the regressors were selected, they were used for the modelling and the one-hour-ahead prediction of the ozone based on the EGP. First, 60 days are needed to initialise the GP model. This period was selected empirically. After that, the evolving GP model is applied. Again, the SE covariance function with the ARD property, white noise covariance function, and a constant mean value are used. The number of optimisation iterations in each EGP step is, in our case, limited to 70. And the maximum size of the active set is determined as 90 data points. The selected values were obtained empirically based on a trade-off between the computational time and the quality of the predictions. The information gain is defined by the maximum Euclid distance of the i-th element to any other. Exponential forgetting, with a forgetting factor of 0.9995, was also determined empirically.

As mentioned before in this section, the one-hour-ahead predictions are made separately at 6:00, 12:00 and 15:00. The predictions after the hour of the first prediction are a combination of iteratively used predictions of the O_3 variable from the previous step and measurements for the other regressors.

When we want to obtain a more realistic picture of the dynamic model prediction, we have to *take into account the uncertainty* of uncertain input data, which in our case is the EGP-model prediction of the O_3 variable in the previous step. The idea of using the uncertainty information is as follows.

We wish to make a prediction at $\mathbf{x}^*$, where the regression vector $\mathbf{x}^*$ contains uncertain input values. Within a Gaussian approximation, the input values can be described by the normal distribution $\mathbf{x}^* \sim \mathcal{N}(\mu_{\mathbf{x}^*}, \mathbf{\Sigma}_{\mathbf{x}^*})$, where $\mu_{\mathbf{x}^*}$ and $\mathbf{\Sigma}_{\mathbf{x}^*}$ are the vector and the matrix of the input mean values and variances, respectively. To obtain a prediction, we need to integrate the predictive distribution $p(y^*|\mathbf{x}^*, \mathcal{D})$, where $\mathcal{D} = \{(\mathbf{x}_i, y_i)|i = 1, \dots, N\} = \{(\mathbf{X}, \mathbf{y})\}$, over the input data distribution, that is

$$p(y^*|\mu_{\mathbf{x}^*}, \mathbf{\Sigma}_{\mathbf{x}^*}, \mathcal{D}) = \int p(y^*|\mathbf{x}^*, \mathcal{D}) p(\mathbf{x}^*) d\mathbf{x}^*, \tag{7}$$

where

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x}^*)}} \exp\left[-\frac{(y^* - \mu(\mathbf{x}^*))^2}{\sigma^2(\mathbf{x}^*)}\right]. \tag{8}$$

Since $p(y^*|\mathbf{x}^*, \mathcal{D})$ is in general a nonlinear function of $\mathbf{x}^*$, the new predictive distribution $p(y^*|(\mu_{y^*}, \mathbf{\Sigma}_{y^*}, \mathcal{D}))$ is not Gaussian and this integral cannot be solved without using an approximation. In other words, when the Gaussian distribution is propagated through a nonlinear model it is not a Gaussian distribution at the output of the model.

Approximations can be roughly divided into numerical methods, for example, Monte Carlo methods and analytical methods. Here, only one of the methods, i.e., *analytical approximation with exact matching of statistical moments*, which is described in (Kocijan, 2016) and references therein, is used. The idea of the method is that the integral of Equation (7) is approximated so that the exact prediction, which is a Gaussian mixture, is approximated with the normal distribution.

4 Results

This section provides results that show whether and approximately when it is reasonable to predict the daily maximum concentrations with iterated one-hour-ahead predictions and with one-day-ahead predictions. What we expected was that from a certain point in the day, the one-hour-ahead predictions with the following iterative predictions should provide more insight. Moreover, predicting with a one-hour data time interval should also provide the estimate of the maximum-concentration time occurrence.

Table 3 provides results of the performance measures for the test data. The description of the performance measures is given in the Appendix.

The content of the table confirms the expectation that the iterated one-hour-ahead predictions deteriorate the further away we make iterations from the time of the first prediction. It seems reasonable in the sense of the predictions' accuracy that one-hour-ahead predictions are used no earlier than a few hours before the expected event, i.e., in the afternoon hours on each day where the maximum is sought.

The scatter plots in Figs. 2 and 3 and the time responses in Figs. 5–7 visually illustrate and confirm the conclusions for Table 3.

Table 3 Performance measures for predictions of the daily maximum concentrations using one-day-ahead and iterated one-hour-ahead predictions: the root-mean-square error (RMSE), the standardised mean-squared error (SMSE), Pearson's correlation coefficient (PCC), the mean fractional bias (MFB), and the factor of the modelled values within a factor of two of the observations (FAC2)

Performance measure	1-day-ahead	1-hour-ahead at 6.00	1-hour-ahead at 12.00	1-hour-ahead at 15.00
RMSE	14.69	27.12	17.36	11.40
SMSE	0.19	0.62	0.27	0.14
MSLL	-0.79	-0.005	-0.56	-0.92
PCC	0.90	0.72	0.87	0.94
MFB	0.025	-0.10	-0.04	-0.03
FAC2	0.98	0.99	0.997	0.998

It is also important to mention the time deviations of the predicted maximums that can be observed from the histograms in Fig. 4. The mean value of the deviations between the predicted and measured values of the daily maximum concentrations are slightly on the negative side. Consequently, if these data are used for alarms, the alarms would be slightly premature.

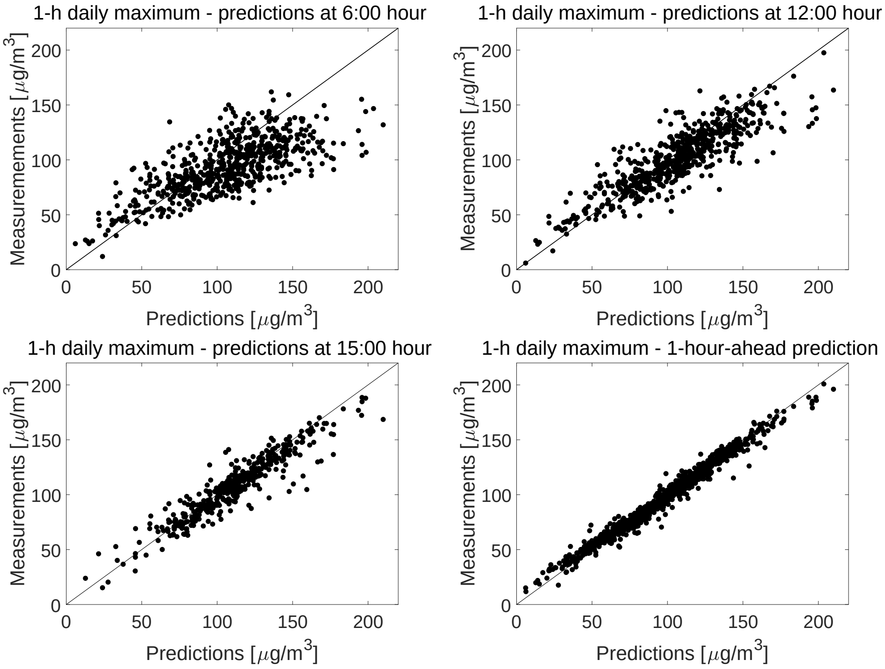

Fig. 2 Comparison of daily maximum values for iterated multi-hour-ahead predictions and one-hour-ahead predictions (bottom right figure) versus observation values

One of the advantages of using GP models is that they provide a distribution of predictions at the model output. This information can be interpreted as a confidence in the predictions. It is clear from Figs. 5–7 that the interval determined

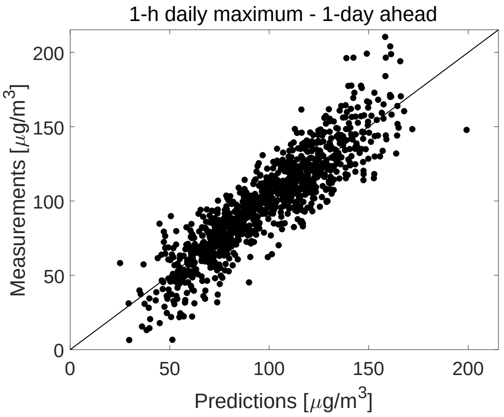

Fig. 3 One-day-ahead predicted values versus observation values for daily maximum ozone concentrations for Koper.

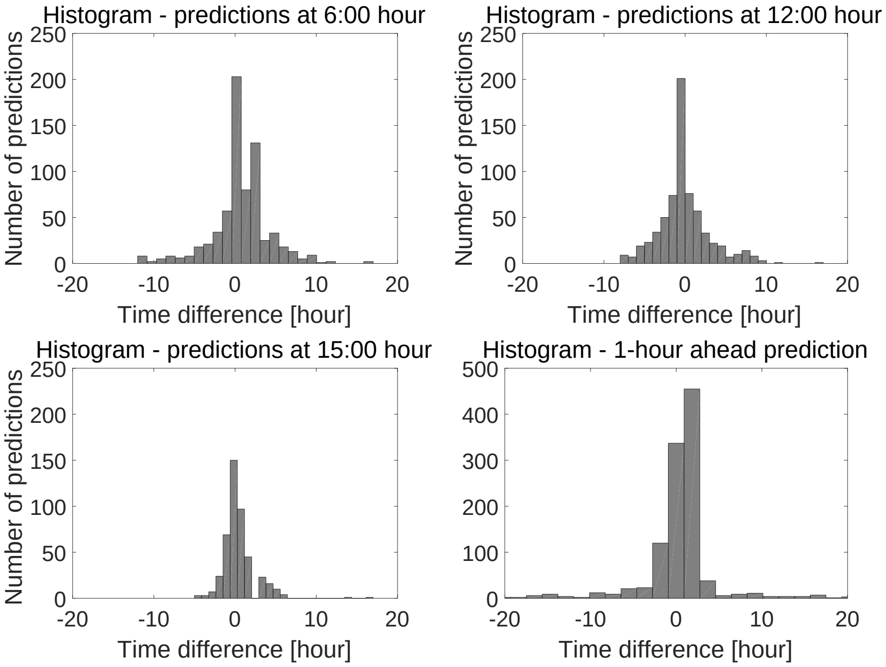

Fig. 4 Comparison of histograms of the difference between the predicted and measured time of maximum ozone values

by 2 × σ that corresponds to the 95 % confidence interval is larger if the predictions have been made earlier in the day. The measured daily maximum values are, however, mostly still within the predicted interval of uncertainty.

How the predictions deteriorate the further away we iterate from the hours of 6.00, 12.00 and 15.00 is shown in Fig. 8.

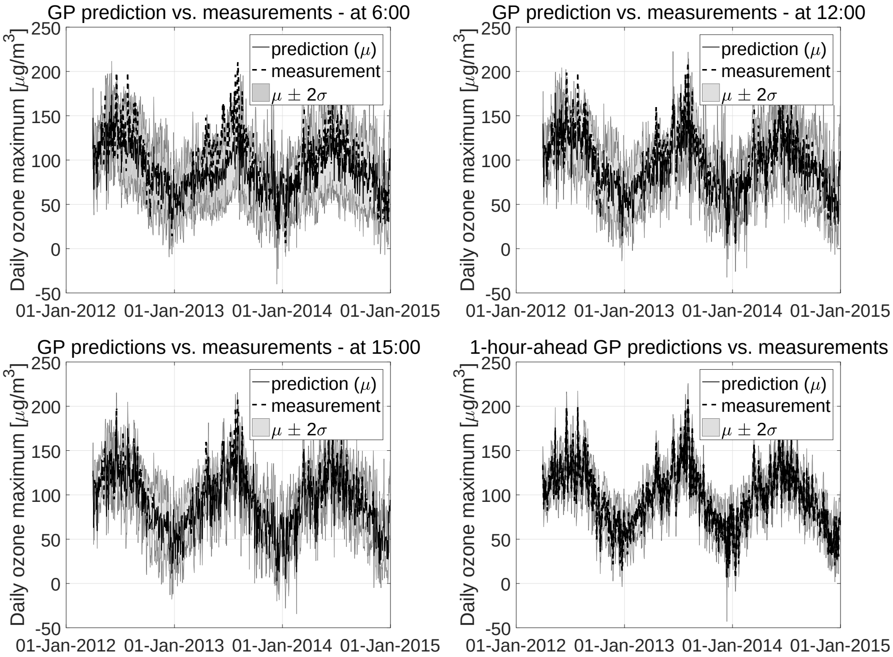

Fig. 5 Comparison of time-series responses for 1-h daily maximum ozone concentrations

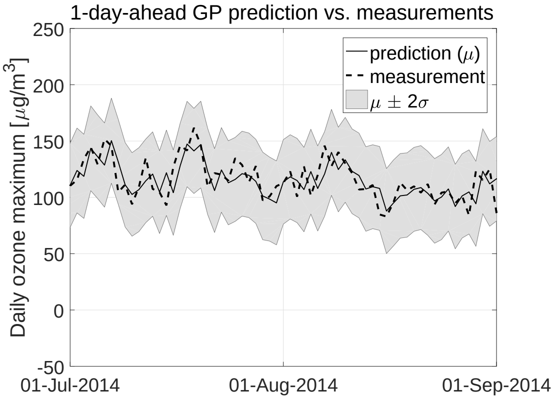

Fig. 6 Zoomed time-series responses for 1-h daily maximum ozone concentrations for Koper

Computations were pursued on personal computer (Central Processing Unit: Intel i7-2600K with 3.4 GHz, RAM memory: 8 GB) with Windows 7 operating systems and Matlab computation software. Calculations of model adaptation and the one-step-ahead prediction for the Evolving Gaussian Process in every time sample of its operation took 0.72 seconds or less, which enables calculations in real time when necessary.

The obtained results are for one location and one modelling methodology only, so the results cannot be considered general. It can be concluded for the case study

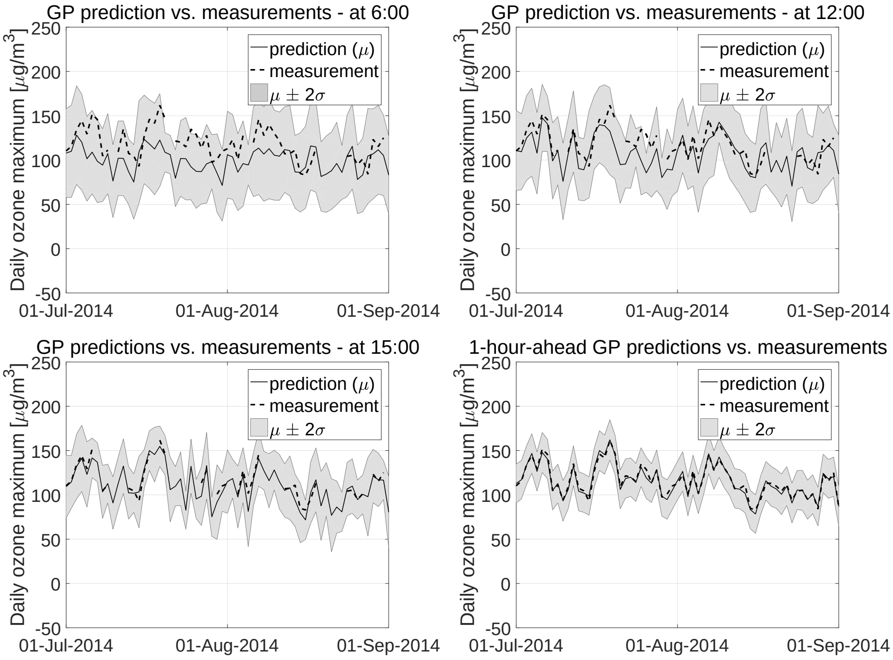

Fig. 7 Comparison of zoomed time-series responses of daily maximum values for iterated multi-hour-ahead predictions and one-hour-ahead predictions (bottom-right figure)

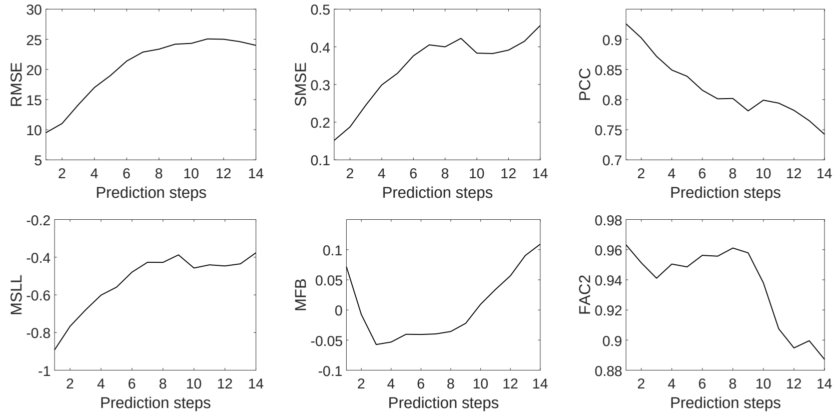

Fig. 8 Comparison of different performance measures for multi-hour-ahead predictions with iterated one-hour prediction step starting at 6.00. The analysis is made with predictions from April 29, 2012 until December 31, 2014.

though, that a combination of models, i.e., an integrated or a hybrid model, that provides the combination of the one-day-ahead prediction with iterated one-hour-ahead predictions in the afternoon hours, may ensure more accurate predictions of the maximum ozone values. This fact can be productively used for providing accurate information for the population and to increase the credibility of the alarm system.

Moreover, the demonstrated modelling method, i.e., the Gaussian process regression, provides a measure of confidence in the prediction. This can be a valuable piece of information that also increases the credibility of the provided prediction.

5 Conclusions

This paper described an empirical assessment of two approaches to predicting the daily maximum values of tropospheric ozone. The first approach was a one-day-ahead prediction and the second was a prediction of maximum values based on a multi-step-ahead iteration of one-hour predictions. The Gaussian Process models were used in modelling, in particular Evolving Gaussian Process models as their on-line version.

The investigation confirmed that iterated one-hour-ahead predictions provide better results in the last few hours before the maximum value occurs. This is a consequence of a deterioration in the predictions' accuracy with increasing iterations. We suggest using the combination of one-day-ahead prediction and iterated one-hour-ahead predictions in the afternoon hours to improve the forecasts of daily maximum ozone values.

In addition, the values of the confidence interval for the model predictions can provide additional information that can increase the credibility of the provided prediction, if used properly. Also, when predicting with the one-hour data time interval the estimate of the maximum-concentration time occurrence is available.

While the investigation confirmed the hypothesis on the use of data time intervals for the prediction, it also opened up further questions regarding the prediction interval for the daily maximum concentrations. Various studies show that empirical models used for ozone prediction provide relatively reliable predictions, mainly for short time predictions. This may be due to an insufficient number of used input variables or variables that do not encompass all the dependencies in the process of ozone formation. This disadvantage can be compensated for with the use of physical models, to a certain extent. However, there is still space for an investigation of empirical-modelling methods that will provide better long-range predictions that can be used for atmospheric modelling.

Acknowledgements The authors acknowledge the financial support from the Slovenian Research Agency (projects No. L2-5475, L2-8174 and No. P2-0001). The Slovenian Environment Agency provided part of the data.

References

Saleh M. Al-Alawi, Sabah A. Abdul-Wahab, and Charles S. Bakheit. Combining principal component regression and artificial neural-networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, 23:396–403, 2008.

Yousif Alyousifi, Nurulkamal Masseran, and Kamarulzaman Ibrahim. Modeling the stochastic dependence of air pollution index data. *Stochastic Environmental Research and Risk Assessment*, Aug 2017.

Robert R. Andrawis, Amir F. Atiya, and Hisham El-Shishiny. Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting*, 27(3):870–886, 2011.

Francesca Bruno and Lucia Paci. Spatiotemporal model for short-term predictions of air pollution data. In *The Contribution of Young Researchers to Bayesian Statistics*, pages 91–94. Springer, 2014.

W. Bukhari and S.-M. Hong. Real-time prediction of respiratory motion using a cascade structure of an extended kalman filter and support vector regression. In *Physics in Medicine and Biology*, volume 59, pages 3555–3573. IOP Publishing, 2014.

José Casals, Miguel Jerez, and Sonia Sotoca. Modelling and forecasting time series sampled at different frequencies. *Journal of Forecasting*, 28(4):316–342, 2009.

Lester Lik Teck Chan, Yi Liu, and Junghui Chen. Nonlinear system identification with selective recursive gaussian process models. *Industrial & Engineering Chemistry Research*, 52(51):18276–18286, 2013.

Mercedes Conde-Amboage, Wenceslao González-Manteiga, and César Sánchez-Sellero. Predicting trace gas concentrations using quantile regression models. *Stochastic Environmental Research and Risk Assessment*, 31(6):1359–1370, Aug 2017.

Weifu Ding, Jiangshe Zhang, and Yee Leung. Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. *Environmental Science and Pollution Research*, 23(19):19481–19494, 2016.

C. Duenas, M.C. Fernandez, S. Canete, J. Carretero, and E. Liger. Stochastic model to forecast ground-level ozone concentration at urban and rural areas. *Chemosphere*, 61:1379–1389, 2005.

EU-Commission. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Communities*, L 152:1–44, 2008. URL `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF`.

Hossam Faris, Mouhammd Alkasassbeh, and Ali Rodan. Artificial neural networks for surface ozone prediction: Models and analysis. *Polish Journal of Environmental Studies*, 23(2), 2014.

S. Faul, G. Gregorčič, G. Boylan, W. Marnane, G. Lightbody, and S. Connolly. Gaussian process modeling of EEG for the detection of neonatal seizures. *IEEE Transactions on Biomedical Engineering*, 54(12):2151–2162, 2007. ISSN 0018-9294.

Yu Feng, Wenfang Zhang, Dezhi Sun, and Liqiu Zhang. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and SVM data classification. *Atmospheric Environment*, 45:1979–1985, 2011.

Bing Gong and Joaquín Ordieres-Meré. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of hong kong. *Environmental Modelling & Software*, 84:290–303, 2016.

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.*, 29 (6):651–661, 2006.

Gregor Gregorčič and Gordon Lightbody. Nonlinear system identification: From multiple-model networks to Gaussian processes. *Engineering Applications of Artificial Intelligence*, 21(7):1035–1055, 2008.

Ulas Im, Roberto Bianconi, Efisio Solazzo, Ioannis Kioutsioukis, Alba Badia, Alessandra Balzarini, Roco Bar, Roberto Bellasio, Dominik Brunner, Charles Chemel, Gabriele Curci, Johannes Flemming, Renate Forkel, Lea Giordano, Pedro Jimnez-Guerrero, Marcus Hirtl, Alma Hodzic, Luka Honzak, Oriol Jorba, Christoph Knote, Jeroen J.P. Kuenen, Paul A. Makar, Astrid Manders-Groot, Lucy Neal, Juan L. Prez, Guido Pirovano, George Pouliot, Roberto San Jose, Nicholas Savage, Wolfram Schroder, Ranjeet S. Sokhi, Dimiter Syrakov, Alfreida Torian, Paolo Tuccella, Johannes Werhahn, Ralf Wolke, Khairunnisa Yahya, Rahela Zabkar, Yang Zhang, Junhua Zhang, Christian Hogrefe, and Stefano Galmarini. Evaluation of operational on-line-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I: Ozone. *Atmospheric Environment*, 115:404–420, 2015. ISSN 1352-2310. doi: http://dx.doi.org/10.1016/j.atmosenv.2014.09.042.

Hyuk Kang and F. C. Park. Motion optimization using Gaussian process dynamical models. *Multibody System Dynamics*, 34(4):307–325, 2015.

Juš Kocijan. *Modelling and Control of Dynamic Systems Using Gaussian Process Models*. Springer International Publishing, Cham, 2016.

Juš Kocijan, Dejan Gradišar, Marija Zlata Božnar, Boštjan Grašič, and Primož Mlakar. On-line algorithm for ground-level ozone prediction with a mobile station. *Atmospheric Environment*, 131:326–333, 2016.

Nikolaos Kourentzes, Fotios Petropoulos, and Juan R Trapero. Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, 30(2):291–302, 2014.

D. J. Leith, M. Heidl, and J. Ringwood. Gaussian process prior models for electrical load forecasting. In *Proceedings of 2004 International Conference on Probabilistic Methods Applied to Power Systems*, pages 112–117, Piscataway, NJ, 2004. IEEE, IEEE.

William E. Leithead, Yunong Zhang, and Kian Seng Neo. Wind turbine rotor acceleration: Identification using gaussian regression. In *Proceedings of 2nd International conference on informatics in control automation and robotics (ICINCO 2005)*, pages 84–91, Setbal, 2005. INSTICC, INSTICC.

B. Likar and J. Kocijan. Predictive control of a gas-liquid separation plant based on a gaussian process model. *Computers and Chemical Engineering*, 31(3):142–152, 2007. doi: 10.1016/j.compchemeng.2006.05.011.

J Liu and D Han. On selection of the optimal data time interval for real-time hydrological forecasting. *Hydrology and Earth System Sciences*, 17(9):3639–3659, 2013.

David John Cameron MacKay. Introduction to Gaussian processes. *NATO ASI Series*, 168:133–166, 1998.

Dejan Petelin, Alexandra Grancharova, and Juš Kocijan. Evolving Gaussian process models for the prediction of ozone concentration in the air. *Simulation Modelling Practice and Theory*, 33(1):68–80, 2013.

Joaquin Quinonero-Candela, Carl Edward Rasmussen, and Christopher K. I. Williams. *Large-Scale Kernel Machines*, chapter Approximation methods for Gaussian process regression, pages 203–223. Neural Information Processing. The MIT Press, Cambridge, MA, USA, September 2007.

Carl Edward Rasmussen and Chris K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Erin M. Schliep, Alan E. Gelfand, and David M. Holland. Alternating Gaussian process modulated renewal processes for modeling threshold exceedances and durations. *Stochastic Environmental Research and Risk Assessment*, Apr 2017.

Jian Qing Shi and Taeryon Choi. *Gaussian process regression analysis for functional data*. Chapman and Hall/CRC, Taylor & Francis group, Boca Raton, FL, 2011.

K. Sud, B. Singh, H. S. Kohli, V. Jha, K. L. Gupta, and V. Sakhuja. Evaluation of different sampling times for best prediction of cyclosporine area under the curve in renal transplant recipients. 34(8):3168–3170, 2002.

Osman Taylan. Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. *Atmospheric Environment*, 150: 356–365, 2017.

R. Žabkar, L. Honzak, G. Skok, R. Forkel, J. Rakovec, A. Ceglar, and N. Žagar. Evaluation of the high resolution WRF-Chem (v3.4.1) air quality forecast and its comparison with statistical ozone predictions. *Geoscientific Model Development*, 8(7):2119–2137, 2015.

Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60:632 – 655, 2012a. ISSN 1352-2310. doi: http://dx.doi.org/10.1016/j.atmosenv.2012.06.031.

Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. Real-time air quality forecasting, part ii: State of the science, current research needs, and future prospects. *Atmospheric Environment*, 60:656 – 676, 2012b. ISSN 1352-2310. doi: http://dx.doi.org/10.1016/j.atmosenv.2012.02.041.

A Performance measures

The following are performance measures used in the study.

- The root-mean-square error – RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(E(\hat{y}_i) - y_i)^2}, \tag{9}$$

  where y_i and $\hat{y}_i$ are the observation and the prediction in the i-th step, respectively, $E(\cdot)$ denotes the expectation, i.e., the mean value, of the random variable, and N is the number of used observations.

- The standardised mean-squared error – SMSE

$$\mathrm{SMSE} = \frac{1}{N}\frac{\sum_{i=1}^{N}(E(\hat{y}_i) - y_i)^2}{\sigma_y^2}, \tag{10}$$

  where σ_y^2 is the variance of the observations.

- The Pearson's correlation coefficient – PCC:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(E(\hat{y}_i) - E(\hat{\mathbf{y}}))(y_i - E(\mathbf{y}))}{N\sigma_y\sigma_{\hat{y}}}, \tag{11}$$

  where $E(\hat{\mathbf{y}})$ is the expectation, i.e., the mean value, of the vector of predictions, and σ_y,$\sigma_{\hat{y}}$ are the standard deviations of the observations and the predictions, respectively.

- The mean fractional bias – MFB:

$$\mathrm{MFB} = \frac{1}{N}\sum_{i=1}^{N}\frac{E(\hat{y}_i) - y_i}{\frac{1}{2}(E(\hat{y}_i) + y_i)}. \tag{12}$$

- The factor of the modelled values within a factor of two of the observations – FAC2:

$$\mathrm{FAC2} = \frac{1}{N}\sum_{i=1}^{N} n_i \quad \text{with} \quad n_i = \begin{cases} 1 & \text{for } 0.5 \le |\frac{E(\hat{y}_i)}{y_i}| \le 2, \\ 0 & \text{else.} \end{cases} \tag{13}$$

RMSE and SMSE are frequently used measures for the accuracy of the predictions' mean values, which are 0 in the case of a perfect model. SMSE is the standardised measure with values between 0 and 1. PCC is a measure of the associativity and is not sensitive to bias. Its value is between -1 and +1, with ideally linearly correlated values resulting in a value 1. MFB is the measure that bounds the maximum bias and gives additional weight to underestimations and less weight to overestimations. Its value is between -2 and +2, with the value 0 in the case of a perfect model. FAC2 indicates the fraction of the data that satisfies the condition from Equation (13). Its value is between 0 and 1, with the perfect model resulting in a value of 1.