Data Assimilation Schemes in Colombian Geodynamics - Cooperative Research Plan for 2017 - 2020 Between Universidad EAFIT and TUDelft, With the Help of Universidad de Antioquia and universidad Nacional de Colombia Sede Medellin

Inicio: 1 de Enero de 2017
Final: 30 de Diciembre de 2020

Technical Report – Data Assimilation schemes, Theoretical aspects

Medellin Air qUality Initiative MAUI

MAUI-RT-05

Entidad Ejecutora	Universidad EAFIT Cra 49 No 7sur - 50 Medellín, Colombia
	Grupo de investigación en modelado matemático – GRIMMAT Grupo reconocido por COLCIENCIAS Categoría A Grupo de investigación en Biodiversidad, Evolución y Conservación - BEC
Responsables	Prof. Olga Lucia Quintero Montoya Prof. Nicolás Pinel Peláez. Investigadores
Entidades Cooperadoras	Department of Applied Mathematics - Tu Delft, Delft The Netherlands TNO
Responsables	Arnold Heemink Arjo Segers

CONTROL DE EDICIÓN Y DISTRIBUCIÓN

Edición	Control*	Nombre y Cargo	Firma	Entidad	Fecha (Día/Mes/Año)
	Creación	O. Lucia Quintero		Universidad EAFIT	

* Especificar tipo de control: Creación - Revisión - Modificación - Distribución.

Contents

ABSTRACT 4

INTRODUCTION 5

THEORETICAL QUESTIONS 7

Backtracking strategy for covariance localization 8

trj-4DVar 11

ABSTRACT

Technical questions about the feasibility of the development of new Data Assimilation schemes within the framework of the Air Quality modeling for Colombia and specifically for the Abulla Valley, are one of the main goals of the research project.

This document contains not only the technical questions to be solved but also, the first approach to the understanding and the development of the related to the recently developed Two-way variance localization and trajectory 4-d Variational methods.

INTRODUCTION

Some basics of Data Assimilation can be found in the Data Assimilation The Ensemble Kalman Filter book from Geir Evensen 2009. Data Assimilation tech- niques were introduced first by Evans in 1993 and quickly adopted and implemented in several applications related to Geosciences and Geophysics such as coastal, oceanographic, model reduced gradient and recently, due to the Eyjafjjallaokull volcano eruption in 2010 applied to Volcanic ash transport through the use of the LOTOS-Euros Model for transport models.

Data assimilation relies on the use of an extension for high dimensional systems of the classical approach for filtering called the Kalman Filter. The assumptions over the dynamic behavior of the system include highly related dynamics written in a high dimensional states vector with normal distribution. The use of other approaches such as particle filtering have been evaluated but the efforts have been made about the possibility of improvement of the information for the models by corrections introduced by data. This is why the Data Assimilation term was adopted because the model will assimilate the data in order to improve its accuracy from observations.

There are at least two types of data assimilation: related to the Ensemble Kalman filter for state estimation and the so called variational methods for the parameter estimation. The former allows for the improvement of the model states trough the use of measurements available at specific timestep. The later provides an approach to the parameter estimation through the solution of an optimization problem integrating multiple observations of data, this suggest the improvement of the optimal solution for the proper estimation of the pdf. Several recent improvements to the Data Assimilation Schemes both Ensemble based or 4D-var have been developed by Barbu 2010, Krymskaya, 2013, Sebacher, 2014, Altaf 2015, Fu et al, 2015, Lu et al, 2015.

Recently, has been a lot of attention regarding the measurement of the impact of the observations (data) over the proper performance of a Data Assimilation Scheme in the enhancement of a model. Several works regarding the quantification of the impact of data in reservoir modeling for subsurface applica- tions (Krymskaya, 2013), inclusion of physical phenomenon into the constraints to the problem (Tijana et al, 2014) or the observation impact analysis in storm surge problems (Verlaan and Sumihar, 2016) show evidence about the feasibility to develop new and improved mathematical analysis tools and its algorithms for Data Assimilation stages regarding the solution of a specific problem.

During the last 3 years, the MahPhys group has been developing and improv- ing the Data Assimilation schemes from at least three points of view, regarding the two types of data assimilation techniques: the variational methods, and sequential methods.

1. The reduced adjoint approach theoretical and idealized test cases (Altaf et al, 2013)
2. The quantification of the impact of data in reservoir modeling through observation sensitivity matrixes (Krymskaya, 2013)
3. Ensemble spurious correlations in Ensemble Kalman Filter (Fu et al, 2015- 2016) (Evans, 2009)

4. Observational spurious correlations solution in 4-d variational Data assim- ilation(Lu et al, 2015-2016)(Beney, 1982) (Jackson, 1991)
5. Conservation of mass and Preservation of possivity in physical meaning variables for Enseble-Type kalman Filter Algorithms (Tijana et al, 2014) and The observation impact analysis methods for Storm Surge - TSBOI-MM algorithm (Verlaan and Sumihar, 2016)
From a practical point of view, the Spurious Correlations have been studied and solved for the case of Data Assimilation in the Eyjafjjallaokull volcano eruption in 2010. For both the case of Ensemble Kalman Filter approaching a solution to the variance localization, so called backtracking localization and the modified traj4D-var to accurately estimates the vertical distribution of effective volcanic ash injection rates from satellite observed columns. The quantification of the impact of data has been evaluated over a model gradient reduced model for reservoir and the reduced adjoint approach was illustrated in a ground water subsurface contaminant model. They have also been working on the application of the LOTOS-EUROS Model for Dust forecasting in China (Jin, 2016).

After carefully reading the latest developments from Prof. Heemink's students, we realized that the answers to the Data Assimilation questions regarding the volcano need to be generalized and formalized through the help of Control Systems formalism. Both approaches solved the problems through the use of model-based knowledge for the use of a backtracking localized strategy to the Ensemble variance localization and the so called trajectory based 4D-var Data Assimilation adjoint free solution of the modified cost function and knowledge based modifications to the 4D-var method.

The developments regarding Data Assimilation Techniques and algorithms applied on the reservoir modeling/history matching in subsurface and storm surge forecasting have been topic of study (Tijana et al, 2014)(Verlaan and Sumihar, 2016).

Consequently several questions arised as research lines for both theoretical and applied fields. So far, we have identified at least seven theoretical and practical aspects to be studied, also a potential problem to start to solve in a joint research.

This document adresses the theoretical challenges proposed by Quintero during the time at TUDELFT in 2016, and also develops further hypothesis regarding their feasibility to be developed within the framework of the actual research of Air Quality modeling in Colombia.

THEORETICAL QUESTIONS

Theoretical questions on Data Assimilation

1. To analyze and/or quantify the impact of the proper selection of a data subset (from a big source of data) to be used in a Data Assimilation scheme in order to develop a weighting strategy or sensitivity analysis to use the environmental information in other systems. Based in previous work of observation sensitivity matrix (Krymskaya et al, 2010) and its properties such as Structure: symmetry, scaling, positive (semi)definiteness; Matrix norm, uncertainty measurement, and evaluation with SVD and trace analysis for the quantification of the impact in modeling. It is necesary perform further research to formulate the criteria for selection of the redun- dant data based on the shape of the curve obtained from the observation matrix eigenvalues or another new one.

2. To answer the question: is it possible, under lineality and stationarity assumptions, to use the observation impact analysis methods developed by Verlaan and Sumihar, 2016 to improve the Data Assimilation Schemes over LOTOS-EUROS model forecasting?. This will be held by studying the impact of the observations at the most recent analysis update under Ensemble based schemes in LOTOS EUROS model for volcanic ash

3. To evaluate the effect of new localization strategies developed by Fu et al, 2015 over the observation sensitivity analysis proposed by (Verlaan and Sumihar, 2016) in LOTOS-EUROS model for volcanic ash data assimilation.

4. The extension of the trj-4DVar (Lu et al, 2015) to bigger systems and mathematically feasible solutions to solve the modified cost function in a new or improved approach adjoint free. We find feasible to try a hybrid scheme for the ill conditioned problem in order to deal with reduced observational noise.

5. The use of a formal sensitivity analysis for the perturbation of the inputs to the two experience based modifications in cost function in the trj-4DVar (Lu et al, 2016 under review), specifically in the penalty term in order to develop a generalized method and probe the stability of the solution.

6. To develop further evidence and formalize the postulate of accuracy of the method for a physical-related window of assimilation, regarding the demonstrated accuracy in assimilation windows that does not compromise the constant concentration of particles in atmosphere (Lu et al, 2016 under review). Our point of view is that dynamically it will change with winds profile.

7. To test the backtracking strategy for the Ensemble Assimilation scheme developed (Fu et a, 2015) against the developed by Liu and Xue (2016) and its approximate ensemble covariance localization.

8. To merge both “spurious” approaches for “multiple observations” prob- lems (variable localization).

BACKTRACKING STRATEGY FOR COVARIANCE LOCALIZATION

Regarding the question number 7: “Testing the backtracking strategy for the Ensemble Assimilation scheme developed (Fu et a, 2015) in LOTOS-EUROS Volcanic ash problem against the developed by Liu and Xue (2016) and its approximate ensemble covariance localization”.

During this report we realized that the mentioned technique suffered a change on the name, I guess this arise from the questions developed during 2016, and in the PhD Thesis “Improving volcanic ash forecasts with ensemble-based data assimilation” by Guangliang Fu, 2017 the term “backtracking” was eliminated and substituted by Two-way tracking covariance localization.

Subsequently the questions about the physical meaning of an operator were partially solved in this thesis. As follows we will develop certain theoretical/practical aspects of the recently introduced data assimilation techniques, looking for the understanding and evaluation of the current conditions of the Air Quality Modeling.

To develop an ensemble-based scheme for data assimilation, it is necessary to define the localization matrix for covariance localization. It is necessary to analyze the characteristics of the physical forecast error covariances. From that, the two-way tracking localized Ensemble kalman Filter (TL-EnKF) was introduced maintaining the correctly specified physical covariances and largely reducing the spuriuous ones.

Basics:

- Because of the limited ensemble size, ensemble-based assimilation methods often produce severe spurious noises between measurements and state variables, implying that some state variables are in fact uncorrelated with an observation but they are wrongly computed as correlated, resulting in a unphysical result for the update.
- Typically, the localization matrix was chosen with ones on the diagonal and with other values different from zero from the diagonal until a specified distance defining a structure as a “non zero band” around the diagonal, defining the filtering scale.
- If the filtering scale is wrongly selected, the uncorrelated covariances can remain, it means that this filtering scale will allow/not allow the physical and non physical covariances. It must be chosen properly.
- For weather, ozone, So2 and CO2 forecast, the physical error covariances are “isotropic” but if something leads to the anisotropy, the length scale for the assimilation can be chosen using the topography as a “vertical” coordinate for a 2d analysis application.
- Depending on the application, the length scale can be according of the interest area, for example for volcano eruption was about 500 km.

Estimation for the physical forecasting error can be found in Fu Thesis, 2017 section 5.2.1 analyzing the properties of the localization matrix. He concluded that for large scale three dimensional atmospheric assimilation, it is usually unrealistic to employ an ensemble size over 100, because the large computational cost.

An anisotropic matrix means that the covariance shape and spread vary in different directions, and this fact implies that the structure of the matrix will vary over space. A helpful insight can be compare the structure with the ECMWF wind field. In our application will be difficult to obtain a good resolution for the wind field in ECMWF for will be very useful the Weather Research Forecast (WRF) model capability to reproduce the wind and meteo variables over Colombia domain.

Covariances can be up-wind dominant and physically it makes sense because errors reduced by assimilation propagate downstream and observations will require larger updates upstream and for the volcano case, the ash covariances will decrease in time, so larger covariances are expected upstream.

Calculating the correlations and standard deviation can be helpful to understand the subyacent problem, because it may be that forecast error covariance is standard deviation dependent. It still remains unclear for the Air quality modeling in Colombia at valley scale. But it seems to be that way for macrodynamics such the punctual source problem.

If we think about the Aburra Valley as a volcano see RT-06, this assumption may be valid, and the LOTOS-EUROS domain can remain in a scale in which the valley represent a very punctual source. It can be recommended for the analysis of the effect and transport of pollutants from the city to the surrounding ecosystems.

The proposed “Two way tracking” method defines an operator for the “dominant processes” so we will look for the dominant processes at two different scales: if we see Aburra Valley as a volcano related to the bigger domain at lower resolution, they can be advection and diffusion but we are not certain about the high resolution dynamics at intra-valley level.

The procedure is straight forward and can be found in section 5.3. some important issues arise when the adjoint matrix for the assimilation is not calculated and an operator was developed for its approximation. The two dominant processes were reversed and it finally was aimed to “roughly capture” the up-wind correlation patterns.

So it arises a question for the air quality problem. Will be useful and not necessary to avoid the calculation of the adjoint matrix? The creation of a mask for localization with a threshold depends on the source of information. i.e remote sensing.

For our research must be interesting to study other variants of the localization matrix such as adaptive radius or radius based on distance functions, and this kind of schemes must be able to reproduce the anisotropic behavior of covariances depending on the dominant processes.

Schemes of data assimilation based on ensemble Kalman filter can also suffer of computational complexity because of the large scale of the model and the introduction of a large ensemble size. Examples of this can be the weather forecast in which the initialization and forecast can make the task even more complex. At lest two things must be taken into account for the computational design of the data assimilation schemes: the first one relies on the possibility of parallelization of the ensemble forecast; secondly it is important to realize that in the analysis stage requires the calculation of the Kalman gain and ensemble covariance matrix so the former step must be ended and the ensembles are combined together. Once this is already taken into account we must be carefull because we face two problems with different model complexities (high or low resolution), observation type (satellite, ground based, remote sensing, which leads to sparsity or density of observations) and the requirements of the framework (accuracy or speed) .

Fu improved the analysis step with an algorithm reformulating the calculation of the analysis matrix, concentrating the calculations within the plume due to advection process. But not for the ozone, So2, Co2 applications, because of they are concentrated in the entire domain. For this specific issue, the domain parallelization (Segers, 2002) must be reviewed.

Nevertheless if we think again in the “Volcano of the aburraes” at medium resolution, Fu algorithm may be useful. It does not affect the localization procedure and localization will not accelerate the assimilation.

Regarding the assimilation of satellite data, an operator (Satellite Observational Operator) was developed to map from 2d to 3d at certain layers and Ensemble Square root Kalman filter was used (EnSKF) and preprocessing of data was performed. So it confirms the approach of our group.

TRJ-4DVAR

The theoretical questions regarding the topic of variational schemes are related to the “Verification of the feasibility for the development of an extension for the trj-4DVar (Lu et al, 2015) in LOTOS-EUROS Volcanic ash problem to bigger systems and mathematically feasible solutions to solve the mod-ified cost function in a new or improved approach adjoint free. We find feasible to try a hybrid scheme for the ill conditioned problem in order to deal with reduced observational noise”.

The challenge for variational data assimilation methods in a punctual source is their incapability to reconstruct the vertical profile of emissions or improve the forecast of the ash concentrations. For our case, the air modeling for Colombia Domain provides several challenges from the initial and boundary conditions to the resolution and focus of analysis i.e intra valley pollution concentration or “volcano of aburraes” effect on surrounding ecosystems. Both of them can serve as source of inspiration for improvements in the variation schemes. For instance as follows we will develop the main topics regarding the questions 4 to 6:

“4. The extension of the trj-4DVar (Lu et al, 2015) to bigger systems and mathematically feasible solutions to solve the modified cost function in a new or improved approach adjoint free. We find feasible to try a hybrid scheme for the ill conditioned problem in order to deal with reduced observational noise.

5. The use of a formal sensitivity analysis for the perturbation of the inputs to the two experience based modifications in cost function in the trj-4DVar (Lu et al, 2016 under review), specifically in the penalty term in order to develop a generalized method and probe the stability of the solution.

6. To develop further evidence and formalize the postulate of accuracy of the method for a physical-related window of assimilation, regarding the demonstrated accuracy in assimilation windows that does not compromise the constant concentration of particles in atmosphere (Lu et al, 2016 under review). Our point of view is that dynamically it will change with winds profile.”

Variational methods such as 4D-var determine an optimal combination of the information to obtain variables or parameters of a model. From a control systems point of view the nonlinear system depends of the inputs, states and outputs and the estate or parameter estimation u, theta and x.

For standard air quality applicarions including souce estimations, the standard 4 d var has been successfully applied (review Elbern et al, 2007 and Meirink et al, 2008).

“A chemica.1 four-dimensional variational data assimilation svstem has been developedand applied for the study of an episodewith enhancedsummerly ozone levels. In this study the optimization parameters are the initial values of the chemical constituents. The case study focuseson an ozone episode over central of surhce observationsof ozone and nitrogen oxides, but also a limited number of ozoneradiosonde records. The four-dimensional data assimilation algorithm iscomposed of the chemistry transportmodel,its adjointmodel with the adjoint versions of the Regional Acid Deposition Model (RADM2) chemical mechanism, horizontal and vertical advection schemes,implicit vertical diffusion, and a limited memory quasi-newton minimization routine. The underlying model of the spatio- te•nporal data assi•nilationschemeis the comprehensivemesoscale-• EUropean Air pollution Dispersionmodel (EURAD), whichis basedon the RADM2 gas phase mechanism. On the basis of a

6 hours data assimilation interval, analyses of the chemical state of the atmosphere were obtained, where the skill is verified in two different ways: (1) improvements of forecasts subsequent to the assimilation procedure and (2) validation of the analyses with observational data which is with held from the variational data assimilation algorithm. It is demonstrated that significant improvements are achieved for short-term forecasts including the afternoon ozone peak abundances. The analysis skill is further corroborated by examinations with retained measurement data.”

“A four-dimensional variational (4D-Var) data assimilation system for inverse modelling of atmospheric methane emissions is presented. The system is based on the TM5 atmospheric transport model. It can be used for assimilating large volumes of measurements, in particular satellite observations and quasi-continuous in-situ observations, and at the same time it enables the optimization of a large number of model parameters, specifically grid-scale emission rates. Furthermore, the variational method allows to estimate un- certainties in posterior emissions. Here, the system is ap- plied to optimize monthly methane emissions over a 1-year time window on the basis of surface observations from the NOAA-ESRL network. The results are rigorously compared with an analogous inversion by Bergamaschi et al. (2007), which was based on the traditional synthesis approach. The posterior emissions as well as their uncertainties obtained in both inversions show a high degree of consistency. At the same time we illustrate the advantage of 4D-Var in reducing aggregation errors by optimizing emissions at the grid scale of the transport model. The full potential of the assimilation system is exploited in Meirink et al. (2008), who use satellite observations of column-averaged methane mixing ratios to optimize emissions at high spatial resolution, taking advan- tage of the zooming capability of the TM5 model.”

But for ash emission rates the problem is ill conditioned due to the information retrieved from satellite, because it is just information about total amounts and do not describe the vertical profile of pollutants. Fortunately, the operator from state to output is known.

Unfortunately, the information from satellite do not allow to follow the source of uncertainty of the model because it can not be retrieved the layer of the error source. So the solution of the optimization problem via gradient, are not necessarily well defined and the gradients are invalid.

In atmospheric transport models, particles are transported because of the meteorological conditions and wind directions, so we demonstrate (Pinel et al, 2017) that the wind carries out information of the emissions from the source to the final deposition. Observations of concentration at certain points could be used to trace back information of emissions through the wind field. So the sensitivity of the state vector of concentrations respect to emissions can be build.

The problem arises when the observations are not only a measurement related one to one to the states, the problem relies on the fact that maybe it could be a combination of states.

Lu, 2017 postulate that satellite observations create strong correlations between states and emission source variables. Preprocessing of data will help to avoid problems of the observation itself, instead of deal with observations uncertainty. This assumptions lead Sha Lu to the formulation of the modified traj4D-variational method by adding better statistics for the correlation of states and sources and she modified the cost function adding the emissions which she called background. The main goal was to develop a series of possible trajectories to be followed in order to determine what kinf of disturbed trajectory provides better results. The trajectories are simple simulations of emission sources (background) and silgth changes in the emission sources (changes in background) for Lotos Euros model she denomined **u** as the source, representing the input to the system. I agree that the control variable must be the emissions rate and developed a path to increase our research capabilities regarding dynamic emissions such as traffic.

The traj 4D var has the advantage of avoind the adjoint calculation so must be implemented for less than 100 states.

We must definitively explore first the ensemble –based 4D var (Liu et al, 2009) which targets to determine a flow dependent matrix, uses Monte Carlo simulation for ensemble generation (I like this most) and is scalable. The previous assumption must be carefully reviewed depending on the focus of the analysis, remember that traj 4D var was developed for vertical profile estimation and maybe we will have a lidar.

There is a regularization term with lack of analysis from Lu, so the equations 4.9 and 4.10 must be carefully reviewed because of the derivative approximation for the term can be useful in some way. For instance, it remains being a particular solution for the vertical profile estimation. Volcano of aburraes can be one source application and due to the known winds profile from (Pinel et al, 2017) deserves to be explored. I will rater prefer a smoothing term in order to reduce oscillations, as the effect of a filter for control purposes.

Mixing satellite data with ground based is definitively our case, resolution and aim must be defined to figure it out how can we approach the solution via variational methods. Selectively, the penalty term for the data was added to the cost function, of course based on preprocessing and prior knowledge. The analysis and knowledge of the team developed from the first data analysis and model runs (Lopez et al, 2017; Rodriguez et al, 2017; Rendón et al, 2017; Posada et al, 2017), provide excellent sources for similar developments. The fact is that the careful selection of restrictive variables from data sources, are more related to the punctual rather than the complete domain problem. Updates of emission sources can help by adding the correction term to the variational scheme.

One of the nicest information that we can use, is the meteorological radar from SIATA, the main problem is the deep learning of the phenomenon and the preprocessing because the amount of information. CALIPSO and CALIOP must be also very useful regarding the Volcano of aburraes hypothesis, not for the Air Quality intra valley monitoring.

So far now the 4th question remains without solution for ash DA but it deserves to be reviewed for the AirQuality problem, because of the intravalley monitoring must be solved by reducing the model to the relevant variables.

The 5th question must be solved by determining the most important source of uncertainty on the emissions, adding information to the inventory and detailing the conditions for simulation. Traffic as dynamic source must be included for the intravalley application in a low scale proble without adjoint, and for the total balance for the Volcano of aburraes hypothesis.

Regarding the accuracy or 6th question, and the sensitivity analysis of the proposed method on the assimilation process and forecast, the development of criteria such as fisher information matrix between the observations and remote sensing information DA process is a formal way to measure the impact of good or bad data. The problem arises when the hessian matrix must be calculated, of course it is feasible for low dimensional problems.

Another criteria introduced is the gradient of the results, providing local and detailed information of the quality of the solution and the computational performance, remember the variational method is solved via gradient.

REFERENCES

Elbern, H and Schmidt, H. ozone episode analysis by four-dimensional variational chemistry data assimilation, J. Geophys. Res 106, 3569 (2001)

Meirink, J. F., Bergamashi, P., Frankerberg, C., dÁmelio, M. T. S., Dlugokencky, E. J., Gatti, L. V., Houweling, S., Miller, J. B., Rockmann, T., Villani, M. G., and Krol, M. C., Four-dimesnioanl variational data assimilation for inverse modeling of atmospheric methane emission: Analysis of SCIAMACHY observations, J. geophys. Res 113, D17301 + 2008.

López et al, 2017, Challenges and opportunities for Open LOTOS-EUROS model to reproduce the Dynamics for Tropical Andes Domain

Pinel et al, 2017. POTENTİAL URBAN POLLUTİON İMPACTS ON PROTECTED AREAS İN COLOMBIA THROUGH ATMOSPHERİC TELECONNECTİONS

Posada et al, 2017. EVALUATION OF THE WRF MODEL FOR THE STUDY OF THE METEOROLOGICAL ENVIRONMENT OVER THE TROPICAL ANDES DURING EL NIÑO

Rendón et al, 2017. MECHANISMS OF AIR POLLUTION TRANSPORT IN URBAN VALLEYS

Rodríguez et al, 2017, CHARACTERIZATION AND ANALYSIS OF SATELLITE AND GROUND

DATA AVAILABLE FOR THE ABURRÁ VALLEY (MEDELLIN

METROPOLITAN AREA) AS INPUTS FOR AIR QUALITY MODELS.