Chapter 1

Introduction

1.1 What is data assimilation?

Loosely speaking, data assimilation may be simply viewed as a method of combining observations with model output. Why do we need data assimilation? Why not just use the observations? While the observations represent an estimate of the current reality, the problem is that we often want to predict the future. For that we need models. But when models are not constrained periodically by reality, they are of little value. Therefore, it is necessary to fit the model state as closely as possible to the observations, before a prediction is made.

When viewed as a method of simply combining different data sources (from observations and from model outputs), data assimilation can be seen as a least squares or regression problem. The method of least squares dates back to Gauss, who may be viewed as the father of estimation theory.

How should different data sources be combined? Well, if one is more accurate than the other, then the more accurate one should be given more weight. If we need to know the accuracy of the data sources, then we need to know something about the stochastic processes that produced the data. Since the data (model and observations) have errors, the underlying processes must be stochastic and not deterministic. Thus data assimilation can also be examined from the point of view of stochastic processes, signal processing or time series analysis. That is, assuming the observed and model variables are the same, the problem may be viewed as one of trying to extract a signal from noisy time series. However, the time series of atmospheric states are incompletely observed, so some interpolation is required to map the signal to the model grid. Thus two basic aspects of data assimilation are: filtering and interpolation.

It is clear that data assimilation can be studied from many viewpoints: estimation theory, signal processing, inverse theory, control theory, etc. Works in a particular field often share a similar vantage point. In this course, we will examine data assimilation from the point of view of estimation theory. As with the other viewpoints, estimation theory has a strong mathematical basis. I have chosen to follow this route because estimation theory naturally leads to the development of the Kalman filter, from which it is easy to derive other methods of data assimilation used in environmental applications.

A primary goal of data assimilation is to produce an "analysis", a model state that closely fits the observations. Thus the analysis should be provided on the model basis (grid or spectral coefficients, for example). The analysis can be used for diagnostic purposes, such as pollution monitoring or budget calculations. In Numerical Weather Prediction (NWP) centres, it is also used for

initiating a forecast. In this case, the analysis should also be compatible with the model dynamics, i.e. be somewhat balanced (some low Rossby number balance) so that the impulsive insertion of the analysis does not result in the generation of spurious gravity wave energy. This process of balancing the initial state is called initialization. A perfect analysis of the real atmosphere might not be compatible with the equations of the numerical forecast model. For example, the model might be hydrostatic, whereas the real atmosphere can be locally nonhydrostatic. Initialization adjusts the analysis to be compatible with the model. It should be noted that the term "data assimilation" typically includes the process of initialization. However, initialization will not be explicitly addressed in this course, so the interested reader is referred to Daley (1991) for a good introduction to the problem of initialization.

1.2 Examples

1.2.1 Navigation

One of the earliest applications which embraced the Kalman filter (newly developed in 1962) was that of designing navigation systems. It is still an important application of the Kalman filter. In a typical, self-contained navigation system, observations are collected and used to produce velocity corrections. The navigation system may be on board a rocket to Mars, a satellite in orbit, a missile over the arctic, a ship at sea or a car. A recent application is the navigation systems using GPS information that provide verbal directions to a car's driver. This type of application typically deals with linear models of small dimension (less than 100). For this type of problem, the Kalman filter is ideal.

1.2.2 Remote sensing

Remote sensing is the indirect measurement of atmospheric parameters from a distance. Satellite data retrievals are important products of space borne instruments. Typically, the actual measurement, say, the radiation at the top of a column of atmosphere, is nonlinearly related to the retrieved product, (a vertical temperature profile in this case). Moreover, there is no unique solution– many possible temperature profiles could result in the same measurement. Therefore, using some background information, such as a model forecast, the observations are combined with the background to produce a retrieval. Methods used include statistical interpolation and variational methods. The models that relate the measured variables to the retrieved variables are usually nonlinear. The dimension of the problem need not be large, but it can be. (See Houghton et al. (1984) for some examples.)

1.2.3 Geophysics

I don't know much about geophysical applications except that inverse theory is an important area of geophysics. There is a text on inverse theory by Tarantola (1987) that is useful for atmospheric as well as geophysicists. The type of problem encountered is when seismic activity is recorded and used to retrieve the structure or velocities within the Earth. The data is then arrival times and the retrieved quantities are densities or velocities.

1.2.4 Pollution source estimation

This is a very new and exciting application of data assimilation. Measurements of pollutants that are released into the atmosphere are very approximate, relying on information from industry and empirical models. If an industrial plant suddenly releases a pollutant, could we use observations of atmospheric constituents to trace the original source of the release? An obvious example is a nuclear accident that has not been reported. If the chemicals are inert, acting effectively like passive tracers, then Kalman smoothers or Four-Dimensional variational methods can be used to trace the original source. Recent examples include Pudykiewicz (1998), Zhang and Heemink (1996), Mulholland and Seinfeld (1995), and Robertson and Persson (1992). The model is typically a pollutant transport model with or without nonlinear chemistry. The data are measurements of mixing ratio. The problem is very underdetermined (not enough data), the models are usually large, complex and nonlinear.

1.2.5 Weather forecasting

Data assimilation has been used to obtain an initial state for integrating a numerical weather prediction model since the 1970's. Early methods included statistical (or optimal) interpolation and analysis corrections. In the 1990's, variational methods became increasingly preferred. With the time evolution of forecast error statistics, these methods become sub-optimal Kalman filters.

The literature in data assimilation for NWP (Numerical Weather Prediction) is very extensive, so I shall only refer to a few recent review articles. A good introduction is found in Daley (1991). Ghil and Malanotte-Rizzoli (1991) provide an overview of atmospheric and oceanic data assimilation. There are many nice introductory articles in the special issue of the Journal of Meteorological Society of Japan (1997).

Because my own background is in data assimilation for NWP, we shall often view the problem from this perspective. Although this is somewhat restrictive, the fact that other fields often follow the developments in the NWP field suggest that the NWP field is sufficiently advanced to serve as an example.

NWP models are very large, complex and nonlinear. Model states are typically on the order of 10^6 or 10^7. Scales of interest range from hours to two weeks. Physical processes such as radiation, boundary layer, and convection are included through complex parameterizations. Apart from the nonlinearity of the dry dynamics (nonhydrostatic primitive equations), the physical parameterizations are also highly nonlinear.

1.3 Underdeterminacy

One of the basic aspects of data assimilation problems for environmental applications is their underdeterminacy with respect to the observation set. Let us examine the NWP problem to illustrate this fact. The typical size of a model state vector is given by the number of gridpoints in the horizontal times the number of levels times the number of prognostic variables. Using current values from the Canadian Meteorological Centre's operational global forecast model (GEM), we obtain:

X - State vector

model	**x × y × z × var.**
CMC global	$400 \times 200 \times 28 \times 4$
TOTAL = N_X	9×10^6

On the other hand, what observations do we have? Of course it varies with time, so let's take some very rough estimates of what is used in a random 6-hour interval.

Y - Observation vector

data	**rep. × item × level**
sondes	$1000 \times 5 \times 15$
satem	$1000 \times 2 \times 15$
SM, ship, buoy	1000×5
aircraft	2000×3
sat. winds	2000×2
TOTAL = N_Y	1.2×10^5

Now it is quite clear the number of knowns (observations) is far smaller than the number of unknowns (model state).

$$\boxed{\frac{N_X}{N_Y} \approx 75}$$

Although this estimate is rather rough, the conclusion is valid for most large scale environmental applications. It is obvious that we cannot simply interpolate from the observations to the model state. The problem of determining the model state is grossly underdetermined. While we can hope that in the future, there will be many more observations, particularly from satellite instruments, computers keep improving in speed and memory so that the models keep growing to fit the available resources. The underdeterminacy problem will likely be with us for a long time.

In the rough estimate above, no account was taken of the spatial distribution of the observations. In reality, where there is lots of data (e.g. over North America) the problem may be overdetermined, and where there is sparse data (the tropics, southern hemisphere, the oceans) it will be underdetermined. Overall though, the problem is generally underdetermined.

How do we deal with the underdeterminacy problem? The obvious solution is to add more information. We can use *prior* information such as a model forecast. Another possibility is to analyse only certain scales of motion (e.g. large scales) assuming that nonlinear scale interactions will allow the smaller scales to be automatically developed. A third possibility is to consider observations over a time period. If the flow is largely advected, information can be spread downstream. If we have an accurate model of the flow's evolution, then all the data in a time interval can be assimilated while the unknowns still remain the same size (the model state at the beginning of the time interval).

1.4 What is the truth?

If we are going to separate the signal from the noise, in data assimilation, we must first decide what the signal is. Let s^e be an estimate of a signal, s, which is in error and let s^t be the *truth* or true atmospheric signal:

$$s^e(x,y,z,t) = s^t(x,y,z,t) + \epsilon^{\text{meas}}(x,y,z,t) \tag{1.1}$$

where s^t is the "true" state and ϵ^{meas} is the measurement error. In this course, measurement error is defined as the sum of errors associated with the measurements provided to the data assimilation system. Since our forecast model is not perfect, what we really want to know is the component of the truth that this model can resolve. Thus, we want the projection of the truth onto our discrete model basis. Let's call this s. Then, the estimate of the signal can be written as:

$$\begin{aligned} s^e(x,y,z,t) &= s(x,y,z,t) + s^t(x,y,z,t) - s(x,y,z,t) + \epsilon^{\text{meas}}(x,y,z,t) \\ &= s(x,y,z,t) + \epsilon^{\text{obs}}(x,y,z,t) \end{aligned} \tag{1.2}$$

where

$$\epsilon^{\text{obs}}(x,y,z,t) = s^t(x,y,z,t) - s(x,y,z,t) + \epsilon^{\text{meas}}(x,y,z,t).$$

Thus, in this course, we choose the signal to be the projection of the truth onto the model basis. Therefore the definition of *signal* or *truth* will depend upon the problem and the error ϵ^{obs} will include both *measurement* and *representativeness* errors. These two errors are usually lumped together and called *observation* error. Thus, representativeness error here refers to the error of representing the model state on a discrete grid as well as the error in interpolating from the model grid to the observation locations. In this section, we have assumed that the observed and model variables are the same. When this is not the case, representativeness error will also include the error of mapping the model variable to the observed one. This is discussed more fully in Ch. 3, section 3.3.

1.5 A Scalar Example with a single observation

To get a general feeling for what data assimilation is all about, let's start with a simple example. The simplest one consists of the estimation of a scalar quantity, say the temperature or ozone mixing ratio at a point in space. This simple example has been used many times in many articles (such as the review articles listed above). The example below comes from Daley (1991), chapter 4.6. Let us assume we also have another source of information, say a model forecast of the variable at the same time and location as the observation. In data assimilation, this is called the **background**, **trial field** or **first guess** field. In data retrieval, this information is called an **a priori** estimate. We can combine these two sources of information linearly, weighting the observations by W:

$$\mathrm{x}^a = \mathrm{x}^b + \mathrm{W}(\mathrm{x}^{\text{obs}} - \mathrm{x}^b). \tag{1.3}$$

The superscripts a and b refer to the analysis and background variables. Now consider the errors involved in this problem. To do that we must first define a *truth*. We always assume that the truth (the true value of the variable) exists, but that it is unknown and unknowable. (Even if we had a direct observation of the truth, the instrument would have some measurement error associated with it, so we would still only have an estimate of the truth. So the truth is a theoretical concept.) Subtract the *true* state from both sides of (1.3).

$$\mathrm{x}^a - \mathrm{x}^t = \mathrm{x}^b - \mathrm{x}^t + \mathrm{W}(\mathrm{x}^{\text{obs}} - \mathrm{x}^t - \mathrm{x}^b + \mathrm{x}^t)$$

The true state has superscript t. If we define the errors as:

$$\begin{aligned} \epsilon^a &= \mathrm{x}^a - \mathrm{x}^t \\ \epsilon^b &= \mathrm{x}^b - \mathrm{x}^t \\ \epsilon^{\text{obs}} &= \mathrm{x}^{\text{obs}} - \mathrm{x}^t \end{aligned}$$

then the analysis equation can be written as:

$$\epsilon^a = \epsilon^b + \mathrm{W}(\epsilon^{\mathrm{obs}} - \epsilon^b) \tag{1.4}$$

If we had many possible realizations of these errors, then we could take an ensemble average:

$$< \epsilon^a >=< \epsilon^b > +\mathrm{W}(< \epsilon^{\mathrm{obs}} > - < \epsilon^b >).$$

If we assume that the errors are unbiased, that is, $< \epsilon^b >=< \epsilon^{\mathrm{obs}} >= 0$, then $< \epsilon^a >= 0$. Thus, if the observation and background errors are unbiased, so is the analysis.

How do we determine the weight W? Obviously, we must use the information that we have. We've already incorporated the data sources, x^a and x^b into our estimate. We've used information about the biases of the errors. Now let us consider using the variances of errors of the data sources by forming an expression for the analysis error. Square (1.4) and take an ensemble average:

$$< (\epsilon^a)^2 >=< (\epsilon^b)^2 > +\mathrm{W}^2 < (\epsilon^{\mathrm{obs}} - \epsilon^b)^2 > +2\mathrm{W} < \epsilon^b(\epsilon^{\mathrm{obs}} - \epsilon^b) > .$$

We would like the analysis error variance to be as low as possible, so minimize $< (\epsilon^a)^2 >$ with respect to W and then solve for W. The derivative is

$$d < (\epsilon^a)^2 > /d\mathrm{W} = 2\mathrm{W} < (\epsilon^{\mathrm{obs}})^2 + (\epsilon^b)^2 > -2 < (\epsilon^b)^2 >= 0$$

where we assumed that $< \epsilon^b\epsilon^{\mathrm{obs}} >= 0$. This means that there is no correlation between the error in the background and the observation. If these are really two independent sources of information, that is an entirely reasonable assumption. However, both data sources are related because they are both a function of the true state. Nevertheless, this dependence is usually ignored and the correlation is assumed to be zero. Now with the definitions,

$$\begin{aligned} (\sigma^{\mathrm{obs}})^2 &= < (\epsilon^{\mathrm{obs}})^2 >, \\ (\sigma^b)^2 &= < (\epsilon^b)^2 >, \end{aligned} \tag{1.5}$$

we may write:

$$\mathrm{W} = \frac{(\sigma^b)^2}{(\sigma^b)^2 + (\sigma^{\mathrm{obs}})^2}. \tag{1.6}$$

$(\sigma^{\mathrm{obs}})^2$ is the observation error variance, and $(\sigma^b)^2$ is the background error variance. Thus we have determined the weight which produces a minimum analysis error variance and this weight depends on the relative accuracies of the observed and background estimates. Because the relative magnitudes appear in the definition (1.6), it is clear that $0 \leq \mathrm{W} \leq 1$. If the observation is perfect, $(\sigma^{\mathrm{obs}})^2 = 0$ and W=1. The observation is given maximum weight. If, on the other hand, the background is perfect, $(\sigma^b)^2 = 0$ and W=0; the observation is ignored. Also, with this particular choice of W, the analysis error variance is

$$< (\epsilon^a)^2 >= \frac{(\sigma^b)^2(\sigma^{\mathrm{obs}})^2}{(\sigma^b)^2 + (\sigma^{\mathrm{obs}})^2} = (\sigma^b)^2(1 - \mathrm{W}) = ((\sigma^b)^{-2} + (\sigma^{\mathrm{obs}})^{-2})^{-1} \tag{1.7}$$

The analysis equation may be written as:

$$\mathrm{x}^a = \mathrm{x}^b + \mathrm{W}(\mathrm{x}^{\mathrm{obs}} - \mathrm{x}^b), \qquad \mathrm{W} = \frac{1}{1+\alpha}.$$

where $\alpha = (\sigma^{\rm obs})^2/(\sigma^b)^2$. The analysis or estimate involves not only the data themselves but also information about the biases and variances of their errors. To better understand the analysis equation, let's consider some special cases. If the observation is very accurate, $(\sigma^{\rm obs})^2 << (\sigma^b)^2$, then $\alpha = 0$, W=1 and $\mathrm{x}^a = \mathrm{x}^{\rm obs}$. Thus, the weight given to the observation is 1, and the analysis is simply taken to be the observed value. On the other hand, if the observation is very inaccurate or the background is very accurate ($(\sigma^{\rm obs})^2 >> (\sigma^b)^2$), then $\alpha >> 1$, W=0 and $\mathrm{x}^a = \mathrm{x}^b$. Thus the weight given to the observation is 0 and the analysis is taken from the background value. Finally, if the observation and background error variances are equal, then W=1/2 and $\mathrm{x}^a = 0.5(\mathrm{x}^b + \mathrm{x}^{\rm obs})$. Thus the analysis is given by an average of the two data sources.

In summary, to estimate a scalar variable based on two data sources, we linearly combined the two pieces of information according to their accuracies. If the error statistics are exactly known, then the solution that provides a minimum of analysis error variance was found.

1.6 The influence of observations over space

Let's reconsider our simple example with a slight complication. Let us now suppose that our observation is not in Toronto, but in Montreal. We will have a background (model forecast) temperature for both cities, but our observation is only available in Montreal. We would like to use this information to get an estimate of the temperature in Toronto. Again, assume that the observation error is unbiased and has variance $(\sigma^{\rm obs})^2$. The background error is also assumed unbiased with variance $(\sigma^b)^2$ at both locations. The background error correlation between the two cities is ρ. As before, we can write an analysis equation:

$$\mathrm{x}^a_T = \mathrm{x}^b_T + \mathrm{W}(\mathrm{x}^{\rm obs}_M - \mathrm{x}^b_M) \tag{1.8}$$

where the subscripts T and M refer to Toronto and Montreal. We can also write the analysis equation in terms of errors, by subtracting the truth from both sides:

$$\epsilon^a_T = \epsilon^b_T + \mathrm{W}(\epsilon^{\rm obs}_M - \epsilon^b_M) \tag{1.9}$$

Typically the observation error includes representativeness error (the instrument may sample a volume far smaller than a model grid volume). This equation is the same as (1.4) except that the background and observation errors are needed in Montreal. As before, we can take the expectation of both sides and note that the analysis error is unbiased if both the background and observation errors are unbiased. To solve for the weights, we again need to use the information at our disposal: the variances of the errors. First we must square both sides of (1.9) and take expectations. Then, on minimizing with respect to W, we can obtain:

$$\mathrm{W} = \frac{< \epsilon^b_T \epsilon^b_M >}{(\sigma^b_M)^2 + (\sigma^{\rm obs}_M)^2} = \frac{\rho \sigma^b_T \sigma^b_M}{(\sigma^b_M)^2 + (\sigma^{\rm obs}_M)^2} \tag{1.10}$$

This weight is similar to (1.6) but now we need to know the background and observation errors at Montreal as well as the correlation between the background errors between Toronto and Montreal. If the observation were actually in Toronto (instead of Montreal), this correlation would be 1, all the M subscripts would be replaced by T's and the weight would be identical to (1.6). If Montreal were so far away that the background error correlation with Toronto's background error is zero, then the observation at Montreal would have no weight and no impact on the estimate of temperature in Toronto. Thus, the information of an observation can be combined with background information, just as before.

1.7 Conclusions

It is clear that data assimilation requires not only the observations and a background, but also knowledge of of error statistics (background, observation, model, etc.) and of our physical knowledge (forecast model, model relating observed to retrieved variables, etc.). The challenge of data assimilation is in combining our stochastic knowledge with our physical knowledge. In particular, for environmental applications, our models are huge, nonlinear and complex. The mathematics of data assimilation is straight-forward for linear models. For nonlinear models, we need to use our physical intuition as well. For large, complex models of any type, the success of any data assimilation algorithm lies is the details. "The devil is in the details" is what is often heard at data assimilation conferences. The choices one makes to render the problem computationally feasible remove the mathematical equivalence of various methods that we shall encounter in the next few weeks. As this course is really an introduction to data assimilation, we will ignore most of the "details" to focus on the mathematical results by considering only linear models. This will allow us to get a bigger picture of what data assimilation is really about, and how all these different methods (OI, 3DVar, 4DVar, KF) are related to each other.

Since we need to know about errors, we need to know something about the underlying stochastic processes that produce these errors. The most general information that we can have about a random variable or random process is given by its probability density function. In the next chapter, we will review some aspects of probability theory which will be needed to develop an understanding of estimation theory. Later on, these results will be extended to the time dimension when we consider stochastic processes.

REFERENCES

1. Cohn, S., 1997: An introduction to estimation theory. In M. Ghil, K. Ide, A. Bennet, P. Courtier, M. Kimoto, N.Nagata, N. Sato (Eds.): *Data assimilation in meteorology and oceanography: Theory and Practice*, Universal Academic Press, 147-178.

2. Daley, R., 1991: *Atmospheric Data Analysis* Cambridge University Press. 457 pp.

3. Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics*, **33**, 141-266.

4. Houghton, H. T., F. W. Taylor and C. D. Rodgers, 1984: *Remote sounding of atmospheres* Cambridge University Press. 343 pp.

5. Mulholland, M. and J. H. Sienfeld, 1995: Inverse air pollution modelling of urban-scale carbon monoxide emissions. *Atmospheric environment*, **29**, 497-516.

6. Pudykiewicz, J. A., 1998: Application of Adjoint tracer transport equations for evaluating source parameters. *Atmospheric environment*, **32**, 3039-3050.

7. Robertson, L. and C. Persson, 1991: On the application of four dimensional data assimilation of air pollution data using the adjoint technique. In: *Proc. 19th Technical Meeting of NATO-CCMS on Air Pollution Modelling and its Applications*, 29 Sept.-4 Oct., 1991. Ierapetra, Crete, pp. 365-373. Plenum Press, N.Y.