Assignment 6: Introduction to Data Assimilation

Due: 19 December 2019 (updated version)

Objectives

In this assignment, you will apply fundamental concepts of data assimilation to highly-simplified yet realistic scenarios to (a) gain experience with applying advanced data assimilation concepts to real-world problems, (b) demonstrate assimilation outcome sensitivity to the specification of the background and observation error variances as well as the ensemble background estimate spread, and (c) examine how linear relationships between variables are used to update ensemble estimates, in spite of the inherent shortcomings of sampling error and atmospheric non-linearity therein.

Quick Reference: 'Point' Data Assimilation

Recall that for any variable x, the least squares combination of an observation x_o and a background x_b to obtain an analysis x_a takes the form:

$$x_a = (1 - k)x_b + kx_o = x_b + k(x_o - x_b)$$

where k is the weighting factor and is equal to the variance of x_b (the background error variance) weighted by the total (background plus observation) error variance:

$$k = \frac{\sigma_b^2}{\sigma_o^2 + \sigma_b^2} \qquad \text{where } 1 - k = \frac{\sigma_o^2}{\sigma_o^2 + \sigma_b^2}$$

The analysis x_a is equal to the background plus an optimally weighted innovation, measuring the departure of the observation x_o from the background estimate x_b.

The resulting analysis error variance takes the form:

$$\sigma_a^2 = \frac{\sigma_o^2 \sigma_b^2}{\sigma_o^2 + \sigma_b^2}$$

Quick Reference: 1-D Ensemble Data Assimilation

Ensemble filters used for atmospheric data assimilation, including the ensemble Kalman filter and ensemble adjustment Kalman filter, apply Bayes' theorem to assimilate and observation and thus update an ensemble of background estimates. To do so, these algorithms assume that the ensemble of background estimates is normally distributed with mean and variance derived from the ensemble estimates themselves, wherein the latter is defined from the departure of the ensemble background estimates from their mean (and thus is representative of the so-called meteorology of the day). The algorithms also assume that the observation can be expressed as a normal distribution with mean equal to the observation value and variance equal to the assumed observation error variance.

A normal distribution for any variable x can be expressed as:

$$\exp\left(\frac{-\left(x-\mu_x\right)^2}{2\sigma_x^2}\right)$$

where μ_x is the mean of x and σ_x^2 is the variance of x.

This distribution can be normalized such that it represents a probability distribution function:

$$\frac{1}{\sigma_x\sqrt{2\pi}}\exp\left(\frac{-\left(x-\mu_x\right)^2}{2\sigma_x^2}\right)$$

The application of Bayes' theorem to ensemble atmospheric data assimilation has the general form:

$$\textit{Posterior Probability} = \frac{\underline{\textit{Prior Probability * Observation Probability}}}{\textit{normalization}}$$

Here, the normalization factor is simply the area underneath the curve cut out by the product in the numerator. Thus, the posterior (i.e., analysis) probability is the normalized product of the prior (or background) and observation probability distributions, which are both normal distributions. Note that the product of any two normal distributions is also a normal distribution!

The normal distribution from the product of two normal distributions has mean and variance of:

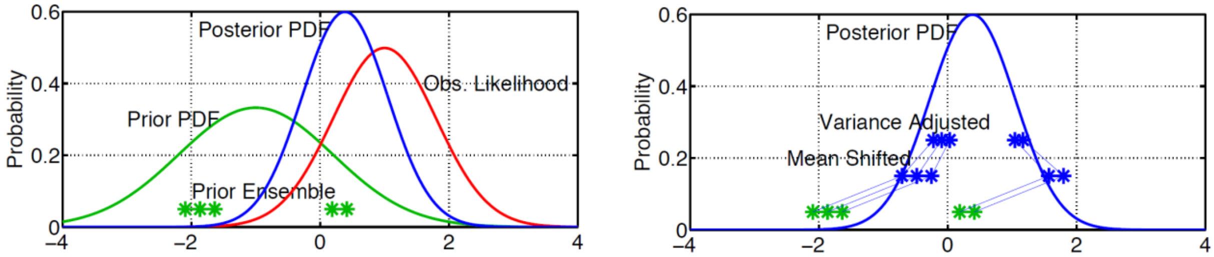
$$\mu_a = \frac{\mu_b\sigma_o^2 + \mu_o\sigma_b^2}{\sigma_o^2 + \sigma_b^2} = \mu_b\,\frac{\sigma_o^2}{\sigma_o^2 + \sigma_b^2} + \mu_o\,\frac{\sigma_b^2}{\sigma_o^2 + \sigma_b^2} = \mu_b\left(1-k\right) + \mu_0 k = \mu_b + k\left(\mu_0 - u_b\right)$$

$$\sigma_a^2 = \frac{\sigma_o^2\sigma_b^2}{\sigma_o^2 + \sigma_b^2}$$

Here, subscripts of *b* indicate background, subscripts of *o* indicate observation, and subscripts of *a* indicate analysis or posterior to make each specific to the ensemble data assimilation application.

Note that both are ***identical*** to the least-squares formulation! Thus, to this point, the 1-D ensemble Kalman filter is ***identical*** to the least-squares formulation as applied to the background ensemble mean μ_b and observation μ_o. How are the individual ensemble member estimates adjusted once the analysis mean has been computed, however? As we will discuss later this semester, the specifics vary between flavors of the ensemble Kalman filter, of which there are several.

The ensemble adjustment Kalman filter is a *deterministic* filter, as there is a clear correspondence between the ensemble member background and analysis distributions. Specifically, the ensemble background estimates are first uniformly (i.e., maintaining their distribution and spread) shifted so that they have a mean equal to the new analysis mean, then uniformly scaled so that they have a variance equal to the new analysis variance. This is depicted visually in the figures below.



Mathematically, this takes the form:

$$\vec{x}_a = \mu_a + \left(\frac{\sigma_a}{\sigma_b}\right)\left(\vec{x}_b - \mu_b\right)$$

wherein the ensemble analysis estimates are equal to the analysis mean plus the departure of each ensemble background estimate from the background mean scaled by the reduction in variance from the background to the analysis.

Quick Reference: Multivariate Ensemble Data Assimilation

In the 1-D ensemble data assimilation case, an observation for a given variable and location is used to update ensemble estimates for that same variable and location. In the more common multivariate case, an observation for a given variable and location is used to update ensemble estimates of *many* variables and locations!

Consider a simplified multivariate scenario, where an observation at one is location used to update an ensemble of estimates for another variable at another location. First, the observation and its variance are used to compute the analysis mean, variance, and ensemble estimates for the ***observed*** variable at its given location. Next, the slope of the linear regression line between the ensemble *background* estimates for the observed variable and the ensemble *background* estimates for the *variable to be updated* is determined. This slope is equal to:

$$\beta = \frac{\mathrm{cov}\left(\vec{x}_{b,o}, \vec{x}_{b,u}\right)}{\mathrm{var}\left(\vec{x}_{b,o}\right)} = \frac{\dfrac{\sum_{i=1}^{n}\left[\left(x_{b,o}^{i} - \mu_{b,o}\right)\left(x_{b,u}^{i} - \mu_{b,u}\right)\right]}{n-1}}{\sigma_{b,o}^{2}}$$

where…

- $\vec{x}_{b,o}$ is the ensemble of background estimates for the observed variable (what we previously called $\vec{x}_b$)
- $\vec{x}_{b,u}$ is the ensemble of background estimates for the variable to be updated
- n is the total number of ensemble members

- $x_{b,o}^{i}$ is the *i*th ensemble member's background estimate for the observed variable
- $\mu_{b,o}$ is the mean of $\vec{x}_{b,o}$ (what we previously called μ_b)
- $x_{b,u}^{i}$ is the *i*th ensemble member's background estimate for the variable to be updated
- $\mu_{b,u}$ is the mean of $\vec{x}_{b,u}$
- $\sigma_{b,o}^{2}$ is the variance of $\vec{x}_{b,o}$ (what we previously called σ_b^2)

If the background estimates are uncorrelated, their covariance and thus the slope will be zero.

Once this slope has been determined, the ensemble analysis estimates for the variable to be updated ($\vec{x}_{a,u}$) is given by:

$$\vec{x}_{a,u} = \vec{x}_{b,u} + \beta\left[\left(\mu_{b,o} - \vec{x}_{b,o}\right) + \left(\frac{\sigma_{a,o}}{\sigma_{b,o}}\right)\left(\vec{x}_{b,o} - \mu_{b,o}\right)\right]$$

Note the similarity to the 1-D ensemble example. In that case, $\vec{x}_{b,o}$ and $\vec{x}_{b,u}$ are the same, as there is only one variable being updated. Because $\beta = 1$ (the background estimate exactly matches itself), the above equation collapses to that for $\vec{x}_a$ above. In the multivariate case, however, we must first find the *adjustment* (not the new values) for the observed variable, then scale the adjustment based on its relationship to the variable to be updated, from the ensemble analysis for the variable to be updated can finally be obtained.

Helpful Resources

Weather Underground provides archived hourly weather observations on their website:

https://www.wunderground.com/history/daily/**SITE**/date/**YYYY-MM-DD**

where **SITE** is replaced by the four-letter station ID (e.g., KMKE), **YYYY** is replaced by the four-digit year, **M** is replaced by the one- or two-digit month, and **D** is replaced by the one- or two-digit day.

Iowa State University provides archived model output statistics (MOS) forecasts on their website:

https://mesonet.agron.iastate.edu/mos/fe.phtml

I strongly recommend only requesting one time at once through the time selection menu.

Finally, the University of Wyoming provides archived radiosonde observations on their website:

http://weather.uwyo.edu/upperair/sounding.html

Questions

In all questions that follow, note that more weight is given to interpretation than to just the assimilation itself. Please review the notes above and try to truly understand the assimilation process and its impact.

1. At 6:52 pm CDT (2352 UTC) 22 September 2017, the 2-m air temperature at Milwaukee, WI was 82.9°F. Assume that the observation is representative of its surroundings and that the thermometer used to measure this temperature is well-calibrated, allowing us to specify the observation error *standard deviation* as 3°F.
    a. (5 pts) The simplest first guess estimate is that given by climatology. Consider the ~2352 UTC 22 September 1981-2010 climatology for Milwaukee. Compute the climatological mean temperature. Compute the background error variance from the 1981-2010 climatology forecast errors. Last, determine *k* and compute the resulting analysis temperature and variance.
    b. (5 pts) A slightly better first guess estimate is that given by persistence, where we use last hour's observed temperature as that at the current time. Use the 5:52 pm CDT 22 September 2017 observation as your first guess. Compute the background error variance from the persistence forecast errors over the 72 h prior to 5:52 pm, using only hourly observations (ending at :52) in doing so. Last, determine *k* and compute the resulting analysis temperature and variance.
    c. (5 pts) A first guess may also come from a numerical model's forecast. Here, we wish to use the 6 h GFS MOS forecast from the 1800 UTC 22 September 2017 run as our first guess. Compute the background error variance from the 6 h 1800 UTC 31 August to 19 September 2017 MOS forecast errors. Last, determine *k* and compute the resulting analysis temperature and variance.
    d. (11.66 pts) Describe the variation in the weighting given to the observation from each method. How does each analysis temperature compare to the observation? Why?
    e. (11.67 pts) Describe how each analysis error variance compares to the background and observation error variances. What does this say about the confidence of the analysis estimate relative to that of the background and the observation?
    f. (11.67 pts) One of the most important yet most challenging aspects of data assimilation is the accurate specification of the background error variance or covariance matrix. Comment on the methods used in (a) – (c) to specify the background error variance. Do these seem appropriate given the background estimate in each? How could they be improved upon, if they could at all? Discuss why.

2. For ensemble data assimilation, an ensemble of background estimates is most commonly derived from a short-term (e.g., 6 h) ensemble forecast. The 1800 UTC 22 September 2017 twenty-member GFS Ensemble forecast is available at on *comet* in the following directory:

/oasis/projects/nsf/wim108/evans36/assignment6

You don’t need to copy all three files in this directory to one of your own, but you will want to make sure that you are in this directory when completing the following steps. The .grib2 file contains the data; the .ctl file is a GrADS control file, and the .idx file is an index file used to map the unorganized data in the file to the describing control file.

If you wanted to extract out the value of the 2-m temperature variable TMP2m for a given location (by latitude and longitude) for a given time for all ensemble members, you might run the following series of commands in GrADS:

```
open gefs.ctl
set time hhZddMONyyyy
set lat ##
set lon ##
set z 1
set e 1 20
set gxout print
d TMP2m
```

Here, hh is replaced by a two-digit hour, dd is replaced by a two-digit day, MON is replaced by a three-letter month identifier, yyyy is replaced by a four-digit year, and the two ##s are replaced by latitude (positive = °N) and longitude (0-360°E – so values in °W need to be handled with care) respectively. The gxout print statement tells GrADS to print the data as text into the terminal/command window. The set e 1 20 statement tells GrADS to consider ensemble members 1-20 within the data file.

Variables found on multiple isobaric levels would exchange set z 1 for set lev ####, where the #### is replaced by a three- or four-digit isobaric level (hPa). The control file provides a full list of the available variables, including their units.

a. (12.5 pts) Extract the 2-m temperature (convert to °F) at 0000 UTC 23 September 2017 (the 6-h forecast) for Milwaukee from all twenty ensemble members. Determine the mean and variance of these estimates. Given the same observation as in question 1, find the mean and variance of the posterior probability distribution function. Describe how the posterior mean and variance compare to both the prior and observation means and variances.
b. (12.5 pts) Compute and list the analysis temperatures. Summarize the changes (sign and magnitude) that occur from the background to the analysis ensemble values.
c. (25 pts) Consider a hypothetical example where ensemble members 4 and 10 instead indicate that rainfall has cooled the forecast 2-m temperature at Milwaukee to 67.1°F and 67.7°F, respectively. Compute the new background mean and variance, the new analysis mean and variance, and the resulting analysis ensemble estimate values. How does the results differ from in question 2(a-b) for both the outliers and non-outliers?

Why do you believe these changes have occurred? Have outlier members adequately been dealt with? To what extent are assuming a normally distributed set of background estimates and linearly adjusting individual ensemble member estimates reasonable for cases such as this? Why?