INTELLIGENT DATA THINNING ALGORITHMS FOR SATELLITE IMAGERY

Bradley Zavodsky[3], Steven Lazarus[2], Xiang Li[1], Mike Lueken[2], Michael Splitt[2], Rahul Ramachandran[1], Sunil Movva[1], Sara Graves[1], William Lapenta[4]

[1]Information Technology and Systems Center, University of Alabama in Huntsville, Huntsville, AL
[2]Florida Institute of Technology, Melbourne, FL
[3]Earth System Science Center, University of Alabama in Huntsville, Huntsville, AL
[4]NASA/Marshall Space Flight Center, Huntsville, AL

ABSTRACT

This paper presents a study on intelligent data thinning for satellite data. In particular, the focus is on the thinning of the Atmospheric Infrared Sounder (AIRS) profiles. A direct thinning method is first applied to a synthetic data set in order to identify optimal data selection strategies. Experiments on synthetic data suggest that a thinned data set should combine homogeneous samples, and high gradient and variance of gradient samples for optimal performance. This result leads to the modification of our previously developed Density Adjustment Data Thinning algorithm (DADT). The modified DADT (mDADT) algorithm is used to thin the AIRS profiles. Experiments are conducted to compare the thinning performances of mDADT with two simple thinning algorithms. Experiment results show that mDADT algorithm performs better than the two simple thinning algorithms, especially over the regions of significant atmospheric features.

1. INTRODUCTION

Despite advances in data assimilation, it remains unclear how to best identify and remove redundant data. Redundant data are defined here as either 1) exhibiting characteristics of linear dependence or 2) being of sufficient density that exceeds the resolution of the assimilation grid. It is common to remove (thin) high spatial and temporal resolution observations—especially data obtained from satellite or radar observing systems. Despite the obvious benefits of high spatial resolution in data sparse regions, this practice occurs because large data volumes can have an adverse affect on the computational costs and functionality of a real-time forecast/analysis system. Operational data reduction methods often tend toward a crude (but computationally efficient) methodology often referred to as subsampling. However, recent work in the area of adaptive thinning such as top-down clustering and thinning through estimation is promising [1]. In this approach, observation removal is contingent upon minimizing the impact on an analysis with error estimates calculated by differencing analyses constructed with and without all of the observations. The observation that causes the smallest increase in analysis error is removed. An advantage of this method is that it directly uses an estimate of analysis quality to drive the data reduction. However, the method is potentially expensive and the degree of optimality remains ambiguous due to both compromises necessary for practical application and the omission of background (first-guess) or observation error in the analysis. Using a simple one-dimensional framework consisting of analyses and thinning, we directly address the issue of optimality. The lessons learned from the synthetic data tests are applied to thermodynamic profiles retrieved from the Atmospheric Infrared Sounder (AIRS) and compared to two unintelligent approaches.

2. SYNTHETIC EXPERIMENT

The following synthetic experiment is designed to determine the optimal thinning strategy using synthetic observations along with explicitly defined truth and background fields. The thinned observations are assimilated using the variational approach described by Lorenc [2]. An idealized 1D truncated Gaussian with 35 observation is sampled in order to obtain the configuration of 5 observations that yields the best analysis. A direct thinning method that takes the total number of possible thinning configurations and selects the configuration that yields the optimal analysis (as determined by the lowest root mean square error between the analysis and truth) is applied. Here, approximately 325,000 unique spatial combinations of 5 observations that can be obtained from a single realization. The first guess field is set to that of the base of the Gaussian function (dashed line Fig. 1). The observations are created by adding white noise to the truth where the observation-to-background error variance is set to 0.25. Analyses were run for each of 4 different length scales (2Δx, 4Δx, 6Δx, and 8Δx). For 5 observations, the best analysis (lowest Mean Square Error, MSE), occurs for a length scale of 4Δx, where Δx is the grid point separation (Fig. 1). For the most part, the optimal observation configurations are those that retain data at the peak, within the gradient, and at points where the gradient changes significantly

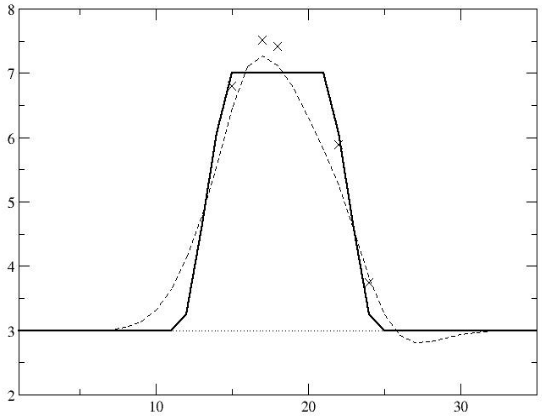

Fig. 1: Truncated Gaussian (solid black curve), 1DVAR analysis (dashed curve), first guess (red dashed line) and optimal observation locations (X's) for analysis length scale of 4Δx. See text for details.

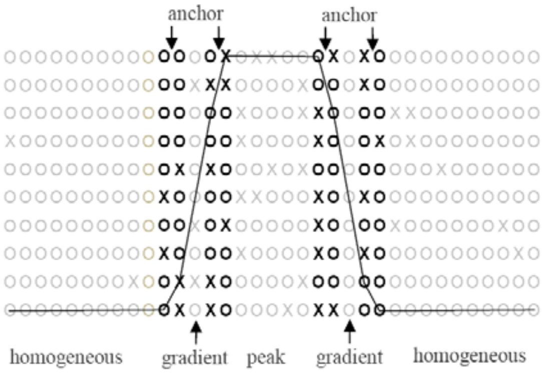

Fig. 2: Each row represents the observation subset with the lowest analysis MSE produced from a set of 325,000 possible configurations of a single realization with an observation error variance of 0.25 and analysis length scale of 4Δx. X's and O's denote retained observations and grid points respectively.

(referred to hereafter as anchor points, bold X's/O's in Fig. 2). In particular, the observations have an affinity for anchor points with 50% of the total observations located at these points versus 26% and18% within the peak and homogeneous regions respectively, and 6% at the gradient (inflection) points (Fig. 2). These results (as well as others not shown) suggest that the thinned data samples should combine homogeneous points, gradient points, and anchor points for optimal performance. Furthermore, optimal thinning performance also depends on key elements of the analysis system itself—including the length scale (L) and the quality of both the background and observations.

3. APPLICATION TO REAL OBSERVATIONS: AIRS PROFILES

Results gleaned learned from the synthetic thinning experiments are applied to three-dimensional thinning of Atmospheric Infrared Sounder (AIRS; [3]) thermodynamic profiles. Each AIRS profile contains a pressure level below which data are of questionable quality, and only data above this pressure level are used for thinning and analyses. The Advanced Regional Prediction Sys-tem (ARPS) Data Analysis System (ADAS [4]) is used to perform analyses using a background from a short-term Weather Research and Forecasting (WRF) model forecast. Error covariances used for the background are standard short-term forecast errors cited in the ADAS documentation, and the errors for the AIRS profiles are based on estimates cited by Tobin et al. [5]. Separate error estimates are used for land and water soundings. For these experiments, all thinning algorithms reduce the total number of AIRS profiles to approximately 10% of the original number of observations.

3.1. Thinning Strategies for Real Observations

For an intelligent thinning algorithm to be successful, it should outperform unintelligent thinning methods. Hence, for comparison purposes, AIRS profiles are thinned using both intelligent and unintelligent thinning strategies. The first of two unintelligent approaches applied here is a simple thinning method that takes every 9th profile within the data set without regard data density. Eight different permutations of the simple thinning are performed because a different data set can be created depending upon where one begins the subselection of observations. A second unintelligent thinning method is applied in which profiles are randomly selected. A search radius is used to ensure that retained observations are thinned at a user-specified distance. Ten different permutations of the random thinning were performed to create an ensemble input for the analysis system.

The intelligent thinning methodology used herein is a modified version of the thinning approach discussed in Ochotta et al. [6] and introduced in Splitt et al. [7] as Density Adjusted Data Thinning (DADT). The DADT systematically builds a thinned set of observations from an initially empty set using a priority queue consisting of variance intensity between observations and their neighbors. The DADT has been modified using guidance from the synthetic experiment described in Section 2 indicating that an optimal thinned data set should consist of a combination of homogeneous, gradient, and anchor samples. The modified DADT (mDADT) considers homogeneous and anchor samples in addition to gradient samples and uses the thermal front parameter (TFP, [8]) to help detect gradient change. Observations targeted for retention are those in regions where the absolute value of the TFP is determined to be significant (i.e., larger than some specified user threshold). Homogeneous samples are then selected in reverse order from the priority queue (i.e. lower variance observations). The three metrics are applied independently with the retained observations removed from the queue following each step. The total number of observations depends on a user-specified observation retention rate. For a given retention rate, the thinning performance will also depend on the proportion of each type of sample in the thinned data set. For the mDADT, the

AIRS profiles were thinned on pressure surfaces creating a pseudo three-dimensional thinning when the levels are recombined. Because ADAS analyzes for specific humidity, the thinning algorithm was applied directly to equivalent potential temperature rather than temperature and moisture separately.

3.2. Results

While it is generally not possible to replicate the quality of a full data analysis by the way of a thinned data set, it is our desire to produce an analysis with a shorter runtime that is as close to the full analysis as possible. Because the optimal analysis length scale will change as a result of the thinning, an average horizontal observation separation is calculated using the distance between each observation and its nearest neighbor. The average distance is then input to a simple linear function derived from the direct method (Section 2), which relates the observation separation to the optimal analysis length scale.

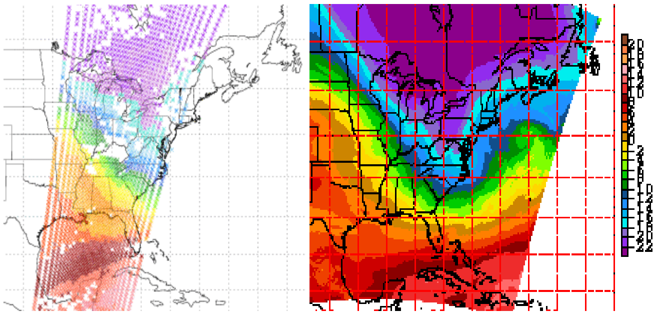

Fig. 3. 700 hPa temperatures from AIRS profiles (left) and the WRF forecast used as analysis first guess (right) showing gradients over upper Midwest and northern Gulf of Mexico.

Figure 4 shows the impact of applying simple, random, and mDADT thinning on a 700 hPa temperature field from 12 March 2005. The first guess field (Fig. 3) exhibits relatively tight gradient regions across the upper midwest and northern Gulf of Mexico. The three thinning algorithms retain observations in different locations with the gradients over Wisconsin and Illinois and over the Gulf of Mexico more clearly depicted in the mDADT thinning (Fig. 4c) than in either the simple (Fig. 4a) or random (Fig. 4b) thinning. Differences between the full analysis and each of the thinned analyses (Figs. 4d-f) indicate the largest discrepancies are produced by the simple thinning (Fig. 3e). The random thinning (Fig. 4e) produces smaller differences than the simple thinning while the mDADT (Fig. 3f) outperforms both techniques. Overall, the differences for all three

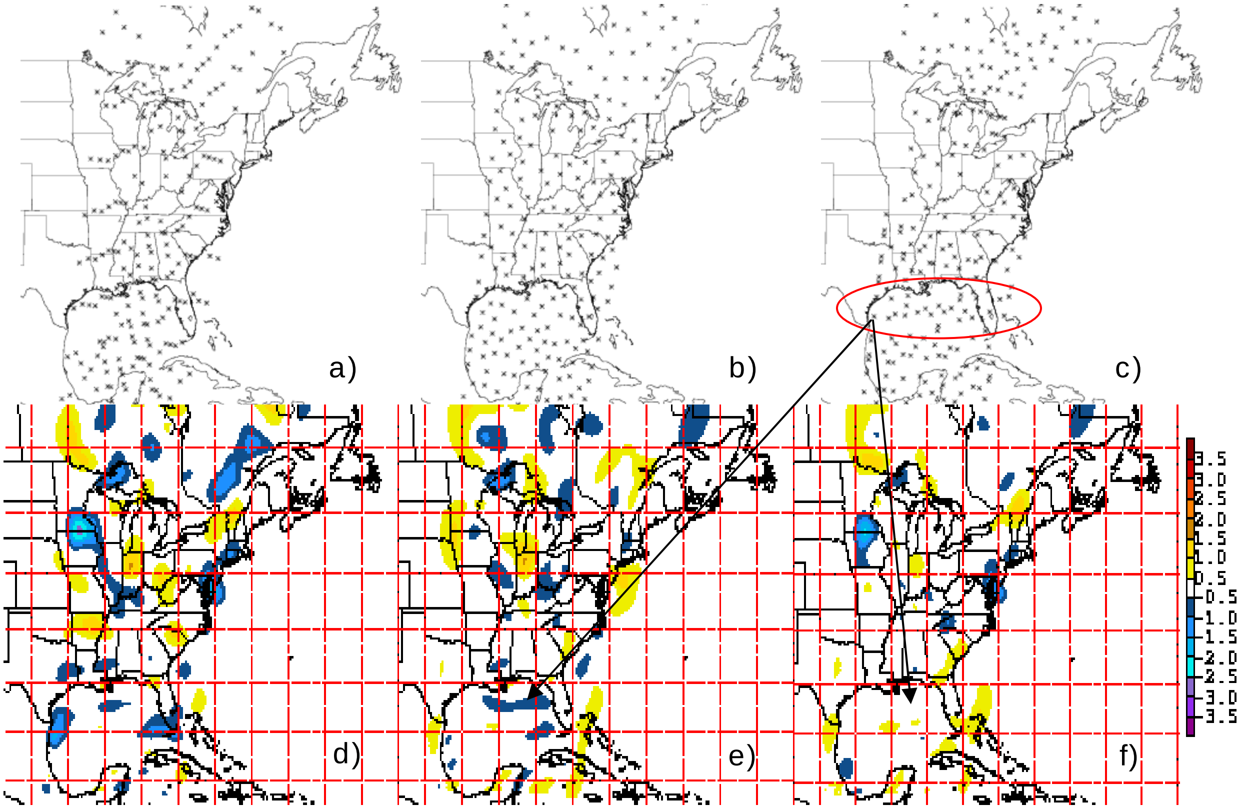

Fig. 4: Thinning results for the 700 hPa temperature valid at 0800 UTC on 12 March 2005. The top row illustrates the location of retained observations for (from left to right) simple subsampling, random subsampling, and mDADT. The bottom row shows the temperature analysis differences (^{o}C) between the full analysis and the analysis obtained from the retained observations in the top row. The simple and random thinning analyses shown are for the best permutations of each.

Table 1: Quantitative assessment of the thinning methods. MSEs compare the full and each thinned analysis. Brackets denote the spread of each metric over multiple runs. For comparison, the full analysis contains 211,232 observations, uses L = 81 km, and has an analysis time of 4,506 seconds.

	Simple	*Random*	*mDADT*
# OBS	[22,939, 23,894]	[22,062, 23,019]	23,572
ALYS time (s)	[1,185, 1,262]	[1,649, 1,763]	1,386
L (km)	[142,144]	[177, 179]	155
T MSE	[0.3292, 0.3653]	[0.3053, 0.3168]	0.3010
q MSE	[1.1266, 1.1604]	[1.0284, 1.0595]	1.0420

techniques are less than ±1.5°C (except over Minnesota for the simple thinning where there is no data retention). The largest differences between the unintelligent and intelligent techniques occur over the upper Midwest, which is coincident with a gradient region. For this particular analysis, the mDADT preserves the gradient in the analysis better than either the simple or random thinning.

To quantitatively assess the differences between the three techniques, a MSE between the full analysis and each thinned analysis has been generated. These results are summarized in Table 1. The MSEs for the simple thinning are much larger than those for both the random and mDADT thinning. The mDADT thinning outperforms all of the randomly-thinned temperature analyses but a few of the random observation permutations do produce a superior moisture analysis. The thinned analyses reduce the computation time by 60-70%; however, the mDADT appears to be the best compromise between speed and retention of a larger number of observations. The increased analysis speed over the random thinning is due to the shorter length scale prescribed to the mDADT thinning. At this time, it is difficult to assess whether the improvement over the random thinning is due to a superior thinning method or to the slightly larger number of observations retained by the mDADT method (# OBS, Table 1). However, the gradient regions appear to be better resolved in the mDADT analysis—indicating intelligent thinning provides a more representative set of observations than either simple or random thinning, especially over regions of potential interest.

4. CONCLUSIONS/FUTURE WORK

A 1D synthetic data test using a direct thinning method suggested that a technique that includes homogeneous, gradient and anchor points is necessary to produce optimal thinning results. As a result, a modified version of the DADT (mDADT) was created that applies a combination of local variance and a TFP parameter to select observations in homogeneous, gradient, and anchor point regions. This algorithm was applied to AIRS profile observations leading to better temperature analyses than either simple or random thinning when compared to an analysis generated from a full data set. In contrast, only some of the moisture analyses generated via the mDADT thinning were better than the random thinning analyses. Efforts continue with regard to fine-tuning the mDADT algorithm. This includes extensive systematic testing using two-dimensional synthetic data sets and demonstration of the algorithm capabilities with respect to real-time data dissemination.

5. REFERENCES

[1] N. Dyn, M. S. Floater, A. Iske, “Adaptive thinning for bivariate scattered data,” *J. Computational and Appl. Math.*, pp. 505-517, 2002.

[2] A.C. Lorenc, “A Global Three-Dimensional Multivariate Statistical Interpolation Scheme,” *Mon. Wea. Rev.*, pp. 701-721, 1981.

[3] H. H. Aumann, M. T. Chahine, C. Gautier, M. D. Goldberg, E. Kalnay, L. M. McMillin, H. Revercomb, P. W. Rosenkranz, W. L. Smith, D. H. Staelin, L. L. Strow, and J. Susskind, “AIRS/AMSU/HSB on the Aqua Mission: Design, Science Objectives, Data Products, and Processing Systems,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 253-264, 2003.

[4] K. Brewster, “Implementation of a Bratseth analysis scheme including Doppler radar”. Preprints, *15th Conf. on Weather Analysis and Forecasting.* Amer. Meteor. Soc., Boston, MA, pp. 596-598.

[5] D. C. Tobin, H. E. Revercomb, R. O. Knuteson, B. M. Lesht, L. L. Strow, S. E. Hannon, W. F. Feltz, L. A. Moy, E. J. Fetzer, and T. S. Cress, “ARM site atmospheric state best estimates for AIRS temperature and water vapor retrieval validation,” *J. Geophys. Res.*, D09S14, pp. 1-18, 2006.

[6] T. Ochotta, C. Gebhardt, V. Bondarenko, D. Saupe, W. Wergen, “On thinning methods for data assimilation of satellite observations,” Preprints, *23rd International Conference on Interactive Information Processing Systems (IIPS)*, Amer. Met. Soc., San Antonio, TX, 2007.

[7] M. Splitt, S. Lazarus, M. Lueken, R. Ramachandran, X. Li, S. Movva, S. Graves, B. Zavodsky, and W. Lapenta, “An Improved Data Reduction Tool in Support of the Real-Time Assimilation of NASA Satellite Data Streams”. Preprints, *12th Conf. on IOAS-AOLS*, Amer. Meteor. Soc., New Orleans, LA, 2008.

[8] R. J. Renard and L. C. Clarke, “Experiments in numerical objective frontal analysis,” *Mon. Wea. Rev.*, pp. 547-556, 1965.