Data Assimilation from Operational and Industrial Applications to Complex Systems

Jean-Philippe Argaud* **Bertrand Bouriquet†** **Julian Hunt**, Cmath (Hon) FIMA†‡

Data Assimilation is a general set of mathematical and numerical methods for computing the optimal estimate of the true state of a system over time. It uses values obtained both from observations and *a priori* models, including information about their errors.

The roots of Data Assimilation go back to the XIX century, when Gauss in 1809 discovered the well-known least square hypothesis, the first systematic method in this field. His aim was to adjust a model to fit observed values in experiments. In his lecture, in 1835, he applied this method to astronomical data.

Superficially, Data Assimilation can be considered as an extension of the least square method. The method of the Best Linear Unbiased Estimation (BLUE) filter is the simplest, both theoretically and computationally [1]. Data Assimilation is also valid for dynamical or time-varying data sets and models. The Kalman Filter is one of the general methods that provides optimal evaluation of data in relation to a time-dependant model. Both the BLUE and the Kalman Filter are very efficient, especially as long as the spaces of observed data or of state variables have a limited size. Other methods are generally used for very large data sets.

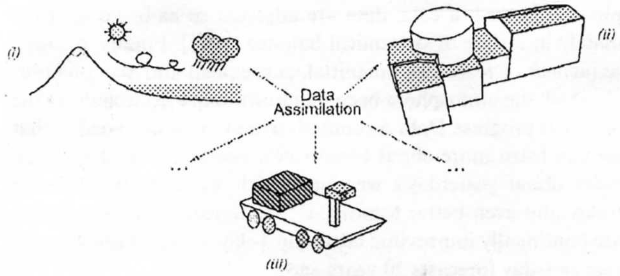

Figure 1: Data Assimilation applications are developing from (i) weather forecasting to (ii) nuclear core simulation and (iii) robotics. It is likely such techniques will spread to other fields of science or technology, and even sociology wherever modelling is required.

Although the developments largely come from control and estimation theory, Data Assimilation's main improvements have resulted from its wide spread use in computing meteorological and ocean models to obtain forecasts for the weather and ocean weather. These are high dimension complex modelling systems for variables computed at a huge number of points (more than 10^7 for typical weather forecasts). They require time varying measurements at as many of these points as possible. In practice, the data (even with satellites probes) are only available at very few (less than about 1% of the computation points). But this is still a very large number of observations. The first modern (in the 1980s) technique developed was 3D-VAR (Three-dimensional Variational Data Assimilation). This method provides optimal estimation at a particular time step, and has to be iterated over time. This technique was upgraded (in the 1990s) to 4D-VAR (Four-dimensional Variational Data Assimilation) in order to take into account information about the process dynamics over the whole time window (up to the computation time). Active research on Data Assimilation and its applications continues particularly to improve its computational efficiency.

There are several fields in science and technology where the effective use of observed but incomplete data is crucial, the most recent applications being in robotics and nuclear power. Those applications as well as meteorology illustrate the importance and value of Data Assimilation as the framework for industrial operations (see Figure 1). In so doing it is becoming a tool for linking different domains in science and technology. But as usual when similar methods are used for different applications, they tend to have different names. However, the similarities and differences between these applications have stimulated the development of more efficient methods in all fields. This short overview of the underlying theory and the common methods used in practice introduces the different approaches to Data Assimilation in various fields of applications. Recent advances and new ideas emerging from these different approaches are compared.

Basic theoretical elements of Data Assimilation

The basic aim of modern Data Assimilation methods is to make optimal estimates of the initial or time developing state of variables in a model, by combining all available information (observations, prior information in the form of a forecast from the model, physical relationships...). The ultimate goal of Data Assimilation is to be able to provide a best estimate of the inaccessible "true" state of a system, and for forecasting its future states, when only partial data is available. The true values of the variables at all the points of the system define the state, which is denoted $\mathbf{x}^t$, with the *t* index for "true". The *a priori* state of the system (denoted by $\mathbf{x}^b$, with *b* for "background") is inferred from measurements (denoted by $\mathbf{y}$). The result of Data Assimilation is called the analysed state $\mathbf{x}^a$, which is the estimation of the true state $\mathbf{x}^t$ we want to find. The detailed explanations of this method can be found in the references [1] and [9], and a scheme of variational methods is presented in Figure 2.

Mathematical relations between these states defined by vectors have to be derived or assumed. As the mathematical spaces of the background and of the observations are not necessarily the same, a bridge between them is needed. This is the so called observation operator H, with its linearization $\mathbf{H}$, that transforms values from the space of the background state to the space of observations. The reciprocal operator is the adjoint of H. For linear processes, the adjoint is the transpose $\mathbf{H}^T$ of $\mathbf{H}$.

Two other ingredients are necessary. The first one is the covariance matrix $\mathbf{R}$ of observation errors ε_o, defined by $\varepsilon_o = \mathbf{y} - H(\mathbf{x}^t)$. It is usually assumed that the errors are unbiased, so that $E[\varepsilon_o] = 0$, where E is the mathematical expectation

* Electricité de France, Research and Development, 1 avenue du Général de Gaulle, 92141 CLAMART Cedex, France
† Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, 42 avenue Gaspard Coriolis, 31057 TOULOUSE Cedex, France
‡ University College London, Gower Street, LONDON, WC1E 6BT, United Kingdom

operator. This can be obtained from the known errors of the unbiased measurements. The second one is the covariance matrix **B** of background errors ε_b, defined by $\varepsilon_b = \mathbf{x}^b - \mathbf{x}^t$. This represents the *a priori* error, assuming it to be unbiased. There are many ways to obtain the *a priori* background error matrices. However, they are commonly included in the output of a model with an evaluation of its accuracy, and/or the result of expert knowledge.

Within this formalism, the analysis $\mathbf{x}^a$ is the Best Linear Unbiased Estimator (BLUE), and is given by the following equation:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{1}$$

where **K** is the gain matrix [1]:

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}. \tag{2}$$

It is worth noting that solving equation (1) is equivalent to minimising the following function:

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + (\mathbf{y} - \mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}), \tag{3}$$

$\mathbf{x}^a$ being the optimal solution.

The minimisation of the formula (3) leads to the same result as the BLUE in the linear case. In a minimisation procedure, it is easy to replace the linearised expression **Hx** by the non linear one $H(\mathbf{x})$. This method is referred to as 3D-VAR in Data Assimilation (see Figure 2).

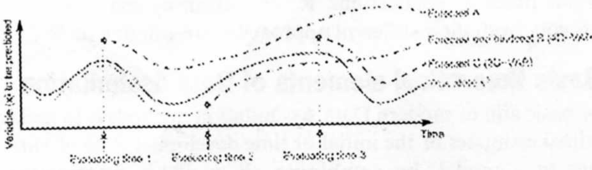

Figure 2: Illustration of the Data Assimilation forecast procedures by adding measurements over time and continual updating forecasts (3D-VAR (B) and 4D-VAR (C)) and hindcast (for 4D-VAR (C) only). Each new measurement is used to re-evaluate and adjust the forecast, bringing it closer to the "true" state, both at later times (blue line) and (for 4D-VAR) earlier times (green line). Note that re-evaluation at time 2 for 4D-VAR also makes use of data at time 1 and earlier.

So far there has been no mention of the intrinsic dynamics of the system. These are naturally introduced in the equation (3), assuming observations are included in the cost function $J(\mathbf{x})$ over a given time window, by discretizing the observation term in equation (3) at several times step indexed by i. Thus, the error function over the whole time window becomes:

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \sum_{i=0}^{n}(\mathbf{y}_i - H_i(\mathbf{x}_i))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_i - H_i(\mathbf{x}_i)) \tag{4}$$

with the actual process dynamics represented by:

$$\mathbf{x} = \mathbf{x}_0 \ \text{and} \ \forall i, \ \mathbf{x}_i = M_{0\rightarrow i}(\mathbf{x}) \tag{5}$$

where $M_{0\rightarrow i}$ is the forecast model for the system from time t_0 to time t_i. Equation (4) is the cost function used for a 4D-VAR method (see Figure 2). It is a nonlinear constrained optimisation problem, which, in the general case, requires extensive computations to be solved.

These equations introduce the essence of Data Assimilation methodology, including how information is gathered and processed. Note how Data Assimilation not only optimises the use of the data and of the numerical models, but also provides considerable scientific understanding of the processes in the system through the greater accuracy of the results.

Introducing some applications

The best known use of modern Data Assimilation is weather forecasting, which has developed rapidly with the improvement of numerical models and the availability of high performance computers. This complex application involves both large numerical models and heterogeneous measurements, from networks of satellites and ground measurements. 4D-VAR is currently the most widely used assimilation method for operational weather prediction. In the linear case, 4D-VAR gives the same result at the end of an assimilation window as a Kalman filter. The minimisation of errors in 4D-VAR can, to some extent, give the same quality of results as a Kalman filter, although the latter was designed for smaller computing resources. As in many complex systems, the dynamics of the atmosphere can be very sensitive to the initial state in certain conditions, and also to model errors. Note that the dynamical equations limit the magnitude of errors in meteorological variables, because the equations effectively correlate the variables over extensive regions within the domain. These errors and their correlations can be further reduced by ensuring that the initial data are compatible with the appropriate physics. In such a case, data are adjusted so as to be approximately in a state of dynamical balance [5, 11]. Finally, through sequences of forecasts, the initial, subsequent and also previous states of the atmosphere become known more accurately as the forecasts progress. Data Assimilation has taught us to realise that we can learn more about tomorrow's weather, if we first learn more about yesterday's weather, which we understand better today and even better tomorrow! Numerical weather forecasts are continually improving; currently 3-day forecasts are as accurate as 1-day forecasts 20 years ago.

These mathematical and computational procedures for making the best evaluation of incomplete data are not limited to the domain of meteorology. An autonomous robot needs to collect as much data as possible to evaluate its position and status in real time. To build precise real-time representations (maps) of its environment, it uses different filtering methods, and switches from one to another to improve the overall system fault-tolerance. Data Assimilation is applied in the same way as in meteorology and oceanography, but here the technique is called filtering and/or data fusion [10]. This approach is being compared with other methods, which do not use Data Assimilation, throught cross validation of the techniques, leading to significant quality improvement.

Another application has emerged in the critical field of modelling and evaluating state of a nuclear reactor core. The operational goal is to evaluate accurately the neutronic state in the core, through neutronic and thermohydraulic coupled equations, which like weather forecasts include parameters which have to be evaluated precisely through complex procedures. The measurements, from in-core instruments, are set up either for periodical checking or are permanently available. There are two ways in which Data Assimilation improves significantly operational security and optimal utilisation of resources (see [6]). In the operational application, information is collected from many instruments in real time, in order to estimate and then control the ongoing state of the whole system. The second one is to improve the design of a reactor model through optimal estimation of their parameters. In both applications, the important aspect is to optimise numerical estimation and to reduce their errors continuously.

Conclusion

Comparing these different applications provides some general insight on Data Assimilation topics. The mathematical methods can be improved through experience acquired in several applications within different fields and in several operational applications. For example, building correlation matrices from limited data and forecasts is often a practical difficulty. Exchanges between application topics have already been shown to be fruitful, e.g. through collaboration between climate modellers and nuclear engineers at CERFACS in Toulouse.

A general conclusion is that Data Assimilation is a powerful tool to improve the accuracy of modelling through the simultaneous use of data and model predictions. Data Assimilation can be applied in many physical domains when quasi-deterministic models exist. The levels of complexity of the Data Assimilation depend on the type of model and on the output requirement, from least squares minimisation in non-time dependent applications, to full stochastic optimisation in the most complex ones. Another advantage of Data Assimilation is that it can be used in domains where approximate models and incomplete data are available.

The optimal use of the information also feeds back into better understanding of the physical process and of the appropriate level of the quality of the modelling, as illustrated in some applications previously mentioned. Data Assimilation techniques can be applied selectively and sequentially to observations or to model improvements, leading for example to improvements in data processing, and in the design of model parametrisations.

Thus Data Assimilation techniques are evolving with new techniques in operational computation, and in design. Striking examples are the recent developments using ensemble methods for very large dimension models and for imprecise systems. Data Assimilation improves the reliability of these models, including those where the models for the same process (e.g. weather) are based on different parametrisations, and where many computations are performed simultaneously to allow for noisy predictions in highly non linear processes [7]. At the same time, Data Assimilation schemes are becoming more accurate by incorporating greater physical understanding into the operational models. They are used in the iterative corrections to the model as part of the optimisation procedure. The increasing size of Data Assimilation calculations, for predictions of fields with a huge number of calculation points, is necessary because the availability of data is not increasing at the same rate. This requires intensive use of high performance computing and optimised numerical algorithms, e.g. at the teraflop level for weather forecasting. Indeed, computations proceed while further data are becoming available.

In all fields of science and technology, including biology and social sciences, the aims are to make better use of the increasing volume and speed of data, even though it is always incomplete. In order to improve the accuracy of the models with increasing power of computation, Data Assimilation is becoming progressively more essential in computation and observation, and also in control and design of complex systems. Another strength of such a technique is that it can process different kinds of data including qualitative data, as well as interconnections between very different kinds of models based on physical, chemical, biological, and increasingly social science.❑

Acknowledgements

The authors thank R. Bannister, M. Fisher and R. Jaulmes for enlightening discussions on their respective topics presented at the ECMI 2008 Conference [12]. The authors also thank A. Weaver for fruitful discussions and exchanges on Data Assimilation, and his careful reviewing of this article.

REFERENCES

1 F. Bouttier, P. Courtier, *Data assimilation concepts and methods*, Meteorological Training Course, ECMWF, 1999
2 R. Daley, *Atmospheric Data Analysis*, Cambridge University Press, 1991
3 G. Evensen, *The Ensemble Kalman Filter: theoretical formulation and practical implementation*, Ocean Dynamics, 53(4), pp. 343-367, 2003
4 E. Kalnay, *Atmospheric Modelling, Data Assimilation and Predictability*, Cambridge University Press, 2003
5 A.C. Lorenc, *Modelling of error covariances by 4D-Var data assimilation*, Quarterly Journal of the Royal Meteorological Society, **129(595)**, pp. 3167-3182, 2003
6 S. Massart, S. Buis, P. Erhard, G. Gacon, *Use of 3D-VAR and Kalman Filter Approaches for Neutronic State and Parameter Estimation in Nuclear Reactors*, Nuclear Science and Engineering, 155(3), pp. 409-424, 2007
7 T. Palmer, R. Hagedorn, *Predictability of Weather and Climate: From Theory to Practice*, Cambridge University Press, 2006
8 O. Talagrand, *Assimilation of Observations, an Introduction*, Journal of the Meteorological Society of Japan, **75(1B)**, pp. 191-209, 1997
9 A. Tarantola, *Inverse Problem: Theory Methods for Data Fitting and Parameter Estimation*, Elsevier, 1987
10 S. Thrun, *Robotic mapping: A survey*, in G. Lakemeyer and B. Nebel (Ed.), Exploring Artificial Intelligence in the New Millenium, Morgan Kaufmann, 2002
11 A. T. Weaver, C. Deltel, E. Machu, S. Ricci, N. Daget, *A multivariate balance operator for variational ocean data assimilation*, Quarterly Journal of the Royal Meteorological Society, **131(613)**, pp. 3605-3625, 2005
12 E. Wilson, J. Norbury, A. Fitt (Ed.), *Proceedings of ECMI 2008 Conference*, Mathematics in Industry Series, Springer, 2009 (to be published)