Nonlinear Data Assimilation for the Regional Modelling of Maximum Ozone Values

MARIJA ZLATA BOŽNAR, BOŠTJAN GRAŠIČ AND PRIMOŽ MLAKAR
MEIS d.o.o., Mali Vrh pri Šmarju, Slovenia

DEJAN GRADIŠAR
Jožef Stefan Institute, Ljubljana, Slovenia

* JUŠ KOCIJAN
Jožef Stefan Institute, Ljubljana, Slovenia and University of Nova Gorica, Nova Gorica, Slovenia

*Corresponding author's address: Juš Kocijan, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: jus.kocijan@ijs.si

ABSTRACT

We present a new method of data assimilation with the aim of correcting the forecast of the maximum values of ozone in regional photo-chemical models for the areas over very complex terrain using multilayer perceptron artificial neural networks. Up until now all these models have been used as a single model for one location when forecasting concentrations of air pollutants. We propose how to construct a much more ambitious model - where the same model - one model - can be used at several locations - the model is spatially transferable and is valid for the whole 2D domain. To achieve these goals we have introduced three novel ideas. The new method improved on average by 10 % in the correlation at measurement-station locations, and improving by roughly 5 % elsewhere.

Keywords: *ozone forecast, data assimilation, complex terrain, neural networks, changing altitudes, geographically transferable artificial neural network model*

Acknowledgement

The authors acknowledge the projects (Development and implementation of a method for on-line modelling and forecasting of air pollution, L2-5475, Assesment of natural and antropogenic processes in micrometeorology of Postojna cave system by numerical models and modern methods of data aquisition and transfer, L2-6762) were financially supported by the Slovenian Research Agency.

We would also like to thank Slovenian Environmental Agency. We have used some of the data from their automatic environmental measuring network.

We would like to thank Arianet s.r.l. from Milan, which provides photochemical model results from the QualeAria system for the area of Slovenia, and Dragana Kokal and Darko Popović for their help with the editing.

1. Introduction

Forecasting ozone concentrations above complex terrain is still a current and somewhat unresolved issue. The established method is to use 3D photochemical models in combination with suitable NWP models (Kim et al., 2010). Despite its usefulness, the method does have certain flaws (Hogrefe et al., 2001).

One of the reported flaws is its biased and inaccurate forecasting of the maximum ozone concentrations for the following day or days (Curier et al., 2012; Dutot et al., 2007; Gong and Ordieres-Meré, 2016; Porter et al., 2015). It is this very information that is important in planning outdoor activities for the parts of the population who are affected by the ozone, experiencing acute or chronic health issues. Studies show the non-linear nature of ozone-health effects relationship (Schlink et al., 2006).

Operational 3D photochemical models are regional by nature. This means that they are intended for notifying populations in large areas. Which is why their spatial resolution is poor, and the local inhomogeneity is not noticeable in the model. However, averaging over spatially large cells usually smooth and lowers the maximum concentrations in the event of inhomogeneous conditions (Božnar et al., 2014a).

One of the possible approaches to improving the forecasting of the key piece of information, the maximum hourly concentration for the following day, is data assimilation. The aim of this data assimilation is to use the additional input data, i.e. local ground measurements or satellite observations, to improve the forecasting of the 3D photochemical model, as is demonstrated in the paper by Messina et al. or Zoogman (Messina et al., 2011; Zoogman et al., 2014). There are several ways of implementing data assimilation.

Examples of these are empirical analysis schemes, least square methods, linear multivariate statistical methods and others (Park et al., 2014). An overview of methods for data assimilation can be found in the literature, e.g. (Kalnay, 2003).

Data assimilation for areas over complex terrain is especially challenging, as the connections between influential variables and forecasts are decidedly nonlinear.

In this paper we will present a new nonlinear data assimilation method of detailed atmospheric conditions in the 2D results of a photochemical model.

The proposed data assimilation method is implemented using an artificial neural network, more precisely a multilayer perceptron (MLP), which is a universal approximator for complex nonlinear systems (Hornik et al., 1989; Kůrková, 1992).

We presented and evaluated the method on the example of forecasting the daily maximum ozone values for the area of Slovenia. Due to Slovenia's complex terrain, at the junction of the Alps, Pannonian Basin, Dinaric and coastal region by the Adriatic, this represents a particularly difficult challenge for all types of modelling of atmospheric events. This applies not only to weather forecasting, but also to the dispersion of air pollution in the atmosphere from local sources and regional photochemical models. However, this makes it representative of the various geographical regions such as the northern coasts of the Mediterranean, Alps, Dinaric and Pannonian Basin.

2. Description of the problem

The 3D photochemical regional model over complex terrain with inhomogeneous conditions for forecasting the ozone has the flaw of underestimating forecasts of the daily maximum concentration of ozone in the troposphere (Božnar et al., 2014a). On the contrary, suitable nonlinear empirical models developed for those locations where there are meteorological measurements and measurements of pollutant concentrations available are able to forecast the daily maximum concentrations much more accurately, but can only be used on measuring stations locations (Gradišar et al., 2016; Grašič et al., 2006; Ibarra-Berastegi et al., 2008; Luna et al., 2014). Artificial neural networks based models constitute an important part of such nonlinear group of models for more than a decade (Abdul-Wahab and Al-Alawi, 2002; Coman et al., 2008; Pelliccioni and Tirabassi, 2006).

We want to find out how nonlinear empirical models can be used to improve results for 2D areas of the troposphere over complex terrain where we use a 3D photochemical regional model to forecast the ozone.

This type of 2D spatially transferable model for data assimilation that has to be developed on the basis of data obtained from measurement stations on air quality and meteorological parameters should be useful for the entire region where we have results from the 3D photochemical model and NWP model and should ensure better forecasting of maximum ozone concentrations.

3. Material and methods

3.1. Test bed

The QualeAria system photochemical model has been operationally available to the public for the entire region of Slovenia and its neighbouring areas for several (Božnar et al., 2014b; QualeAria, 2016), and which was developed in the Arianet research company in collaboration with other Italian institutions (Zanini et al., 2005). We extensively evaluated the system for Slovenia using long data sets of measured data (Božnar et al., 2014b). Evaluating the QualeAria system gave good, solid results, however, the challenge of forecasting key maximum concentrations is still not resolved (Mircea et al., 2014).

For Slovenia, the QualeAria system forecasts photochemical and other pollutants in the atmosphere at a resolution of 12km horizontally and at intervals of 1 hour for 2 days in advance (the forecast for the current day and the following day). The results are regularly available on the KOoreg project website (MEIS, 2012; Mlakar et al., 2012).

Locally measured meteorological data and air quality data from 12 measuring stations from the national measurement network were used to improve these forecasts (SEA Data portal, 2016). Data from this measurement network were used in the form of average hourly values. Twelve measuring stations can be considered a good number for such a small area as Slovenia, and therefore can provide a high-quality evaluation of the ozone forecasting improvements.

Figure 1. Geographical location of the selected air quality stations (Source: Public information of Slovenia, Surveying and Mapping Authority of the Republic of Slovenia, Map of the Republic of Slovenija 1:500000, first edition 1997)

The stations are spread out in a very diverse variety of locations that can be initially classified according to their altitude (from the coast to the Alps, at an altitude of 1853m). They have also been positioned in large towns (Ljubljana, Celje, Nova Gorica, Maribor, Koper), with some also in small towns (Trbovlje, Zagorje, Murska Sobota, Hrastnik), two are at higher altitudes (Otlica and Krvavec), and one is in a sparsely populated remote inland agricultural area (Iskrba). The locations are shown in Figure 1.

3.2. *Modeling tools*

Data assimilation was performed using the Multilayer Perceptron Artificial Neural Network (MLPANN), which is an example of a feedforward neural network.

MLPANN is a mathematical structure capable of the approximation of a random non-singular nonlinear function of several independent variables (Hornik et al., 1989; Kůrková, 1992). The principles of ozone formation depending on other meteorological and air quality variables in the terrestrial atmosphere are definitely such function.

The complexity of MLPANN use is merely that the solution is not analytically achievable, but must be achieved with the use of iterative methods that minimize the function of the criteria – the mean squared error of prediction in the given set of known data.

The MLPNN establishes the principles – connections between independent variables and the dependent variable – based on the information given in known measured values of these variables. In our case those are the known values of the meteorological measurements, meteorological forecasts and measurements of the concentration of pollutants in the air and concentrations predictions from the regional model with a rough spatial resolution. In the MLPANN terminology we call such variables, features. The set of the values of these features at the same location and at the same time is called a pattern. The set-up of the new data assimilation model can be done based on the numerous patterns available from different locations across Slovenia, where there are stations for measuring the quality of the outside air and meteorological variables, thus this is the typical example of a problem, which is very suitable for modelling with the MLPANN.

With the MLPANN based model for the data assimilation method we would like to improve and localize the forecast of maximum ozone concentrations over complex terrain, which is produced with a general regional model with a rough spatial resolution.

The key steps of the construction of the MLPANN model based on the given set of patterns described with features are: the selection of features, the selection of patterns, the definition of a suitable MLPANN topology, which must include the hidden level with nonlinear transfer function neurons, the process of learning with the appropriate optimization algorithm and the optimization and testing of the built model. You can learn more about the theoretical background of these steps in the literature (Božnar, 1997; Lawrence, 1993; Mlakar, 1997). In the continuation of this text we will explain the implementation of these steps on a practical example.

4. Theory: methodology of non-linear data assimilation

However, in this paper we have made a key breakthrough by using the appropriate methodology of model use on the basis of MLP – we have created a spatially transferrable model for nonlinear data assimilation with which we have significantly improved the basic ozone forecasts, which are the result of the FARM 3D photochemical model in the QualeAria system over the whole model domain.

Our first goal was to improve the forecasting of the maximum one-hour concentrations of ozone for the following day for each of the selected locations where measurements and results from the NWP and QualeAria FARM model were available. The method is based on our previous methods (Kocijan, 2016; Kocijan et al., 2015; Petelin et al., 2015), but which we have significantly upgraded with additional input features (regressors) and a new approach to model development with which we have achieved spatial transferability between various locations where air pollution measurements are carried out.

Instead of creating a model for each measurement location, we created a model that would work for all locations.

Our second, even more ambitious goal was to create a 2D spatially transferrable model of nonlinear data assimilation that would work for the ground level of entire modelled area, therefore even where there are no air pollution or meteorological measurements, but there are improved numerical weather forecasts available in a detailed spatial and time resolution.

This type of spatially transferable model has to be developed on the basis of data obtained from measurement stations on air quality and meteorological parameters, but can be used for all other areas where we have the results from the 3D photochemical model and NWP in a more detailed resolution. In other words, this means that we can significantly improve the entire 2D ground level photochemical model result for all the modelled areas, all ground cells of this basic model, and in this way we can achieve the true spatial transferability of the nonlinear data assimilation model.

Whereby, the key feature of this new model should be highlighted, namely the fact that it is based on measured laws of ozone behaviour in combination with local meteorological characteristics at the selected measurement location itself, and not on the linear interpolation and Kriging-related methods (Stein, 1999) that depend on the Euclidean position of the locations in view of the neighbouring locations, which are unsuitable from the point of view of their theoretical basis for inhomogeneous conditions above complex terrain, such as it is dealt with in this paper.

We tested the validity of both resulting models using independent data of the measuring stations described in the test bed chapter.

4.1. The general procedure of constructing an MLP-based model

The construction of an MLP-based model has the following steps: features determination, patterns selection, determining the network topology, training the model for the selected measured data by optimizing the model parameters, and testing – validating – using independent data measured in nature and which were not used in the model development process (Mlakar and Božnar, 1997). These steps are described in more detail in the subchapters that follow.

To achieve the model's transferability to the entire domain of the photochemical model on the basis of the selected learning locations only, the generalizing capability of the universal approximator is taken advantage of. This capability means that the MLP-based model is able to also forecast the value of the output variable for input variable values that somewhat deviate from known learning patterns, therefore for similar patterns. Learning patterns for training the model must be selected in such a way that the physical patterns from any location on the domain will be similar to at least some of the learning patterns.

4.1.1 Features

Features or regressors must be suitably informative. This means that they comprise of those quantity values that affect the formation of the ozone. As ozone-forming laws are not entirely known, especially over complex terrain, systematic searching methods can be used. An overview of methods can be found in the literature, e.g. (Guyon and Elisseeff, 2003; May et al., 2011).

4.1.2 Learning patterns

Learning patterns or regression vectors are measured values or known numeric variable forecasts included in the input and output features. They must be selected so that they contain the key characteristics of the geographically and meteorologically varied locations. In theory, it would be best if they included all the locations for which the model is being developed. As this is not possible, it is necessary to select the most representative locations that can be chosen on the basis of experience or algorithms.

In addition to the known patterns for training the model, we also need known patterns for evaluating the model. These have to be independent known patterns that were not used in the model development process.

In the Test Bed chapter we have described the measuring station locations. We used data from three of these stations for training the model for the proposed nonlinear data assimilation procedure. We then tested the model using independent data. These are data that were not used in the model development process; they come from the same stations but for other time periods than those that were used for learning. However, the additional test using other completely independent stations and their data was the key point.

4.2. Input features for nonlinear data assimilation

The initial selection of features is based on past studies (Grašič et al., 2006; Kocijan, 2016; Petelin et al., 2015). The selection of previously known and evaluated features (Table 1) is supplemented with basic ground meteorology, pollutant concentration measurements and photochemical model forecasts for these variables, both for the present time and for the following day.

We also used meteorological parameter forecasts in significantly more detailed spatial resolution than the QualeAria system originally uses. The MEIS system for weather forecasting based on WRF and GFS input data is used (Božnar et al., 2011). To make an improvement we use the morning numerical weather forecast for the current day and the following day. The end result – a 2D forecast of the maximum ozone concentration – is given in a detailed spatial resolution, just as the WRF forecasts are, where the cells are three times shorter horizontally than those in the QualeAria system (from 12km × 12km to 4km × 4km). Better spatial resolution is another improvement when forecasting maximum ozone concentrations using our proposed method.

As the number of potential additional features is manageable, we evaluated them by systematically adding and taking away, using the forward selection method (May et al., 2011) until we reached the best possible selection of features.

Table 1: Input features of the initial model (Kocijan, 2016; Petelin et al., 2015).

Code	**Parameter**				**Source**
O3(*k*)	ozone concentration	1-h average	measured daily maximum	present	AMS
pGlSolRad(*k*+1)	global solar radiation	1-h average	forecasted daily maximum	predicted	WRF
pTemp(*k*+1)	air temperature	1-h average	forecasted daily maximum	predicted	WRF
Temp(*k*)	air temperature	1-h average	measured daily maximum	present	AMS
GlSolRad(*k*)	global solar radiation	1-h average	measured daily maximum	present	AMS
pRelHum(*k*+1)	relative humidity	1-h average	forecasted daily maximum	predicted	WRF
pNOx(*k*+1)	nitrogen oxides	1-h average	forecasted daily maximum	predicted	QualeAria
pPressure(*k*+1)	air pressure	1-h average	forecasted daily maximum	predicted	WRF
pPressure(*k*)	air pressure	1-h average	forecasted daily maximum	present	WRF
pO3(*k*+1)	ozone concentration	1-h average	forecasted daily maximum	predicted	QualeAria

AMS ... Automatic measuring station
WRF ... The Weather Research & Forecasting Model
QualeAria ... Forecast system for the Air Quality
k+1 ... the day for which the forecast is made
k ... the day before the day for which the forecast is made

We first expanded the basic set of features from Table 1 with the feature – model forecast of diffuse solar radiation. Diffuse solar radiation and direct solar radiation form the global solar radiation. The amount of direct solar radiation is proportional to the UV radiation, which is key for the formation of ozone (Finlayson-Pitts and Pitts, 1999). This is the main reason that leads us to the idea of it being possible to use complementary diffuse solar radiation to improve ozone forecasting. NWP models already have a global solar radiation forecast built in and it is considered in the 3D diffuse photochemical model. The local forecast of direct solar radiation is a potentially good feature to improve ozone forecasts. But as there is no direct solar radiation model available in NWP model, a complementary feature – diffuse solar radiation is used. That is the output of the statistical MLP model for the 2D distribution of diffuse solar radiation, which uses values from the NWP model as its input (Božnar et al., 2016). The model forecasts diffuse solar radiation for the entire area of Slovenia in the KOoreg system in a spatially horizontal resolution of 4km and in a temporal interval of up to two days in advance and half hour resolution. Additional confirmation of the choice of features were the known models for global and diffuse solar radiation for clear sky conditions (Badescu et al., 2012) where the rule of connection between the ozone and solar radiation that reaches the ground is used in the opposite direction.

However, since the goal is to achieve spatial transferability of the model for the proposed data assimilation, we had to also find a solution for the normalization/relativization of the meteorological input features which have significantly varied ranges depending on altitude or spatial locations due to the variety of climate conditions. Examples of this type of input feature are air temperature and air pressure. Instead of absolute air pressure, we used air pressure reduced to sea level (Pugh, 1996). Instead of air temperature, we used potential temperature (Bolton, 1980).

The final set of features for locations for which meteorological measurements and ozone concentration measurements, as well as QualeAria FARM 3D photochemical model and NWP WRF model results are available in detailed resolution are shown in Table 2.

Table 2: Input feature of the basic type of transferrable model

Code	**Parameter**				**Source**
O3(*k*)	ozone concentration	1-h average	measured daily maximum	present	AMS
pGlSolRad(*k*+1)	global solar radiation	1-h average	forecasted daily maximum	predicted	WRF
pPotTemp(*k*+1)	**potential** air temperature	1-h average	forecasted daily maximum	predicted	WRF
PotTemp(*k*)	**potential** air temperature	1-h average	measured daily maximum	present	AMS
GlSolRad(*k*)	global solar radiation	1-h average	measured daily maximum	present	AMS
pRelHum(*k*+1)	relative humidity	1-h average	forecasted daily maximum	predicted	WRF
pNOx(*k*+1)	nitrogen oxides	1-h average	forecasted daily maximum	predicted	QualeAria
pPressure(*k*+1)	**mean sea-level** air pressure	1-h average	forecasted daily maximum	predicted	WRF
pPressure(*k*)	**mean sea-level** air pressure	1-h average	forecasted daily maximum	present	WRF
pO3(*k*+1)	ozone concentration	1-h average	forecasted daily maximum	predicted	QualeAria
pDifSolRad(*k*+1)	**diffuse solar radiation**	daily sum	forecasted daily sum energy	predicted	ANN-WRF

AMS ... Automatic measuring station
WRF ... The Weather Research & Forecasting Model
QualeAria ... Forecast system for the Air Quality
ANN-WRF ... Artificial Neural Network model of diffuse solar radiation based on WRF inputs
k+1 ... the day for which the forecast is made
k ... the day before the day for which the forecast is made

The set from Table 2 is used to develop a basic type of model that only works for locations where there are concentration measurements and meteorological measurements available.

The end goal is to implement the data assimilation for all other ground level 2D areas for which only QualeAria FARM photochemical model results are available and measurements not. This is why we developed the upgraded model type. The difference between the basic and the upgraded model is in used input features. It only uses FARM and WRF forecasts for its input features (Table 3). It should be emphasized that for this model, solely for the training process, ozone data and other air quality and meteorological parameters from the selected representative training locations where there are measuring stations were nonetheless used. These data are not used for forecasting. The MLP model draws on information from known data that describe the studied rule. If it is successfully trained, the rule that is dealt with can be used in a general way.

Table 3: Input feature of the upgraded type of transferrable model

Code	**Parameter**				**Source**
pO3(*k*)	ozone concentration	1-h average	measured daily maximum	present	**QualeAria**
pGlSolRad(*k*+1)	global solar radiation	1-h average	forecasted daily maximum	predicted	WRF
pPotTemp(*k*+1)	**potential** air temperature	1-h average	forecasted daily maximum	predicted	WRF
PotTemp(*k*)	**potential** air temperature	1-h average	measured daily maximum	present	**WRF**
GlSolRad(*k*)	global solar radiation	1-h average	measured daily maximum	present	**WRF**
pRelHum(*k*+1)	relative humidity	1-h average	forecasted daily maximum	predicted	WRF
pNOx(*k*+1)	nitrogen oxides	1-h average	forecasted daily maximum	predicted	QualeAria
pPressure(*k*+1)	**mean sea-level** air pressure	1-h average	forecasted daily maximum	predicted	WRF
pPressure(*k*)	**mean sea-level** air pressure	1-h average	forecasted daily maximum	present	WRF
pO3(*k*+1)	ozone concentration	1-h average	forecasted daily maximum	predicted	QualeAria
pDifSolRad(*k*+1)	**diffuse solar radiation**	daily sum	forecasted daily sum energy	predicted	ANN-WRF

WRF... The Weather Research & Forecasting Model
QualeAria ... Forecast system for the Air Quality
ANN-WRF ... Artificial Neural Network model of diffuse solar radiation based on WRF inputs
k+1 ... the day for which the forecast is made
k ... the day before the day for which the forecast is made

4.3. Selection of patterns for nonlinear data assimilation

In this chapter, the key new step in pattern selection is presented which ensures that we can develop just one model that is spatially transferable and because of that it is not necessary to develop a model for each location separately, as was in the articles that deal with the model for the same area (Božnar et al., 1993; Kocijan, 2016; Kocijan et al., 2015; Mlakar and Božnar, 2011; Petelin et al., 2015).

We achieved spatial transferability by constructing a model for the proposed nonlinear data assimilation on the basis of the data for three measuring stations all together and not for each one separately. The MLP's characteristic of being a universal approximator with generalizing capabilities is used.

We used data from the stations in Koper (on the Adriatic coast), Krvavec (a high-altitude station in the Alps) and Murska Sobota (in the Pannonian Basin, on one of the rare flat areas in Slovenia). These three stations illustrate the key differences between the characteristics of modelling the ozone above Slovenia (an urban, coastal, suburban flat area, and high-altitude station). Because MLP model is sensitive to the selected stations care needs to be taken how to select these three or more stations.

We developed the model using the training data from these stations, which are combined into one common set of learning patterns. The same principle of combining patterns from various stations into one common training set was first used for the basic model with basic features, and secondly for the upgraded model.

4.4. Training the MLP model for data assimilation

To train the basic and upgraded model we used the following MLP configuration, which is based on the number of neurons depending on the number of patterns, as was determined in our preliminary study (Grašič et al., 2006) where there were 50 neurons. We reduced this number by trial and error so that we achieved rapid training without the excessive memorizing of learning patterns.

Both final MLPs in our study share the same properties:

- input features as listed in Tables 2 and 3;
- 20 neurons in hidden layer based on a tangent sigmoid activation function;
- output layer with one neuron based on a linear activation function to reproduce output feature;
- iterative backpropagation training algorithm (gradient descent with a momentum optimization algorithm; momentum = 0.1; learning rate = 0.1).

4.5. *Model development and validation data*

To develop and independently evaluate the models for the proposed data assimilation method, we have data on the features for the following periods at our disposal:

Table 4: Model development and evaluation periods

Purpose	**Period**
Model development	1 January 2012 – 31 December 2012 (1 year)
Model evaluation	1 January 2013 – 31 December 2014 (2 years)

The first year's data were used to train and optimize the models, while the remaining two years of data are used to independently test the models. This means of pattern selection was also used in the previous study (Petelin et al., 2015).

For training the basic model we used 1,000 learning patterns from three different locations and 1,030 learning patterns from the same three locations for the upgraded model. The number is not equal to the sum of all the days for the selected period, as it was necessary to exclude some of the intervals where not all the measurements or forecasts were available. This is normal in measuring networks and operational modelling systems.

5. Results and discussion

In the present paper we propose how to construct a model for nonlinear data assimilation - where the same model - one model - can be used at several locations - the model is spatially transferable and is valid for the whole 2D domain. To achieve this we propose two approaches:

- how to construct a model that can be used on an arbitrary location in a 2D domain where air quality and meteorological measurements are available – basic model;

- how to construct a model that can be used over the whole 2D domain under the condition that a numerical weather forecast is available in a better resolution than was used originally by the photo-chemical regional model – upgraded model.

In both cases, only one model should be constructed and not several models, one for each location under examination. To achieve these goals we have introduced the following novel ideas:

- a new method of model training patterns selection;
- normalized values of air temperature and pressure (potential temperature and mean sea level pressure) were used as input features for the data assimilation model; this enables the model to be used regardless of the altitude;
- new feature - diffuse solar radiation - adds additional information and improves results;
- the multilayer perceptron artificial neural network based model was used as a method of non-linear data assimilation for 2D areas (and not as a model for the direct forecasting of pollutant concentration at one location - as is the case in the known models since our first publication of the SO_2 prediction model in the Atmospheric Environment in 1993 (Božnar et al., 1993)).

For the proposed data assimilation basic and upgraded model, estimators for the basic output from the QualeAria FARM model for the output feature of forecasting one-hour maximum ozone concentrations for the following day were used as the starting point for comparison.

Both models for the proposed data assimilation method were first tested on the three stations that were used in the training process. The evaluation was carried out on completely independent patterns from these stations that were not used for training. Then both models were tested on the remaining stations for which measurements were available. These other stations were not used in training the models.

The evaluation of the quality of data assimilation can only be implemented for the measuring station locations, even for the upgraded model which can be used for the entire 2D output area from the FARM model. On the basis of this result we predict that a similar improvement will occur on the entire 2D area. We reason this prediction with the fact that we have a relatively large number of stations available in Slovenia that are distributed throughout parts of the territory that have very different characteristics.

We used the following evaluators for the statistical processing of the validation results (Badescu et al., 2012; Gradišar et al., 2016; Kocijan et al., 2016):

- The root mean square error (RMSE)
- The normalized mean squared error (NMSE)
- The coefficient of determination (R^2)
- Pearson's correlation coefficient (PCC)
- The mean fractional bias (MFB)
- The factor of the modelled values within a factor of two of the observations (FAC2).
- The normalized mean bias error (MBE) MBE [%]
- The coefficient of variation based on the root mean square value CV(RMSE) [%].

In Tables 5, 6 and 7, we have presented the evaluator values for ozone forecasts from the QualeAria model, the basic MLP model and the upgraded MLP model for data assimilation.

Scatter diagrams on Figures 2 to 5 illustrate the matching measurements and forecasts obtained through data assimilation.

Table 5: Evaluators for the QualeAria model ozone forecast for all locations using data not used in model development.

Station	**RMSE [μg/m³]**	**NMSE**	**R²**	**PCC**	**MFB**	**FAC2**	**MBE [%]**	**CV (RMSE) [%]**
Celje	30.685	0.161	0.652	0.807	-0.164	0.91	20.71	35.72
Hrastnik	19.669	0.068	0.692	0.832	-0.100	0.96	12.32	24.44
Iskrba	20.583	0.063	0.644	0.802	-0.152	0.99	15.54	22.98
Koper	25.691	0.094	0.744	0.863	-0.236	0.97	20.97	27.27
Krvavec	40.268	0.209	0.517	0.719	-0.430	0.85	33.99	37.19
Ljubljana	29.412	0.153	0.703	0.838	-0.201	0.91	22.43	34.46
Maribor	20.611	0.086	0.694	0.833	-0.053	0.93	10.47	27.70
Murska Sobota	21.213	0.080	0.695	0.834	-0.139	0.96	15.10	26.10
Nova Gorica	26.007	0.111	0.741	0.861	-0.117	0.94	16.96	30.30
Otlica	26.302	0.099	0.659	0.812	-0.283	0.93	23.11	27.65
Trbovlje	17.760	0.060	0.682	0.826	-0.007	0.95	5.51	23.76
Zagorje	18.344	0.066	0.678	0.823	0.068	0.92	0.83	25.50

Table 6: Evaluators for the basic MLP model for air quality automatic measuring station locations ozone forecast using data not used for training the MLP model.

Station	**RMSE [μg/m³]**	**NMSE**	**R²**	**PCC**	**MFB**	**FAC2**	**MBE [%]**	**CV (RMSE) [%]**
Celje	17.720	0.041	0.808	0.899	0.100	0.94	-4.28	20.63
Hrastnik	15.788	0.035	0.787	0.887	0.123	0.95	-9.13	19.62
Iskrba	14.032	0.023	0.728	0.853	0.073	0.99	-5.52	15.67
Koper	12.984	0.019	0.841	0.917	0.034	0.99	-1.48	13.78
Krvavec	11.669	0.012	0.761	0.872	-0.031	1.00	3.04	10.78
Ljubljana	17.179	0.038	0.822	0.907	0.116	0.94	-6.17	20.12
Maribor	16.433	0.044	0.819	0.905	0.165	0.92	-10.39	22.09
Murska Sobota	14.980	0.032	0.798	0.893	0.103	0.95	-6.45	18.43
Nova Gorica	16.935	0.037	0.824	0.908	0.110	0.94	-5.12	19.73
Otlica	12.536	0.018	0.726	0.852	0.000	1.00	0.86	13.18
Trbovlje	17.700	0.049	0.771	0.878	0.174	0.94	-13.54	23.68
Zagorje	18.719	0.058	0.786	0.887	0.218	0.91	-16.24	26.02

Table 7: Evaluators for the upgraded MLP model for the 2D area of the entire Slovenia ozone forecast using data not used for training the MLP model.

Station	**RMSE [µg/m³]**	**NMSE**	**R²**	**PCC**	**MFB**	**FAC2**	**MBE [%]**	**CV (RMSE) [%]**
Celje	21.736	0.058	0.763	0.873	0.170	0.91	-9.53	25.29
Hrastnik	22.795	0.066	0.750	0.866	0.240	0.92	-21.06	28.32
Iskrba	20.256	0.047	0.620	0.788	0.149	0.97	-12.95	23.37
Koper	15.854	0.027	0.806	0.898	0.058	0.98	-2.37	16.59
Krvavec	19.500	0.035	0.527	0.726	-0.096	1.00	8.41	18.00
Ljubljana	19.638	0.048	0.801	0.895	0.159	0.92	-8.92	22.85
Maribor	24.760	0.088	0.773	0.879	0.300	0.87	-24.48	33.05
Murska Sobota	19.512	0.050	0.784	0.885	0.201	0.93	-15.76	24.05
Nova Gorica	22.174	0.059	0.807	0.898	0.205	0.90	-12.79	25.80
Otlica	17.433	0.031	0.597	0.773	0.004	1.00	1.08	17.57
Trbovlje	28.171	0.109	0.700	0.837	0.326	0.89	-30.39	37.67
Zagorje	30.670	0.134	0.744	0.863	0.388	0.85	-36.28	42.66

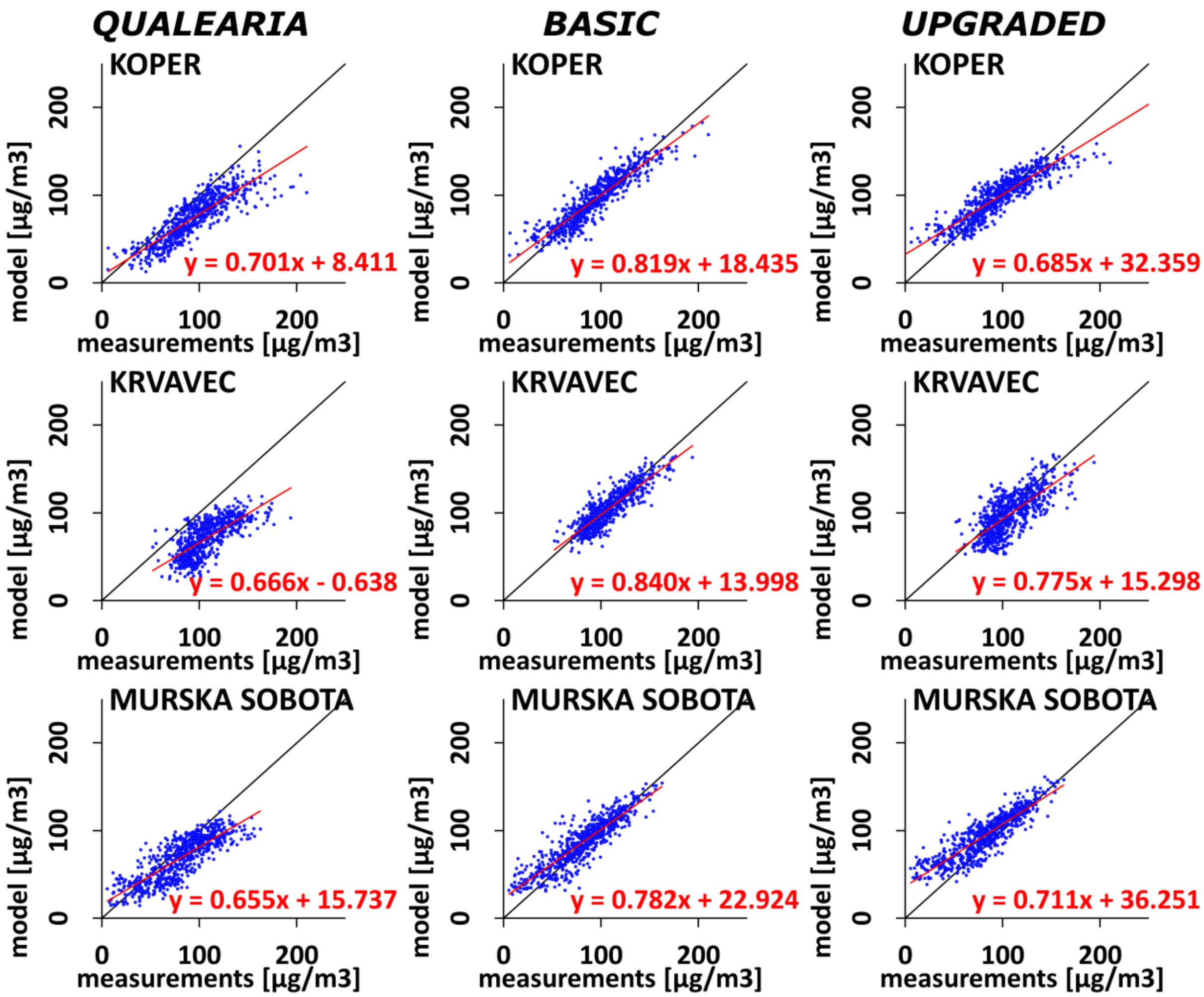

Figure 2: Comparison of modelling results of the 1^{st} group of stations (Koper, Krvavec and Murska Sobota, these three stations were used to train both MLP models) from three models (first column photochemical model QualeAria, second column MLP Basic Model, third column MLP Upgraded Model)

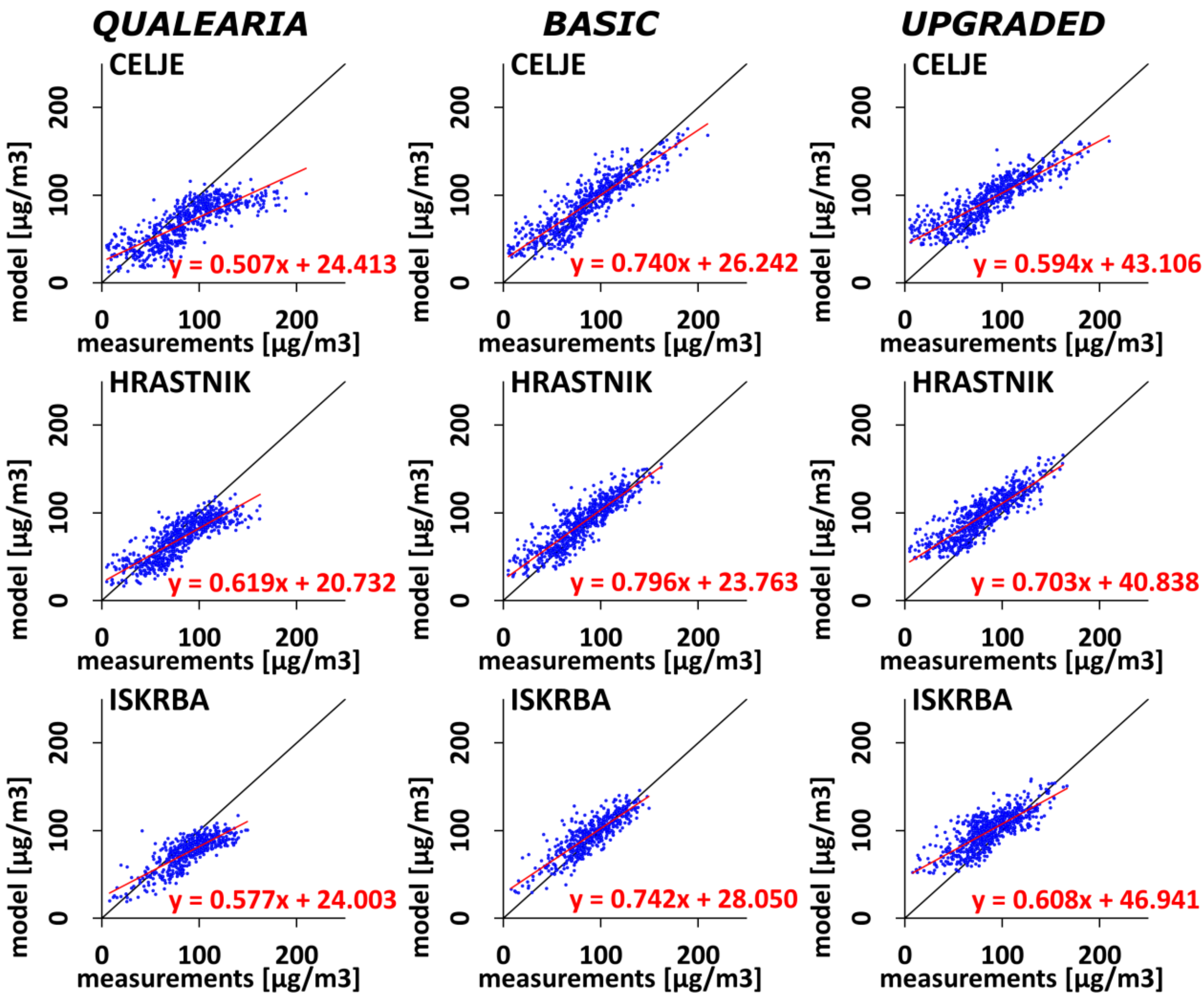

Figure 3: Comparison of modelling results of the 2^{nd} group of stations (Celje, Hrastnik and Iskrba) from three models (first column photochemical model QualeAria, second column MLP Basic Model, third column MLP Upgraded Model)

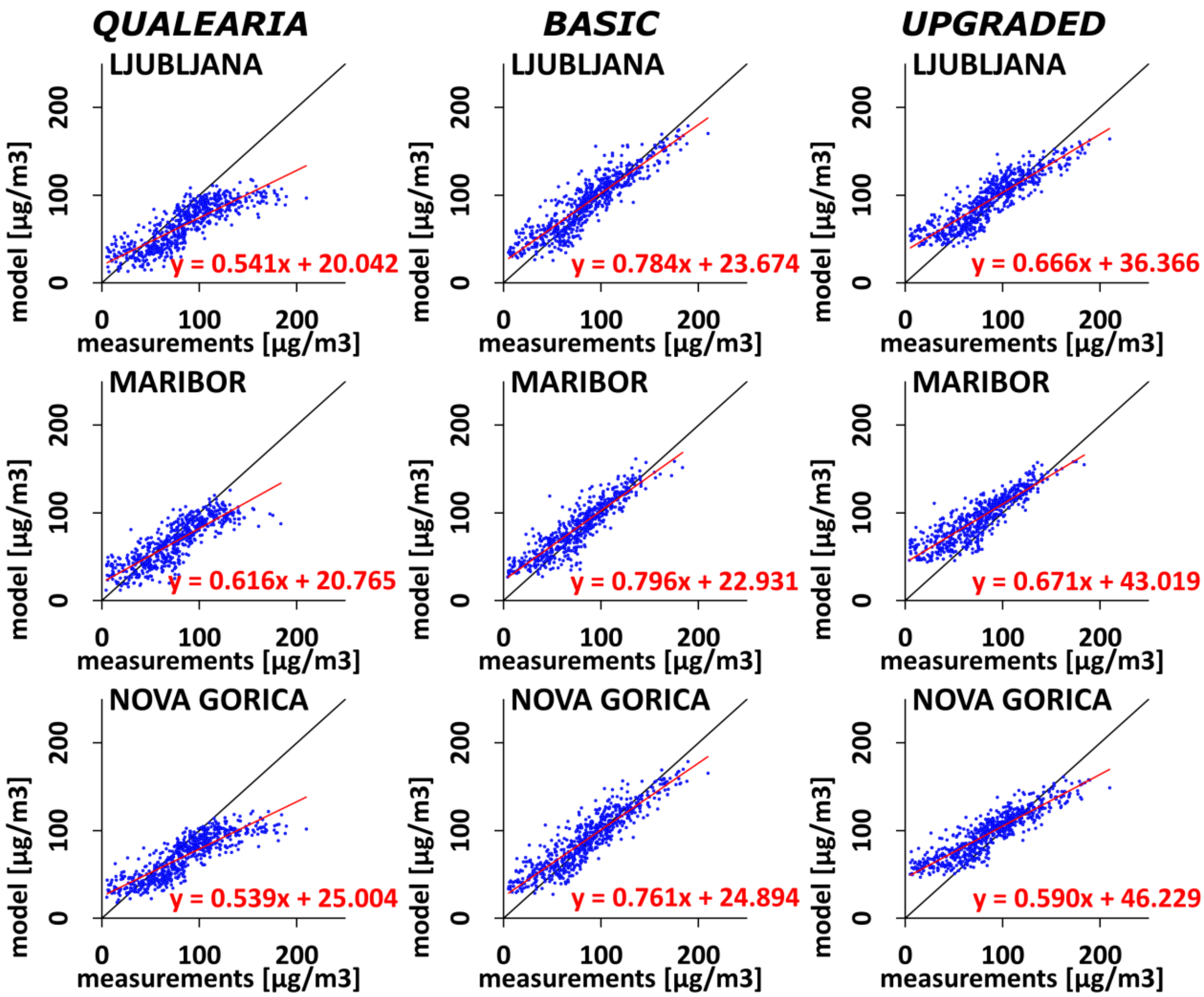

Figure 4: Comparison of modelling results of the 3^{rd} group of stations (Ljubljana, Maribor and Nova Gorica) from three models (first column photochemical model QualeAria, second column MLP Basic Model, third column MLP Upgraded Model)

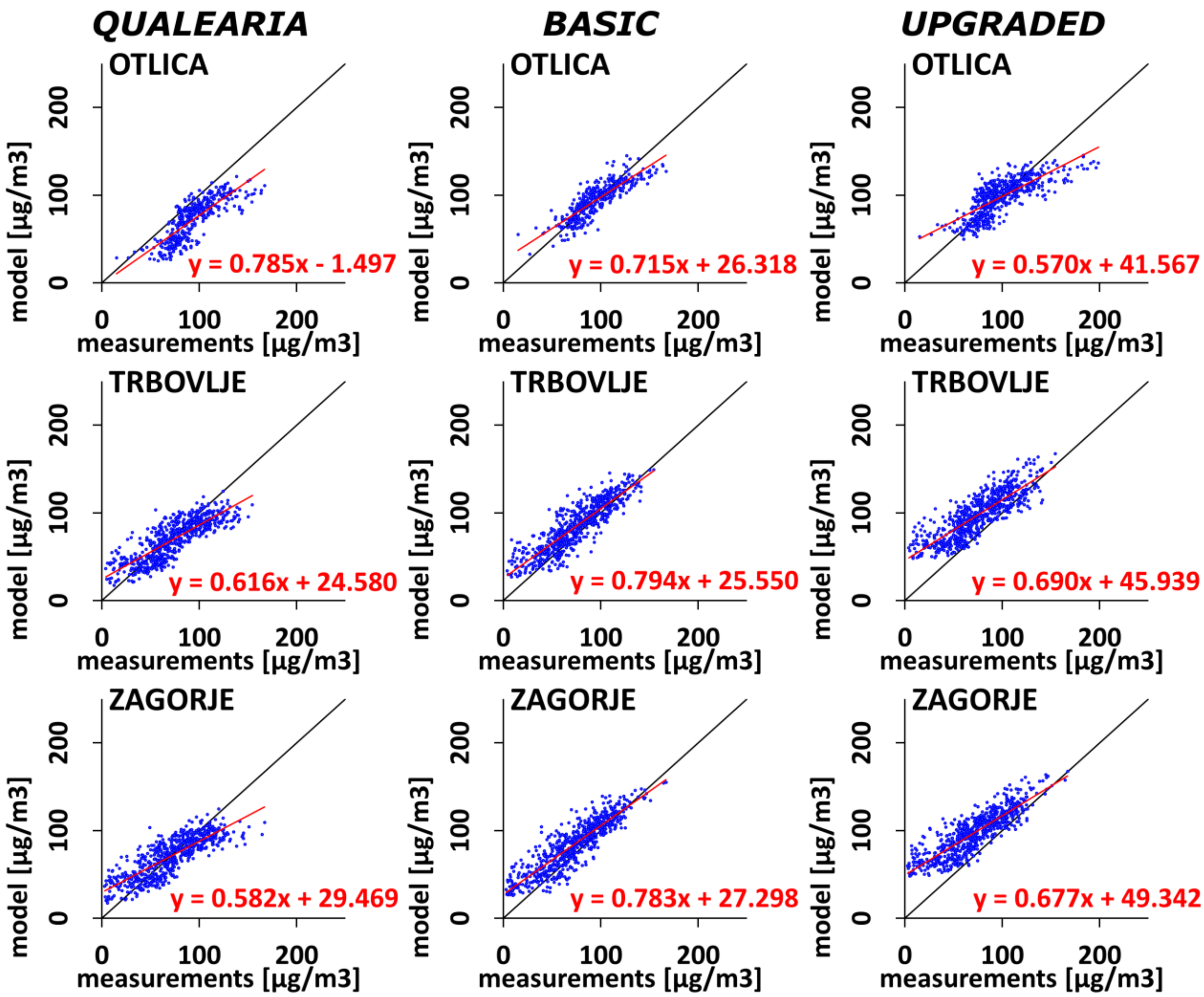

Figure 5: Comparison of modelling results of the 4th group of stations (Otlica, Trbovlje and Zagorje) from three models (first column photochemical model QualeAria, second column MLP Basic Model, third column MLP Upgraded Model)

It is clear from Figure 2 and the corresponding evaluator values from Table 5, 6 and 7 that nonlinear data assimilation significantly improves the results of the QualeAria photochemical model.

The main result is visible in Figures 3, 5 and 5 and the corresponding evaluator values from Table 5, 6 and 7 where we can see that the proposed nonlinear data assimilation is successful throughout the entire ground-level 2D domain. An example of the 2D one-hour maximum ozone ground concentrations is presented on Figure 6.

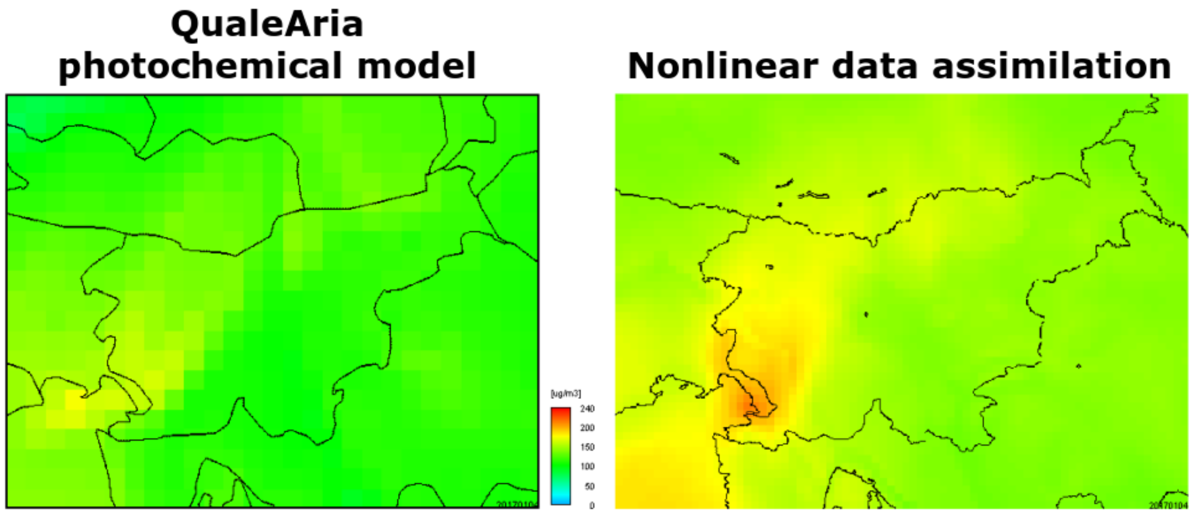

Figure 6: Example of the one-hour maximum ozone ground concentrations forecasted by QualeAria FARM model (left figure) and proposed nonlinear data assimilation upgraded model (right figure)

The basic model for automatic air quality measuring station locations indicates a great improvement in ozone forecasting, while the upgraded model for the 2D area of entire Slovenia predicts slightly worse, but still shows a significant improvement. The greatest improvement was achieved with the highest ozone values. The new method improved on average by 10% in the correlation at measurement station locations, and improving by roughly 5% regarding the remaining intermediate ground level area of the modelling domain.

This was the goal, and in doing so we confirmed the proposed concept on this example. With great likelihood we can predict that the behaviour of the proposed method will give comparable results on other complex terrains.

6. Conclusions and future work

In this paper, a method of nonlinear data assimilation of measured and forecasted data on atmospheric characteristics into 2D ground level results of ozone forecasts from a 3D photochemical model. New method was tested using the example of ozone forecasting in Slovenia. The same methodology might be used for other pollutants concentrations including other than maximal values forecasting but this assumptions are yet to be investigated.

Using the proposed method of nonlinear data assimilation we significantly improved the key forecasting of the maximum one-hour ozone concentration for one day in advance. This forecast of maximum ozone concentration is crucial for planning daily activities for the sensitive section of the population where high concentrations cause a worsening of their state of health.

We developed a model for nonlinear data assimilation based on MLP so that it has a key new feature – spatial transferability. We were able to successfully integrate the general information of the link between the concentration of ozone with the remaining meteorological characteristics of the atmosphere and the characteristics of atmospheric pollution. We extracted this information from the measured data from only three measuring stations in Slovenia, while the model proved to work for all 12 stations in Slovenia which we have at our disposal to use for validation. The stations are located over very complex terrain of Slovenia – a junction of Alpine mountains, Pannonian plains and the Mediterranean sea.

According to our knowledge of the field, this is the first example in the world of an MLP-based model for the ozone which is spatially transferable and successfully improves the results of photochemical models for the ground level. But on a similar atmospheric problem of investigating solar radiation, we have just successfully developed spatially transferable model of diffuse solar radiation using similar methodology (Božnar et al., 2016).

The basic model of nonlinear data assimilation is useful for locations where meteorological and ozone measurements are available in real time. This model demonstrates a significant improvement in ozone forecasting. The upgraded model of nonlinear data assimilation is useful for all ground level cells in the QualeAria FARM photochemical model for the entire area of Slovenia and does not require additional measurements. Due to its greater generality, the upgraded model achieves somewhat smaller, but still significant, improvements in ozone forecasting.

However, it is both models that demonstrate a key new capability – spatial transferability – which originates from the means of development as we have described it.

Regarding our future work, we foresee the testing of both models also for neighbouring countries where we have the QualeAria system and measuring data at our disposal. We want to evaluate the level of universal spatial transferability of the two nonlinear data assimilation models developed.

References

Abdul-Wahab, S.., Al-Alawi, S.., 2002. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environ. Model. Softw. 17, 219–228. doi:10.1016/S1364-8152(01)00077-9

Badescu, V., Gueymard, C.A., Cheval, S., Oprea, C., Baciu, M., Dumitrescu, A., Iacobescu, F., Milos, I., Rada, C., 2012. Computing global and diffuse solar hourly irradiation on clear sky. Review and testing of 54 models. Renew. Sustain. Energy Rev. 16, 1636–1656. doi:10.1016/j.rser.2011.12.010

Bolton, D., 1980. The Computation of Equivalent Potential Temperature. Mon. Weather Rev. 108, 1046–1053. doi:10.1175/1520-0493(1980)108<1046:TCOEPT>2.0.CO;2

Božnar, M., 1997. Pattern selection strategies for a neural network-based short term air pollution prediction model. Intell. Inf. Syst. 1997. IIS ’97. Proc. 340–344. doi:10.1109/IIS.1997.645285

Božnar, M., Lesjak, M., Mlakar, P., 1993. A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. Atmos. Environ. Part B. Urban Atmos. 27, 221–230. doi:10.1016/0957-1272(93)90007-S

Božnar, M.Z., Grašic, B., Mlakar, P., 2014. The problem of limit values exceedances detection in complex terrain using measurement and models, in: HARMO 2014 - 16th

International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Proceedings. pp. 287–291.

Božnar, M.Z., Grašič, B., De Oliveira, A.P., Soares, J., Mlakar, P., 2016. Spatially transferable regional model for half-hourly values of diffuse solar radiation for general sky conditions based on perceptron artificial neural networks. Renew. Energy. doi:10.1016/j.renene.2016.11.013

Božnar, M.Z., Mlakar, P., Grašič, B., 2011. Short-term fine resolution WRF forecast data validation in complex terrain in Slovenia, in: HARMO 2011 - Proceedings of the 14th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes.

Božnar, M.Z., Mlakar, P., Grašič, B., Calori, G., D'Allura, A., Finardi, S., 2014. Operational background air pollution prediction over Slovenia by QualeAria modelling system - validation. Int. J. Environ. Pollut. 54, 175–183.

Coman, A., Ionescu, A., Candau, Y., 2008. Hourly ozone prediction for a 24-h horizon using neural networks. Environ. Model. Softw. 23, 1407–1421. doi:10.1016/j.envsoft.2008.04.004

Curier, R.L., Timmermans, R., Calabretta-Jongen, S., Eskes, H., Segers, A., Swart, D., Schaap, M., 2012. Improving ozone forecasts over Europe by synergistic use of the LOTOS-EUROS chemical transport model and in-situ measurements. Atmos. Environ. 60, 217–226. doi:10.1016/j.atmosenv.2012.06.017

Dutot, A.-L., Rynkiewicz, J., Steiner, F.E., Rude, J., 2007. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. Environ. Model. Softw. 22, 1261–1269. doi:10.1016/j.envsoft.2006.08.002

Finlayson-Pitts, B.J., Pitts, J.N., 1999. Chemistry of the Upper and Lower Atmosphere: Theory, Experiments, and Applications. Elsevier Science.

Gong, B., Ordieres-Meré, J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong. Environ. Model. Softw. 84, 290–303. doi:10.1016/j.envsoft.2016.06.020

Gradišar, D., Grašič, B., Božnar, M.Z., Mlakar, P., Kocijan, J., 2016. Improving of local ozone forecasting by integrated models. Environ. Sci. Pollut. Res. doi:10.1007/s11356-016-6989-2

Grašič, B., Mlakar, P., Božnar, M.Z., 2006. Ozone prediction based on neural networks and Gaussian processes. Nuovo Cim. C Geophys. Sp. Phys. C 29, 651–661. doi:10.1393/ncc/i2006-10011-5

Guyon, I., Elisseeff, a, 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. doi:10.1162/153244303322753616

Hogrefe, C., Rao, S.T., Kasibhatla, P., Hao, W., Sistla, G., Mathur, R., McHenry, J., 2001. Evaluating the performance of regional-scale photochemical modeling systems: Part II—ozone predictions. Atmos. Environ. 35, 4175–4188. doi:10.1016/S1352-2310(01)00183-2

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366. doi:10.1016/0893-6080(89)90020-8

Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argandoña, J., 2008. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. Environ. Model. Softw. 23, 622–637. doi:10.1016/j.envsoft.2007.09.003

Kalnay, E., 2003. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press.

Kim, Y., Fu, J.S., Miller, T.L., 2010. Improving ozone modeling in complex terrain at a fine grid resolution: Part I – examination of analysis nudging and all PBL schemes associated with LSMs in meteorological model. Atmos. Environ. 44, 523–532. doi:10.1016/j.atmosenv.2009.10.045

Kocijan, J., 2016. Modelling and Control of Dynamic Systems Using Gaussian Process Models, Advances in Industrial Control. Springer International Publishing, Cham. doi:10.1007/978-3-319-21021-6

Kocijan, J., Gradišar, D., Božnar, M.Z., Grašič, B., Mlakar, P., 2016. On-line algorithm for ground-level ozone prediction with a mobile station. Atmos. Environ. 131, 326–333.

Kocijan, J., Hančič, M., Petelin, D., Božnar, M.Z., Mlakar, P., 2015. Regressor selection for ozone prediction. Simul. Model. Pract. Theory 54, 101–115. doi:10.1016/j.simpat.2015.03.004

Kůrková, V., 1992. Kolmogorov’s theorem and multilayer neural networks. Neural Networks 5, 501–506. doi:10.1016/0893-6080(92)90012-8

Lawrence, J., 1993. Introduction to Neural Networks. California Scientific Software, Nevada City, CA, USA.

Luna, A.S., Paredes, M.L.L., de Oliveira, G.C.G., Corrêa, S.M., 2014. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. Atmos. Environ. 98, 98–104. doi:10.1016/j.atmosenv.2014.08.060

May, R., Dandy, G., Maier, H., 2011. Review of Input Variable Selection Methods for Artificial Neural Networks, in: Artificial Neural Networks - Methodological Advances and Biomedical Applications. InTech. doi:10.5772/16004

MEIS, 2012. KOoreg: Regional Air Pollution Control Prognostic and Diagnostic Modelling System [WWW Document]. URL http://www.kvalitetazraka.si (accessed 11.21.16).

Messina, P., D’Isidoro, M., Maurizi, A., Fierli, F., 2011. Impact of assimilated observations on improving tropospheric ozone simulations. Atmos. Environ. 45, 6674–6681. doi:10.1016/j.atmosenv.2011.08.056

Mircea, M., Ciancarella, L., Briganti, G., Calori, G., Cappelletti, A., Cionni, I., Costa, M., Cremona, G., D’Isidoro, M., Finardi, S., Pace, G., Piersanti, A., Righini, G., Silibello, C., Vitali, L., Zanini, G., 2014. Assessment of the AMS-MINNI system capabilities to simulate air quality over Italy for the calendar year 2005. Atmos. Environ. 84, 178–188. doi:10.1016/j.atmosenv.2013.11.006

Mlakar, P., 1997. Determination of features for air pollution forecasting models. Intel. Inf. Syst. 1997. IIS ’97. Proc. 350–354. doi:10.1109/IIS.1997.645291

Mlakar, P., Božnar, M., 2011. Artificial Neural Networks - a Useful Tool in Air Pollution and Meteorological Modelling, in: Advanced Air Pollution. InTech. doi:10.5772/20824

Mlakar, P., Božnar, M., 1997. Perceptron neural network-based model predicts air pollution. Intell. Inf. Syst. 1997. IIS ’97. Proc. doi:10.1109/IIS.1997.645288

Mlakar, P., Božnar, M.Z., Grašič, B., Tinarelli, G., 2012. Zasavje canyon regional online air pollution modelling system in highly complex terrain - Description and validation. Int. J. Environ. Pollut. 50, 22–30.

Park, S.-Y., Lee, S.-H., Lee, H.W., 2014. Assimilation of wind profiler observations and its impact on three-dimensional transport of ozone over the Southeast Korean Peninsula. Atmos. Environ. 99, 660–672. doi:10.1016/j.atmosenv.2014.09.082

Pelliccioni, A., Tirabassi, T., 2006. Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. Environ. Model. Softw. 21, 539–546. doi:10.1016/j.envsoft.2004.07.015

Petelin, D., Mlakar, P., Božnar, M.Z., Grašič, B., Kocijan, J., 2015. Ozone forecasting using an online updating Gaussian-process model. Int. J. Environ. Pollut. 57. doi:10.1504/IJEP.2015.074494

Porter, P.S., Rao, S.T., Hogrefe, C., Gego, E., Mathur, R., 2015. Methods for reducing biases and errors in regional photochemical model outputs for use in emission reduction and exposure assessments. Atmos. Environ. 112, 178–188. doi:10.1016/j.atmosenv.2015.04.039

Pugh, D.T., 1996. Tides, surges and mean sea-level (reprinted with corrections), Marine and

Petroleum Geology. John Wiley & Sons Ltd. doi:10.1016/0264-8172(88)90013-X

QualeAria, 2016. Prototype of Air Quality Forecasting System for Italian Territory [WWW Document]. AriaNet Srl. ENEA (Italian Natl. Agency New Techonologies, Energy Sustain. Econ. Dev. Italy. URL http://www.aria-net.eu/QualeAria (accessed 8.16.16).

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E., 2006. Statistical models to assess the health effects and to forecast ground-level ozone. Environ. Model. Softw. 21, 547–558. doi:10.1016/j.envsoft.2004.12.002

SEA Data portal, 2016. SEA Data portal: Public data portal [WWW Document]. Slov. Environ. Agency. URL http://www.arso.gov.si/en/air/data/amp/ (accessed 8.16.16).

Stein, M.L., 1999. Statistical interpolation of spatial data: Some theory for kriging. Springer Science & Business Media.

Zanini, G., Pignatelli, T., Monforti, F., Vialetto, G., Vitali, L., Brusasca, G., Calori, G., Finardi, S., Radice, P., Silibello, C., 2005. The MINNI Project : An Integrated Assessment Modeling System For Policy Making, in: MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand. pp. 2005–2011.

Zoogman, P., Jacob, D.J., Chance, K., Worden, H.M., Edwards, D.P., Zhang, L., 2014. Improved monitoring of surface ozone by joint assimilation of geostationary satellite observations of ozone and CO. Atmos. Environ. 84, 254–261. doi:10.1016/j.atmosenv.2013.11.048