An Improved Data Reduction Tool in Support of the Real-Time Assimilation of NASA Satellite Data Streams

Michael Splitt[1], Steven Lazarus, Mike Lueken
Florida Institute of Technology, Melbourne, FL

Rahul Ramachandran, Xiang Li, Sunil Movva, Sara Graves
Information Technology and Systems Center, UAH, Huntsville, AL

Bradley Zavodsky
Earth System Science Center, UAH, Huntsville, AL

William Lapenta
MSFC/NASA, Huntsville, AL

1. INTRODUCTION

Efforts to reduce the size of large data streams such as radar or satellite observations are essential in operational meteorology. These sources are computationally expensive to assimilate and may contain observations with spatially correlated error (Bergman and Bonner, 1976; Liu and Rabier 2002; Ochotta et al. 2005). The high-volume data may also be redundant or at a resolution higher than that of the analysis grid itself. As a result, a systematic reduction of data not only improves analysis efficiency but can also potentially increase analysis quality. Here, the term 'redundant' refers to observations that provide little in the way of information with respect to an analysis. For example, observations may be unnecessary in the presence of a good first guess field or in regions of quiescent or relatively nondescript conditions. We present continuing work from a NASA funded project designed to examine data thinning via the application of automated Intelligent Data Thinning (IDT, Ramachandran 2005) and Density Adjusted Data Thinning (DADT) algorithms, both developed at the University of Alabama in Huntsville (UAH). The algorithms are designed to retain information-dense regions of a data set while removing unnecessary data. Information-rich data targeted for retention are within regions of high spatial variance.

The goal of this work is to adapt these tools for operational meteorological applications. The IDT algorithm is applied to uniformly-gridded synthetic data and evaluated using traditional data assimilation approaches. Results from sensitivity tests are presented and optimal thinning strategies are discussed for some simple synthetic scenarios. The DADT is applied to real, non-uniformly-gridded satellite observations and examined for observation retention and analysis impact.

[1] Corresponding author address: Florida Institute of Technology, 150 W University Boulevard Melbourne, FL 32901. E-mail: msplitt@fit.edu.

2. DATA THINNING

The evaluation of the thinning algorithms is two-fold and involves 1.) observation retention issues and 2.) an indirect measure of the impact as manifest through analyses. Each of the analysis approaches used here updates a first-guess field with a correction or "analysis increment" based on the observations. This correction consists of a weighted combination of innovations, which are differences between the observation and the background field. Because the weights depend on the relative error of the observations and background field, evaluation is difficult in terms of real data because the error characteristics are unknown. The synthetic tests circumvent this problem as well as provide a truth field for direct evaluation. In terms of 1.) above, the variance-driven algorithms (IDT and DADT) might retain a significant amount of redundant information if applied in observation space. As a result, both methods are also tested in innovation space.

In contrast to the IDT algorithm, the sub-sampling (SS) approach exhibits no preferential observation selection but is straight-forward and computationally efficient and thus commonly used in operational meteorology. The SS approach often thins observations to a specified minimum distance separation and/or retains every n^{th} observation. In reality, some observation values may be more important as they provide more information to a data analysis system. In part, this motivates the development of techniques that can differentiate between regions of high information content and those that contain redundant data. The IDT algorithm was originally developed for applications to contiguous data (i.e., no missing values). Obviously this is problematic for real data which typically contain gaps. As a result, two distinct versions of IDT emerged

from this work including a global version, IDTG and a local version, IDTL. Neither IDT version is equipped to handle non-uniformly-gridded data sets, such as satellite observations, so a third DADT algorithm has been created for these applications. Each of these techniques is described in the following section.

2.1 IDTG: The Global IDT Algorithm

A snapshot of the observation (or innovation) values can be treated as an image with pixel intensities equal to the observation values at the corresponding grid points. The problem of finding regions of high information content thus translates to identifying 'anomalous' regions in the corresponding image. For a multimodal pixel distribution, pixels that form the tails of each mode are identified to be most deviant from the mean of all the pixels, contribute the most to the cumulative variance of the region, and are thus targeted for subsampling at a higher retention rate.

For each mode, the statistics of the pixels that are close to the mean are calculated. These sets of pixels are referred to as the background regions and are thinned for a low rate of data retention. All other image regions are considered to be heterogeneous and deemed to have high information content and thus are sub-sampled at a higher retention rate. The IDTG algorithm recursively decomposes the image into a tree structure with a root node that comprises the complete image. A region at level 'L' of the tree is identified as homogeneous if it passes *both* statistical similarity tests (T-Test and F-Test) by comparing region means and variances to the background values. If either of the tests fails, the region is decomposed into two sub regions defined as level 'L+1'. The splitting process continues, recursively dividing the target regions into smaller sub regions. The splitting process terminates when either a region is found to be homogeneous or deemed to be too small for additional splitting. In the latter case, it is considered to be heterogeneous and thus sampled at a higher data retention rate. The minimum sub region size is currently set to 3-by-3 pixels, however users can set a larger size. For regions larger than the minimum, an 'optimal' splitting point along the length (X) or width (Y) is determined. This optimal splitting point is defined as the location for which the combined cumulative variance of the two sub regions is a minimum. This approach is similar to the least-squares approximation described by Wu (1993). The outcome of the recursive decomposition process is a mixture of both homogeneous and heterogeneous regions that are sub sampled at different retention rates accordingly.

2.2. IDTL: The Local IDT Algorithm

One challenge for the IDTG algorithm described above is calculating the background statistics for a multimodal image as it is often difficult to determine the number of modes in a probability density function. As a result, we have developed a variant of the IDTG algorithm which follows a similar recursive process that decomposes an image into homogeneous sub regions. However, instead of comparing the statistical similarity between a sub region and the backgrounds to determine whether or not to split the region, a more local approach (referred to here as IDTL) is taken whereby two sub regions at tree level 'L+1', are compared to determine whether or not the region at level 'L' should be split. Splitting occurs for level 'L' only if the two sub-regions are determined to be statistically different based on the T-test and F-test. As previously mentioned, the desired outcome of the decomposition process is to partition an image into as many homogeneous sub regions as possible (thereby removing redundant data).

2.3 DADT: A Density Adjusted Thinning Algorithm

Because satellite data are not on a uniformly-spaced grid (image), an intelligent data thinning algorithm was developed for irregularly-gridded data sets. Although both the uniform-grid based and non-uniform-grid based algorithms are information retention oriented, non-uniform-grid based algorithms cannot apply the image decomposition approach employed by the former. The non-uniformly-gridded thinning algorithm is referred to here as Density Adjusted Data Thinning (DADT). For this approach, the information content that an observation contains is measured by the intensity variance of neighboring observations. Each observation is evaluated and ranked based on its information content and placed in a descending order priority queue. Observations that are at the top of the queue and outside the scope of observations in the thinned data set, are retained. The scope of an observation is defined by a circular area of radius R. Adjusting the parameter R determines the spatial pattern and local density of the thinned data set with a larger R producing a more uniform field of thinned observations. Like IDT, the DADT is an iterative algorithm creating a thinned data set by successively adding optimal available sample points from the original data set until a desired number of points are retained. The value of R is iteratively decreased by a tunable value (ΔR). A smaller value of ΔR produces a more uniformly thinned set of observations.

3. EXPERIMENTS

Figure 1 depicts a flow diagram of the experiment set-up. As previously reported in Ramachandran et al. (2007), the IDTG showed improvement over both a box variance (BV) and subsample (SS) thinning approach when applied to two-dimensional synthetic data. The

same data are used in the next section to examine two key parameters and their impact on data retention. A condition of optimality (i.e., minimum analysis error) is then discussed for a 1D analytic function. The DADT is then applied to non-uniformly-gridded satellite data in two pseudo-3D applications.

3.1 Synthetic Data: IDTG

IDTG consists of several key parameters which affect the thinning rate of the observations. The sensitivity of IDTG to two of these parameters—the confidence interval (CI) and the background—is examined here. Figure 2 depicts this parameter space for both the observations and innovations with the contours representing the number of retained observations. The background (y-axis) is defined as the percentage of pixels used to create the global mean and variance to be used in the statistical similarity tests. The thinned data are the same as those used in Ramachandran et al. (2007) in which the truth, observations, and first-guess field along with their errors are known explicitly. The observation errors are spatially uncorrelated and the observation-to-background error variance is 0.25. For a given background pixel percentage, as the CI increases the number of observations retained decreases–a direct result of the increased width of the interval. In other words, when the observations pass the similarity test, the region is identified as homogeneous, and the splitting process is terminated. However, the gradient is not linear with most of the observation reduction at the high end of the CI scale (CI > 0.98). This is especially true when fewer pixels are used to compute the background statistics. When the number of pixels used to estimate the background statistics are increased, for a fixed CI, the mean and variance approach that of their true global values and, as a consequence, the likelihood that a sub region will fail the similarity tests decreases. This sensitivity virtually disappears for high CIs because most of the observations have already been thinned.

The innovation CI/background parameter space is similar to that derived for the observation thinning. However, the innovation isopleths are, for the most part, more steeply sloped. More importantly, for a given CI/background parameter combination, there are fewer observations retained in innovation space. This result is expected given that both the observations and background are of relatively good quality. For the case of a degraded first-guess field, the rate of thinning is decreased in innovation space (not shown). Analyses are currently underway to identify the parameters that minimize both the number of observations retained and analysis error.

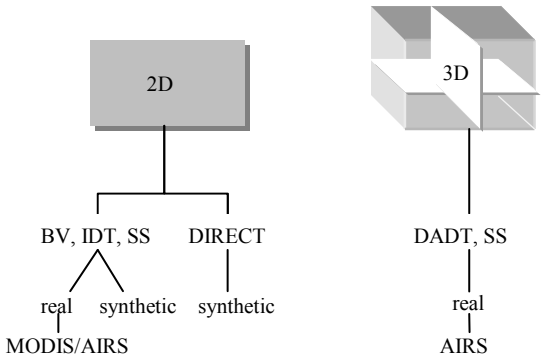

Fig. 1. Experiment configuration. Thinning algorithms include: Box Variance (BV), Intelligent Data Thinning (IDT), and sub-sampling (SS). See text for details.

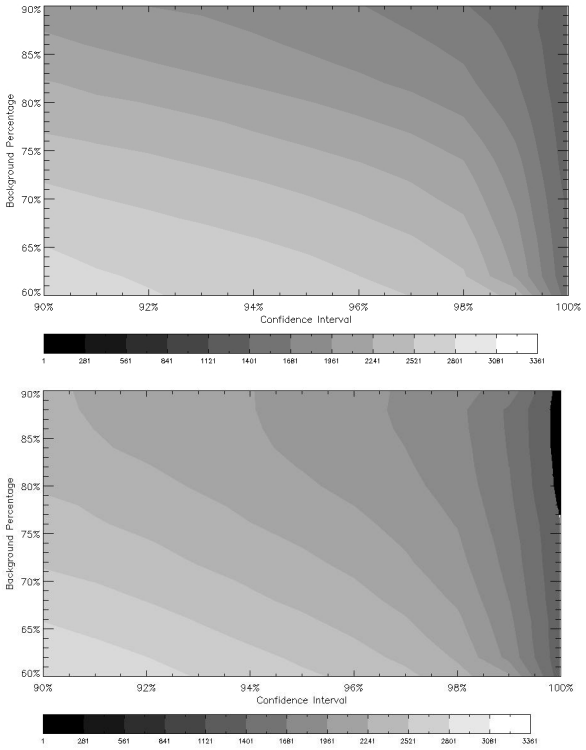

Fig. 2. Number of observations retained (shading) as a function of Confidence Interval (%) and pixels used to estimate the background statistics (%) for observation thinning (top) and innovation thinning (bottom).

3.2 Synthetic Data: The Direct Method

Despite comparisons with less sophisticated approaches, the optimal observation configuration is not clear. Optimal configuration is defined here as the observation configuration that minimizes the analysis error for a given number of observations. Although this error metric may produce considerably different results than those obtained via four-dimensional data assimilation, it is instructive to determine the observation distribution produced by the analysis with

the lowest root mean square error (rmse). An idealized truncated 1D Gaussian (Fig. 3) with 35 points (observations) is assumed to be the truth field and is sampled for the unique combination of five points that yields the best analysis. We refer to this thinning method as DIRECT (see Fig. 1). To expedite the analyses (there are approximately 325K possible unique combinations of five observations), a simple successive correction algorithm (Barnes, 1964) is applied here. There is no first-guess field and the observations are assumed to have no error. Figure 3 shows results for three different length scales (2Δx, 4Δx and 20Δx, where Δx is the analysis grid spacing). The placement of the observations depends implicitly on the analysis length scale—with the optimal observation locations encroaching on the gradient regions as the scale decreases. These results indicate that the thinning algorithm should depend on not only observation density and grid resolution but should also be coupled to the implicit scales of the analysis. It is not clear yet whether this approach will help guide improvements in the IDT and DADT algorithms; however, it does provide information regarding how closely the optimal retention agrees with the IDT selection.

3.3 Real Data: AIRS Temperature Assimilation

Temperature and water vapor profiles are derived from the radiances measured by the Atmospheric Infrared Sounder (AIRS) instrument and the Advanced Microwave Sounding Unit (AMSU) on the Aqua EOS platform. Temperature soundings obtained from Version 5 of the AIRS retrieval algorithm are used in this case study. Each sounding contains approximately 54 vertical levels between 1013.25 and 100 hPa. Globally, the AIRS Version 4 retrieved profiles—compared to rawinsondes collocated in time and space—exhibit RMS errors of 1 K in 1-km layers for temperature and 10-15% RH in 2-km layers for moisture (Tobin et al. 2006, Divakarla et al. 2006). Although Version 5 profiles have not yet been validated, it is expected that the relative validation errors will be similar to (or better than) those presented for Version 4 (Susskind, personal communication). Each AIRS profile contains a specific pressure level below which data is of decreased quality. For this study, only the temperature and moisture data above this maximum pressure level are used in the analyses. Levels below the maximum pressure level were designated with a missing data value and not considered in the DADT thinning algorithm. A plot of the 497-hPa level of AIRS temperature for 12 March 2005 is shown in Figure 4a. In the figure, gaps in the data represent areas where the AIRS data have been removed by the AIRS quality indicators.

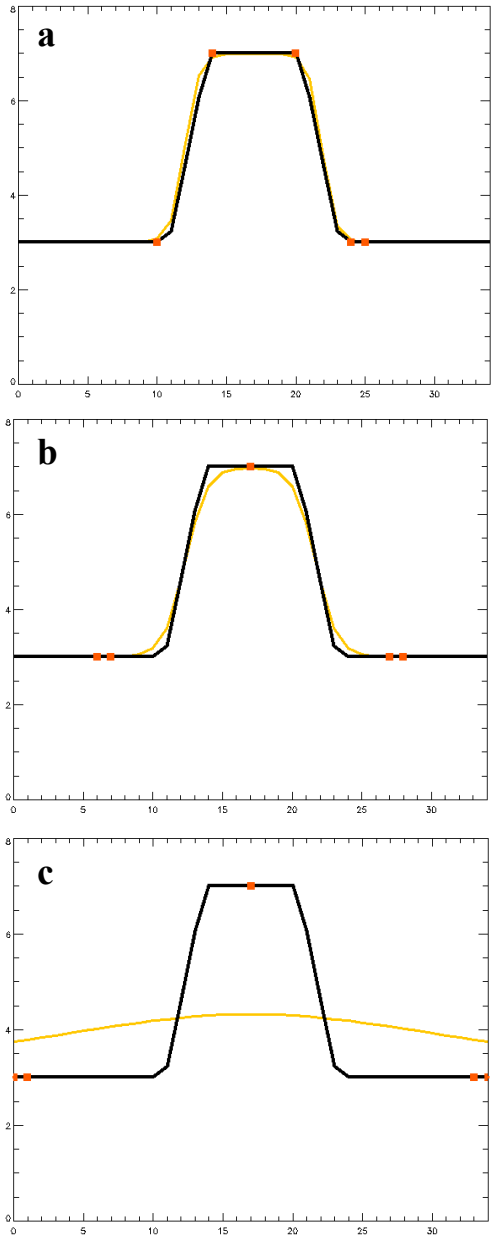

Fig. 3. Truncated Gaussian (black curve), Barnes analysis (orange curve) and optimal observation locations (red circles) for analysis length scales of a.) 2Δx, b.) 4Δx, and c.) 20Δx. See text for details.

For the 12 March 2005 case study examined herein, the background field for the analysis is an 8-hour forecast from the Weather Research and Forecast (WRF) model (Skamarock et al. 2005) initialized at 00 UTC by the 40-km resolution North American Model (NAM). A short forecast is used as the background instead of an analysis because of the asynoptic time of the AIRS overpass (0742 UTC). The WRF output is mapped to the ARPS Data Analysis System (ADAS; Brewster 1996) grid. The ADAS domain is nearly identical to that of the WRF with the exception of small

differences in the pressure levels. The error covariances used for the background are standard short-term forecast errors cited in the ADAS documentation, and the error tables used for the AIRS profiles are based on estimates cited in validation experiments by Tobin et al. (2006). Separate error estimates are used for land and water soundings. Three iterations of the ADAS Bratseth scheme are performed with horizontal scaling factors of 150, 120, and 100 km, respectively. Based on the vertical resolution of the data and the layer-averages that each AIRS level represents, the vertical length scale is set to 750 m for the first two iterations and then reduced to 400 m for the final iteration.

3.3.1 Psuedo-3D thinning: Pressure Levels

The DADT algorithm was applied to each pressure level of the AIRS observations using the methodology described in Section 2.3 with R and ΔR values of 10 and 2 degrees respectively. This approach is essentially a 2D application of the DADT because information between vertical levels is not shared. However, it does produce a 3D field of thinned observations that can be applied to test both computation time and analysis fidelity. Figure 4b shows the observations that are retained by the DADT algorithm at 497 hPa. Retention is highest in the gradient regions over Illinois and Indiana and over Florida and the northern Gulf of Mexico). Elsewhere, the retained data are more evenly-spaced.

Both the full and DADT-thinned data were assimilated into ADAS as rawinsonde data with the errors and scale factors previously described. The full data set contains 2197 54-level profiles (118,638 unique pieces of information) and takes 6885 seconds to complete an analysis. The thinned data set, which retains 405 observations per level for a total of 21,870 unique pieces of information, has a run time of only 846 seconds—an 88% decrease in run time! A profile of the average analysis increment at each analysis level (Fig. 5) indicates that, for the most part, the analysis fidelity is preserved for the thinned data. A systematic variation of the R and ΔR parameters, in tandem with the analyses, should provide insight regarding the algorithm sensitivity. Additionally, a proper coupling of the thinning algorithm to the analysis length scale, as discussed previously, would also likely improve the results shown here.

3.3.2 Pseudo-3D thinning: Pressure and Vertical Levels

An alternative approach for obtaining a pseudo-3D field of AIRS observations is to apply IDTL, independently, to both pressure and vertical cross sections. Examples are presented here for which a temperature innovation field is successively thinned in

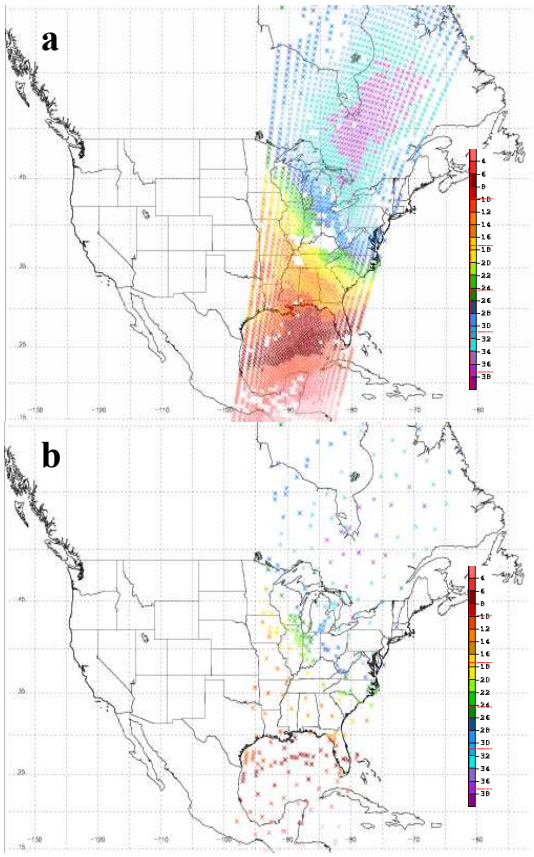

Fig. 4. AIRS temperature data for the 497 hPa level valid at 0742 UTC 12 March 2005 for the a) full data set and b) thinned data.

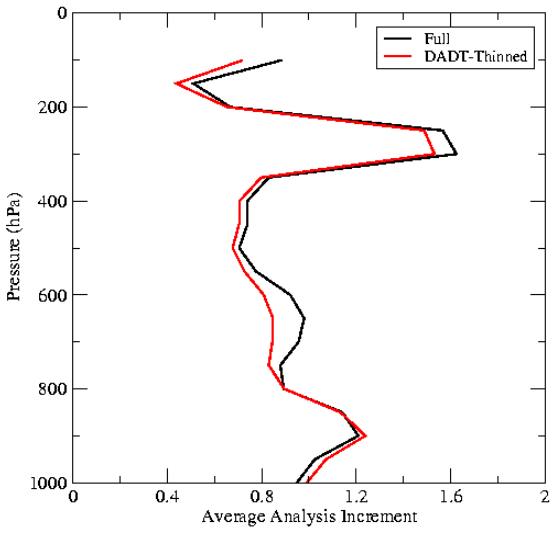

Fig. 5. Average analysis increment of full (black) and thin (red) analyses for the 12 March 2005 case study.

two dimensions by slicing the AIRS data vertically in the along-swath and cross-swath directions as well as at AIRS pressure levels. Innovations are produced by interpolating the NCEP Global Reanalysis temperatures to AIRS locations for the 12 March 2005 case study (Fig. 6). Innovations are used for this experiment because of the large temperature variability of vertical atmospheric profiles. For this experiment, IDTL was run on 30 vertical cross-swath sections, 135 along-swath vertical cross sections, and 54 pressure cross-sections (a total of 219 independent 2D IDTL applications). An along-swath vertical cross-section, delineated by the dashed black line in Fig. 6, is shown in Fig. 7. Superimposed on the figure are the full set of innovations and the thinned innovations obtained by applying IDTL in: 1.) a single along-swath vertical cross-section (red boxes) and 2.) in 54 pressure

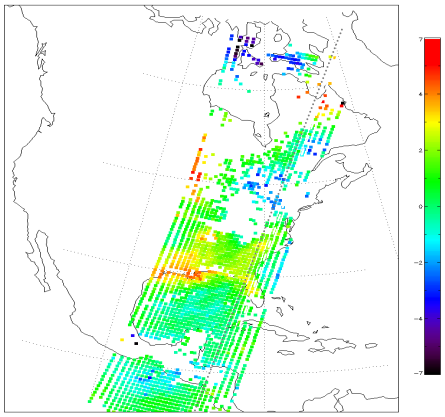

Fig. 6. 852 hPa temperature innovations (^{o}C, AIRS minus NCEP Global Reanalysis) valid 0700 UTC 12 March 2005. Black dashed line denotes cross-section shown in Fig. 7.

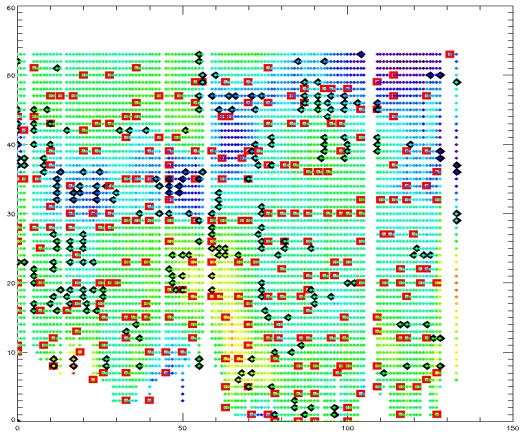

Fig. 7. Vertical along-swath cross-section indicated by the dashed black line in Fig. 6. Red boxes denote observations retained from 2D thinning in the along-swath cross section and the black diamonds represent the thinned data from a collection of 54 pressure cross sections. See text for details.

level cross-sections (black diamonds). Clearly, there is little overlap in the retained observations indicating that the IDTL depends on how the algorithm is applied in two-dimensions. There is however, coherence between adjacent vertical cross-sections (not shown). At this time it is not clear how best to apply three dimensional thinning. One possible approach is to combine the uniquely retained observations in the various swath directions. This will be tested and compared against the thinning by pressure level described in the previous section for the DADT. However, the results of this test suggest the need for a true 3D version of IDT and DADT.

4. DISCUSSION/FUTURE WORK

For the synthetic data presented, the thinning efficiency in innovation space is tied to the quality of the background field and observations. When the first-guess field is a good approximation of the truth and the observation error is small, the IDT (IDTG) removes a greater number of observations in innovation space due to redundancy. One-dimension simulations using on the order of 250K analyses of a truncated Gaussian function indicate that, for a given number of observations, the optimal observation distribution depends on the analysis length scale.

A test of thinned AIRS L2 thermodynamic profiles using a thinning algorithm for non-uniformly-gridded data produced an analysis that required significantly less computation time yet maintained analysis quality. Successive applications of the IDT in two dimensions to AIRS temperature profiles suggest that a three-dimensional version of the IDT is likely necessary – at the very least to ensure consistency in the selection of observations.

Using both artificial and real data, we continue to test, tune and develop both versions for operational applications. Albeit not shown here, the IDTL is also being tested on sea surface temperature (SST) from the Moderate-resolution Imager Spectroradiometer (MODIS).

5. REFERENCES

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.*, **3**, 396-409.

Bergman, K. H. and W. D. Bonner, 1976: Analysis Error as a Function of Observation Density for Satellite Temperature Soundings with Spatially Correlated Errors. *Mon. Wea. Rev.*, **104**, 1308-1316.

Bratseth A. M., 1986: Statistical interpolation by means of successive corrections. *Tellus*, **38A**, 439-447.

Brewster, K., 1996: Implementation of a Bratseth analysis scheme including Doppler radar data. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Amer. Meteor. Soc., Norfolk, VA, 92-95.

Divakarla, M. G., C. D. Barnet, M. D. Goldberg, L. M. McMillin, E. Maddy, W. Wolf, L. Zhou, and X. Liu, 2006:Validation of Atmospheric Infrared Sounder temperature and water vapor retrievals with matched radiosonde measurements and forecasts. *J. Geophys. Res*., **111**, D09S15, 20 pp.

Liu, Z.-Q. and F. Rabier, 2002: The interaction between model resolution, observation density in data assimilation: A one-dimensional study. *Q. J. R. Meteorol. Soc.*, **128**, 1367-1386.

Ochotta, T., C. Gebhardt, D. Saupe, and W. Wergen, 2005: Adaptive thinning of atmospheric observations in data assimilation with vector quantization and filtering methods. *Q. J. R. Meteorol. Soc.*, **131**, 3427-3437.

Ramachandran R., X. Li, S. Movva, S. Graves, S. Greco, D. Emmitt, J. Terry, and R. Atlas, 2005: Intelligent Data Thinning Algorithm for Earth System Numerical Model Research and Application. Preprints, *21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Amer. Met. Soc., San Diego, CA.

R. Ramachandran, X. Li, S. Movva, S. Graves, M. Splitt, S. Lazarus, M. Luekenn, B. Zavodsky, and W. Lapenta, 2007: An Improved Data Reduction Tool in Support of the Real-Time Assimilation of NASA Satellite Data Streams. *11th Symposium on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface*, San Antonio TX, Amer. Met. Soc., January 13-18, 2007.

Tobin, D. C., 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, 2006: ARM site atmospheric state best estimates for AIRS temperature and water vapor retrieval validation. *J. Geophys. Res.*, **111**, D09S14, 18 pp.

Wu, X., 1993: Adaptive split-and-merge segmentation based on piecewise least-square approximation. *IEEE Trans.*, **15**, 808-815.