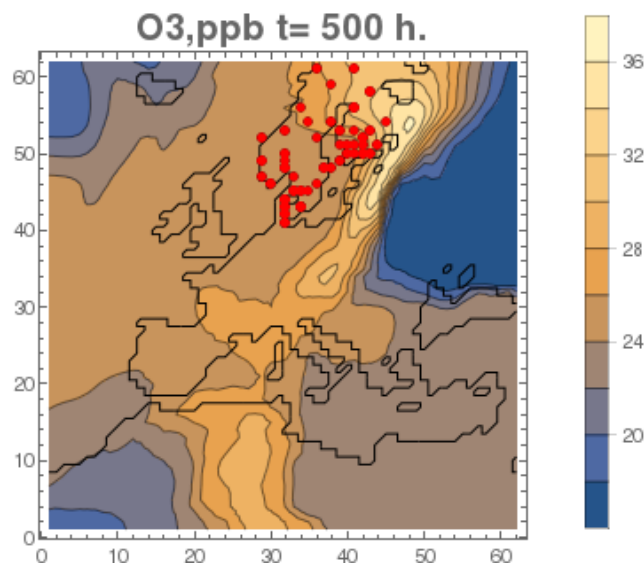




## Scientific Report 14-03

# Chemical Data Assimilation Experiment of Real Data based on Online Integrated Enviro-HIRLAM Model Output

*Alexey Penenko<sup>1,2</sup>, Vladimir Penenko<sup>1</sup>, Roman Nuterman<sup>3,4,5</sup>,  
Alexander Baklanov<sup>3,6</sup>, Alexander Mahura<sup>3</sup>*



<sup>1</sup> Institute of Computational Mathematics and Mathematical Geophysics SB RAS, ICM&MG SB RAS, prospekt Akademika Lavrentjeva 6, 630090, Novosibirsk, Russia

<sup>2</sup> Novosibirsk State University, NSU, Pirogova Str. 2, 630090, Novosibirsk, Russia

<sup>3</sup> Danish Meteorological Institute, DMI, Lyngbyvej 100, DK-2100, Copenhagen, Denmark

<sup>4</sup> University of Copenhagen, Nørregade 10, PO Box 2177, 1017 Copenhagen K, Denmark

<sup>5</sup> Tomsk State University, TSU, Lenin Ave., 36, 634050, Tomsk, Russia

<sup>6</sup> World Meteorological Organization, WMO, 7bis av. la Paix, CP2300, CH-1211, Geneva, Switzerland



# Colophon

**Serial title:**  
Scientific Report **14-03**

**Title:**  
Chemical data assimilation experiment of real data based on online integrated Enviro-HIRLAM model output

**Subtitle:**  
---

**Author(s):**  
Alexey Penenko, Vladimir Penenko, Roman Nuterman, Alexander Baklanov, Alexander Mahura

**Other contributors:**  
---

**Responsible institution:**  
Danish Meteorological Institute

**Language:**  
English

**Keywords:**  
Discrete-Analytical schemes, splitting method, fine-grained data assimilation, chemical data assimilation, variational approach, convection-diffusion-reaction model

**Url:**  
[www.dmi.dk/dmi/sr14-03.pdf](http://www.dmi.dk/dmi/sr14-03.pdf)

**Digital ISBN:**  
978-87-7478-657-3 (on-line)

**ISSN:**  
1399-1949 (on-line)

**Version:**  
-

**Website:**  
[www.dmi.dk](http://www.dmi.dk)

**Copyright:**  
Danish Meteorological Institute  
Application and publication of data and text is allowed with proper reference and acknowledgement



## Content:

Abstract .....	4
1. Introduction.....	4
2. Data Assimilation Algorithm.....	5
2.1. Transport and Transformation Model.....	5
2.2. Fine-Grained Data Assimilation to the Model.....	7
3. Input Data Analysis.....	10
4. Chemical Data Assimilation .....	12
5.1. Data Assimilation Scenario.....	12
5.2. Data Assimilation Results.....	13
5.2.1. Site-wise division.....	13
5.2.2. Temporal difference.....	17
5.2.3. Species exclusion experiment.....	18
Conclusion .....	18
Acknowledgements.....	19
References.....	19

## Abstract

Results of numerical experiments with chemical data assimilation algorithm of *in situ* concentration measurements on real data scenario have been presented. The algorithm is based on the variational approach and splitting scheme. This allows avoiding iterative direct problems solution for transport and transformation model and the algorithm becomes a “real-time algorithm”. In order to construct test scenario, meteorological data has been taken from Enviro-HIRLAM output, initial conditions from MOZART model output and measurements from Airbase database.

## 1. Introduction

We consider the following classes of problems associated to the inverse modeling:

- **Direct problems:** System’s behavior has to be forecasted and studied with a mathematical model.
- **Inverse problems:** Model parameters must be adjusted to fit model forecasts to the corresponding measurement data. It may take to solve series of direct problems with various model parameters.
- **Data assimilation problems:** A forecast has to be improved (on-line) by adjusting model parameters with incoming measurement data. It may take to solve series of inverse problems with various measurement data.

In this work we present data assimilation algorithm for convection-diffusion part of atmospheric chemistry model. To construct a data assimilation algorithm, the following properties should be taken into account:

- Atmospheric composition is being changed rapidly, therefore current and future system state is of interest. Chemical weather forecast in “Real-time”.
- Stiff chemical kinetics equations (different time scales), various chemical mechanisms and their nonlinear behavior.
- Uncertainties are not only in initial conditions but also in model coefficients (reaction rates) and in emission rates.
- High dimensionality ( $\approx 10^7$ ) of modern atmospheric chemistry transport models due to high number of spatial variables and different species, imposes requirements to the computational performance.
- Relatively small number of chemical species in a small number of spatial points can be measured.
- Data assimilation algorithms must be embedded in existing models.
- Multi-disciplinary study.

A review and examples of chemical data assimilation algorithms can be found in [1, 2, 3]. Summarizing them, we would like to emphasize that unlike data assimilation in meteorology initial states in the chemical data assimilation are to be “forgotten” due to diffusion process. Meanwhile the emission rates and model coefficients play a significant role as the sources of uncertainty in the chemical data assimilation. In our work we use source-term uncertainty to perform data assimilation.

The report is a continuation of [4]. The current work has the following purposes:

- To compile a complete realistic scenario to evaluate data assimilation algorithms presented in the previous report [4].
- In order to apply a Data Assimilation algorithm to real data, one has to prepare a consistent set of parameters:
  - ✓ Meteorological conditions for transport and transformation of model parameters.
  - ✓ Chemical background for initial and boundary conditions.

- ✓ Measurement data have to be reduced to a common form and divided into assimilated and reference sets.
- The data assimilation algorithm has to be applied to the assimilated set and compared to a reference one.

## 2. Data Assimilation Algorithm

### 2.1. Transport and Transformation Model

Let us consider a spatial-temporal domain:

$$\vec{x} = (x_1, x_2, x_3) \in \Omega = [0, l_1] \times [0, l_2] \times [0, l_3], t \in [0, T], \Omega_T := \Omega \times [0, T].$$

bounded by  $\delta\Omega_T = \delta\Omega \times [0, T]$ . In the domain we consider atmospheric chemistry transport and transformation model for different substances like contaminants, heat, moisture, radiation, etc.

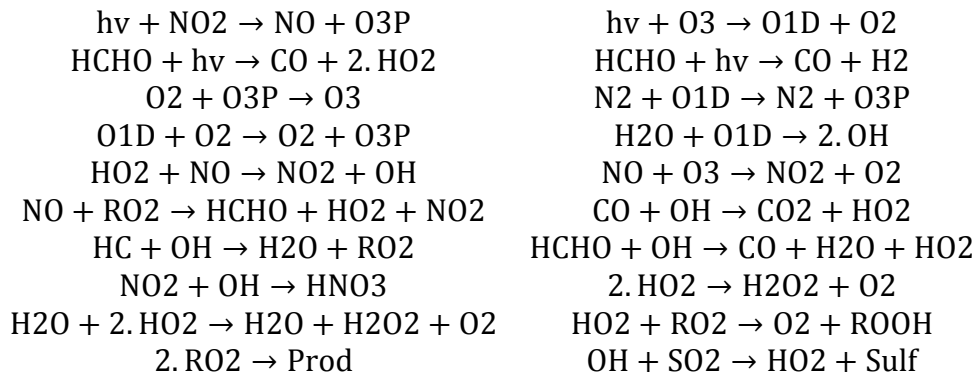
$$\begin{aligned} L\bar{\phi} &\equiv \frac{\partial \bar{\phi}(\vec{x}, t)}{\partial t} + \text{div}(\vec{u} \bar{\phi}(\vec{x}, t) - \mu(\vec{x}, t) \text{grad} \bar{\phi}(\vec{x}, t)) = \\ &= S(\bar{\phi}(\vec{x}, t)) + \bar{f}(\vec{x}, t) + \bar{r}(\vec{x}, t), (\vec{x}, t) \in \Omega_T, \end{aligned} \quad (1)$$

$$\mu(\vec{x}, t) \frac{\partial \bar{\phi}(\vec{x}, t)}{\partial \vec{n}} + \beta(\vec{x}, t) \bar{\phi}(\vec{x}, t) = \bar{g}(\vec{x}, t), (\vec{x}, t) \in \partial\Omega_T, \quad (2)$$

$$\bar{\phi}(\vec{x}, 0) = \bar{\phi}^0(\vec{x}), \vec{x} \in \Omega. \quad (3)$$

Here  $\bar{\phi}(\vec{x}, t)$  is a state function that has physical meaning of concentrations fields at point  $(\vec{x}, t) \in \Omega_T$ , e.g.  $\phi_l(\vec{x}, t)$  corresponds to the concentration of  $l^{\text{th}}$  substance at point  $(\vec{x}, t)$ . Here  $l = 1, \dots, N_c$ ,  $N_c$  is the number of considered substances. Vector  $\vec{u}(\vec{x}, t) = (u_1(\vec{x}, t), u_2(\vec{x}, t), u_3(\vec{x}, t))$  denotes "wind speed",  $\mu(\vec{x}, t) = \text{diag}(\mu_1(\vec{x}, t), \mu_2(\vec{x}, t), \mu_3(\vec{x}, t))$  is a diagonal diffusion tensor,  $S: \mathbb{R}^{N_c} \rightarrow \mathbb{R}^{N_c}$  is a transformation operator,  $\vec{n}$  is the boundary outer normal direction,  $\bar{f}(\vec{x}, t)$ ,  $\bar{g}(\vec{x}, t)$ ,  $\bar{\phi}^0(\vec{x})$  - a priori data for the sources and initial data,  $\bar{r}(\vec{x}, t)$  is a control function (uncertainty), that is introduced in the perfect model structure to assimilate data. As it is for  $\bar{\phi}(\vec{x}, t)$ , each entry of  $\bar{f}(\vec{x}, t)$ ,  $\bar{g}(\vec{x}, t)$ ,  $\bar{r}(\vec{x}, t)$ ,  $\bar{\phi}^0(\vec{x})$  vectors corresponds to a quantity attributed to  $l$ -th substance at point  $(\vec{x}, t)$ .

Transformation operator  $S$  is defined by the chemical kinetics system of 22 reacting species from [5, 6] augmented with the  $SO_2$  reaction taken from the CMAQ model [7]:



Reaction rates have been taken from [5] and depend on time, i.e, photochemistry is considered.

This kinetics system can be presented in the production-destruction operator form:

$$S_l(\bar{\phi}(\vec{x}, t)) = -P_l(\bar{\phi}(\vec{x}, t))\phi_l(\vec{x}, t) + \Pi_l(\bar{\phi}(\vec{x}, t)), l = 1, \dots, N_c, \quad (4)$$

$$P_l, \Pi_l: \mathbb{R}_+^{N_c} \rightarrow \mathbb{R}_+. \quad (5)$$

**Direct problem:** With given  $\vec{f}$ ,  $\vec{g}$ ,  $\vec{\phi}^0$ ,  $\vec{r}$  determine  $\vec{\phi}$  from (1)-(3). Exact solution  $\vec{\phi}_*$  is a solution of direct problem corresponding to "unknown" emissions  $\vec{r}_*$ .

We consider all the functions and model parameters to be smooth enough for the solutions to exist and further transformations to make sense.

For the numerical solution let us introduce uniform temporal grid  $\omega_t = \{t^j\}_{j=1}^{N_t}$  on  $[0, T]$  with step size  $\tau$  and  $N_t$  points and uniform spatial grids  $\omega_\beta$  with  $N_\beta$ ,  $\beta = 1, 2, 3$  grid points on  $\Omega$ ,  $\omega = \omega_1 \times \omega_2 \times \omega_3$ . Let  $Q(\omega)$  be the space of real grid functions on  $\omega$ . Direct problem can be efficiently solved with splitting method. Let us consider additive-averaged splitting scheme (analogous to [8]) on the intervals  $t^j \leq t \leq t^{j+1}$ . The splitting is done with respect to physical process (advection-diffusion and transformation processes) and advection-diffusion part is further split with respect to spatial dimensions. Finally we have 4 parallel stages for the step partition  $\sum_{\beta=1}^4 \gamma_\beta = 1$  and sources partition  $\vec{f} = \sum_{\beta=1}^4 \vec{f}_\beta$ .

- Convection-diffusion processes ( $\beta = 1, 2, 3$ )

$$\begin{aligned} \gamma_\beta \frac{\partial \vec{\phi}_\beta(\vec{x}, t)}{\partial t} + \frac{\partial}{\partial x_\beta} (u_\beta(\vec{x}, t) \vec{\phi}_\beta(\vec{x}, t)) - \frac{\partial}{\partial x_\beta} \left( \mu_\beta(\vec{x}, t) \frac{\partial}{\partial x_\beta} \vec{\phi}_\beta(\vec{x}, t) \right) \\ = \vec{f}_\beta(\vec{x}, t) + \vec{r}_\beta(\vec{x}, t), \quad (\vec{x}, t) \in \Omega \times [t^{j-1}, t^j], \\ \mu(\vec{x}, t) \frac{\partial \vec{\phi}_\beta(\vec{x}, t)}{\partial \vec{n}} + \beta(\vec{x}, t) \vec{\phi}_\beta(\vec{x}, t) = \vec{g}_a(\vec{x}, t), \quad (\vec{x}, t) \in \partial\Omega_\beta \times [t^{j-1}, t^j], \\ \vec{\phi}_\beta(\vec{x}, t^{j-1}) = \vec{\phi}(\vec{x}, t^{j-1}), \quad \vec{x} \in \Omega, \end{aligned}$$

where  $\partial\Omega_\beta = \{\vec{x} \in \partial\Omega \mid x_\beta = 0 \mid x_\beta = l_\beta\}$ . This initial value problem can be approximated with implicit matrix form:

$$\gamma_\beta \frac{\vec{\phi}_\beta^j - \vec{\phi}^{j-1}}{\tau} + L_\beta \vec{\phi}_\beta^j = \vec{r}_\beta^j + \vec{f}_\beta^j, \quad (6)$$

$$L_\beta(\vec{\phi}) := \{L_\beta(\vec{\phi}_l)\}_{l=1}^{N_c}. \quad (7)$$

Here  $\vec{\phi}^j \in Q(\omega)^{N_c}$  stands for the solution on the  $j$ -th time layer,  $\vec{r}^j \in Q(\omega)^{N_c}$  is the uncertainty on the  $j$ -th time layer and  $L_\beta: Q(\omega) \rightarrow Q(\omega)$  are approximated advection-diffusion operators from (1) corresponding to spatial dimensions.

- Chemical reaction processes ( $\beta = 4$ )

$$\begin{aligned} \gamma_\beta \frac{\partial \vec{\phi}_\beta(\vec{x}, t)}{\partial t} + \text{diag}(\vec{P}(\vec{\phi}_\beta(\vec{x}, t))) \vec{\phi}_\beta(\vec{x}, t) \\ = \vec{\Pi}(\vec{\phi}_\beta(\vec{x}, t)) + \vec{f}_\beta(\vec{x}, t) + \vec{r}_\beta(\vec{x}, t), \\ (\vec{x}, t) \in \Omega \times [t^{j-1}, t^j], \\ \vec{\phi}_\beta(\vec{x}, t^j) = \vec{\phi}(\vec{x}, t^j), \quad \vec{x} \in \Omega, \end{aligned}$$

or in the entry-wise form

$$\begin{aligned} \gamma_\beta \frac{\partial \phi_{\beta l}(\vec{x}, t)}{\partial t} + P_l(\vec{\phi}_\beta(\vec{x}, t)) \phi_{\beta l}(\vec{x}, t) \\ = \Pi_l(\vec{\phi}_\beta(\vec{x}, t)) + f_{\beta l}(\vec{x}, t) + r_{\beta l}(\vec{x}, t), \\ (\vec{x}, t) \in \Omega \times [t^{j-1}, t^j], \quad l = 1, \dots, N_c. \end{aligned}$$

where  $P_l$  is the destruction rate functional and  $\Pi_l$  is the production functional. In [9,10,11] a family of unconditionally monotonic schemes have been built, from the first to fourth order of accuracy. One of the single stage schemes is equivalent to the known QSSR scheme [12]:

$$\phi_{\beta l}^j(\vec{p}) = A_l(\vec{\phi}^{j-1}(\vec{p})) + B_l(\vec{\phi}^{j-1}(\vec{p})) r_{\beta l}^j(\vec{p}), \quad l = 1, \dots, N_c, \quad \vec{p} \in \omega,$$

$$A_l(\vec{\phi}^{j-1}(\bar{p})) = \phi_l^{j-1}(\bar{p})e^{-P_l(\vec{\phi}^{j-1}(\bar{p}))\tau}$$

$$+ \frac{1 - e^{-P_l(\vec{\phi}^{j-1}(\bar{p}))\tau}}{P_l(\vec{\phi}^{j-1}(\bar{p}))\tau} (\Pi_l(\vec{\phi}^{j-1}(\bar{p})) + f_{\beta l}^j(\bar{p}))\tau,$$

$$B_l(\vec{\phi}^{j-1}(\bar{p})) = \frac{1 - e^{-P_l(\vec{\phi}^{j-1}(\bar{p}))\tau}}{P_l(\vec{\phi}^{j-1}(\bar{p}))\Delta t} \tau.$$

In vector form

$$\vec{\phi}_\beta^j = \vec{A}(\vec{\phi}^{j-1}) + \text{diag}(\vec{B}(\vec{\phi}^{j-1}))\vec{\gamma}_\beta^j. \quad (8)$$

- Next step approximation

$$\vec{\phi}^j = \sum_{\beta=1}^4 \gamma_\beta \vec{\phi}_\beta. \quad (9)$$

Advantage of the scheme is that individual processes for each dimension are evaluated independently (in parallel).

## 2.2. Fine-Grained Data Assimilation to the Model

In order to assimilate measurement data, we have to connect measured quantities with the model variables. This is formally done with the measurement operator  $H$ :

$$I(t) = H(t, \vec{\phi}^*(., t)) + \vec{\eta}(t), \quad t \in [0, T], \quad (10)$$

where  $I(t)$  are measurement data,  $\vec{\phi}^*(., t)$  is "true" (or exact) solution,  $\vec{\eta}(t)$  is measurement data uncertainty.

**Data assimilation problem:** Determine  $\phi(., t)$  for  $t > t^*$  with (1)-(3), (10) and functions  $f_a, g_a, \phi_a^0, I$  defined on  $0 < t \leq t^*$ .

In the work we consider  $N_M$  *in situ* measurements at the domain grid points  $\{(\bar{x}_M^m, t_M^m)\}_{m=1}^{N_M} \subset \omega \times \omega_t$ . Hence the  $m$ -th measurement is defined by the vector

$$\xi_m = \{(\bar{x}_M^m, t_M^m, l_M^m, I_m, \sigma_M^m)\}, \quad m = 1, \dots, N_M.$$

where  $\bar{x}_M^m$  is the spatial coordinate of the measurement,  $t_M^m$  is the moment of measurement,  $l_M^m$  is the number of substance measured,  $I_m$  is the resulting concentration and  $\sigma_M^m$  is the standard variation of the measurement. According to the data assimilation problem statement in a time-step  $t^j$  we can use only measurements with  $t_M^m \leq t^j$ . Let us define the set of indices

$$\theta^j = \{1 \leq m \leq N_M | t_M^m = t^j\}.$$

The corresponding measurement operator

$$H^j \vec{\phi} = \{\phi_{l_M^m}(\bar{x}_M^m, t_M^m)\}_{m \in \theta^j},$$

$$I^j = \{I_m\}_{m \in \theta^j}, \quad \sigma^j = \{\sigma_M^m\}_{m \in \theta^j}.$$

A function  $\vec{\eta}(t)$  is from a set of admissible values that describe error estimate for measurement data. The error  $\vec{\eta}$  is considered to be bounded in (weighted) Euclidean norm in the measurements space

$$\|\vec{\eta}(t)\|_{\sigma^j} = \sqrt{\sum_{m \in \theta^j} \left(\frac{\eta_m}{\sigma_M^m}\right)^2} \leq \delta_{\vec{\eta}}.$$

Variational data assimilation provides the solution to a data assimilation problem as the minimum of the functional with the constraints imposed by the model. The functional usually combines measurement data misfit with a norm of a control variable:

$$J^j(\vec{\phi}, r) = \|H^j \vec{\phi} - I^j\|_{\sigma_j}^2 + \alpha \|\vec{r}\|^2,$$

where  $\|\cdot\|$  is the norm of a Hilbert space over  $Q(\omega)^{N_c}$  and  $\langle \cdot, \cdot \rangle$  is the corresponding inner product,  $\alpha$  is the regularization (assimilation parameter), which selects whether the solution will be closer to the direct model solution or will reproduce measurements better. In the paper on the time step  $t^j$  we update only the control variable  $\vec{r}^j$  for this time step. In the context of assimilating data to the split model we distinguish between the two approaches:

- In conventional approach [13, 14] one takes minimum of  $J^j(\vec{\phi}^j, \vec{r}^j)$  as the solution with constraints (6), (8), (9) and  $\vec{r}_\beta = \gamma_\beta \vec{r}$ , i.e., control functions of different splitting stages are connected.
- In the fine-grained approach [15, 16, 17] the same data are assimilated to different parts of model and the results are coupled afterwards. We seek for the minimum of the functional

$$J_f^j(\{\vec{\phi}_\beta^j, \vec{r}_\beta^j\}_{\beta=1}^4) = \sum_{\beta=1}^4 J^j(\vec{\phi}_\beta^j, \vec{r}_\beta^j)$$

on constraints (6), (8) with the independent  $\vec{r}_\beta$ .

Using the method of Lagrange multipliers to solve minimization problem with equality constraints, we can construct augmented functional:

$$\bar{J}_f(\{\vec{\phi}_\beta^j, \vec{r}_\beta^j\}_{\beta=1}^4) = \sum_{\beta=1}^4 J^j(\vec{\phi}_\beta^j, \vec{r}_\beta^j) + \sum_{\beta=1}^3 \left\langle \gamma_\beta \frac{\vec{\phi}_\beta^j - \vec{\phi}^{j-1}}{\tau} - L_\beta \vec{\phi}^j - \vec{r}_\beta^j - \vec{f}_\beta^j, \vec{\psi}_\beta^j \right\rangle + \langle \vec{\phi}_4^j - \vec{A}(\vec{\phi}^{j-1}) - \text{diag}(\vec{B}(\vec{\phi}^{j-1})) \vec{r}_4^j, \vec{\psi}_4^j \rangle$$

We can see that components of  $\bar{J}_f(\{\vec{\phi}_\beta^j, \vec{r}_\beta^j\}_{\beta=1}^4)$  corresponding to different  $\beta$  are independent hence stationary point coordinates can be found independently.

In order to present an algorithm of finding a stable point for the convection-diffusion part, we need further elaboration of operator  $L$ . As a result of splitting we can consider equations (6) independent for each coordinate line in both dimensions. The algorithm is the same for any coordinate line and here we will describe the algorithm applied the  $q$ -th grid line along X axis for a given  $1 \leq q \leq N_y$  and  $l$ -th substance field. Let  $\phi_i^j = ((\phi_x)_l)_{iq}^j$ ,  $f_i^j = ((f_x)_l)_{iq}^j$ ,  $r_i^j = ((r_x)_l)_{iq}^j$ ,  $1 \leq i \leq N_x =: N$ . For the sake of computational efficiency we use approximations of (1) that produce tridiagonal matrix systems:

$$-a_i \phi_{i+1}^j + b_i \phi_i^j = \phi_i^{j-1} + \tau r_i^j + \tau f_i^j, i = 0, \quad (11)$$

$$-a_i \phi_{i+1}^j + b_i \phi_i^j - c_i \phi_{i-1}^j = \phi_i^{j-1} + \tau r_i^j + \tau f_i^j, i = 1, \dots, N-1, \quad (12)$$

$$b_i \phi_i^j - c_i \phi_{i-1}^j = \phi_i^{j-1} + \tau r_i^j + \tau f_i^j, i = N. \quad (13)$$

In this term the assimilated state is the solution of the minimization problem

$$J(\phi^j, r^j) \tau = \left( \sum_{i=1}^{N-1} \left( \frac{\phi_i^j - I_i^j}{\sigma_i} \right)^2 M_i^j + \alpha \sum_{i=1}^{N-1} (r_i^j)^2 \right) \tau,$$



WRT (11)-(13) where  $M_i^j$  is the spatial-temporal measurement mask (i.e.  $M_i^j$  equals 1 if  $\bar{x}_{iq} \in X_M^j$  and 0 otherwise, in other words, it is equal to 1 if there is a measurement data at point  $\bar{x}_{iq}$ ),  $I_i^j$  is measurement data at point  $\bar{x}_{iq}$  (if there is a measurement) and  $\sigma_i$  is measurement device standard deviation of the measurement in point  $\bar{x}_{iq}$  (if there is a measurement). Introducing Lagrange multipliers, we obtain augmented functional:

$$\bar{J}_f(\phi^j, r^j, \phi^{*j})\tau = J(\phi^j, r^j)\tau + \sum_{i=0}^{N-1} (-a_i\phi_{i+1}^j + b_i\phi_i^j - c_i\phi_{i-1}^j - \phi_i^{j-1} - \tau r_i^j - \tau f_i^j)\phi_i^{*j}.$$

Taking the first variations of the augmented functional equal to zero, we obtain the following equations:

$$\nabla_{\phi_i^{*j}} \bar{J}_f(\phi^j, r^j, \phi^{*j}) = 0,$$

which is equivalent to (11)-(13).

$$\nabla_{\phi_i^{j+1}} \bar{J}_f(\phi^j, r^j, \phi^{*j}) = 0$$

is equivalent to

$$\begin{aligned} -c_{i+1}\psi_{i+1}^j + b_i\psi_i^j &= -\frac{2M_i^j}{\sigma_i^2}(\phi_i^j - I_i^j)\tau, \quad i = 0, \\ -c_{i+1}\psi_{i+1}^j + b_i\psi_i^j - a_{i-1}\psi_{i-1}^j &= -\frac{2M_i^j}{\sigma_i^2}(\phi_i^j - I_i^j)\tau, \quad i = 1, \dots, N-1, \\ b_i\psi_i^j - a_{i-1}\psi_{i-1}^j &= -\frac{2M_i^j}{\sigma_i^2}(\phi_i^j - I_i^j)\tau, \quad i = N. \end{aligned}$$

and

$$\nabla_{r_i^{j+1}} \bar{J}_f(\phi^j, r^j, \psi^j) = 0$$

is equivalent to

$$2\alpha r_i^j - \psi_i^j = 0, \quad i = 0, \dots, N.$$

The systems obtained can be merged into tridiagonal matrix equation [15,16,17]

$$\begin{aligned} -A_i\Phi_{i+1}^j + B_i\Phi_i^j &= F_i^j, \quad i = 0, \\ -A_i\Phi_{i+1}^j + B_i\Phi_i^j - C_i\Phi_{i-1}^j &= F_i^j, \quad i = 1, \dots, N-2, \\ B_i\Phi_i^j - C_i\Phi_{i-1}^j &= F_i^j, \quad i = N-1, \end{aligned}$$

$$A_i = \begin{pmatrix} a_i & 0 \\ 0 & c_{i+1} \end{pmatrix}, \quad B_i = \begin{pmatrix} b_i & -\frac{\tau}{2\alpha} \\ \frac{2M_i\tau}{\sigma_i^2} & b_i \end{pmatrix}, \quad C_i = \begin{pmatrix} c_i & 0 \\ 0 & a_{i-1} \end{pmatrix},$$

$$\Phi_i^j = \begin{pmatrix} \phi_i^j \\ \psi_i^j \end{pmatrix}, \quad F_i^{j+1} = \begin{pmatrix} \phi_i^{j-1} + \tau f_i^j \\ \frac{2M_i\tau}{\sigma_i^2} I_i^j \end{pmatrix},$$

which is solved by the direct matrix sweep method.

For the transformation step data assimilation the algorithm is the same for any grid point  $\bar{p} \in \omega$ . For brevity let  $\vec{\phi}^j = \vec{\phi}_4^j(\bar{p}) \in \mathbb{R}^{N_c}$ ,  $\vec{r}^j = \vec{r}_4^j(\bar{p}) \in \mathbb{R}^{N_c}$ ,  $\vec{\psi}^j = \vec{\psi}_4^j(\bar{p}) \in \mathbb{R}^{N_c}$  and the

result is sought as the stationary point of the augmented functional

$$\bar{J}(\vec{\phi}^j, \vec{r}^j) = \sum_{l=1}^{Nc} \left( \frac{\phi_l^j - I_l^j}{\sigma_l} \right)^2 M_l^j + \alpha \sum_{l=1}^{Nc} (r_l^j)^2 + \sum_{l=1}^{Nc} \left( \phi_l^j - \phi_l^{j-1} e^{-P_l(\vec{\phi}^{j-1})\Delta t} - \frac{1 - e^{-P_l(\vec{\phi}^{j-1})\Delta t}}{P_l(\vec{\phi}^{j-1})\Delta t} (\Pi_l(\vec{\phi}^{j-1}) + r_l^j)\tau \right) \psi_l^j.$$

where  $M_l^j$  is equal to 1 if l-th substance is measured at point  $\vec{p}$  at moment  $t^j$  and zero otherwise. This minimum is given by the formula:

$$\phi_l^j = \frac{1}{1+Z} \left( \phi_l^{j-1} e^{-P_l(\vec{\phi}^{j-1})\tau} + \frac{1 - e^{-P_l(\vec{\phi}^{j-1})\tau}}{P_l(\vec{\phi}^{j-1})\tau} \Pi_l(\vec{\phi}^{j-1})\tau \right) + \frac{Z}{1+Z} I_l^j,$$

$$Z = \frac{M_l^j}{\alpha(\sigma_l)^2} \left( \frac{1 - e^{-P_l(\vec{\phi}^{j-1})\tau}}{P_l(\vec{\phi}^{j-1})\tau} \right)^2.$$

As we can see, the resulting algorithm can be implemented without iterations.

### 3. Input Data Analysis

In our model we use ppb as basic concentration units. To compile the scenario to test Data Assimilation, we have used the following datasets:

- Airbase - European atmospheric composition measurement data [18]. We have used data from 22 Scandinavian (DK, SE, FI, NO) measurement sites. There are 64001 in total  $SO_2, O_3, NO_2, CO, NO$  measurements for the considered period of time. Concentration data for substances  $SO_2, O_3, NO_2, NO$  is given in  $\mu\text{g}/\text{m}^3$  and for  $CO$  in  $\text{mg}/\text{m}^3$ . Conversion formula

$$\text{Concentration} = 8314 \frac{\text{Temp} \cdot \text{conc}}{\text{MW} \cdot \text{Pres}},$$

where Concentration is concentration in [ppb],  $\text{conc}$  is concentration [ $\mu\text{g}/\text{m}^3$ ],  $\text{Temp}$  is temperature [K],  $\text{MW}$  is molecular weight in [ams],  $\text{Pres}$  is pressure in [Pa]. And

$$\text{Concentration} = 1000 * 8314 \frac{\text{Temp} \cdot \text{conc}}{\text{MW} \cdot \text{Pres}},$$

where  $\text{conc}$  is concentration [ $\text{mg}/\text{m}^3$ ].

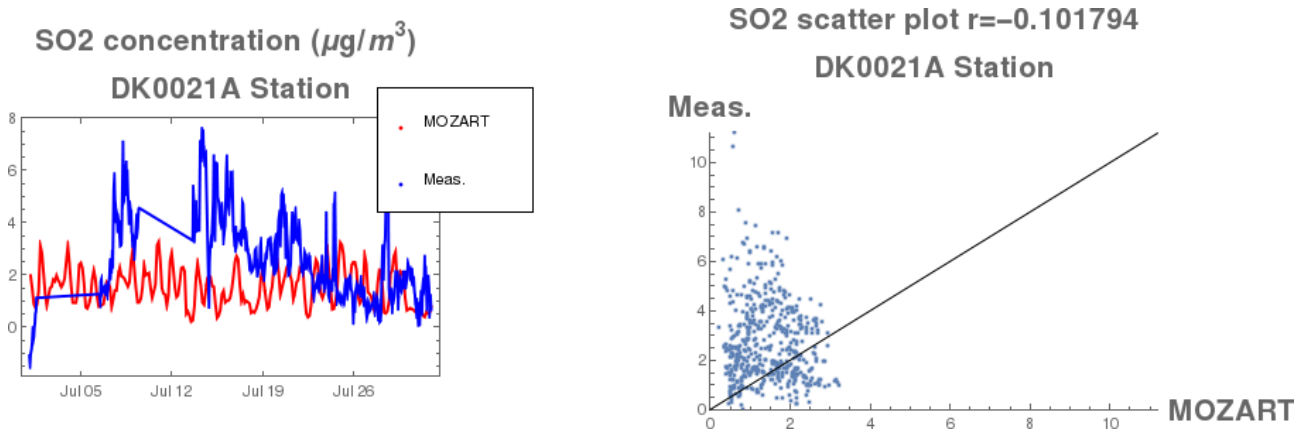
- MOZART IFS data is used as initial conditions. Dataset contains  $O_3, NO, NO_2, HNO_3, OH, H_2O_2, CO, SO_2$  concentration fields in [mole/mole] Conversion formula [19]:

$$\text{Concentration} = \text{vmr} * 10^9.$$

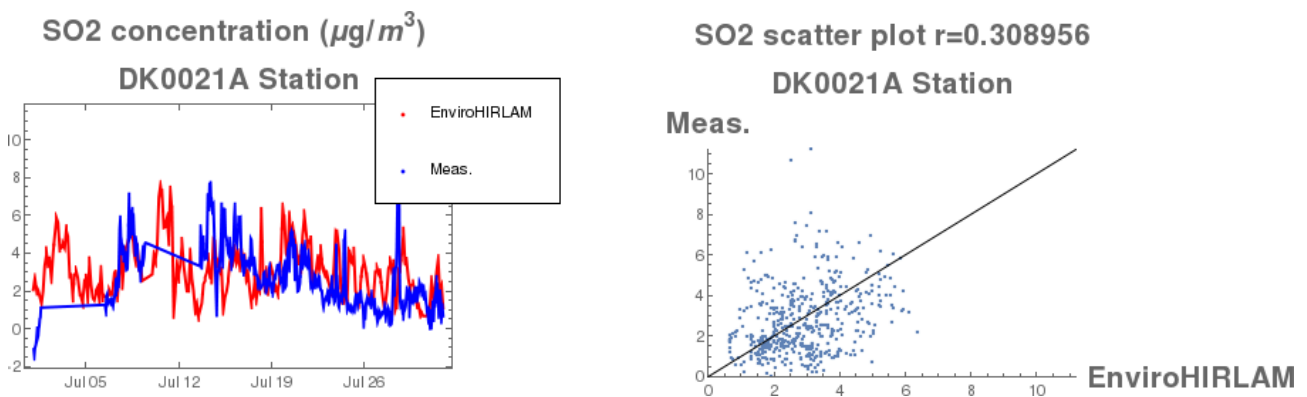
where Concentration is concentration in [ppb],  $\text{vmr}$  is volume mixture ratio in [mole/mole].

- Enviro-HIRLAM output has been used for meteorological data [20].

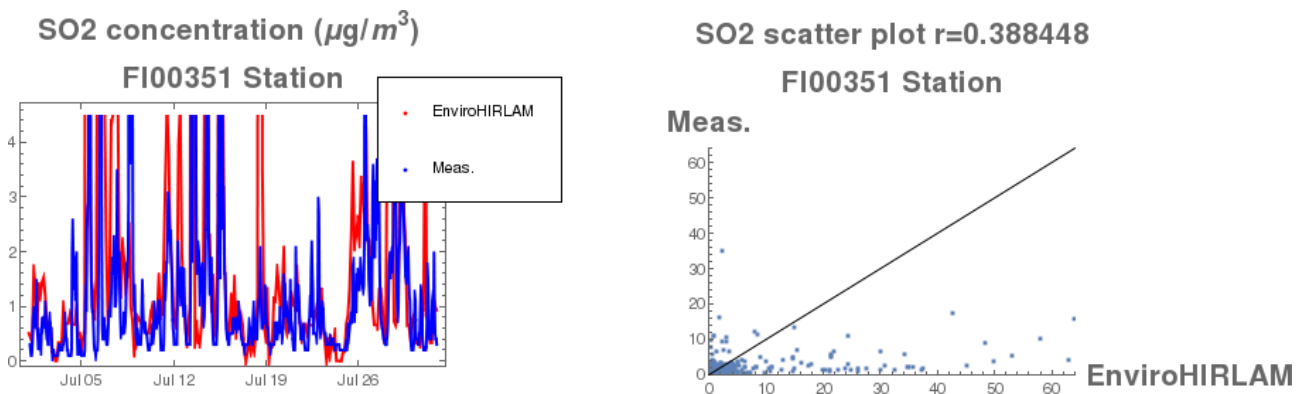
In order to check data consistency, we have conducted some basic correlation study. We've chosen  $SO_2$  since its concentrations were present in both datasets from MOZART and Enviro-HIRLAM. In figures 1 and 2 we can see that models performance can vary. In this case Enviro-HIRLAM did better estimate ( $r = 0.31$ ) than MOZART ( $r = -0.1$ ) for measurement site DK0021A. Time series are presented to the left and scatter plots to the right.



**Figure 1.** SO<sub>2</sub> Time series comparison at the measurement site (left). Scatter plot of measurements vs MOZART-IFS model output (right).

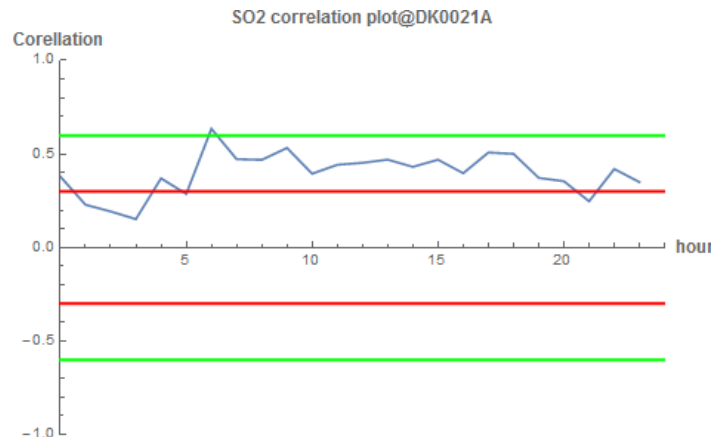


**Figure 2.** SO<sub>2</sub> Time series comparison at a measurement site (left). Scatter plot of measurements vs EnviroHIRLAM model output (right).

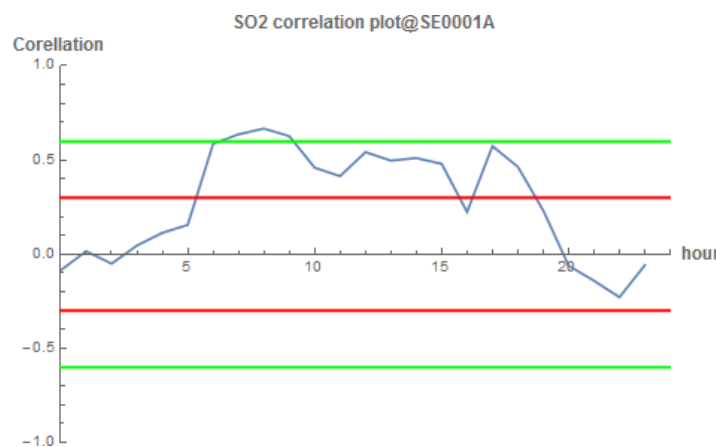


**Figure 3.** SO<sub>2</sub> Time series comparison at a measurement site (left). Scatter plot of measurements vs EnviroHIRLAM model output (right).

The second numerical experiment was devoted to estimation of the correlation with respect to time of a day. In figures 4 and 5 we can see that the correlation is better for the day-time (from 5 am to 8 pm) and slightly drops in the night-time.



**Figure 4.** Measurements vs Enviro-HIRLAM model output correlation with respect to time of day (for July 2010).



**Figure 5.** Measurements vs Enviro-HIRLAM model output correlation with respect to time of day (for July 2010).

## 4. Chemical Data Assimilation

### 5.1. Data Assimilation Scenario

In our numerical experiments we have done the following simplifications with respect to the data assimilation problem statement:

- The model used is a 2D model;
- Zero Neumann boundary conditions are used (to be taken from MOZART);
- Simple transformation mechanism (22 chemical reactions);
- Transformation rates depend only on time of a day (important to account for pressure and temperature), revised parameters will be useful;
- Available emission databases are needed (annual update is desirable for operational use);
- A simplified diffusion coefficient is used.

In the data assimilation, we have used, the domain which is defined by temporal and spatial grids:

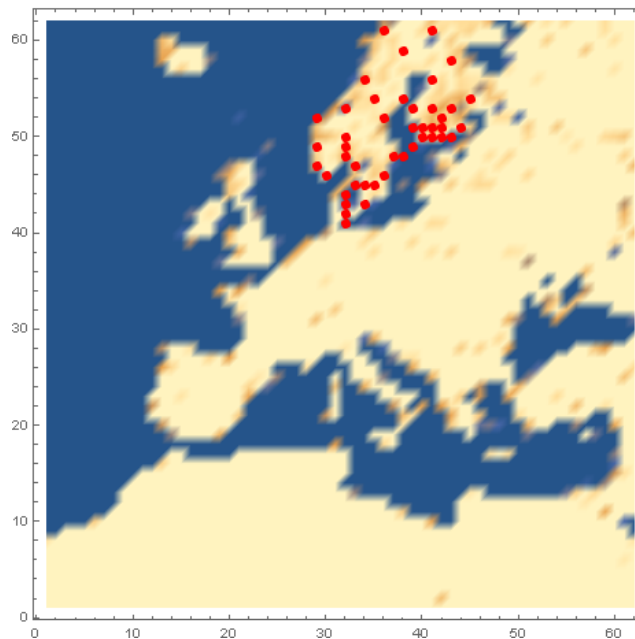
- Temporal grids:
  - Physical time span: 1 July 2010 - 1 August 2010.
  - Temporal grid for transport processes
 
$$Lt = 3720(\text{grids}) * 720(\text{s}) = 31\text{day}.$$
  - Nested temporal grid for transformation processes is 100 times finer:

$$Lt = 372000(\text{grids}) * 7.2(s) = 31\text{day}.$$

- Spatial grids: (5 times coarser than Enviro-HIRLAM grid in each dimension):  
 $Lx = 61(\text{grids}) * 83(\text{km}) = 5086\text{km},$   
 $Ly = 61(\text{grids}) * 81(\text{km}) = 4962\text{km}.$

Transport model parameters have been taken from the Enviro-HIRLAM meteorological output. We have used wind speed at 10m components (U10, V10). Diffusion coefficient is evaluated from the wind speeds.

Measurement Data have been taken from AirBase - the European Air quality database [5]. We have used 64001 SO<sub>2</sub>, CO, NO, O<sub>3</sub>, NO<sub>2</sub> measurements in total from 22 Scandinavian (DK - Denmark, SE - Sweden, FI - Finland, NO - Norway) measurement sites (Fig.6).



*Figure 6. Computation domain and measurement sites locations (red dots).*

To study impact of both transformation and data assimilation processes, we have considered four configurations encoded by:

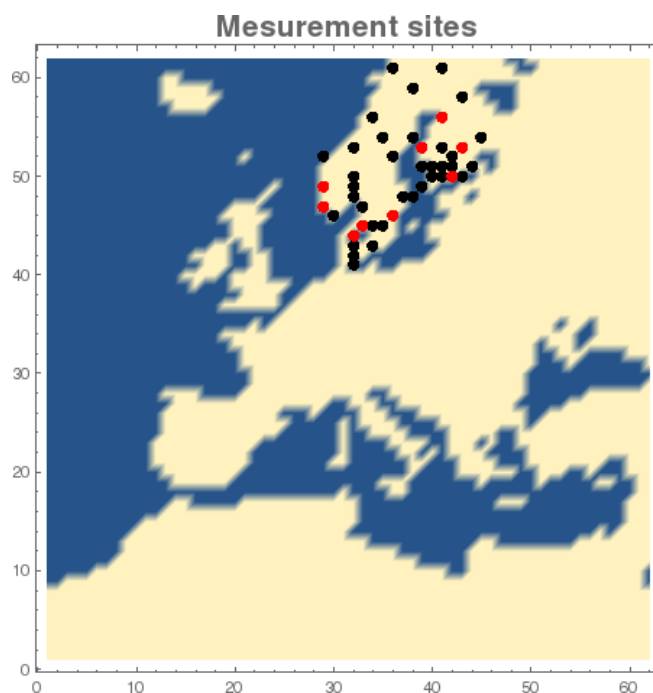
- **DATrspT(rns):** Transformations with data assimilation to transport processes;
- **TrspTrns:** Transformations with transport processes without data assimilation;
- **DATrsp:** Data assimilation to transport processes without transformation;
- **Trsp:** Transport processes (with neither transformation nor data assimilation).

## 5.2. Data Assimilation Results

To compare configurations, some part of measurements are assimilated the rest is used as reference. The choice of division criteria defines a numerical experiment. Results are compared with respect to correlation coefficient and RMSE.

### 5.2.1. Site-wise division

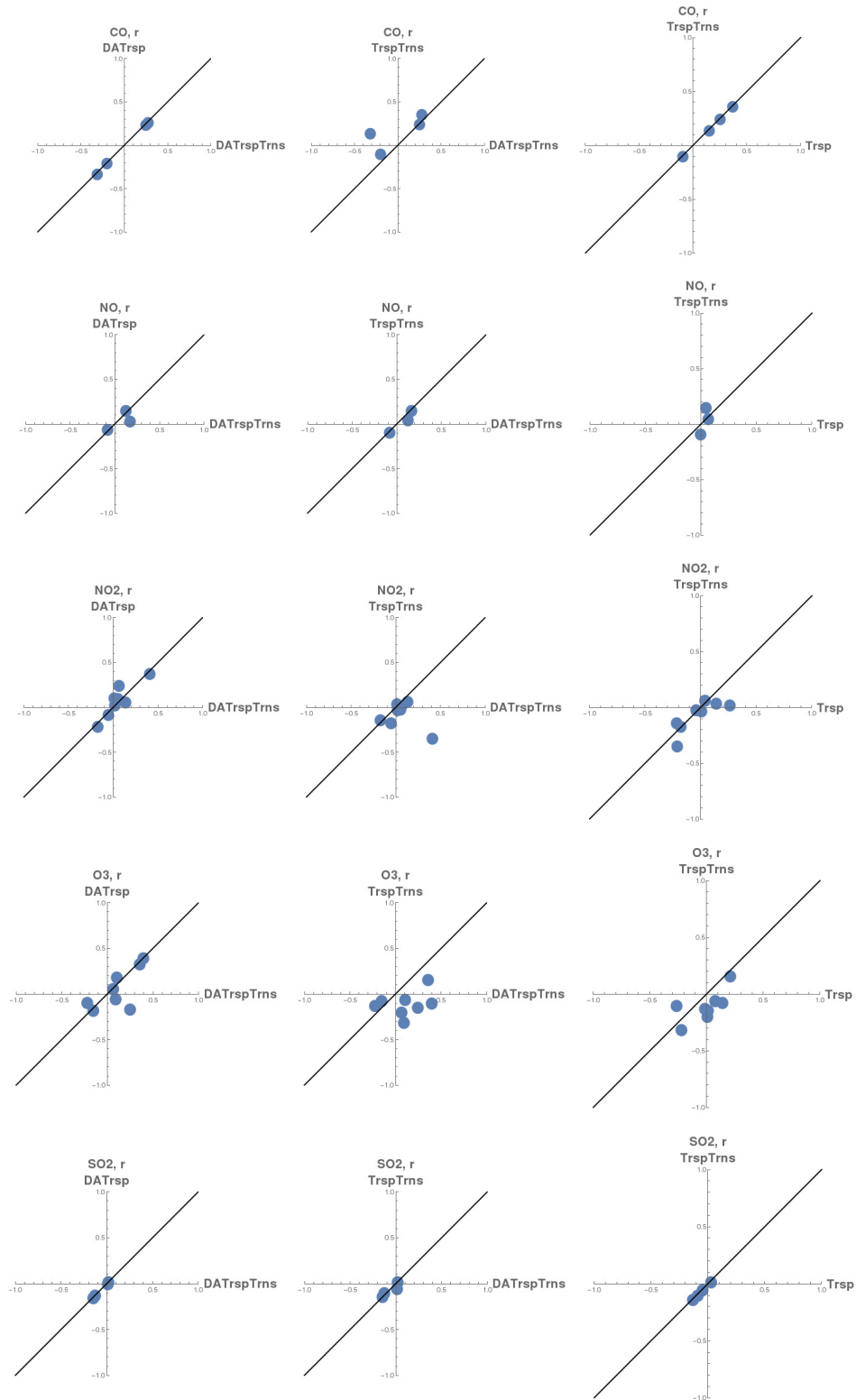
In this experiment we assimilated 80% of data on sites and used the rest of data for reference (Fig.7).



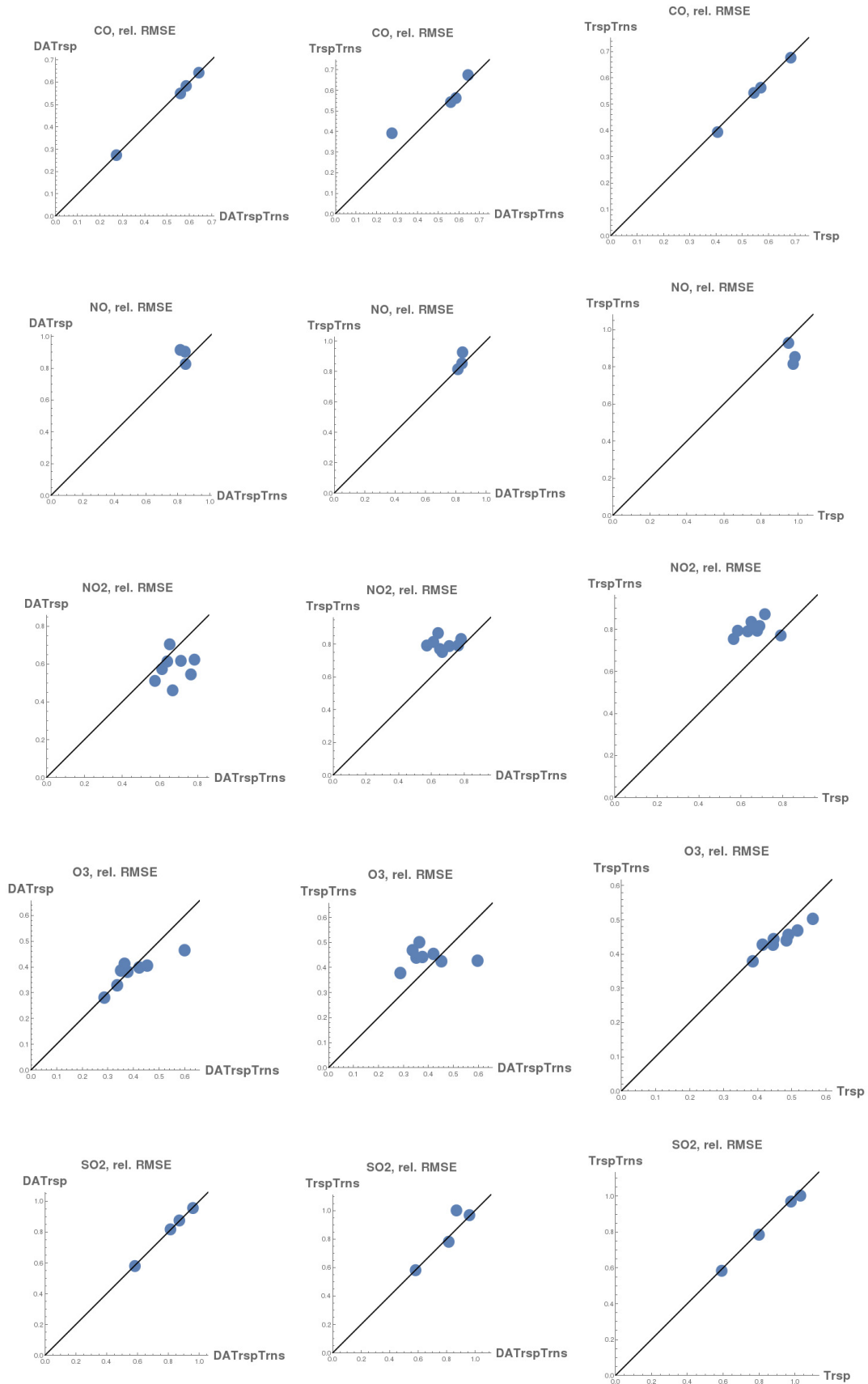
**Figure 7.** Assimilated measurement sites locations (black dots) and reference measurement sites locations (red dots).

In Fig. 8 site-wise comparison of correlation coefficient with reference measurement data for different species is presented. Correlation coefficients of the corresponding configuration have been put to axes named after these configurations. Dots correspond to measurement sites. Left column of graphs compares DATrsp (without transformation) with DATrspTrns for different substances. The more dots are below bisector line the better DATrspTrns works. Central column compares model without data assimilation (TrspTrns) with the model with data assimilation (DATrspTrns). The right column compares the performance of direct models forecast (without DA) for the model with transformation (TrspTrns) and model without (Trsp).

After we have analyzed the figures, we can conclude that the improvement of DA assimilation algorithm can be identified for  $O_3$ . In Figs. 9 we presented site-wise comparison of RMSE with reference measurement data for different species. On the axes we have put RMSE of the corresponding configuration. Analyzing the figures we can conclude that the improvement of DA assimilation algorithm can be identified for  $NO_2$  and  $O_3$ .



**Figure 8.** Comparison of correlation coefficients for different sites (dots) and for different configurations from left to right. Left figure is a comparison of DATrsp and DATrspTrns results (comparison of DA algorithm without transformations and with transformations). Central figure is a comparison of TrspTrns and DATrspTrns (benefit of Data Assimilation). Right figure is a comparison of TrspTrns to Trsp (comparison of direct model win transformations and without).



**Figure 9.** Comparison of relative RMSE for different sites (dots) and for different configurations from left to right. Left figure is a comparison of DATrsp and DATrspTrns results (comparison of DA algorithm without transformations and with transformations). Central figure is a comparison of TrspTrns and DATrspTrns (benefit of Data Assimilation). Right figure is a comparison of TrspTrns to Trsp (comparison of direct model



win transformations and without).

### 5.2.2. Temporal difference

Division criteria: Assimilate data for  $t < T/2$  and compare the model output to measurement data for  $t > T/2$ .

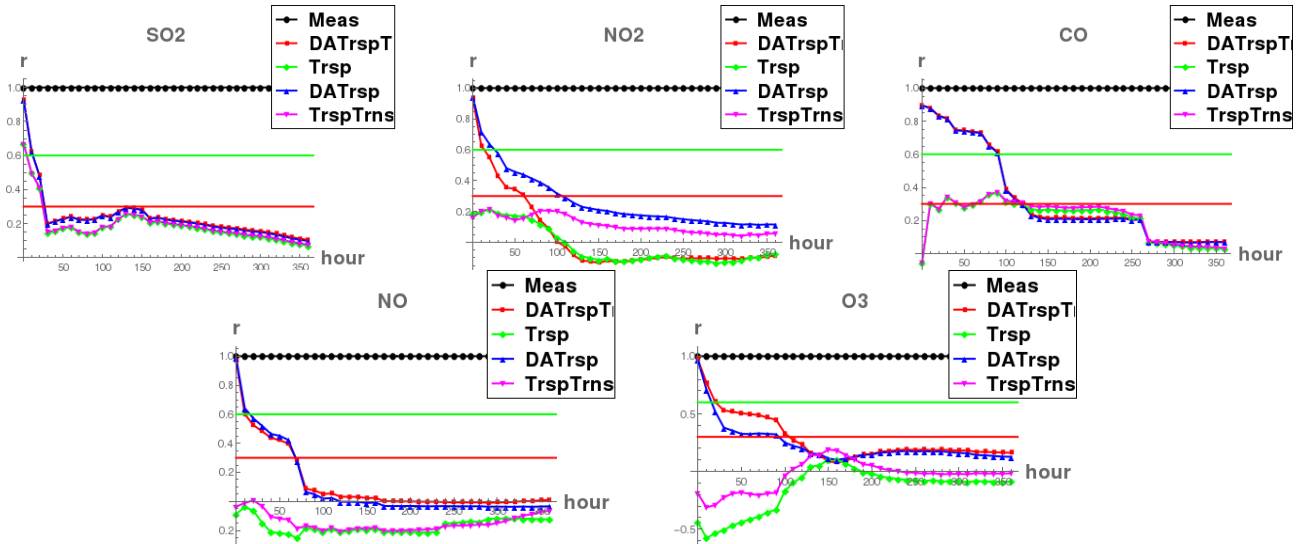


Figure 10. Correlation decrease with time after the last assimilated measurement.

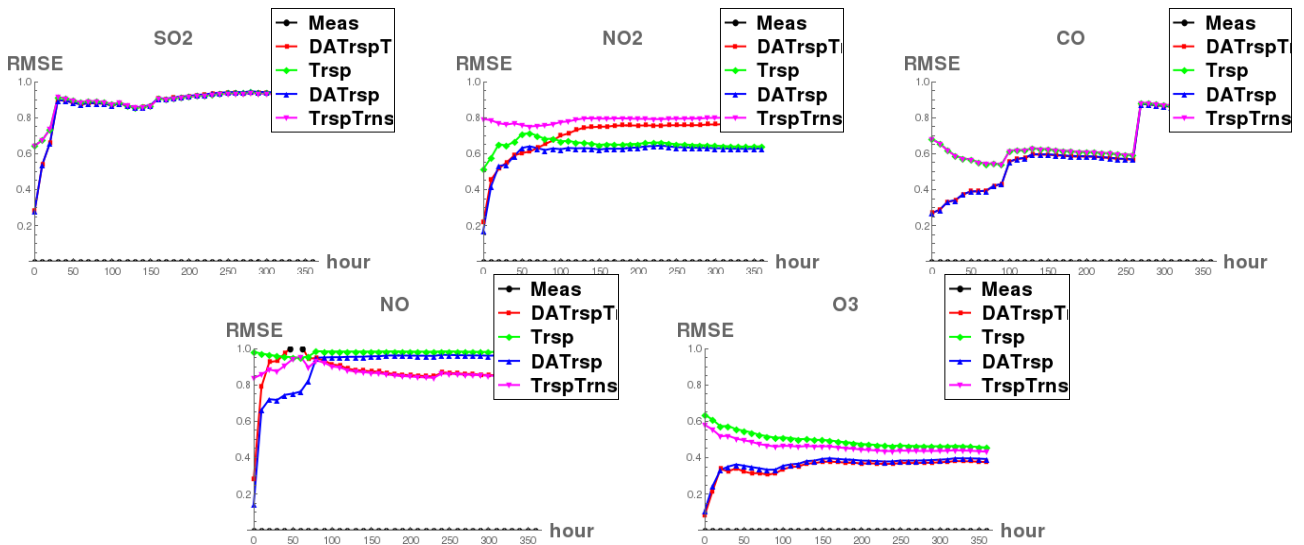


Figure 11. RMSE increase with time after the last assimilated measurement.

In Fig. 10 we presented results of comparison to time extending subsets of data from the reference set. From Fig. 10 we can conclude that the models with data assimilation (red and blue curves) provide better performance as the forecast tool compared to the direct models (Green and Magenta curves) for all the species except for SO<sub>2</sub>. Based on this we can draw a general conclusion on the advantage of data assimilation in this case.

As we can see from the figures, the model with chemistry provides slower degradation of the correlation coefficient for O<sub>3</sub>, while it is reverse for NO<sub>2</sub>. For the rest of substances correlation degradation rate is almost identical.

Looking at RMSE degradation in Fig. 11 we can confirm the conclusion of data assimilation benefits. Comparing DATrsp and DATrspT, we can state that RMSE degradation rate (its increase) for DATrspT for all species is higher than for DATrsp. Probably it is because of more uncertain and nonlinear nature of DATrspT with chemical transformations.

### 5.2.3. Species exclusion experiment

Division criteria: Assimilate all the data except for data of the selected substance.

Species	Meas	DATrspTrns	Trsp	DATrsp	TrspTrns
CO	1.	0.0420364	0.041718	0.041718	0.042036
NO	1.	0.370355	-0.068419	-0.068419	-0.035124
NO <sub>2</sub>	1.	0.458236	0.063321	0.063321	-0.019848
SO <sub>2</sub>	1.	-0.014101	-0.016041	-0.016041	-0.014101
O <sub>3</sub>	1.	0.119529	-0.058031	-0.058031	-0.029600

*Table 1. Comparison of results for different excluded species.*

As we can see from Table 1, the best performance for DA with chemical transformations has been obtained for NO<sub>2</sub> and NO. This can be explained by a strong link between the two species. In case the concentration of one of them is known, the other can be reconstructed. Less improvement has been obtained for O<sub>3</sub> (below significant correlation level of 0.3). For CO and SO<sub>2</sub> the results of all configurations have been almost the same.

## Conclusion

Combination of splitting and data assimilation schemes let us construct computationally effective algorithms for data assimilation of *in-situ* measurements to convection-diffusion models.

A complete data assimilation scenario has been compiled with meteorological data from Enviro-HIRLAM model, initial concentration data from the MOZART model and *in-situ* measurement data from AirBase.

We carried out series of numerical experiments in which we tested data assimilation algorithms on different divisions of measurement data into assimilated and reference datasets. Data assimilation was able to improve modeling results with imperfect (approximate) models and model parameters. The advantage of data assimilation algorithm that includes chemical transformations was identified for ozone concentrations modeling.

Among the future steps to improve data assimilation results we can identify:

- Inclusion of more realistic boundary conditions is required;
- Additional tuning is essential for coefficients of chemical reactions;
- Quality control of chemical data measurements at stations is recommended for excluding "extreme" data;
- Revision of implementation procedure/steps for chemical model is required;
- Additional evaluation of monthly and seasonal variability is needed.



## Acknowledgements

Authors are thankful to DMI colleagues for constructive comments and discussions. Authors appreciate Irina Gerasimova's help with the text of report. The work has been partially supported by COST Action ES1004 (<http://eumtechem.info>) STSM Grant #21654, RFBR 14-01-31482 mol\_a and 14-01-00125, Programs # 4 Presidium Russian Academy of Science (RAS) and # 3 MSD RAS, integration projects Siberian Branch RAS #8 and #35.

## References

- [1] A. Sandu and C. Tianfeng. *Chemical data assimilation - an overview*. Atmosphere, 2:426–463, 2011.
- [2] Baklanov A. et. al. *Online coupled regional meteorology chemistry models in Europe: current status and prospects*. Atmos. Chem. Phys., 14:317–398, 2014.
- [3] Elbern H., Strunk A., Schmidt H., Talagrand O., *Emission rate and chemical state estimation by 4-dimensional variational inversion*. Atmos. Chem. Phys. 2007. Vol.7. P.3749-3769.
- [4] Penenko, A., Penenko, V., Nuterman, R., Mahura, A. *Discrete-Analytical Algorithms for Atmospheric Transport and Chemistry Simulation and Chemical Data Assimilation*. DMI Scientific Report 14-02 (2014) [www.dmi.dk/dmi/sr14-02.pdf](http://www.dmi.dk/dmi/sr14-02.pdf)
- [5] William R. Stockwell and Wendy S. Goliff *Comment on "Simulation of a reacting pollutant puff using an adaptive grid algorithm" by R. K. Srivastava et al.* JOURNAL OF GEOPHYSICAL RESEARCH, 2002 V 107 No D22.
- [6] Gery, M.W. G.Z. Whitten, J.P. Killus, and M.C. Dodge, *A photochemical kinetics mechanism for urban and regional scale computer modeling*, J. Geophys. Res., 94, 12, 952-12,956,1989
- [7] Byun, D. W. and Schere, K. L.: Review of the governing equations, computational algorithms and other components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, Applied Mechanics Reviews, 59(2), 51–77, 2006.
- [8] A.A. Samarskii and P.N. Vabishchevich. *Computational Heat Transfer*. Vol.1,2. Wiley, Chichester, 1995.
- [9] V. Penenko, E. Tsvetova, *Variational methods for construction of monotone approximations for atmospheric chemistry models*, Journal of Computational and Applied Mathematics, 2013, V 6, No 3. 210-220.
- [10] V. Penenko, E. Tsvetova, *Discrete-analytical methods for the implementation of variational principles in environmental applications* Journal of Computational and Applied Mathematics 226 (2009) 319-330
- [11] V.V. Penenko, E.A. Tsvetova, A.V. Penenko., *Variational approach and Euler's integrating factors for environmental studies* Computers and Mathematics with Applications, (2014) V.67, Issue 12, P. 2240–2256.
- [12] Hesstvedt, E.; Hov, O.; Isaacsen, I. *Quasi-steady-state-approximation in air pollution modelling: comparison of two numerical schemes for oxidant prediction*. Int. J. Chem. Kinet. 1978, 10, 971–994.
- [13] V.V. Penenko *Methods of numerical modeling of atmospheric processes*. Hydrometeoizdat (1981), Leningrad, pp.352
- [14] MARCHUK G.I., ZALESNY V.B. *A numerical technique for geophysical data assimilation problems using Pontryagin's principle and splitting-up method* Russian Journal of Numerical Analysis and Mathematical Modeling. 1993. Vol. 8. №. 4. C. 311–326.
- [15] A. Penenko *Some theoretical and applied aspects of sequential variational data assimilation (In Russian)*, Comp. tech. v.11, Part 2, (2006) 35-40.
- [16] V. V. Penenko *Variational methods of data assimilation and inverse problems for studying the atmosphere, ocean, and environment* Num. Anal. and Appl., 2009 V 2 No 4, 341-351.
- [17] Penenko, A.V. and Penenko V.V. (2014), *Direct data assimilation method for convection-diffusion models based on splitting scheme*. Computational technologies, 19 Issue 4, 69 (In Russian).
- [18] <http://acm.eionet.europa.eu/databases/airbase>.
- [19] Daniel J. Jacob, Introduction to Atmospheric Chemistry, Princeton University Press, 1999, <http://acmg.seas.harvard.edu/people/faculty/djj/book/bookchap1.html>
- [20] Nuterman R. et. al *High resolution forecast meteorology and chemistry for a Danish Urban Area* EMS 2011.