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

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

Rahul Ramachandran[1], 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

Reduction of information in data sets for meteorological data assimilation systems is motivated by the volume of data provided in remote sensing platforms such as satellite and radar systems. While remote sensing systems provide a significant source of real-time data over data-sparse regions, the data are high-volume and may be redundant which can lead to an unnecessary increase in computational burden. Significant reduction in repetitive data can increase analysis quality due to the improvement in numerical convergence resulting from an increased number of iterations (Purser et al., 2000). Reduction of observations can take one of the following three general strategies: a) data thinning, b) creation of super-observations, or c) a combination of the two methods. The impact of data thinning reduction methods on simple analysis systems are investigated here to assess the benefit of the data thinning on analysis quality and cost.

2. 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, several trivial sub-sampling methods are evaluated including: evenly spaced subsampling equating to one third, one sixth and one ninth of the full set of observations. Additional evaluation was conducted with randomly selected (spatially) sets of observations to match similar observations numbers used in the other techniques. The BV and IDT methods were tested additionally on thinning of innovations (where an innovation is the difference between the observation and the background field), since two of the analysis schemes used in our tests operate in “innovaton space”.

General sub-sampling approaches assign equal priority to all the observations when there is no predisposition for choosing one observation over the other. But, some observation values can be more important in that they provide more information to a data analysis. This motivates the identification of regions with high information content and retention of a higher percentage of observations from these regions.

2.1 *Box Variance (BV) Method*

The BV method (Zavodsky et al., 2006) divides the analysis domain into boxes with 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. All other data points within that box are eliminated. However, if the variance is greater than the threshold, no thinning occurs within the box (i.e. all observations/innovations are retained). Herein, the observation variance threshold is set to 0.50 and innovation threshold is set to 0.40. These threshold values have been selected to produce a comparable number of observations to those retained by the IDT. The weaknesses of this methodology are that the thinning is dependent on the box size and the user defined threshold.

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

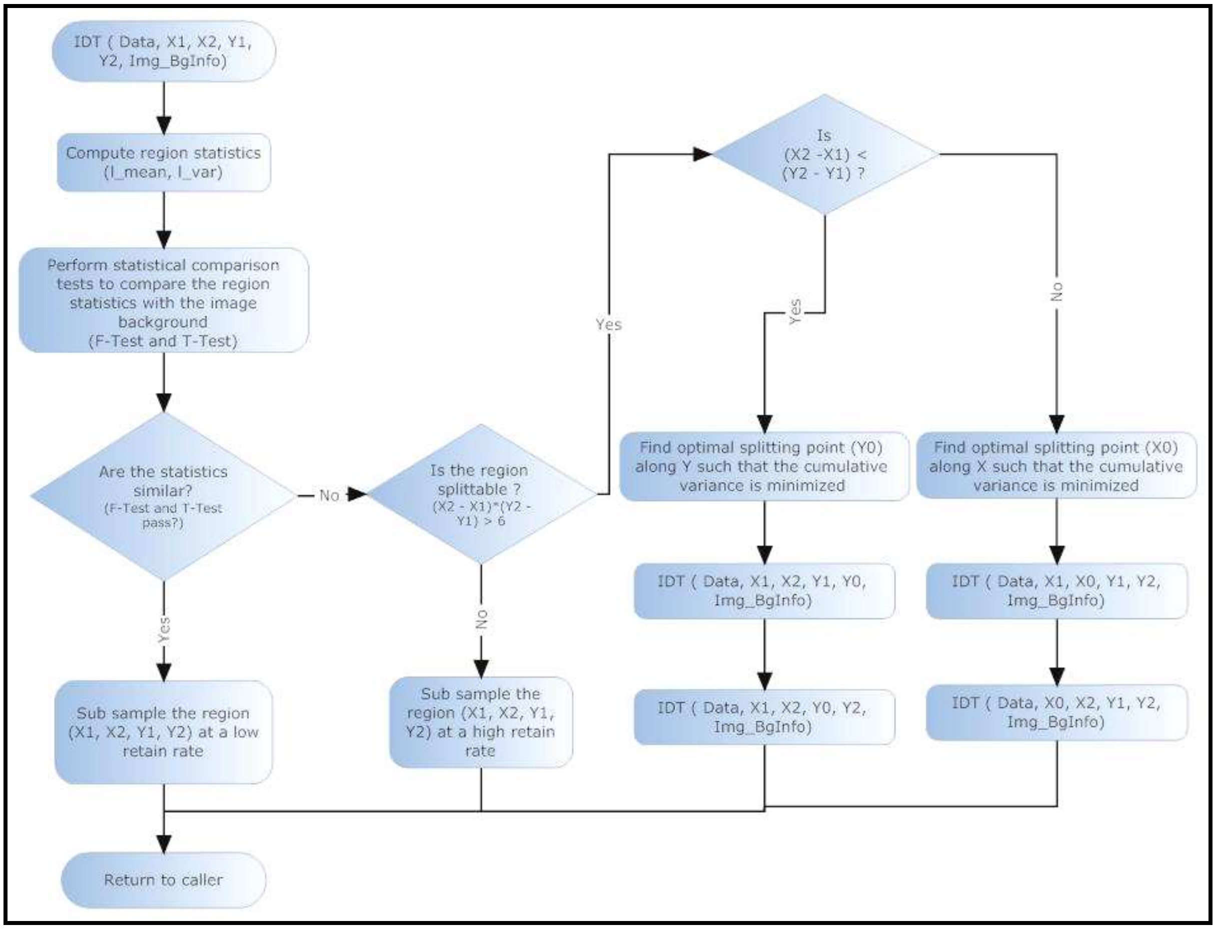
IDT (Data, X1, X2, Y1, Y2, Img_BgInfo)
Compute region statistics (I_mean, I_var)
Perform statistical comparison tests to compare the region statistics with the image background (F-Test and T-Test)
Are the statistics similar? (F-Test and T-Test pass?)
No
Is the region splittable ? (X2 - X1)*(Y2 - Y1) > 6
Yes
Is (X2 -X1) < (Y2 - Y1) ?
Yes
No
Find optimal splitting point (Y0) along Y such that the cumulative variance is minimized
Find optimal splitting point (X0) along X such that the cumulative variance is minimized
IDT (Data, X1, X2, Y1, Y0, Img_BgInfo)
IDT (Data, X1, X0, Y1, Y2, Img_BgInfo)
IDT (Data, X1, X2, Y0, Y2, Img_BgInfo)
IDT (Data, X0, X2, Y1, Y2, Img_BgInfo)
Yes
No
Sub sample the region (X1, X2, Y1, Y2) at a low retain rate
Sub sample the region (X1, X2, Y1, Y2) at a high retain rate
Return to caller

Fig. 1. Decision tree for the Intelligent Data Thinning (IDT) Algorithm.

2.2 *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 now 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. These deviant pixels contribute the most to the cumulative variance of the region and are 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 have high information content and are sub sampled at a higher retention rate. The IDT algorithm (Ramachandran et al., 2005) 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.

One step of the recursive IDT algorithm is depicted in the flowchart shown in Fig. 1. Simple description statistics (mean and variance) are computed for the region. 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 (1993). If the region is too small, it cannot be split, so it is sub sampled at a higher retention rate.

3. EXPERIMENT DESIGN

3.1 *Truth, Background and Observations*

The experiments were designed so that the truth, background, and observational fields are specified and thus known explicitly. While this is a viable approach for synthetic data, it is not necessarily replicable in real-world settings where background error statistics are generally unknown. The truth field (175 x 175 grid) was intended to replicate a temperature field associated with a peninsula whereby two regions of strong gradient separate regions of relatively little temperature gradient.

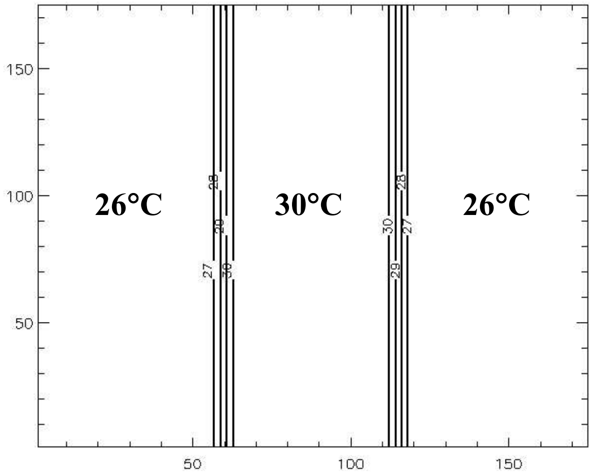

Fig. 2 Truth field. The temperature distribution is intended to replicate a peninsula with strong coastal gradients. Contours of temperature are every 1°C.

The synthetic background field was created to provide spatial correlation statistics that were known and that did not violate the theoretical statistical assumptions of optimal interpolation. The relevance of the results obtained using well defined synthetic data versus data from an operational setting will depend on the quality of the background field which, in practice, is often less well behaved. Additionally, operational data sets may include data from platforms with varying observational error as opposed to the uniform observation error considered here.

The background field was generated following the work of Evensen (1994). 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. 3). The perturbation field created with this method has no knowledge of the high temperature gradient regions. As a result, this approach produces a background field that contains the same error decorrelation in these regions as any other region throughout the grid. Thus, the perturbation fields were smoothed in the regions of the temperature gradient to provide a more realistic background field. This adjustment did not significantly change the resulting variance and decorrelation statistics for the full domain.

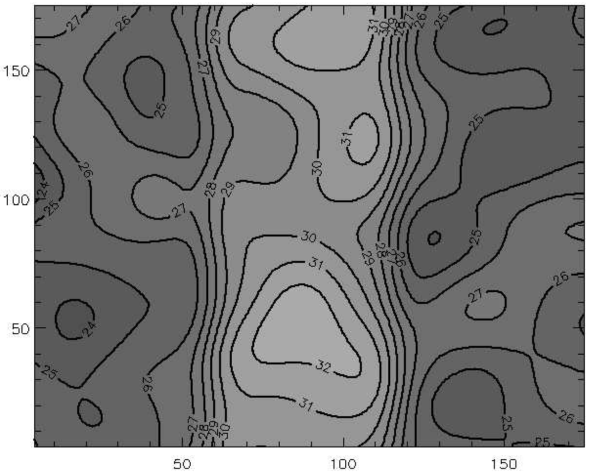

Fig. 3 Background field with observational error variance of 1 deg^2 and a decorrelation length of 25 grid points.

Observational temperature data were created across the analysis domain on a two-dimensional grid (a 58 x 58 grid with spacing of 3 times the length of the analysis grid). Spatially uncorrelated error (white noise) was introduced into the observations with a variance of 0.25. The error in the observations was set significantly less than the background, a plausible and desirable scenario. The various thinning strategies (a total of 7) were applied to these sets of observations. Fig. 4 depicts the full set of observations along with two of the thinned data sets. The BV and IDT box thinning strategies were applied to both the observational data and the innovation data. The innovation data set was created by interpolating the background field to the observation locations. Both the BV and IDT algorithms key on the variability amongst the observations in the gradient regions. The number of data points retained in the gradient regions is substantially higher.

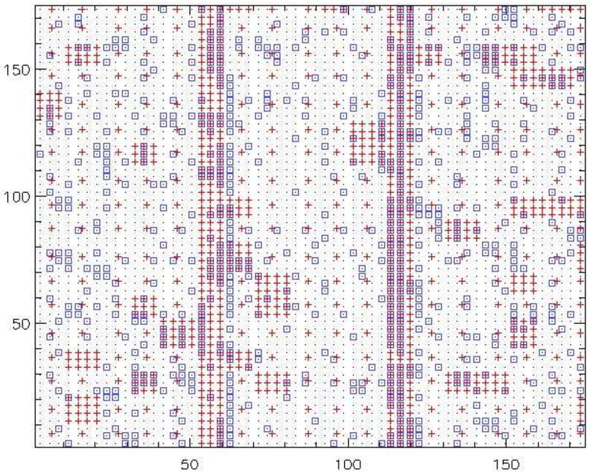

Fig. 4 Observation locations for the full set of observations (grayl dot), the Box Variance method conducted on observations (blue box) and the IDT algorithm conducted on observations (red X).

3.2 *Analysis Schemes*

3.2.1 Bratseth Analysis

The Bratseth analysis is a successive corrections scheme that converges to optimal interpolation (OI) with sufficient iterations (Bratseth, 1986) with an advantage in computational speed over other analysis methods (Lazarus et al, 2002 and Deng et al, 2006), Specification of the background field and observational errors are required for this technique. The implementation used for our experiments followed the method of Kalnay (2003).

To ensure convergence of this iterative method, the analysis was run (with the full set of observations) until the resultant root-mean-square error (RMSE) difference between the *n*- and *(n+1)*-iteration runs was less than 0.0001. While this does not necessarily guarantee convergence to OI, it is reasonable to assume that subsequent iterations will not significantly improve the analysis. For the background field described herein, 87 iterations are required for convergence.

The background error covariance is assumed to be Gaussian with a magnitude set to 1.0 in accordance with the prescribed background error, and the observations covariance matrix is set to 0.25, which is consistent with the random error assigned to the observations. The analyses are performed using a spatial scaling factor of 25 units, which is taken from the specification of the decorrelation length field used to generate the background field.

3.2.2 Kriging Analysis

Kriging analysis is an interpolation scheme that originates from geostatistics and is comparable to optimal interpolation methods (Cressie, 1990). The implementation of Kriging used here is termed Ordinary Kriging. We chose an exponential covariance function (variogram) for modeling the error covariance with values set to be consistent with the decorrelation length and error prescribed in the background and observations. 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 via which to directly assess the impact of the background field on the analyses.

4. RESULTS

The RMSE results of the 3 analysis systems using the full observational data set and the 7 filtered data sets are contained in Table 1. As expected, Kriging without use of the background field produced the highest analysis error in all cases. The Bratseth method produced the lowest RMSE in all cases. Interestingly, the lowest error was the Bratseth method using the IDT filtered data and was even lower than using the full data set. It is believed that the high RMSE using the Bratseth scheme with the full observation set is related to the weighting of observations in high data density regions as opposed to a problem with analysis convergence.

RMSE averaged over the full domain is somewhat misleading in terms of the quality of the analysis. Hence, Table 2 lists the RMSE in the regions of the strong temperature gradient. The lowest RMSE in the gradient regions for the full set of observations was produced by the Kriging analysis using the full set of observations. The Bratseth technique performed the poorest in the gradient regions for the full set of observations, but overall, the lowest RMSE in the gradient regions was produced by the Bratseth analysis with use of the IDT-filtered observations. The difference between use of the observations for the filtering techniques versus use of the innovations for the filtering is more clearly seen within the gradient regions. In the gradient regions, the RMSE is higher for all analysis schemes with the use of the innovation based BV and IDT filtering methods versus the observation based BV and IDT filtering methods.

Table 1. RMSE (full domain) for the three analysis schemes using 8 different thinning strategies.

Method (#obs)	Kriging NB	Kriging	Bratseth
Full (3364)	0.0643	0.0607	0.0583
Sub_3 (400)	0.1503	0.1055	0.0907
Sub_6 (100)	0.7837	0.3656	0.2196
Sub_9 (49)	1.6730	0.6111	0.3951
BV_obs (931)	0.1253	0.1235	0.0867
BV_ino (1068)	0.1549	0.1248	0.0612
IDT_obs (721)	0.1582	0.1129	0.0567
IDT_ino (950)	0.1352	0.0811	0.0490

Table 2. RMSE (gradient regions) for the three analysis schemes using 8 different thinning strategies.

Method (#obs)	Kriging NB	Kriging	Bratseth
Full (3364)	0.0971	0.0820	0.2909
Sub_3 (400)	0.5353	0.2483	0.3238
Sub_6 (100)	1.4342	0.6179	0.5376
Sub_9 (49)	2.0103	0.7620	0.8083
BV_obs (931)	0.2218	0.2045	0.0917
BV_ino (1068)	0.3702	0.2810	0.1452
IDT_obs (721)	0.2987	0.1505	0.0960
IDT_ino (950)	0.4136	0.2020	0.1315

Table 2. RMSE (gradient regions) for the three analysis schemes used as a function of 8 different thinning strategies.

Although the error statistics in the two tables are somewhat informative, direct comparison is difficult as the error is not normalized for the varied number of observations. The RMSE plotted in Figures 5 and 6 on a log/log plot following Anderson et al. (2005). The display of information in this way helps assess the quality as a function of computational expense which is a nonlinear function of the number of observations. The best error results tend towards the lower left corner of the plot (low RMSE and low number of observations) while the poorest results tend towards the upper right (higher RMSE and higher number of observations).

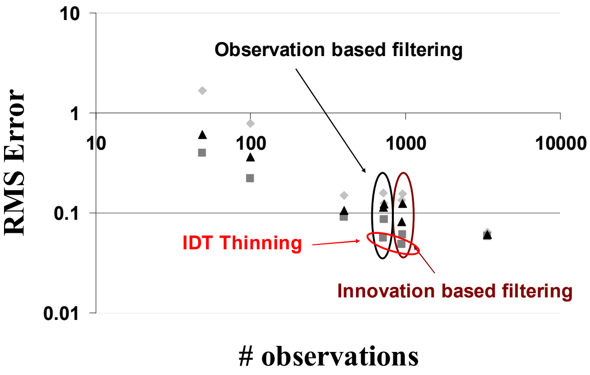

Figure 5. RMSE for the full domain for Kriging without a background field (gray diamonds), Kriging with a background field (black triangles) and Bratseth (gray squares). Observation-based filtering is delineated by the black ellipse; innovation-based filtering is surrounded by the maroon ellipse.

The results as shown in Figure 5 indicate that for the full domain, the IDT thinning algorithms (circled in red) produced the best results using the Bratseth analysis scheme. The use of innovations in the BV and IDT filtering algorithms produced slightly better results for the full domain; the BV and IDT filtering schemes based on the observations produced better results in the gradient regions (Fig. 6). It is interesting that fewer observations are retained using observation-based filtering (shorter computation time), and the errors are smaller for each analysis type.

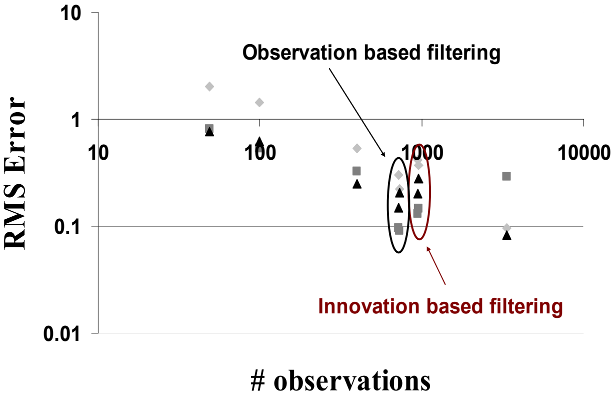

Figure 6. As in Figure 5, but for RMSE in the gradient regions.

The initial findings indicate that the analysis error is both a function of the observational filtering and the analysis system. The differences between the Bratseth and Kriging analyses are indicative of variability within optimal interpolation schemes, as they are not anticipated to converge to the same solution. As anticipated, the Kriging without a background field had the highest RMSE.

The innovation based data thinning degraded the analysis results even though two of the analysis systems work in innovation space. Thinning algorithms that operated on the observational data performed the best. Albeit limited, this result is positive in the sense that data reduction does not depend on the particular analysis system and thus could be applied at the "front end" to avoid bandwidth limitations etc. Overall, the IDT data thinning algorithm outperformed the other methods for this set of tests.

5. FUTURE WORK

Evaluation of the data filtering algorithms in the synthetic data environment will continue in order to address the issue of the optimal thinning of data. It is not clear, for example, to what degree the thinning algorithms performance is tied to the quality of the background field and observations. Also, systematic modification of the thresholds in the BV and IDT algorithms will be conducted to assess the sensitivity of the filtering algorithms to these parameters and to determine the optimal settings for the given analysis systems.

The next phase of the work will be to evaluate the thinning algorithms on real data streams from 1) sea surface temperature from the Moderate-resolution Imager Spectroradiometer (MODIS) direct broadcast and 2) temperature and water vapor profiles derived from the Atmospheric Infrared Sounder (AIRS) instrument aboard the Aqua EOS platform. The application of algorithms will be extended to discontinuous data streams (e.g. satellite sea surface temperature swaths with gaps due to cloud cover).

REFERENCES

Anderson, J.L., B. Wyman, S. Zhang, and T. Hoar, 2005: Assimilation of surface pressure observations using an ensemble filter in and idealized global atmospheric prediction system. *Journal of the Atmospheric Sciences*, **62**,2925-2938.

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

Deng, X. and R. Stull, 2005: A Mesoscale analysis method for surface potential temperature in mountainous and coastal terrain. *Mon. Wea. Rev.,***133**, 389-408.

Cressie, N., 1990: The origins of Kriging. *Mathematical Geology*, **22**, 230-252.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**, 10143-10162.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge Press, 341 pp.

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

Purser, R. J., D.F. Parrish and M. Masutani 2000: Meteorological observational data compression; an alternative to conventional 'Super-Obbing'. NCEP Office Note 430. Available online at: http://www.emc.ncep.noaa.gov/mmb/papers/purser/on430.pdf

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.

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

Zavodsky, B., S. Lazarus, R. Ramachandran, and X. Li, 2006: 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.