Evaluation of Data Reduction Algorithms for Real-Time Analysis

STEVEN M. LAZARUS AND MICHAEL E. SPLITT
Florida Institute of Technology, Melbourne, Florida

MICHAEL D. LUEKEN
NCEP/Environmental Modeling Center, Camp Springs, Maryland

RAHUL RAMACHANDRAN, XIANG LI, SUNIL MOVVA, AND SARA J. GRAVES
Information Technology and Systems Center, University of Alabama in Huntsville, Huntsville, Alabama

BRADLEY T. ZAVODSKY
Marshall Space Flight Center, Huntsville, Alabama

(Manuscript received 21 April 2009, in final form 9 December 2009)

ABSTRACT

Data reduction tools are developed and evaluated using a data analysis framework. Simple (nonadaptive) and intelligent (adaptive) thinning algorithms are applied to both synthetic and real data and the thinned datasets are ingested into an analysis system. The approach is motivated by the desire to better represent high-impact weather features (e.g., fronts, jets, cyclones, etc.) that are often poorly resolved in coarse-resolution forecast models and to efficiently generate a set of initial conditions that best describes the current state of the atmosphere. As a precursor to real-data applications, the algorithms are applied to one- and two-dimensional synthetic datasets. Information gleaned from the synthetic experiments is used to create a thinning algorithm that combines the best aspects of the intelligent methods (i.e., their ability to detect regions of interest) while reducing the impacts of spatial irregularities in the data. Both simple and intelligent thinning algorithms are then applied to Atmospheric Infrared Sounder (AIRS) temperature and moisture profiles. For a given retention rate, background, and observation error, the optimal 1D analyses (i.e., lowest MSE) tend to have observations that are near regions of large curvature and gradients. Observation error leads to the selection of spurious data in homogeneous regions of the intelligent algorithms. In the 2D experiments, simple thinning tends to perform better within the homogeneous data regions. Analyses produced using AIRS data demonstrate that observations selected via a combination of the simple and intelligent approaches reduce clustering, provide a more even distribution along the satellite swath edges, and, in general, have lower error and comparable computational requirements compared to standard operational thinning methodologies.

1. Introduction

In the practical world of data assimilation (DA), satellite and radar data are commonly reduced by removing a portion of the data or by combining high spatial and temporal resolution observations. Despite the obvious benefits of high spatial resolution in data-sparse regions, large data volumes can have an adverse affect on the computational costs and functionality of a real-time forecast–analysis system. Data reduction can also positively impact an analysis system if the observation error is spatially correlated but is assumed not to be (e.g., Bergman and Bonner 1976; Liu and Rabier 2003; Bondarenko et al. 2007). As a result, operational data reduction is a common practice and often tends toward a simple and computationally efficient methodology referred to as subsampling. In this approach, data are

1 This approach is also referred to as “stepwise,” “simple,” or “uniform” thinning (Ochotta et al. 2005, 2007).
retained by systematically selecting one of every $N$ observations regardless of its relative importance to an analysis or forecast. Conversely, targeting observing systems can be thought of as belonging to a sophisticated, but rather expensive, thinning methodology whereby information from a model is used to identify regions where uncertainty would have large impacts on the resulting forecast (e.g., Rabier et al. 2008). In the latter, the goal is to retain a portion of the data that optimizes the assimilation by identifying a data subset that maximizes the information content while simultaneously reducing the number of observations. Simple data thinning often involves a degree of information sharing whereby the retained observations are replaced by combining multiple observations into a single datum typically referred to as a superob (Lorenc 1981; Purser et al. 2000). Simple thinning, which also includes random observation sampling, is a nonadaptive strategy in that the observation values themselves are not considered in the thinning process (e.g., Bondarenko et al. 2007). Regardless of the approach, it is useful to quantify the impacts of thinning on an analysis–forecast system, especially since the impacts may not be optimal. Furthermore, intelligent observation reduction may even result in an inferior analysis compared to that of nonadaptive subsampling strategies (Bondarenko et al. 2007). Here, we distinguish between four-dimensional variational data assimilation (4DVAR), for which optimality involves improvements in both the analyses and forecasts, and a stand-alone system where the goal is to maximize analysis fidelity in the presence of data compression. It is the latter metric that we adopt here. The obvious advantage of the latter is that the process is not necessarily tied to a particular forecast system and thus the data reduction protocol can be invoked at the source—thereby liberating both bandwidth and computational resources (Purser et al. 2000).

Despite advances in data assimilation, assessing the value of observations to an analysis remains somewhat difficult, especially in an operational setting. Purser et al. (2000) describe the application of information theory to quantify the impacts of combining dense observations into superob. The theory, based on the concept of information entropy (Shannon 1948), provides a measure of the contribution of observations to the probabilistic estimate of the atmospheric state. Entropy theory has also been applied by Peckham (1974), Rodgers (1976), Eyre (1990), Xu (2007), and Xu et al. (2009). Recent work in the area of adaptive thinning such as top-down clustering and thinning through estimation is also promising (Ochotta et al. 2005; Dyn et al. 2002). In the top-down approach of Ochotta et al. (2005), observations are grouped based on spatial position and measurement values into “clusters” with each cluster having a single representative measurement referred to as the cluster mean. The cluster mean is defined as the observation that minimizes, in a least squares sense, the distance between itself and all other cluster elements. The observation inside each cluster that is closest to a subcluster mean is retained. The top-down approach is attractive from a computational standpoint because it proceeds by iteratively inserting data beginning with a single observation. In contrast, estimation thinning systematically removes observations beginning with the entire dataset (Ochotta et al. 2005). Thinning is contingent upon minimizing the impacts of data removal based on a local approximation of an observation—with error estimates calculated by differencing the observation and the best fit. An advantage of this method is that it directly uses a local estimate of an observation to drive the data reduction. However, the method is potentially expensive and the degree of optimality, within the context of reducing analysis error, remains ambiguous due to 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. A two-dimensional synthetic dataset and accompanying observations are manufactured with known observation and background errors. These data are mined using various thinning methodologies, assimilated into an analysis scheme, and compared against analyses generated using different subsampling methods. We then apply the thinning algorithms to temperature and moisture profiles retrieved from the Atmospheric Infrared Sounder (AIRS) instrument and the Advanced Microwave Sounding Unit (AMSU) on the Aqua Earth Observing System (EOS) platform. Finally, a brief summary and discussion is presented.

2. Thinning methodologies

A description of the different thinning approaches is presented in this section. Each of these methods can be applied in observation or innovation space (i.e., observation minus first guess). Although observation thinning will depend on the first-guess field (section 3), here the goal is to develop a stand-alone thinning algorithm that is independent of the background; thus, thinning is only considered for observation space. The thinning algorithms are constructed using a combination of simple and adaptive sampling methodologies—the latter of which depend on first- and second-order derivatives. This approach is motivated by our desire to better represent high-impact weather features (e.g., fronts, jets, cyclones, etc.) that are often poorly resolved in coarse-resolution forecast models. In contrast to targeting, which relies on
model error growth to identify relevant observations (e.g., Snyder 1996; Langland et al. 1999), the work here is designed to provide a set of initial conditions that best describe the current state of the atmosphere. For example, within the context of cyclogenesis, targeting often identifies regions upstream of the feature of interest as dynamically relevant rather than the feature itself. It is interest in the latter that drives the development of the intelligent thinning algorithms presented here.

a. The direct method

The direct method determines the optimal solution by testing all possible thinned datasets for a given retention number. A sample space of thinned observation subsets, each with \( n \) uniquely distributed elements, is created from a population of \( q \) observations. The number of unique subsets \( N \), each consisting of \( n \) observations, is determined by dividing the total number of permutations of the population by the number of equivalent permutations obtained by permuting the elements of the observation subset; that is,

\[
N = \frac{q!}{(n!)(q-n)!}.
\]

(1)

An optimal observation distribution is identified by running each of the thinned subsets through an analysis system and selecting the set with the lowest mean-square error determined from the difference between the analysis and either a known function (truth) or an analysis using the entire dataset.

b. Intelligent methods

1) Modified density

The modified density adjusted data thinning (mDADT) approach, similar to the top-down estimation error analysis (EEA) described by Ochotta et al. (2007), builds a subset of thinned observation from the ground up by systematically adding data to an initially empty set. The mDADT algorithm uses a sequential combination of a thermal front parameter (TFP, Renard and Clarke 1965), homogeneity, and local variance (LV) measures. The TFP is defined by

\[
\text{TFP} = -\nabla |\nabla \theta| \cdot \nabla \theta / |\nabla \theta| = -\nabla |\nabla \theta| \cdot \hat{e}_{|\nabla \theta|},
\]

(2)

where \( \theta \) is the potential temperature and \( \hat{e}_{|\nabla \theta|} \) is a unit vector in the direction of \( \nabla \theta \). The TFP is the directional derivative of the gradient of \( \theta \) in the direction of \( \hat{e}_{|\nabla \theta|} \) (Renard and Clarke 1965). Observations targeted for retention are those in regions where the absolute value of the TFP is larger than some specified user threshold. The TFP was originally developed as a means by which to objectively analyze (automate) frontal boundaries on a weather map. Fronts are important synoptic-scale weather features and are often poorly resolved, especially by coarse-resolution models. Given its higher order, the TFP is calculated for interior grid points only.

The LV for an observed sample, \( s \), is defined as follows:

\[
\text{LV}(s) = \frac{\sum_{t=N(s)}^{t=N(s)} |f(s) - f(t)|/d(s, t)}{\sum_{t=N(s)}^{t=N(s)} 1/d(s, t)}.
\]

(3)

where \( t \) is a neighboring observation of \( s \), \( N \) is the set of \( t \), and \( d(s, t) \) is the distance between \( s \) and \( t \) and \( f(s) \) and \( f(t) \) are the observed values at \( s \) and \( t \), respectively. The LV metric measures the intensity variance of the neighborhood surrounding an observation. Gradient regions and noisy observations are characterized by large LV values, and homogeneous regions are characterized by small LV values. In the mDADT method, the thinned observations are selected using the TFP, which represents second-order intensity derivatives, and the LV for which high (low) values depict gradient (homogeneous) regions. Observations are ranked and sorted for each metric, individually. The metrics are applied independently, with the retained observations removed from the queues following each iteration. Here, independent is synonymous with the concept of sampling with replacement in that observations are not removed until the end of an iteration. Hence, observation retention does not depend on the order in which the metrics are applied; thus, observations may be selected more than once, resulting in fewer observations than the targeted thinning rate. An iteration-dependent (decreasing) blackout radius \( R \) determines which observations in the queue are added to the thinned subset on successive passes through the data by establishing a local region around the highest-ranked observation within which all other observations are excluded. An initial blackout radius \( R_0 \) is automatically determined and depends on a user-specified thinning rate \( r \) and the desired percentage of homogeneous observations \( h \):

\[
R_0 = \frac{1}{\sqrt{r \times h}}.
\]

(4)

For each iteration, \( R \) is incrementally reduced by an amount equal to half the average distance between adjacent observations in the full dataset. The process continues with the next highest-ranked observation, with the caveat that it must lie outside of the blackout radius. As a result, observations are not necessarily selected in descending order of importance. Thus, some high-ranking observations will be passed over for selection and will
remain in the queue for future iterations. Observations thinned during previous iterations are removed permanently from the queue. This process prevents large data voids in the homogeneous regions of the thinned data subsets with larger $R$ indicating a more homogeneous distribution. The subset of thinned data is an accumulation of observations retained in all iterations. The iteration process stops when the number observations retained equals the selected thinning rate, which is defined as the percentage of observations that are retained from the full dataset. The observation retention percentage for each sampling methodology is user specified and their sum defines the desired thinning rate. The initial black-out radii for the TFP and gradient regions are set to be $0.5 \times R_0$. The iteration process continues until the desired percentage of observations is attained for each metric. Here, mDADT differs from the top-down EEA method in that no optimization process is performed during the iterative selection of thinned observations in mDADT. The degree of fit is calculated once for all observations before the thinning process begins. Therefore, the mDADT method takes significantly less time than the top-down EEA approach of Ochotta et al. (2005).

2) DENSITY BALANCED

The mDADT method produces thinned datasets with relatively uneven observation distributions that favor regions of large TFP or LV values. Although higher sample densities can improve the analyses over regions of interest, the sample density over the rest of an image is reduced for a given data thinning rate. Because variations in data density can be an issue within the context of data assimilation (e.g., Trapp and Doswell 2000), we designed another thinning method: density-balanced data thinning (DBDT). This method uses the same metrics as mDADT for data selection; however, in order to ensure a more uniform observation distribution, $R$ is fixed and not reduced iteratively as it is in mDADT. In addition, the observation selection process is order dependent; that is, observations are removed following each sampling (TFP, gradient, homogeneous) within an iteration step. Hence, unlike mDADT, the observations selected will depend on the order in which the samples are selected. Here, we begin with significant TFP samples, followed by the selection of observations with large LV, and then homogeneous samples. Although $R$ is fixed, it is selected by trial and error and is determined by selecting the realization closest to the desired thinning rate.

3. Synthetic thinning and analysis experiments

To quantitatively evaluate the different thinning algorithms, a series of synthetic experiments are conducted for which the truth, background, and observational fields are specified and thus known explicitly. The experiments are also designed to provide guidance for both algorithm development and real-data applications. Tests are performed using both one- and two-dimensional idealized cases.

The thinned observations are assimilated using the variational (VAR) methodology described by Lorenc (1986). To circumvent the large storage requirements and to mitigate computational costs, VAR uses both localization and a recursive filter (Hayden and Purser 1995), thereby avoiding a direct inversion of the background error covariance matrix. Here, the observation operator is a linear interpolator from model-to-observation space. To aid in convergence, a preconditioner is used (e.g., Gao et al. 2004; Huang 2000) after which the functional is minimized using the conjugate gradient method (e.g., Gao et al. 2004) with the maximum number of iterations set to 100 for the experiments described here. The filter is approximated using fourth-order autoregression to replicate a Gaussian shape (Purser et al. 2003).

a. One-dimensional experiment: Truncated Gaussian

Serving as a precursor to the two-dimensional problem presented in section 2b, an idealized 1D truncated Gaussian (Fig. 1, solid line) with 35 points is assumed to be the truth field. A single observation realization is created by adding white noise to the truth with the observation-to-background error variance ratio set to $1/4$ (0.25). The observations are sampled for a unique combination of five that yield the best analysis as defined by

![Fig. 1. Truncated Gaussian (solid black curve), 1DVAR analysis (dashed curve), first-guess (gray dotted curve), and optimal observation locations (\times's) for analysis length scale of $4\Delta x$, where $\Delta x$ is the grid spacing. See text for details.](image)
the lowest mean-square error (MSE) between the analysis and the true Gaussian. From Eq. (1), there are approximately 325,000 unique spatial combinations of five observations obtained from a single data realization (\( q = 5 \), \( n = 5 \)). The first-guess field is set to that of the base of the function (Fig. 1, dotted line). For an optimal analysis, the analysis length scale is given by the background error decorrelation length scale. We ran 325-K analyses for each of the length scales of 2, 4, 8, and 16 \( D_x \), where \( D_x \) is the analysis grid spacing. In these simple experiments, the optimal observations cluster around the function peak with decreasing length scale. The best analysis, which occurs for the 4 \( D_x \) length scale (Fig. 1), places observations at the peak, within the gradient, and at points where the gradient changes (referred to hereafter as anchor points).

To gauge the representativeness of the observation selection, 10 additional realizations are generated, each with a single observation set and 325,000 analyses, using the same error ratio (Fig. 2, bottom 10 rows). Each row represents the observation locations for the best analysis of each realization. Of these, 50% of the observations were located at anchor points; 26% and 18% within the peak and homogeneous regions, respectively; and 6% at the gradient (inflection) points. If the observation error is reduced from 0.25 to 0.05, the percentage of anchor points retained by the direct method remains close to 50%, while all but one of the gradient points are selected in the three realizations shown (Fig. 2, top 3 rows). For this case, no observations are retained in the peak region.

DBDT observation selections from the “best” analysis (lowest MSE) and for 100% TFP with no observation error are also shown in Fig. 2. The best analysis observation subset is chosen from more than 70 realizations in which the retention percentage of one of the sampling methods is systematically varied while holding the other two fixed. Although the retention percentages selected were 50% for gradient, 5% TFP and, 45% homogeneous, in reality, 40% (60%) of the observations selected are TFP (homogeneous). In contrast, for the 10 realizations with an error variance ratio of 0.25, the DIRECT method selects on average 5%, 50%, and 45% gradient, TFP, and homogeneous observations, respectively. Hence, for the same retention percentages, DBDT yields fewer than the desired number of anchor points due to the large observation error. In the absence of observation error, DBDT has no difficulty identifying TFP points (100% TFP; Fig. 2). Hence, in the presence of a low signal-to-noise ratio, simple thinning may be the best option.

We examine the impacts of the quality of the background field on the observation selection by improving the first-guess field by replacing the constant value with a smoothed version of the truncated Gaussian. The number of retained anchor points increases to 85% of the selected observations with 10% in the peak and 5% in the homogeneous regions (Fig. 3). Interestingly, there are no gradient points selected for these cases regardless of the observation error. In the gradient regions, observations add little in the way of analysis information because of the improved first guess. Conversely, the increase in the number of anchor points selected reflects the addition of key observations at the base of the truncated Gaussian where there are now large differences between the observations and the first-guess field. These results suggest that, within the context of data assimilation, a more robust thinning approach should consider innovation space.
Five observations are also selected via a simple thinning approach in which every seventh observation is retained from the full data, using seven unique subsampled datasets generated from the same set of full observations. Again, the background field is equal to the base value of the function. The MSEs for both the direct and simple thinning are shown in Table 1. For the simple thinning, the brackets denote the spread in the analysis MSE from the seven unique subsampled datasets while, for the direct method, the brackets represent the range in the optimal MSE from the 10 observation realizations shown in the bottom rows of Fig. 2. The analysis MSE for the direct method is substantially less than that obtained from simple thinning. Figure 4 shows the lowest analysis MSE (optimal thinning) versus the number of retained observations for the direct method for an analysis length scale of $4\Delta x$. The analysis error asymptotes near five observations for these simple experiments, suggesting that a 15% observation retention rate is sufficient to retain much of the full data analysis fidelity. Assuming that observation errors are uncorrelated, the number of observations for which there is a steep increase in analysis error will depend, in part, on the wavelength of the feature and will shift toward greater (fewer) observations as the wavelength of the feature decreases (increases). The simple thinning mean and standard deviation are provided (Fig. 4, the open square and error bar). While the direct approach is expensive and thus impractical, an observation thinning strategy should, at the very least, yield a lower MSE than that produced via simple thinning.

### TABLE 1. MSE as a function of analysis length scale and constant error variance of 0.25 for the direct, simple, and DBDT thinning methods applied to the 1D truncated Gaussian with a constant background field. Parentheses denote the minimum and maximum analysis MSEs obtained from multiple realizations (see text for details).

<table>
<thead>
<tr>
<th>Length scale ($\Delta x$)</th>
<th>MSE direct</th>
<th>MSE simple</th>
<th>MSE DBDT (CUR)</th>
<th>MSE DBDT (TFP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.16</td>
<td>(1.75, 2.33)</td>
<td>(0.68, 1.17)</td>
<td>(0.61, 2.50)</td>
</tr>
<tr>
<td>4</td>
<td>(0.06, 0.16)</td>
<td>(0.41, 0.86)</td>
<td>(0.61, 2.89)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.19</td>
<td>(0.94, 1.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIG. 4. Analysis MSE vs the number of observations retained for the truncated Gaussian shown in Fig. 1. The filled squares correspond to the lowest MSEs obtained from a pool of approximately $N$ [Eq. (1)] analyses with 1, 3, 5, and 35 observations, respectively; each with an error variance ratio of 0.25 and an analysis length scale of $4\Delta x$. Curve shown is a best-fit power law. The open square and the error bar represent the mean analysis error $\pm \sigma$, respectively, for 7 simple thinned realizations and an observation retention rate of 5.

b. Two-dimensional experiments: Warm peninsula

For the two-dimensional case, the truth field is designed to replicate an idealized temperature pattern that might be associated with a heated peninsula or warm ocean current (Fig. 5a). The background field is generated following the work of Evensen (1994), whereby a pseudorandom two-dimensional field of perturbations is created using a prescribed error covariance and decorrelation length scale (set to 1 and $25\Delta x$, respectively, where $\Delta x$ is the grid spacing and the error variance units depend on the square of the analysis variable, which is unspecified here). When added to the truth, the perturbation field created via this method produces a background field with homogeneous and isotropic error characteristics. However, to provide a more realistic background field, the truth field is first smoothed and then added to the perturbations (Fig. 5b). This adjustment does not appreciably impact the resulting variance and decorrelation statistics for the full grid. The truth, background, and analysis all have the same grid dimensions of $175\times175$. Observations (3364 total) are generated systematically within the analysis domain using an observation separation distance equal to 3 times the length of the analysis grid spacing. Observations are created by introducing uncorrelated error (white noise), with a specified variance, to the truth. The error variance ratio (observation to background) is set to 0.25. Each of the thinning strategies discussed in section 2 is applied to the synthetic observations and the retained data are assimilated into a two-dimensional variational data assimilation (2DVAR) algorithm.

Using the results obtained in section 3a, which indicate a significant decrease in MSE as the percentage of observations retained increases from 10% to 15% (Fig. 4), we target an observation retention rate in this range. A subsample rate of every third observation retains approximately 12% of the total observations. The intelligent
thinning algorithms are tuned using internal parameters (as discussed in section 2) to yield this approximate retention rate. For simple thinning, an equivalent retention rate is obtained by selecting every third observation in the $x$ and $y$ directions. The number of observations retained for the thinned datasets is given by the parenthetical numbers in Table 2. Multiple analyses were generated for simple thinning, and mDADT (where one of the three metrics is held fixed while varying the remaining two—producing over 70 realizations). Gradients are often poorly represented in a model forecast and thus can be enhanced by a good analysis (Anderson et al. 2005). Hence, results are shown, in Table 2, for the analyses that produce the minimum MSE within the gradient region (Fig. 5a, white dashed lines) as well for those that yield the minimum MSE for the full domain. The accompanying optimal analysis length scales ($L$) are also shown for both regions.

Figure 6a depicts the analysis for the case where all observations are assimilated while the remaining panels (Figs. 6b and 6c) are the analyses that produce the lowest MSE within the gradient region for the simple and mDADT thinning algorithms, respectively. The total numbers of observations retained over the full domain are comparable (399 versus 400; see Table 2). However, the numbers of retained gradient observations differ with 40 and 62 for the simple and mDADT algorithms, respectively (see Table 2). If the initial blackout radius [Eq. (4)] is too large, the selectively thinned dataset may omit key observations, while a small initial radius may lead to densely clustered observations. For the warm peninsula, an initial search radius of zero (eight) grid units minimizes the gradient (full domain) MSE. The smaller search radius leads to a preferential selection of observations within the gradient region and a correspondingly large MSE over the full domain for this algorithm (not shown). Drawing for the gradients favors smaller analysis length scales (Table 2) but can lead to degraded statistics within the homogeneous regions due to overfitting. As a result, the mDADT analysis shown in Fig. 6c is selected based on error statistics taken over a slightly expanded gradient area (i.e., larger than that shown in Fig. 5a) that includes a larger portion of the adjacent homogeneous regions. The expansion results in somewhat degraded gradient region statistics while the full domain statistics have improved significantly as a result of reducing the overfit. In addition, the optimal gradient region length scales are larger for these analyses and, more importantly, the thinned observation distribution is very different with a preference for TFP observations over that of gradient and a more even observation distribution throughout the analysis domain. For the 12% retention rate shown, the

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE (full)</th>
<th>MSE (full best)</th>
<th>$L$ (full)</th>
<th>MSE (grad)</th>
<th>$L$ (grad)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FULL</td>
<td>0.052 (3364)</td>
<td></td>
<td>13</td>
<td>0.073 (290)</td>
<td>8</td>
<td>All obs</td>
</tr>
<tr>
<td>SIM</td>
<td>0.097 (400)</td>
<td>0.033</td>
<td>11</td>
<td>0.315 (40)</td>
<td>6</td>
<td>Simple</td>
</tr>
<tr>
<td>mDADT</td>
<td>0.180 (399)</td>
<td></td>
<td>13</td>
<td>0.120 (62)</td>
<td>10</td>
<td>Mod. density adj.</td>
</tr>
</tbody>
</table>
intelligent algorithm performs better than simple thinning in the gradient region. The full domain statistics are dominated by the homogeneous regions where simple thinning prevails (Table 2).

Signal detection in the presence of observation error can be problematic. As a result, a simple filter was applied to the observations prior to thinning. This resulted in a shift in the locations identified as having maximum TFP and thus the retained observations were at less than optimal locations, which degrades the analyses.

4. Thinning real data

Results gleaned from the synthetic experiments are applied to thinning real satellite data in the form of temperature and moisture profiles from AIRS. AIRS is a hyperspectral atmospheric sounder with a spatially varying footprint resolution ranging from 50 km at nadir to 100 km at the scan edges (Aumann et al. 2003). The full dataset consists of only the highest quality AIRS data, as defined by the AIRS version 5 quality indicators (Susskind et al. 2006). Each thinned dataset is a subset of this full set of profiles and is described in detail in the next section. The quality indicators flag questionable data due to possible cloud contamination, often leaving gaps in the data. The variable spatial resolution and missing data can be problematic for some thinning algorithms (e.g., Koch et al. 1983). In addition, AIRS data are thinned using the equivalent potential temperature derived from the temperature and moisture profiles. Equivalent potential temperature is used here because it combines elements of both temperature and moisture into a single variable that can be used for thinning. The thinning algorithms are applied at each AIRS pressure level and are strictly two-dimensional. Results from the 700-hPa pressure level are shown as these results are representative of other tropospheric levels.

Analyses are performed using the Advanced Regional Prediction System (ARPS) Data Analysis System (ADAS; Brewster 1996). ADAS is a univariate successive correction analysis scheme that converges to optimal interpolation without the need for large matrix inversions. The error covariance structure is modeled as being Gaussian and isotropic and the observation errors are assumed to be uncorrelated (Lazarus et al. 2002). ADAS is employed, in part, because it is used operationally at select National Weather Service Weather Forecast Offices. The background field for the analysis is the Weather Research and Forecasting (WRF; Dudhia et al. 1998; Skamarock et al. 2005) model initialized using the 0000 UTC 40-km North American Mesoscale Model (NAM-212) and run to the time of the assimilated AIRS overpass. The background errors are standard short-term forecast errors cited in the ADAS documentation while the errors for the AIRS profiles are based on estimates cited by Tobin et al. (2006). Separate error estimates are used for land and water soundings due to the characteristics of the AIRS retrieval algorithm and its sensitivity to land emissivity. The background and observation errors for temperature and relative humidity used for this study are shown in Fig. 7. The error decorrelation length scale is estimated using the mean of the distance between each observation and its nearest neighbor with longer length scales selected for more sparsely populated data. The length-scale estimate is applied to the final (fourth) analysis iteration, with the first three passes using 4, 3, and 2 times this value, respectively. Four analysis passes appears to be sufficient as the ADAS analysis converges for the full AIRS profile dataset for this number of iterations.

As the truth is not explicitly known for real data, each thinning algorithm is evaluated using an average root-mean-square error (RMSE) calculated from the difference between 1) an analysis using the full AIRS dataset
and an analysis using the datasets obtained from each thinning algorithm (FT) and 2) each thinned analysis and full analysis at the location of withheld observations (DO). As previously discussed, the full data analysis may be suboptimal because the observation error correlation is not taken into account. As a result, an additional metric, referred to as squared analysis increments (SAI), is calculated. The SAI gauges the amount of change imparted on the background field by the observations by squaring the difference between the analysis and first-guess field at observation locations. Each of these metrics is consistent with the thinning concept that removed data should minimally impact an analysis.

### a. Simple thinning strategies for real observations

The AIRS profiles are thinned using two of the intelligent techniques described in section 2 (mDADT and DBDT) and compared to three simple methods: random (RAN), subsample (SUB), and superob (SUP; DiMego 1988). For consistency with the synthetic tests, each thinning algorithm is tuned to retain approximately 11% of the full AIRS profile dataset. For RAN, observations are randomly selected with the distance between them limited by a user-specified threshold. A distance of 120 km is used here. The RAN approach is attractive in that it tends to produce evenly spaced observations in the presence of irregularly spaced data. To ensure that a representative analysis error is generated, nine realizations are created via the RAN method. The RAN selection methodology follows that of the bootstrap approach in which the length of the data subsets is preserved but the datasets are not necessarily unique (Efron and Tibshirani 1993). In contrast, the SUB approach thins by partitioning the $30 \times 135$ profile swath into groups of 3 cross-track by 3 along-track profiles for a total of 9 profiles per group. From each group, the AIRS profile deemed to be the best, per the quality indicators, is retained under the assumption that these profiles are least likely to contain cloud contamination. If none of the subset profiles have an observation at a given level, it is assigned a missing value. In lieu of selecting the best profile, the SUP method averages the nine profiles from each group and replaces the center sounding with the mean of the group.

### b. Results

#### 1) Observation retention

AIRS temperature and moisture profiles, valid 0730–0748 UTC 12 March 2005, are combined to produce the 700-hPa equivalent potential temperature ($\theta_e$) swath shown in Fig. 8a. The full data swath contains 3076 profiles. Figure 8b is the 700-hPa equivalent potential temperature for the same day but for the 0800 UTC WRF model forecast used as the analysis first guess. Both images indicate the presence of a midlatitude trough over the eastern United States, with an associated 700-hPa baroclinic zone from Manitoba southeastward to Georgia and extending east off of the mid-Atlantic states. A second moisture-related theta-e gradient is located over the southern Gulf of Mexico (GOM). The mDADT gradient and the TFP parameters that correspond to Fig. 8a are shown in Fig. 9. Relatively large values of TFP are found north of the Yucatan, along the northern GOM, around the base of the trough, and southeast of Hudson Bay.
Large values of TFP can also be seen over Illinois. Regions of enhanced gradient include the central GOM, as well as eastern Iowa and Illinois, extending southeast to Tennessee and North Carolina. The region of enhanced gradient associated with the base of the trough is located north of the maximum in TFP, which captures changes in the gradient. Figure 10 depicts the mDADT observation retention for rates of 0.9 for TFP and gradient-based thinning, respectively, for a zero blackout radius. The thinned observation set for this particular parameter selection, which represents an approximate saturation with respect to the mDADT metrics, is shown in order to demonstrate the functionality of the algorithm for real data. The retained observations are consistent with their respective metrics shown in Fig. 9 with relatively large differences for both metrics south and southeast of Hudson Bay. Large differences are also present over the northern GOM, a region with few gradient observations, while the west-central GOM is dominated by gradient observations. As expected, gradient observations are concentrated directly within the theta-e gradient regions.

To optimize the analysis, mDADT parameters are chosen to avoid overfitting the observations. Although the ADAS takes into account variations in data density, irregularly spaced observations can be problematic. The mDADT observations are thinned with an initial blackout radius ~1.47° latitude. Although nadir and off-nadir samples have different spatial resolutions, they are considered identical for the purposes of applying the blackout radius. Analysis results are presented for thinning percentages of 30%/60%/10% and 5%/90%/5% (gradient/TFP/homogeneous). The percentages are not necessarily optimal but were selected, in part, to illustrate algorithm sensitivity to edge effects, data density, and data gaps. The low retention percentage chosen for the homogeneous regions will produce an inferior analysis for the full-domain statistics; hence, results are shown for both the “regions of interest” (boxes in Fig. 11) and the full domain. Within the regions of interest, the full analysis contains approximately 800 observations while the thinned datasets range from 86 to 119 (Table 3). Considering the simple thinning approaches only, within the regions of interest, the RAN technique produces more evenly distributed observations, as it does not directly depend on the spatial variability of the data associated with the satellite viewing angle. In contrast, the SUP and SUB methods preserve the along-swath observation distribution of the native data. Otherwise, neglecting the regions of missing data over the Ohio Valley and the
Yucatan Peninsula, each of the simple techniques produces a fairly uniform observation distribution. The mDADT provides a mix of observation types that are selected near changes in, or within, gradient regions and retains few observations outside the regions of interest. TFP observations straddle the gradient across the central GOM as well as a second enhanced gradient region in the far southern GOM, which extends east to the

FIG. 9. The 700-hPa AIRS $\theta_e$ thinning metrics valid at 0730–0748 UTC 12 Mar 2005 including the (a) local variance (gradient) and (b) TFP. See text for details.

FIG. 10. Retained observations by type for mDADT (zero blackout radius) with retention percentages of 5% homogeneous, and (a) 90% TFP and 5% gradient, and (b) 5% TFP and 90% gradient. Homogeneous, TFP, and gradient observations are represented by the plus, asterisk, and open diamond symbols, respectively (as in Fig. 6).
Yucatan and into the western Caribbean. As expected, the overall data density is higher within the regions of interest, and the mDADT observation selection is consistent with those retained for the saturated cases, that is, 90% TFP and 90% gradient images with no blackout radius (Fig. 10). DBDT reduces the observation clustering observed in eastern Tennessee and over the Yucatan and provides a more even observation distribution in homogeneous regions particularly along the satellite swath edges. Because the initial blackout radius for mDADT is slightly less than that for DBDT, DBDT selects more TFP-based observations in the regions of interest for the 30%/60%/10% case (Figs. 11d and 11e, asterisks).

2) ANALYSES

Within the context of data assimilation, the motivation for producing an automated intelligent data thinning algorithm is to reduce the computation time and reduce the potential harmful observation error correlation. Table 3 presents quantitative results from full and thinned analyses. The total run times, which include both the thinning and analysis, are reduced by as much as 75% from that of the full analysis. With the exception of DBDT, the intelligent algorithms are not more expensive than their simple thinning counterparts. Despite attempts to constrain the overall number of observations for the different thinning methods, the numbers retained vary in the regions of interest. To account for this spread, the RMSE is normalized by multiplying the RMSE by the ratio of the number of observations to the maximum number of observations retained. The SAI is also normalized by dividing by the full-analysis SAI, which represents the maximum analysis impact for the assimilated swath. Hence, the best thinned analyses will have normalized FT errors near one. Results for these error measures are presented in Table 3 for the full domain and two regions of interest. The DBDT algorithm performs the best for all metrics—with the exception of the
regional FT statistic for the 30%/60%/10% thinning percentages. For the simple thinning methods, there is little difference between SUB and SUP while RAN produces improved results for the FT metric only. However, for the entire analysis domain, analyses produced by the SUP algorithm are generally superior to those generated using SUB or RAN (on the order of 0.1°C for the non-normalized DO and FT metrics; not shown). The FT difference fields corresponding to Table 3 are shown in Fig. 12. Only one of nine realizations is shown for RAN (Fig. 12a). The simple thinning approaches (Figs. 12a–c) perform relatively well but do encounter problems, especially with respect to spatial inhomogeneities in the satellite swath. An exception is RAN, which is less prone to data gaps. In particular, the areas of enhanced differences are associated with missing data over south Florida and from Kentucky north to Michigan (Fig. 12a). There are also regions of large differences along the northwest swath edge from Iowa northeast to southern Ontario. The mDADT algorithm (Fig. 12d) struggles with internal and external data boundaries with edge effects evident over Iowa and the northeast U.S. coast into southern Quebec, Canada. These problems are mitigated by the DBDT algorithm, and the resulting analyses (Figs. 12e and 12f) are improved over both the regions of interest as well as the full domain, despite fewer observations (Table 3). The superior performance of the DBDT algorithm reflects a design that attempts to take advantage of the best aspects of both the intelligent and simple algorithms, that is, a quasi-uniform data distribution while simultaneously retaining observations deemed to be relevant.

5. Summary

The goal of this study is to produce a computationally efficient analysis of a quality that falls somewhere between the best possible and that produced by simple thinning. A brute force approach using a one-dimensional problem and all possible observation permutations is mined for the best analysis possible for a given observation error and retention rate. The observations selected depend on the first-guess field as well as characteristics of the assimilated data including observation error, curvature, and gradient features. In particular, there is a tendency to select observations at points in which the gradient changes (so-called anchor points) constituting 50%–80% of the observations retained depending on the background. These results are used to guide the development and tuning of the intelligent algorithms. A two-dimensional problem with known error characteristics was also constructed. In part, the two-dimensional experiments are used to evaluate the retention percentage parameters (TFP, gradient, and homogeneous) within the mDADT algorithm. Results indicate that simple thinning tends to perform better over the relatively uninteresting homogeneous data regions. Information gleaned from the synthetic experiments was used to create a new thinning algorithm designed to take advantage of the mDADT metrics while reducing the impacts of spatial irregularities in the data. This new algorithm (DBDT), its intelligent thinning predecessor, and simple thinning were each applied to combined AIRS temperature and moisture profiles. The observations selected using the DBDT approach produced superior analyses that outperform

<table>
<thead>
<tr>
<th>Method</th>
<th>DO RMSE (°C)</th>
<th>FT RMSE (°C)</th>
<th>SAI</th>
<th>Run time (s)</th>
<th>No. of OBS</th>
<th>L (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full analysis</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
<td>244</td>
<td>793</td>
<td>80</td>
</tr>
<tr>
<td>RAN</td>
<td>0.75</td>
<td>0.32</td>
<td>0.87</td>
<td>56</td>
<td>100</td>
<td>147</td>
</tr>
<tr>
<td>SUB</td>
<td>0.74</td>
<td>0.42</td>
<td>0.88</td>
<td>56</td>
<td>99</td>
<td>146</td>
</tr>
<tr>
<td>SUP</td>
<td>0.74</td>
<td>0.43</td>
<td>0.86</td>
<td>58</td>
<td>100</td>
<td>149</td>
</tr>
<tr>
<td>mDADT 30%/60%/10%</td>
<td>0.85</td>
<td>0.43</td>
<td>0.88</td>
<td>57</td>
<td>108</td>
<td>152</td>
</tr>
<tr>
<td>mDADT 5%/90%/5%</td>
<td>1.48</td>
<td>1.00</td>
<td>0.83</td>
<td>57</td>
<td>86</td>
<td>152</td>
</tr>
<tr>
<td>DBDT 30%/60%/10%</td>
<td>0.67</td>
<td>0.33</td>
<td>0.90</td>
<td>105</td>
<td>88</td>
<td>152</td>
</tr>
<tr>
<td>DBDT 5%/90%/5%</td>
<td>0.62</td>
<td>0.25</td>
<td>0.92</td>
<td>106</td>
<td>87</td>
<td>152</td>
</tr>
<tr>
<td>RAN</td>
<td>1.30</td>
<td>0.56</td>
<td>0.82</td>
<td></td>
<td>269</td>
<td></td>
</tr>
<tr>
<td>SUB</td>
<td>1.30</td>
<td>0.60</td>
<td>0.84</td>
<td></td>
<td>259</td>
<td></td>
</tr>
<tr>
<td>SUP</td>
<td>1.19</td>
<td>0.53</td>
<td>0.84</td>
<td></td>
<td>258</td>
<td></td>
</tr>
<tr>
<td>mDADT 30%/60%/10%</td>
<td>1.22</td>
<td>0.65</td>
<td>0.94</td>
<td></td>
<td>271</td>
<td></td>
</tr>
<tr>
<td>mDADT 5%/90%/5%</td>
<td>1.18</td>
<td>0.55</td>
<td>0.93</td>
<td></td>
<td>258</td>
<td></td>
</tr>
<tr>
<td>DBDT 30%/60%/10%</td>
<td>1.16</td>
<td>0.48</td>
<td>0.95</td>
<td></td>
<td>255</td>
<td></td>
</tr>
<tr>
<td>DBDT 5%/90%/5%</td>
<td>1.11</td>
<td>0.36</td>
<td>0.98</td>
<td></td>
<td>256</td>
<td></td>
</tr>
</tbody>
</table>

JUNE 2010 L A Z A R U S E T A L . 849
standard operational thinning methodologies. DBDT was shown to reduce observation clustering and provides a more even observation distribution along the satellite swath edges. DBDT, which combines adaptive and non-adaptive thinning strategies, appears to be fairly robust to parameter selection. Furthermore, the DBDT algorithm is computationally efficient, which makes its application plausible within an operational framework. Given the degrees of freedom inherent in the more sophisticated algorithms, the results presented do not necessarily reflect the best analyses possible but are instead intended to demonstrate their utility for operational applications. For example, it might be possible to apply a gradient detection algorithm such as the DBDT to determine the location of clouds for radiance assimilation. Assuming that a gradient in brightness temperatures can delineate cloud edges, such an algorithm could be used to better define clear scenes in the observations.

Acknowledgments. This research was supported by funding under NASA Grant NNG06GG18A.
REFERENCES


Tellus, 61A, 144–153.