Adaptive thinning of atmospheric observations in data assimilation with vector quantization and filtering methods

By T. OCHOTTA[1]*, C. GEBHARDT[2], D. SAUPE[1] and W. WERGEN[2]

[1] *University of Konstanz, Germany*

[2] *Deutscher Wetterdienst, Offenbach, Germany*

(Received 23 May 2005; revised 5 October 2005)

SUMMARY

In data assimilation for numerical weather prediction, measurements of various observation systems are combined with background data to define initial states for the forecasts. Current and future observation systems, in particular satellite instruments, produce large numbers of measurements with high spatial and temporal density. Such datasets significantly increase the computational costs of the assimilation and, moreover, can violate the assumption of spatially independent observation errors. To ameliorate these problems, we propose two greedy thinning algorithms, which reduce the number of assimilated observations while retaining the essential information content of the data. In the first method, the number of points in the output set is increased iteratively. We use a clustering method with a distance metric that combines spatial distance with difference in observation values. In a second scheme, we iteratively estimate the redundancy of the current observation set and remove the most redundant data points. We evaluate the proposed methods with respect to a geometric error measure and compare them with a uniform sampling scheme. We obtain good representations of the original data with thinnings retaining only a small portion of observations. We also evaluate our thinnings of ATOVS satellite data using the assimilation system of the Deutscher Wetterdienst. Impact of the thinning on the analysed fields and on the subsequent forecasts is discussed.

KEYWORDS: Clustering Numerical weather prediction Redundant satellite data

1. INTRODUCTION

An important task in data assimilation for numerical weather prediction (NWP) is the effective exploitation of large amounts of data produced by current and future observation systems, in particular satellite instruments. The high spatial and temporal density of satellite data is potentially highly valuable for estimating an initial state in the numerical forecast process. However, the theoretical framework of commonly used assimilation schemes and the constraints of operational NWP require a careful preprocessing of the data. A larger number of assimilated observations increases the computational costs, occupies more disk space, and leads to more time-consuming transmission of data. Moreover, a high spatial and/or temporal data density can severely violate the assumption of independent observation errors made in most operational and experimental assimilation schemes (variational, Kalman filter, optimal interpolation). The error correlations are unknown a priori, but even taking into account estimations of these correlations in the assimilation system would require more complex observation error statistics leading to an additional increase in computational costs.

Therefore, we need efficient methods that reduce the amount of data and extract the essential information content. The quality of analyses and forecasts should be preserved or even be improved. The most intuitive and most commonly used approach of data reduction is a thinning by selecting observations within predefined regions (e.g. one observation per 75 km × 75 km box) or at specified intervals (e.g. every third observation). Liu and Rabier (2002) investigated the theoretical properties of such a non-adaptive approach on the assumption of a known true state of the atmosphere and determined an optimal relation between model grid resolution and observation density, which minimizes the effect of correlated observation errors. While such a theoretical

* Corresponding author: Department of Computer and Information Science, University of Konstanz, 78457 Konstanz, Germany. e-mail: tilo.ochotta@uni-konstanz.de

consideration can provide a very good benchmark for a suitable mean observation density, thinning algorithms may be improved further by adapting to the sampling density of the local atmospheric situation.

An approach to introduce adaptiveness to data thinning is presented by Ramachandran *et al.* (2005). The authors use a quadtree structure to partition the dataset recursively. The original observations are then approximated by the set of remaining cell centroids. The drawback of the quadtree approach is that the cell borders are fixed in zonal and meridional directions due to the given subdivision order, which may lead to many cells with only few original data points.

We propose two adaptive thinning algorithms using simplification methods from geometry processing in computer graphics and by clustering algorithms. Both algorithms produce reduced datasets as approximations of the original data, but they differ in methodical aspects and in the criteria for selecting the observations to be retained in the simplification. The first approach ('top-down clustering') is a thinning by iterative point insertion, which builds clusters of observations that are similar in spatial position and value according to a simple metric. The simplified dataset consists of the centres of the clusters. In contrast to the work of Ramachandran *et. al* (2005), our clustering provides approximations in which the cells adapt to structures in the signal, and thus yield better approximations than using fixed straight lines. The second approach ('estimation error analysis') identifies and iteratively removes the most redundant observations. The degree of redundancy of an observation is modelled to be inversely proportional to the estimation error of its reconstruction by an estimation filter applied to all retained observations in its neighbourhood.

We apply the two methods to Advanced TIROS Operational Vertical Scanner (ATOVS) satellite data, which are processed in a one-dimensional variational (1D-Var) assimilation scheme to retrieve profiles of atmospheric temperature and humidity. These profiles are input data for the optimal interpolation at the German Weather Service (DWD).

This paper is organized as follows. In the next section we discuss the formal framework for simplification of a given observation set by iterative point removal and iterative point insertion. We then introduce the thinning methods that we have implemented in detail, namely top-down clustering and estimation error analysis. In sections 3 and 4 we provide implementation details and present experimental results. Finally, we summarize our findings and address future work.

2. Proposed thinning algorithms

Given a dataset, P_0, of n observations our goal is to extract m points ($m \ll n$) that yield a good approximation of the original data. We consider two basic concepts of simplification, namely, iterative point insertion and iterative point removal. The first approach starts with the empty dataset and identifies and iteratively inserts available points such that the resulting intermediate point sets yield the best possible approximations of the complete set of observations, using a suitable error metric. The latter approach is given by iterative point removal, where we start with the full dataset, P_0, and iteratively remove the observation which is most redundant with respect to some error measure.

The methods that we discuss in this paper work in an adaptive manner in the sense that the error metric not only considers the spatial distribution of the data when removing or inserting observations, but also the observation values. In the end we obtain thinnings with higher data density in regions in which the variance of the original signal is large.

In this section, we discuss the two thinning algorithms that we have devised and implemented for thinning of ATOVS satellite data, in particular differences of bias-corrected multichannel brightness temperature between observation and first guess. We consider the full observation set P_0 that holds the positions of measurements as points in 3D space $\mathbb{R}^3$. More precisely, we transform geographic coordinates (λ, ϕ) to Cartesian coordinates $(x, y, z)^{\mathrm{T}}$, yielding the unit sphere $\mathbb{S}^2$ embedded in $\mathbb{R}^3$. For each observed point, an observation value is given, i.e. a k-dimensional vector that holds the measured multichannel brightness temperature differences. We formally define a function $f : P \rightarrow \mathbb{R}^k$ that gives the observation value for an observation position $p \in P$.

(*a*) *Top-down clustering*

The idea behind top-down clustering is to group observations with similar spatial positions and measurement values into clusters which are approximated by one representative measurement, i.e. the mean of the cluster (closest point to all cluster elements). A similar approach is widely used in vector quantization for lossy image compression (Gersho and Gray 1992).

For our purpose, we develop a clustering method that works in two phases, namely, cluster splitting and point relaxation. We start by approximating the full dataset P_0 by the cluster mean with respect to some distance measure. More specifically, we consider a cluster $\mathcal{C}$ with cluster elements $p \in \mathcal{C}$, $p = (x, y, z)^{\mathrm{T}}$ that groups the observations at the positions p. Given a cluster $\mathcal{C}$ we define its cluster centroid $\overline{p} = |\mathcal{C}|^{-1} \sum_{p \in \mathcal{C}} p$ and the mean observation value $\overline{v} = |\mathcal{C}|^{-1} \sum_{p \in \mathcal{C}} f(p)$. We now define a distance metric for cluster elements that combines spatial distance of measurement positions with distance of measurement values $d_f : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ with

$$d_f(p, q) = (\|p - q\|^2 + \alpha^2 \|f(p) - f(q)\|^2)^{\frac{1}{2}}, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean metric. The factor α scales the domain of measurement values relative to point coordinates. The metric $d_f(p, q)$ simultaneously takes into account the distances in space and measured value between two observations using a suitably chosen scaling factor. A similar approach has been proposed by Riishøjgaard (1998) when specifying flow-dependent background error correlations. The cluster mean is then defined as observation $\widehat{p}$ that minimizes the sum of squared distances to all cluster elements $q \in \mathcal{C}$:

$$e(\mathcal{C}, p) = \sum_{q \in \mathcal{C}} d_f(p, q)^2, \quad \widehat{p} = \arg \min_{p \in \mathcal{C}} e(\mathcal{C}, p). \tag{2}$$

We define $e(\mathcal{C}) := e(\mathcal{C}, \widehat{p})$ as the cluster error that provides a measure for the approximation quality of $\mathcal{C}$.

We start by setting $\mathcal{C}_0 := P_0$ and $\mathcal{U} := \{\mathcal{C}_0\}$, i.e. all observations are in one cluster. In the splitting phase we subdivide any cluster $\mathcal{C} \in \mathcal{U}$ with an error $e(\mathcal{C})$ that is larger than a given threshold $t > 0$. For this purpose we use principal component analysis (PCA) in order to split $\mathcal{C}$ across its major principal axis through the cluster centroid (Fig. 1). For cluster $\mathcal{C}_0 = P_0$, we compute the 3×3 covariance matrix

$$\mathbf{S} = (p_1 - \overline{p}, \ldots, p_n - \overline{p}) \cdot (p_1 - \overline{p}, \ldots, p_n - \overline{p})^{\mathrm{T}},$$

and solve the eigensystem $\mathbf{S}x_i = \lambda_i x_i$, $i \in \{0, 1, 2\}$ and $\lambda_0 \geqslant \lambda_1 \geqslant \lambda_2$. We build two new clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ that replace $\mathcal{C}_0$ in $\mathcal{U}$, $\mathcal{C}_1 = \{p \in \mathcal{C}_0 \mid (p - \overline{p})^{\mathrm{T}} x_0 \leqslant 0\}$ and $\mathcal{C}_2 = \mathcal{C}_0 \setminus \mathcal{C}_1$.

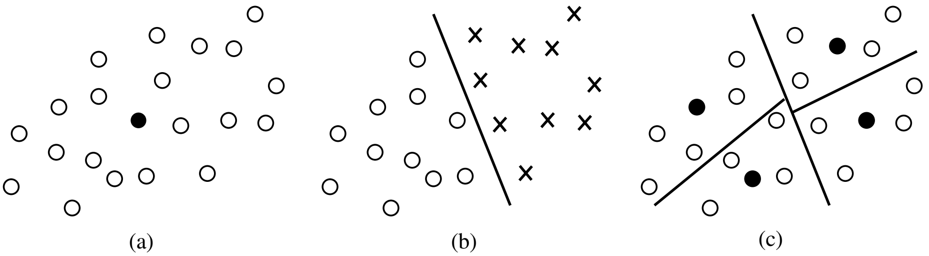

Figure 1. Concept of top-down clustering. (a) Observations are grouped to a cluster with a cluster centre (filled dot); (b) when the associated cluster error is too large, the cluster is split up by Principal Component Analysis, providing two new clusters; (c) this procedure is repeated until all cluster errors are below a given threshold, $t > 0$. The set of centroids is the reduced observation set.

The procedure of cluster splitting is repeated until all clusters in $\mathcal{C} \in \mathcal{U}$ satisfy $e(\mathcal{C}) \leq t$.

Having m clusters in $\mathcal{U}$ yields a total approximation error of $\sum_{\mathcal{C} \in \mathcal{U}} e(\mathcal{C})$. In the second phase of the clustering method, we can further decrease this error by applying so-called Lloyd iterations (Lloyd 1982), that work as follows. Each cluster element $p \in \mathcal{C}_i$ is reassigned to the cluster $\mathcal{C}_j$ for which the distance to the cluster mean $\widehat{p} \in \mathcal{C}_j$ is minimal with respect to d_f. This may change the means for affected clusters, and thus requires the recomputation of the means (2). This process is repeated until convergence.

After cluster splitting and relaxation, the approximation of value $f(p)$ of an original observation $p \in P_0$ is given by the value $f(q)$ of an observation in the thinned dataset P_i, where q is the nearest neighbour of the given point p, using the adaptive error metric, $q = \arg\min_{q \in P_i} d_f(p, q)$.

(*b*) *Thinning through estimation*

Top-down clustering produces approximations of P_0 by iteratively inserting points in regions in which the cluster error is large. A dual approach is given by starting with the full dataset and removing measurements which are redundant. Our method associates to a thinned observation set P_i an approximation of all observation values $f(p)$, $p \in P_0$, and iteratively removes points from P_i which cause the least degradation in approximation quality. Our approximation function $\widetilde{f}_{P_i} : \mathbb{R}^3 \to \mathbb{R}^k$ (k is the dimension of observation values) is given by a linear filter,

$$\widetilde{f}_{P_i}(x) = \frac{1}{Z(P_i, x)} \sum_{p \in P_i} f(p) \cdot w_h(\|x - p\|) \tag{3}$$

with

$$w_h(s) = \exp(-s^2/h^2),$$

and $Z(P_i, x) = \sum_{p \in P_i} w_h(\|x - p\|)$ provides normalization. The weighting function w_h is monotonically decreasing and assigns larger weights to points near x. The parameter h defines the spatial scale of w_h. The resulting approximation is a smooth function that we will evaluate for points $p \in P_0$ (see Fig. 2).

For the input set P_0 we consider the function $\widetilde{f}_{P_0}$ as a reference for the approximations $\widetilde{f}_{P_i}$ of the thinned observations P_i. Note that $\widetilde{f}_{P_0}$ is not an interpolation of the observation values of the full dataset, but rather a linearly filtered version. We define the

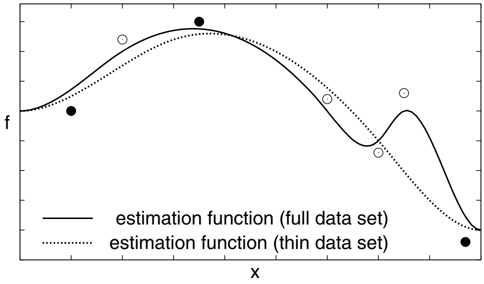

Figure 2. Concept of estimation error analysis. For the full observation set P_0 (all dots) and the simplified set P_i (filled dots), smooth estimation functions $\widetilde{f}_{P_0}$ (solid line) and $\widetilde{f}_{P_i}$ (dotted line) respectively, are constructed. The estimation error is defined by the normalized sum of squared differences between $\widetilde{f}_{P_0}$ and $\widetilde{f}_{P_i}$ evaluated at all $p \in P_0$. See text for definitions.

mean squared error of a thinning P_i of P_0 as

$$E_{\text{mse}}(P_i) = \frac{1}{|P_0|} \sum_{p \in P_0} \|\widetilde{f}_{P_0}(p) - \widetilde{f}_{P_i}(p)\|^2. \tag{4}$$

For $i = 0, 1, 2, \ldots, n-1$, we remove the point $p_{j_i} \in P_i$ of the thinning P_i defined by

$$p_{j_i} = \arg\min_{p \in P_i} E_{\text{mse}}(P_i \setminus \{p\})$$

and set

$$P_{i+1} := P_i \setminus \{p_{j_i}\}.$$

Thus, the removing of the observation at location p_{j_i} causes the least increase in mean squared error E_{mse}.

3. Implementation

For top-down clustering, we use a distance metric for which observation values are scaled by a factor α in order to combine spatial distances with value differences (1). We choose α such that a spatial distance of 100 km corresponds to a difference in measurement values of 1 K, i.e. $\alpha = 100\,\text{km}\,\text{K}^{-1}$. We found this value by comparing the estimation error (4) for various thinnings, and selecting the curve with the best overall performance.

The cluster mean $\widehat{p}$ in Eq. (2) is calculated efficiently as the element $p \in \mathfrak{C}$ closest to the cluster centroid $\overline{p}$ with respect to the corresponding squared distance $\|p - \overline{p}\|^2 + \alpha^2 \|f(p) - \overline{v}\|^2$.

The parameter h in (3) controls the degree of smoothing of the estimation filter. We propose to choose h such that the filtered observation values $\widetilde{f}_{p_0}(p)$ are within 5% of the range of physical measurement errors on average. In other words, $|P_0|^{-1} \sum_{p \in P_0} \|\widetilde{f}_{P_0}(p) - f(p)\| \leq 0.05 \cdot E_{\max}$, where $E_{\max}$ denotes the bound on the physical measurement error in any of the observation channels.

Since the weighting function $w_h(\|x - p\|)$ decreases exponentially with the square of the distance $\|x - p\|$, one may restrict the summation in (3) for computing $\widetilde{f}_{P_i}(x)$ to those observation locations p that are not farther from x than, say $r = 3h$. For $p \in P_i$ consider

$$\Delta E_{P_i}(p) := E_{\text{mse}}(P_i \setminus \{p\}) - E_{\text{mse}}(P_i). \tag{5}$$

The point removal in the estimation error analysis works by selecting the point $p_{j_i} \in P_i$ with minimal $\Delta E_{P_i}(p_{j_i})$ for removal. In our implementation, we maintain a priority queue of observation points $p \in P_i$, starting with $i = 0$, which is sorted by increasing error increments $\Delta E_{P_i}(p)$. For each point removal step, we discard the first priority queue element and update the error increments $\Delta E_{P_i}(p)$ to the new differences $\Delta E_{P_{i+1}}(p)$. Recall that removing a point $p_{j_i} \in P_i$ affects the approximation $\widetilde{f}_P(x)$ only within a range r of p. Therefore, error increments $\Delta E_{P_i}(p)$ do not change for all points $p \in P_i$ that are sufficiently distant from the removed point p_{j_i}. More precisely, an update is required for $p \in P_i \setminus \{p_{j_i}\}$ if and only if there is an observation point $q \in P_i$ with $\|p_{j_i} - q\| \leq r$ and $\|p - q\| \leq r$.

4. Experimental results

In this section, we present experimental results of the proposed thinning strategies, which we apply to satellite temperature measurements. The input data consists of brightness temperatures, measured by the Advanced Microwave Sounding Unit-A (AMSU-A) on the NOAA-15 and NOAA-16 satellites. In particular, we consider measurements over sea in eight channels as typically used in the experimental analysis and forecast system of the DWD. The values $f(p) \in \mathbb{R}^8$, $p \in P_0$ are innovations, i.e. differences between bias-corrected measured brightness temperature and first guess, which vary from -1 to 1 K per channel (Fig. 3). As the reference method, we use stepwise thinning that retains every third point in the zonal and meridional directions, yielding a spatial density of observations of about 70 to 140 km in the thinned dataset.

Figure 4 shows a plot for differences in data density for the clustering and stepwise thinning. Red regions indicate that the clustering or estimation method retains more points than stepwise thinning in a window of 100 km × 100 km. Blue regions correspond to lower data density for clustering. In Fig. 4(a), we observe two effects. Firstly, on eastern boundaries of the two tracks at 120°W, there are more measurements retained by the clustering method. This can be explained by the fact that in these regions the original observation set contains fewer observations per area unit due to scan line properties of the measurement device, i.e. the sampling density of observations decreases towards the far end of the track because the camera looks at the surface under a non-right angle. Stepwise thinning coherently removes observations in these areas without considering this spatial effect, while the clustering method can overcome this problem by preserving more points through modelling larger cluster errors. The same effect can be observed at western track boundaries at 30°E. Secondly, the clustering method retains more points in regions in which satellite tracks overlap, such as at 135°W, 50°S due to its capability of preserving observations with large variance in measurement values, even when they are spatially close.

The same two effects can be observed for the density difference between the estimation method and stepwise thinning (Fig. 4(b)).

Figure 5 shows error curves with respect to mean squared estimation error (4) for top-down clustering and estimation error analysis with $h = 50$ km (3). In our experiments we have found that the clustering method provides better results than the stepwise thinning strategy. Estimation error analysis was designed to minimize the error shown in Fig. 5 and therefore provides best results down to about 5000 remaining observations. For larger h, i.e. $h = 100$ km, estimation error analysis performs even better, also for thinnings with fewer remaining observations (not shown).

We also present first results of experimental analysis–forecast cycles using top-down clustering, estimation error analysis, and stepwise thinning over the period of 27

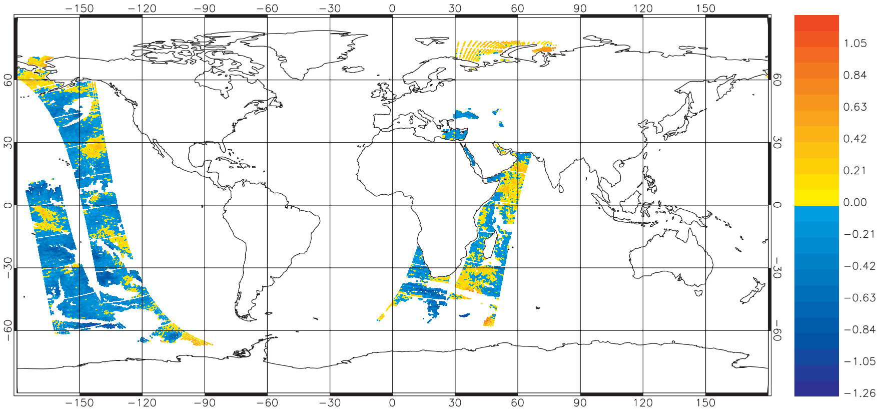

Figure 3. Full dataset: differences (K) between observed brightness temperatures and first guess for channel 5 of the AMSU-A device on the NOAA-15 satellite on 21 July 2004; dataset consists of 27 367 measurements over 2230–0130 UTC.

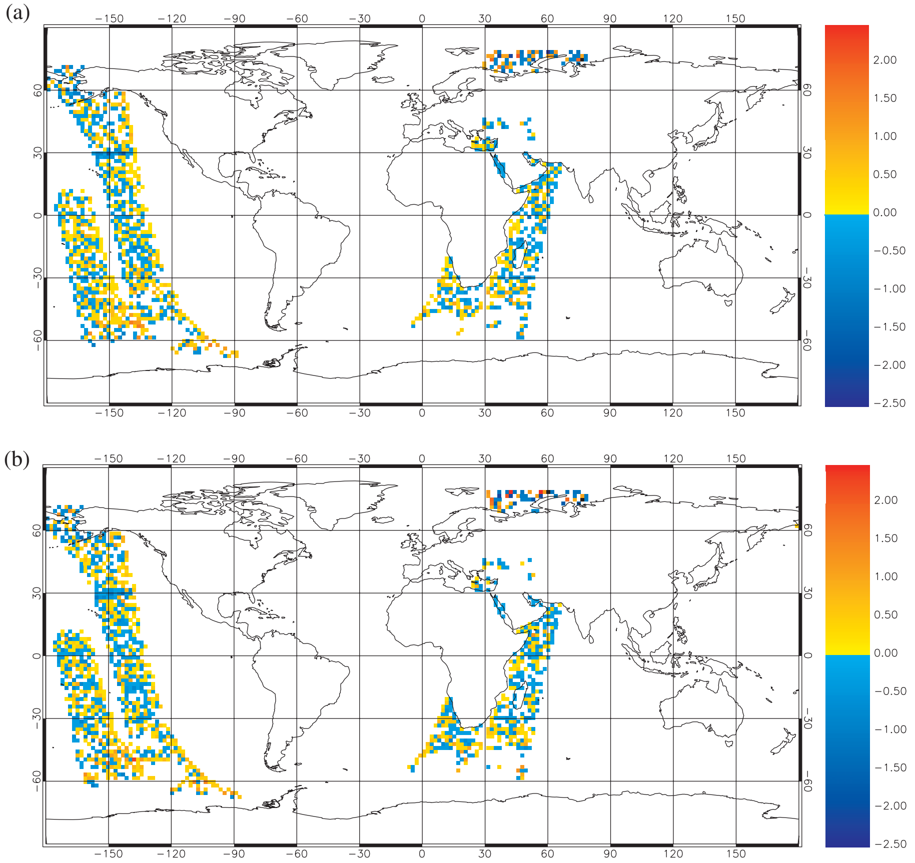

Figure 4. Differences in data density (a) between top-down clustering and stepwise thinning, and (b) between estimation error analysis and stepwise thinning. The colour denotes the average difference of numbers of observations retained in 100 km × 100 km boxes by the compared methods.

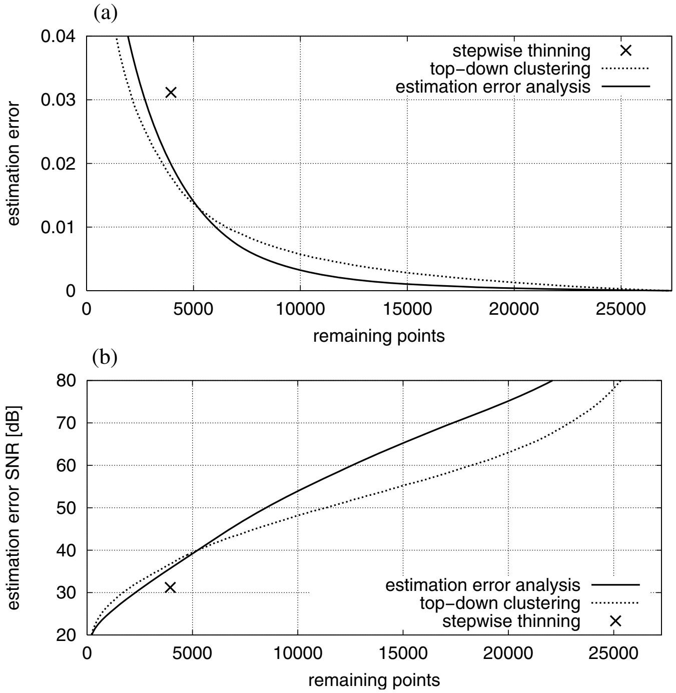

Figure 5. Comparison of top-down clustering, estimation error analysis, and stepwise thinning, with respect to estimation error (4): (a) mean squared estimation error, and (b) signal-to-noise ratio (SNR, dB) of mean squared estimation error that gives the error relative to the mean absolute value of the signal.

$$\mathrm{SNR} = 10 \log \frac{\sum_{p\in P_0} \|\tilde{f}_{P_0}(p)\|^2}{\sum_{p\in P_0} \|\tilde{f}_{P_0}(p)-\tilde{f}_{P_i}(p)\|^2}.$$

December 2004 to 22 January 2005. All experiments are carried out with the global model GME of DWD with 40 km horizontal resolution and 40 vertical levels (Majewski *et al.* 2002) and assimilation by optimal interpolation. Compared to the computational costs of the 1D-Var scheme of several minutes, the time for the top-down clustering is negligible, requiring only about 0.7 seconds. (Stepwise thinning costs less than 0.1 seconds.) The processing time for estimation error analysis is significantly higher at about 110 seconds. This is due to the numerous updates of estimation terms (5) needed for every point removal.

Figure 6 shows the northern hemispheric mean anomaly correlations for the 500 hPa geopotential as a function of forecast time for all three thinning approaches. Both the estimation error analysis and the clustering method show slightly higher anomaly correlations compared to the stepwise thinning, with the largest positive impact for forecast lead times of more than 60 hours. While the difference between estimation error analysis and stepwise thinning increases almost linearly over the last four days of the forecast time, the difference between clustering and stepwise thinning increases moderately at the beginning and with a higher rate for days 5 to 7. The difference between estimation error analysis and clustering is much smaller than the difference between stepwise thinning and each of the two other methods.

Compared to the consistent improvements achieved by the two adaptive methods in the northern hemisphere, the experimental results show ambiguous behaviour for the southern hemisphere (Fig. 7). There are only very small differences between estimation

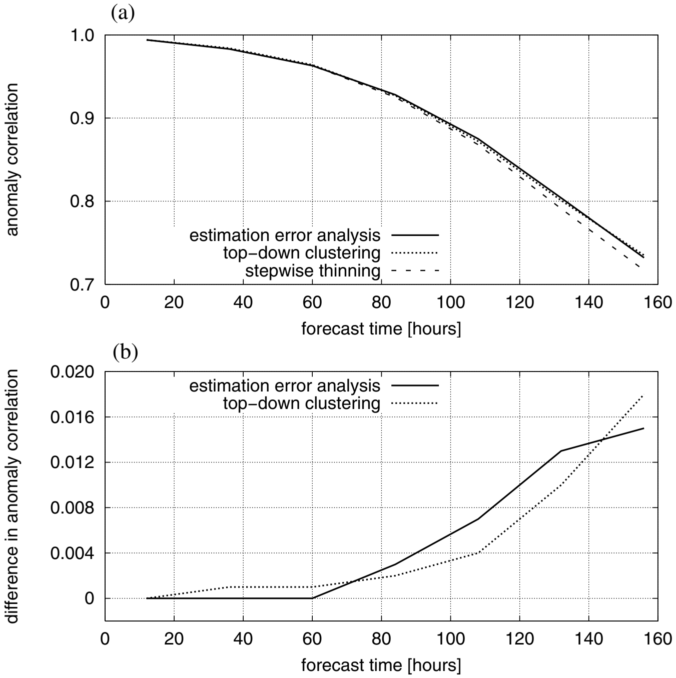

Figure 6. (a) Anomaly correlations of 500 hPa geopotential for the northern hemisphere based on 27 156-hour forecasts started at 12 UTC from 27 December 2004 to 22 January 2005 from estimation error analysis (solid line), top-down clustering (dotted), and stepwise thinning (dashed). Scores are calculated at intervals of 24 hours starting at the 12-hour forecast. (b) is as (a), but for estimation error analysis minus stepwise thinning (solid) and clustering minus stepwise thinning (dotted).

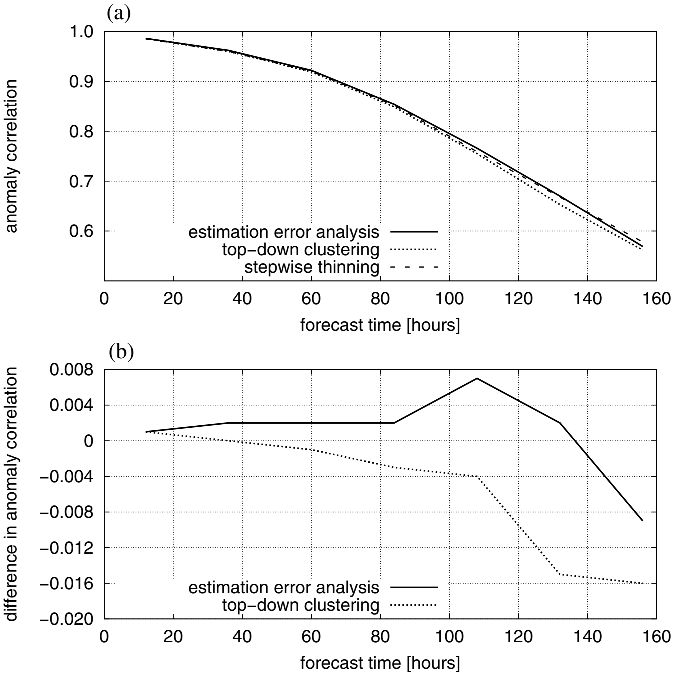

Figure 7. As Fig. 6, but for the southern hemisphere.

error analysis and stepwise thinning, but with slightly higher anomaly correlations for the estimation error approach, except for the difference between low anomaly correlations at forecasts for 156 hours. The positive impact of the estimation error analysis is clearly smaller on the southern hemisphere than on the northern. The southern hemispheric anomaly correlations resulting from the cluster method are lower than those related to stepwise thinning. The negative impact is comparable in magnitude to the positive impact seen in Fig. 6.

Since all presented differences of forecast scores are small and rely on a relatively small number of forecasts, we investigated in more detail the results for forecast lead times of more than 60 hours. We found that all positive and negative impacts are not caused by strong outliers but that the average scores shown in Figs. 6 and 7 are representative for most of the 27 forecasts.

5. Conclusion and future work

We have presented methods for thinning of large sets of meteorological observations for data assimilation in weather forecast systems. In the clustering scheme, representative observations are inserted iteratively into a growing set of observations. When the desired dataset size is reached, a relaxation procedure improves the approximation of the original full set of observations. The approximation quality is modelled using a metric that combines spatial distance of observation locations with measurement values. In the thinning by removal of observations from the full set, we propose to use an estimation of the quality of thinned sets of observations to iteratively identify and remove the most redundant observation. In both methods, more measurements are retained in regions with large variance in the original signal, while fewer remain in homogeneous regions.

Our applications of the adaptive thinning methods to AMSU-A data resulted in improvements over the stepwise thinning in terms of spatial distribution of the thinned data (see Fig. 4) and in terms of the approximation of the full dataset according to the mean squared estimation error (4) (see Fig. 5). The clustering and estimation error methods were applied to the same type of observations within the analysis–forecast cycles over a period of about four weeks. The results showed that better forecasts can be achieved with clustering and estimation error analysis in the northern hemisphere than by forecasts relying on the stepwise thinning. In the southern hemisphere, estimation error resulted in marginally improved scores. However, the clustering method produced a negative impact with respect to stepwise thinning. This raises the need for more detailed studies of the clustering method in future work, e.g. with respect to the tuning of the metric (1) or the capability of clustering to preserve the variance of the innovations.

In future work we plan to apply the thinning methods to other types of observational data. Furthermore, we can improve the speed of the estimation error analysis by performing a reduced number of updates of estimation error terms at every point removal. First experiments show that we can reduce the total costs for thinning down to 3 seconds (accepting a slight loss in estimation quality), which is a significant speed-up compared with 110 seconds for the full evaluation.

We expect further improvements with the currently developed point insertion version of the estimation error analysis. Such an approach significantly reduces the number of iterative evaluations of estimation error terms. The current implementation of the estimation error analysis with point removal is a clear improvement over stepwise thinning but is also a trade-off between computational feasibility and optimality with respect to the theoretical properties which can affect the thinning at high numbers of iterations. This will be overcome by a point insertion estimation error approach.

The comparison of the methods when thinning to even smaller datasets is an important part of future studies as well. Less input data in the global analysis is computationally beneficial and can be necessary to avoid spatially correlated observation errors. However, the importance of incorporating the observed values themselves in the thinning approach increases with decreasing numbers of retained observations to preserve the essential information content of the data.

Also, thinning methods may be evaluated in simulations using the framework of Liu and Rabier (2002), when extended to spherical datasets.

ACKNOWLEDGEMENTS

We thank Reinhold Hess, Gerhard Paul and Andreas Rodin for fruitful discussions. We also thank the reviewers for helpful comments. This work was supported by the DFG Graduiertenkolleg 1042 ‘Explorative Analysis and Visualization of Large Information Spaces’.

REFERENCES

Gersho, A. and Gray, R. M. 1992 *Vector quantization and signal compression.* Kluwer Academic Publishers

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

Lloyd, S. P. 1982 Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, **28,** 129–137

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

Ramachandran, R., Li, X., Movva, S., Graves, S., Greco, S., Emmitt, D., Terry, J. and Atlas, R. 2005 ‘Intelligent Data Thinning Algorithm for Earth System Numerical Model Research and Application’. In Proceedings of 21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, San Diego. American Meteorol. Soc, Boston, USA

Riishøjgaard, L. P. 1998 A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus*, **50A,** 42–57