19th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
3-6 June 2019, Bruges, Belgium

APPLICATION AND INTERCOMPARISON OF THREE DATA ASSIMILATION METHODS FOR THE EVALUATION OF AIR QUALITY ON THE ÎLE-DE-FRANCE AREA

Chi Vuong Nguyen[1], Lionel Soulhac[1], Perrine Charvolin[1], Cyril Joly[2] and Olivier Sanchez[2]

[1]Laboratoire de Mécanique des Fluides et d'Acoustique, Université de Lyon, CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Claude Bernard Lyon I, Ecully, France
[2]Airparif, Paris, France

Abstract: For several years, data assimilation techniques have been used in the field of air quality to combine the results of numerical models with field measurements, in order to produce the best possible estimate of the pollutants concentration field. Although these methods have been used extensively on a regional scale (Blond *et al.*, 2003; Silver *et al*., 2012; Zhang *et al*., 2012), they have rarely been applied on an urban scale, using models taking buildings into account in a simplified way (ADMS Urban, SIRANE) and using observations in proximity of traffic. The need to produce high-resolution air quality maps, as well as the rapid development of micro-sensors, is leading us towards the application of data assimilation methods on an urban scale.

During this work, a new data assimilation code, named YODA, has been developed. This module integrates three data assimilation approaches studied in Nguyen (2017), namely Bias Adjustment Technique (BAT), Best Linear Unbiased Estimator (BLUE) and Source Apportionment Least Square method (SALS).

These data assimilation methods have been applied to evaluate the air quality in the Île-de-France region (150 km x 120 km) in collaboration with the Airparif air quality agency. Overall, the results show that these methods statistically improve the estimates of air quality. They reduce hourly errors of estimates and increase correlations. Note that the improvement is not spatially and temporally constant.

In this case study, the best results are mainly obtained with the BLUE method. Results after data assimilation can also be worse than the initial estimates and sometimes the errors, although reduced, are still important. Finally, it is important to note that the BLUE method can lead to concentration fields that are not physically consistent.

Key words: *data assimilation, urban air quality.*

INTRODUCTION

For several years, data assimilation techniques have been used for assessing air quality by combining the results of numerical models with field measurements, in order to produce the best possible estimate of the pollutants concentration. Although these methods have been used extensively on a regional scale (Blond *et*

al., 2003; Zhang *et al*., 2012), they have rarely been applied on an urban scale, using models taking buildings into account in a simplified way (ADMS Urban, SIRANE) and using observations in proximity of traffic. The need to produce high-resolution air quality maps, as well as the rapid development of micro-sensors, is leading us towards the application of data assimilation methods on an urban scale.

This study compares three data assimilation approaches namely Bias Adjustment Techniques (BAT), Best Linear Unbiased Estimator (BLUE) and Source Apportionment Least Square (SALS). These three data assimilation techniques, studied in Nguyen (2017) and implemented in the new data assimilation code YODA, are applied to estimate air quality on an urban scale in the Île-de-France region.

DATA ASSIMILATION METHODS

Data assimilation (DA) methods are designed to optimally combine measured and modelled data in order to improve the system state estimate (Kalnay, 2003; Swinbank et al., 2003). This section presents the three data assimilation used in this study, namely the BAT, BLUE, and SALS methods.

Problem statement

The objective of the DA methods is to determine the true state of a system which is by definition unknown. This state is represented by the state vector $\boldsymbol{x^t}$ (t stand for true) called true state. To estimate this true state, the DA methods rely on measured and modelled data. The modelled data forms the a priori estimate of the system state represented by the state vector $\boldsymbol{x^b}$ (b stand for background) called background. $\boldsymbol{x^t}$ and $\boldsymbol{x^b}$ have a size n. Likewise, the observations are represented by the observation vector $\boldsymbol{y}$ which has a size m. By combining the observations and the background, the DA techniques lead to the best possible estimation represented by the state vector $\boldsymbol{x^a}$ called analysis which has also a size n.

To compare the observation vector and the state vectors, it is necessary to define the so called observation operator to pass from system space (space relative to state vectors) to the observations space. This operator is represented by the matrix $\mathbf{H}$ which has a size m x n. Therefore, the equivalent of the true state, the background, and the analysis in the observation space are $\mathbf{H}\boldsymbol{x^t}$, $\mathbf{H}\boldsymbol{x^b}$, and $\mathbf{H}\boldsymbol{x^a}$ respectively.

Bias Adjustment Techniques (BAT)

The Bias Adjustment Techniques consist in estimating the bias of the background with respect to the true state and removing it to result in an unbiased analysis. With these methods, the observations are supposed to be unbiased with respect to true state ($\overline{\boldsymbol{y} - \mathbf{H}\boldsymbol{x^t}} = 0$). BAT can be classified into two approaches: i) additive and ii) multiplicative BAT. In this study, only the multiplicative BAT approach is used. This variant has the advantage of guaranteeing positive estimates. With this alternative, the analysis is determined at each time step with equation (1), where $\overline{\boldsymbol{x}}$ represents the average of the vector $\boldsymbol{x}$.

$$\boldsymbol{x^a} = \boldsymbol{x^b}\frac{\overline{\boldsymbol{y}}}{\overline{\mathbf{H}\boldsymbol{x^b}}} \tag{1}$$

Best Linear Unbiased Estimator (BLUE)

The Best Linear Unbiased Estimator method is a statistical interpolation method which determines the analysis with respect to background and observations errors. The observation errors correspond to errors from the instrument and from the operator $\mathbf{H}$ modelling. With this method, the analysis is expressed with the equation (2), where $\boldsymbol{B}$ and $\boldsymbol{R}$ are the background error covariance matrix and observation errors covariance matrix.

$$\boldsymbol{x^a} = \boldsymbol{x^b} + \boldsymbol{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\boldsymbol{B}\mathbf{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}\left(\boldsymbol{y} - \mathbf{H}\boldsymbol{x^b}\right) \tag{2}$$

In this study, the observations errors between two different points p_i and $p_{j\neq i}$ are supposed to be uncorrelated. So the matrix $\boldsymbol{R}$ is diagonal (non-diagonal term equal to zero). Moreover we assume that the probability distribution of the observation errors is gaussian and that 95 % of these errors are inferior to some percentage of the mean measured concentration. Consequently the observation errors variance, R_{ii}, is modelled with equation (3), where $y_{i,t}$ is the observation associated to the i-th monitoring station at time

t, δ_i is the associated uncertainty and T is the number of time steps. For this case study, the uncertainty is set to 15 % as suggested by Union, P. (2008) for the NO_2 monitoring stations.

$$R_{ii} = \left(\frac{\delta_i}{1.96T} \sum_{t}^{T} y_{i,t} \right)^2 \tag{3}$$

In this study, we suppose that the strongest correlation of the background errors associated to p_i and $p_{j \neq i}$ is when these points are impacted by the same events (Blond *et al*. 2003). So, the background errors covariance B_{ij} is modelled as a function of the background correlation coefficient ρ_{ij}^b and variances $\sigma_i^{2,b}$ and $\sigma_j^{2,b}$ with equation (4), where the parameters α, ρ_0 and L_ρ represent an adjustment coefficient, a characteristic correlation coefficient, and a characteristic correlation length respectively. These parameters are determined by seeking those that satisfy the χ^2 diagnostic (Tilloy *et al*., 2013). Several parameters combinations can satisfy the χ^2 diagnostic. In this case, the chosen combination is that which leads to a smaller quadratic error with a leave-one-out cross validation.

$$B_{ij} = \alpha \sqrt{\sigma_i^{2,b} \sigma_i^{2,b}} \rho_0 \exp\left(\frac{\rho_{ij}^b - 1}{L_\rho} \right) \tag{4}$$

Source Apportionment Least Square (SALS)

The Source Apportionment Least Square method (Nguyen *et al*., 2018) supposes that uncertainties associated to modelled estimates from air quality models are mainly due to emission estimate errors. So, the SALS method consists of correcting (indirectly) emission data in order to improve estimates from air quality models. This correction is achieved by modulating, in an optimal way, the contributions of the sources. The analysis at each time step with the SALS method is defined in equation (5), where $\boldsymbol{x_g^b}$ represents the background associated to the contribution of the sources in group g, λ_g is the modulation coefficient associated to the sources in group g and G is the number of source groups.

$$\boldsymbol{x^a} = \sum_{g}^{G} \lambda_g \boldsymbol{x_g^b} \tag{5}$$

With this DA method, the best estimate is that which minimizes the quadratic error in respect to observations which are supposed to be perfect ($\boldsymbol{y} = \mathbf{H}\boldsymbol{x^t}$). So, the coefficients λ_g are evaluated at each time step by minimizing the cost function J defined by the equation (6).

$$J = (\boldsymbol{y} - \boldsymbol{x^a})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{x^a}) \tag{6}$$

This method assumes that the source's group contribution is positive (or null) and as a result the minimization of the cost function J is made with Lawson & Hanson (1974) method which guarantees positive coefficients λ_g.

CASE STUDY

Description of the study case

The study case consists on evaluating air quality on the Île-de-France region (150 km x 120 km), more specifically the hourly NO_2 concentrations. The study period extends from 1st December 2016 to 30th June 2017. On this period, 35 monitoring stations have provided NO_2 measurements. For this study, the background (for the assimilation) comes from numerical dispersion model simulations performed by a coupling of ADMS Urban for the local scale contribution and CHIMERE for the regional scale contribution. The background is the sum of the contributions from the traffic emission, the surface distributed sources and the background concentration (concentration due to pollution coming from outside the domain). These three contributions are used to apply the SALS method.

These data (measurements and simulations results) constitute the *input* for the data assimilation. The performances of the three data assimilation methods are assessed by means of a leave-one-out cross-validation. The purpose of this approach is to evaluate uncertainties from data assimilation results on an area without measurements.

The quality of the estimates is assessed by means of 3 statistical indices: Bias, Root Mean Square Error (RMSE), and the correlation coefficient (Corr). These indices are defined in the Table 1 where c_m and c_p represent the measured and predicted concentrations respectively.

Table 1. Statistical indices to assess the quality of the estimates

Bias	RMSE	Corr
$\overline{c_m - c_p}$	$\sqrt{\overline{(c_m - c_p)^2}}$	$\frac{\overline{(c_m - \overline{c_m})(c_p - \overline{c_p})}}{\sqrt{\overline{(c_m - \overline{c_m})^2}\,\overline{(c_p - \overline{c_p})^2}}}$

Results

Figure 1 compares the statistical indices obtained before (relative to the background) and after cross-validation with the three DA methods at each monitoring station. Generally, the bias for the background stations is made worse by the DA methods but improved for the traffic stations by about 20 %. The BLUE method is the only one which improves both initial negative and positive bias. Globally, the DA methods also improve the RMSE by about 20 %. The BLUE method generally leads to the best RMSE. Moreover the worst RMSE is lower/better with this method. Likewise, the DA methods improve the correlation coefficient by about 10 %. Once again, the correlation coefficients are generally slightly better with the BLUE method. Note that this method is the only one which improves the RMSE and the correlation coefficient for all the monitoring stations.

Globally, these results indicate that the DA methods statistically improve the hourly NO_2 concentration estimates and that the results of the BLUE method are slightly better. Regardless of the DA method, the improvement/deterioration are not spatially constant. Additionally, the estimates of the DA methods can also be worse than the background's ones and the errors, although reduced, can still be important.

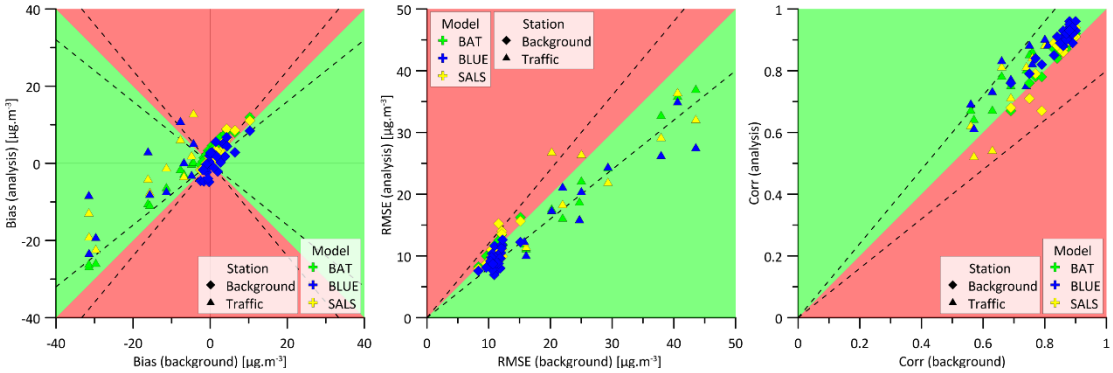

Figure 1. Comparison of the Bias, the RMSE, and the coefficient correlation before (background) and after (analysis) cross-validation with the three DA methods. Green and red areas indicate an improvement and a deterioration respectively. The dash line points out an improvement/deterioration by about 20 %.

Figure 2 shows the NO_2 hourly concentration field for the 2nd February 2017 at midnight on the Île-de-France region modelled with the background and the BLUE method. These results indicate that the BLUE method can significantly modify the spatial variability of the concentrations (it is also true for the BAT and SALS method but it is not shown here). Note that the concentrations along some traffic roads are lower than the concentrations in the surrounding area with the BLUE method. This is not physically consistent since the NO_2 concentrations should be greater along the roads themselves rather than in the surrounding area. This behaviour is a consequence of the modelling of the $\boldsymbol{B}$ matrix (eq. (4)).

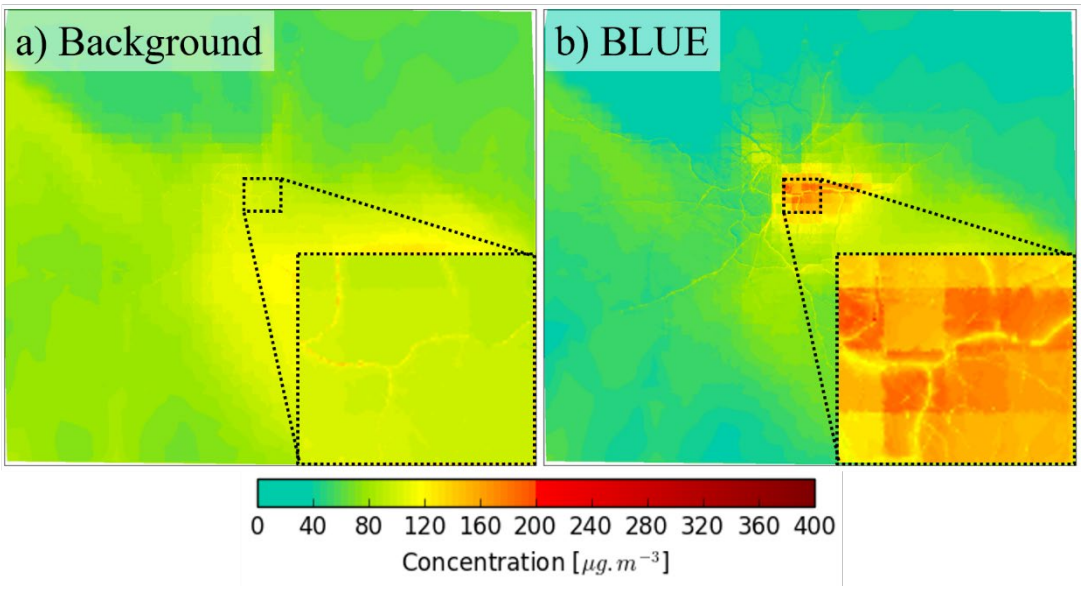

Figure 2. Hourly NO_2 concentration field the 02/12/2016 at midnight on Île-de-France region (150 km x 120 km) estimated with the background and the BLUE method

CONCLUSION

This study aims to compare the performances of three DA methods, namely the BAT, BLUE, and SALS methods. These DA methods are assessed in a case study with the aim of evaluating hourly NO_2 concentrations on an urban scale in the Île-de-France region. Overall, the results show that these DA methods can statistically improve the estimates of air quality. They reduce hourly errors of the estimates and increase correlations. Note that the improvement is not spatially and temporally constant.

In this case study, the best results are generally obtained with the BLUE method. The estimates of the DA methods can also be worse than the initial estimates and the errors, although reduced, can still be important. Finally, it is important to note that the BLUE method can lead to incoherent concentration fields.

ACKNOWLEDGMENTS

This study was funded by Airparif air quality agency. This work was carried out as part of the FUI project FAIRCITY supported by the Auvergne-Rhône-Alpes Region.

REFERENCES

Blond, N., Bel, L., & Vautard, R. (2003). Three-dimensional ozone data analysis with an air quality model over the Paris area. Journal of Geophysical Research: Atmospheres, 108(D23).

Kalnay, E. (2003). Atmospheric modeling, data assimilation and predictability. Cambridge university press.

Nguyen, C. V. (2017). Assimilation de données et couplage d'échelles pour la simulation de la dispersion atmosphérique en milieu urbain (PhD, Ecole Centrale de Lyon, France).

Nguyen, C. V., Soulhac, L., & Salizzoni, P. (2018). Source Apportionment and Data Assimilation in Urban Air Quality Modelling for NO2: The Lyon Case Study. *Atmosphere*, *9*(1), 8.

Swinbank, R., Shutyaev, V., & Lahoz, W. A. (Eds.). (2003). Data Assimilation for the Earth System (Vol. 26). Springer Science & Business Media.

Tilloy, A., Mallet, V., Poulet, D., Pesin, C., & Brocheton, F. (2013). BLUE-based NO2 data assimilation at urban scale. Journal of Geophysical Research: Atmospheres, 118(4), 2031-2040.

Union, P. (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Union.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. Atmospheric Environment, 60, 656-676.