1.2 THE COMPARISON OF SEVERAL STATISTICAL RETRIEVAL METHODS USED IN RETRIEVING TEMPERATURE AND MOISTURE VERTICAL PROFILES

Kun Zhang* and P. Dong
Beijing Piesat Information & Technology Limited

1. INTRODUCTION

The rapid development of various observation techniques has made great contributions to the improvement of weather forecast technology. With the capacity to provide meteorological observation data covering the whole regions, meteorological satellite has been playing an important role in the areas of weather analysis and forecast, climate research, environment monitoring, disaster prevention and mitigation, etc. Among all the research contents, how to obtain atmospheric temperature and moisture information from satellite observation data is one of the critical issue in the field of remote sensing (Zhang et al., 2014). The obtaining process, which is the so called retrieval, can be divided into physical and statistical according to the methodology it uses. physical and statistical retrieval methods both have advantages and disadvantages when considering the performance and computational complexity.

Due to the timeliness and implementation simplicity, statistical regression is widely used as a kind of statistical retrieval method. In this paper, we will focus on other statistical retrieval methods: support vector regression(SVR) and neural network(NN), along with statistical regression. Experiments were carried out based on FY3 satellite observations and ERA-Interim's temperature and moisture profiles. Results indicated that, in terms of retrieval precision, neural network and support vector regression both are much better then statistical regression.

2. DATA AND METHODOLOGY

Assuming the vector composed by satellite observations and other information such as satellite scanning angle as the forecast factor $\boldsymbol{x}$, and the temperature and moisture profiles to be retrieved as $\boldsymbol{y}$. Statistical retrieval methods usually establish a function $y=\varphi(x)$ between $\boldsymbol{x}$ and $\boldsymbol{y}$, and φ is obtained based on a prepared training data set $X=[x_1,x_2,\cdots x_m]^T$ and $Y=[y_1,y_2,\cdots y_m]^T$. The data set of FY3 and ERA-Interim from Nov. 19th, 2014 to Nov. 20th, 2014 is treated as the training data, while the data set of Nov. 21st is used for validation. Depending on different kinds of weather situations and underlying surface, those data are divided into 5 kinds: ocean, clear sky; ocean, cloudy, no rain; ocean, rainy; land, clear sky; land, cloudy. The difference between each statistical retrieval methods are the ways they use to obtain φ . In the next part we will discuss the fundamental of three methods: statistical regression, neural network, and support vector regression respectively.

2.1 *Statistical Regression*

Statistical regression method (linear) establishes a linear relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$: $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$. $\boldsymbol{A}$ is calculated by solving an optimization problem:

$$\min \| \boldsymbol{A}\boldsymbol{X}^T-\boldsymbol{Y}^T \|_F^2 \tag{1}$$

It's easy to know that the solution to this problem is:

$$\boldsymbol{A}=\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2}$$

Smith and Wolf (1976) came up with an improved method based on the eigenvectors of covariance matrix of the training set. This method can reduce the dimension and the illness caused by the correlation between different channels. Assuming $\boldsymbol{M}$ is the covariance matrix of $\boldsymbol{X}$, $\boldsymbol{U}$ is the matrix composed by several eigenvectors of $\boldsymbol{M}$ (eigenvectors of some of the largest eigenvalues), and $\boldsymbol{Z}=(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{U}$, where

**Corresponding author address:* Kun Zhang, Beijing Piesat Information & Technology Limited, Beijing, MO 100195; email: zhang_kun@bjtu.edu.cn

$\overline{X} = [\overline{x}, \overline{x}, \cdots \overline{x}]^T$, $\overline{x} = \sum_{i=1}^{m} x_i / m$. Then the regression coefficient A is:

$$A = (Y - \overline{Y})^T Z (Z^T Z)^{-1} \quad (3)$$

where $\overline{Y} = [\overline{y}, \overline{y}, \cdots \overline{y}]^T$, $\overline{y} = \sum_{i=1}^{m} y_i / m$. And the relation between x and y is:

$$y = \overline{y} + A U^T (x - \overline{x}) \quad (4)$$

the same process also can be taken on Y to reduce the dimension and correlation of the parameters to be retrieved.

2.2 *Neural Network*

Neural network is composed of large numbers of artificial neurons. Each artificial neuron accepts inputs $x_1, x_2, ..., x_n$ from other neurons, and each input has a corresponding weight w_i . The output y can be expressed by the following equation:

$$y = f(\sum_{i=1}^{n} w_i x_i + b) = f(w^T x) \quad (5)$$

where $w = (w_1, w_2, ..., w_n, b)^T$, $x = (x_1, x_2, ..., x_n, 1)^T$, b is a threshold, f is the activation function. One of the most popular neural network is the back propagation neural network, which stacks several layers of artificial neurons together. Each layer accepts the outputs of the previous layer as inputs and transmits the calculated data to the next layer. Assuming W is the vector composed by all weights, B is the vector composed by all thresholds. F is the whole neural network, the training process of the neural network can be considered as solving an optimization problem:

$$\min_{W, B} \| F(X, W, B) - Y \| \quad (6)$$

where X and Y are the training input and output respectively. This problem can be solved with the back propagation algorithm, which is the origin of this neural network's name. Hornik (1989) demonstrated that three layers BP neural network can simulate any continuous, bounded input-output mapping to an arbitrary degree of accuracy. But we don't know how many neurons the network needed, and more layers has a natural advantage when simulating complex nonlinear mappings. For choosing the number of neurons, one strategy is to set few neurons at first, then gradually increase the number and analysis the performance of the network on a fixed training set, there are also some empirical formulas such as $\sqrt{n_1 + n_2} + a$, where n_1 and n_2 are the numbers of neurons of the input and output layer, respectively, a is a constant between 1 and 10.

2.3 *Support Vector Regression*

The basic principle of support vector regression is to maximize the generalization ability when it satisfies the preset regression accuracy. Assuming the training set is $(x_i, y_i), i = 1,2,..., m$. ε is a preset accuracy. Then the linear regression problem can be seen as to obtain a hyperplane, which satisfies:

$$| w^T x_i + b - y_i | \le \varepsilon, i = 1,2,..., m \quad (7)$$

because the distance between point (x_0, y_0) and hyperplane $y = w^T x + b$ is $| w^T x_0 + b - y_0 | / \sqrt{1 + \| w \|_2^2}$, so if we want more points satisfy (7) in practical application, function $\min \| w \|_2^2$ can be set as the objective function. Thus we obtain an optimization problem:

$$\min \frac{1}{2} \| w \|_2^2 + C * \sum_{i=1}^{m} (\xi_i + \xi_i^*)$$
$$s.t. \begin{cases} y_i - w^T x_i - b \le \varepsilon + \xi_i \\ w^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \\ i = 1,2,..., m \end{cases} \quad (8)$$

where *C* is a parameter, $\xi_i \ge 0, \xi_i^* \ge 0, i = 1,2,..., m$ are slack variables. This problem is not easy to solve, but Cristianini and Shawe-Taylor (2000) demonstrated that it is equivalent to it's dual problem:

$$\max_{a_i,a_i^*,r_i,r_i^*} -\frac{1}{2}\sum_{i,j=1}^{m}(a_i-a_i^*)(a_j-a_j^*)<\boldsymbol{x}_i,\boldsymbol{x}_j> \tag{9}$$

$$-\sum_{i=1}^{m}(a_i+a_i^*)\varepsilon+\sum_{i=1}^{m}(a_i-a_i^*)y_i$$

$$s.t.\begin{cases}0\le a_i\le C\\ 0\le a_i^*\le C\\ \sum_{i=1}^{m}(a_i-a_i^*)=0\\ i=1,2,...,m\end{cases}$$

and this problem can be efficiently solved by sequential minimal optimization algorithm (Platt, 1998). The dual problem also has another advantages: input vectors $\boldsymbol{x}$ are always in the form of inner product $<\boldsymbol{x}_i,\boldsymbol{x}_j>$, This feature is the key point of using kernel function. All of the above are talking about the linear regression, when it comes to the nonlinear case, one direct approach is mapping the input data $\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_m$ into a high dimensional feature space (Hilbert space) $\Omega(\boldsymbol{x}_1),\Omega(\boldsymbol{x}_2),...,\Omega(\boldsymbol{x}_m)$, then the problem comes back to the linear case. But $\Omega(\boldsymbol{x}_i)$ usually has a very high dimension which makes the problem very complicate, even unable to solve. Fortunately, there is a kind of function called kernel function which satisfies $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=<\Omega(\boldsymbol{x}_1),\Omega(\boldsymbol{x}_2)>$, thus the dimension explosion problem can be solved perfectly, we can keep the idea of mapping data into high dimensional space, but all the calculations are just done in the original space. There are some popular used kernel functions, for example: polynomial kernel function $K(\boldsymbol{x}_1,\boldsymbol{x}_2)=(<\boldsymbol{x}_1,\boldsymbol{x}_2>+1)^d$, Gauss kernel function $K(\boldsymbol{x}_1,\boldsymbol{x}_2)=\exp(-\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2/(2\sigma^2))$.

3. RESULTS

In the experiment of neural network, we considered two kinds of network structure: one hidden layer and two hidden layers. For the one hidden layer network, we use 20 hidden nodes, and for the two hidden layers network, we use 15 and 5 hidden nodes for each layer, respectively. Fig.1 and Fig.2 show the RMSE of retrieved temperature and moisture profiles under these two network structures with a clear sky on land and on ocean respectively. From Fig.1 and Fig.2 we can know that when retrieving temperature profiles, single hidden layer and two hidden layers have about the same effect. But when it comes to the moisture profiles which have a high nonlinearity, two hidden layers is obviously superior to single hidden layer.

Fig.3 shows the comparison of these three methods on ocean with a clear sky (for the neural network we use two hidden layers). As you can see, when retrieving temperature profiles, the RMSE distribution of neural network and support vector regression are almost the same, while retrieving moisture profiles, neural network has a better precision. And these two methods both are much more better then statistical regression. Fig.4 shows the comparison of these three methods on land with a clear sky. From Fig.4 we can see that both neural network and support vector regression have a similar RMSE distribution of retrieved temperature and moisture profiles. For temperature profiles, these two methods are superior to statistical regression to a certain extent. For moisture profiles, these two methods are much more better then statistical regression.

Table.1 shows the average RMSE of retrieved temperature and moisture profiles of all layers. From Table.1 we can see that neural network and support vector regression both have a better performance compared with statistical regression under each condition, which provides new ways for operational use.

4. SUMMARY

In this paper we briefly narrated the principle of three statistical retrieval methods: statistical regression, neural network, and support vector regression. Experiments are carried out using FY3 observations and ERA-interim's profile data to test and compare these three methods. Results indicated that: 1. for neural network method, two hidden layers performs better in retrieving moisture profiles then one hidden layer, while in retrieving temperature profiles, they both have a similar effect; 2. neural network and support vector regression both are much more better then statistical regression under all weather and underlying conditions.

But neural network and support vector regression also have disadvantages, their performance depend on the parameters to a large extent, the parameters should be choosing carefully. Our future work will concentrate on the sensitivity to parameters and ways to choose the parameters.

5. REFERENCES

Zhang, J., et al., 2014: Ensemble retrieval of atmospheric temperature profiles from AIRS. *Advances in Atmospheric Sciences,* **31.3**, 559.

Smith, W. L., and H. M. Woolf, 1976: The use of eigenvectors of statistical covariance matrices for interpreting satellite sounding radiometer observations. *Journal of the Atmospheric Sciences,* **33.7**, 1127-1140.

Hornik, K., M. Stinchcombe, and H. White, 1989: Multilayer feedforward networks are universal approximators. *Neural networks,* **2.5**, 359-366.

Cristianini N., J. Shawe-Taylor, 2000: An introduction to support vector machines and other kernel-based learning methods[M]. *Cambridge university press.*

Platt, J. C., 1998: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in Kernel Methods-support Vector Learning,* **208**, 212-223.

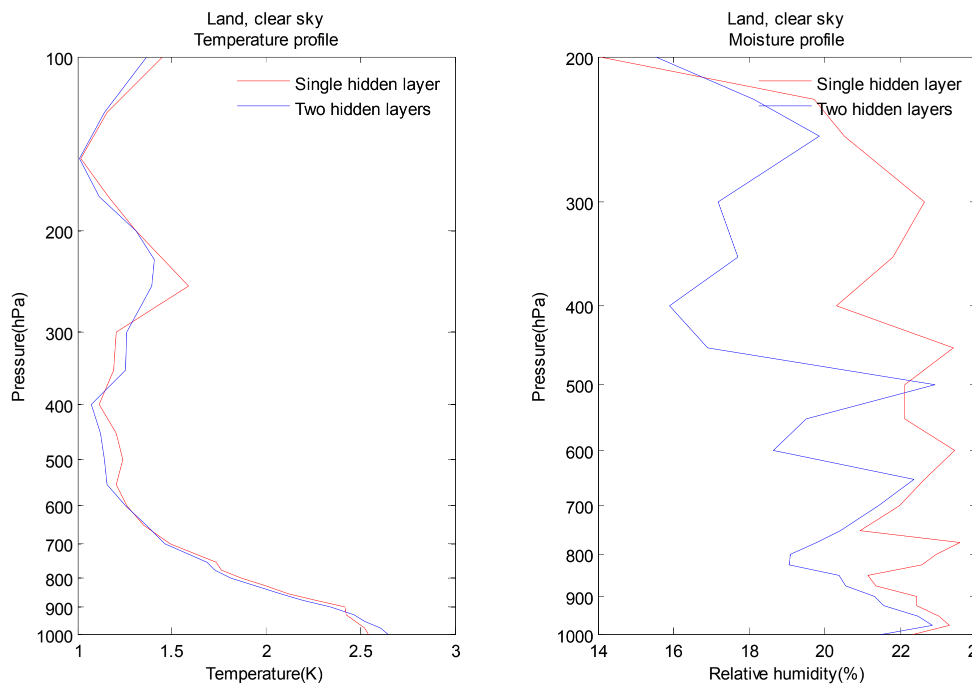

Figure 1. RMSE of retrieved temperature and moisture profiles on land with a clear sky under two network structures.

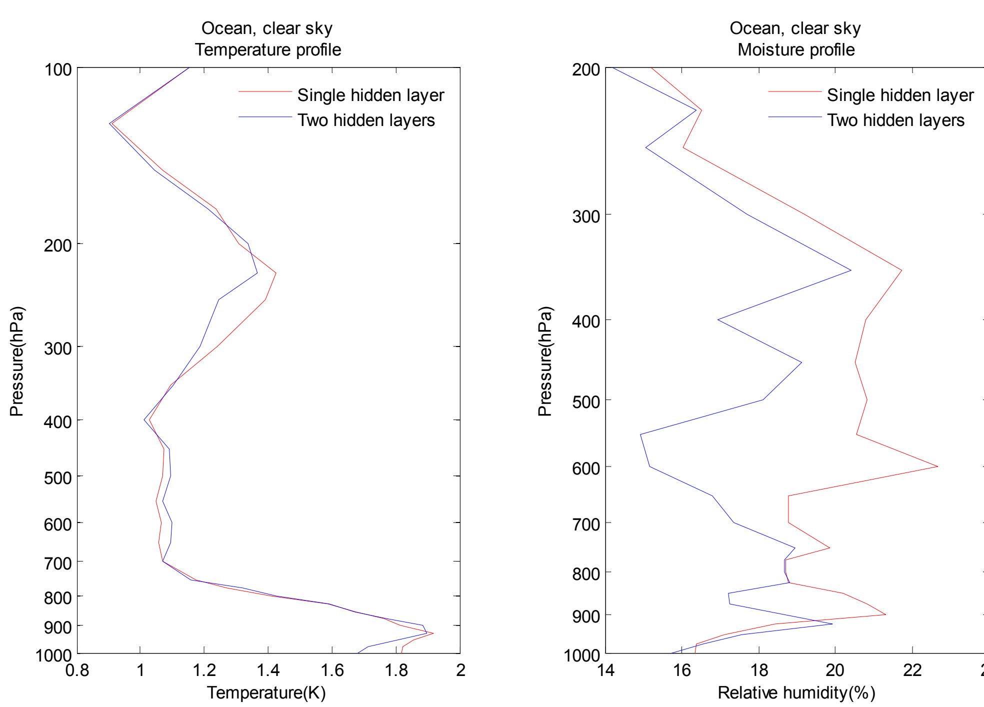

Figure 2. RMSE of retrieved temperature and moisture profiles on ocean with a clear sky under two network structures.

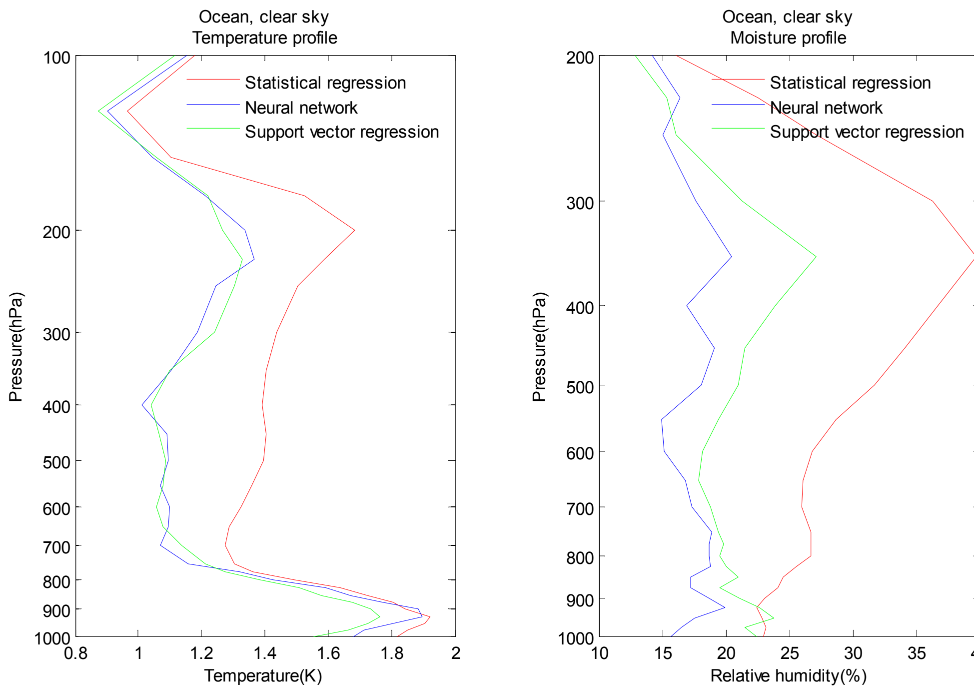

Figure 3. RMSE of retrieved temperature and moisture profiles on ocean with a clear sky under three methods.

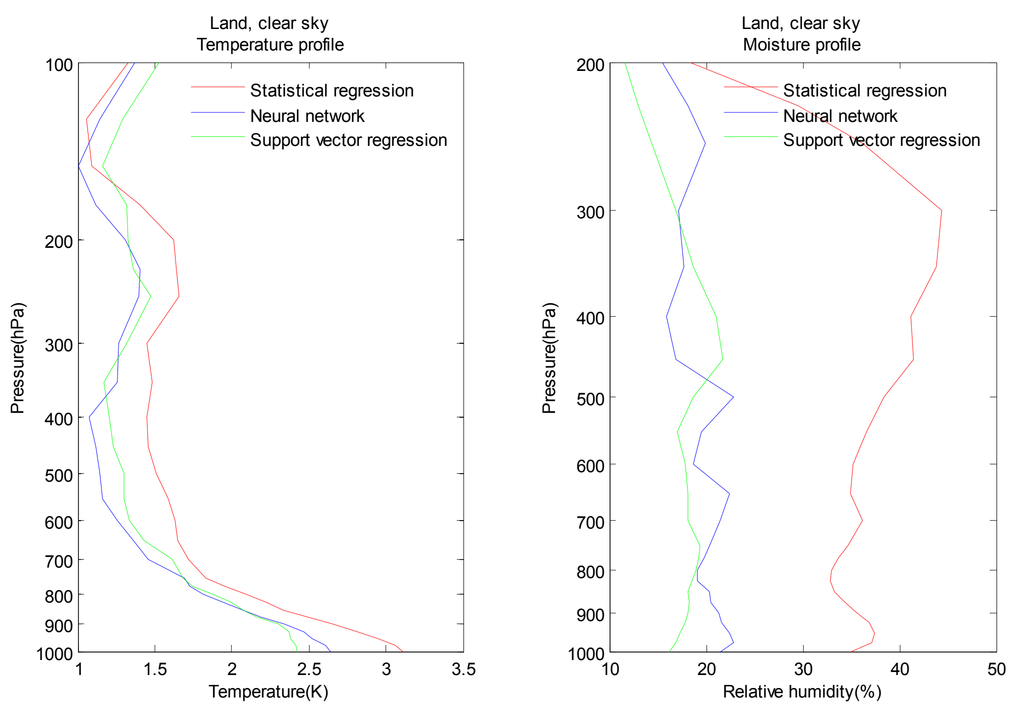

Figure 4. RMSE of retrieved temperature and moisture profiles on land with a clear sky under three methods.

Table 1. Average RMSE of temperature and moisture profiles of three methods under all conditions

		Average RMSE of temperature profiles（K）	Average RMSE of moisture profiles (%)
Ocean, clear sky	Statistical regression	1.50	27.0
	NN	1.33	17.4
	SVR	1.30	20.2
Ocean, cloudy, no rain	Statistical regression	1.70	32.4
	NN	1.45	20.6
	SVR	1.36	23.7
Ocean, rainy	Statistical regression	1.80	54.3
	NN	1.46	29.7
	SVR	1.40	24.7
Land, clear sky	Statistical regression	1.90	35.6
	NN	1.62	19.8
	SVR	1.66	17.6
Land, cloudy	Statistical regression	2.37	51.8
	NN	1.89	24.2
	SVR	1.80	24.9