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

Rahul Ramachandran[1], Xiang Li, Sunil Movva, Sara Graves
Information Technology and Systems Center, UAH, Huntsville, AL

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

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

William Lapenta
MSFC/NASA, Huntsville, AL

I. INTRODUCTION

Today's research and operational forecast models and data assimilation systems have difficulty ingesting and utilizing large volumes of satellite data, in part due to prohibitively large computational costs, time constraints and bandwidth issues. To address this problem, NASA recently funded a project aimed at refining, testing and customizing an existing automated Intelligent Data Thinning (IDT) algorithm, developed at the University of Alabama in Huntsville (UAH), in conjunction with commonly used data assimilation systems for numerical weather prediction models. The most significant measure of a successful data reduction algorithm is its ability to retain valuable information – that which has maximum impact on the model forecast – while simultaneously reducing the data volume. The IDT algorithm is specifically designed to retain information-dense regions of a data set while removing redundant data. This recursive simplification algorithm, is based on the computer graphics concept of data decimation, retains data within regions of high spatial frequency (large variances), while subsampling regions of low spatial frequency (low variances) to thin the data.

The goal of this project is to test, refine and customize the existing IDT algorithm in order to transition it into a deliverable data reduction tool useful for real-time applications with a wide variety of dense NASA satellite data streams in operational, research, and private industry communities. Here, we present results from sensitivity analysis on IDT with selected synthetic data sets and various assimilation methodologies.

II. DATA THINNING STRATEGIES

Two non-trivial approaches to data thinning are evaluated: a) the box variance (BV) method, and b) the intelligent data thinning (IDT) method. Additionally, a simple sub-sampling method is evaluated whereby the number of retained observations is systematically varied to one third, one sixth and one ninth of the full set of observations. Evaluation was also conducted with randomly selected (spatially) sets of observations to match similar observation numbers used in the other techniques. The evaluation is two-fold and involves 1.) observation retention issues for the two thinning methodologies 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 an "increment" or correction. This correction consists of a weighted combination of innovations (an innovation is the difference between the observation and the background field). The non-trivial data thinning methods applied here both rely on variability as a means by which to selectively filter the input. In the absence of observation or background error it is possible that either of the IDT or BV approaches might still retain a significant amount of redundant information if applied in observation space. As a result for comparison purposes, the BV and IDT methods are also tested in innovation space.

There is no preferential selection of particular observations for the sub-sampling methodology, but it is straight-forward and computationally efficient and is thus commonly used in operational meteorology. 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 non-trivial thinning techniques are briefly described in the following sections.

[1] Corresponding author address: Information Technology and Systems Center, University of Alabama in Huntsville, Huntsville, AL 35899. E-mail: rramachandran@itsc.uah.edu

A. *Box Variance (BV) Method*

The BV method [1] divides the analysis domain into boxes with a prescribed 10×10 grid-space length. Each box is marked as containing high information content if the variance of the observations (or innovations) is higher than a predetermined, user-defined threshold specific to that particular data set. If the variance is less than the threshold, the observation (or innovation) whose value is closest to the mean value of all the data within the box is retained. However, if the variance is greater than the threshold, no thinning occurs within the box (i.e., all observations/innovations are retained). The observation (innovation) variance threshold is not fixed in the experiments presented here, but rather is selected so as to closely match the number of observations retained in the IDT and sub-sampling experiments.

B. *Intelligent Data Thinning (IDT)*

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 'abnormal' regions in the corresponding image. For a multimodal pixel distribution, pixels that form the tails of each mode are most deviant from the mean of all the pixels, contribute the most to the cumulative variance of the region, and are thus identified for subsampling at a higher retention rate.

For each mode, we compute the statistics of the pixels that are close to the mean. These sets of pixels are called the background regions and are thinned for a low rate of data retention. All other regions in the image are deemed to have high information content and are sub sampled at a higher retention rate. The IDT algorithm [2] recursively decomposes the image into a tree structure. The root node of the tree is the complete image. Each region in level 'L' of the tree is decomposed into two regions of level 'L+1' if it fails the statistical similarity tests that compare the region with the background, thus, recursively splitting the target regions into smaller sub regions while leaving the background regions intact.

The statistics involved include the mean and variance which are computed for the sub-region. Two statistical similarity tests (F-Test and T-Test) are performed using the computed statistics to check if the region is similar to one of the backgrounds. The F-Test provides a similarity measure between the variances, and the T-Test provides a similarity measure between the means. If the region is similar in terms of mean as well as variance, we sub-sample the region to retain less data. Otherwise, the region is tested for a sub-region of interest (i.e. high information content) in order to split the region. If the region is large enough, an optimal splitting point along the length (X) or height (Y) is found, and the region is decomposed into sub regions at this point leaving two uniform and differing regions. This optimal splitting point is selected at a position that reduces the cumulative variance within each region—if they are represented by their means—in an approach similar to the least-square approximation described by Wu [3]. If the region is too small, it cannot be split, so it is sub sampled at a higher retention rate.

III. Experiment Design

A. *Truth, Background and Observations*

The truth and, background are specified on a 175 x 175 grid, and observations are randomly located within this grid. The truth field (Fig. 1) was chosen to mimic a temperature field associated with a heated peninsula or warm ocean current whereby two regions of strong gradient separate regions of relatively little temperature gradient.

The synthetic background field was generated such that the spatial error correlation statistics were known and consistent with theoretical statistical assumptions of optimal interpolation. Albeit useful, the degree to which our synthetic data experiments mimic the real world will depend, in part, on the error covariance statistics which are not typically well known. Additionally, the experiments performed here are highly idealized where we have assumed homogenous, isotropic, and uncorrelated observation error.

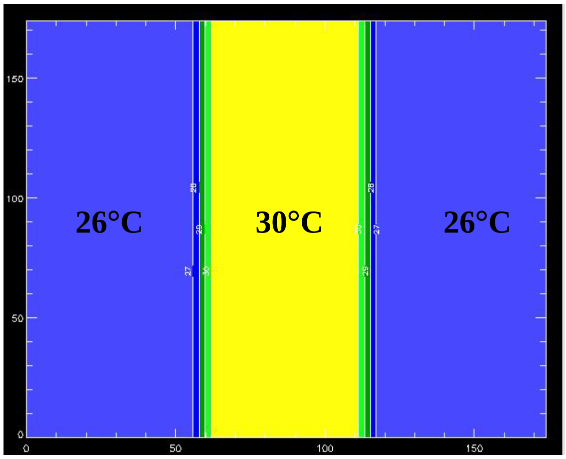

Fig. 1. Truth field. The pseudo temperature distribution is intended to replicate a peninsula or ocean current with strong peripheral gradients. Temperature contours are 1°C.

The background field was generated following the work of [4]. A pseudo-random two-dimensional field of perturbations from the truth was prescribed using a variance of 1 and decorrelation length of 25 grid points (Fig. 2). The perturbation field created with this method has no knowledge of the high temperature gradient regions. These perturbations were then added to a smoothed truth field to create the background. As a result, this approach produces a background field that contains the same error decorrelation in all regions of the grid. Although the smoothing of the truth violates the assumption of isotropic error, this adjustment does not significantly impact the resulting variance and decorrelation statistics for the full domain.

Observations were generated systematically within the analysis domain with an observation separation distance 3 times the length of the analysis grid spacing. Spatially uncorrelated error (white noise) was introduced into the observations with a variance of 0.25. Various thinning strategies (a total of 6) were applied to these sets of observations. Fig. 3 depicts the full set of observations along with two of the thinned data sets. The BV algorithm was applied to both the observations and innovations while the IDT was applied to the observations only. Innovations were generated by interpolating the background field to the observation locations. Both the BV and IDT algorithms respond to variability such as that in the gradient regions or that introduced by error. Overall, the number of data points retained in the gradient regions is substantially higher.

B. Analysis Schemes

Bratseth

The Bratseth analysis is a successive corrections scheme that converges to optimal interpolation (OI) with sufficient iterations [5]. The iterative approach is both computationally feasible and economical compared to the traditional OI and variational approaches, [6] and [7]. Similar to the OI approach, the Bratseth method requires an estimation of the background and observation error covariances. It is assumed that the Bratseth method converges when the average difference between successive iterations is less than 10^{-4}. While this does not necessarily guarantee convergence to OI, it is reasonable to assume that subsequent iterations will not significantly improve the analysis. Using the full data, 87 iterations are required for convergence.

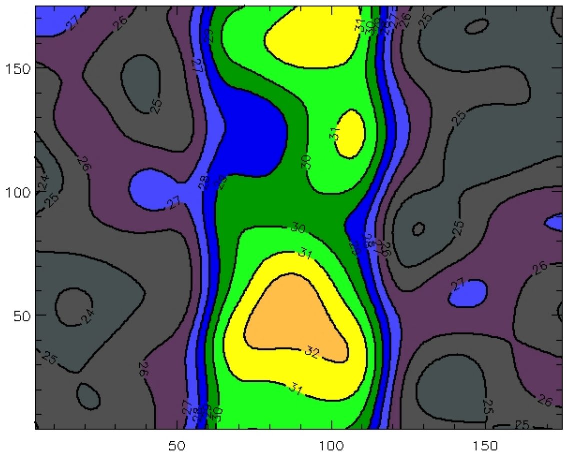

Fig. 2. Background field with observational error variance of 1 deg^2 and a decorrelation length of 25 grid points.

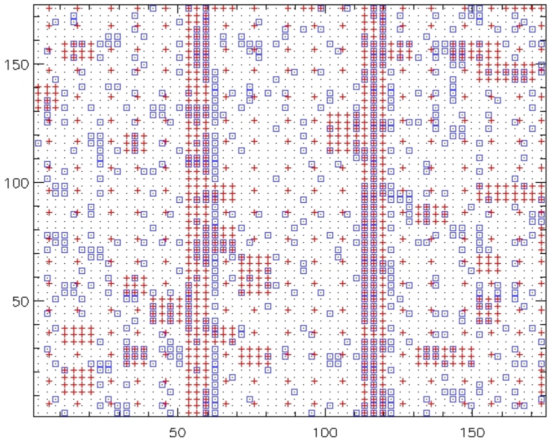

Fig. 3. Observation locations for the full set of observations (gray dots), the Box Variance (BV) method conducted on observations (red X's) and the IDT algorithm using observations (blue boxes).

The background and observation error covariance are set to 1.0 and 0.25 respectively, and a third parameter, the error decorrelation length scale, is set to 25. Although unrealistic (i.e., these parameters, in general, are unknown), the values chosen here are in accordance with the prescribed errors.

Kriging

Kriging is an interpolation scheme that originates from geostatistics and, like the Bratseth, is comparable to OI [8]. The implementation of Kriging used here is termed Ordinary Kriging. We chose an exponential covariance function (variogram) for modeling the error covariance that is consistent with the Bratseth method. The Kriging analysis was conducted in two modes: 1) without use of a background field (observations only) and 2) with use of background field. The Kriging using the mode 1 approach was chosen to provide a benchmark by which to directly assess the impact of the background field on the analyses.

Variational

A two-dimensional variational (2DVAR) analysis approach, that uses localization to reduce the size of the background error covariance matrix, is applied [9]. The background error covariance is modeled using a recursive filter [10]. The variational problem is given by the functional J,

$$J = \frac{1}{2}\left(x - x_b\right)^T B^{-1}\left(x - x_b\right) + \frac{1}{2}\left[H(x) - y_o\right]^T R^{-1}\left[H(x) - y_o\right], \qquad (1)$$

where x is the analysis variable, x_b is the background, B and R are the background and observation error covariance matrices respectively, *H(x)* maps the background field to the observation space, and y_o is the observation vector. Upon preconditioning, Eq. (1) can be rewritten as

$$J = \frac{1}{2} q^T q + \frac{1}{2}\left[HCq - d\right]^T R^{-1}\left[HCq - d\right], \qquad (2)$$

where $J_{inc}(x)$ is the incremental change in the analysis variable, q is the preconditioned analysis variable, C is the preconditioner, and d is the innovation vector. Eq. (2) is differentiated with respect to the control (analysis) variables and then minimized through a limited memory conjugate gradient algorithm [9]. The error covariance for both the background and observations are identical to the Bratseth and Kriging techniques.

IV. Results

A. Synthetic Data

Domain Root Mean Square Error (rmse) for the 4 analysis systems using the full observational data set and the 6 filtered data sets are presented in Table 1. As expected, Kriging without use of the background field produced the highest analysis error in all cases. The 2DVAR method produced the lowest rmse in all cases and, neglecting the full observation analyses, the lowest error (for all methods) is associated with the sub-sample 3 thinning. The IDT error is comparable (within 0.002). Because we are interested in resolving the gradients, the rmse is also shown for the gradient region only (Table 2). Interestingly, 2DVAR performs consistently for all thinning methodologies while the Kriging and Bratseth analyses are degraded for the sub_6 and sub_9 experiments. Overall, BV_obs (see table caption) performs the best (i.e., lowest rmse) for all techniques except 2DVAR, which yields a slightly larger error than that of the BV_ino. It is worth pointing out that, for these experiments, the BV_obs approach retains the greatest number of observations within the gradient region (see Fig. 3). The difference between the observation and innovation thinning is more evident within the gradient regions for the BV methods, while in the gradient regions the rmse is higher for all analysis schemes (except 2DVAR) for the BV_ino experiment.

Observation Retention Issues

Although the error statistics in the two tables are somewhat informative, direct comparison is difficult as the error does not take into account the number of observations. A log/log plot of the rmse versus number of observations is shown in Figures 4 and 5 for the full domain and gradient regions respectively [11]. These diagrams display information in a way that helps assess the thinning/analysis quality as a function of computational expense. As illustrated, the relationship is a nonlinear function of the number of observations. The desired results (low rmse/low number of observations) tend towards the lower left corner of the plot.

Table I
RMSE (full domain) for the four analysis schemes using full data (Full) and 6 different thinning strategies including sub-sample 3, 6, and 9 (Sub_3, Sub_6, Sub_9), box variance with observations and innovations (BV_obs, BV_ino), and IDT with observations (IDT_obs). Number of observation retained in parentheses.

Method (#obs)	**Kriging NB**	**Kriging**	**Bratseth**	**2DVAR**
Full (3364)	0.1042	0.0507	0.0583	0.0481
Sub_3 (400)	0.1681	0.0890	0.0907	0.0636
Sub_6 (100)	0.3790	0.2203	0.2196	0.0808
Sub_9 (49)	0.7635	0.3972	0.3951	0.1015
BV_obs (732)	0.2008	0.1315	0.1348	0.0712
BV_ino (788)	0.1833	0.1014	0.1042	0.0667
IDT_obs (721)	0.1698	0.0938	0.0946	0.0658

Table II
Same as in Table 1 except for gradient region only.

Method (#obs)	**Kriging NB**	**Kriging**	**Bratseth**	**2DVAR**
Full (3364)	0.4941	0.2591	0.2909	0.2891
Sub_3 (400)	0.6791	0.3181	0.3238	0.3325
Sub_6 (100)	1.1469	0.5382	0.5376	0.3451
Sub_9 (49)	2.1382	0.8018	0.8083	0.3625
BV_obs (732)	0.5258	0.2283	0.2654	0.3167
BV_ino (788)	0.6786	0.3250	0.3479	0.3014
IDT_obs (721)	0.5490	0.2469	0.2722	0.3048

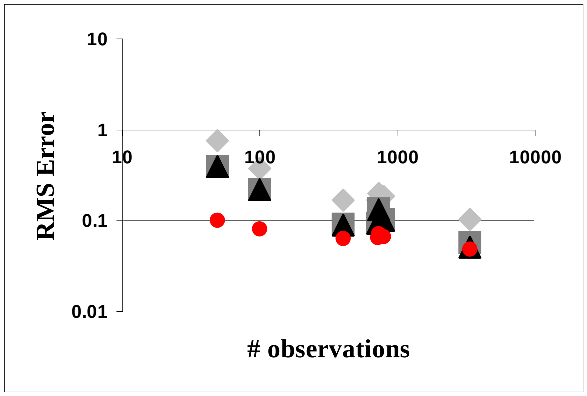

Fig. 4. RMSE for the full domain for Kriging without a background field (gray diamonds), Kriging with a background field (black triangles), Bratseth (gray squares), and 2DVAR (red dots).

As discussed previously, for the full domain, the 2DVAR analyses outperform the other techniques, especially when the number of observations are significantly reduced (Fig. 4). As the number of observations retained increases, the discrepancy between the various analysis techniques disappears. The Kriging without background rmse is actually greater than the background rmse for some of the thinned experiments (gray diamonds). Interestingly, the quality of the 2DVAR approach in the gradient region appears to be relatively insensitive to variation in the number of observations (Fig. 5). These findings suggest that the analysis error is both a function of the observational filtering and the analysis system.

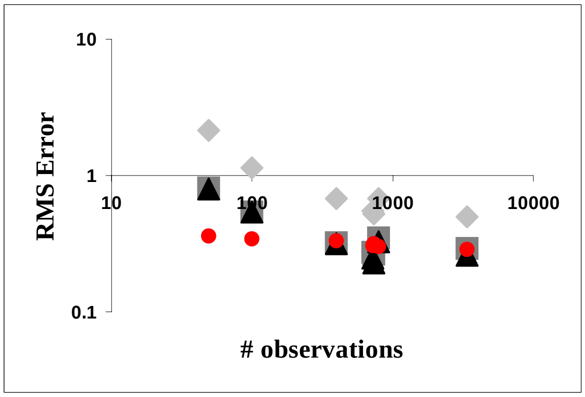

Fig. 5. Same as Figure 4, but for the gradient region only.

V. Discussion/Future Work

Synthetic Data Continued

Some additional evaluation of the data filtering algorithms using synthetic data remains. It is not clear, for example, to what degree the thinning algorithms performance is tied to the quality of the background field and observations. Table III depicts an observation/background quality matrix. Thus far, the observation thinning appears to be the best approach (with the exception of 2DVAR, Table II). However, the combination of quality observations and a degraded first-guess field (the green shaded box, Table III) is biased in favor of observation-based thinning. The bias is an artifact of smoothing the truth to create the background which increases the analysis error in the gradient region (i.e., the analysis draws to spurious innovations). As previously mentioned, the impact of the smoothing on the true error statistics is minimal when averaged over the entire domain, which indicates that the background error characteristics are clearly not homogeneous as assumed here. Of particular interest is the opposite configuration, i.e., poor observations and a quality background field – a combination that should favor innovation space thinning. Additionally, when both the observations and first-guess are of decent quality, we anticipate that the innovation space filtering will be the best approach as the innovation variability will be significantly less than that of the observations alone.

Additional synthetic experiments in which the gradient in the background field is displaced rather than smoothed (a realistic scenario) and experiments where the observation error is correlated will also be performed.

Table III
Experiment matrix for varying combinations of quality in the observations (OBS) and background field (BG).

	Good OBS	**Bad** OBS
quality BG	**?**	**?**
degraded BG	**obs better**	**?**

Real Data Experiments

We have begun to apply the thinning algorithms to sea surface temperature (SST) from the Moderate-resolution Imager Spectroradiometer (MODIS) direct broadcast. A case with minimal cloud cover has been chosen for evaluation (Fig. 6). This work will be expanded to include temperature and water vapor profiles derived from the Atmospheric Infrared Sounder (AIRS) instrument aboard the Aqua EOS platform.

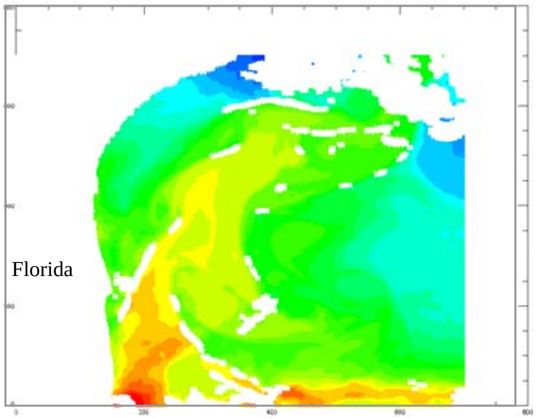

Fig. 6. MODIS-AQUA sea surface temperatures valid 1825 UTC 20 May 2006.

REFERENCES

[1] B. Zavodsky, S. Lazarus, R. Ramachandran, and X. Li, “Evaluation of an Innovation Variance Methodology for Real-Time Data Reduction of Satellite Data Streams,” Preprints, *18th Conference on Probability and Statistics in the Atmospheric Sciences*, Amer. Met. Soc., Atlanta, GA, January 2006.

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

[3] Wu, X., “Adaptive split-and-merge segmentation based on piecewise least-square approximation,” *IEEE Trans.*, vol. 15, pp. 808-815, 1993.

[4] G. Evensen, “Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics,” *J. Geophys. Res.*, vol. 99, pp. 10143-10162, 1994.

[5] A. M. Bratseth, “Statistical interpolation by means of successive corrections. *Tellus*, vol. 38A, pp. 439-447, 1986.

[6] S. M. Lazarus, C.M. Ciliberti, J.D. Horel, and K. Brewster, “Near-real-time applications of a mesoscale analysis system to complex terrain,” *Wea. Forecasting*, vol. 17, pp. 971-1000, 2002.

[7] X. Deng and R. Stull, “A Mesoscale analysis method for surface potential temperature in mountainous and coastal terrain,” *Mon. Wea. Rev*.,vol. 133, pp. 389-408, 2005.

[8] N. Cressie, “The origins of Kriging,” *Mathematical Geology*, vol. 22, pp. 230-252, 1990.

[9] J. Gao, M. Xue, K. Brewster, and K. K. Droegemeier, “A three-dimensional variational data analysis method with recursive filter for doppler radars,” *J. Atmos. Oceanic. Technol.*, vol. 21, pp. 457-469, March 2004.

[10] C.M. Hayden and R.J. Purser, “Recursive filter objective analysis of meteorological fields: applications to NESDIS operational processing,” *J. Appl. Meteor*, vol. 34, pp. 3-15, January 1995.

[11] J.L. Anderson, B. Wyman, S. Zhang, and T. Hoar, “Assimilation of surface pressure observations using an ensemble filter in and idealized global atmospheric prediction system,” *J. Atmos. Sci.*, vol. 62, pp. 2925-2938, 2005.