Assignment 6: Introduction to Data Assimilation

Due: 5 December 2017

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:

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

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 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. In other words, the background error variance is defined by the departure of the ensemble background estimates from the background mean (assumed to be "truth" for this purpose), such that as long as the underlying assumptions are reasonable it reflects the meteorology of the day. These algorithms further assume that the observation can be expressed in terms of 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:

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

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

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

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

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

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

Here, the normalization factor is simply the area underneath the curve cut out by the product in the numerator.

The product of any two normal distributions is itself a normal distribution. Thus, the posterior PDF is the PDF of the normal distribution given by the product in the right-hand-side numerator.

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

$$\mu_a = \frac{\mu_b\sigma_o^2 + \mu_o\sigma_b^2}{\sigma_o^2 + \sigma_b^2} \qquad \sigma_a^2 = \frac{\sigma_o^2\sigma_b^2}{\sigma_o^2 + \sigma_b^2}$$

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.

The ensemble adjustment Kalman filter updates the individual ensemble estimates after computing the posterior probability distribution function by adjusting them so that they have equal mean and variance to the posterior probability distribution function. The resulting adjustment vector has the form:

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

The first right-hand side term is the departure of each background estimate from the analysis mean. The second right-hand side term is related to the departure of each background estimate from the background mean. This departure is a measure of the spread of the background estimates. This is then weighted by the ratio of the analysis and background standard deviations. In the rare instances where the two are equal, the adjustment simplifies to μ_a - μ_b – adjust each member by the difference between the analysis and background means.

The ensemble of analysis estimates $\vec{x}_a$ is equal to the ensemble of background estimates $\vec{x}_b$ plus the adjustment increment vector, i.e.,

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

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 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 error variance is used to compute the background estimate adjustment vector 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 (at its location) and the ensemble background estimates for the variable to be updated (at its location) is determined. This slope is equal to:

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

where…

- $\vec{x}_{b,o}$ is the ensemble of background estimates for the observed variable
- $\vec{x}_{b,u}$ 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 $\vec{x}_{b,o}$
- $x_{b,u}^i$ 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 $\vec{x}_{b,o}$

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

Once this slope has been determined, it is multiplied by the background estimate adjustment vector determined for the observed variable. This product is the background estimate adjustment vector for the variable to be updated. Adding this to the ensemble background estimates for this variable allows one to obtain the posterior analysis for the variable to be updated.

Helpful Resources

Weather Underground provides archived hourly weather observations on their website:

https://www.wunderground.com/history/airport/**SITE**/**YYYY**/**M**/**D**/DailyHistory.html

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. (3 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. (3 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. (3 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. (7 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. (7 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. (7 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:

   http://derecho.math.uwm.edu/classes/NWP/assignments/Assignment6/

   You will want to download all three files in this directory. The .grib2 file contains the data; the .ctl file is a GrADS control file, while the .idx file is an index file used to map the data in the file (which is not self-describing) 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) 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. (7.5 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. (7.5 pts) Adjust each background estimate and list the prior and analysis temperatures. Summarize the changes (sign and magnitude) associated with the adjustment.
c. (15 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 posterior mean and variance, and the resulting posterior 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?

3. The observed 500 hPa height at Bismarck, ND at 0000 UTC 23 September 2017 was 5720 m. Assume that the observation is representative of its surroundings and that the instrument used to obtain the observation is well-calibrated, allowing us to specify an observation error *standard deviation* of 20 m.
    a. (8 pts) Physically, what relationship – if any – would you expect between the 500 hPa height at Bismarck, ND and 2-m temperature at Milwaukee? Would you expect their correlation magnitude to be small or large, and the correlation to be direct or inverse? Why? Feel free to include a hypothetical weather setting with your answer if you find it helps with your explanation, although such a setting is not necessary.
    b. (4 pts) Extract the 500 hPa height at 0000 UTC 23 September 2017 (6-h forecast) for Bismarck, ND for all twenty GFS Ensemble members. Determine the mean and variance of these estimates. Given the observation noted above, find the posterior mean and variance. Describe how the posterior mean and variance compare to the prior mean and variance.
    c. (4 pts) Compute the background adjustment as in question 2(b) and list the values of the adjustments. Compute the linear correlation coefficient between background Bismarck 500 hPa height and Milwaukee 2-m temperature estimates, using the data from question 2(a) before assimilation for the latter. Compute the slope of the linear regression line between the background estimates. Compute the adjustment for the 2-m temperature background estimates and determine the new analysis estimates

as well as the analysis mean and variance. How does the analysis variance compare to the corresponding background variance?

d. (8 pts) Describe how the ensemble estimates of 2-m temperature were updated in light of the distant 500 hPa height observation. How does the extent to which they were updated compare to your expectations of their physical relationship in 3(a)?

e. (8 pts) In this question and the conceptual framework that underlies it, we assumed a linear relationship between the two variables considered. Consider the wide range of variable types that may be assimilated. Do you expect linear relationships to be appropriate for all such types? Provide one example for each of where it is and is not reasonable. Given that most ensembles only have twenty to fifty members, do you expect the relationships between variables – linear or otherwise – to be well-sampled by the ensemble? How might these concerns impact assimilation quality?

f. (8 pts) From questions 1(e), 2(a), and 3(c), you have likely identified a consistent relationship between the analysis and background variances over all applications. What is this relationship? Most ensemble adjustment Kalman filters will assimilate thousands to tens of thousands of observations at once. How might this relationship affect analysis spread characteristics after assimilating thousands of observations? Discuss the impact this could have on the weight given to observations during the assimilation process.