2B.3

ON THINNING METHODS FOR DATA ASSIMILATION OF SATELLITE OBSERVATIONS

T. Ochotta[1*] C. Gebhardt[2] V. Bondarenko[1] D. Saupe[1] W. Wergen[2]

[1] *University of Konstanz, Germany*
[2] *Deutscher Wetterdienst, Offenbach, Germany*

ABSTRACT

Thinning of observational data sets is an essential task in assimilation of satellite data for numerical weather forecast. In this work we modify and improve the scheme of so-called estimation error analysis (EEA). EEA is an adaptive thinning method that iteratively removes observations from a given data set, guided by a special approximation error measure evaluated at all original observation points. We propose EEA variants that differ in methodological and performance aspects, such as the Grid-EEA method, where errors are evaluated on a regular grid on the globe. Moreover, in the Top-Down EEA, we propose to construct the thinnings by an iterative point insertion strategy, which leads to improved performance since the number of insertion steps is typically much smaller than the number of corresponding removal operations in EEA. We also provide an efficient implementation of the proposed methods yielding a significant acceleration of the standard EEA approach.

1. INTRODUCTION

Data assimilation combines observational data with a background model to produce initial states of the atmosphere for numerical weather forecasts. Current and future satellite instruments produce large amounts of measurements that are integrated into prediction systems operating in periodic time intervals, e.g., every 3 hours. These data sets show very different characteristics with respect to data volume as well as spatial and temporal density, which demands for a very careful preprocessing of the data.

Concerning data density there are two major aspects to be considered when integrating the observations into operational data assimilation schemes. Firstly, a high data density leads to high computational costs and to occupation of much disk space. Secondly, a high data density may violate the assumption of independent measurement errors, which is made in most practical assimilation schemes. The error correlations are not known in advance and implementing estimations of these correlations would lead to a significant increase in computational complexity. Moreover, in the work of Liu and Rabier (2002) it was shown that there is a connection between the observation density and the resolution of the model grid. Their theoretical study shows that the analysis quality decreases, if the density of the observational data set is too large and error correlations are neglected as it is the case in many operational schemes. These restrictions motivate thinning methods, which have the goal to reduce the large satellite data sets, to reduce spatial error correlations between the observations, and to retrieve the essential information content of the data for optimal use in data assimilation.

In this work we revisit an adaptive observations thinning scheme that was recently proposed in Ochotta et al. (2005). The estimation concept approximates measurement values on the sphere by a continuous estimation function, where the value at any position on the sphere is given by a weighted average of the observation values in a local neighborhood. Considering the full observational data set and a subset, this approach allows for defining an approximation error measure by considering differences between the two corresponding estimation functions. The estimation error analysis (EEA) method constructs a thinned data set by iteratively removing observations from the full data set, such that at each step, the degradation in global estimation error is minimal. The differences in the estimation function are thereby evaluated at the positions of the original observations. The procedure is terminated when a desired number of retained observations is reached.

In this work we extend the EEA concept by introducing variants, which differ in methodological and performance aspects. In the first proposed variant (Grid-EEA), we employ a regular grid on the globe

*Corresponding author address: Tilo Ochotta, University of Konstanz, Fachbereich Informatik & Informationswissenschaft, Fach M697, 78457 Konstanz, Germany; e-mail: ochotta@inf.uni-konstanz.de

to evaluate the estimation function. This approach is motivated by the fact that in most cases, the observations are distributed non-uniformly. The traditional EEA is based on evaluation of errors at the positions of the observations of the original data set, which consequently produces a bias of the error function due to uneven sampling densities. Depending on the resulting grid resolution, the Grid-EEA algorithm provides thinnings of varying accuracy at different run times.

While the EEA algorithm adopts the concept of iterative point removal to obtain a thin data set, we propose another EEA variant, in which the observations are iteratively inserted starting with the empty set. This approach is preferable in settings where only a small part of the original data set is retained, e.g., 10%, as it is the case with the satellite soundings in the assimilation process at the Deutscher Wetterdienst (DWD). For these thinnings, the resulting number of insertion operations is consequently much smaller than the number of observation removal steps in the traditional EEA method, which makes the method faster and more accurate.

We finally focus on the problem of an efficient implementation of the methods. The introduced variants rely on observation removal or insertion operations. We propose to organize all observations that are candidates for removal or insertion operations in efficient data structures, which allow sorting the observations according to their redundancy degree. We present two variants of processing these data structures during the thinning process, leading to different behavior in accuracy and computational complexity.

The paper is organized as follows. In the next section we state the problem of data thinning and revisit related work. In section 3 we describe the proposed EEA variants, followed by a presentation of experimental results in section 4. We conclude our work and suggest directions for future improvements in section 5.

2. PROBLEM STATEMENT AND RELATED WORK

In this section, we follow the notations in Ochotta et al. (2005). Our input data set consists of ATOVS (Advanced TIROS Operational Vertical Sounder) satellite data, in particular differences of multichannel brightness temperature between bias-corrected observations and their first guess. We consider the full observation set P_0 that holds the positions of n measurements as points in three-dimensional space $\mathbb{R}^3$. More specifically, we transform geographic coordinates (λ, ϕ) to cartesian coordinates $(x,y,z)^T$. For each observed point, an observation value is given, i.e. a k-dimensional vector that holds the measured multichannel brightness temperature differences. We define a function $f : P_0 \to \mathbb{R}^k$ that gives the observation value for an observation position $p \in P_0$.

Given the full data set P_0, the goal of data thinning is to find a subset $P_i \subset P_0$ with $|P_i| = n - i$, $i = 1, \ldots, n-1$, that approximates P_0 well. The approximation quality can be measured by a real-valued function $E : 2^P \times 2^P \to \mathbb{R}^+ \cup \{0\}$, where 2^P denotes the set of subsets of P. The *optimal thinning* is then given by the set $P_i \subset P_0$, satisfying $E(P_i, P_0) \le E(P_i', P_0)$ for all subsets $P_i' \subset P_0$ with $|P_i'| = |P_i|$. This is a hard optimization problem, since the well-known NP-hard Rucksack problem can be seen to be a special case of this optimization. Therefore, we cannot expect to find an algorithm for computing the optimal thinning in polynomial time. Instead heuristic methods can be used, e.g., iteratively removing points from the full data set until a desired data density is reached.

The most common technique for reducing the observational data set is given by uniform thinning. In most cases, the measurements that are obtained by satellite instruments are arranged in a grid-like structure on the globe, which suggests to establish thinning by selecting every n-th point in zonal and meridional direction. Although this uniform approach yields robust data reduction, it is not aware of the variances of observational values in the data set. The goal is to find adaptive thinnings that retain a higher density in regions with large gradients or other significant structures in the input signal.

One approach to establish this behavior was proposed in Ramachandran et al. (2005). For a given observational data set, their methods produces a spacial octree, in which each node corresponds to a rectangular cell in 3D-space. The observations therein are represented by a single reference point, given by the cell center. The subdivison of each octree cell is ruled by a statistical test, which models the corresponding approximation quality.

2.1. *Estimation Error Analysis (EEA)*

In Ochotta et al. (2005) it was shown that high-quality thinnings of observational data sets can be obtained using the concept of iterative point removal. The invented EEA method is based on an estimation filter, which provides a continuous function on the sphere for approximating the measurement values. An estimation of the observational value for a given location with respect to a given

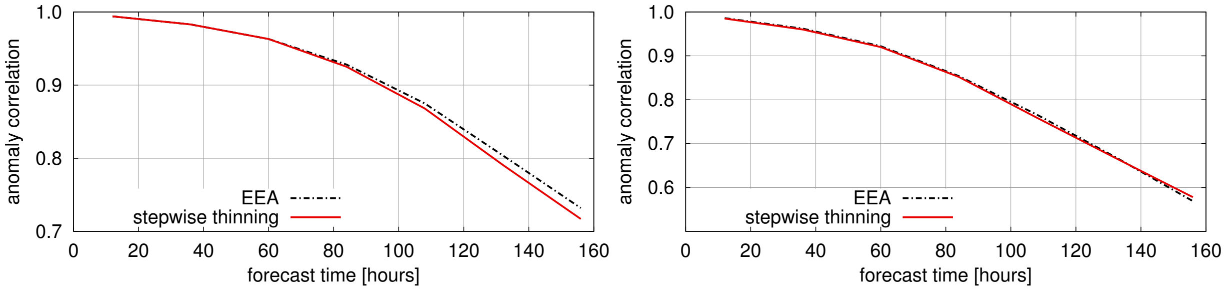

Figure 1: Anomaly correlations of 500 hPa geopotential based on 27 subsequent 156h-forecasts started at 12 UTC on Dec. 27, 2004 for the northern hemisphere (left) and the southern hemisphere (right).

data set P_i is computed by the normalized weighted sum of observations in P_i,

$$\tilde{f}_{P_i}(x) = \frac{\sum_{p \in P_i} f(p) \cdot w_h(\|x-p\|)}{\sum_{p \in P_i} w_h(\|x-p\|)}, \qquad (1)$$

where $w_h(s) = e^{-s^2/h^2}$ is a positive, exponentially decreasing weighting function, which assigns larger weights to points near x. The parameter h defines the spatial scale of w_h. Given the input set P_0, the function $\tilde{f}_{P_0}$ serves as a reference for the approximations $\tilde{f}_{P_i}$ of the thinned observation set P_i.

The mean squared error of a thinning $P_i \subset P_0$ is defined by averaging the squared differences of the estimation function with respect to the full data set P_0 and the simplified set P_i. These differences are evaluated at the positions in P_0,

$$E_{mse}(P_i) = \frac{1}{|P_0|} \sum_{p \in P_0} \|\tilde{f}_{P_0}(p) - \tilde{f}_{P_i}(p)\|^2. \qquad (2)$$

The EEA method aims at finding a minimum of the mean squared error. The estimation function is used to define the redundancy degree for each observation in P_0. If replacing an observation leads to a sufficiently small change in the estimation function, the observation is supposed to be redundant with respect to its neighboring observations. Following this concept, the EEA method removes for $i = 0, 1, 2, \ldots, n-1$, the point $p_{j(i)} \in P_i$ of the thinning P_i, given by

$$p_{j(i)} = \arg\min_{p \in P_i} E_{mse}(P_i \setminus \{p\}). \qquad (3)$$

The resulting data set is given by

$$P_{i+1} := P_i \setminus \{p_{j(i)}\}. \qquad (4)$$

Applying this procedure yields a sequence of nested subsets P_i of P_0,

$$P_0 \supset P_1 \supset P_2 \supset \cdots \supset P_n = \emptyset,$$

each of which provides an approximation of the original observation set with $n-i$ retained observations. Throughout this paper we refer to this method as the *traditional variant of the EEA* or short EEA, as in Ochotta et al. (2005).

Figure 1 shows the mean anomaly correlations for the 500 hPa geopotential depending on forecast time for the EEA method in comparison to the reference method of *stepwise thinning*, which corresponds to retaining every third observation in zonal and meridional direction of the original data set. While the EEA algorithm shows slightly higher anomaly correlations for long term forecasts compared to the non-adaptive stepwise thinning for the northern hemisphere, no comparable behavior is observed for the southern hemisphere.

3. PROPOSED EEA VARIANTS

3.1. *EEA on a Regular Grid*

The EEA thinning algorithm iteratively selects the observation with the least degradation on the global estimation function $\tilde{f}$. The function $\tilde{f}$ is thereby evaluated at the positions of the observations in the full data set P_0. Since in most cases, the observations are distributed *non-uniformly*, this leads to a bias in the estimation error due to the uneven sampling densities in P_0.

In the first variant of the EEA algorithm, we propose to employ a regular grid on the globe to evaluate the estimation function $\tilde{f}$. We apply the grid of the GME that is used for the model state vectors at the DWD. The grid of the GME is defined by recursive subdivisions of the 20-sided icosahedron. The subdivision of each triangle is performed by halfing the edges, which results in four new triangles, whereby each vertex in the grid is projected on the sphere.

The constructed grid contains $10n_i^2 + 2$ vertices, where n_i is the number of intervals on one side of the icosahedron. The number n_i also corresponds to the number of subdivision steps l, $n_i = 2^l$. The grid of the GME provides a near-uniform discretiza-

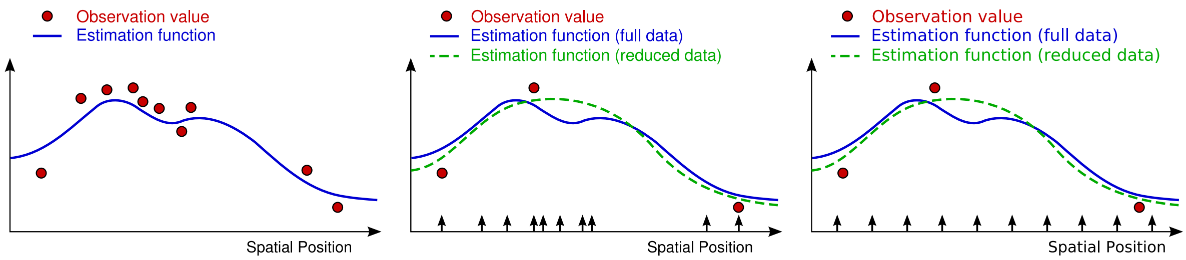

Figure 2: Concept of estimation error analysis (EEA); a smooth curve is associated to a given observations set (left); thinning based on evaluation of the estimation function leads to a bias of the error due to uneven sampling (middle); we propose to use a regular grid instead to overcome this problem (right).

tion of the sphere, although it contains triangles fans of varying size and shape. For details, see Majewski et al. (2002).

Let G_l be the set of vertices of a given GME grid of subdivision level l. According to Eq. (2) we use the estimation error $E_{mse}(P_i, G_l)$, which is based on evaluating squared differences between two estimation functions on the grid of the GME with a specified subdivision level l,

$$E_{mse}(P_i, G_l) = \frac{1}{|G_l|} \sum_{q \in G_l} \|\tilde{f}_{P_0}(q) - \tilde{f}_{P_i}(q)\|^2. \quad (5)$$

The selection of the observation $p_{j(i)}$ to be removed in each step $i = 0, 1, 2, \ldots, n-1$, (where n again denotes the total number of observations in the full data set P_0) is then defined as

$$p_{j(i)} = \arg\min_{p \in P_i} E_{mse}(P_i \setminus \{p\}, G_l). \quad (6)$$

The regular grid G_l facilitates a near uniform sampling of the estimation function $\tilde{f}_{P_i}$, see Fig. 2, and hence, allows for a more accurate evaluation of the difference between two estimation functions. The parameter l defines the resolution of the grid and can either be adjusted by the user or in coherence with the model resolution that is used in the assimilation.

3.2. *Top-Down EEA*

The EEA as well as the Grid-EEA are greedy thinning methods that employ the point removal strategy. As denoted in section 2, the greedy property is given by the fact that they work iteratively by evaluating a number of potential candidates and selecting the best one, e.g., the candidate with the least increase in total error. This concept leads to *locally optimal* solutions, meaning that the thinning P_i is found as the best set $P_i \subset P_{i-1}$. The goal, however, is to find the *globally optimal* thinning $P_i \subset P_0$ with $E_{mse}(P_i) \leq E_{mse}(P_i')$ for all $P_i' \subset P_0$, $|P_i'| = |P_i|$. This global optimality is not guaranteed for the thinnings in this work that uses the greedy strategy.

As a consequence, the iterative point removal method may select non-optimal points leading the algorithm away from the globally optimal solution. Considering the data sets in this work, the density of the thinned observation sets is at roughly 10% of the density of the full data set. Taking this into account, we propose thinning by top-down estimation analysis (Top-Down EEA), in which observations are iteratively added starting with the empty set. In line with the findings in subsection 3.1, we build the Top-Down EEA upon a regular grid G_l. Considering Eq. (3) and (4), the observation $p_{j(i)}$ at step $i = n-1, n-2, \ldots, 1$ is selected as follows:

$$p_{j(i)} = \arg\min_{p \in P_0 \setminus P_i} E_{mse}(P_i \cup \{p\}, G_l),$$

yielding the observations set

$$P_{i-1} := P_i \cup \{p_{j(i)}\}.$$

The advantage of this approach is that the number of point insertion operations for sparse thinnings is much smaller than the number of point removal steps in the traditional EEA (see subsection 2.1). This leads to an acceleration of the thinning procedure as well as to an improvement in approximation quality due to reduced error accumulation.

3.3. *Implementation and Acceleration*

The computation of an estimation value $\tilde{f}_{P_i}(x)$ (1) requires for the summation of many (weighted) contributions. Since the weighting function $w_h(s)$ is exponentially decreasing, we restrict the summation in (1) to observations in P_i within a local neighborhood of radius r around x. This yields for each observation $p \in P_i$, a region of influence where estimation terms $\tilde{f}_{P_i}(x)$ are dependent on the observation value $f(p)$. It follows that the removal of this observation only affects $\tilde{f}_{P_i}(x)$, if and only if $\|x - p\| \leq r$, using the euclidean norm.

This approach may lead to special cases, where no observations are available in the prescribed r-neighborhood, e.g., for very sparse data sets. We then degenerate the estimation filter to a nearest neighbor interpolation. More precisely, for these samples, we grow r until one observation is within the r-neighborhood and set the estimation value to the corresponding observation value.

We utilize the r-neighborhoods to construct a graph to hold depencies between grid points, for which the estimation function is evaluated, and observations that are removal candidates. The increase in total error when removing observation $p_{j(i)} \in P_i$ (6) can easily be evaluated by computing the estimation function for a relatively small number of grid samples with and without involving $p_{j(i)}$.

The point removal procedure is implemented using a priority queue, which is a data structure to efficiently find elements with the highest associated priority across a sequence of operations. In our case, the elements correspond to observations, the priority value of each is given by the inverse of the associated error increment. The elements are sorted in descending order, starting with the element with the highest priority. The removal of an observation then corresponds to the removal of the first element in the priority queue, followed by an updating procedure, in which error increments of affected observations are recomputed. In particular, the removal of an observation $p \in P_i$ requires for recomputation of the error increments of observations $s \in P_i$, if and only if there is a grid point $q \in G_l$ with $\|p-q\| \leq r$ and $\|s-q\| \leq r$.

Lazy Evaluation The benefits of using the priority queue are reduced computational costs, since only the error increments of affected observations need to be recomputed for each removal step, while the increments of all other observations remain untouched. However, depending on the parameter r, the number of required recomputations per observation removal step can become large, e.g., up to 100 for the data sets we use. The resulting computational costs may consequently slow down the thinning procedure.

To accelerate the process we propose to establish the so-called lazy evaluation scheme, in which the updating procedure is slightly modified in order save computational costs. In contrast to the full updating procedure, no error increments are updated after removing an observation. Instead, there is a validity test when an observation is to be removed. More specifically, considering $p \in P_i$ the observation that is the candidate for the next removal operation, we check if all dependent observations are still in P_i, e.g., have not yet been removed. If this is the case, the error increment of p is still valid and p can be removed. Otherwise, the error increment has to be recomputed and the corresponding element is reinserted into the priority queue.

The benefit of the lazy evaluation is a reduced number of evaluated error increments at each point insertion or removal step, leading to reduced computational costs. The drawback is a slight loss in accuracy that follows from the fact that the search space of possible observation candidates is smaller than for the full updating procedure, which consequently produces a slightly different sequence of processed observations during the thinning.

4. RESULTS AND DISCUSSION

We present experimental results for an ATOVS satellite data set containing 27367 observations with observation values as differences between bias-corrected measured brightness temperatures and first guess in eight channels. Measurements are considered over sea as implemented in the experimental analysis and forecast system of the DWD. As reference method we implement the *stepwise thinning* scheme, which corresponds to retaining every third point in zonal and meridional direction in the full data set P_0. The stepwise thinning is used in the operational service of the DWD.

The proposed algorithms in this paper can operate as individual components of the EEA. We combine them with each other and consider the following EEA variants:

traditional EEA; Iterative point removal and evaluating $\tilde{f}$ at the positions of the observations in P_0 (as in Ochotta et al. (2005));

Top-Down EEA; Iterative point insertion and evaluating $\tilde{f}$ at the positions of the observations in P_0;

Grid EEA; Iterative point removal and evaluating $\tilde{f}$ on the regular grid G_l;

Top-Down Grid EEA; Iterative point insertion and evaluating $\tilde{f}$ on the regular grid G_l;

Moreover, these four variants can be carried out using either the full updating procedure or the lazy evaluation approach.

The approximation quality of the methods is compared in terms of the grid-based mean-squared estimation error E_{mse}, Eq. (5), for which we use a grid with a dense sampling (G_8) to obtain an accurate error evaluation. Note that the grid resolution may be differently used for the thinning process and for the computation of the approximation quality (5), e.g.,

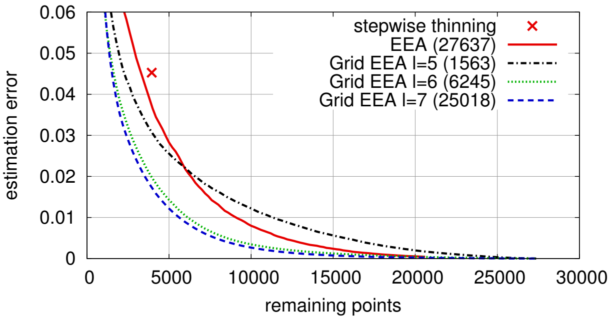

Figure 3: Comparison between the traditional EEA and the grid-variant based on the error $E_{mse}(P_i, G_8)$ (5); the grid-based thinning is carried out using grid of varying resolutions l; in brackets the number of grid samples that are considered during the thinning.

we can choose a smaller l for the thinning for acceleration purposes.

Figure 3 shows the resulting error $E_{mse}(P_i, G_8)$ for various grid resolutions. Considering the number of original observations, applying a grid with reduced resolution, e.g., $l = 6$ with 6245 samples, already leads to satisfying approximation quality. Increasing the grid resolution marginally improves the performance, i.e $l = 7$.

Results of applying the top-down variant are shown in Fig. 4. We consider two scenarios, firstly, applying the top-down strategy to the traditional EEA in Ochotta et al. (2005), where we observe a significant improvement in approximation quality (Fig. 4(top)), and secondly, using the regular grid G_6 as proposed in this paper, where a smaller quality gain for thinnings with a small number of retained observations is obtained. We found that this behavior is typical, meaning that there is an evidence that the grid variant of the EEA approach based on point removal provides good data approximation even for strong thinnings. Note that in Fig. 4(top), we used the error measure based on evaluation of the differences in estimation function at the positions of the original points in P_0, see Eq. (2).

Table 1 shows timings for various thinnings achieved with the proposed methods. We can basically observe two effects. Firstly, for the top-down approach, the run time is proportional to the number of retained observations, while the variants based the point removal strategy show inverse proportionality. Secondly, the regular grids can be leveraged to accelerate the thinning process by using smaller resolutions l, e.g., the number of samples for evaluation of $\tilde{f}$ on the grid G_6 is much smaller (6245 samples) than for the traditional EEA approach (27637 samples). Lastly, the lazy evaluation technique significantly reduces the run times, e.g., in order to construct a thinning with 3000 retained observations, the Grid EEA method needs 63 seconds using the full updating procedure, while the lazy evaluation scheme needs only 0.78 seconds.

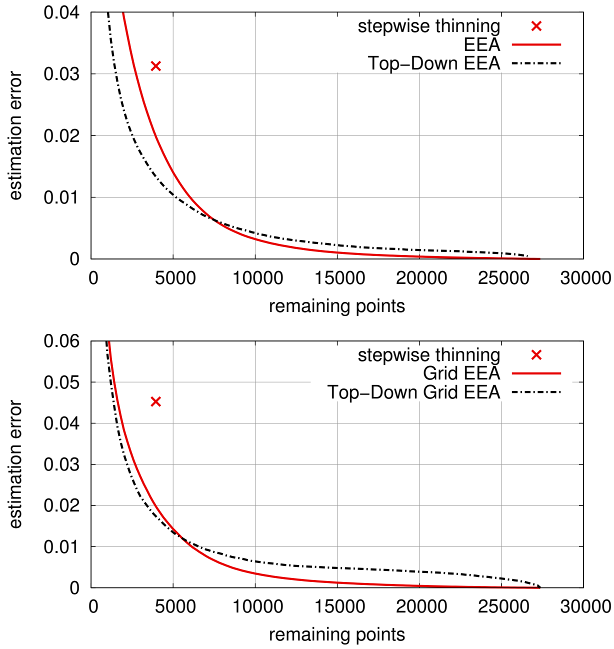

Figure 4: The top-down approach improves the approximation quality of the traditional EEA (top); applying the regular grid reduces the quality gain (bottom).

Focusing on the approximation quality of applying the lazy evaluation method, we observe that the quality loss is negligible for the point removal strategy, Fig. 5(top). The top-down approach, however, leads to a noticable increase in approximation error for thinnings that retain more than 15% of the original observations. The acceleration technique in combination with the top-down approach is therefore only suitable, if a small number of observations if desired.

5. CONCLUSIONS AND FUTURE WORK

We presented methods for thinning of observational data sets that are delivered by current satellite instruments. We build our methods upon the previously proposed estimation error analysis scheme, in which observations were iteratively removed according to a redundancy measure. We proposed variants of this thinning approach by applying a uniform sampling strategy for the evaluation of the corresponding estimation function. We furthermore extended the technique to applying an iterative point

EEA $n-i=$	run time [seconds] full updating 1000	2000	3000	5000	15000	25000
traditional	120	120	110	98	93	25
Top-Down	13	24	36	58	140	180
Grid	81	69	63	57	**45**	**11**
Grid Top-Down	**4.9**	**9.6**	**14**	**23**	55	68

EEA $n-i=$	run time [seconds] lazy evaluation 1000	2000	3000	5000	15000	25000
traditional	4.6	4.4	4.3	4.0	2.4	0.7
Top-Down	3.4	3.9	4.1	4.5	5.8	7.1
Grid	0.87	0.83	0.78	0.74	**0.44**	**0.14**
Grid Top-Down	**0.47**	**0.56**	**0.62**	**0.72**	0.98	1.2

Table 1: Timings in seconds for the proposed EEA methods achieved by using the full updating procedure for priority queue maintenance (top); applying the lazy evaluation strategy leads to a drastic acceleration (bottom); the bold timings indicate the fastest method for the corresponding thinning.

insertion strategy instead of removing observations from the full data set, and moreover, we discussed an acceleration technique for time-efficient processing of very large data sets. Results show that our modifications lead to an improved overall approximation quality, as well as a drastic acceleration of the estimation error approach.

For a practical application, we recommend to use the estimation error analysis scheme based on regular grids. Depending on the input data density and the desired number of retained observations, the top-down approach using the full updating procedure has shown the best approximation quality at tolerable computational costs. Future satellites are expected to deliver data sets with a much higher density, and moreover, there is a demand on further decreasing the desired number of retained observations. For these scenarios, the lazy evaluation strategy might become more meaningful to accelerate the data reduction.

In future work we plan to analyze the performance of the proposed methods with respect to long term forecasts. Moreover, thinning techniques can be discussed that deeply analyze the influence of observations and their distribution within the assimilation process.

ACKNOWLEDGMENTS

We thank Reinhold Hess, Gerhard Paul, and Andreas Rodin for helpful discussions throughout this project. This work was supported by the DFG Graduiertenkolleg 1042 'Explorative Analysis and Visualization of Large Information Spaces'.

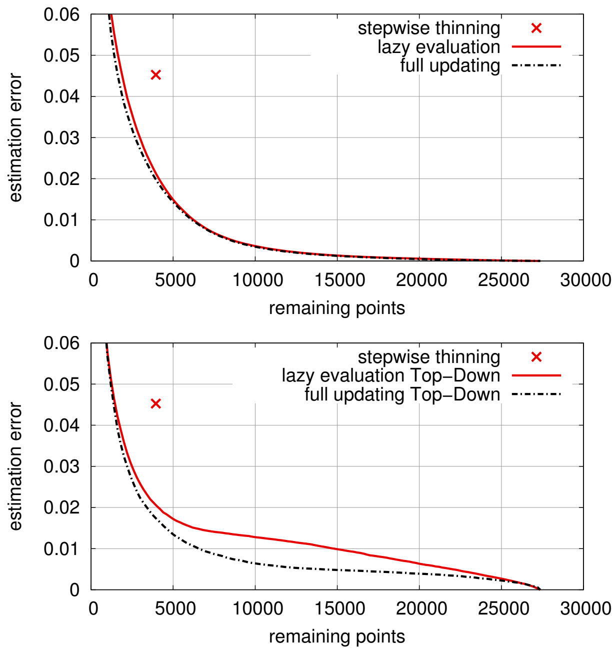

Figure 5: The performance impact of the lazy evaluation strategy is negligible for the Grid EEA using the point removal strategy (top), while there is a significant quality loss using the point insertion approach for thinnings with more than 15% retained observations (bottom).

References

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

Majewski, D., D. Liermann, P. Prohl, B. Ritter, M. Buchhold, T. Hanisch, and G. Paul, 2002: The operational global icosahedral-hexagonal grid point model GME – operational version and high resolution tests. *Monthly Weather Review*, **130**, 319–338.

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

Ramachandran, R., X. Li, S. Movva, S. Graves, S. Greco, D. Emmitt, J. Terry, and R. Atlas, 2005: Intelligent data thinning algorithm for earth system numerical model research and application. *Proc. 21st Intl. Conf. on IIPS*.