APPLICATION OF PRINCIPAL COMPONENT ALGORITHM TO THE DIRECT ASSIMILATION OF AIRS IN THE WRF 4D-VAR SYSTEM

Yu Yi[i 1,2], Zhang Weimin[1], Huang Qunbo[1], Ye Minhua[2], Sun jing[2]

1. Computer Science, National University of Defense Technology, Changsha, China
2. NO.94865 Troops of Chinese People's Liberation Army, Hangzhou, China
yuyi2019@nudt.edu.cn wmzhang104@163.com hqb09@163.com xyz01326@163.com 1073679954@qq.com

***Abstract*--The use of Principal Component (PC) algorithm is explored for the efficient representation observations from high-resolution infrared sounders for the purposes of data assimilation into numerical weather prediction (NWP) models. A new version of the fast radiative transfer model has been developed that exploits principal component analysis and then implemented into the WRF 4D-Var data assimilation system, thus allow the investigation of the direct assimilation of PC scores from Atmospheric Infrared Sounder (AIRS). Testing of a prototype system where 119 AIRS spectra replaced by only 20 PC scores show significant computational saving with no detectable loss of skill in the resulting analyses or forecasts. The methodologies implemented in this regard are examined and the potential for future increased use of the data are explored.**

***Keywords*-Principal Component Analysis; AIRS; PCRT; Data Assimilation**

I. INTRODUCTION

On 4th May 2002 the AIRS (Atmospheric InfraRed Sounder, Aumann et al., 2003) instrument was launched on the NASA EOS-AM (Aqua) satellite. AIRS measures the radiance emitted from the Earth in 2378 channels covering the spectral interval from 650 to 2665 cm^{-1} at a resolution of a around 1 cm^{-1}. Operational assimilation of AIRS radiance began at the European Centre for Medium-Range Weather Forecasts (ECMWF) in October 2003, this being the first assimilation of data from a kilochannel infrared sounder into an operational numerical weather prediction (NWP) system [1]. The assimilation of high-resolution radiances measured by AIRS has produced a significant positive impact on forecast quality [2]. The kilochannel infrared sounder AIRS can provide information with high accuracy and vertical resolution though the use of the large number of low-noise channels. In the last few years, infrared sounders with thousands of channels have become a major part of many NWP centers [3].

The high volume of data resulting from these observations presents many challenges, particularly in the areas of data transmission, data storage and assimilation [4]. Processing the complete AIRS spectrum is inefficient and would impose an infeasible burden in the assimilation system [1]. The operational use of AIRS radiance at ECMWF is restricted to a selection of temperature sounding channel in the long-wave region of the spectrum and to a very limited number of humidity sounding channels in the main infrared water vapour band [5]. Up to 155 AIRS channels were chosen to be assimilation.

In principle, to exploit the full information content of AIRS, the number of channels used in the assimilation could be increased to cover the full spectrum. NWP users are limited to assimilating less than the full AIRS spectrum by the prohibitive computational cost, but there are fewer pieces of independent information in a AIRS spectrum than channels by around tow orders of magnitude [6]. Channels that are similar in content are highly correlated to each other. There is thus a need to find a more efficient way of communicating the measured information to the analysis system than

simply increasing the number of channels. Various methods for compressing the available information have been suggested to do this, including the use of principal component analysis (PCA) which is the subject of this paper.

The outline of the paper is as follow. Sections 2 and 3 outline respectively the theoretical properties of PCA and discuss the features of PC data derived from AIRS radiance spectrum. Section 4 present the methodology of direct assimilation of PC scores using the WRF 4D-Var system. Finally, conclusions are given in section 5.

II. THE THEORY OF PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a classical statistical method for the efficient encapsulation of information from voluminous data [7], and it allows the reduction of the dimensionality of a problem by exploiting the linear relationship between all the variables contained in a multivariate dataset. The reduction of the dimension of the set is obtained by replacing the original set of correlated variables with a smaller number of uncorrelated variables called principal components. Since the new derived variables retain most of the information contained in the original dataset, they can be used to efficiently represent the data.

Suppose our dataset consists of *l* spectra of *n* radiances into an *l* by *n* data matrix ***R***. The dataset can then be represent by the vector population r = $(r_1, r_2, \ldots . r_n)^T$ (here T denotes the transpose). If ***C*** is then *n* by *n* covariance matrix arranged as row vectors in descending order according to the magnitude of their eigenvalues, thus:

$$C = \frac{1}{n_{obs}} RR^T = A\Lambda A^T \qquad (1)$$

Where A is the matrix of principal components and Λ their associated eigenvalues.

The principal component, *p*, related to the observed radiance can be written as:

$$p = A\, r \qquad (2)$$

The eigenvectors represent the directions of maximum variance in the data; consequently, each PC gives the linear combination of the variables that provides the maximum variation. The PCs are orthogonal, hence uncorrelated (although this does not imply that they are statistically independent), and the values associated to each spectrum are known as PC scores. If λ_i is the eigenvalue associated with the i^{th} eigenvector, then the value of $\lambda_i / \sum_{i=1}^{n} \lambda_i^2$ gives the proportion of variation explain by the i^{th} PC. Since the matrix A is orthogonal, its inverse is equal to its transpose and we can write:

$$r = A^T p \qquad (3)$$

We can then reduce the dimension of the problem by replacing the n original variables with the first *m* PCs. In many cases, the choice of the number of dimensions is based on the total variation accounted for by the leading PCs and it will in general depend on specific aspects of the original dataset.

For any new observed radiance spectrum, r^{obs}, we can compute the equivalent PC scores by projecting the radiances upon the full set of eigenvectors derived from the covariance matrix of the training dataset. As discussed above, less than n eigenvectors are typically required to reduce most of the information in the observed spectrum. Therefore, we can compute a vector of m truncated observed PC scores, p^{obs}:

$$p_i^{obs} = \sum_{k=1}^{n} A_{i,k} r_k^{obs} \qquad (4)$$

Where i=1,m. The truncated PC scores may be regarded as efficient encapsulation of the original observation that may be used for storage, transmission or indeed assimilation.

Direct assimilation of principal components requires a suitably fast and accurate radiative transfer model, this model can calculate principal components directly rather than first 2378 individual radiances. We can develop a PCA-based fast RT model by the main steps: the generation of a training

database of LBL radiances and the development of an algorithm to predict accurate PC scores for any atmospheric profile not included in the training dataset.

III. AIRS SPECTRAL BAND AND PC GENERATION

For the purpose of investigation of the suitability of PCA as a mechanism to efficiently present this information to an assimilation system, it is assumed that we have access to the full AIRS measured spectrum.

AIRS is an infrared radiometer on Auq with 2378 channels, covering the infrared part of the spectrum in three bands, 650.0-1136.6 cm^{-1}, 1217.0-1613.9 cm^{-1} and 2181.5-2665.2 cm^{-1}. AIRS is flow together with an Advanced Microwave Sounding Unit (AMSU)-A, and 3 $\times$3 AIRS fields of view (FOVs) are sampled per AMSU-A FOV. We only consider FOVs for which all channels currently considered for assimilation are used in the assimilation system, i.e. only FOVs for which all active channels are diagnosed as cloud-free and are not removed by other quality-control procedure.

Up to 119 AIRS channels are used in the assimilation configuration used in this study; most of these are in the long-wave CO2 band . The use of the water-vapour band is restricted to seven AIRS channels. Cloud screening aims to identify clear channels based on evaluating FG-departure signature. The scheme is applied to temperature- sounding signatures. For the water-vapour band, the cloud-screening is linked to the results from the temperature-sounding channels. Up to 48 stratospheric AIRS channels are used over land and not sensitive to the surface.

In this study we only consider the assimilation of PC scores generated from the 119 AIRS channels chosen by method presented above. The conversion of radiance to PC scores is carried out using Eq.4, i.e. by projecting the radiance vector on the fixed basis of synthetic eigenvectors utilized in the PC based fast radiative transfer model used in the assimilation. Note the eigenvectors are derived from the 119 channels described above.

IV. PC ASSIMILATION METHODOLOGY

A. The architecture of PC scores 4D-Var system

The basic idea of the four-dimensional (4D) variational assimilation [4] for atmospheric data X is to optimally combine observations over short time window and the corresponding model short-term forecast X_B on the model grid.

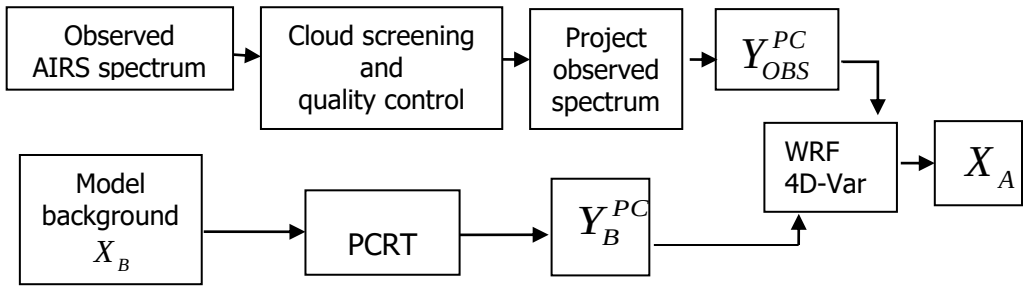

Figure1. The flow diagram of direct PC scores assimilation

The atmospheric state (X) are used as input to the observation operator PCRT to compute model equivalent of m PC scores, $Y_B^{PC}(X)$. If we ignore the time integration of the forecast model to the observations, the cost function to be minimized is essentially:

$$J(X) = [X - X_B]^T B^{-1} [X - X_B] + [Y_{OBS}^{PC} - Y_B^{PC}(X)]^T O^{-1} [Y_{OBS}^{PC} - Y_B^{PC}(X)] \quad (5)$$

Here, X_B is the accuracy of the background estimate of the atmospheric state and is described by the error covariance. The error covariance *O* describes the accuracy of the observations and associated. The atmospheric states X_A referred to as *analysis* derived from the minimization of the cost function. The departures of the analysis X_A from X_B are referred to as analysis increments defined at the start of the 4D-Var window.

The methodology adopted in this paper for the direct 4D-Var assimilation of PC scores is shown in Figure 1.The observed AIRS spectrum are first screened for the presence of clouds and contaminated spectrum are discarded. Then project the clear spectrum on to the 119 channel basis described previously, to produce a vector of PC scores Y_{OBS}^{PC}. Each vector of observed PC scores has length n=119, but we assimilate only the first m of these (where m < n, and in ranked order according with the descending eigenvalues). All the channels are completely clear data and none of the eigenvectors correspond to cloud signals.

B. The observation operator PCRT model

The observation operator is a PCA-based fast radiative transfer model named PCRT model which performs rapid and accurate simulation of PC scores of AIRS radiance using a multiple linear regression scheme. The training of the PCRT model includes the steps as follow:

1) Select the atmospheric profile training set for diverse atmospheric situations.
2) Input the atmospheric profile to the standard fast radiative transmittance (RT) model which is based on the computation of a database of line-by-line (LBL) spectra.
3) The standard fast RT model computes the polychromatic radiance from the database of LBL spectra.
4) Select a number of polychromatic radiances to consist the predictors which is profile-dependent.
5) Use the accurate LBL model to calculate a large dataset of synthetic noise-free clear sky radiance.
6) Compute the regression coefficients by using the PC scores obtained from the eigenvectors of the covariance matrix of the large dataset derived from the step 5).
7) Input the predictors from the step 4) to the multiple linear regression scheme and use the linear combination of the predictors to express the PC scores.

The key step 6) is based on principal component analysis and the function is implemented as follows:

```
IF (LHOOK) CALL DR_HOOK('PCRT_PCSCORES',
0_jpim,ZHOOK_HANDLE)
 k = 1
 DO prof = 1, nprofiles
  DO i = 1, nchannels/nprofiles
   chan = chanprof(i)%chan
   DO j = 1 + (prof-1) * npcscores/nprofiles, prof *
          npcscores/nprofiles
     chan_pc = chanprof_pc(j)%chan
     pccomp%pcscores(j) = pccomp%pcscores(j) +
&& coef_pccomp%pcreg(opts%ipcreg)%coefficients(i,
chan_pc) * radiancedata%clear(k) /
coef_pccomp%noise(chan)
   ENDDO
   k = k + 1
  ENDDO
 ENDDO
```

Figure2. The algorithm based on PCA.

C. The radiance reconstruction error

We can reconstruct any spectrum to a given accuracy using a truncated number of principal components, *p*. This number will in general depend on the number of observations, *n*, and variables, *m*, in the dataset. In practical terms, we require that the value of p is such that for each channel, *i*, the radiance reconstruction error, ε_i, is below the instrument noise, η_i:

$$\varepsilon_i = \sqrt{\frac{\sum_{j=1}^{n} \left(\frac{1}{m}\sum_{k=1}^{m} A_{k,i} y_{k,j} - \frac{1}{p}\sum_{k=1}^{p} A_{k,i} y_{k,j}\right)^2}{n}} < \eta_i \quad (6)$$

The value of p can tuned to achieve filtering of the observation. It is argued that the atmospheric signal is more highly correlated across the spectrum than the random instrument noise [8]. As the atmospheric signal can be represent by the high rank eigenvectors (i.e. those with larger eigenvalues) and the instrument noise represented by the low rank eigenvectors, using PCA can separate variations of the atmospheric signal from variations of the random instrument noise. Generally the number of principal components, p, is far smaller than the number of variables, m, using PCA reduces the dimension of observation information. In this paper the atmospheric signal from 119 AIRS channels is encapsulated to 20 PC scores.

V. CONCLUSIONS

Testing of the principal component based prototype system shows significant computational savings with no detectable loss of skill in the resulting analysis of forecasts. Performance tests indicate that the WRF 4D-Var minimization requires 21% less computer resources (elapsed CPU time) when 20 PC scores are used instead of 119 AIRS radiance-achieving an 6 fold reduction in data volume and 21% reduction in the overall cost of assimilation. This figure represents the computational efficiency is significant improved, but could be improved even further. Indeed in some respects the assimilation of PC scores leads to marginal improvements over the traditional radiance based assimilation.

REFERENCES

[1] Collard AD. 2007. Assimilation of AIRS and IASI at ECMWF. *In Proceeding of ECMWF Seminar on Recent development in the use of satellite observations in NWP, 3-7 Sept 2007*. ECMWF Report, Reading, U.K.

[2] McNally AP, Watts PD, Smith JA, Engelen R, Kelly GA, Thepaut JN, Matricardi M. 2006. The assimilation of AIRS radiance data at ECMWF. *Q. J. R. Meteorol*. 132:935-957.

[3] Kelly G, Thepaut J-N. 2007. 'Evaluation of the impact of the space component of the Global Observing System through Obsering System Experiments'. *ECMWF Newsltetter* No.112.

[4] Collard AD, McNally AP, Hilton FI, Healy SB, Atkinson NC. The use of principal component analysis for the assimilation of high-resolution sounder observations for numerical weather prediction. *Q. J. R. Meteorol. Soc*. 136: 2038-2050, October 2010 Part B.

[5] Martricardi M. A principal component based version of the RTTOV fast radiative transfer model. *Q. J. R. Meteorol. Soc*. 136: 1823-1835, October 2010 Part A.

[6] Huang H-L, Smith WL, Woolf HM. 1992. Vertical resolution and accuracy of atmospheric infrared sounding spectrometers. *J. Appl. Meteorol.* 31:265-274.

[7] Jolliffe IT. 2002. Principal Component Analysis. Springer: New York.

[8] Martricardi M, McNally AP. The Direct Assimilation of Principal Components of IASI Spectral in the ECMWF 4D-Var. *Q. J. R. Meteo. Roc.* No.690, December 2012.