9.2 EVALUATION OF AN INNVOATION VARIANCE METHODOLOGY FOR REAL-TIME DATA REDUCTION OF SATELLITE DATA STREAMS

Bradley T. Zavodsky*
Earth System Science Center, University of Alabama in Huntsville, Huntsville, Alabama

Steven M. Lazarus
Department of Marine and Environmental Systems, Florida Institute of Technology, Melbourne, Florida

Rahul Ramachandran and Xiang Li
Information Technology and Systems Center, University of Alabama in Huntsville, Huntsville, Alabama

1. INTRODUCTION

Remote sensing data have become important sources of meteorological observations for numerical weather prediction. Radiance, atmospheric profiles, and integrated quantities derived from satellites provide a significant source of real-time data over data-sparse regions. While these observations contain valuable information, their voluminous nature can be problematic. In particular, some of the data may actually be redundant for a data assimilation system because they do not necessarily add information to the analysis. Despite increasing computational resources, the real-time assimilation of large remote sensing data sets into mesoscale models remains somewhat impractical. As the use radar and satellite data in operational analysis systems continues to increase, effectively handling the large volume of data becomes an ever-increasing challenge. In addition, the reduction in data burden so streamlines an analysis that it permits additional iterations operationally thereby improving the overall quality of the analysis (Purser et al. 2000).

An issue associated with data reduction is the retention of second order features such as gradients (e.g. Lorenc 1981, Hillger and Purdom 1990, and Purser et al. 2000). A successful thinning algorithm should reduce data redundancy while maintaining analysis fidelity. The final analysis should, in general closely resemble the analysis produced from its full-data counterpart but should be computed in a fraction of the time. Within a given domain (at a given time) there may be widespread and/or isolated regions of redundant observations. Hence, one challenge of data reduction is knowing where it is appropriate to thin the data. Here, we propose using innovations (background minus observation) to identify whether or not observations themselves are redundant. By using innovations rather than the observations alone, we attempt to directly account for the 'quality' of an analysis first-guess field with respect to the observations.

Herein, we compare three different methodologies for data removal: subsampling every 7th observation (a technique common to many operational systems), a box variance method, and a variance F-test method. We use a synthetic data set developed to simulate satellite data on a faux domain with topographical (i.e., land/water) features that crudely approximate the Florida peninsula where relatively strong temperature (and moisture) gradients can occur along the coastlines. In part, our goal is to evaluate the various thinning algorithms by comparing analyses from each of the methods to the truth field. It is worth pointing out that this approach may not necessarily be optimal in the context of producing a set of initial conditions that minimize forecast error. Rather, the approach presented here attempts to minimize analysis error.

2. THE SYNTHETIC DOMAIN

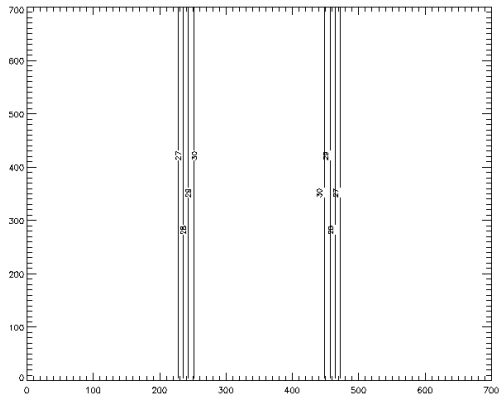

Fig. 1. Truth field for synthetic domain.

* *Corresponding author address:* Bradley T. Zavodsky, 320 Sparkman Dr., Huntsville, AL 35806; e-mail: Brad.Zavodsky@nasa.gov.

To demonstrate how a thinning algorithm might work with real data, we use a synthetic data set on a domain that resembles a typical 2-m land/sea temperature gradient scenario over the Florida peninsula (see Fig. 1). The domain is a 700 × 700 unit box with a 4-unit-resolution analysis grid (i.e., 30,625 grid points). The simulated peninsula is relatively warm and flanked by sharp gradient regions adjacent cooler and homogeneous surroundings representative of the ocean. Coastal regions are those located between 220 and 250 units and 450 and 480 units along the abscissa. The gradient is 7.5^{o}/1 unit. Ocean regions are areas located less than 220 units and greater than 480 units on the abscissa, and the land region is located between 250 and 450 units.

Although unrealistic, for simplicity we assume that the synthetic background field is constant everywhere with a value consistent with that of the ocean region. Approximately 7,800 observations with resolution of 8 units are evenly spaced across the domain to simulate the high-density common in remote sensing data (see Fig. 2a). The observations are defined by sampling the truth field and adding random Gaussian error of ± 1^{o}.

3. THE THINNING ALGORITHMS

As previously mentioned, three distinct thinning methodologies are tested. A comparison of these methods follows. Fig. 2 shows observation locations of the full data set and each of the thinned data sets used herein.

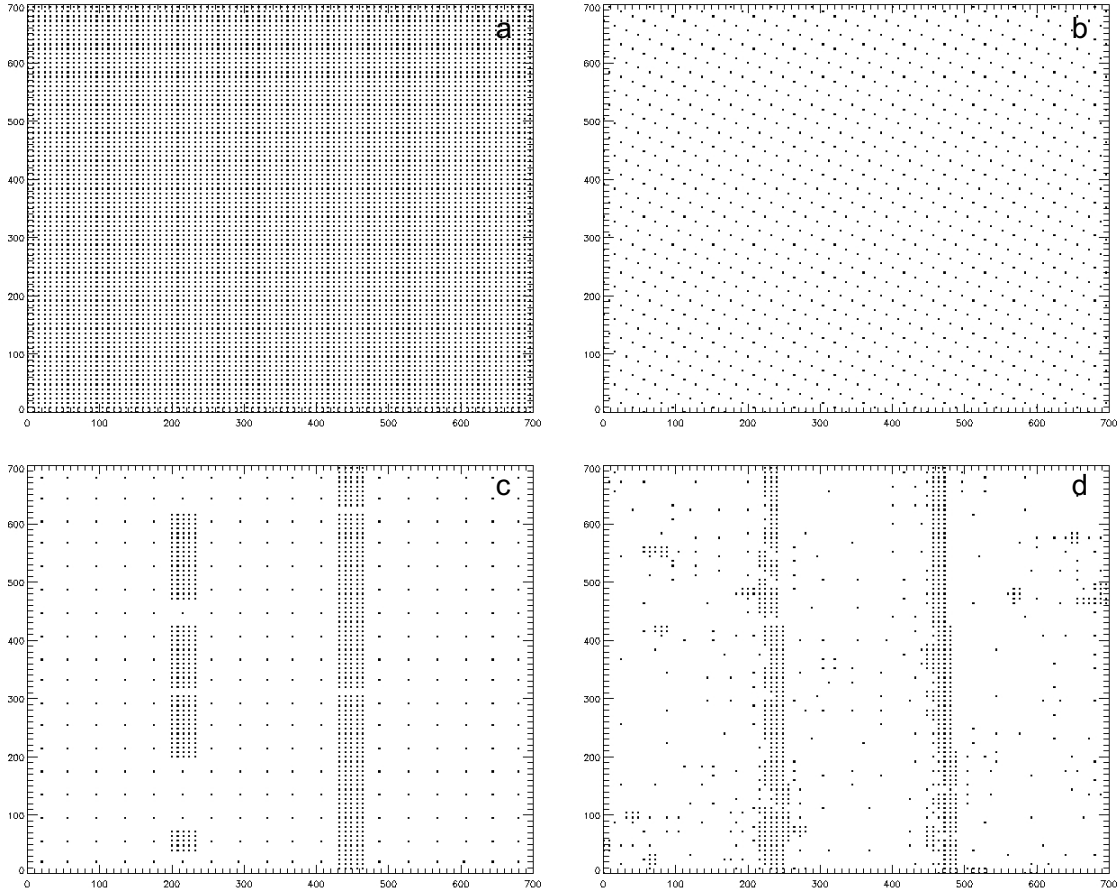

Fig. 2 Data fields used for testing: a) full data set (7,744 points), b) thinning by subsampling method (1,107 points), c) thinning by box variance method (980 points) (variance threshold = 0.465), and d) thinning by IDT (1,000 points) (significance level = 0.1; variance threshold = 0.099).

3.1 *Subsampling Method*

The subsampling thinning approach systematically retains every 7th observation (e.g. Fig. 2b).

3.2 *Box Variance Method*

For the box variance method, the synthetic domain is divided into 324 boxes. The algorithm

prescribes a thinning radius to each box based on the variance of the innovations within that box whereby observations located within the given radius are combined to form a superobservation. The superobservation is created by systematically combining (via linear interpolation to an equidistant point) pairs of observations that are the statistical average of the two original values. Once an observation is used to create a superobservation, it is removed from the data set. This process continues until no observation (or superobservation) pairs are within the predetermined thinning radius.

In the presence of a gradient (and constant background field), we want to retain more of the observations as to better resolve the gradient. As a result, in areas of high variance, the thinning radius is small, and in areas of lower variance, the thinning radius is such that only one representative observation will remain for each box. In an effort to ensure comparable statistics for each of the methods, the box innovation variance threshold (i.e. the variance level whereby observation removal is triggered) is selected to retain approximately 1000 observations. Fig. 2c shows the thinned field for the box variance method.

3.3 *Intelligent Data Thinning Algorithm*

The Intelligent Data Thinning (IDT) algorithm (Ramachandran et al. 2005) searches for regions with high spatial frequency (large variances) and keeps all the data points from these regions. Subsampling is performed on regions with low spatial frequency (low variances) to thin the data to a representative point by recursively dividing the data into four quadrants. For each quadrant, the algorithm then calculates an objective measure. If the objective measure is greater than the user-specified threshold, the algorithm continues by dividing that quadrant into four sub-quadrants, and repeats this procedure for each of the sub-quadrants. This process continues until one of two criteria are met: 1) if the objective measure is less than the user threshold, then the algorithm terminates that recursive path and the center data point of the quadrant is used as the representative thinned value, or 2) the recursion reaches the lowest level where the quadrant contains just four points. For the latter condition, the algorithm saves all four points.

The objective measure used in the IDT algorithm is the statistical F-test, which evaluates the hypothesis that two sample distributions have different variances by evaluating the null hypothesis that their variances are consistent. A hypothetical data region with a variance based on the product of the user-specified threshold and the global mean is compared against the variance of each of the data quadrants during the recursion. The algorithm calculates the F-test probability using the size of the quadrant for the degrees of freedom. If the F-test probability is within the acceptable limit, the null hypothesis holds, meaning the variances are similar. In this case, the algorithm thins the quadrant and the recursion terminates. In the case where the null hypothesis fails, the algorithm continues to perform the recursion on the next level of quadrant decomposition.

4. KALNAY ANALYSIS

All analyses are performed using an assimilation scheme described by Kalnay (2003). An analysis is produced from weighted corrections of a background (i.e., first guess) field. The Kalnay analysis is a variation of the Bratseth (1986) approach whereby the observation values are iterated using

$$d_v = Ad_{v-1} + d_o, \tag{1}$$

where d_v is the v^{th} iteration of the innovation vector and d_{v-1} and d_o are the previous and initial innovation vectors respectively. A is a weighting matrix comprised of elements

$$a_{ij} = \delta_{ij} - \left(b_{ij} + \delta_{ij} r_{ij} \right) / m_{jj}, \tag{2}$$

where δ_{ij} is the Kronecker delta (i.e. 1 if i = j, 0 if i ≠ j), b_{ij} and r_{ij} are the background and observation error covariances respectively, and m_{jj} is an observation density matrix defined by

$$m_{jj} = \sum_{k=1}^{p} \left| b_{jk} + \delta_{jk} r_{jk} \right|, \tag{3}$$

where p is the number of observations. Once the correction vector, d_n, has been calculated, the grid point analysis is obtained in one pass by

$$\phi_g^a = \phi_g^b + b_{g1} \Lambda\ b_{gp} \begin{pmatrix} 1/m_{11} & 0 & \Lambda & 0 \\ 0 & 1/m_{22} & \Lambda & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \Lambda & 1/m_{pp} \end{pmatrix} \begin{pmatrix} d_1 \\ \mathrm{M} \\ d_p \end{pmatrix}, \quad (4)$$

where ϕ_g^b and ϕ_g^a are the background and resulting analysis values respectively for a given grid point.

We use 20 iterations to produce a correction vector in the experiments shown here. The error covariance matricies are assumed to be Gaussian with the error variance estimated directly *via* an average of the squared differences between the truth and background (or truth minus the pseudo observations). This field is assumed to be both spatially homogeneous and constant for each analysis and is set to 0.0140. Defining the error covariance matrix in this manner is applicable to synthetic data but unrealistic for real-world applications, as truth is never known. Analyses are performed using a spatial scaling factor of 60 units.

5. RESULTS

To quantify our results, each thinned analysis is compared to the truth field. Fig. 3 shows difference fields between the truth field and each analysis where the observations are assigned a ±1° Gaussian error. The root mean square (RMS) error statistic between the truth the background, and full and thinned analyses is shown in Table 2.

Because the background field is set to a constant value (equal to that of the truth over the ocean region), there is no RMS error between the background and truth in the ocean region. For the other regions (i.e., the coastal and land), the background RMS errors are much larger than that of the analyses. These regions are as defined in Section 2. Ideally, the full data set will produce the best analyses (i.e., smallest RMS error). This is the case except for the analysis over the coastal region where the IDT thinning method actually yields the best analysis. The smaller RMS error for the IDT thinned data is an artifact of a combination of the large scaling factor (set to 60 units) and the resulting data distribution, which results in a "smoother" analysis over the gradient region for the full data experiment.

In the gradient region, the IDT method has the smallest RMS error followed by the subsampling and the box variance methods. One might expect that because of the relatively coarse observation distribution produced by the subsampling algorithm (every 7th observation) that the gradient will not be well resolved. However, the box variance method yields the largest RMS errors because there are large gaps whereby data are not retained by the algorithm (e.g., Figs. 2 and 3). Also, the location of both gradients is shifted 20 units in the box variance method. In contrast to the box variance method, the IDT data correctly identifies the gradient and is dense with no gaps in the gradient region. The box variance method could be improved by 1) thinning using smaller boxes to better capture the location of the gradient and/or 2) increasing the variance threshold to add more observations in the gradient region.

The IDT does not perform as well as the other methods in non-gradient regions (i.e., ocean and land) where the other thinned have small differences (–0.25° to 0.25°; e.g., Fig. 3) between the truth and analysis. The IDT method retains less than half of the observations over the ocean and land regions compared to the subsampling method (see Table 1). Even though the innovations are correctly identified as redundant in non-gradient regions, the data reduction should be (but is not) accompanied by a rescaling of the analysis parameters (i.e., we retain a constant error covariance and scaling factor). Ideally, as the data are thinned, the individual observations should receive more weight. Additionally, the analysis parameters are clearly not spatially homogeneous nor are the errors necessarily isotropic as assumed here. These issues are currently being investigated in order to tune the analysis to optimally account for the reduction in data density over each region.

Table 1 Observation distribution by region for each thinning method.

	Ocean	Coastal	Land
Full	4840	704	2200
Subsampling	692	101	314
Box Variance	361	274	345
IDT	317	561	122

6. CONCLUSIONS/FUTURE WORK

Remote sensing data are important sources of meteorological observations over data sparse areas. The large amount of data make these data a

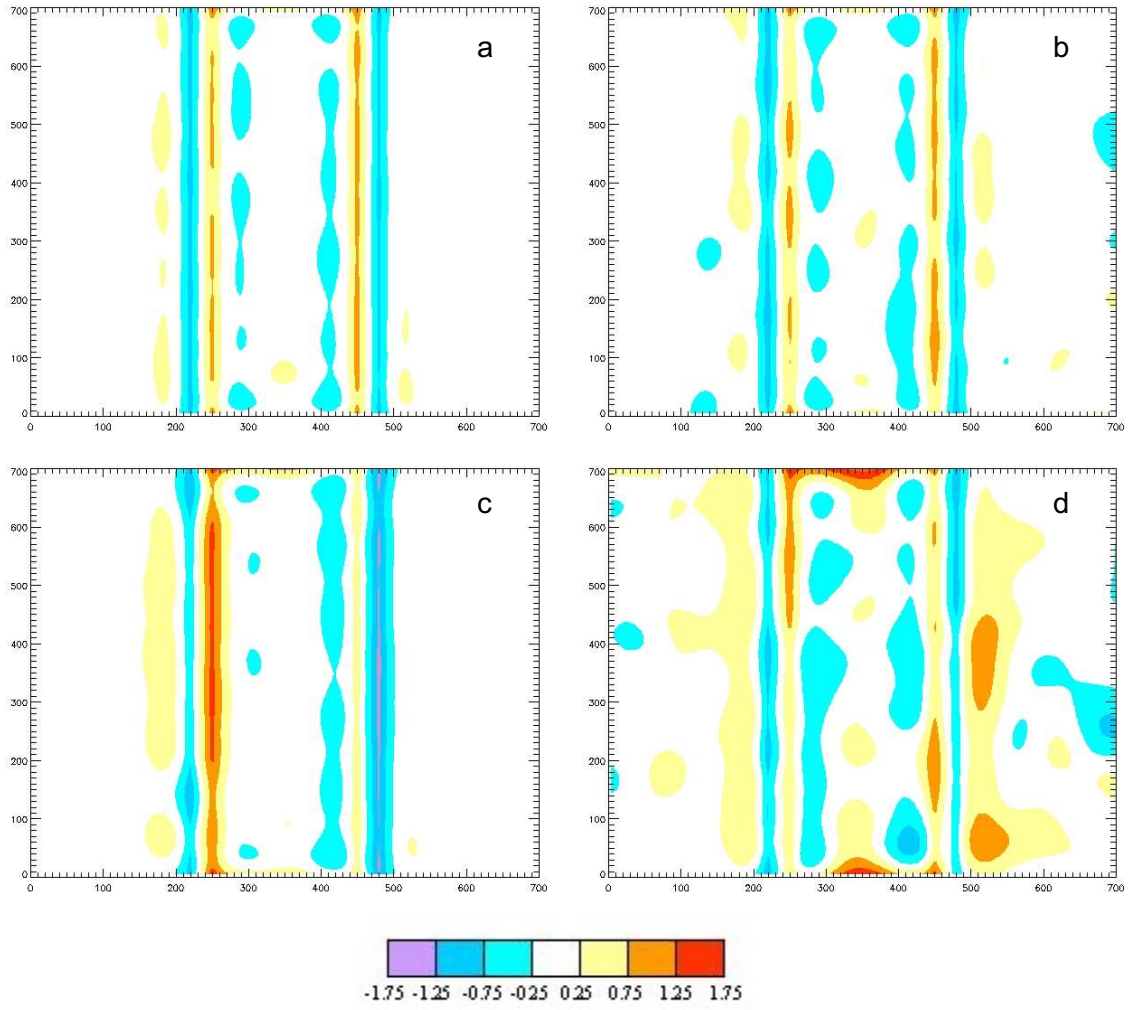

Fig. 3 Difference fields (analysis minus truth) for a) the full data analysis, b) the subsampling method, c) the box variance method, and d) the IDT method. Each analysis was performed using observations with ±1^{o} Gaussian error.

Table 1 Root mean square errors between the different analyses and the truth field stratified by thinning methodology and region.

		Background *vs.* Truth	Full *vs.* Truth	Thinned *vs.* Truth
Subsampling	Ocean	0.0000	0.1768	0.1992
	Coastal	2.3043	0.5197	0.5246
	Land	4.0000	0.2719	0.2802
Box Variance	Ocean	0.0000	0.1768	0.2207
	Coastal	2.3043	0.5197	0.7144
	Land	4.0000	0.2719	0.3472
IDT	Ocean	0.0000	0.1768	0.3282
	Coastal	2.3043	0.5197	0.4944
	Land	4.0000	0.2719	0.4220

viable candidate for data compression. Three different compression algorithms, which depend on the quality of the background field (and observations) and the 'weather of the day', are tested over a synthetic peninsula in an attempt to identify redundant information. To evaluate impact of the loss of information as a result of the data reduction, a relatively straightforward statistical analysis

is applied. Our goal, in part, is to expedite the analysis process for operational applications while simultaneously maintaining analysis fidelity. Because we have created synthetic data, we are able to directly gauge the quality of the resulting analyses. In reality, there is no truth field and thus the evaluation metrics are less certain (e.g., some component of a model forecast). Unfortunately (but not surprising) our results indicate that it can be problematic to gauge the quality of "thinned" analyses by using the full analysis (the latter of which in practicality is all we have).

The impact of various data compression methods on the resulting analyses is regionally dependent. Over the coastal region—where meteorological gradients are common—the IDT algorithm produces an analysis with the lowest RMS error when compared against the truth. In contrast, this same method produces the analysis with the largest RMS errors over both the ocean and land (i.e., relatively constant innovation) regions. The poorer performance of the more sophisticated IDT approach over the homogeneous regions is a direct result of failing to adjust the analysis parameters (i.e., error covariance and length scale) which were intentionally held fixed here in order to simplify these initial experiments.

Future work will focus on three different aspects including 1) tuning the analysis parameters, 2) the introduction of spatially varying analysis parameters, and 3) the transition to actual satellite data. In terms of the latter, we intend to apply the methods to data from the AIRS instrument aboard the Aqua EOS platform.

REFERENCES

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

Hillger, D.W. and J.F.W. Purdom, 1990: Clustering of satellite sounding radiances to enhance mesoscale meteorological retrievals. *J. Appl. Meteor.*, **29**, 1344-1351.

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

Lorenc, A.C., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, **109**, 701-721.

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.