Optimal Interpolation (OI) Assimilation of Station Precipitation Data into the ERA40 Precipitation Estimates

Author: J.L. McCreight (mccreigh@nsidc.org)

Introduction

New estimates of total monthly precipitation have been generated for the time period of January 1979 through July 2002 at points above 45 degrees north. These estimates combine the spatially complete "background" field of ECMWF ERA-40 precipitation with in-situ observations of precipitation which vary in availability in both space and time. The methodology used to combine these "background" and observed fields is Optimal Interpolation (OI) assimilation (e.g. Bouttier and Courtier, 1999).

The ERA-40 analysis already includes some data assimilation and represents a best guess at precipitation over the Arctic (Serreze et al, 2005). ERA-40 precipitation data include large-scale rain rate, convective rain rate and snowfall rate. These data were interpolated to an equal area grid from the N80 grid using bilinear interpolation. The new grid is known as the EASE 100km grid (Brodzik and Knowles, 2002), each cell has the square root of area ~=100.27 km. Extensive observations of precipitation in the Arctic (and above 45 degrees north) have been compiled into a database spanning the years 1950-2003. This data set is informally known as the "Big Merge" and is used in Serreze and Etringer (2003) and Serreze et al 2005. With the big merge are included scant observations from Russian drifting Ice stations in the Arctic Ocean. (Some day perhaps these will be officially merged too.) Reported monthly totals of station precipitation were adjusted following the seasonally and spatially variable precipitation gauge correction factors of Legates and Willmont (1990).

Since the observation data is total precipitation at monthly resolution, total monthly ERA-40 precipitation (summing over convective rainfall, large-scale rain fall and snowfall) is found for each month and the assimilation described below is performed specifically on total monthly precipitation. While the data may be left at this level, we have disaggregated the total monthly precipitation back to it's constituent variables (large-scale rainfall, convective rainfall, and snowfall) and also disaggregated it in time, back to the original ERA-40 6-hourly resolution. Our disaggregation scheme is a very simple one, we follow how the total precipitation was initially distributed or weighted for the monthly ERA-40 and apply this to the new magnitude of total precip. In other words, the temporal distribution of the precipitation variables is left unaltered except for the total area under the curves (ie the height of the curves) while the ratios of the total precipitation between the 3 variables remain constant.

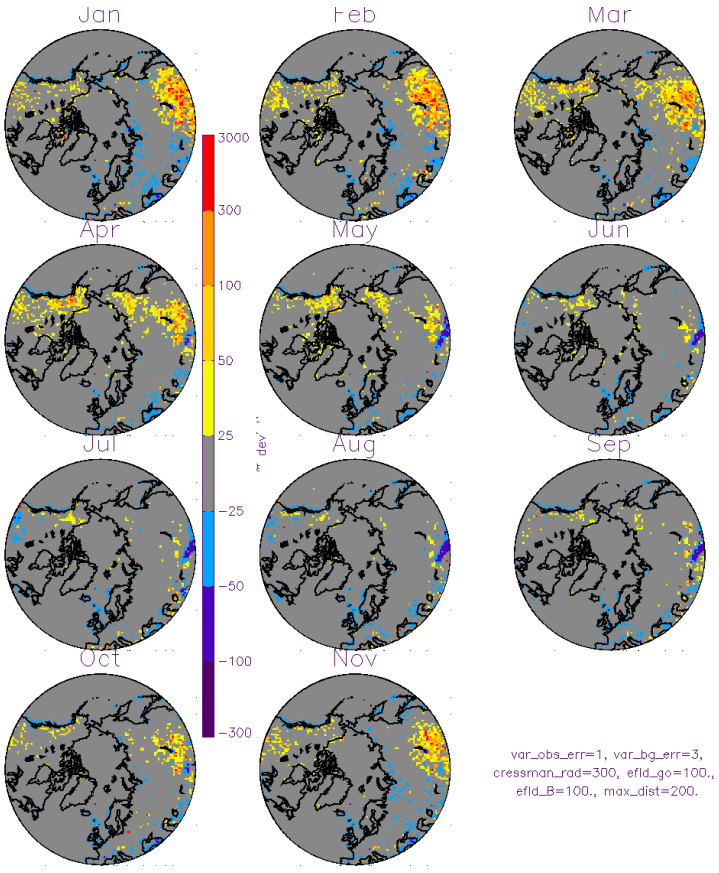

Figure 1. Percent difference of time averaged assimilated precip compared ERA-40 precip for assimilation A.

This study focuses on the satellite era, post

1978, when ERA-40 is believed to be more reliable. The analysis is computed on the same, equal-area grid to which the ERA-40 data have been interpolated in the study of Slater et al (2007). The assimilated product is essentially an update of the background ERA-40 data towards the observations when and where they are available.

The assimilated field confirms the ERA-40 biases found by Serreze et al. (2005) as can seen by comparing figures 1 and 2 (full size png files accompany this document, externally) to figure 3 (this is a copy of figure 7 of Serreze et al (2005)). Where observations are available on a regular basis, biases in ERA40 can be estimated by the average difference of ERA-40 and the assimilated product over the time period of study. In areas where observations do not exists, the ERA-40 values are left unaltered. In areas where data occur infrequently, the background also remains relatively unaltered over the period of study and no real bias can be inferred.

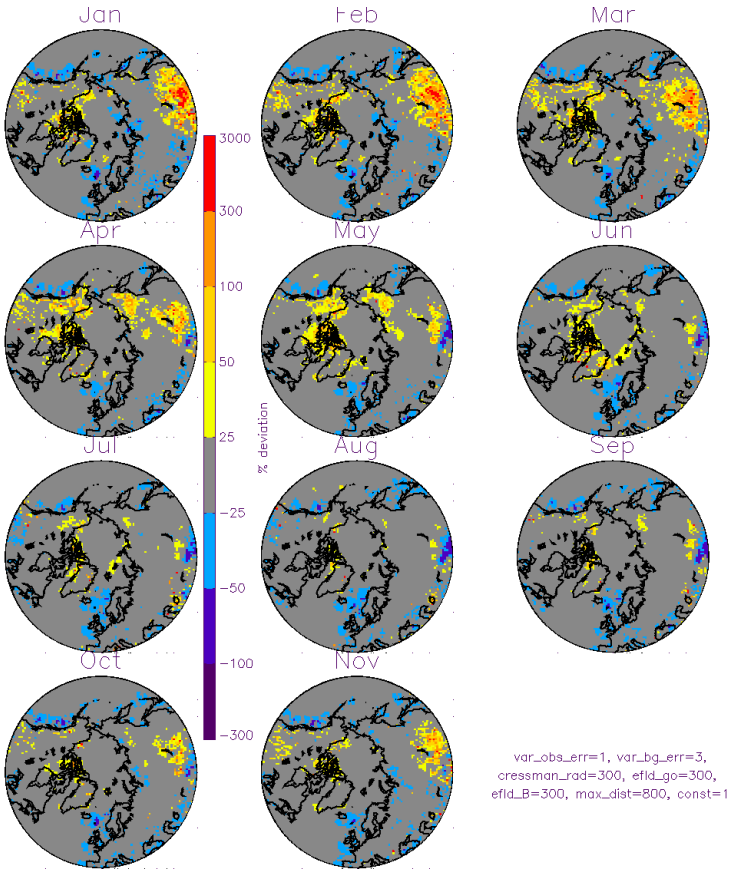

Figure 2. Percent difference of time averaged assimilated precip compared to ERA-40 for assimilation B.

Methodology

This section pertains to Optimal Interpolation (OI) assimilation, other simple details of the method have been outlined in the introduction. Optimal Interpolation assimilation derives an "analysis" or assimilated value by adjusting the background value at a given time according to its error with N selected observations at that time. The adjustment to the background value is termed the analysis increment. In the following equation x_a is the resulting analysis and x_b is the background value (ERA-40). The analysis increment depends on several quantities; y is a vector of N selected observations, H represents an interpolation operator which translates the background value onto the N observations, and W is the set of optimal weights (or gain) of the errors used to adjust the background value by linear combination of the N differences on the observations:

$$x_a = x_b + W^T(y - H[x_b]) \qquad (*)$$

The interpolation operator, H, is subjectively determined by the analyst. In our case, Cressman interpolation was used to interpolated the background field onto the observations as the background is evenly distributed in space.

A basic assumption of OI is that only a few observations are important for determining the analysis increment. The intuitive idea is that only observations near a grid point should matter for assimilation at that point and that there is no correlation (or teleconnection) with points far away. This is made formal by our choice of spatial structure function, which describes the correlation of any two points in

space based on their separation. Here we use a Gaussian form

$$u_{ij} = \exp\left(-\frac{1}{2}\left[\frac{d_{ij}}{d_{ef}}\right]^2\right)$$

where u_{ij} is the correlation between points *i* and *j*, their distance is denoted d_{ij} and the parameter d_{ef} is the e-folding distance. The distance at which observations may be considered insignificantly correlated to a background value is determined by some threshold of correlation which is in turn determined by the choice of e-folding distance. One can choose either a distance or a correlation threshold at which to exclude observations. We have used several choices for d_{ef} ranging from 100km to 500km. The value of 300km is used by ECMWF in their land surface data assimilation (I assume this includes precip). For precipitation, this value represents assimilation at the synoptic scale.

The structure function plays a central role in OI. To determine the "optimal"weights, *W*, intervariable correlation is used to account for the effects of "clustering" in the data, hence the name OI. For example, say there were 10 observations reported within 1km of each other at 10 km due north of our assimilation point. Say there were also 1 observation reported 10km due south of the point. The structure function deems that the 10 observations to the north are all very highly correlated and adjusts their weights so as not to overemphasize this region and neglect the single observation to the south.

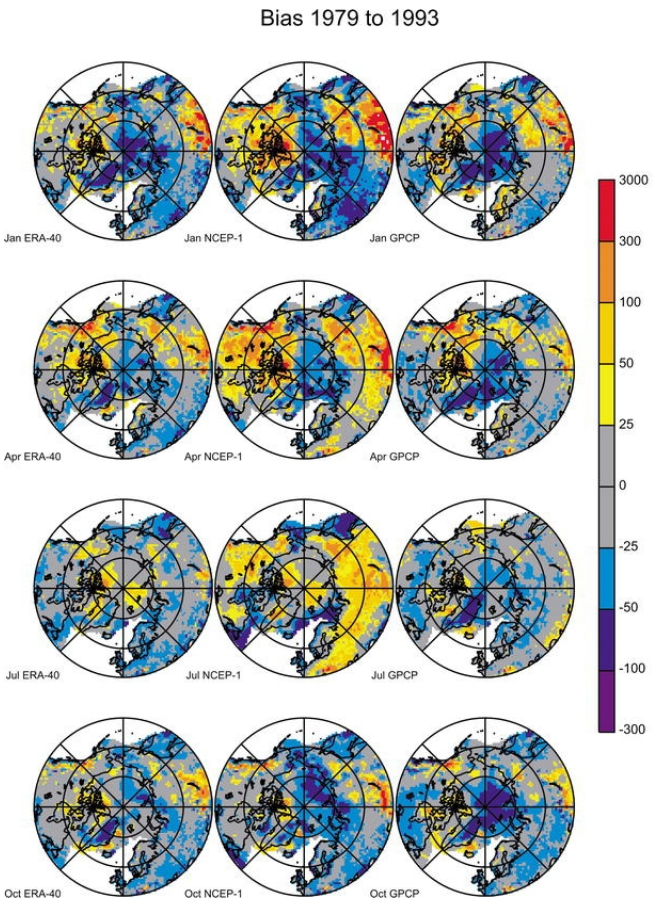

Figure 3. Biases in reanalysis products of Arctic precipitation, taken from Serreze et al, 2005, figure 7.

To specify clustering in the data, both grid point to observation distances (correlations implied immediately via the structure function) and inter-observation distances (correlations) are computed. Grid point to observation correlations are given in the vector *b* (length *N*) and inter-observation correlation is provided by the matrix *B* (dimension *NxN*). These correlations are transformed to covariance by multiplying each by the (scalar) background error, σ_b^2 . To express the variance in the observations themselves, we form the *O* matrix, $I\sigma_o^2 = O$, where *I* is the identity matrix and σ_o^2 is the (scalar or vector) error associated with the measurement of the observations. Now the optimal weights can be considered as the linear transformation that maps observation covariances, *B*+*O*, on to grid to observation covariances, *b*. Alternatively, the weights may be seen as the grid to observation covariances normalized by the observation covariances:

$$(B+O)W = b$$
$$W = (B+O)^{-1} b$$

In this way, *W* accounts for clustering in the observations when weighting the differences on the observations for determining the analysis increment (at point of assimilation) in equation (*). The optimal weights are similar to those of ordinary Kriging in that *W* is a "best linear estimator." Here the term best means that variance in the

interpolation error is minimized by W. This is achieved by the appropriate linear combination of differences on the observations assuming a spatial structure function.

In formulating the covariances, the background error, σ_b^2 , and the observation error, σ_o^2 , were introduced. While these numbers are separate, only their ratio affects change in *W*. This ratio specifies the amount of confidence placed on the background versus the observations. The free parameters in OI assimilation include interpolation method, *H*, structure function, *u*, and the error covariances, σ_b^2 and σ_o^2 . Error covariances are the most problematic to estimate as changing the ratio exhibits significant control over the results of the assimilation. If there is some objective measure of the background and measurement errors, then this becomes a less arbitrary parameter. However, when assimilating at large scales and for such a spatially heterogeneous phenomena as precipitation, one is confronted by the representativeness problem of the estimates which has to be factored into the errors. Even if we know the actual rain gauge error, this is not an appropriate for estimating the observation error at scales of hundreds of kilometers. In practice, one subjectively makes decisions about the free parameters of OI and produces results for several different scenarios, effectively probing the sensitivity of the analysis to these. The analyst draws conclusions from this variety of results.

Results

Results for 2 assimilations are shown in figures 1 and 2 (full size png versions of these are included). The figures depict the time averaged (1979-2002) percent difference in precipitation at each pixel for the individual assimilation runs as compared with the ERA-40. (Assimilation A corresponds to the data in directory BM_assim.07.06.15.12.46.07 and assimilation B to those in BM_assim.07.06.22.11.01.09.) The parameters of each assimilation are sketched in the bottom right hand corner of each figure and these are reproduced in table 1. The relative trust in the observations as compared with the background values is the same in both OI assimilations at 3:1, which is the ratio background error variance to observation error variance. This ratio was deemed a reasonable way to weight the competing fields and to generate a product with substantial new information. However, the effect of the structure function, which essentially determines the spatial scale of representativeness of each assimilated observation, is investigated in the 2 different assimilations. In assimilation A (figure 1) the e-folding distances of both structure functions (between observations, ie efld_B, and between the background and observed points, ie efld_go, as these are referenced in the figures) are set to 100km. In assimilation B, these

Assimilation and code name	Observation error variance	Background error variance	Cressman interpolation radius	E-folding distance for grid-obs covariance	E-folding distance for observation covariance
A: 07.06.15.12.46.07	1	3	300km	100km	100km
B: 07.06.22.11.01.09	1	3	300km	300km	300km

Table 1: Summary of parameter values used in the included precipitation assimilations.

values are both set to 300km. (The max_dist parameter effectively sets a cut-off when searching for points to participate in the assimilation at a give point. This is accordingly adjusted in these two assimilations.) In varying the e-folding distances between the assimilations, we investigate the effect of assuming the observed values are more locally representative (A) versus them being more synoptically representative. In the later case, the observations affect a more widespread change from the background field, which is evidenced in the figures. In addition to being more widespread in their influence, the

magnitude of the adjustment (xa-xb) in the later assimilation (B) will tend to be greater in areas of sparse observations. This is because of increased covariance at the same distance. However, this can be effect can be offset by clustering of competing observations and the increase in the intra-observation covaraince. So, while the spatial effect of the change of parameters is fairly obvious in the figures, the impact of of the parameters on the adjustment is less obvious and has to be understood in terms of how the observations are clustered.

The results show good agreement the biases found by Serreze et al (2005) in their figure 7, reproduced here as figure 3 (also included as a jpg file). In areas with data on an inter-annual basis, the assimilation reproduces the biases. Areas of no change (gray) in the assimilation figures 1 and 2 indicate one of two things, either that there is insufficient data over the period of study to significantly adjust the background field in the time averaged way presented, or that there is actually very little bias in the ERA-40 product relative to the observations. A map of the temporal density of data would help separate these two cases.

Summary

The Optimal Interpolation assimilation of an expansive set of precipitation observations ("The Big Merge") into the ERA-40 precipitation product provides a way to mathematically generate new, best-guess fields of precipitation over the Arctic. The results express the findings of Serreze et al (2005) in a form which can be used for analysis or modeling.

References

Bouttier, F. and Courtier, P. 1999. Data assimilation concepts and methods. *ECMWF Meteorological Training Course Lecture Series*, 59 pp.

Brodzik, M. J. and K. W. Knowles. 2002. "EASE-Grid: a versatile set of equal-area projections and grids" in M. Goodchild (Ed.) *Discrete Global Grids*. Santa Barbara, CA, USA: National Center for Geographic Information & Analysis.

Legates, D. R., and C.J. Willmott, 1990. Mean Seasonal and spatial variability in gauge-corrected, global precipitation. *Int. J. Climatol.*, **10**, 111-127

Serreze, M.C. and Etringer, A.J., 2003. Precipitation characteristics of the Eurasian Arctic drainage system. *International Journal of Climatology,* 23, 1267-1291.

Serreze, M.C., Barrett, A. and Lo, F., 2005. Northern high latitude precipitation as depicted by atmospheric reanalyses and satellite retrievals. *Monthly Weather Review,* 113, 3407-3430.

Slater, A. G., T. J. Bohn, J. L. McCreight, M. C. Serreze, and D. P. Lettenmaier (2007), A multimodel simulation of pan-Arctic hydrology, *J. Geophys. Res.*, 112, G04S45, doi:10.1029/2006JG000303