ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 32, MARCH 2015, 277–286

Data Selection Using Support Vector Regression

Michael B. RICHMAN*[1], Lance M. LESLIE[1], Theodore B. TRAFALIS[2], and Hicham MANSOURI[3]

[1]*School of Meteorology and Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma,* 73072*, USA*

[2]*School of Industrial and Systems Engineering, University of Oklahoma, Norman, Oklahoma,* 73019*, USA*

[3]*Power Costs, Inc.,* 301 *David L. Boren Blvd., Suite* 2000*, Norman, Oklahoma* 73072*, USA*

(Received 17 April 2014; revised 11 September 2014; accepted 18 September 2014)

ABSTRACT

Geophysical data sets are growing at an ever-increasing rate, requiring computationally efficient data selection (thinning) methods to preserve essential information. Satellites, such as WindSat, provide large data sets for assessing the accuracy and computational efficiency of data selection techniques. A new data thinning technique, based on support vector regression (SVR), is developed and tested. To manage large on-line satellite data streams, observations from WindSat are formed into subsets by Voronoi tessellation and then each is thinned by SVR (TSVR). Three experiments are performed. The first confirms the viability of TSVR for a relatively small sample, comparing it to several commonly used data thinning methods (random selection, averaging and Barnes filtering), producing a 10% thinning rate (90% data reduction), low mean absolute errors (MAE) and large correlations with the original data. A second experiment, using a larger dataset, shows TSVR retrievals with MAE < 1 m s^{-1} and correlations $\geqslant 0.98$. TSVR was an order of magnitude faster than the commonly used thinning methods. A third experiment applies a two-stage pipeline to TSVR, to accommodate online data. The pipeline subsets reconstruct the wind field with the same accuracy as the second experiment, is an order of magnitude faster than the non-pipeline TSVR. Therefore, pipeline TSVR is two orders of magnitude faster than commonly used thinning methods that ingest the entire data set. This study demonstrates that TSVR pipeline thinning is an accurate and computationally efficient alternative to commonly used data selection techniques.

Key words: data selection, data thinning, machine learning, support vector regression, Voronoi tessellation, pipeline methods

Citation: Richman, M. B., L. M. Leslie, T. B. Trafalis, and H. Mansouri, 2015: Data selection using support vector regression. *Adv. Atmos. Sci.*, **32**(3), 277–286, doi: 10.1007/s00376-014-4072-9.

1. Introduction

The quantity of geophysical data is increasing at a rapid rate. Hence, it is essential to identify and/or select features that preserve relevant information in the data. Data selection has as its two main aims the removal of redundant and faulty data. Here, the emphasis is on redundant data, so the terms data selection and data thinning will be used interchangeably. Redundant data arise from two main sources: when the data density is greater than the spatial and temporal resolution of the analysis grid and when the data are not linearly independent. Penalties for retaining redundant data are the (possibly massive) increase in computational cost, the failure to satisfy key assumptions of the data analysis scheme (Lorenc, 1981) and the increased risk of overfitting (particularly for problems with high dimensions).

The need for data selection is exemplified by satellite observations. Satellites are among the most important contributors of observations to the data selection process and, hence, to the analysis. Notably, satellites provide high-resolution observations over data poor regions, especially the oceans and sparsely populated land areas. Historically, data redundancy issues led to the development of data selection approaches that were simple and cost effective. These included: allocating the observations to geographical grid boxes and then averaging the data in each box to produce so-called super-observations, or "superobs" (Lorenc, 1981; Purser et al., 2000); the selection of observations, in both meridional and zonal directions, with random sampling of the observations (Bondarenko et al., 2007); and the use of filters, such as the Barnes scheme (Barnes, 1964). Owing to their simplicity, and because they are non-adaptive, such strategies are referred to as unintelligent data selection techniques. For example, they do not specify targeted areas of interest or weight the data according to their contribution to minimizing differences between the thinned and non-thinned data.

Recently, various intelligent data selection strategies have emerged (e.g., Lazarus et al., 2010). Such approaches are effective in identifying and removing redundant data and have

* Corresponding author: Michael B. RICHMAN
Email: mrichman@ou.edu

other desirable features. One example is the Density Adjusted Data Thinning (DADT; Ochotta et al., 2005; 2007), and its successor, the modified DADT (mDADT; Lazarus et al., 2010). The intelligent data selection schemes are adaptive, as they attempt to retain those observations that are less highly correlated with other observations, but contribute more significantly to the retention of the information content in the observations (e.g., they employ metrics based on gradients and/or curvature of the fields). Intelligent data selection schemes usually require definitions of redundancy measures, and their sampling strategies iteratively remove observations that fail to meet the metric threshold criteria.

The present work develops an entirely different, kernel-based, intelligent data selection technique using Support Vector Machines (SVMs). SVMs require neither *a priori* specification of metrics nor of thinning rates. SVMs are alternatives to artificial neural networks, decision trees and Bayesian networks for classification and prediction tasks (Schölkopf and Smola, 2002) used in supervised learning, such as statistical classification and regression analysis. Although SVMs were introduced several decades ago (Vapnik, 1982), they have been investigated extensively by the machine learning community only since the mid-1990s (Shawe-Taylor and Cristianini, 2004).

SVMs require solving a quadratic programming problem with linear constraints. Therefore, the speed of the algorithm is a function of the number of observations (data points) used during the training period. Hence, the SVM solution to problems comprised of numerous data points is computationally inefficient. Several methods have been proposed to ameliorate this problem. Platt (1999) applied Sequential Minimal Optimization (SMO), to break the large quadratic programming problem into a series of smallest analytically solvable problems. A faster SMO SVM algorithm, advantageous for real-time or online prediction or classification for large scale problems, was suggested by Bottou and LeCun (2004). Musicant and Mangasarian (2000) applied a linear program SVM method to accommodate very large datasets. Bakır et al. (2004) selectively removed data using probabilistic estimates, without modifying the location of the decision boundary. Other techniques used online training to reduce the impact of large data sets. Bottou and LeCun (2005) showed that performing a single epoch of an online algorithm converges to the solution of the learning problem. Laskov et al. (2006) develop incremental SVM learning with the aim of providing a fast, numerically stable and robust implementation. Support Vector Regression (SVR) uses the kernel approach from SVM to replace the inner product in regression. It is discussed extensively by Smola and Schölkopf (1998). SVM techniques have been applied to small-scale meteorological applications, such as rainfall and diagnostic analysis fields supporting tornado outbreaks. These include the studies of Son et al. (2005), Santosa et al. (2005), Trafalis et al. (2005), and in satellite data retrievals, by Wei and Roan (2012). The present study seeks to further enhance SVR in two respects: (1) by applying a Voronoi tessellation (Bowyer, 1981) to reduce the size of the large observational data sets and, (2) adopting a pipeline methodology (Quinn, 2004) to improve the computational efficiency of the data selection scheme.

In section 2, large-scale problems using satellite datasets are described. In section 3, it is shown how Voronoi tessellation reduces the size of the large observational data sets, and how a pipeline SVM methodology substantially enhances the computational efficiency of the data selection scheme. The results are presented in section 4. Finally, conclusions are discussed in section 5.

2. Data

This study employs data from the WindSat microwave polarimetric radiometry sensor (Gaiser et al., 2004). WindSat provides environmental data products, including latitude, longitude, cloud liquid water, column integrated precipitable water, rain rate, and sea surface temperature. WindSat measurements over the ocean are used operationally to generate analysis fields and also as input to numerical weather prediction models of the U.S. Navy, the U.S. National Oceanic and Atmospheric Administration (NOAA) and the United Kingdom Meteorological Office. As a polarimetric radiometer, WindSat measures not only the principal polarizations (vertical and horizontal), but also the cross-correlation of the vertical and horizontal polarizations. The cross-correlation terms represent the third and fourth parameters of the modified Stokes vector (Gaiser et al., 2004). The Stokes vector provides a full characterization of the electromagnetic signature of the ocean surface and the independent information needed to uniquely determine the wind direction (Chang et al., 1997).

To illustrate the data selection procedure introduced herein, it suffices to explore a single data type, namely, sea surface wind (SSW) speeds and directions. For SSW data, it is necessary to account not only for random errors but also for spatially correlated errors. Typical ascending swaths for a 24 hour sample of WindSat data provide ~ 1.5 million observations. Given this massive number of data points, oversampling of wind data can severely degrade the analysis and, consequently, the model forecasts.

Three experiments were carried out using different WindSat datasets. The first experiment was designed to assess, on a relatively small sample, the accuracy and computational efficiency of a Voronoi tessellation followed by SVR to thin the WindSat data. Hereafter, this sequential combination of Voronoi tessellation followed by SVR will be referred to "TSVR". Two hours of WindSat data from 1 January 2005 were chosen in the region 127°W to 145°E longitude and 23° to 42°N latitude, providing 13 540 observations for the data selection process. Additionally, TSVR was compared to three commonly used data thinning techniques (simple averaging, random selection and a Barnes filter) to assess the relative accuracy and computational efficiency of each method. A second experiment used 226393 observations to determine if the accuracy and computational efficiency gains by TSVR were preserved with a much larger dataset. The third experi-

ment employs a pipeline methodology (section 3.3) as it has been employed successfully to achieve much higher computational efficiency (e.g., Ragothaman et al., 2014). Such an approach is expected to enhance real-time processing of an on-line stream of WindSat data.

3. Learning Machine Methodologies

3.1. *Voronoi Tessellation*

Experiments show that the standard SVR algorithm loses computational efficiency when analyzing more than several thousand observations (Platt, 1999). Since the WindSat data sets used in this study are in excess of this, and can exceed 10^6 observations, direct application of SVR is not feasible. Methods have been proposed to reduce this problem (e.g., Platt, 1999; Musicant and Mangasarian, 2000). Voronoi tessellation partitions a plane with p points into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other. The cells are called polytopes (e.g., Voronoi polygons). They were employed by Voronoi (1908) and have been applied in diverse fields, such as computer graphics, epidemiology, geology, and meteorology. As shown in Fig. 1, the tessellation is achieved by allocating the data points to a number of Voronoi cells (Du et al., 1999; Mansouri et al., 2007; Gilbert and Trafalis, 2009; Helms and Hart, 2013). The process uses the Matlab "voronoi" function (Matlab, 2012).

As mentioned above, for a discrete set, S, of points in $\Re^n$ and for any point $\boldsymbol{x}$, there is one point of S closest to $\boldsymbol{x}$. More formally, let X be a space (and S a nonempty subset of X) provided with a distance function, d. Let C, a nonempty subset of X, be a set of p centroids $(\boldsymbol{P}_c), c \in [1,p]$. The Voronoi cell, or Voronoi region, V_c, associated with the centroid $\boldsymbol{P}_c$ is the set of all points in X whose distance to $\boldsymbol{P}_c$ is not greater than their distance to the other centroids, $\boldsymbol{P}_j$, where j is any index different from c. That is if $D(\boldsymbol{x},A) = \inf\{d(\boldsymbol{x},\boldsymbol{a})|\boldsymbol{a} \in A\}$ denotes the distance between the point x and the subset A, then $V_c = \{\boldsymbol{x} \in X | d(\boldsymbol{x},\boldsymbol{P}_c) \leqslant d(\boldsymbol{x},\boldsymbol{P}_j), \text{ for all } j \neq c\}$.

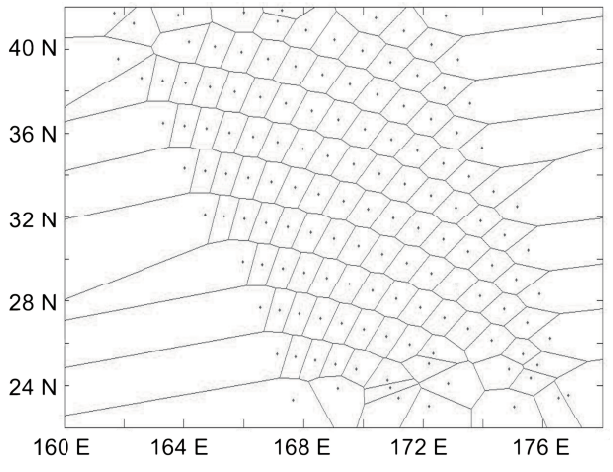

Fig. 1. Voronoi tessellation for a subset of data over the Pacific Ocean.

In general, the set of all points closer to $\boldsymbol{P}_c$, than to any other point of S, is called the Voronoi cell for $\boldsymbol{P}_c$. The set of such polytopes is the Voronoi tessellation corresponding to the set S. In two dimensional space, a Voronoi tessellation can be represented as shown in Fig. 1. Since the number of data points inside each Voronoi polygon is much less than for the full data set, the computational time is reduced greatly. Moreover, further efficiency can be gained by using parallel computing, solving a set of Voronoi polygons simultaneously.

3.2. *Support Vector Regression*

In SVR, it is assumed that there is a data source providing a sequence of l observations and no distributional assumptions are made. Each observation (data point) is represented as a vector with a finite number n of continuous and/or discrete variables that can be denoted as a point in the Euclidean space, $\Re^n$. Hence, the l observations are data points in the Euclidean space $\Re^n$.

The l observations are divided into p cells using Voronoi tessellation. The methodology consists of making each kth observation a seed or "centroid" for a Voronoi cell $V_c, \forall c \in [1,p]$. The parameter k is set such that $p = \lfloor 1/k \rfloor$. Hence, for a larger k, fewer cells will be generated. Each cell V_c will be composed of data points represented by $\boldsymbol{x}_{i,c} \in \Re^n, \forall i \in [1,l]$. In regression problems, each observation $\boldsymbol{x}_{i,c}$ is related to a unique real valued scalar target denoted by $y_{i,c}$. The couplets $(\boldsymbol{x}_{i,c}, y_{i,c})$ in $\Re^{n+1}$ are a set of points that have a continuous unknown shape that is not assumed to follow a known distribution. The objective of support vector regression (SVR) is to find a machine learning prediction function (in our application, this is an estimation at a particular time, t, rather than a forecast at time $t + \Delta t$), denoted by f_c for each cell V_c such that the differences between $f_c(\boldsymbol{x}_{i,c})$ and the target values, $y_{i,c}$, are minimized.

In the present study, the target is either the u-or the v-component of the winds. By introducing, for each observation $\boldsymbol{x}_{i,c}$, a set of positive slack variables, $\xi_{i,c}$, which are minimized, the following set of constraints for the regression problems are generated for each cell V_c:

$$\begin{cases} |f_c(\boldsymbol{x}_{i,c}) - y_{i,c}| \leqslant \xi_{i,c} & \forall i \in [1,l] \\ \xi_{i,c} \geqslant 0 & \forall i \in [1,l] \end{cases} . \tag{1}$$

For linear regression, in the SVM literature, f_c belongs to a class of functions denoted by F, such that:

$$F := \{\boldsymbol{x} \in \Re^n \mapsto \langle \boldsymbol{w}_c \cdot \boldsymbol{x} \rangle + b_c, \|\boldsymbol{w}_c\| \leqslant B_c\} , \tag{2}$$

where b_c is the bias term, $B_c > 0$ is a constant that bounds the weight space, $\boldsymbol{w}_c = \sum_{j=1}^{l} \alpha_{j,c}, \boldsymbol{x}_{j,c}$, and $\alpha_{j,c} \in \Re \forall j \in [1,l]$.

In the case of nonlinear regression, the class of functions, F, is changed to allow for linear regression in Hilbert space to where the observations $\boldsymbol{x}_{i,c}$ will be mapped. This is achieved by introducing a nonnegative definite kernel $k : \Re^n \times \Re^n \rightarrow \Re$, to induce a new Hilbert space H and map $\varphi : \Re^n \rightarrow H$ such

that $k(\boldsymbol{x},\boldsymbol{y}) = \langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y}) \rangle_H$ for any $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\Re^n$. Hence, F becomes:

$$F := \{\mathbf{x} \in \Re^n \mapsto \langle \boldsymbol{w}_c \varphi(\boldsymbol{x}) \rangle_H + b_c \| \boldsymbol{w}_c \|_H \leqslant B_c\} \ , \quad (3)$$

where $\boldsymbol{w}_c = \sum_{j=1}^{l} \alpha_{j,c} \varphi(\boldsymbol{x}_{j,c})$, and $\alpha_{j,c} \in \Re \forall j \in [1,l]$. Explicit knowledge of H and φ is not required. Therefore, the set of constraints Eq. (1) becomes:

$$\begin{cases} \left| \sum_{j=1}^{l} \alpha_{j,c} k(\boldsymbol{x}_{j,c}, \boldsymbol{x}_{i,c}) + b_c - y_{i,c} \right| \leqslant \xi_{i,c} & \forall i \in [1,l] \\ \xi_{i,c} \geqslant 0 & \forall i \in [1,l] \end{cases} . \quad (4)$$

SVM allows for an objective function that reduces the slack variables and the expected value of $|f_c(\boldsymbol{x}_{i,c}) - y_{i,c}|$. To achieve that objective, minimize the quantities b_c, $\xi_{i,c}$, and $\|\boldsymbol{w}_c\|_H$.

Thus, $\|\boldsymbol{w}_c\|_H^2 = \langle \sum_{j=1}^{l} \alpha_{j,c} \varphi(\boldsymbol{x}_{j,c}) \cdot \sum_{j=1}^{l} \alpha_{j,c} \varphi(\boldsymbol{x}_{j,c}) \rangle_H = \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_{i,c} \alpha_{j,c} \langle \varphi(\boldsymbol{x}_{i,c}) \cdot \varphi(\boldsymbol{x}_{j,c}) \rangle_H = \boldsymbol{\alpha}_c^{\mathrm{T}} \boldsymbol{K}_c \boldsymbol{\alpha}_c$, where $(\boldsymbol{K}_c)_{ij} = k(\boldsymbol{x}_{i,c}, \boldsymbol{x}_{j,c})$. The quadratic problem to be solved is:

$$\begin{array}{l} \min_{\alpha_c, \xi_c, b_c} \boldsymbol{\alpha}_c^{\mathrm{T}} \boldsymbol{K}_c \boldsymbol{\alpha}_c + C \boldsymbol{\xi}_c^{\mathrm{T}} \boldsymbol{\xi}_c + b_c^2 \\ \text{subject to : } |\boldsymbol{K}_c \boldsymbol{\alpha}_c + b_c \boldsymbol{1} - \boldsymbol{y}_c| \leqslant \boldsymbol{\xi}_c \end{array} , \quad (5)$$

where $C > 0$ is a positive trade-off constant that penalizes the non-zero values of the $\xi_{i,c}$, $\boldsymbol{1}$ is a $l \times 1$ vector of ones, and $\boldsymbol{y}_c$ is the vector with elements $y_{i,c}$.

The optimal solution $(\boldsymbol{\alpha}_c^* b_c^*)$ of Eq. (5) yields the following prediction function:

$$f_c : \boldsymbol{x} \mapsto \sum_{i=1}^{l} \alpha_{i,c}^* k(\boldsymbol{x}_{i,c}, \boldsymbol{x}) + b_c^* \ . \quad (6)$$

The vectors for $\boldsymbol{x}_{i,c}$ which the values of $\alpha_{i,c}$ are nonzero are called support vectors.

From Eq. (3) a kernel is required. In this work, several kernels were tested for their ability to select a smaller number of observations with a minimum loss of information. Those tested were:
the linear kernel,

$$k(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x} \ , \quad (7)$$

the radial basis function kernel (RBF),

$$k(\boldsymbol{x}_i, \boldsymbol{x}) = e^{-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{\sigma^2}} \ , \quad (8)$$

the polynomial kernel of degree q,

$$k(\boldsymbol{x}_i, \boldsymbol{x}) = \left(1 + \frac{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}}{g}\right)^q \ , \quad (9)$$

and the sigmoidal kernel,

$$k(\boldsymbol{x}_i, \boldsymbol{x}) = \tanh(a(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}) + \theta) \ , \quad (10)$$

where σ, g, a and θ are scaling constants.

3.3. *Pipeline TSVR*

To improve the efficiency of the TSVR, a pipeline methodology (Quinn, 2004) is introduced to allow for an online stream of meteorological satellite data. The pipeline approach is appropriate for such data because the satellite samples a swath of new wind data as it orbits. Within each Voronoi polygon, the pipeline is applied to the variables used to estimate the winds by TSVR. A two-stage pipeline (with 50% overlap as shown in Fig. 2) is applied that fetches and preprocesses new data as old data is executing in the CPU. Figure 2 illustrates the pipeline, showing that the orbital swath is divided into discrete steps and how these new data are incorporated into the TSVR process. Figure 2 shows the pipeline window of width of four CPU time units, ingesting the data set. At each step, the most recent data are included in the window, while the oldest data are released. Next, the window moves to the right by one-half step. Hence, instead of thinning all the data within a window, the cells outside the window are dropped and new Voronoi cells are formed that contain only the new data. If this overlapping approach were not adopted, the data would have to be ingested, preprocessed and analyzed prior to moving on to the next batch of data, thereby reducing the efficiency of the process.

3.4. *Measures of differences between non-thinned and thinned data*

Mean squared differences (commonly referred to as MSE), mean absolute differences (MAE), as well as the correlation between the original (non-thinned) and thinned satellite observed winds are employed to measure the quality of the thinned observations. MSE, MAE and correlations are defined in Wilks (2011). These are commonly applied metrics to measure differences between two fields.

4. Results

4.1. *Results of the first experiment*

The main objective of this experiment is to assess the feasibility of the TSVR, and to determine the most effective kernel, using a small sample (13 540 observations) of WindSat data. Support vectors are used for the reproduction of the wind field after data selection. Because of the intelligent adaptive capability of the TSVR, fewer than 8% of the observed satellite data were needed to reconstruct the wind field. To quantify the accuracy of the reconstructed winds using TSVR, the thinned winds are compared to the non-thinned observations. From Eq. (3), a kernel must be selected to generate the support vectors and reconstructs the wind fields. Table 1 shows metrics (MSE, MAE and correlations) for the kernels defined in section 3.1. The various

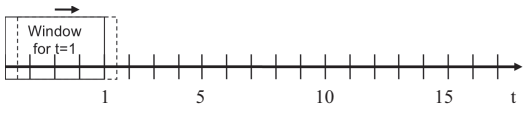

Fig. 2. Pipeline thinning showing the moving data window.

Table 1. MAE, MSE and correlation metrics comparing the differences between observed and thinned *u*- and *v*-components of wind for different SVR kernels. The kernel selected (RBF 1) is in bold font.

	u-component			*v*-component		
	MAE (m s^{-1})	MSE (m^2 s^{-2})	Correlation	MAE (m s^{-1})	MSE (m^2 s^{-2})	Correlation
RBF .5	1.25	8.27	0.90	0.88	3.07	0.98
RBF 1	**1.15**	**5.99**	**0.91**	**0.72**	**2.71**	**0.98**
RBF 5	1.30	6.72	0.90	0.99	3.29	0.98
RBF 10	1.48	8.32	0.87	1.09	3.71	0.98
RBF 20	1.60	9.30	0.85	1.22	4.25	0.98
RBF 50	1.86	10.95	0.82	1.62	5.94	0.97
RBF 100	2.03	12.41	0.80	1.89	7.34	0.96
Linear	2.06	12.84	0.79	1.96	7.82	0.95
Poly 12	2.06	12.82	0.79	1.96	7.81	0.95
Poly 13	2.05	12.80	0.79	1.96	7.80	0.95
Sig 11	2.06	12.85	0.79	1.96	7.82	0.95

kernels tested were: linear; seven radial basis functions with the σ parameter varying from 0.5 to 100; polynomials with $g = 1$ and of orders (q) 2 and 3; and sigmoidal with the two scale parameters (a, θ) set to 1. The smallest differences between thinned and non-thinned wind data were obtained for the RBF kernel, with a *u*-component MAE (MSE) of 1.05 m s^{-1} (5.99 m^2 s^{-2}), which are 44% (53%) reductions in the discrepancies, respectively, obtained from any non-RBF kernel. For the *v*-component, the corresponding reductions for the RBF kernel, compared to a non-RBF kernel, were even larger at 63% (65%). The variances explained (correlations squared) are 82.8% and 96.0% for the *u*-and *v*-components, representing improvements of 33% and 6%, respectively, over any non-RBF kernel. Therefore, the RBF kernel with parameter 1 is used for all subsequent TSVR analyses.

Figure 3 shows frequency counts of the reconstructed wind errors for the 13540 observations thinned by TSVR. For the *u*-component (Fig. 3a), 77% (87%) of the discrepancies of the magnitudes are $\leqslant$ 1 m s^{-1} (2 m s^{-1}), which is at or below the accepted observation error for these data (Quilfen et al., 2007). Similar discrepancies were found for the *v*-component (Fig. 3b). Both distributions are highly leptokurtic, illustrating the efficacy of TSVR. Figure 4 presents the thinned (Figs. 4a, c) and non-thinned (Figs. 4b, d) satellite wind field contours for the *u*- and *v*-components. The close spatial correspondence of the patterns for each component is consistent with the large positive correlations in Table 1 for the RBF 1 kernel.

For the present problem, most of the support vectors have alpha values near zero (Fig. 5), thus they have an insignificant contribution to the final solution. From Eq. (6), those support vectors with zero or near-zero alpha values are ignored, providing further data reduction. For the present analysis, Figure 5 illustrates the large data reduction capability of SVR for these data. From the available 13540 data points, only $\sim$ 1000 support vectors ($<$ 8%) are required to reconstruct the wind vector field with the aforementioned high level of accuracy. Specifically, for each Voronoi cell, the satellite data points inside each cell are used to train the SVR. Fewer than 8% of the observations were support vectors and are retained; therefore, the thinning rate is $>$ 92%. The $<$ 8% support vectors had an MAE of 0, the MAE of the other $>$ 92% data

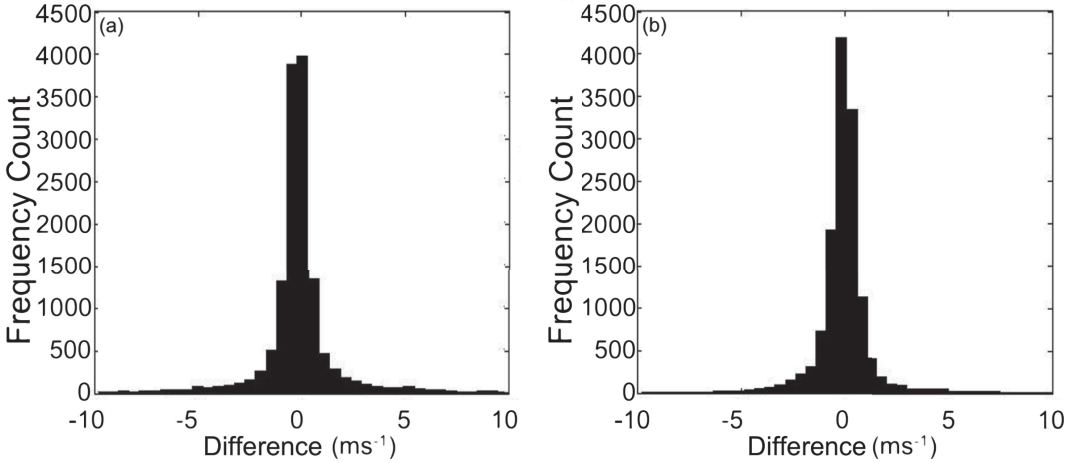

Fig. 3. Frequency counts of the wind speed discrepancies (m s^{-1}) between the original non-thinned data and the thinned data (a) for the *u*-component and (b) for the *v*-component of the sea surface winds.

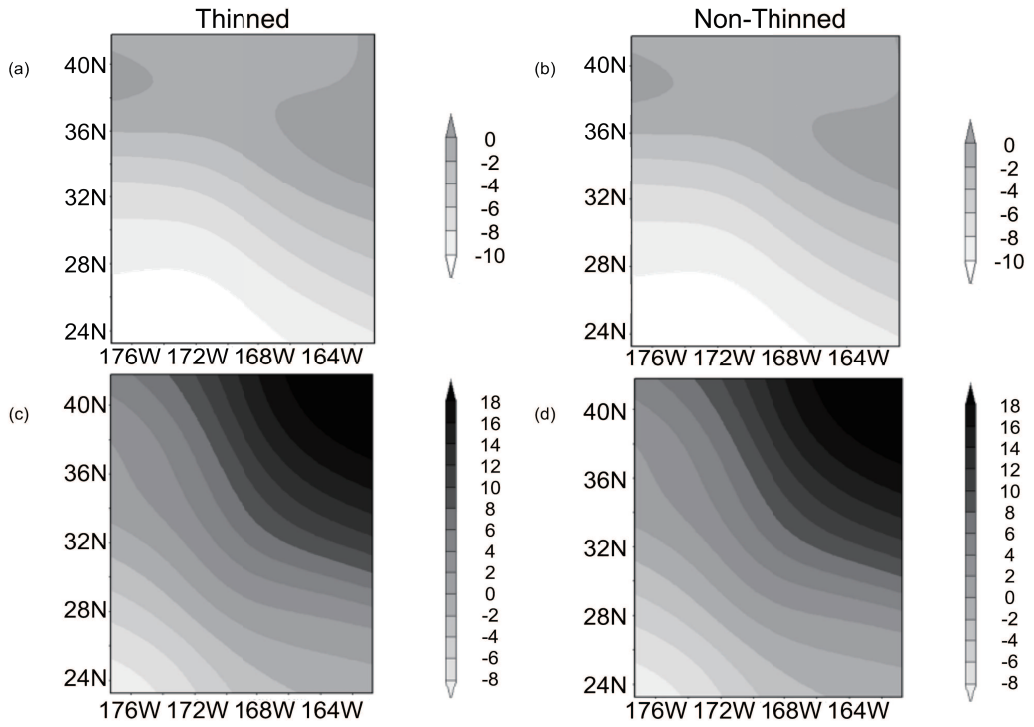

Fig. 4. Contour maps of the *u*-and *v*-components (in m s^{-1}) of the (a, c) thinned and (b, d) non-thinned wind fields.

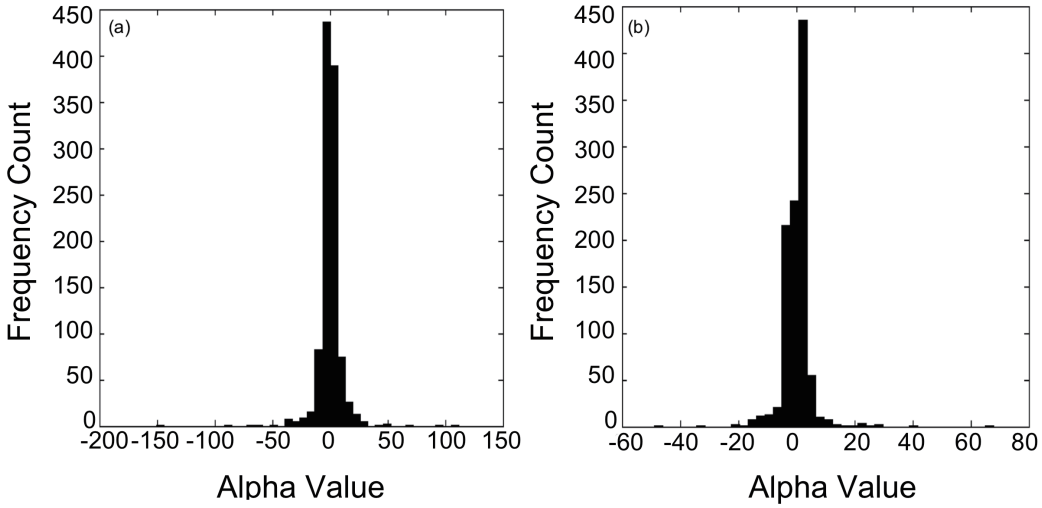

Fig. 5. Distributions of the alpha values for the support vectors of (a) the *u*-component and (b) the *v*-component of the winds.

points was calculated using only $< 8\%$ of the support vectors. Since the percentage of support vectors is a function of the complexity of the data field, it will vary according to the spatial and temporal data structure.

4.2. *Results of the second experiment*

Given the large data reduction and high level of accuracy in reproducing the wind fields provided by TSVR, as found in section 4.1, a considerably larger sample (226393 data points) was drawn to assess the scalability of TSVR and to compare it to several commonly used data thinning techniques. For these commonly used techniques, the observations were assigned to cells of *h* degrees latitude and longitude. For random sampling, a single observation was selected. For the other schemes, all data were used. The accuracy of these data selection methods is shown in Figs. 6a-d (MAE, MSE) and Fig. 7 (correlation). The MAE for the *u*-component (Fig. 6a) shows that, as the width of the data cells decreases, the discrepancies decrease for both averaging and random selection. The accuracy of Barnes filtering improves

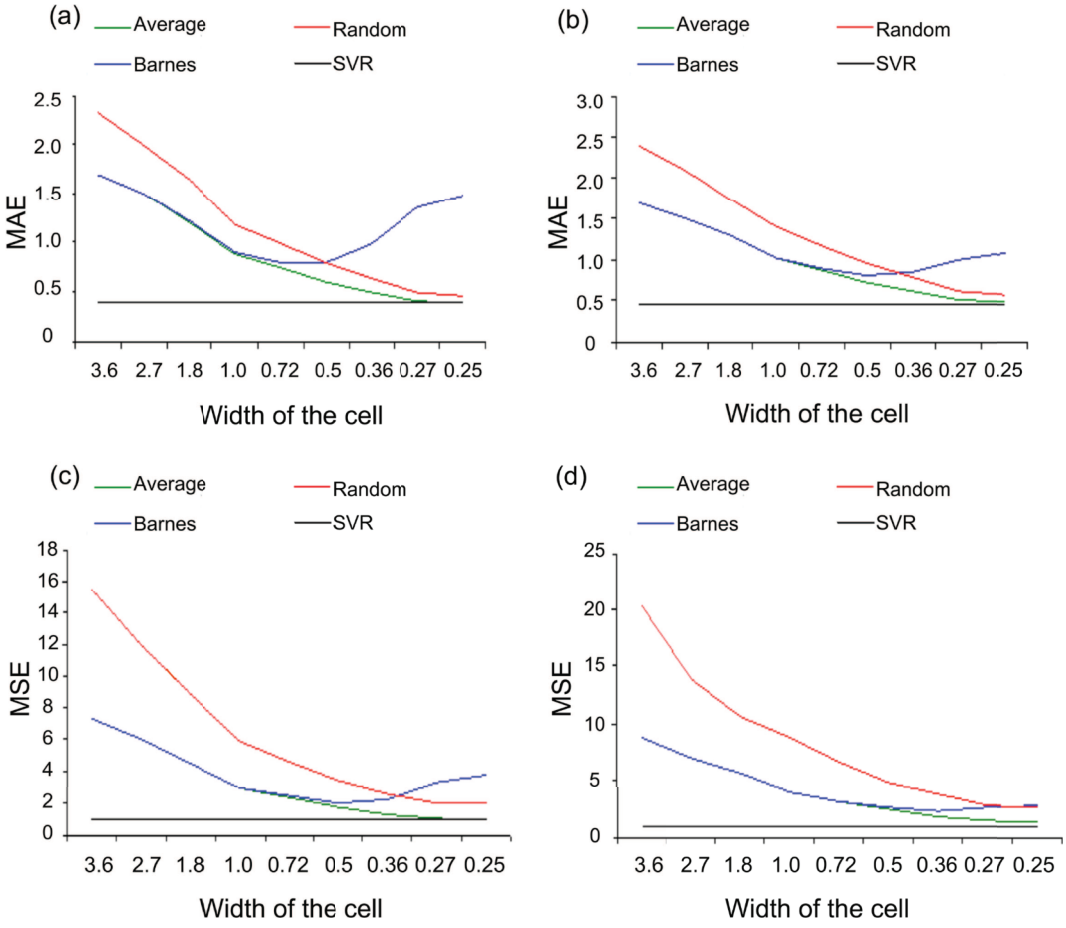

Fig. 6. Mean absolute differences (MAE) and mean squared differences (MSE) between the thinned and non-thinned *u*-components (a, c) and *v*-component (b, d) of the wind (in m s^{-1}) for the averaging, random, Barnes and TSVR thinning methods.

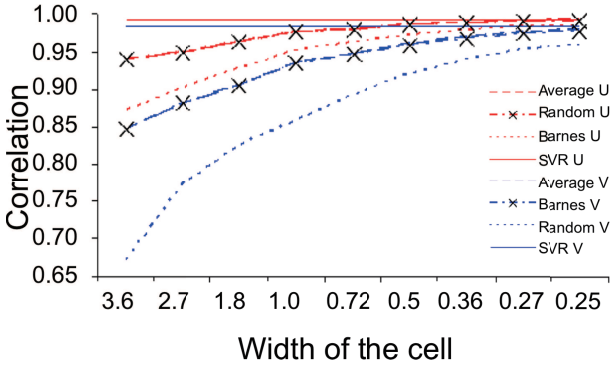

Fig. 7. Correlations between the thinned and non-thinned data for the averaging, random, Barnes and TSVR thinning methods.

as the cells decrease in size and reaches a minimum at a cell width of approximately 0.7 degrees; beyond that, insufficient data density produces increasingly inaccurate results. As the Voronoi tessellation is applied to TSVR, the cells do not change and hence the accuracy remains constant. For the *v*-component (Fig. 6b), similar behavior is noted for all techniques. TSVR is the most accurate thinning technique with MAE ~ 0.5 m s^{-1}. The MSE values (Figs. 6c, d) are larger than the corresponding MAE values; however, the ranking of the techniques remains the same, with the random sampling being least accurate, averaging and Barnes giving similar results and the TSVR producing the most accurate thinning. The correlation between the thinned and non-thinned winds is calculated for the same data selection methods (Fig. 7). As the cell width decreases, the correlations for the u-components, given by the three commonly used techniques move closer to the TSVR value, but never exceed it. Despite these large correlations at small cell widths, the larger MAE and MSE of the three commonly used techniques indicate less accurate thinning for those methods. The *v*-component correlations for the other methods are considerably lower than those for TSVR (Fig. 7). Moreover, the high correlations obtained with the three commonly used data selection methods is achieved at the expense of a loss of computational efficiency (Fig. 8), as the TSVR requires approximately 250 seconds to thin these data at the aforementioned accuracy (correlation of 0.99 and 0.98 for the TSVR) versus over 1000 seconds for the other three techniques. For this experiment, the percentage of data required to obtain this level of accuracy for the TSVR is $\sim 10\%$. In comparison, the thinning rates of the three commonly used methods, to achieve accuracy close

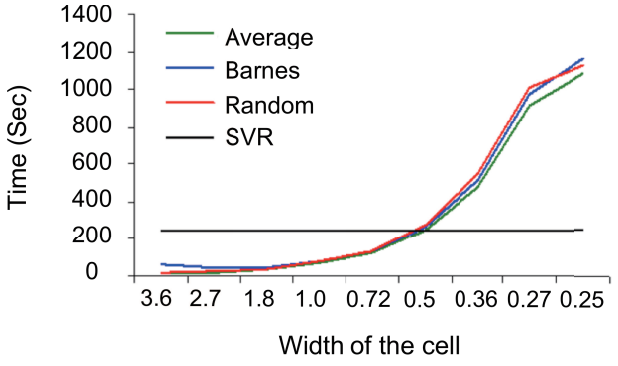

Fig. 8. Computation time as a function of cell width (in degrees) for the average, random and Barnes thinning solutions versus TSVR.

to that of the TSVR, is much larger ($\sim 26\%$).

4.3. *Results of the third experiment*

Using TSVR, computation times can be decreased by buffering in a series of subsets of data and calculating the support vectors of each sample. This process is known as pipeline thinning (Fig. 2). To investigate the gain in computational efficiency of the pipeline approach, compared to TSVR without a pipeline, a sample of 120983 data points was drawn from the 1.5 million observations. The results for the regular and pipeline TSVR are very similar, with MAE magnitude differences (Fig. 9a, b) of $\leqslant 0.05$ m s^{-1} and the MSE differences of $\leqslant 0.1$ m^2 s^{-2} (Fig. 9c, d). The correlations between the reconstructed and observed winds for the regular versus pipeline methods (Fig. 9e, f) show trivial differences in the second decimal point, at most. It is notable that the correlations for the *u*-component are, for both the regular and pipeline methods, ~ 0.97 (Fig. 9e) and, for the *v*-component, ~ 0.99 (Fig. 9f), indicating the very close correspondence between the thinned and the non-thinned data. The computation time for the pipeline TSVR is less than that for the regular TSVR. The computational efficiency gain arises as, for the first CPU time step (Fig. 10; $t = 1$), all the data within the window are thinned; however, for $t > 1$, using pipeline TSVR, only the new data are thinned. For both the pipeline and non-pipeline TSVR approaches, the time needed to thin the data for the first period was ~ 145 seconds. However, for periods 2–13, the average thinning time was ~ 142 seconds for the regular TSVR, decreasing by an order of magnitude to 13 seconds for the pipeline TSVR approach (Fig. 10). Therefore, the pipeline TSVR approach requires just 9% of the time of the non-pipeline TSVR method, while providing

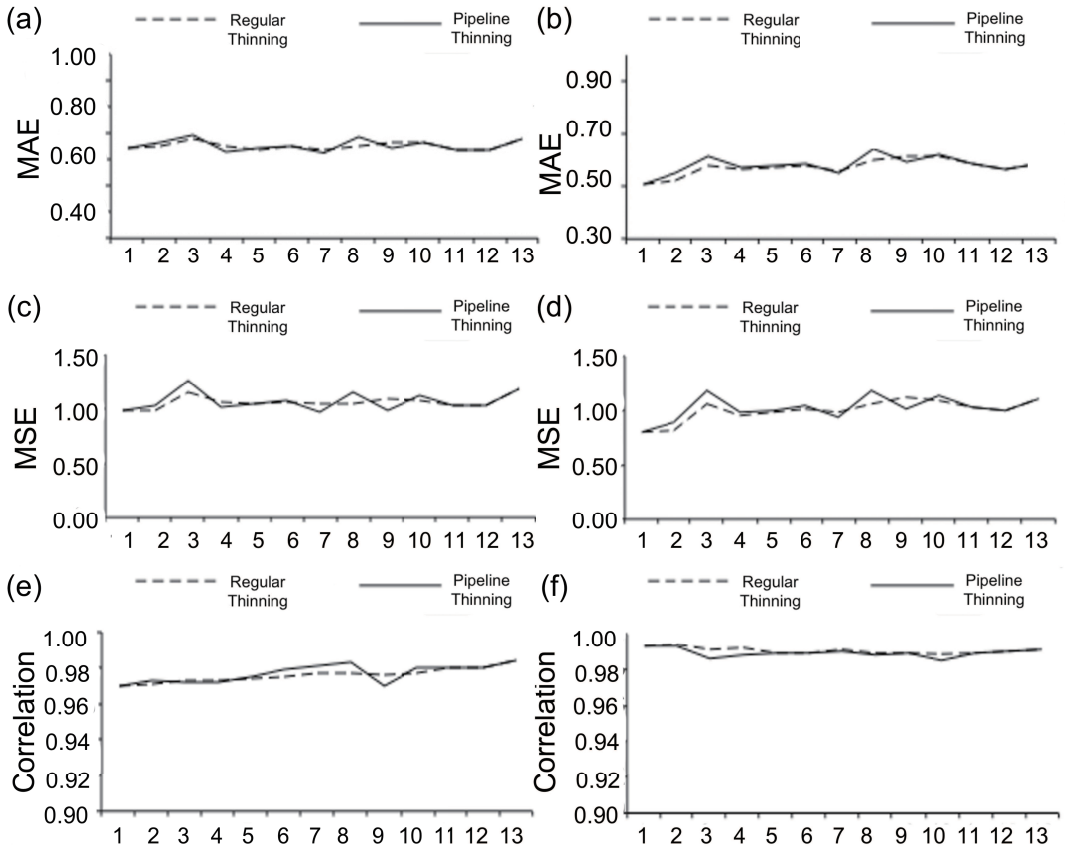

Fig. 9. Mean absolute differences (MAE), mean squared differences (MSE) (in m s^{-1}) and correlations between the thinned and non-thinned *u*-components (a, c, e) and *v*-component (b, d, f) of the wind regular SVR thinning versus the pipeline TSVR thinning. The data subset is shown on the horizontal axis.

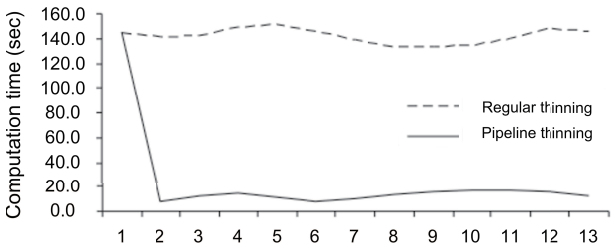

Fig. 10. Regular TSVR Thinning versus pipeline TSVR thinning computation times. The data subset is shown on the horizontal axis.

almost identical accuracy.

5. Conclusions

The removal of redundant data is commonly known as data thinning. In this study, the application is the thinning of *u*-and *v*-components of the winds estimated from WindSat. The number of observations is reduced through a combination of Voronoi tessellation and support vector regression (TSVR). Here, hundreds of thousands of observations are assigned to several thousand Voronoi cells to optimize the wind retrieval accuracy. For each cell, separate TSVR analyses were conducted, for the *u*-and *v*-components of the winds. The number of Voronoi cells can be adapted, consistent with the complexity of the field, by increasing or decreasing their number. The process can be extremely efficient if the process is parallelized by assigning the SVR calculation inside each Voronoi cell to a separate CPU.

The results of the thinning experiments yielded decidedly encouraging results. The TSVR requires fewer than 8%–10% of the WindSat data to produce a highly accurate estimate of the wind field (MAE < 1 m s^{-1} and the correlation $\geqslant +0.98$). In comparison, commonly used techniques, such as random selection, averaging and a Barnes filter, are computationally efficient, but have poor retrieval accuracy at coarse spatial resolution. However, at high spatial resolution, as the accuracy of the three commonly used techniques approaches that of TSVR, the computational times for the other thinning methods exceed those of the TSVR approach by a factor of ~ 4.

High retrieval accuracy is a requirement for meaningful analysis. Of the thinning techniques examined, only TSVR offers this combination of providing extremely high retrieval accuracy with the shortest clock time. To determine whether the computational efficiency of the TSVR approach could be improved further, a pipeline thinning methodology was applied to the TSVR, reducing the clock time from 150 to 15 seconds. Therefore, for any application requiring ingesting and preprocessing online data, followed by thinning, the pipeline TSVR methodology is advantageous. In this study, it is not only the most accurate of all methods tested but is also the fastest, by up to two orders of magnitude.

Acknowledgements. The authors wish to acknowledge NOAA Grant NA17RJ1227 and NSF Grant EIA-0205628 for providing financial support for this work. The third author was partly supported by RSF Grant 14-41-00039. The opinions expressed herein are those of the authors and not necessarily those of NOAA or NSF. The authors wish also to thank Kevin HAGHI, Andrew MERCER and Chad SHAFER for their assistance with several of the figures.

REFERENCES

Bakır, G. H., L. Bottou, and J. Weston, 2004: Breaking SVM complexity with cross-training. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, MIT Press, 81–88.

Barnes, S. L, 1964: A technique for maximizing details in numerical weather-map analysis. *Journal of Applied Meteorology*, **3**, 396–409.

Bondarenko, V., T. Ochotta, and D. Saupe, 2007: The interaction between model resolution, observation resolution and observations density in data assimilation: A two-dimensional study. Preprints, 11*th Symp. On Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface*, San Antonio, TX, Amer. Meteor. Soc., P5.19. [Available online at http://ams.confex.com/ams/pdfpapers/117655.pdf.]

Bottou L., and Y. LeCun, 2004: On-line learning for very large datasets. *Applied Stochastic Models in Business and Industry*, **21**, 137–151.

Bowyer, A., 1981: Computing Dirichlet tessellations. *Comput. J.*, **24**, 162–166.

Chang, P., P. Gaiser, K. St. Germain, and L. Li, 1997: Multi-Frequency Polarimetric Microwave Ocean Wind Direction Retrievals. Proceedings of the International Geoscience and Remote Sensing Symposium 1997, Singapore. [Available online at http://www.nrl.navy.mil/research/nrl-review/2004/featured-research/gaiser/#sthash.IskB3x9l.dpuf.]

Du Q., V. Faber, and M. Gunzburger, 1999: Centroidal Voronoi tessellations: applications and algorithms. *SIAM Review*, **41**, 637–676.

Gaiser, P. W., K. M. St. German, E. M. Twarog, G. A. Poe, W. Purdy, D. Richardson, W. Grossman, W. L. Jones, D. Spencer, G. Golba, J. Cleveland, L. Choy, R. M. Bevilacqua, and P. S. Chang, 2004: The WindSat space borne polarimetric microwave radiometer: Sensor description and early orbit performance. *IEEE Trans. on Geosci. and Remote Sensing*, **42**, 2347–2361.

Gilbert, R. C., and T. B. Trafalis, 2009: Quadratic programming formulations for classification and regression. *Optimization Methods and Software*, **24**, 175–185.

Helms, C. N., and R. E. Hart, 2013: A polygon-based line-integral method for calculating vorticity, divergence, and deformation from nonuniform observations. *J. Appl. Meteor. Climatol.*, **52**, 1511–1521.

Laskov, P., C. Gehl, S. Krüger, and K.-R. Müller, 2006: Incremental support vector learning: Analysis, implementation and applications. *Journal of Machine Learning Research*, **7**, 1909–1936.

Lazarus, S. M., M. E. Splitt, M. D. Lueken, R. Ramachandran, X. Li, S. Movva, S. J. Graves, and B. T. Zavodsky, 2010: Evaluation of data reduction algorithms for real-time analysis. *Wea. Forecasting*, **25**, 511–525.

Lorenc, A. C., 1981: A three-dimensional multivariate statistical

interpolation scheme. *Mon. Wea. Rev*., **109**, 1177–1194.

Mansouri, H., R. C. Gilbert, T. B. Trafalis, L. M. Leslie, and M. B. Richman, 2007: Ocean surface wind vector forecasting using support vector regression. In C. H. Dagli, A. L. Buczak, D. L. Enke, M. J. Embrechts, and O. Ersoy, editors, *Intelligent Engineering Systems Through Artificial Neural Networks*, **17**, 333–338.

MATLAB, 2012: MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States. [Available online at http://nf.nci.org.au/facilities/software/Matlab/techdoc/ref/voronoi.html.]

Musicant D. R., and O. L. Mangasarian, 2000: Large scale kernel regression via linear programming. *Machine Learning*, **46**, 255–269.

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

Ochotta, T., C. Gebhardt, V. Bondarenko, D. Saupe, and W. Wergen, 2007: On thinning methods for data assimilation of satellite observations. Preprints, 23*rd Int. Conf. on Interactive Information Processing Systems* (*IIPS*), San Antonio, TX, Amer. Meteor. Soc., 2B.3. [Available online at http://ams.confex.com/ams/pdfpapers/118511.pdf.]

Platt, J, 1999: Using sparseness and analytic QP to speed training of support vector machines. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems* 11, MIT Press, 557–563.

Purser, R. J., D. F. Parrish, and M. Masutani, 2000: Meteorological observational data compression: An alternative to conventional "super-obbing". *NCEP Office Note* 430, 12pp. [Available online at http://www.emc.ncep.noaa.gov/mmb/papers/purser/on430.pdf.]

Quilfen, Y., C. Prigent, B. Chapron, A. A. Mouche, and N. Houti, 2007: The potential of QuikSCAT and WindSat observations for the estimation of sea surface wind vector under severe weather conditions, *J. Geophys. Res. Oceans*, **112**, 49–66.

Quinn, M. J. 2004: *Parallel Programming in C with MPI and openMP*. Dubuque, Iowa: McGraw-Hill Professional, 544pp.

Ragothaman, A., S. C. Boddu, N. Kim, W. Feinstein, M. Brylinski, S. Jha, and J. Kim, 2014: Developing ethread pipeline using saga-pilot abstraction for large-scale structural bioinformatics. BioMed Research International, 2014. 1–12, doi: 10.1155/2014/348725.

Santosa, B., M. B. Richman, and T.B. Trafalis, 2005: Variable selection and prediction of rainfall from WSR-88D radar using support vector regression. *Proceedings of the* 6*th WSEAS Transactions on Systems*, **4**, 406–411.

Schölkopf, B., and A. Smola, 2002: *Learning with Kernels.* MIT Press, 650pp.

Smola, A. J., and B. Schölkopf, 1998: *A Tutorial on Support Vector Regression* Royal Holloway College, NeuroCOLT Technical Report (NC-TR-98-030), University of London, UK. [Available online at http:// svms.org/tutorials/SmolaScholkopf1998.pdf.]

Shawe-Taylor, J., and N. Cristianini, 2004: *Kernel Methods for Pattern Analysis*. Cambridge University Press, 478pp.

Son, H-J, T. B. Trafalis, and M. B. Richman, 2005: Determination of the optimal batch size in incremental approaches: An application to tornado detection, *Proceedings of International Joint Conference on Neural Networks, IEEE*, 2706–2710.

Trafalis, T. B., B. Santosa, and M. B. Richman, 2005: Feature selection with linear programming support vector machines and applications to tornado prediction, *WSEAS Transactions on Computers*, **4**, 865–873.

Vapnik, V., 1982: *Estimation of Dependences Based on Empirical Data*. Springer, 505pp.

Voronoi, G., 1908: Recherches sur les paralléloèdres Primitives. *J. Reine Angew. Math.* **134**, 198–287 (in French).

Wei, C.-C., and J. Roan, 2012: Retrievals for the rainfall rate over land using special sensor microwave imager data during tropical cyclones: Comparisons of scattering index, regression, and support vector regression. *J. Hydrometeor*, **13**, 1567–1578.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed., Elsevier, 676 pp.