Inverse theory week 9: Data assimilation – a quick introduction. Part 2.

Hugh C. Pumphrey

November 15, 2007

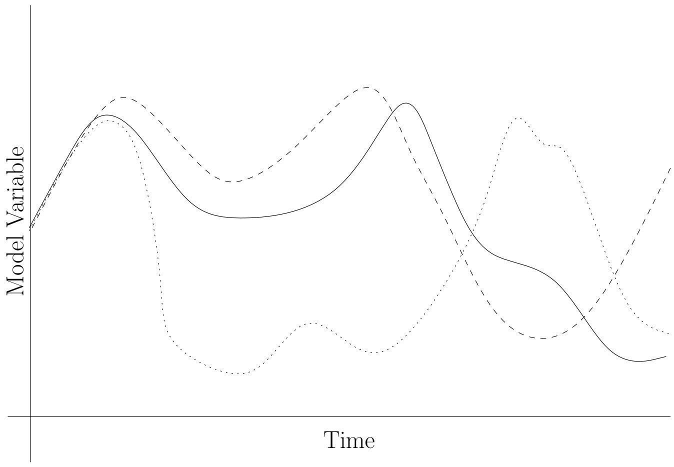

Figure 1: Very schematic suggestion of how an atmospheric quantity (real or modelled) might behave if it started off in three nearly identical states. The three cases are similar for a while, but then diverge.

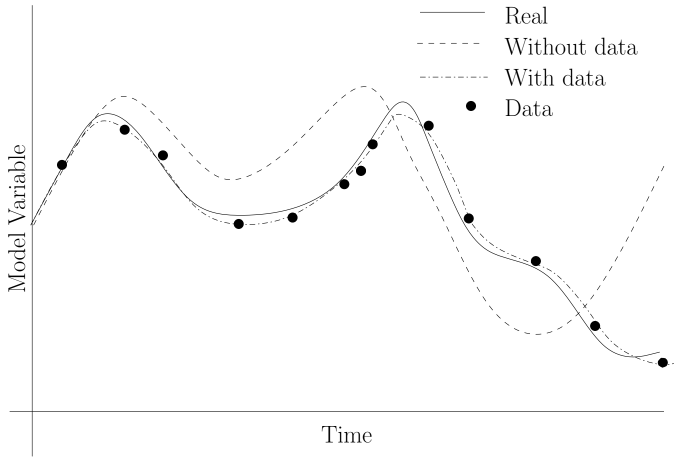

Figure 2: The idea of assimilation. A model will drift away from reality by itself. With a continuous injection of real data, it will stay in step with the real atmosphere.

1 Introduction

We have considered retrieval theory for several weeks, and have briefly examined how the atmosphere may be represented by a global circulation model. We note that:

- With the satellite measurements alone, we get (say) temperature at the measurement times and locations.
- With a GCM (on its own) we get temperatures, pressures, winds and humidities at every grid point, at every time step. But it won't do the same thing as the real weather for very long, because the equations which a GCM solves are *nonlinear*.

To get the best of both worlds, we would like to use measurements to persuade a GCM to keep tracking the real weather and not to wander off and do its own thing. But inserting measurements into a GCM is not as easy as it sounds.

2 Assimilation – the basic idea

We have already seen that for many nonlinear systems, the solution depends very sensitively on the initial conditions. Figure 1 suggests how a nonlinear system (or a model of it) might behave.

For a GCM to predict the weather for even a short time, it needs to be initialised with a good estimate of the state of the atmosphere. For it to continue to predict the weather, we would need to keep introducing information about the real atmosphere. By doing this, we should be able to keep nudging the model to keep it in line with what the real atmosphere is doing. The information about the real atmosphere could come from a variety of sources: satellite soundings, weather balloons, ground measurements and so forth. The idea is presented schematically in Figure 2. Note that the data and model together are presenting you with a useful value for the model variable at times when you have no directly measured data. This is the real value of assimilation, but with the atmosphere it is even better than it looks in the figure. One reason is that the data may be unevenly and sparsely spread over the globe, but the model fields are defined at every grid point. Another reason is that the remotely sensed data may only be for temperature, but if we nudge the model temperature to agree with the data, the model winds and pressures will also become more realistic.

3 Early assimilation schemes

Meteorological balloon launches and ground station measurements are traditionally made at fixed times of day, usually midday and midnight GMT. Early assimilation schemes were based on this assumption. A typical scheme might be as indicated in figure 3. The introduction of the data is not easy to get right. If we simply replace the

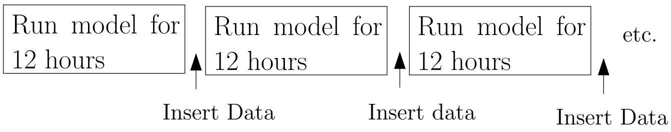

Figure 3: Typical early assimilation scheme. Data are introduced at certain synoptic times and the model runs freely between those times

temperature profile at a grid point with values taken from a nearby measurement, the GCM tends to react badly to the interference. This is because the resulting field disagrees with the equations of the GCM. Typically the GCM's response will contain large and unreaslistic oscillations. When these oscillations settle down, the model fields will be little changed from the values they would have had without the data. The term "rejection" is used for this reluctance of the model to accept the measurements

To avoid rejection and oscillations, the temperature data would be interpolated onto the model grid, and values for the other model variables would be estimated from it where possible. The resulting fields would be combined with the model fields at the end of the previous day, using a weighted mean, to provide a starting point for the next day.

4 4d-var: a modern assimilation scheme

The introduction of remotely sensed measurements was a serious problem for traditional assimilation schemes, for two reasons.

1. The measurements are not direct measurements of one of the model variables
2. The measurements are made continuously through the day, instead of at fixed synoptic times.

Point 1 can be addressed by retrieving a profile of the model variable (usually temperature) using standard retrieval theory. In the early days of satellite data, point 2 was addressed by assimilating the measurements at the nearest standard synoptic time. This is clearly a gross approximation. Modern assimilation schemes take account of the time of day of the measurements. A common scheme is called 4-dimensional variational assimilation, or 4D-Var.

Variational schemes are based on the minimization of a cost function. We consider the entire state of the GCM, that is, the winds, temperatures, pressures and humidities at every grid point, to be our state vector $\mathbf{x}$. We now ask: what is the model state at the start of our day, t_0, which would cause the model to have the best match to our data points? We typically try to introduce one day's worth of data at a time, repeating the whole process each day. The idea is illustrated in Figure 4. To achieve this, we consider the state at t_0 to be our "state vector" and find a suitable cost function of it which we can minimize. The cost function for 4D-var is:

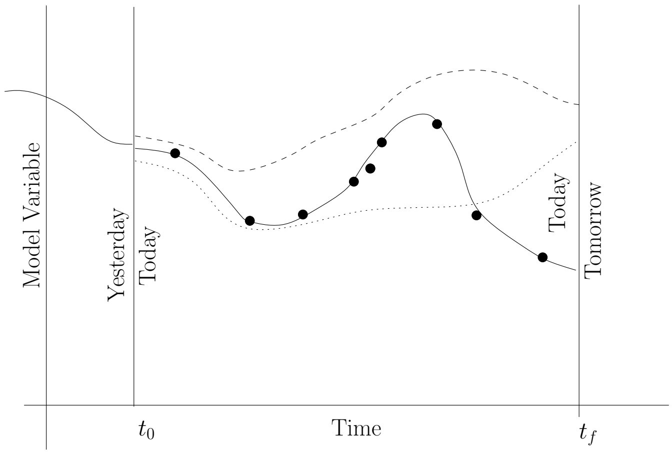

Figure 4: Illustration of a variational scheme. We want the value of the model state at t_0 which causes the model state to pass as close to our data points as possible as it is run forward.

$$
\begin{aligned}
J(\mathbf{x}(t_0)) &= (\mathbf{x}(t_0) - \mathbf{x}_B)^T \mathbf{S}_B^{-1} (\mathbf{x}(t_0) - \mathbf{x}_B) + \\
&\quad \sum_{t_0}^{t_f} (\mathbf{y}(t) - F(\mathbf{x}(t)))^T \, \mathbf{S}_E^{-1} \, (\mathbf{y}(t) - F(\mathbf{x}(t)))
\end{aligned}
$$

The formula looks very familiar to us: it is almost the same as the cost function in the MAP formula.

The first term penalizes the difference between the initial state and an a priori estimate $\mathbf{x}_B$. This is usually taken to be the state of the model at the end of the previous day. In assimilation circles it is usually called the background field, hence the subscript B. Evaluating this term is not trivial, because $\mathbf{x}_B$ can have over 10^6 elements in a large GCM. That means that $\mathbf{S}_B$ has 10^{12} elements and would require nearly 4 Terabytes of memory just to store it. It is usually necessary to assume that $\mathbf{S}_B$ is sparse in some way (i.e. that a large percentage of it contains 0s) and to store only the non-zero parts.

As in the MAP retrieval formula, the second term penalizes the differences between the measurements $\mathbf{y}$ and the estimate you would make of the measurements from the state vector. In this case, the forward model involves an interpolation from the GCM's grid points to the position at which the measurement was made. If the measurement was a direct measurement by a weather balloon, that is all that is needed. If the measurement is a remotely sensed one, then F should ideally contain an interpolation to the measurement point, followed by a radiative transfer calculation. Until very recently this was too expensive and it was normal to retrieve temperature profiles from measurements and assimilate the retrieved profiles. Direct assimilation of radiances is preferable, though.

The measurement forward model, F , is not all that we need to calculate the second term. We also need to run the GCM forwards from the start of the day to find $\mathbf{x}(t)$ for every time at which a measurement was made. The process of minimizing J is a very complex one – we can not write down a simple formula like the MAP formula. The essential aim is exactly the same, but on a *much* larger scale. The problem is nonlinear, but an inverse Hessian or Marquadt-Levenberg approach is too expensive. A steepest descents approach is often used – more iterations are needed, but each iteration is cheaper than for a Marquadt-Levenberg scheme.

Further reading

- Ross Bannister, *Elementary 4-D Var*, `http://www.met.rdg.ac.uk/~ross/Documents/Var4d.html` Describes 4D-var in greater detail than is possible here, including a description of how the cost function is minimized.
- Clive Rodgers, *Inverse methods for atmospheric sounding*, Chapter 8 provides a brief summary of assimilation from the inverse theory point of view.
- R. Daley, *Atmospheric Data Analysis*, Rather an old-fashioned book nowadays, not much on variational methods. Covers pre-variational methods in some detail, explaining why you can't just shove in the data and expect it to work.