Available online at www.sciencedirect.com

ScienceDirect

Remote Sensing of Environment 112 (2008) 1846–1855

Remote Sensing of Environment

www.elsevier.com/locate/rse

Spacebased estimation of moisture transport in marine atmosphere using support vector regression

Xiaosu Xie [a,*], W. Timothy Liu [b], Benyang Tang [b]

[a] M/S 300-323, 4800 Oak Grove Dr., Jet Propulsion Laboratory, Pasadena, CA 91109, USA
[b] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Received 27 March 2007; received in revised form 6 September 2007; accepted 9 September 2007

Abstract

An improved algorithm is developed based on support vector regression (SVR) to estimate horizontal water vapor transport integrated through the depth of the atmosphere (Θ) over the global ocean from observations of surface wind-stress vector by QuikSCAT, cloud drift wind vector derived from the Multi-angle Imaging SpectroRadiometer (MISR) and geostationary satellites, and precipitable water from the Special Sensor Microwave/Imager (SSM/I). The statistical relation is established between the input parameters (the surface wind stress, the 850 mb wind, the precipitable water, time and location) and the target data (Θ calculated from rawinsondes and reanalysis of numerical weather prediction model). The results are validated with independent daily rawinsonde observations, monthly mean reanalysis data, and through regional water balance. This study clearly demonstrates the improvement of **Θ** derived from satellite data using SVR over previous data sets based on linear regression and neural network. The SVR methodology reduces both mean bias and standard deviation compared with rawinsonde observations. It agrees better with observations from synoptic to seasonal time scales, and compare more favorably with the reanalysis data on seasonal variations. Only the SVR result can achieve the water balance over South America. The rationale of the advantage by SVR method and the impact of adding the upper level wind will also be discussed.

Keywords: Moisture transport; Water cycle; Support vector regression; Remote sensing

1. Introduction

Ocean is the main reservoir of heat and water on Earth. The never-ending recycling process in which a small fraction of water is continuously removed from the ocean as excess evaporation over precipitation into the atmosphere, redistributed through atmospheric circulation, deposited as excess precipitation over evaporation on land, and returned to the ocean as river discharge, is critical to the existence of human life and the variability of weather and climate.

The moisture transport integrated over the depth of the atmosphere is

$$\boldsymbol{\Theta} = \frac{1}{g}\int_0^{p_s} q\mathbf{u}dp \qquad (1)$$

where g is the acceleration due to gravity, p is the pressure, p_s is the pressure at the surface, q and $\mathbf{u}$ are the specific humidity and wind vector at a certain level. Bold symbols represent vector quantities. **Θ** links the water reservoir in the ocean to those over ice and land, and the divergence of which is the difference between evaporation (E) and precipitation (P) at the ocean surface, assuming the change in atmospheric storage is small.

In situ observations for the major components of the hydrological cycle are sparse and intermittent, which severely restrains our understanding of the water cycle. Efforts to retrieve **Θ** from rawinsonde observations in various regions and time periods can be traced back several decades ago (e.g., Starr & White, 1955; Rasmusson, 1967; Rosen et al., 1979; Peixoto et al., 1981; Bryan & Oort, 1984). In addition to in situ measurements, numerical weather prediction (NWP) data products have also been used to study atmospheric moisture transport and hydrological budget (e.g., Roads et al., 1992; Trenberth & Guillemot, 1995; Mo & Higgins, 1996; Cohen et al., 2000).

Spacebased observations, with improved coverage and resolution, should improve the estimation of **Θ**. As suggested

* Corresponding author.
E-mail address: xiaosu@pacific.jpl.nasa.gov (X. Xie).

0034-4257/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2007.09.003

by Liu (1993), $\mathbf{\Theta}$ can be written as the product of an equivalent velocity ($\mathbf{u}_e$) and W, where

$$W = \frac{1}{g}\int_0^{p_s} q dp \tag{2}$$

is the precipitable water, and $\mathbf{u}_e = \mathbf{\Theta}/W$, by definition, is the depth-average velocity weighted by humidity. Using W derived from the Scanning Multichannel Microwave Radiometer by Liu (1987), two methods to compute monthly mean $\mathbf{\Theta}$ over tropical oceans were demonstrated. Heta and Mitsuta (1993) assumed $\mathbf{u}_e$ equals to the 850 mb wind vector produced by the numerical model that assimilated cloud drift wind, while Liu (1993) related $\mathbf{u}_e$ to surface wind vectors derived from satellite data using a polynomial regression. With the availability of surface equivalent neutral wind vector ($\mathbf{u}_s$) from QuikSCAT, Liu and Tang (2005) produced $\mathbf{\Theta}$ for tropical and subtropical oceans at twice daily resolution, by relating $\mathbf{u}_e$ to $\mathbf{u}_s$ through neural network (NN). Based on the same physical rationale, Xie and Liu (2005) used nonparametric regression (Hardle, 1990), which is essentially linear regression (LR), to estimate $\mathbf{\Theta}$ and showed that the annual mean of $\nabla\cdot\mathbf{\Theta}$ bears similar large-scale patterns as that of E-P. Both terms show the major climatological features over the tropical and subtropical oceans.

In this study, an improved method based on SVR (e.g., Vapnik, 1995, 1998; Schölkopf, 1997; Joachims, 1999; among many others), will be presented. The method is described in Section 2 and the training data in Section 3. Validation of the new estimate and the comparison with the linear regression method and the data set produced by Liu and Tang (2005) are shown in Section 4.1 and 4.2. An earlier version of $\mathbf{\Theta}$ derived by SVR was used to show continental water balance in South America by Liu et al. (2006); the result is updated in Section 4.3. Discussion is given in Section 5.

2. Support vector regression

Support vector machines (SVMs) have several major applications, i.e., classification, regression, and time series prediction. The SVMs for classification have been widely used, such as in handwritten digital recognition, object recognition, speaker identification, face detection in image, and text categorization. The SVMs for regression are referred as support vector regression (SVR) henceforth, which is a statistical tool to derive the relationship between input and output. A comprehensive tutorial of SVMs for pattern recognition can be found in Burges (1998) and a tutorial of SVR can be found in Smola and Schölkopf (2004). In the past decade, SVMs have become increasingly popular due to their good generalization performance. Many studies demonstrated that SVMs match or outperform other machine learning methods (e. g., Blanz et al., 1996; Müller et al., 1997; Drucker et al., 1997; Meyer et al., 2003).

The theoretical basis of SVMs originates from statistical learning theory (Vapnik, 1995, 1998). SVMs essentially translate a nonlinear problem to a linear one by mapping the input into a high-dimensional feature space, and fit a linear model in the feature space (Boser et al., 1992; Vapnik 1995; Cherkassky & Mulier, 1998). The solution of the optimization is unique. Through the introduction of kernel function, the mapping to the feature space is implicit, and computations are performed directly in the input space rather than in the feature space. The concept of SVR is summarized below through mathematical expressions.

The description here is concise. Details can be referred to Smola and Schölkopf (2004) and Tang and Mazzoni (2006). The training samples are denoted as $(x_i,\mathbf{\Theta}_i)\{i=1,\ldots,N\}$, where x_i is the input and $\mathbf{\Theta}_i$ the output. A kernel mapping function $\Phi(x_i)$ transforms the input into the high-dimensional feature space. Then a linear model is fitted to the data in the feature space by a convex optimization problem:

$$\text{minimize}\frac{1}{2}\ \|\ \mathbf{w}\ \|^2 + C\sum_i\left(\xi_i + \xi_i^*\right) \tag{3}$$

$$\text{subject to}\begin{cases}[\mathbf{w}\cdot\Phi(x_i)+b]-\mathbf{\Theta}_i\le\varepsilon+\xi_i\\ \mathbf{\Theta}_i-[\mathbf{w}\cdot\Phi(x_i)+b]\le\varepsilon+\xi_i^*\\ \xi_i\ge0,\ \xi_i^*\ge0\end{cases} \tag{4}$$

where $\mathbf{w}$ is the weight vector in the feature space, ξ_i and ξ_i* are called slack variables, measuring deviation of the estimated model output from the target data. The output is represented by a linear function in the feature space: $f(x_i)=\mathbf{w}\cdot\Phi(x_i)+b$, where b is a bias term determined during the SVR training. ε is the precision parameter. The goal of SVR is to find a function $f(x_i)$ with the deviation from the target $\mathbf{\Theta}_i$ no larger than ε for all the training data. In this sense, only those data points with deviations larger than ε will contribute to the error function (see support vectors defined in the next paragraph). Bigger ε leads to fewer support vectors. C is the penalty factor, which determines the trade off between the model complexity and the degree to which deviations larger than ε are tolerated. If C is too large (small), it may cause overfitting (underfitting).

Introducing the Lagrangian multipliers α_i and α_i*, Eqs. (3) and (4) become

$$\text{maximize}: \begin{array}{l} -\frac{1}{2}\sum_{i,j}\left(\alpha_i^*-\alpha_i\right)\left(\alpha_j^*-\alpha_j\right)\Phi(x_i)\cdot\Phi(x_j)\\ -\varepsilon\sum_i\left(\alpha_i^*+\alpha_i\right)+\sum_i\mathbf{\Theta}_i\left(\alpha_i^*-\alpha_i\right)\end{array} \tag{5}$$

$$\text{subject to}\sum_{i,j}\left(\alpha_i^*-\alpha_i\right)=0\ \text{and}\ \alpha_i,\alpha_i^*\in[0,C] \tag{6}$$

where the weight vector $\mathbf{w}$ is described as a linear combination of the following training pattern

$$\mathbf{w}=\sum_i\left(\alpha_i^*-\alpha_i\right)\Phi(x_i) \tag{7}$$

The inputs with nonzero $\alpha_i{}^*-\alpha_i$ are called support vectors, which are those inputs whose model outputs $f(x_i)$ differ from the target data $\mathbf{\Theta}_i$ by at least ε. The number of support vectors is usually much less than that of the training data samples, which leads to a sparse representation.

By substituting Eq. (7), the output $f(x)$ with an input x can be expressed as

$$f(x)=\mathbf{w}\cdot\Phi(x)+b=\sum_i\left(\alpha_i^*-\alpha_i\right)\Phi(x_i)\cdot\Phi(x)+b \tag{8}$$

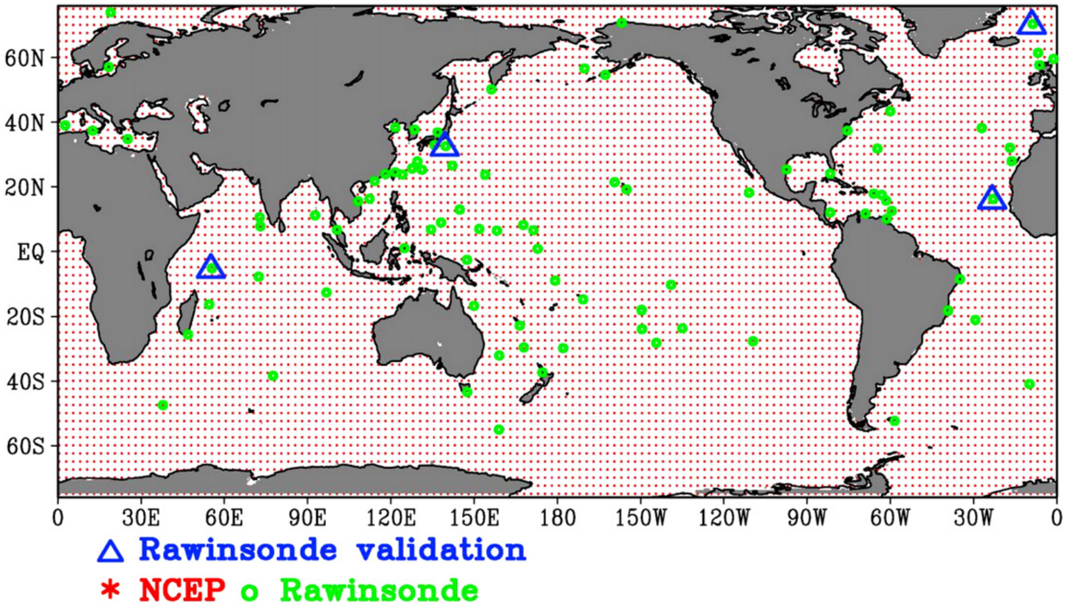

Fig. 1. Distribution of rawinsonde stations, superposed on the NCEP 2.5° grids. The blue triangles mark those rawinsonde stations for validation.

The nonlinear transformation $\Phi(x_i)$ is not required to be explicit. The dot product of the two mapped vectors in Eqs. (5) and (8) is calculated by a kernel function

$$K(x_i, x) = \Phi(x_i) \cdot \Phi(x) \tag{9}$$

There are many types of kernel functions. In this study, the commonly used radial basis function is adopted

$$K(x_i, x) = \exp\left[-\gamma \ ||x - x_i \ ||^2\right] \tag{10}$$

where γ is a kernel parameter.

The accuracy of SVR depends on the selection of the hyperparameters (C and ε) and the kernel parameter (γ). The initial values of the parameters are empirically estimated from the training data based on previous studies. Then only one parameter varies until the optimized correlation between the trained output and the target data is found. In fact, there is a range of the optimized parameters where the correlation coefficient is not sensitive. In this algorithm, C=10, ε=0.2, and γ=0.05.

3. Training data

The target data $\mathbf{\Theta}$ are daily averages computed from both rawinsonde observations and the National Center for Environmental Prediction (NCEP) reanalysis. The NCEP data are included in the training samples to represent a global coverage, because the rawinsonde measurements are sparse over the ocean

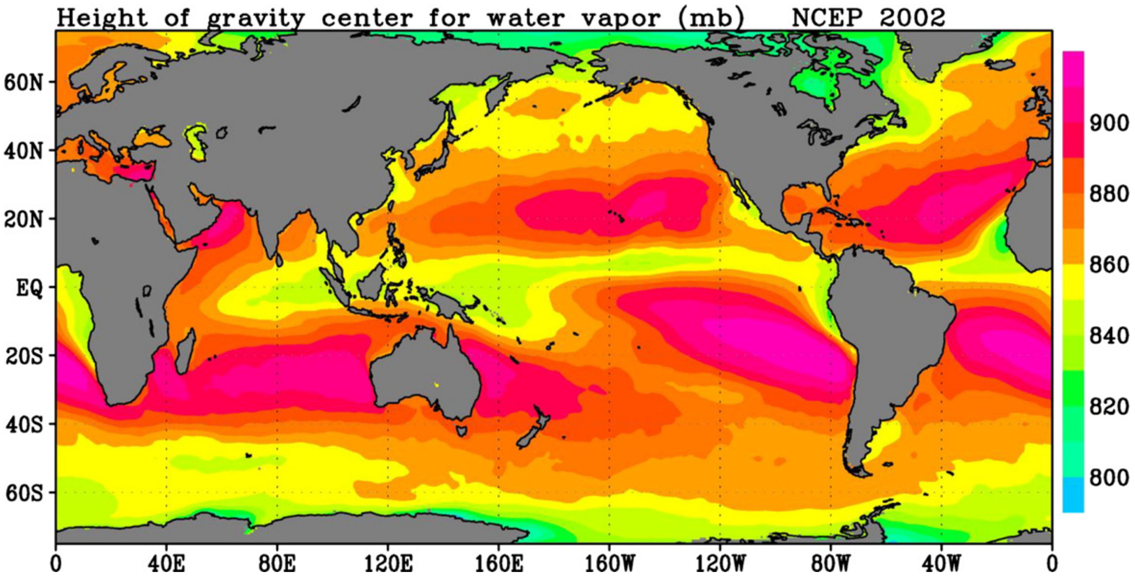

Fig. 2. Pressure of gravity center for humidity estimated from the NCEP reanalysis, averaged for 2002.

Table 1a
Mean and standard deviation of the difference between Θ from rawinsonde and Θ from satellite data using the three algorithms

	Mean $\Delta\Theta_x$, kg/m/s	Mean $\Delta\Theta_y$, kg/m/s	S.D. $\Delta\Theta_x$, kg/m/s	S.D. $\Delta\Theta_y$, kg/m/s
SVM	−2.75	−8.58	69.83	60.16
LR	−2.00	−4.15	125.22	80.25
NN	−28.56	−19.00	135.19	97.54

Table 1b
Correlation coefficient between Θ from rawinsonde and Θ from satellite data using the three algorithms

	$\Delta\Theta_x$, kg/m/s	$\Delta\Theta_y$, kg/m/s
SVM	0.948	0.867
LR	0.826	0.739
NN	0.804	0.552

and with limited geographic coverage (Fig. 1). One year of data in 2001 were used as training samples. 90 rawinsonde stations were selected through data quality control, of which 74 stations are within 40°S–40°N in the three ocean basins and 16 stations are located at higher latitudes. **Θ** computed from the NCEP reanalysis is evenly distributed at 2.5° by 2.5° grids between 75°S and 75°N every 10 days. Distribution of the rawinsonde stations is shown in Fig. 1, on top of the NCEP grids. From these data sets, a total of 26,547 data points were selected as the training data set, with 10,000 points randomly selected from the rawinsonde data between 40°S and 40°N, 6547 points poleward of 40° latitude, and 10,000 points randomly selected from the NCEP reanalysis.

The input data include daily averages of $\mathbf{u}_s$ measured by QuikSCAT (Liu, 2002), $\mathbf{u}_{850}$ from the combined cloud drift wind averaged between 800 mb and 900 mb derived from MISR (Horvath & Davies, 2001) and geostationary satellites (Hayden & Pursor, 1995), W from SSM/I (Wentz, 1997), the day of the year, longitude and latitude. Time and longitude take the forms of sine and cosine because of their periodicity. The input parameters are interpolated to the locations of the target data. The input parameters and the target data (x), except for time and longitude, are normalized as: $x' = (x - \bar{x})/\sigma$, where $\bar{x}$ and σ are the mean and standard deviation of x.

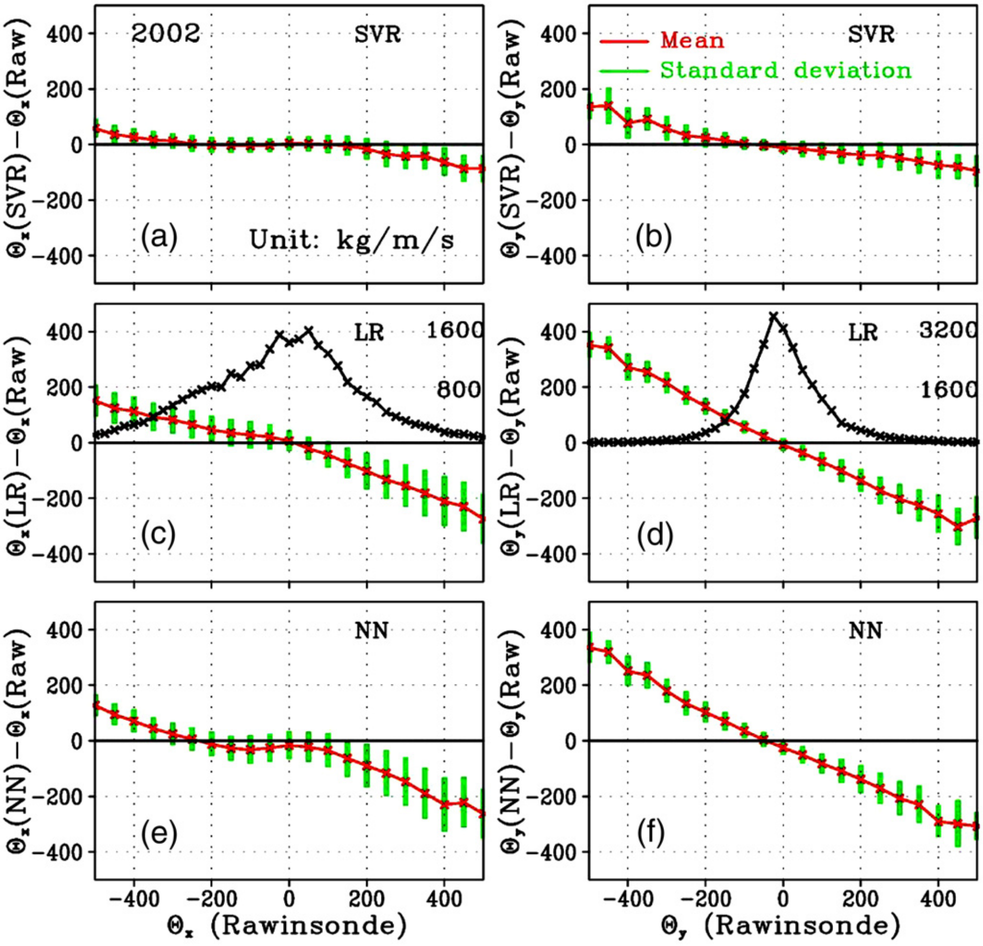

Fig. 3. Bin average of $\Theta_{SVR}-\Theta_R$, as a function of Θ_R, for (a) zonal component, and (b) meridional component. Standard deviation is superimposed on each bin average as error bars. (c) and (d) are the same as (a) and (b), except for $\Theta_{LR}-\Theta_R$. The black curves represent number distributions based on Θ_R. (e) and (f) are the same as (a) and (b), except for $\Theta_{NN}-\Theta_R$.

The methodologies of Liu (1993) and Heta and Mitsuta (1993), as discussed in Section 1, are combined by adding $\mathbf{u}_{850}$ to $\mathbf{u}_s$ as an input parameter. Although humidity peaks near the surface and decreases rapidly with height, the gravity center of the humidity profile is lifted up to the top of the boundary layer over the ascending areas and suppressed to a lower level over the descending regions (Fig. 2). In addition, the dominant mode of vertical humidity variability was found to peak at the top of the boundary layer (Liu et al., 1991), which also suggests the importance of winds at this level in addition to the surface winds.

4. Analysis

Four years, 2001–2004, of Θ_{SVR} were produced using SVR. The same training data were used to produce Θ_{LR}, using multivariate linear regression. They are compared with Θ_{NN}, produced and released to the science community by Liu and Tang (2005), using neural network based on different training data. We did not attempt to optimize the Θ_{NN} algorithm in this study. The daily values are first evaluated in Section 4.1 with rawinsonde data in 2002 (independent of the training data), as validation. The distribution of rawinsonde is limited, and monthly mean global distributions were compared with Θ_{NCEP} computed from the NCEP reanalysis in Section 4.2.

4.1. Comparison with daily rawinsonde data

There are 28,408 rawinsonde observations (Θ_R) collocated with the satellite-based data. Table 1 shows that the mean difference between Θ_R and all the model estimates (Θ_{SVR}, Θ_{LR} and Θ_{NN}) are small. The standard deviation of the difference is smaller for Θ_{SVR} than those for Θ_{LR} and Θ_{NN}. As indicated by the correlation coefficients between model estimates and the rawinsonde observations, Θ_{SVR} is better than the other two data sets. The bin-averages in Fig. 3 show that all three satellite-based data sets overestimate compared to observations at low amplitudes and underestimate at high amplitudes, as expected. The numbers of data samples at the two ends are small, as indicated by the Θ_R distributions in Fig. 3c and d. The mean difference and standard deviation from the rawinsondes are much lower for Θ_{SVR} among the three data sets, particularly for the zonal component.

Performance of each method is also evaluated by time series comparison at individual rawinsonde stations. Θ_{SVR} has the best performance across the broad range of time scales. For better figure clarity, Θ_{LR} is not shown. Comparisons at two stations, representing different wind regimes, are shown as examples (Figs. 4 and 5). Θ_{SVR} captures not only the seasonal variation but also the high-frequency variability of Θ_R, with amplitude in good agreement with observations. For instance,

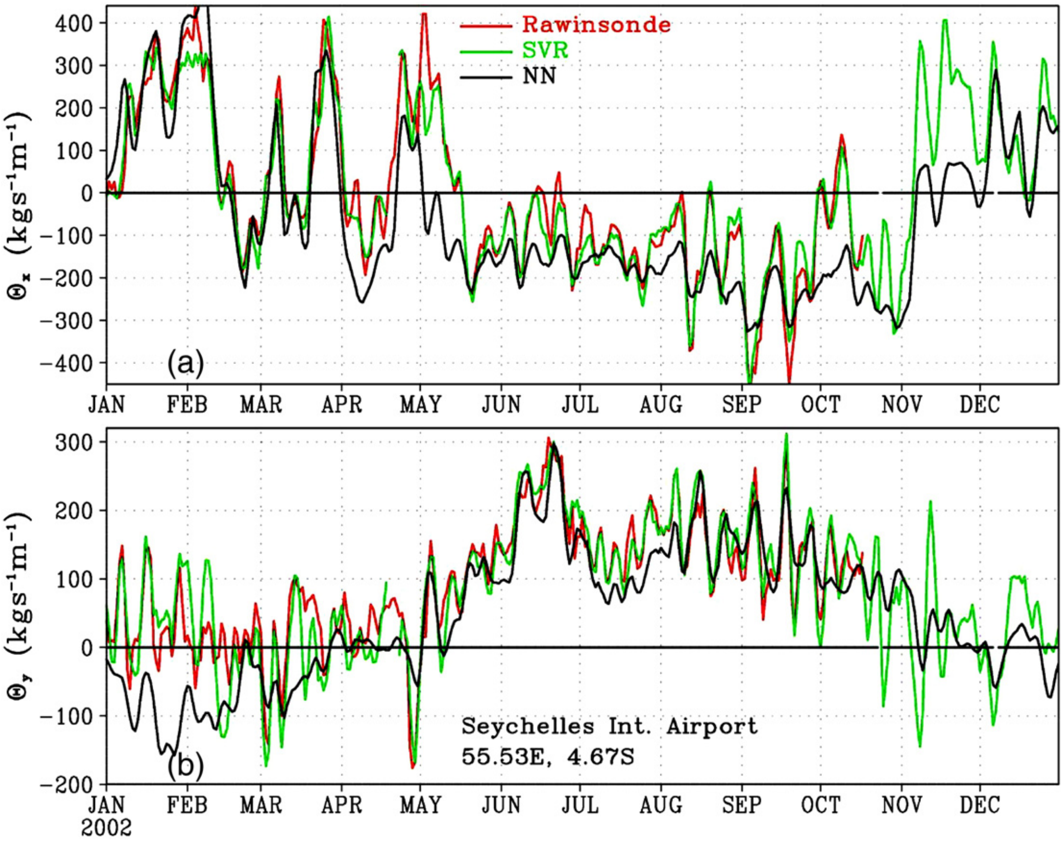

Fig. 4. Comparison of daily Θ derived from satellite data using SVR (green curve) and NN (black curve) with Θ derived from rawinsonde (red curve) for (a) zonal and (b) meridional components at Seychelles International Airport (55.53°E, 4.67°S). A 3-day running mean is applied.

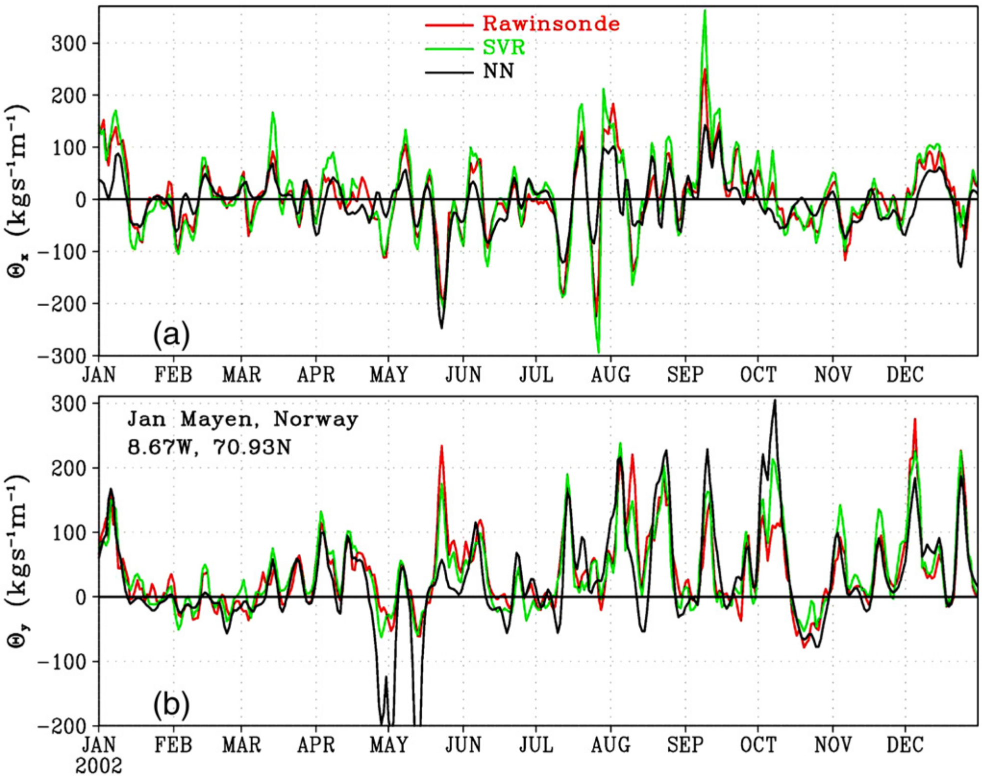

Fig. 5. Same as Fig. 4, except for the rawinsonde station at Jan Mayen, Norway (8.67°W, 70.93°N).

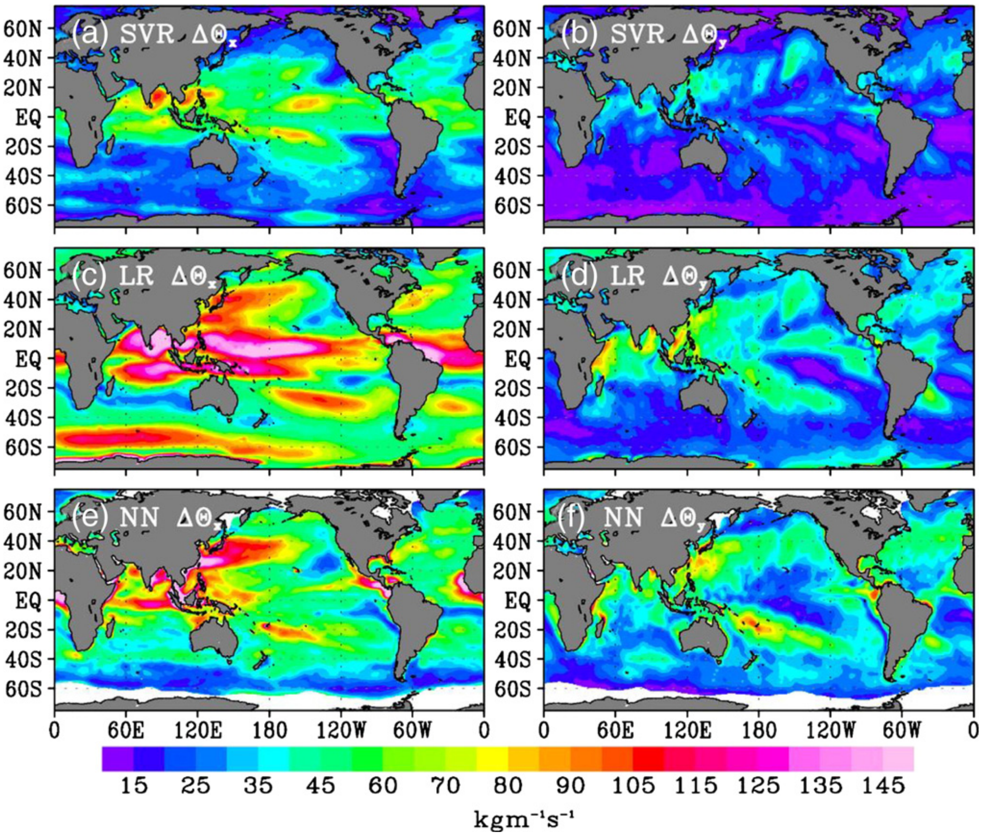

Fig. 6. RMS differences of monthly **Θ** derived from satellite data using SVR and from NCEP reanalysis for (a) zonal and (b) meridional components. (c) and (d) are the same as (a) and (b), except for LR. (e) and (f) are the same as (a) and (b), except for NN.

the time series at Seychelles International Airport (Fig. 4) demonstrates the reversal of the Indian summer and winter monsoon. The station at high latitude (Fig. 5 for Jan Mayen, Norway) does not show any clear seasonal cycle but has higher frequency variations. Θ_{LR} (not shown in the figures) and Θ_{NN} often lack sufficient skill in capturing the high and low extreme values, which generally produces a narrower range. Θ_{LR} is particularly poor in high latitude compared with the rawinsonde observations. Θ_{NN} is sometimes deficient in high-frequency variability, such as the meridional component during January–April 2002 at Seychelles International Airport (Fig. 4). The meridional component of Θ_{NN} shows unrealistic amplitudes in April and May in 2002 at Jan Mayen, Norway (Fig. 5).

4.2. Comparison with monthly NWP estimates

Among the root-mean-square (RMS) differences between the three data sets derived from satellite data and from NCEP reanalysis, the SVR method gives the lowest values in general (Fig. 6). Relatively large RMS in the zonal component of Θ_{SVR} is found in the Bay of Bengal and South China Sea, where strong seasonal variations of the monsoon circulation dominate. Θ_{LR} has much larger RMS differences than the other two data sets, especially for the zonal component in the tropics and the southern ocean. Fig. 7 shows that the temporal variation of Θ_{NCEP} has significant correlation with Θ_{SVR} over global oceans (with correlation coefficient of over 0.9 in most areas) for the 4-year monthly mean time series. The zonal components of Θ_{LR} and Θ_{NN} show low correlation with Θ_{NCEP} to the west of Japan and to the east along the Kuroshio extension. In this region, the time series in Fig. 8 shows that the low frequency variation of Θ_{SVR} agrees with Θ_{R} marginally better than Θ_{NCEP}, while Θ_{LR} and Θ_{NN} fail to capture the peaks of westerly in the summers of 2003 and 2004. The meridional component of Θ_{LR} and Θ_{NN} does not agree with Θ_{NCEP} very well in the temporal variability in the eastern tropical Atlantic off the African coast and the eastern Pacific off Baja California and off Peru (Fig. 7d, f). The zonal component of Θ_{LR} also shows low correlation with Θ_{NCEP} in the eastern Pacific and Atlantic, south of the equator, and Gulf Stream. The reason is that Θ_{LR} poorly predicts the seasonal cycle in these regions; it fails to capture the strong easterlies during the boreal spring in the eastern Pacific and Atlantic, and the strong westerlies during the boreal summer in the Gulf Stream. There is no rawinsonde station in these regions, except in the Gulf Stream. The zonal component of Θ for the three satellite-based data sets and NCEP reanalysis all capture the seasonal cycle of the rawinsonde observations, but those derived from LR and NN have lower amplitudes. For the meridional components south of Baja California and west of Cape Verde of Africa, all data sets are poor compared to observations, because the amplitudes and the annual cycles are weak, as demonstrated in Fig. 9.

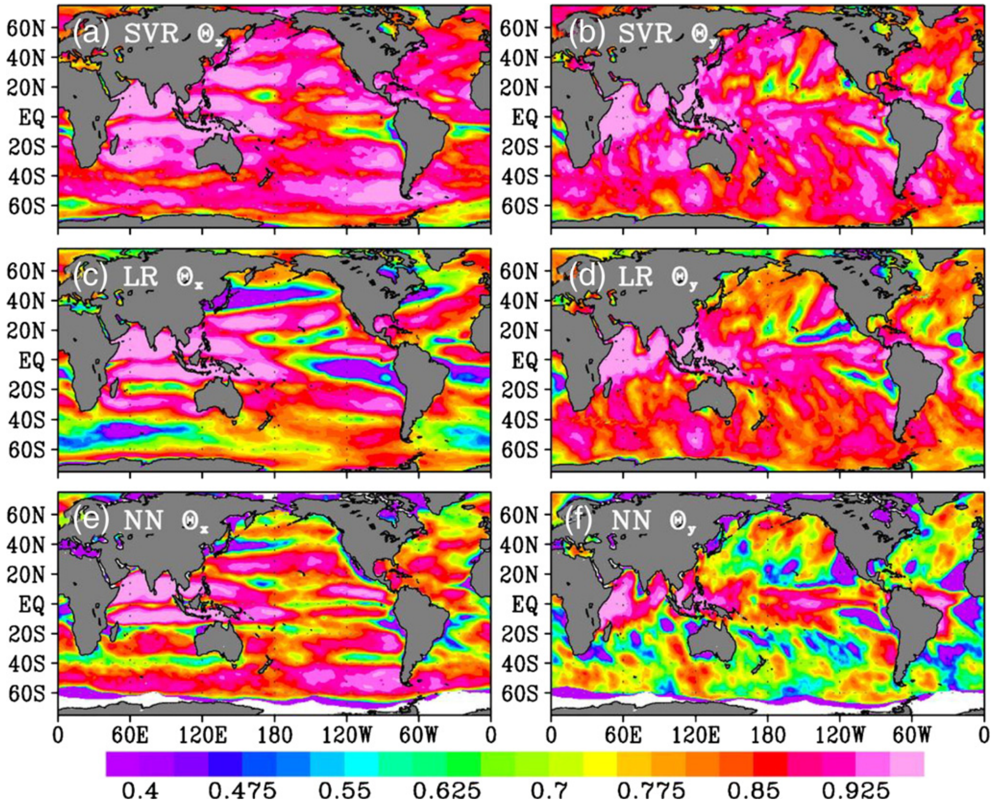

Fig. 7. Correlation coefficients between monthly Θ derived from satellite data using SVR and from NCEP reanalysis for (a) zonal and (b) meridional components. (c) and (d) are the same as (a) and (b), except for LR. (e) and (f) are the same as (a) and (b), except for NN.

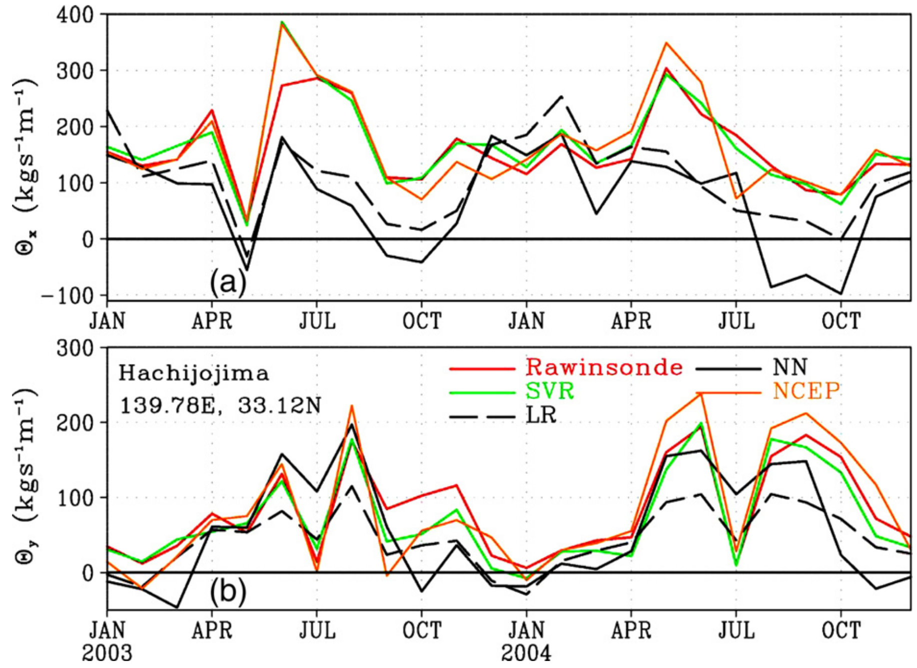

Fig. 8. Comparison of monthly **Θ** derived from satellite data using SVR (green curve), LR (dashed black curve) and NN (solid black curve) with **Θ** derived from rawinsonde (red curve) and the NCEP reanalysis (orange curve) for (a) zonal and (b) meridional components at Hachijojima (139.78°E, 33.12°N).

4.3. Water balance over South America

The best validation of the remote sensing technique is, perhaps, through application. Liu et al. (2006) showed the approximate balance of the mass change rate ($\partial M/\partial t$) measured by Gravity Recovery and Climate Experiment (GRACE), for the continent of South America, with the normal component of **Θ** integrated across the entire coast lines of South America ($\int\boldsymbol{\Theta}$) minus the climatologic river runoff from the continent (R), in agreement with conservation principle.

$$\frac{\partial M}{\partial t} = \int \boldsymbol{\Theta} - R \tag{11}$$

The **Θ** used by Liu et al. (2006) is derived by the same SVR technique, except without the use of $\mathbf{u}_{850}$ as input parameter.

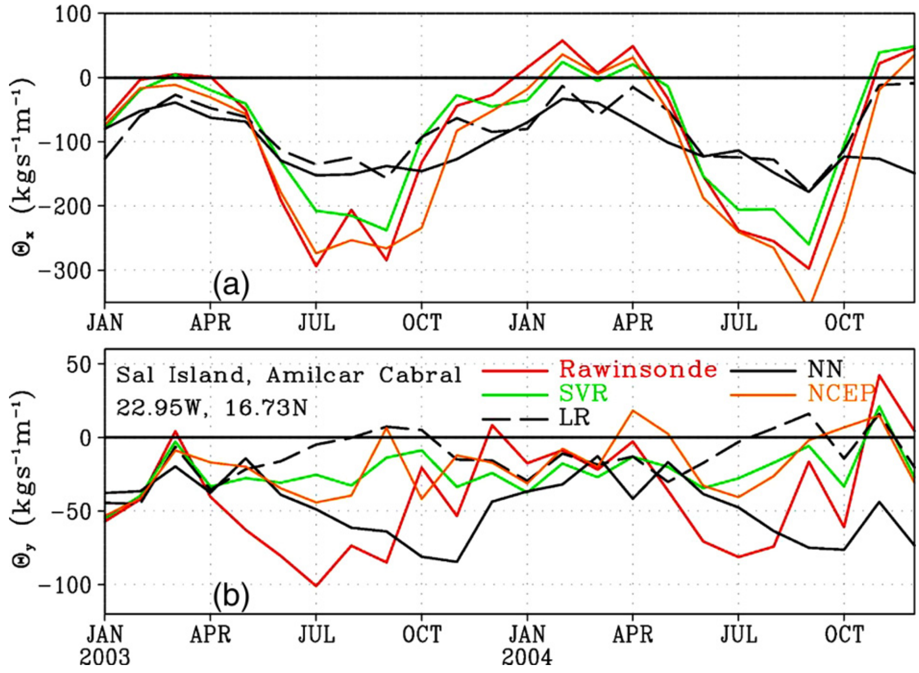

Fig. 9. Same as Fig. 8, except for rawinsonde station at Sal Island, Amilcar Cabral (22.95°W, 16.73°N).

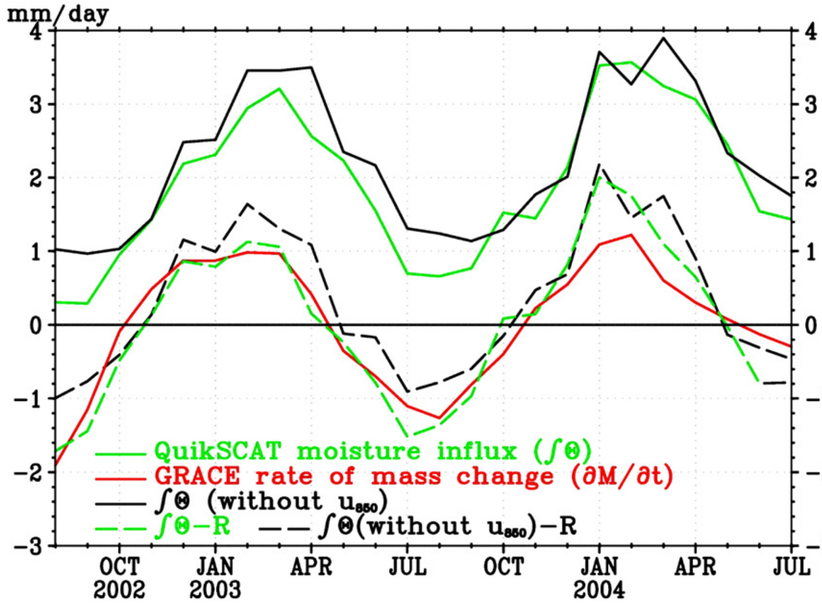

Fig. 10. Annual variation of hydrologic parameters over South America: mass change rate $\partial M/\partial t$ (red line), total moisture transport across coastlines into the continent $\int\boldsymbol{\Theta}$ estimated from satellite data using SVR (solid green line), $\int\boldsymbol{\Theta}-R$ (dashed green line, where R is climatological river discharge), $\int\boldsymbol{\Theta}$ estimated from the earlier version of SVR without $\mathbf{u}_{850}$ (solid black line), and the respective $\int\boldsymbol{\Theta}-R$ (dashed black line).

Using both $\boldsymbol{\Theta}_{\mathrm{LR}}$ and $\boldsymbol{\Theta}_{\mathrm{NN}}$ result in too much moisture inflow into the continent to balance $\partial M/\partial t$ and R. The addition of $\mathbf{u}_{850}$ as input parameter slightly improve the balance as shown in Fig. 10.

5. Discussion

In this study, an improved approach using SVR to estimate $\boldsymbol{\Theta}$ over the global ocean from combined satellite observations, including $\mathbf{u}_{\mathrm{s}}$ measured by QuikSCAT, $\mathbf{u}_{850}$ derived from MISR and geostationary satellites, and W from SSM/I, is presented. The statistical model is constructed through SVR by mapping the input parameters, which are $\mathbf{u}_{\mathrm{s}}$, $\mathbf{u}_{850}$, W, time and location, to the target data calculated from rawinsondes and the NCEP reanalysis. The results are validated against independent rawinsonde observations and compared with data sets produced based on LR and NN. The SVR algorithm reduces both the mean bias and the standard deviation from the observations, outperforming the other two data sets. It captures not only the seasonal changes but also the synoptic and intraseasonal variations of the observations; the agreement is in both phase and amplitude, and in both low and high latitude oceans. A common deficiency in the previous data sets is their inability to produce the extreme values, and therefore they do not have the full spectrum of variability in the observations.

The monthly mean $\boldsymbol{\Theta}$ derived from the three methods is also compared with the NCEP reanalysis. Although NWP reanalysis may have considerable uncertainties in simulating water cycle, they are often used to estimate moisture transport. Overall $\boldsymbol{\Theta}_{\mathrm{SVR}}$ correlates better with the NCEP reanalysis than $\boldsymbol{\Theta}_{\mathrm{LR}}$ and $\boldsymbol{\Theta}_{\mathrm{NN}}$ in seasonal variations. It also has the smallest RMS differences from the reanalysis data. The RMS bias of $\boldsymbol{\Theta}_{\mathrm{LR}}$ in the zonal component is much larger than the other data sets. $\boldsymbol{\Theta}_{\mathrm{LR}}$ also shows large differences from the NCEP reanalysis for the zonal component over the Kuroshio and Gulf Stream extensions, and in the eastern Pacific and Atlantic south of the equator. $\boldsymbol{\Theta}_{\mathrm{NN}}$ has larger correlation discrepancies from the NCEP reanalysis in the meridional component. All three methods show poor correlation for the meridional components south of Baja California and west of Cape Verde of Africa, when the amplitude is weak and without clear annual cycles.

The advantage of SVR method in the retrieval of $\boldsymbol{\Theta}$ from spacebased observations is clearly demonstrated. The rationale of the better performance by SVMs is discussed in many studies (e.g., Cortes & Vapnik, 1995; Gretton et al., 2001). It is briefly summarized here. First, the approach is relatively easy to use, because there are only a few parameters to adjust. The sparse setting of SVR, with the data training only based on the support vectors, avoids overfitting of the training data. By using the standard quadratic programming algorithms (Vapnik, 1995), only one global optimum is achieved. Mapping inputs into the high-dimensional feature space and introducing kernel function can solve the nonlinear relationship between inputs and outputs by turning a nonlinear regression to a linear fitting.

The present algorithm includes u_{850} as an additional input parameter, which marginally improves the water balance over South America, compared with an earlier algorithm using SVR. Because W can be measured by spacebased microwave radiometers, retrieval of $\boldsymbol{\Theta}$ essentially becomes the problem of estimating $\mathbf{u}_{\mathrm{e}}$. Directly mapping $\mathbf{u}_{\mathrm{s}}$ to $\mathbf{u}_{\mathrm{e}}$ only uses the vertically coherent part of wind profile. It is not sufficient over those areas when the surface wind is decoupled from the upper level and adding winds above the boundary layer should improve the methodology. The impact of adding the upper level wind will be evaluated comprehensively over the global ocean, and optimization of the SVR algorithm will be further investigated.

Acknowledgment

This study was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration (NASA). It was supported jointly by the Energy and Water System, the Physical Oceanography, the Precipitation Science, and the Ocean Vector Wind programs of NASA.

References

Blanz, V., Schölkopf, B., Bülthoff, H., Burges, C., Vapnik, V., & Vetter, T. (1996). Comparison of view-based object recognition algorithms using realistic 3D models. *Artificial neural networks ICANN'96* (pp. 251−256).

Boser, B., Guyon, I., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of 5th annual workshop on computer learning theory, Pittsburgh, PA* (pp. 144−152).

Bryan, F., & Oort, A. (1984). Seasonal variation of the global water balance based on aerological data. *Journal of Geophysical Research*, *89*, 11,717−11,730.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121−167.

Cherkassky, V., & Mulier, F. (1998). *Learning from data.* New York: John Wiley and Sons.

Cohen, J. L., Salstein, D. A., & Rosen, R. D. (2000). Interannual variability in the meridional transport of water vapor. *Journal of Hydrometeorology*, *1*, 547−553.

Cortes, C., & Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, *20*, 273−297.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155−161.

Gretton, A., Doucet, A., Herbrich, R., Rayner, P. J. W., & Schölkopf, B. (2001). Support vector regression for black-box system identification. *Statistical signal processing, proceedings of the 11th IEEE signal processing workshop* (pp. 341−344).

Hardle, W. (1990). *Applied nonparametric regression: Cambridge University Press 333 pp.*

Hayden, C. M., & Pursor, R. J. (1995). Recursive filter objective analysis of meteorological fields: Applications to NESDIS operational processing. *Journal of Applied Meteorology*, *34*, 3−15.

Heta, Y., & Mitsuta, Y. (1993). An evaluation of evaporation over the tropical Pacific Ocean as observed from satellites. *Journal of Applied Meteorology*, *32*, 1242−1247.

Horvath, A., & Davies, R. (2001). Feasibility and error analysis of cloud motion wind extraction from near-simultaneous MISR measurements. *Journal of Atmospheric and Oceanic Technology*, *18*, 591−680.

Joachims, T. (1999). Making large-scale SVR learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—Support vector learning* : MIT Press.

Liu, W. T. (1987). *1982–1983 El Nino atlas—Nimbus-7 microwave radiometer data.JPL Publication*, *vol. 87-5*. (pp.): Jet Propulsion Laboratory 68 pp.

Liu, W. T. (1993). Ocean surface evaporation. In R. J. Gurney, J. Foster, & C. Parkinson (Eds.), *Atlas of satellite observations related to global change* (pp. 265−278). Cambridge: Cambridge University Press.

Liu, W. T. (2002). Progress in scatterometer application. *Journal of Oceanography*, *58*, 121−136.

Liu, W. T., & Tang, W. (2005). Estimating moisture transport over ocean using spacebased observations from space. *Journal of Geophysical Research*, *110*, D10101. doi:10.1029/ 2004JD005300

Liu, W. T., Tang, W., & Niiler, P. P. (1991). Humidity profiles over ocean. *Journal of Climate*, *4*, 1023−1034.

Liu, W. T., Xie, X., Tang, W., & Zlotnicki, V. (2006). Spacebased observations of oceanic influence on the annual variation of South American water balance. *Geophysical Research Letters*, *33*, L08710. doi:10.1029/2006GL025683

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*, 169−186.

Mo, K. -C., & Higgins, R. W. (1996). Large-scale atmospheric moisture transport as evaluated in the NCEP/NCAR and the NASA/DAO reanalyses. *Journal of Climate*, *9*, 1531−1545.

Müller, K. -R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997). Predicting time series with support vector machines. *Proceedings of international conference on artificial neural networks* (pp. 999).

Peixoto, J. P., Salstein, D. A., & Rosen, R. D. (1981). Intra-annual variations in large-scale moisture fields. *Journal of Geophysical Research*, *86*, 1255−1264.

Rasmusson, E. M. (1967). Atmospheric water vapor transport and the water balance of North America. *Monthly Weather Review*, *95*, 403−426.

Roads, J. O., Chen, S. -C., Kao, J., Langley, D., & Glatzmaier, G. (1992). Global aspects of the Los Alamos general circulation model hydrological cycle. *Journal of Geophysical Research*, *97*, 10,051−10,068.

Rosen, R. D., Salstein, D. A., & Peixoto, J. P. (1979). Variability in the annual fields of large-scale atmospheric water vapor transport. *Monthly Weather Review*, *107*, 26−37.

Schölkopf, B. (1997). *Support vector learning.*München: R. Oldenbourg Verlag Doktorarbeit, TU Berlin. http://www.kernel-machines.org/

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199−222.

Starr, V. P., & White, R. M. (1955). Direct measurement of the hemispheric poleward flux of water vapor. *Journal of Marine Research*, *14*(3), 217−225.

Tang, B., & Mazzoni, D. (2006). Multiclass reduced-set support vector machines. *Proceedings of the 23th international conference on machine learning (ICML)*.

Trenberth, K. E., & Guillemot, C. J. (1995). Evaluation of the global atmospheric moisture budget as seen from analyses. *Journal of Climate*, *8*, 2255−2272.

Vapnik, V. N. (1995). *The nature of statistical learning theory: Springer.*

Vapnik, V. N. (1998). *Statistical learning theory.* New York: Wiley.

Wentz, F. J. (1997). A well-calibrated ocean algorithm for special sensor microwaver/imager. *Journal of Geophysical Research*, *102*(C4), 8703−8718.

Xie, X., & Liu, W. T. (2005). Hydrological budget in the Tropical Pacific. *16th conference on climate variability and change* : Amer. Meteor. Soc. http://ams.confex.com/ams/pdfpapers/85220.pdf