Thinning WindSat Data Using Support Vector Regression

M. Richman[1], L. Leslie[1], H. Mansouri[2], C. Shafer[1]

[1]School of Meteorology

[2]School of Industrial Engineering

The University of Oklahoma

JCSDA 6th Workshop on Satellite Data Assimilation

June 10-11, 2008

Holiday Inn BWI

Outline

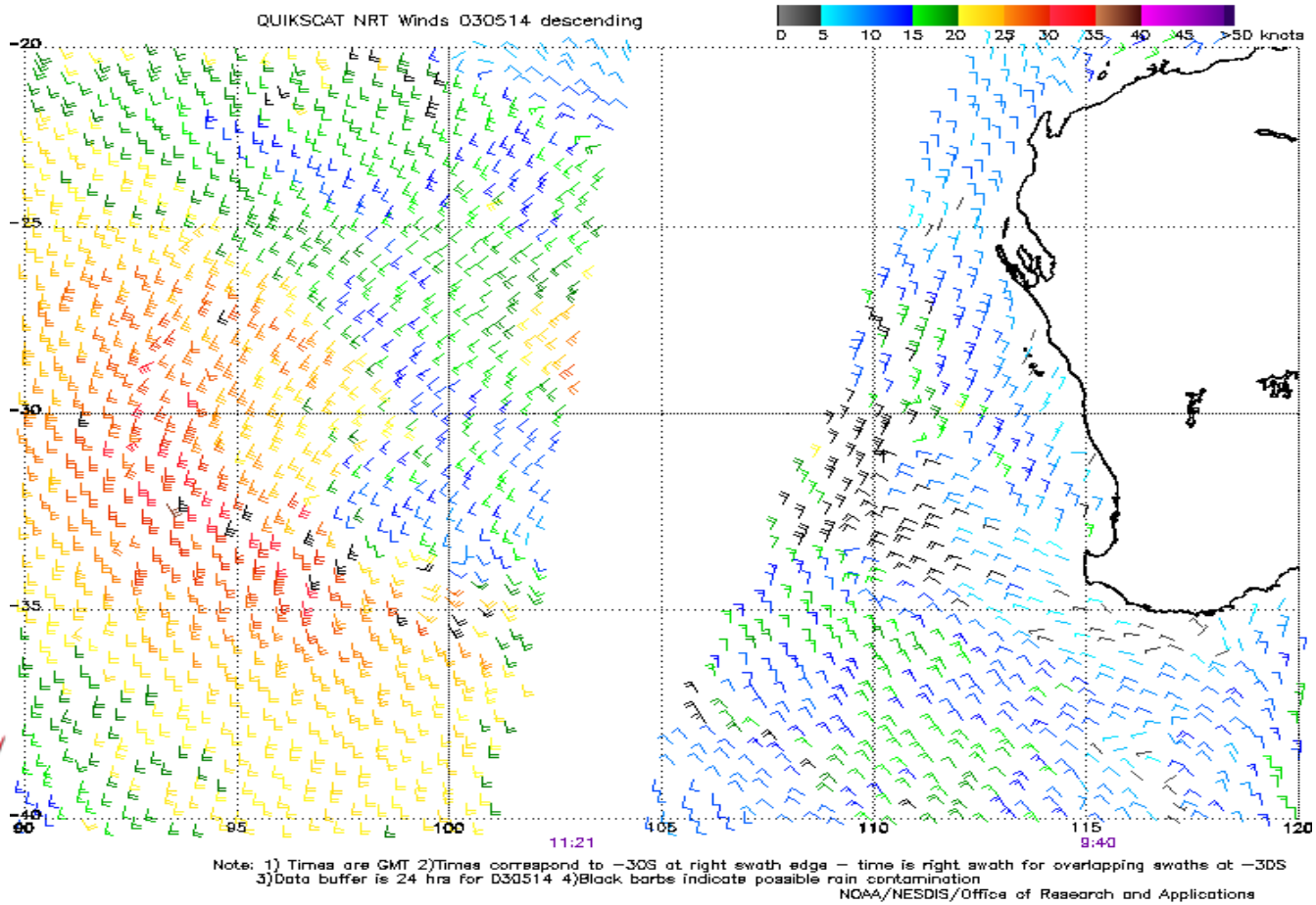
- Motivation
- Objectives
- SVR
- Thinning WindSat Data
- On-line Thinning
- Conclusion and Future work

Motivation

- The amount of wind data provided by satellites is *very large* and critical for accurate and timely weather forecasts
- *Consequently*, any type of analysis of these data should use subsets
- Traditionally, superobbing is the main method used to thin these data (Barnes scheme, average, random…)
- Support Vector Regression (SVR) is an attractive alternative

Impact of Ocean Winds on Forecasts

Quikscat scatterometer wind vectors valid 0940 and 1121UTC 14 May 2003that were ingested into the initial analysis valid 12UTC 14 May 2003.



Forecast Improvement Statistics

Summary of predictions of the model runs after 36 hours, valid 0000 UTC 16 May 2003

	Central Pressure of low (hPa)	Primary position error of low at 00UTC (km)	Albany to Geraldton pressure gradient (hPa)	Maximum wind speed (m/s)
Analysis	984	0	20.1	25
No ocean wind data	987	150	15	22
Ocean wind data included	984	100	20	25

Objectives

- Thin WindSat data
- Ameliorate the quality of the subsets (thinning rate and reconstruction accuracy)
- Compare SVR and the other thinning methods
- Improve the thinning procedure to allow faster on-line implementation

Support Vector Regression

- SVR is a supervised learning algorithm
- It is used in tasks such as statistical classification and nonlinear regression analysis [did a thorough overview of the theory last year; it maps the input space into a higher dimensional Hilbert space (feature space). Mercer kernels are used]
- It consists of solving a quadratic programming problem
- Over a few thousand data points SVR tends to be very slow and requires a lot of core memory

SVR continued

Data points x_i and the desired targets (observations) y form a data-target pair (x_i,y_i) s.t. $f(x_i) \sim y_i$. The prediction function f belongs to a class of functions denoted by F, s.t.

$F: = \{x \text{ in } R^n \rightarrow \langle w \cdot x \rangle + b : ||w|| \leq B, \text{ where } B > 0, w = \sum_j \alpha_j x_j$

$\langle w \cdot x \rangle$ is mapped into Hilbert space (H) with a Mercer kernel as Φ, and the problem becomes

$F: = \{x \text{ in } R^n \rightarrow \langle w \cdot \Phi(x) \rangle_H + b : ||w||_H \leq B, \text{ where } w = \sum_j \alpha_j \Phi(x)_j$

The problem has constraints that are kernelized through quadratic programming that leads to the optimal prediction function:

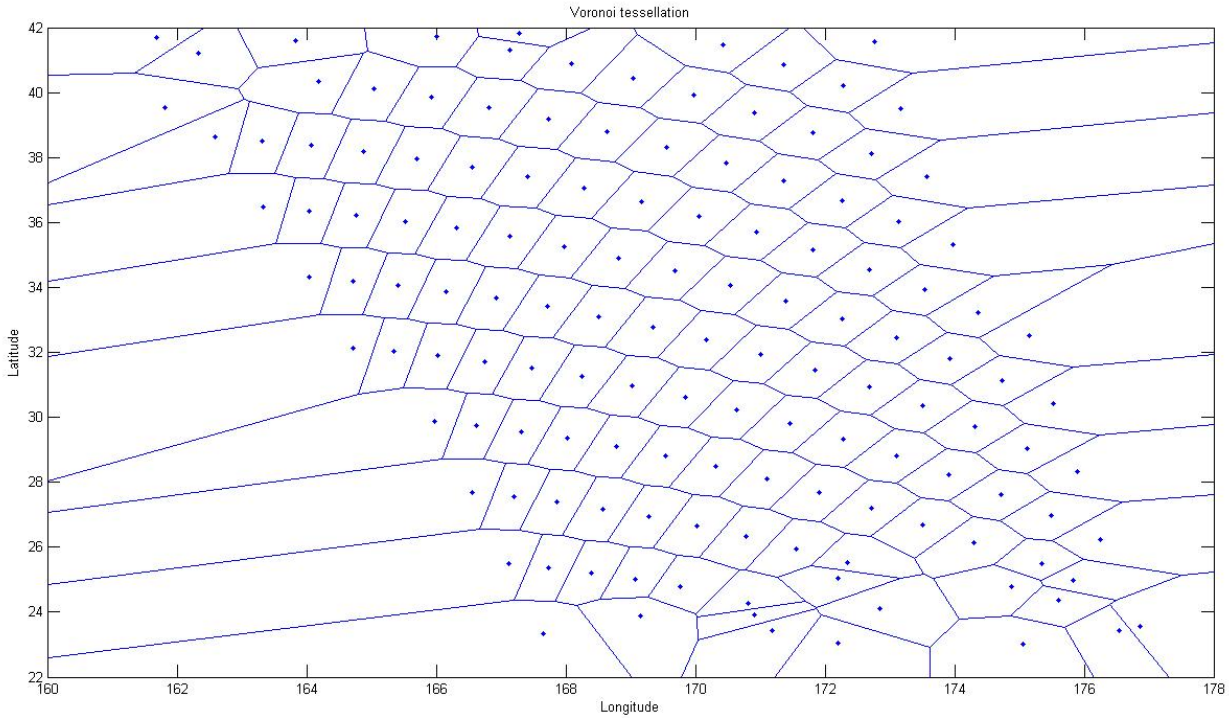
$f: = \{x \rightarrow \sum_j \alpha_j^* k(x_j,x) + b^*$ Note, the vectors x_j for which α_j^* are nonzero are called “support vectors”

Voronoi Tessellation

- Instead of thinning all the data in one optimization step, they are split into subsets and optimized separately with multiple SV

- The Voronoi Tessellation consists of decomposing a data set into subsets using a discrete set of points

Voronoi Diagram

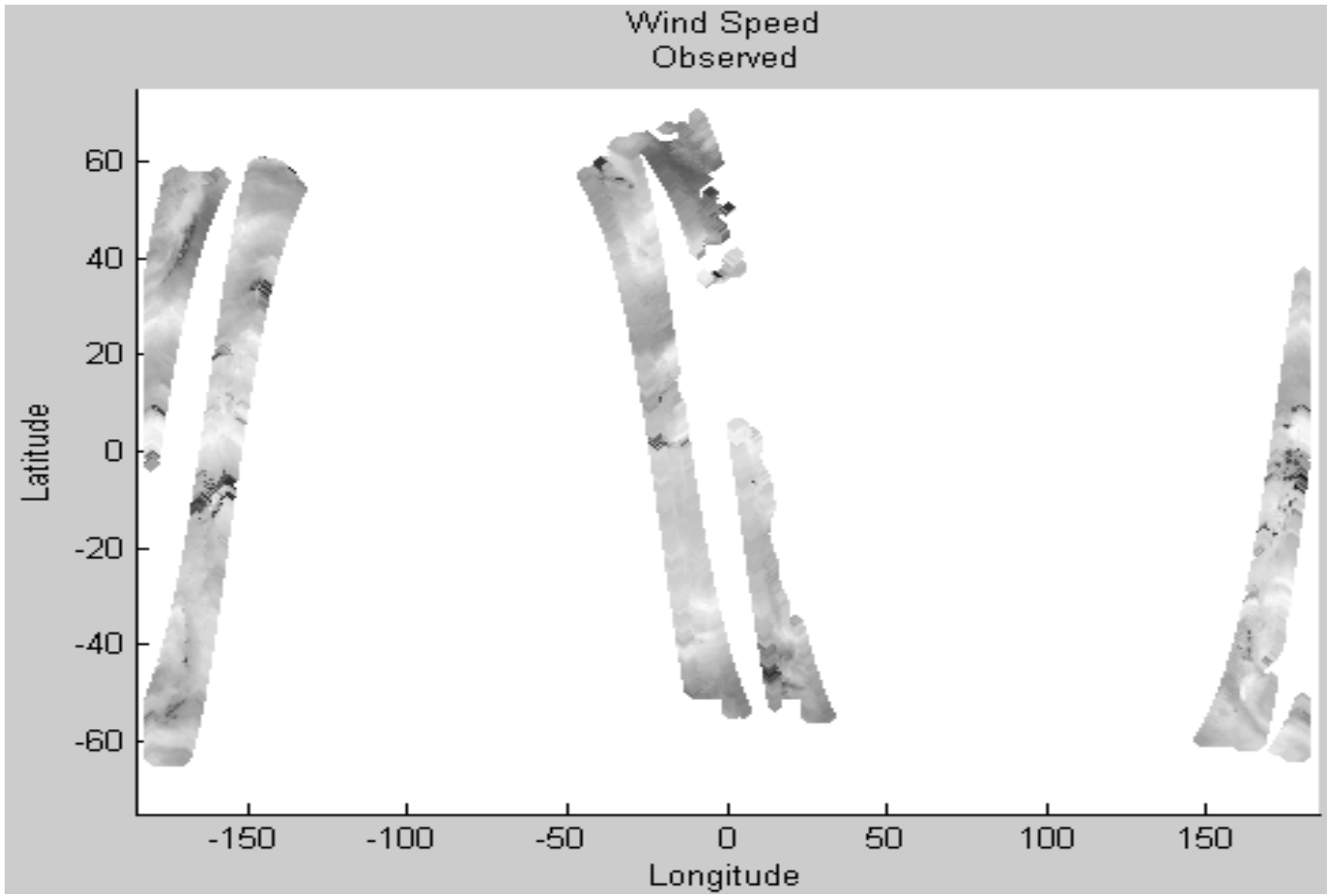


The thinning can be made quicker through parallelization

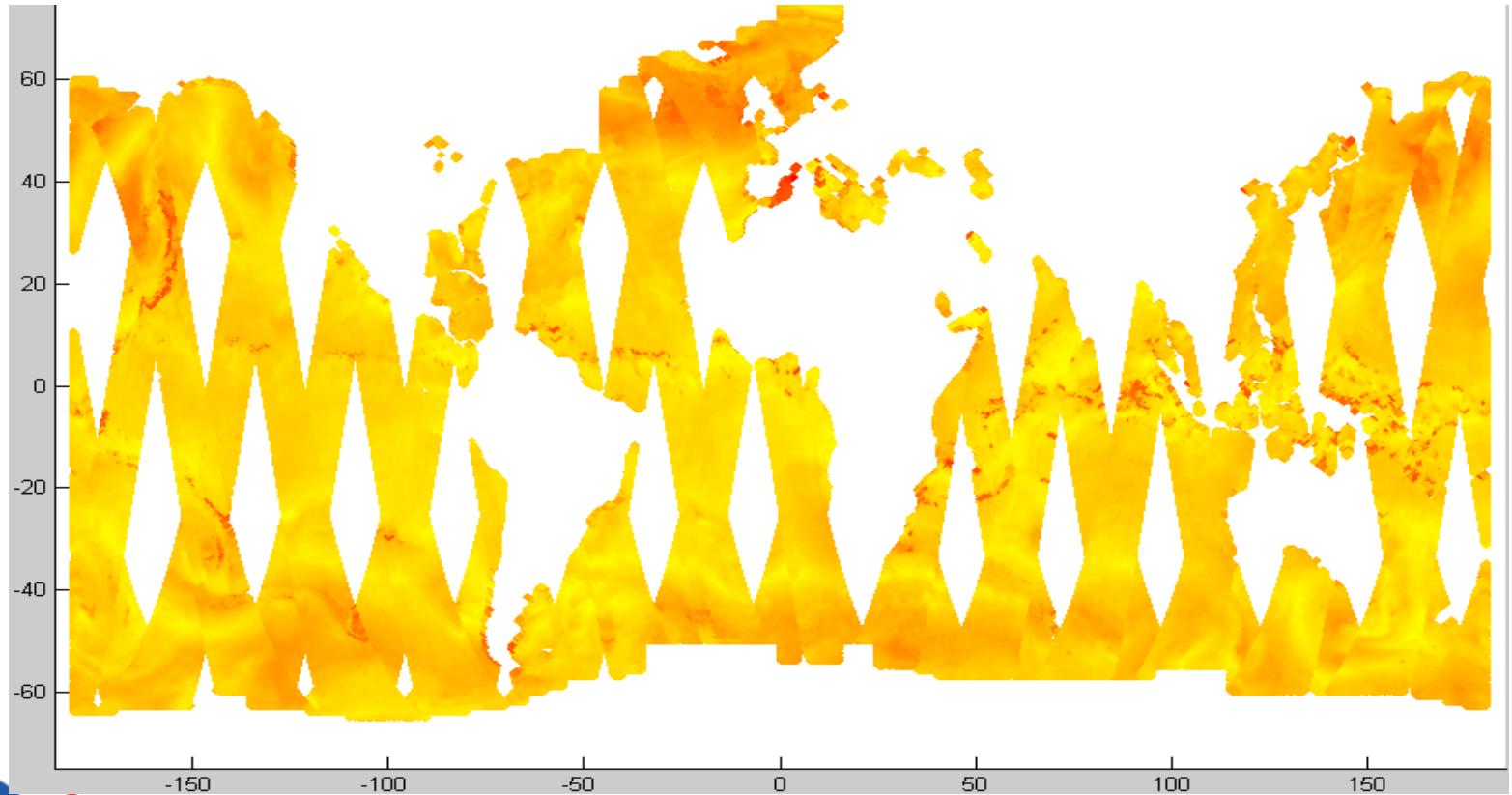
Thinning WindSat Data
-Data Set-

- During the experiment, we used data collected on January 13, 2005 from 2:45 to 6:18 (GMT)
- The data set consists of 226,393 points
- 2 SVR analysis are needed: u and v components of the wind speed vector
- The input vector contains longitude, latitude, SST, water vapor, cloud liquid water, rain rate

Thinning WindSat Data
-Data Set-



Thinning WindSat Data
-24 hours Data-

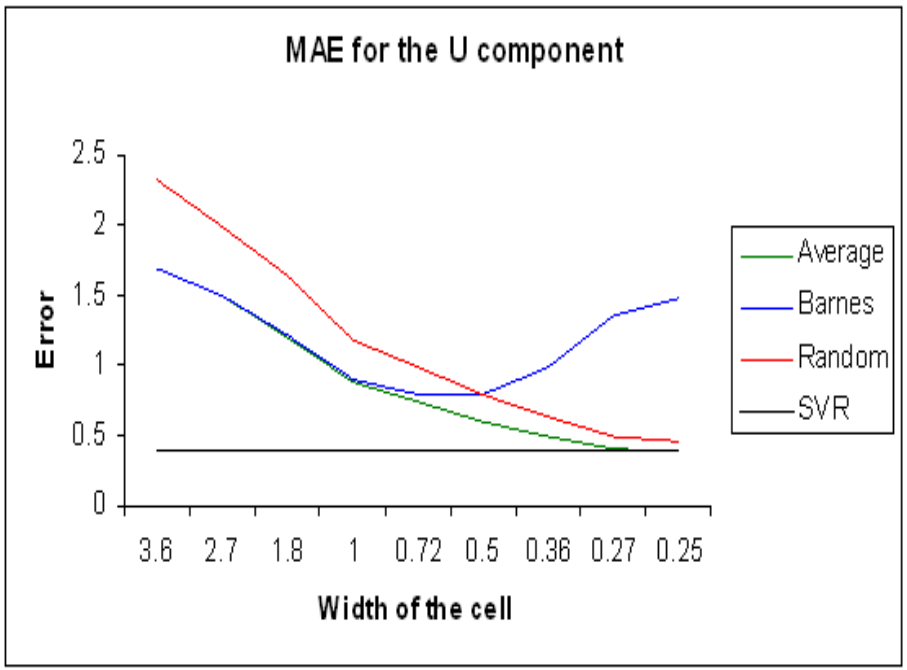


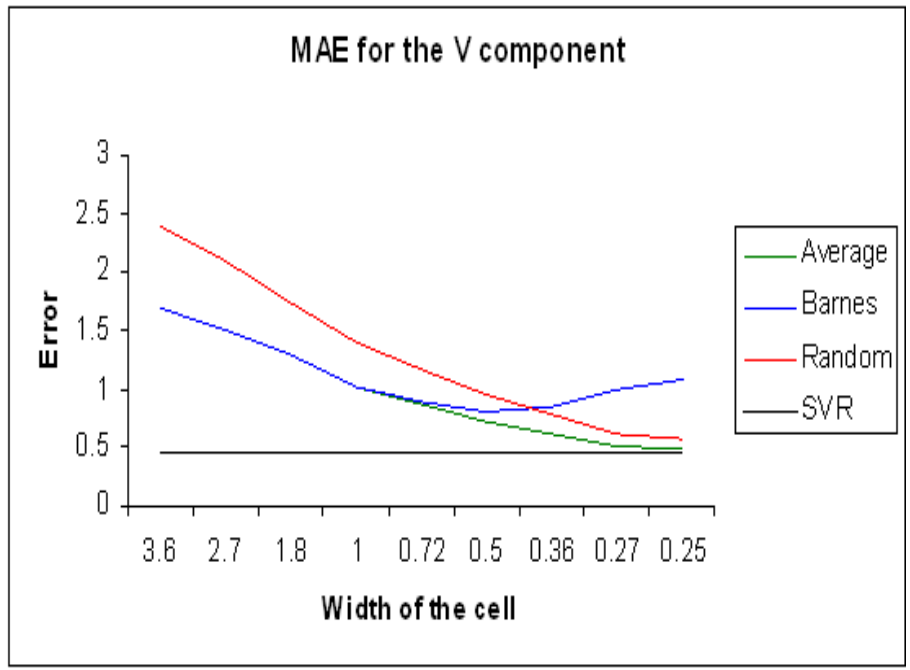
Over 1.5 million data points

Thinning WindSat Data -Experiments-

- The surface of the earth to cells of *h^o latitude* × *h^o longitude*
- All data points will be distributed to these cells
- One data point will be selected from each nonempty cell (randomly, Barnes scheme, average)
- The selected data points are used to construct the subset (the bigger is h, the smaller is the subset)

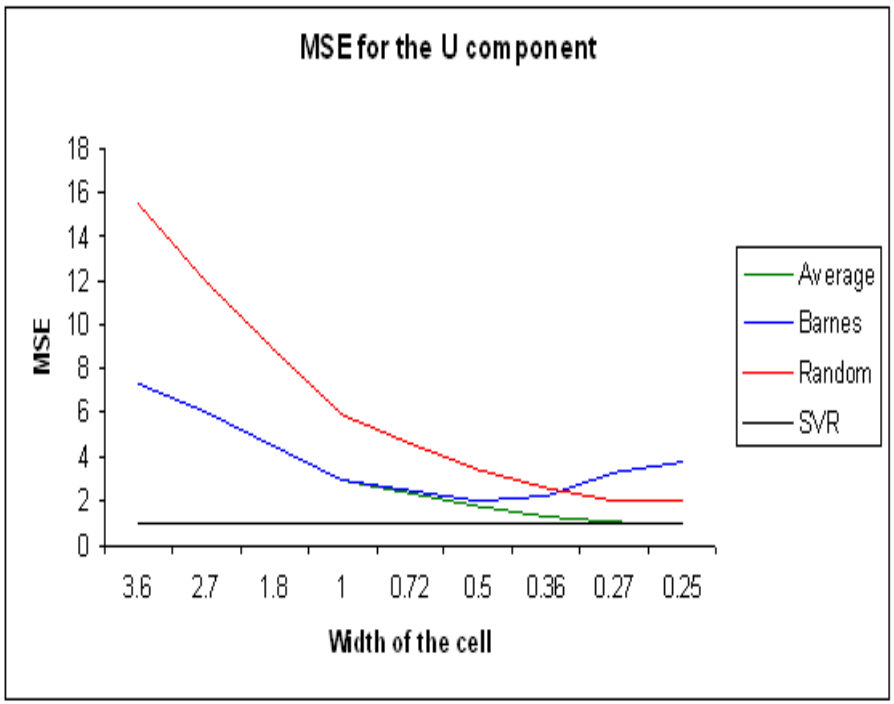
Thinning WindSat Data
-Results: MAE (m/s)-

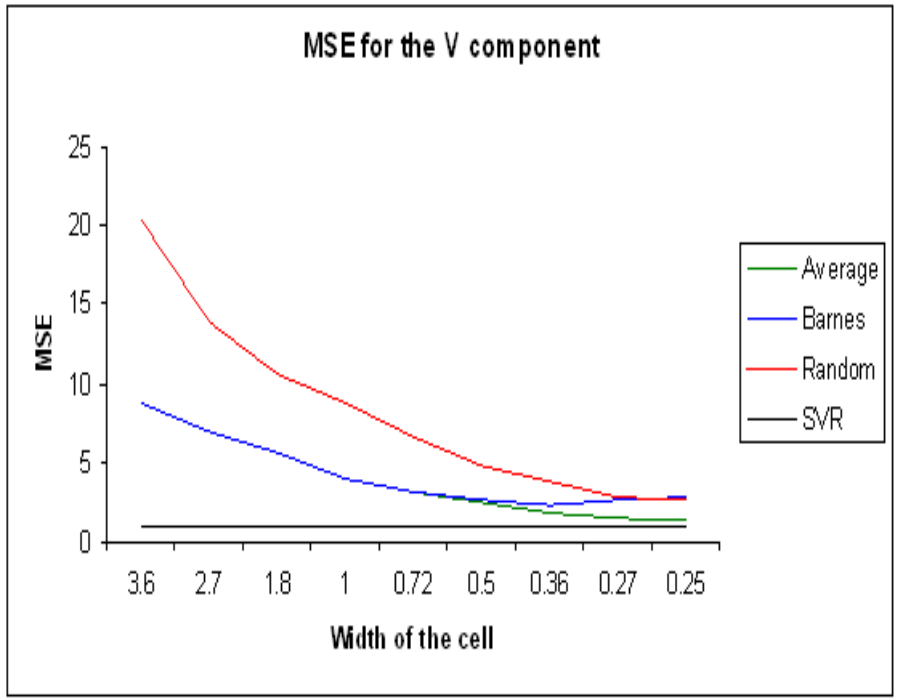




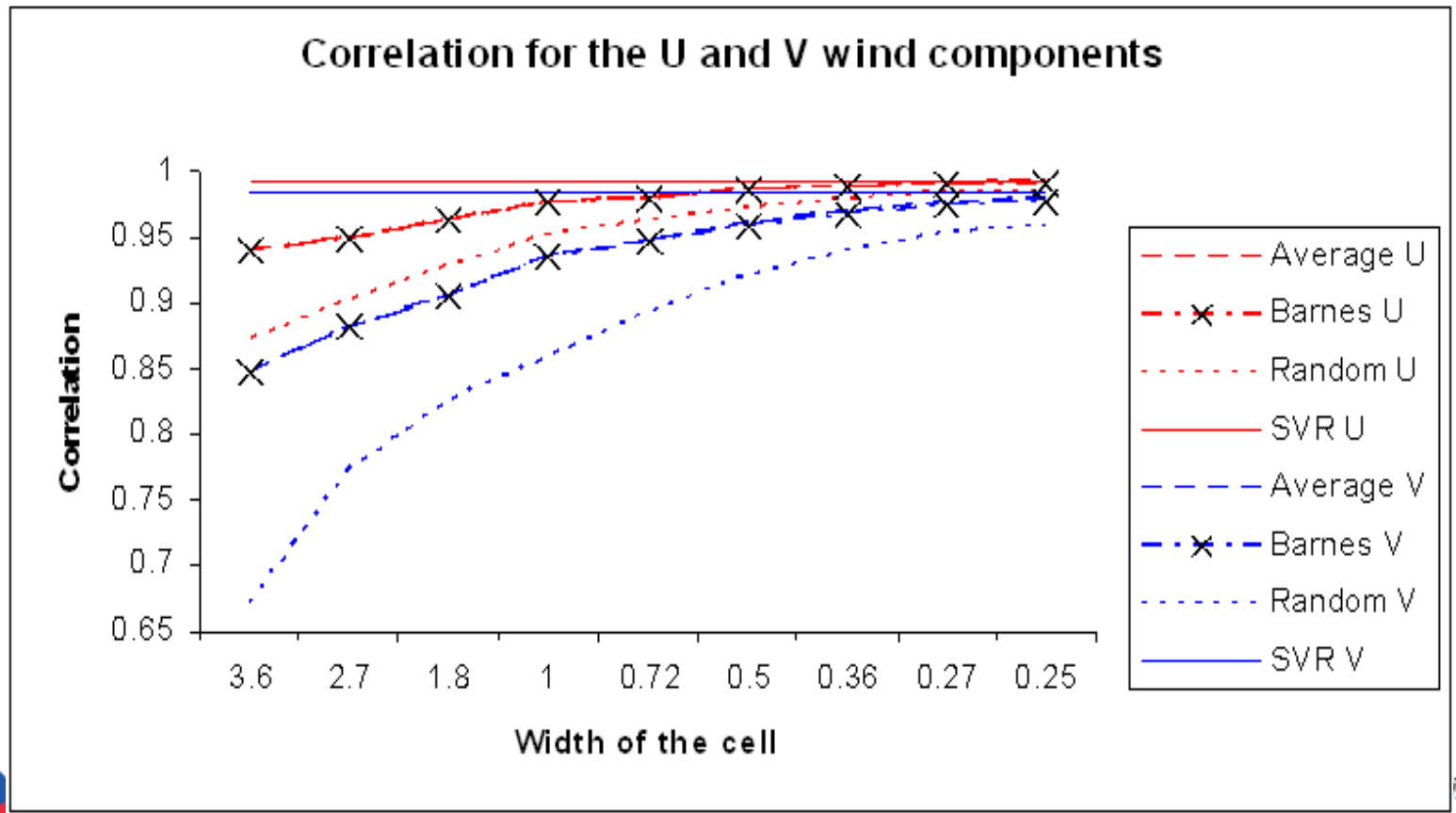
Width of the cell is in degrees

Thinning WindSat Data
-Results: MSE (m^2/s^2)-

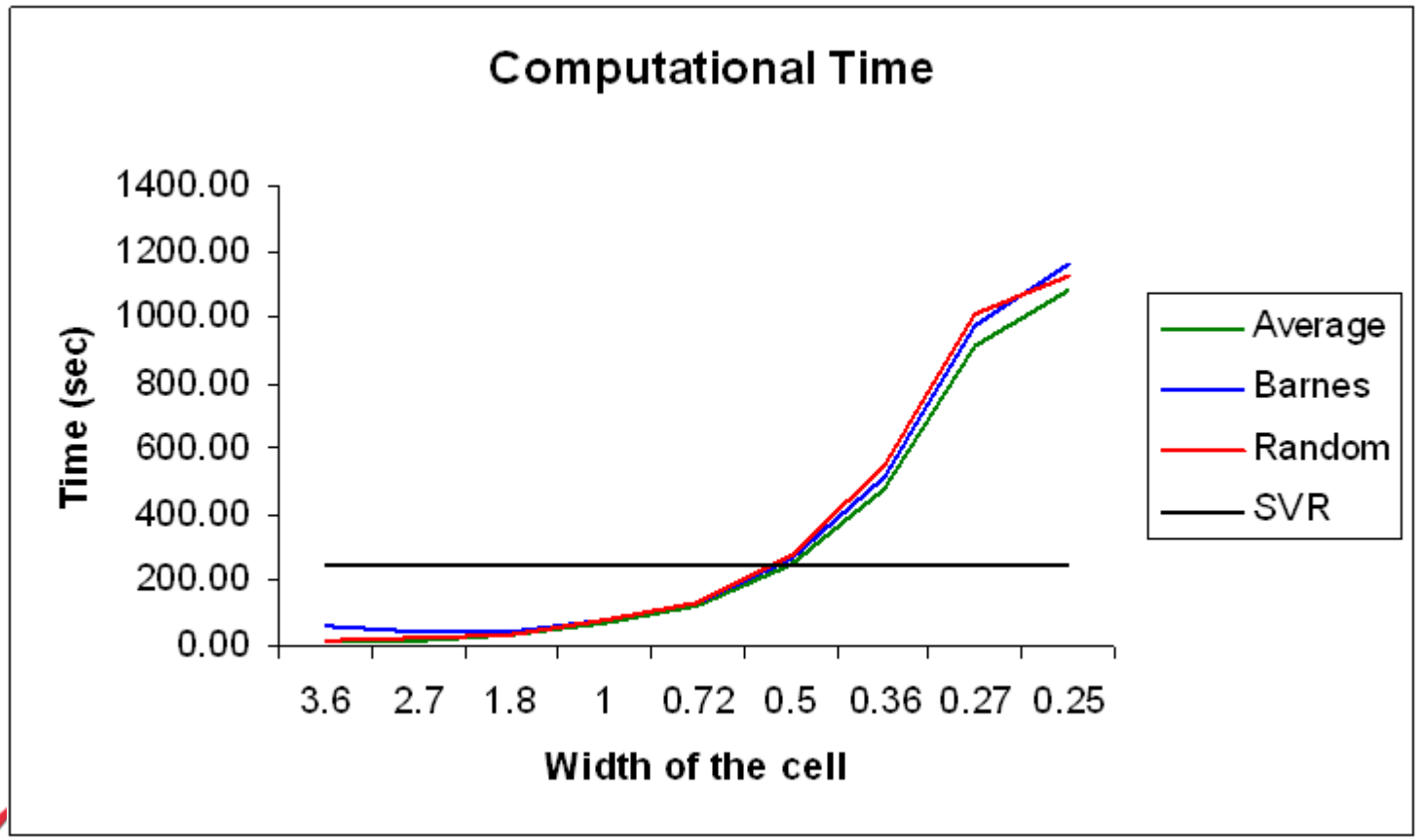




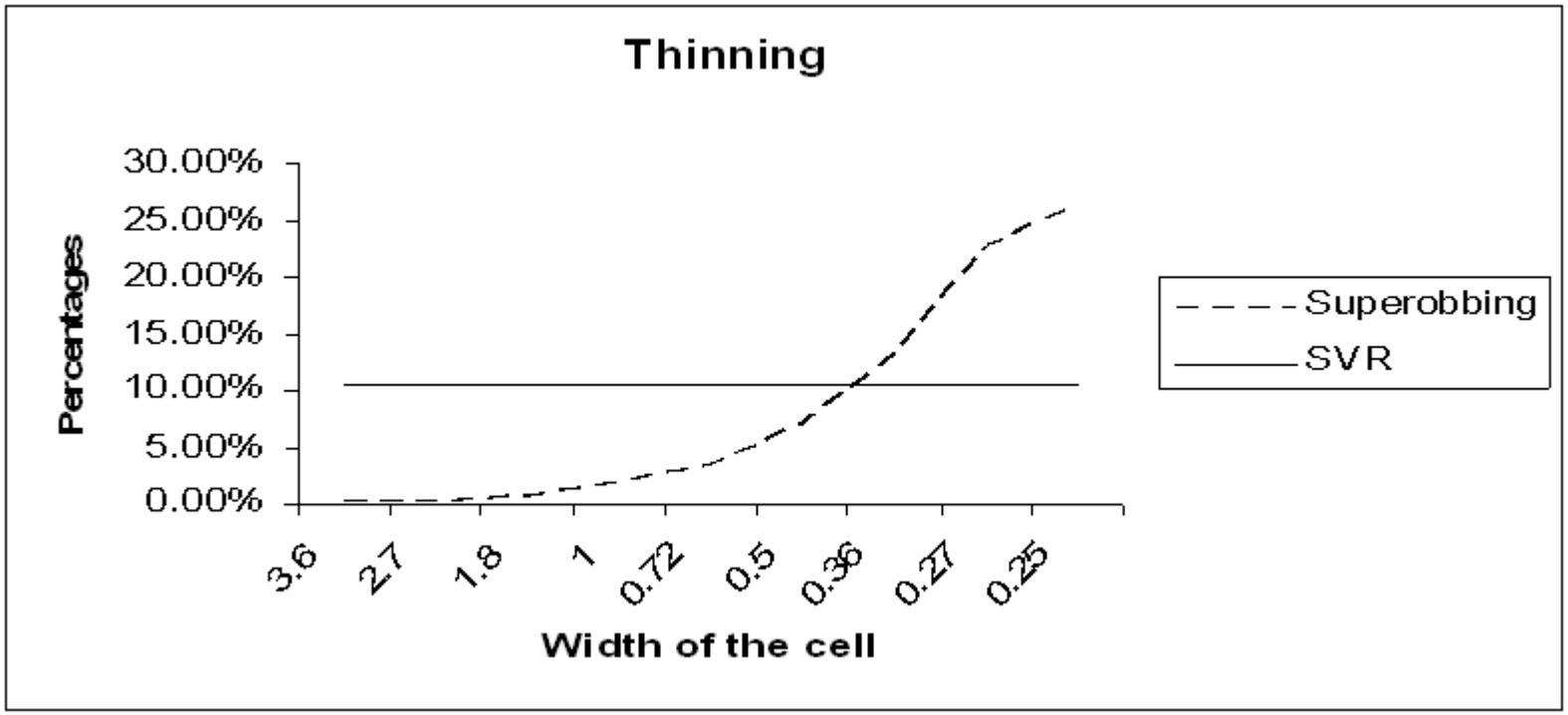
Thinning WindSat Data
-Results: Correlation-



Thinning WindSat Data
-Results: Computation Time (s)-



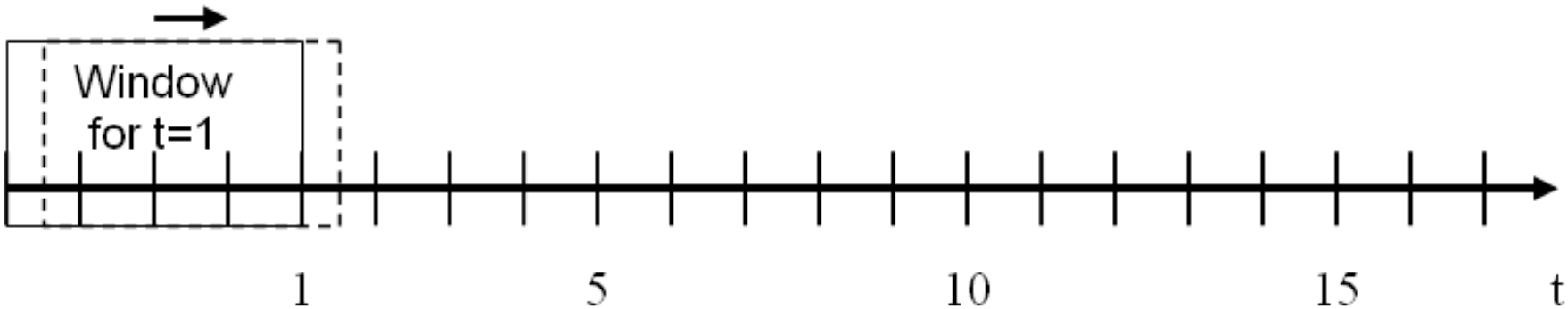
Thinning WindSat Data
-Results: Thinning Percentage-



Thinning percentage = (size of the subset)/(size of the original data set)

On-Line Thinning
-Methodology-

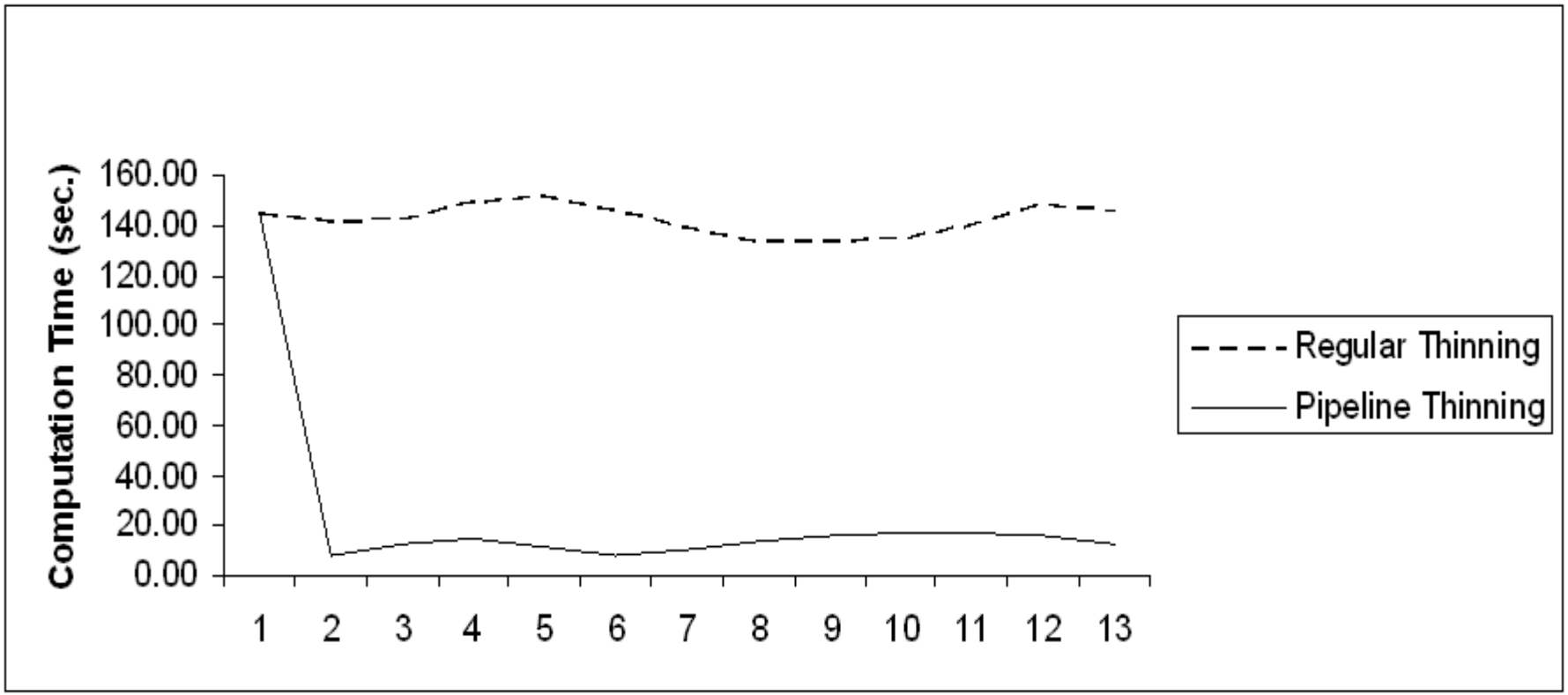
- Develop a pipeline method based on SVR and Voronoi tessellation
- Manage an on-line stream of the WindSat data
- Thin WindSat data on-line



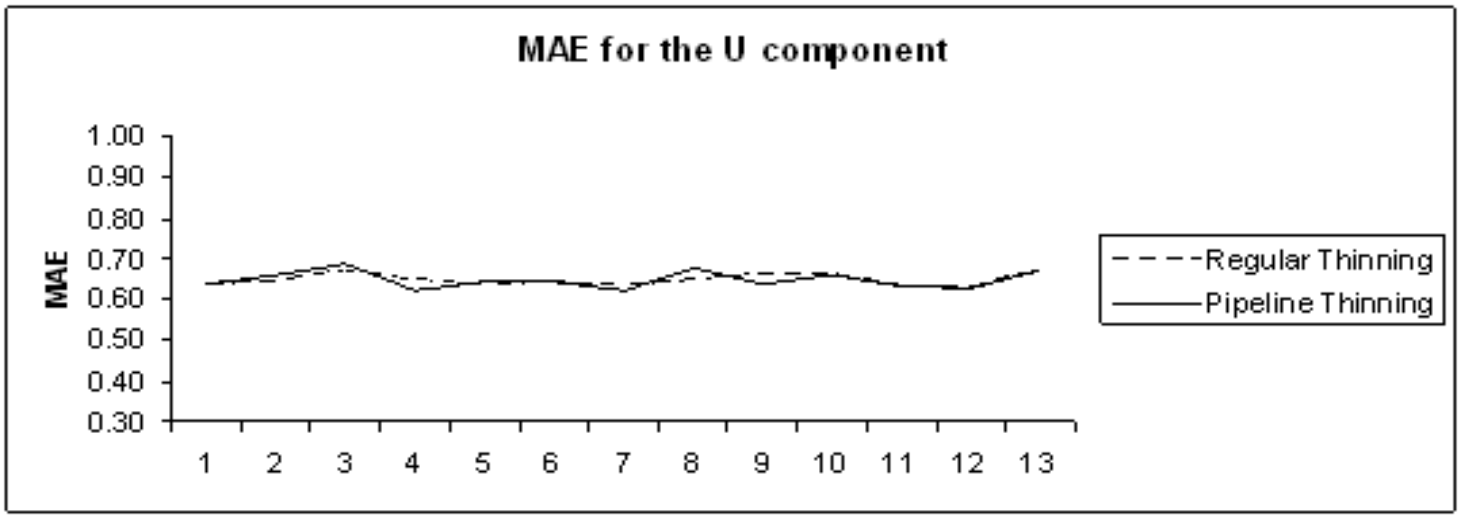
On-Line Thinning
-Experiment-

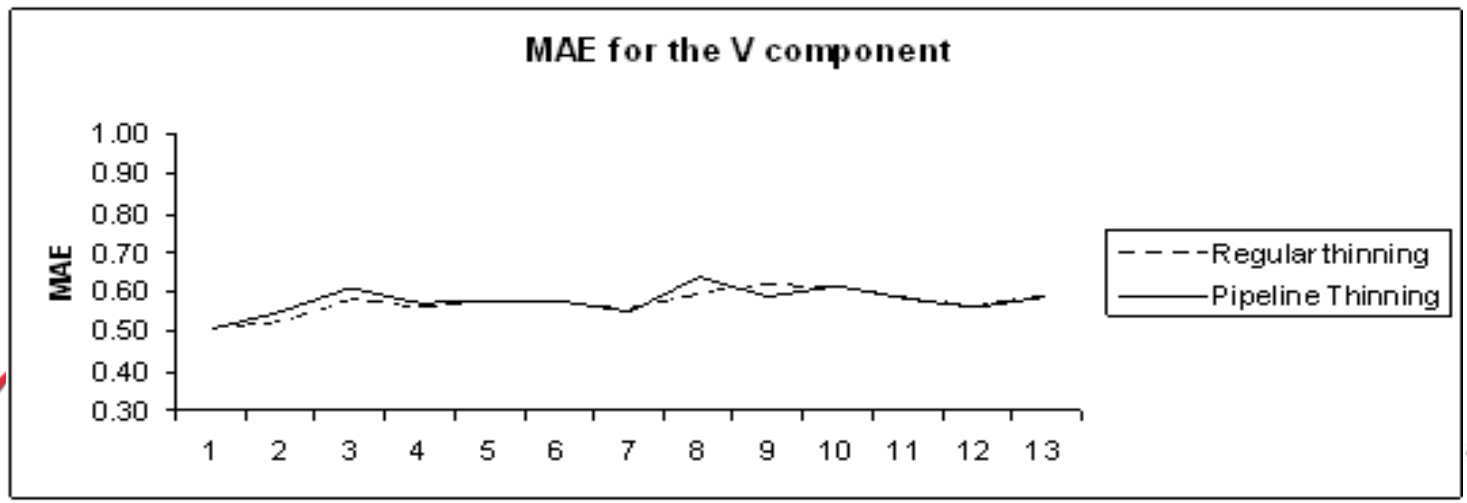
- WindSat data were thinned for 13 periods
- The data set has 120,983 data points
- The most up to date data are used during the thinning
- Results obtained using the pipeline were compared to results obtained thinning the whole data for each period

On-Line Thinning
-Results: Computation Time (s)-

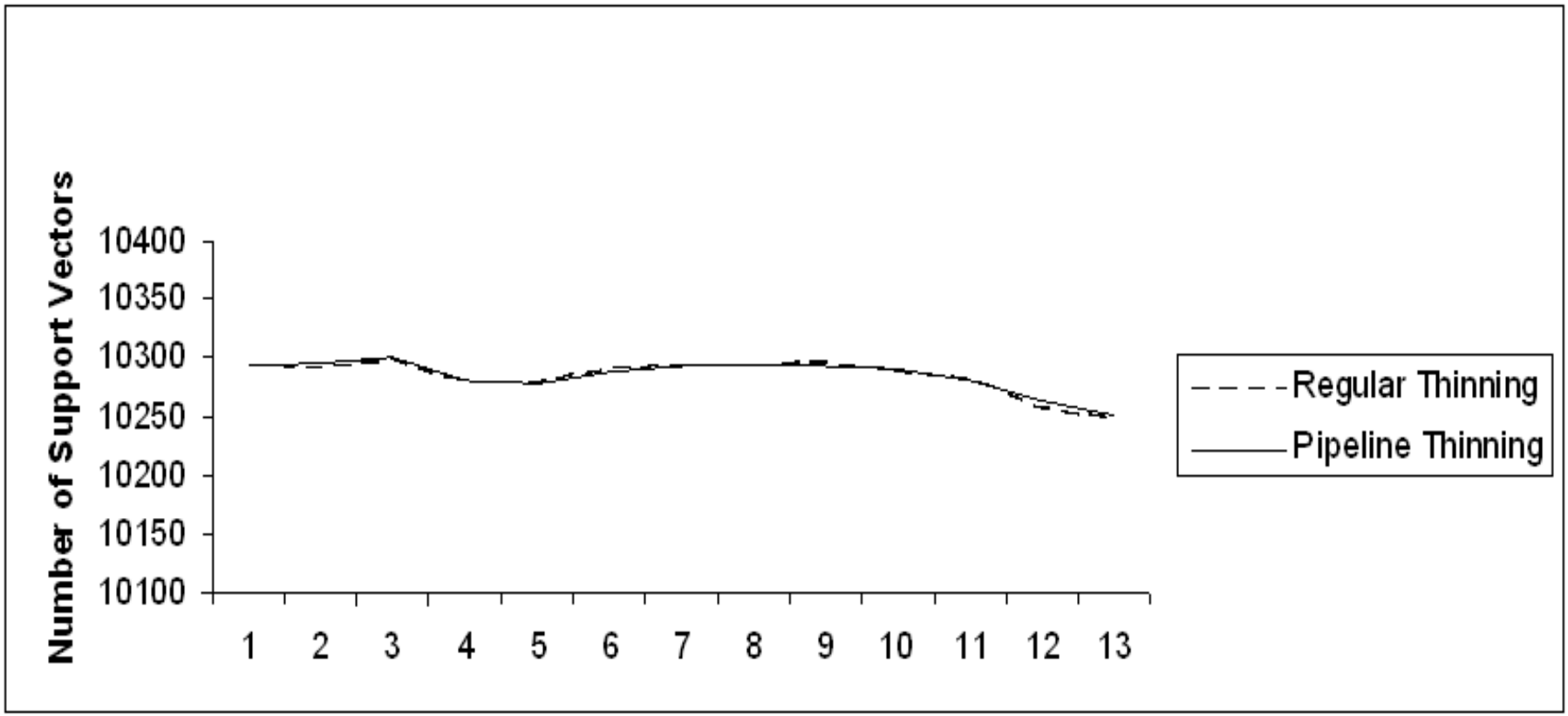


On-Line Thinning
-Results: MAE (m/s)-

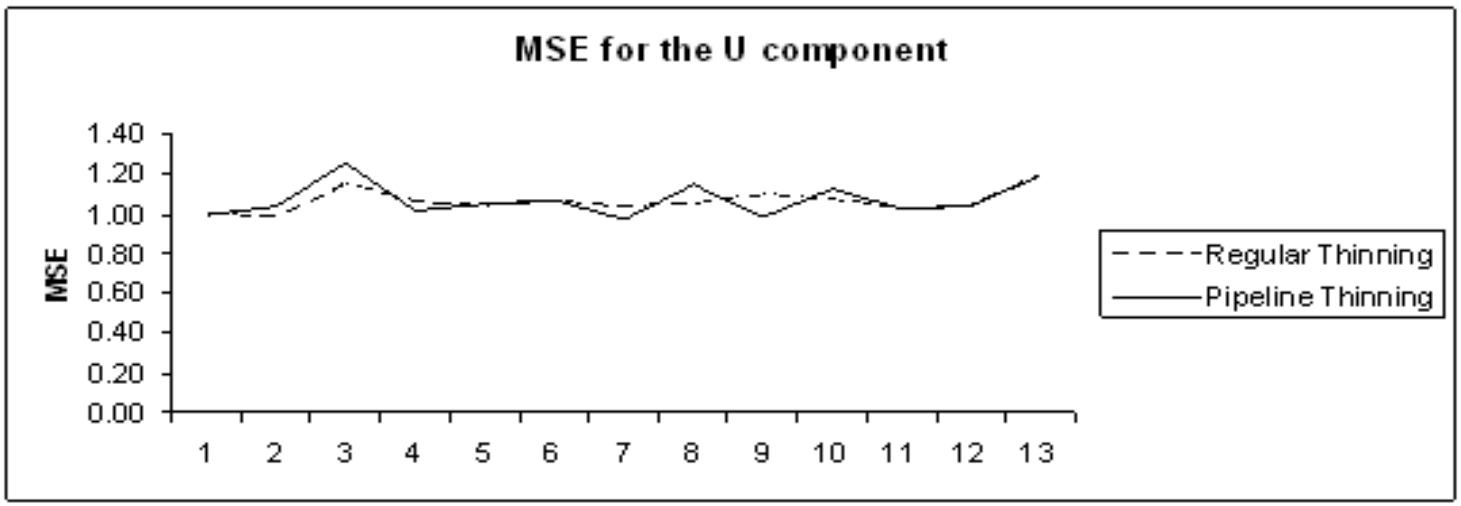


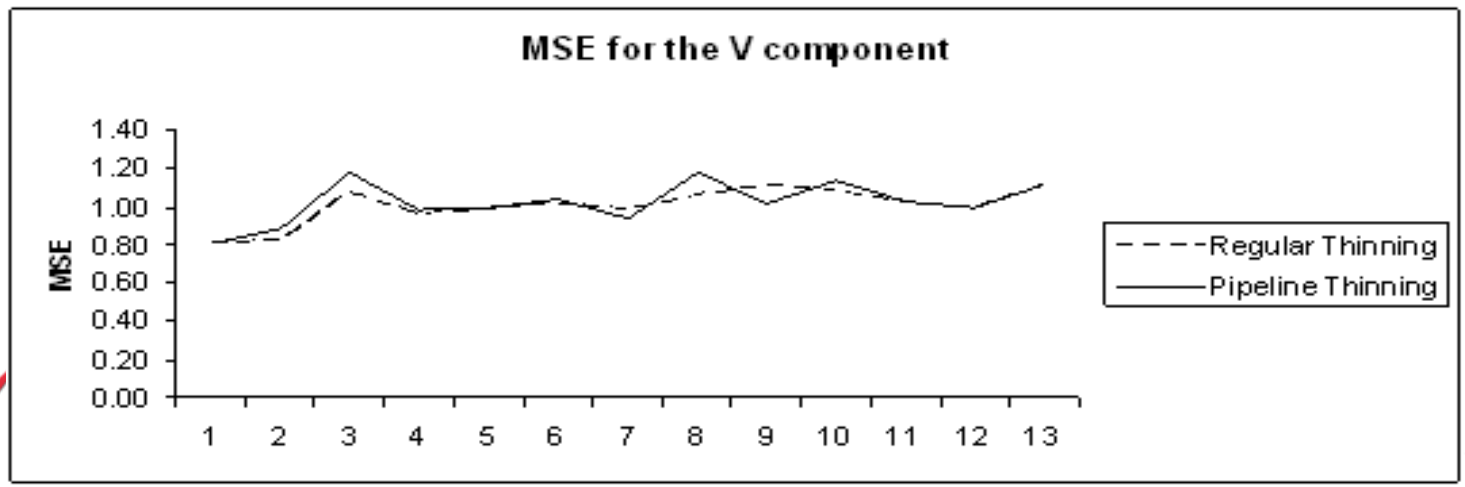


On-Line Thinning
-Results: Thinning Percentage-

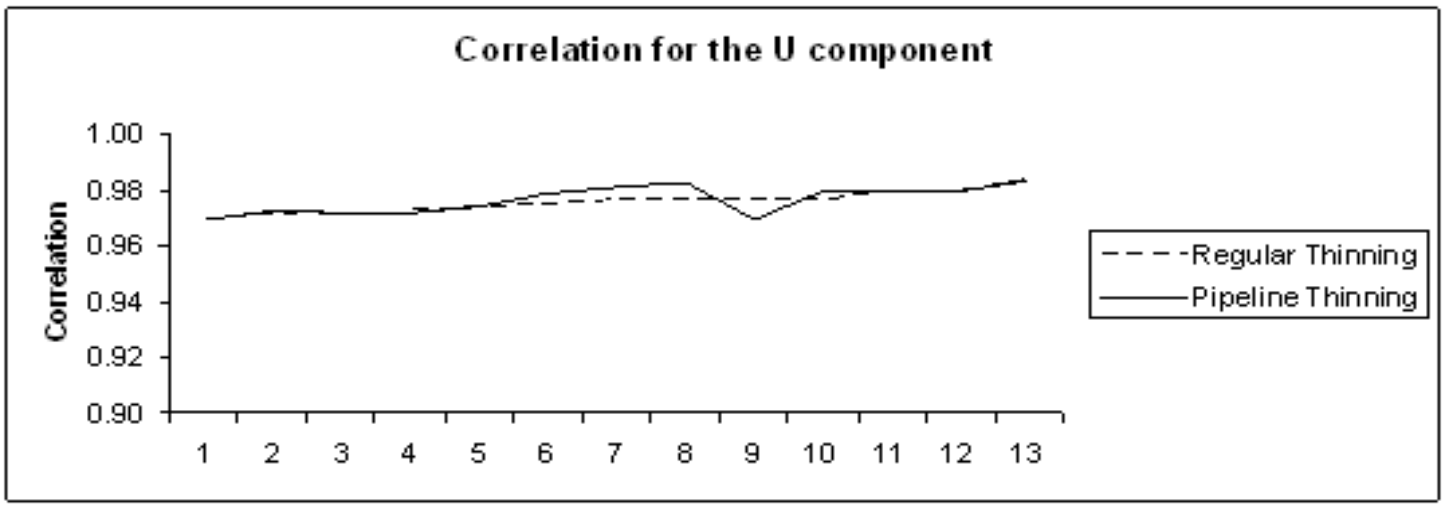


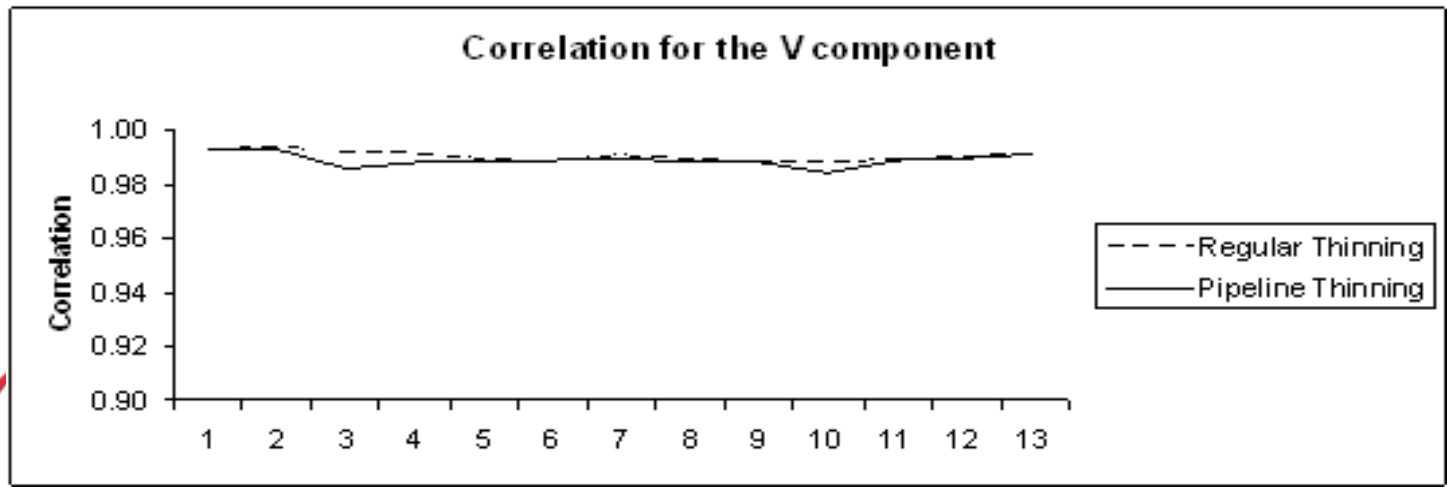
On-Line Thinning
-Results: MSE (m^2/s^2)-





On-Line Thinning
-Results: Correlation-





Conclusion

- Only 10.4% of the data were needed to reconstruct the wind field with high accuracy (correlation coefficients are > .99 for the u- and the v-components). To obtain the same accuracy as SVR, the thinning rate of the other superobbing methods was 26.3% and the computation time was over four times that of SVR
- After the first period, the time needed to thin using a pipeline is around 9% of the time needed to thin using the whole data set. The pipeline yielded subsets that had the same accuracy as thinning the whole data set

Future work

- Thinned data will be assimilated and their impact on model forecasts assessed in collaboration with Dr. John LeMarshall of the Australian Bureau of Meteorology and scientists at the University of Wisconsin

Publications from the research

- **Mansouri, H., Trafalis, T.B., Gilbert, R., Leslie, L.M., Richman, M.B., "Ocean Surface Wind Vector Forecasting Using Support Vector Regression", *Intelligent Engineering Systems Through Artificial Neural Networks,* 17 (2007): 333-338.**

- **Mansouri, H., Richman, M.B., Trafalis, T.B., Leslie, L.M., "A Pipeline Support Vector Regression Method to Thin Large Sea Data On-Line", accepted at *Intelligent Engineering Systems Through Artificial Neural Networks* (2008).**

- **Mansouri, H., Leslie, L.M., Trafalis, T.B., Richman, M.B., "Thinning Satellite Data Using Support Vector Regression", In progress.**

Thinning WindSat Data Using Support Vector Regression

PIs and Co-PIs: M. Richman and L. Leslie

NWP Center Collaborators: JCSDA

Accomplishments

- **Only 10.4% of the data were needed to reconstruct the wind field with high accuracy (correlation coefficients are > .99 for the u- and the v-components)**
- **To obtain the same accuracy as SVR, the thinning rate of the other superobbing methods was 26.3% and the computation time was over four times that of SVR.**

Future Plan

Thinned data will be assimilated and their impact on model forecasts assessed in collaboration with Dr. John LeMarshall and scientists at the University of Wisconsin.