Towards ensemble assimilation of hyperspectral satellite observations with data compression and dimension reduction using principal component analysis

Yinghui Lu and Fuqing Zhang

Department of Meteorology and Atmospheric Science, and Center for Advanced Data Assimilation and Predictability Techniques, The Pennsylvania State University, University Park, Pennsylvania

Submitted to *Monthly Weather Review for publication as an article*

December 2018

Corresponding author: Professor Fuqing Zhang, Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, PA 16802.

E-mail: fzhang@psu.edu

Early Online Release: This preliminary version has been accepted for publication in *Monthly Weather Review*, may be fully cited, and has been assigned DOI 10.1175/MWR-D-18-0454.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2019 American Meteorological Society

Abstract

Satellite-based hyperspectral radiometers usually have thousands of infrared channels which contain atmospheric state information with higher vertical resolution compared to observations from traditional sensors. However, the large numbers of channels can lead to computational burden in satellite data retrieval and assimilation. Furthermore, most of the channels are highly correlated and the pieces of independent information contained in the hyperspectral observations are usually much smaller than the number of channels. Principal Component Analysis (PCA) was used in this research to compress the observational information content contained in the Atmosphere Infrared Sounder (AIRS) channels to a few leading principal components (PCs). The corresponding PC scores were then assimilated into a PCA-based ensemble Kalman filter (EnKF) system. In this proof-of-concept study based on simulated observations, hyperspectral brightness temperatures were simulated using the atmospheric state vectors from convection-permitting ensemble simulations of Hurricane Harvey (2017) as input to the Community Radiative Transfer Model (CRTM). The PCs were derived from a pre-existing training dataset of brightness temperatures calculated from convection-permitting simulation over a large domain in the Indian Ocean representing generic atmospheric conditions over tropical oceans. The EnKF increments from assimilating many individual measurements in the brightness temperature space were compared to the EnKF increments from assimilating significantly fewer numbers of leading PCs. Results showed that assimilating about 10 to 20 leading PCs could yield increments that were nearly indistinguishable to that from assimilating hyperspectral measurements from orders of magnitude larger number of hyperspectral channels.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

1. Introduction

Hyperspectral infrared radiometers such as the Atmosphere Infrared Sounder (AIRS; Aumann et al. 2003) on the Earth Observing System (EOS) Aqua platform and the Infrared Atmospheric Sounding Interferometer (IASI; Hilton et al. 2012) on MetOp series provide retrieval profiles with better accuracy and higher vertical resolution than traditional broadband radiometers such as the Advanced Baseline Imager (ABI) onboard GOES-16 (Schmit et al. 2005, 2017) because of their higher spectral resolution and broader spectral coverage compared to traditional sounders (Huang et al. 1992). Assimilating selected temperature sounding channels and humidity sounding channels from these hyperspectral observations provides positive impact on both global models and regional models (McNally et al. 2006; Collard and McNally 2009; Guidard et al. 2011; Xu et al. 2013). Although it is tempting to assimilate all the channels to exploit the full information content in these hyperspectral observations, its high computational cost can be prohibitive, especially in the near-real time operational environment. Moreover, although thousands of channels are available on these instruments (e.g., 2378 channels on AIRS and 8461 channels on IASI), most channels are highly correlated, resulting in much fewer pieces of independent information than the number of channels (Huang et al. 1992).

Carefully selecting a subset of all the channels that contains as much independent information as possible is one way to balance computational cost and information content obtained (e.g., Goldberg et al. 2003; Susskind et al. 2003). Alternatively, the full information content (or part of it) of hyperspectral observations can be compressed using dimension reduction techniques such as the Principal Component Analysis (PCA;

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

Goldberg et al. 2003; Huang and Antonelli 2001; Aires et al. 2002), which is also called Empirical Orthogonal Function (EOF; Hannachi et al. 2007; Monahan et al. 2009; North 1984). Besides data compression, another benefit from these PCA-based approaches is noise reduction. It can be argued that variations of atmosphere signals are more correlated across the hyperspectral channels while variations of random noises are less correlated. Thus, after PCA, the atmosphere signals are more likely to be represented by the leading principal components (i.e., those with larger eigenvalues) while noises are more likely to be represented by the lower-rank principal components (i.e., those with smaller eigenvalues). A truncated number of leading principal components (PCs), with significantly fewer variables than the number of channels, can be used to reproduce the original hyperspectral observations with reduced noise (Huang and Antonelli 2001; Antonelli et al. 2004; Turner et al. 2006).

PCA-compressed data have been used in data assimilation approaches in various ways. Collard et al. (2010) assimilated reconstructed IASI radiances from the principal components. Matricardi and Mcnally (2014) directly assimilated PC scores instead of radiances over clear-sky condition in the European Centre for Medium-range Weather Forecasts (ECMWF) 4D-Var data assimilation system. They showed that directly assimilating 20 PC scores instead of 165 IASI radiances leads to significant computational savings with no detectable loss of skills.

PCA-based radiative transfer forward models have been developed to efficiently simulate hyperspectral radiances or brightness temperatures (BTs) from atmospheric states (temperature, water vapor, and trace gas profiles, etc.). Examples are the PCA-based version of Radiative Transfer for TOVS (PC_RTTOV, Matricardi, 2010) and the Principal

Component-based Radiative Transfer Forward Model (PCRTM) developed by Liu et al. (2006). These PCA-based radiative transfer models directly simulate PC scores using pre-trained PCs. If channel radiances or brightness temperatures are needed, they can be reconstructed using PC scores and the pre-trained PCs provided by the PCA-based forward models. The overall execution speed is several to tens of times faster than channel-based fast forward models. If PC scores are directly used in subsequent retrieval or data assimilation (e.g., Matricardi and McNally 2014), extra time savings can be obtained since the reconstruction of brightness temperatures or radiances is not necessary.

The typical approach of infrared satellite data assimilation used by operational centers is to only assimilate radiances not affected by clouds (Geer et al. 2018), which includes not only cloud-free conditions, but also channels unaffected by clouds (e.g., stratospheric sounding channels in cloudy conditions) For cases where clouds and precipitation are ubiquitous, all-sky data assimilation shows positive impact on the perdition skill especially for short-range forecasting (Zhang et al. 2016; Minamide and Zhang 2019, 2018). Wu et al. (2017) showed that PCRTM could be used in physical retrieval for IASI cloudy scene analysis, illustrating possible use of PCA-based radiative transfer models in data assimilation under all-sky conditions. These findings indicate that assimilating hyperspectral observations using PCA-based data assimilation method under all-sky conditions is worth exploring, in particular for the ensemble-based data assimilation techniques.

In this paper, we explore the possibility of directly assimilating PC scores instead of original hyperspectral observations under the Ensemble Kalman Filter (EnKF) framework for all-sky conditions. The outline of this paper is as follows. Section 2 discusses the

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

methodology including a brief review of PCA and PC scores in satellite data applications, and the proposed approach in direct assimilation of PC scores with the EnKF. Section 3 describes data compression and dimension reduction of hyperspectral measurements with PCA. Section 4 shows data assimilation experiments with EnKF update using PC scores as innovation vector. Section 5 provides summary and discussion.

2. Methodology

2.1 Brief review of PCA and PC scores in satellite data applications

Principal component analysis is a widely used method for dimension reduction, data compression, and noise reduction. Its application in data application has been described in a number of papers with slightly different approaches and notations (e.g., Collard et al. 2010; Matricardi and McNally 2014). Here we briefly summarize PCA in the context of hyperspectral infrared observations as follows.

Consider a training dataset consists of n_{obs} hyperspectral observation samples $\mathbf{y}_i (i = 1, 2, \dots, n_{obs})$, each with n_{ch} channels, where the boldface non-italic font represents a multidimensional vector. In principle, the hyperspectral observations can be any multidimensional variables, such as radiances or normalized radiances respect to instrument noise. For frequencies where the majority of observed radiances are from thermal emissions of the atmosphere and the Earth surface, such as longwave infrared and microwave, radiances are often converted to brightness temperatures. Usually, PCA are trained using a large dataset that represents the range of variations in atmospheric conditions, including absorber amounts (e.g., water vapor and trace gas), temperature profiles, and surface parameters (Matricardi, 2010), and it can be assumed that $n_{obs} > n_{ch}$.

The mean of all hyperspectral observations of the training dataset is $\overline{\mathbf{y}^{tr}} = \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} \mathbf{y}_i$, which is a vector with dimension n_{ch}, and each element of $\overline{\mathbf{y}^{tr}}$ is the mean value of all n_{obs} observations for a given channel. The deviations of all hyperspectral observations from the mean are arranged into a $n_{ch} \times n_{obs}$ matrix $\mathbf{Y} = (\mathbf{y}'_1, \mathbf{y}'_2, \ldots, \mathbf{y}'_{n_{obs}})$, where each column of $\mathbf{Y}$ is the deviation of one hyperspectral observation sample from the training dataset mean $\overline{\mathbf{y}^{tr}}$, i.e., $\mathbf{y}'_i = \mathbf{y}_i - \overline{\mathbf{y}^{tr}}$, and each row of $\mathbf{Y}$ contains n_{obs} observation deviations at each channel. The covariance matrix $\mathbf{C}$ of matrix $\mathbf{Y}$ can be written as

$$\mathbf{C} = \frac{1}{n_{obs}-1} \mathbf{Y}\mathbf{Y}^{\mathrm{T}}, \quad (1)$$

where the superscript T denotes matrix transpose. Through eigenvalue decomposition, covariance matrix $\mathbf{C}$ can be decomposed as

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}, \quad (2)$$

where $\mathbf{\Lambda}$ is a diagonal matrix with all the eigenvalues λ_i of $\mathbf{C}$ arranged by their magnitude, and $\mathbf{U}$ is an orthogonal matrix consists of corresponding eigenvectors of $\mathbf{C}$ as column vectors. These column vectors of matrix $\mathbf{U}$ are the *principal components* (PCs) that lay along the directions of maximum variance of the dataset. With the assumption that $n_{obs} > n_{ch}$, the dimension of $\mathbf{U}$ is $n_{ch} \times n_{ch}$ and the eigenvectors can be expected to be nonzero.

The *PC score* $\mathbf{z}$ of a multidimensional vector $\mathbf{y}$ that represents one hyperspectral observation, either in the training dataset or not, can be obtained by first subtracting the mean hyperspectral observation of *the training dataset* $\overline{\mathbf{y}^{tr}}$ and then projecting the deviation $\mathbf{y} - \overline{\mathbf{y}^{tr}}$ on the directions of the PCs obtained from the training dataset, i.e.,

$$\mathbf{z} = \mathbf{U}^{\mathrm{T}}(\mathbf{y} - \overline{\mathbf{y}^{tr}}). \quad (3)$$

From Eq. (3), the PCA process can be interpreted as a linear combination of the original

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

hyperspectral channels into a set of synthesized "super channels", where each PC (each column of matrix $\mathbf{U}$) contains the weights of all hyperspectral channels for the corresponding "super channel". Then the PC scores are the "measurements" at these super channels. Note that the PCs and mean hyperspectral observation $\overline{\mathbf{y}^{tr}}$ trained from the training dataset should always be used together. Consequently, the hyperspectral observation can be reconstructed from the PC scores and the PCs by

$$\mathbf{y} = \mathbf{U}\mathbf{z} + \overline{\mathbf{y}^{tr}} . \quad (4)$$

The variance explained by the i^{th} PC is given by the eigenvalue associated with it, λ_i. The proportion of variance explained by the i^{th} PC, or the *variance ratio*, is given by $\lambda_i / \sum_j \lambda_j$. As will be shown later, λ_i decreases rapidly with increasing i, indicating the majority of variance in the hyperspectral observations can be expressed using a smaller number of leading PCs. Using only the first n_{pc} PCs and corresponding PC scores leads to a lossy compression of the original hyperspectral observation:

$$\tilde{\mathbf{y}} = \mathbf{U}_{npc}\mathbf{z}_{npc} + \overline{\mathbf{y}^{tr}}, \quad (5)$$

where $\mathbf{U}_{npc}$ is a $n_{ch} \times n_{pc}$ matrix with the first n_{pc} PCs as its columns and $\mathbf{z}_{npc}$ is a column vector with the first n_{pc} PC scores. When more PCs are used, the reconstruction error, i.e., the difference between the reconstructed and original hyperspectral observation, is smaller. However, noises in the lower rank PCs will also be included in the reconstructed hyperspectral observation so that smaller reconstruction error is not necessarily better. The optimal choice of n_{pc} may differ from application to application since it depends on the noise characteristics of the original hyperspectral observations as well as storage and transmission limitations. For example, Turner et al. (2006) discussed the number of PCs

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

used for noise reduction of the ground-based Atmospheric Emitted Radiance Interferometer (AERI).

2.2 Proposed approach in direct assimilation of PC scores with the EnKF

The EnKF update equation at the analysis step is adopted from the formulations presented in Houtekamer and Zhang (2016, their Eqs. 1 and 2):

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}[\mathbf{y}^o - \mathcal{H}\mathbf{x}^f], \tag{6}$$

$$\mathbf{K} = \mathbf{P}^f\mathcal{H}^{\mathrm{T}}(\mathcal{H}\mathbf{P}^f\mathcal{H}^{\mathrm{T}} + \mathbf{R})^{-1}, \tag{7}$$

where $\mathbf{x}^a$ is the updated estimate of atmosphere state from the prior estimate of the atmosphere state $\mathbf{x}^f$ using the extra information from the new observation $\mathbf{y}^o$, $\mathbf{K}$ is the Kalman gain matrix, $\mathcal{H}$ is the forward operator that performs mapping of the model state to observation state, $\mathbf{R}$ is the observation error covariance, and $\mathbf{P}^f$ is the background error covariance. In EnKF, ensemble-based approximations of $\mathbf{P}^f\mathcal{H}^{\mathrm{T}}$ and $\mathcal{H}\mathbf{P}^f\mathcal{H}^{\mathrm{T}}$ are used (Houtekamer and Zhang 2016, their Eqs. 6 and 7):

$$\mathbf{P}^f\mathcal{H}^{\mathrm{T}} \equiv \frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}\left(\mathbf{x}_i^f - \overline{\mathbf{x}^f}\right)\left(\mathcal{H}\mathbf{x}_i^f - \overline{\mathcal{H}\mathbf{x}^f}\right)^{\mathrm{T}}, \tag{8}$$

$$\mathcal{H}\mathbf{P}^f\mathcal{H}^{\mathrm{T}} \equiv \frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}\left(\mathcal{H}\mathbf{x}_i^f - \overline{\mathcal{H}\mathbf{x}^f}\right)\left(\mathcal{H}\mathbf{x}_i^f - \overline{\mathcal{H}\mathbf{x}^f}\right)^{\mathrm{T}}, \tag{9}$$

where

$\overline{\mathbf{x}^f} = \frac{1}{N_{ens}}\sum_{i=1}^{N_{ens}}\mathbf{x}_i^f$ and $\overline{\mathcal{H}\mathbf{x}^f} = \frac{1}{N_{ens}}\sum_{i=1}^{N_{ens}}\mathcal{H}\mathbf{x}_i^f$.

In the case of hyperspectral observations, $\mathcal{H}\mathbf{x}_i^f$ is the simulated hyperspectral observations from the i^{th} ensemble member, i.e., $\mathbf{y}_i^f$. $\overline{\mathcal{H}\mathbf{x}^f}$ is the mean simulated hyperspectral observation of *the ensemble*. If the PC scores corresponding to $\mathbf{y}_i^f$ calculated using Eq. (3) is $\mathbf{z}_i^f$, with the help of Eq. (4), Eqs. (8) and (9) can be written as

$$\mathbf{P}^f\mathcal{H}^{\mathrm{T}} = \left[\frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}\left(\mathbf{x}_i^f - \overline{\mathbf{x}^f}\right)\left(\mathbf{z}_i^f - \overline{\mathbf{z}^f}\right)^{\mathrm{T}}\right]\mathbf{U}^{\mathrm{T}}, \quad (10)$$

$$\mathcal{H}\mathbf{P}^f\mathcal{H}^{\mathrm{T}} = \mathbf{U}\left[\frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}\left(\mathbf{z}_i^f - \overline{\mathbf{z}^f}\right)\left(\mathbf{z}_i^f - \overline{\mathbf{z}^f}\right)^{\mathrm{T}}\right]\mathbf{U}^{\mathrm{T}}. \quad (11)$$

Comparing the right-hand-size of Eqs. (8) and (10), as well as the Eqs. (9) and (11), if a PCA-based forward model $\mathcal{H}_{pc}$, which is based on the same eigenvector basis $\mathbf{U}$ as in Eq. (3), is used to perform mapping from the model state to PC scores such that $\mathbf{z}_i^f = \mathcal{H}_{pc}\mathbf{x}_i^f$, Eqs. (10) and (11) can be written as

$$\mathbf{P}^f\mathcal{H}^{\mathrm{T}} = \mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}, \quad (12)$$

$$\mathcal{H}\mathbf{P}^f\mathcal{H}^{\mathrm{T}} = \mathbf{U}\mathcal{H}_{pc}\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}. \quad (13)$$

Since the matrix $\mathbf{U}$ is an orthogonal matrix, the equation $\mathbf{U^T U} = \mathbf{U U^T} = \boldsymbol{I}$ holds, where $\boldsymbol{I}$ is identity matrix. The Kalman gain matrix can be written as

$$\mathbf{K} = \mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\left(\mathbf{U}\mathcal{H}_{pc}\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}} + \mathbf{R}\right)^{-1} = \mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\left[\mathbf{U}\left(\mathcal{H}_{pc}\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}} + \mathbf{U}^{\mathrm{T}}\mathbf{R}\mathbf{U}\right)\mathbf{U}^{\mathrm{T}}\right]^{-1}$$

$$= \mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\left(\mathcal{H}_{pc}\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}} + \mathbf{U}^{\mathrm{T}}\mathbf{R}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}, \quad (14)$$

and Eq. (6) can be written as

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}\mathbf{U}[\mathbf{z}^o - \mathbf{z}^f]. \quad (15)$$

Substituting Eq. (14) into Eq. (15) yields

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}_{pc}[\mathbf{z}^o - \mathbf{z}^f], \quad (16)$$

$$\mathbf{K}_{pc} = \mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}\left(\mathcal{H}_{pc}\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}} + \mathbf{U}^{\mathrm{T}}\mathbf{R}\mathbf{U}\right)^{-1}. \quad (17)$$

Comparing the right-hand-side of Eq. (10) and Eq. (12), as well as Eq. (11) and Eq. (13), and using $\mathbf{z}_i^f = \mathcal{H}_{pc}\mathbf{x}_i^f$, $\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}$ and $\mathcal{H}_{pc}\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}}$ can also be calculated from the ensemble

$$\mathbf{P}^f\mathcal{H}_{pc}^{\mathrm{T}} = \left[\frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}\left(\mathbf{x}_i^f - \overline{\mathbf{x}^f}\right)\left(\mathcal{H}_{pc}\mathbf{x}_i^f - \overline{\mathcal{H}_{pc}\mathbf{x}^f}\right)^{\mathrm{T}}\right], \quad (18)$$

$$\mathcal{H}_{pc}\mathbf{P}^{f}\mathcal{H}_{pc}^{\mathrm{T}} = \left[\frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}\left(\mathcal{H}_{pc}\mathbf{x}_i^{f} - \overline{\mathcal{H}_{pc}\mathbf{x}^{f}}\right)\left(\mathcal{H}_{pc}\mathbf{x}_i^{f} - \overline{\mathcal{H}_{pc}\mathbf{x}^{f}}\right)^{\mathrm{T}}\right]. \quad (19)$$

Equation (16-19) are very similar to the original EnKF update Eqs. (6-9), with the observation variable changed from hyperspectral observations to PC scores, the forward operator changed to a PCA-based radiative transfer model that directly calculate PC scores, and the observation error covariance changed to $\mathbf{R}_{pc} = \mathbf{U}^{\mathrm{T}}\mathbf{R}\mathbf{U}$. This indicates assimilating all the PC scores using Eqs. (16-19) and assimilating hyperspectral observations using Eqs. (6-9) should give the same analysis increment $\mathbf{x}^{a} - \mathbf{x}^{f}$.

Goldberg et al. (2003) chose to save and distribute 200 PC scores to assure that there was sufficient information to reconstruct observed AIRS observations with over 2000 channels. As will be shown later in this paper, the variance of hyperspectral observations explained by the leading PCs are orders of magnitude larger than the lower-rank PCs. This means the leading PC scores (leading dimensions of $\mathbf{z}$) are expected to be much larger in value than the lower-rank PC scores (lower-rank dimensions of $\mathbf{z}$). The matrix elements corresponding to the lower-rank PC scores of the variance term in the Kalman gain matrix, i.e., $\mathcal{H}_{pc}\mathbf{P}^{f}\mathcal{H}_{pc}^{\mathrm{T}} + \mathbf{U}^{\mathrm{T}}\mathbf{R}\mathbf{U}$, are expected to be dominated by instrument noise. As a result, assimilating the lower-rank PC scores have much less contribution to the EnKF increment compared to the leading PC scores. Thus, assimilating only the leading PC scores can provide similar result to that when all the channels of the hyperspectral observations are assimilated.

The most obvious benefit from this PCA-based EnKF method that assimilates truncated PC scores is the huge savings in computation time. The speedup includes both time savings in the forward modeling step and the EnKF analysis step. AIRS have over 2000 channels.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

If assimilating less than 20 leading PCs can provide similar result compared to assimilating over 2000 channels, tens of times speedup can be expected.

3. Data compression and dimension reduction of hyperspectral satellite measurements with PCA

The focus of this work is to demonstrate that directly assimilating a number of leading PC scores yields similar increments to the model states compared to assimilating a much larger number of hyperspectral channels. In other words, it is to show that the truncated version of Eqs. 16 and 17 yields similar increments to Eqs. 6 and 7. As such, full data assimilation cycle was not performed. Instead, EnKF priors from previous data assimilation experiments for Hurricane Harvey (2017) described in Chapter 6 of Minamide (2018) was used, which is described in Section 3.1.

When directly assimilating PC scores in an EnKF system, there are generally two approaches to map atmosphere states into PC scores. One way is to simulate the hyperspectral observations using traditional channel-based radiative transfer models, then project the simulated hyperspectral observations to pre-trained PC components. By following this approach, the data assimilation system benefits from the computational savings in the analysis step due to reduction in dimensionality and from noise reduction by discarding the noise-affected lower-rank PCs. The drawback of this approach is that the computational cost can be prohibitive if large number of channels are need, especially for operational systems. Another way to obtain PC scores is to directly map from atmospheric states to PC scores by using a PCA-based radiative transfer model as forward operator $\mathcal{H}_{pc}$ such as PCRTM and PC_RTTOV. Additional computational savings in radiative transfer

calculations may be achieved with the latter approach. Since simulating hyperspectral observations was necessary in this research (i.e., used in Eqs. 6 and 7), we chose to follow the first approach. Another benefit of this approach is that the result can be independent of the choice of PCA-based radiative transfer model and its underlying assumptions and constraints.

AIRS brightness temperatures (BTs) were simulated using the Community Radiative Transfer Model (CRTM; Han et al. 2006) using modeled atmospheric and surface states as input. Principal components should be trained using a dataset that is independent of the EnKF assimilation experiment. In this work, simulated AIRS BTs from the simulation of Dynamics of the MJO (DYNAMO) field campaign described in Ying and Zhang (2017) is used to train the PCA, which is also described in details in Section 3.1; the selection of AIRS channels used in this study is described in Section 3.2; and the training of PCA and calculation of PC scores is described in Section 3.3.

3.1. *Generating hyperspectral observations*

EnKF inputs from previous data assimilation experiments for Hurricane Harvey (2017) described in Chapter 6 of Minamide (2018) was used to evaluate the PCA-based EnKF data assimilation method. The ensemble data assimilation experiment had 60 ensemble members, which were initialized at 0000 UTC 22 Aug 2017 with perturbations added using WRFDA CV3 option and integrated for 12 hours for spin-up. Minimum sea level pressure and all-sky GOES-16 channel 8 observations were assimilated every hour after 1200 UTC 22 Aug 2017 for three days till 1200 UTC 25 Aug 2017. The intermediate resolution (9 km) domain was used because it directly predicted hydrometer mixing ratios and had a

relatively large geographical coverage. The WRF single-moment 6-class mixed-phase microphysics scheme was used (WSM6; Hong and Lim 2006). EnKF priors at two different analysis times were used: one at 1400 UTC 22 August 2017 and the other one at 0200 UTC 25 August 2017 representing the initial stage and the well-developed stage of Harvey (2017) respectively. The WRF model used had 42 vertical levels with model top at 10 hPa. More details about model configuration was described in Chapter 6 of Minamide (2018).

Principal components were trained using simulated hyperspectral observations from the DYNAMO period, which represents general atmospheric conditions over tropical oceans with different types of cloudy and clear conditions, different moisture profiles, and a wide range of surface temperatures. The WRF simulation for DYNAMO had a large domain covering 20°N to 20°S and 50°E to 120°E with 9 km horizontal resolution and 44 vertical levels with model top at 20 hPa. Hydrometer properties were directly predicted by the WRF double-moment (WDM) scheme (Lim and Hong 2010). More detailed information in model configuration was described by Ying and Zhang (2017). Hyperspectral BTs simulated using CRTM from the WRF model output at 0000 UTC 21 October 2011 were calculated. Then Gaussian random noise was added to each channel with standard deviation equal to instrument noise level of that channel. This noise-added hyperspectral BTs were used to train PCA.

When using modeled atmospheric states as inputs to CRTM, typical effective radius values of hydrometers were used: 16.8 μm for cloud water drops, 25 μm for cloud ice, 1000 μm for rain, 500 μm for snow, and 500 μm for graupel. Only ocean grid points were used in this study to avoid additional uncertainties in land surface temperature and emissivity (that will be explored in future studies). The zenith angles for AIRS cross-track footprints

range from near 0° (footprint close to nadir) to about 57° (outermost footprint) in each scan line. This angle is important for AIRS TB calculations since it influences ocean surface emissivity and effective path length of the atmosphere. To represent this variability, the zenith angles were randomly drawn from a uniform distribution from 0° to 57°. For the 60 ensemble members for Hurricane Harvey (2017), the same zenith angles were used at the same location for all ensemble members, while different zenith angles were used at different locations.

Figure 1 shows simulated BTs of AIRS channel with wavelength 12.183 μm within the atmosphere window. Panel (a) shows simulated BTs of the DYNAMO case; panel (b) shows the 1st ensemble member of the early stage Harvey (at 1400 UTC 22 August 2017); and panel (c) shows the 1st ensemble member of the developed stage Harvey (at 0200 UTC 25 August 2017). It is clearly shown that the largest variability in window channel BTs is due to variations in clouds. Warmer BTs usually correspond to clear sky while colder BTs usually correspond to clouds of different levels. The three circles in panel (b) labeled with B1 through B3 are three locations with cold, medium, and warm BTs for the 1st ensemble member of the early stage Harvey case respectively, indicating high-cloud, mid-cloud, and clear-sky conditions. Note that other ensemble members at these locations could have different sky conditions than the 1st ensemble member. Figure 2 shows histograms of simulated BTs of AIRS channel with wavelength 12.183 μm for all the 60 ensemble members at the three locations B1 through B3. At location B1, most ensemble members had cold BTs smaller than 240 K, indicating most members had high cloud. At location B2, most ensemble members had warm BTs (larger than 270 K) while some members had

339 medium BTs (between 240K and 270K). Location B3 is similar to B2, but with more warm
340 BTs and less medium BTs.
341
342 3.2. *Channel selection*
343 Although using all the channels can exploit the full information content in hyper-
344 spectral observations, not all channels were used in this demonstrative work. Firstly,
345 channels with noise equivalent temperature difference (NEDT) larger than 1K were
346 excluded. Secondly, the model tops of the WRF simulations used in this study only
347 extended to 10 hPa or 20 hPa. Simulated BTs for channels sensitive to atmosphere layers
348 close to or above the model tops are not accurate and thus should be excluded, which
349 includes some temperature and water vapor sounding channels. As a result, only 1670
350 channels with wavelengths between 4.44 μm to 14.503 μm, with similar wavelength range
351 to Lin et al. (2017), were used in the current study.
352 Another choice to make is that whether all channels within this wavelength range
353 should be used in data assimilation. It is known that the number of pieces of independent
354 information in hyperspectral observations is much smaller than the number of channels.
355 Figure 3 shows the inter-channel correlations between different channels. Channels with
356 wavelength 8~13 μm (roughly corresponding to channel number 400~1330) directly sense
357 the earth surface or cloud top and are highly correlated with each other. They have largest
358 variations in BTs among various sky conditions because of large differences between cloud
359 top temperatures and earth surface temperatures. The water vapor channels with
360 wavelengths about 6~8 μm (channel number 1300~1864) and the temperature sounding
361 channels with wavelengths about 13~14.5 μm (channel number 137~400) close to CO_2

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

absorption bands have smaller BTs than the window channels but are also highly correlated with the window channels. One reason is that when the cloud top is higher than the peak of the weighting functions of these sounding channels, radiances emitted by the cloud also have large contribution to the measured radiances by these sounding channels. Another reason is that correlations exist between atmospheric state variables, such as surface temperature and atmospheric temperature. As such, correlations between BTs of window channels and sounding channels also exist over clear-sky condition.

In this research, classical fast forward radiative transfer model is used. Simulating BTs for all AIRS channels and then calculating PC scores require significant computational resources. Since BTs of AIRS channels were highly correlated, the 324-channel subset suggested by Goldberg et al. (2003) and Susskind et al. (2003) may be used to balance the information content and computational requirement, which has 199 of the selected channels within the frequency range specified in this work. We performed PCA on both the full-spectral-resolution BTs (1670 channels) and the subset BTs (199 channels) over the DYNAMO domain. Figure 4(a) shows the variance explained by each PC for both the full-spectral-resolution BTs (red line) and the subset BTs (blue line). For both full-spectral-resolution and selected-channel BTs, variance explained by each PC decreases rapidly, especially for the first few PCs. If 1 K^2 was used as variance threshold, corresponding to 1 K white noise in the observation, the full-spectral-resolution BTs had 19 principal components with variance larger than the threshold, while the subset BTs had 14 PCs.

Figure 4(b) shows that the explained variance ratio of both the full-spectral-resolution BTs (red) and the subset BTs (blue) corresponding to each PC. The full-spectral-resolution BTs had less PCs than the subset BTs with explained variance ratio above a certain level,

indicating variance in the full-spectral-resolution BTs were more strongly concentrated to the first few leading PCs, possibly because that many highly correlated window channels are excluded in the 324-channel subset. Figure 4(c) shows the first PCs of the full-spectral resolution BTs (red) and subset BTs (blue), while Figure 4(d) shows the second PCs. Note that the amplitude difference was because the PCs were scaled to have unit vector length. Although the number of channels were different, the shapes of the corresponding PCs were similar for the full-spectral-resolution and subset BTs. As such, the subset BTs is used in this study instead of full-spectral-resolution BTs to save computational time in the forward radiative transfer calculations while retain most information in the hyperspectral observations.

3.3. *The training of PCA and calculation of PC scores*

When selecting AIRS channels used in this study, noisy channels with NEDT values larger than 1 K were already excluded. Most of the 199 selected channels have similar noise level lower than 0.5 K. As such, unlike other research that performs PCA on noise normalized radiances to avoid errors from noisy channels, we performed PCA on BTs directly. The simulated hyperspectral BTs of the DYNAMO case were used as training dataset to train the PCA. The domain of DYNAMO case had 444 grids along latitudinal direction and 777 grids along longitudinal direction. Hyperspectral BTs from every fourth ocean grid point were used to train the PCA, resulting in a training dataset with 74,612 hyperspectral samples, each with BTs of the 199 selected channels. The choice of the training dataset in the current demonstrative work was not intended to be optimal. With all hyperspectral BTs simulated from a single time at 0000 UTC 21 October 2011, they could

not represent seasonal or diurnal variations of the atmospheric states. However, this deficiency could be compensated to some extent by the large geographical coverage of the domain.

PC scores of the hyperspectral BTs simulated from the 60-member ensemble WRF run for Hurricane Harvey (2017) were calculated using the mean hyperspectral BTs and PCs trained from the training dataset following Eq. 3, where **y** was the hyperspectral BTs simulated from the two analysis times of the Harvey case mentioned in Section 3.1, and $\overline{\mathbf{y}^{tr}}$ was the mean hyperspectral BTs of *the training dataset* from DYNAMO. Then these differences in hyperspectral BTs were projected onto the PCs trained from *the training dataset* to calculate PC scores.

Using more PCs leads to higher reconstruction accuracy but higher computational cost. It is beyond the scope of this work to find the optimal number of PCs to use. Rather, we explore the influence of PC truncation on reconstruction error and EnKF increments. Figure 5 shows the reconstruction error when the first 10 (panels a-c), 15 (panels d-f), or 20 (panels g-i) leading PCs were used to reconstruct the hyperspectral BTs simulated from the 1st member of Harvey (2017) at 0200 UTC 25 August 2017, i.e., the well-developed stage. The three panels (a, d, g) in the first column show the residual error of a water vapor sounding channel with wavelength 7.493 μm, panels (b, e, h) in the middle column show the residual error of a window channel with wavelength 10.213 μm, and panels (c, f, i) in the right column show the residual error of a temperature sounding channel with wavelength 13.796 μm. Reconstruction error values for 1400 UTC 22 August 2017 was similar and not shown.

Root mean square of reconstruction error over the entire domain of the DYNAMO case and the 1st member of Harvey (2017) case at the two aforementioned analysis times for all the selected channels are shown in Figure 6, with the three panels corresponding to using 10, 15, and 20 PCs respectively. When more PCs were used, the reconstruction errors were smaller. The error level difference between 15 PCs and 20 PCs was small, indicating information gained from using even more PCs can be limited.

It is shown in Figure 1(c) that the range of variation of the window channel was about 80 K, with BTs about 220 K over hurricane eyewalls and about 300 K over clear-sky ocean. This large variation was successfully captured by the 20 leading PCs with residuals similar to or below instrument noise levels. Residuals in water vapor sounding channels and temperature sounding channels showed similar characteristics, indicating the majority of information content in the hyperspectral BTs could be captured using about 20 leading PC components.

4. Data assimilation experiments with the EnKF update using PC scores as the innovation vector

The PCA-based EnKF assimilation method was evaluated by comparing the EnKF increment $\mathbf{x}^a - \mathbf{x}^f$ obtained from assimilating PC scores and from directly assimilating simulated AIRS hyperspectral BTs. The 40th ensemble member of the two analysis times of the Harvey case was used as 'truth' because its temperature profiles and water vapor profiles had relatively large departure from the ensemble mean at the three locations B1 through B3 in Figure 1(b). Gaussian noises with standard deviations equal to instrument noise levels for all channels were added to the simulated hyperspectral BTs from the 40th

ensemble member to generate synthetic hyperspectral observations. The other 59 members were used as ensemble members of the EnKF system.

Figure 7(a-c) shows the simulated hyperspectral BTs from all the ensemble members at the locations labeled as B1, B2, and B3 in Figure 1(b) respectively. Each colored thin line shows simulated hyperspectral BTs from one ensemble member. The color of each line was determined by the brightness temperature at the same window channel used in Figure 1 from each ensemble member at the specific location. As a result, lines with the same color in Figure 7 (a-c) do not necessarily correspond to the same ensemble member. Although the three locations were selected as example of high-cloud, mid-cloud, and low-cloud in the 1st member respectively, different ensemble members could have different scene types, as is shown in Figure 2. Nevertheless, most ensemble members at B1 had relatively higher clouds, while more members at B2 and B3 had clear sky conditions. The black lines show the mean hyperspectral BTs of the *training dataset*. It should not be confused with the BTs calculated using ensemble mean atmospheric states ($\mathcal{H}\mathbf{x}^f$; black dashed lines, used in EnKF), or the mean hyperspectral BTs of the *ensemble* (not shown), since it is the mean hyperspectral BTs of the *training dataset* that should be subtracted from the hyperspectral BTs from the ensembles when calculating their PC scores.

Figure 7(d-f) show the corresponding 20 leading PC scores of the ensemble members at the three locations respectively. Lines with the same colors in Figure 7(a & d), Figure 7(b & e), and Figure 7(c & f) correspond to the same ensemble members respectively. The black dashed line in Figure 7 (d-f) show the 20 leading PC scores of the ensemble mean. The absolute values of PC scores decreased rapidly with increasing PC index, which agreed well with the rapid decrease in explained variance shown in Figure 4(a).

When AIRS hyperspectral BTs were assimilated, Eqs. (6)-(9) were used to calculate EnKF increments. Although instrument noise levels for most AIRS channels are smaller than 0.5K, observation error for all the channels is assumed as 1K in this experiment to account for other possible error sources including representative error and forward radiative transfer model error. As such, the observation error covariance **R** was a diagonal matrix with all its diagonal elements equaling to 1 K^2.

When PC scores were assimilated, EnKF increments were calculated using a truncated version of Eqs. (16-19) that only a number of leading PCs and PC scores were used. The observation error covariance changed to $\mathbf{R}_{pc} = \mathbf{U}_{npc}^{\mathrm{T}}\mathbf{R}\mathbf{U}_{npc}$, where $\mathbf{U}_{npc}$ was a truncated version of matrix **U** as in Eq. (5). Since $\mathbf{U}_{npc}$ was a $n_{ch} \times n_{pc}$ matrix, the size of matrix $\mathbf{R}_{pc}$ was $n_{pc} \times n_{pc}$. When **R** was a diagonal matrix with all its diagonal elements equaling to 1 K^2 as assumed above, $\mathbf{R}_{pc}$ was also a diagonal matrix with all its diagonal elements equaling to 1 K^2 but with a much smaller size $n_{pc} \times n_{pc}$.

Figure 8 shows the EnKF increments of temperature profiles (panels a-c) and of water vapor profiles (panels d-f) when AIRS hyperspectral BTs were assimilated (black lines) and when PC scores were assimilated (colored lines) at 1400 UTC 22 August 2017. Different choices of n_{pc} values from 2 PCs to 30 PCs were tested. With increasing number of PC scores assimilated, the increments from assimilating PC scores became closer to that from assimilating AIRS hyperspectral BTs. When more than 15 PC scores were assimilated, the increments from assimilating PC scores and from assimilating AIRS hyperspectral BTs were almost indistinguishable for both the temperature profiles and water vapor profiles.

Figure 9 shows the root-mean-square (RMS) of the difference between increments from assimilating AIRS hyperspectral BTs and PC scores calculated over all layers at the three

499 locations when different number of PCs were assimilated at 1400 UTC 22 August 2017 at
500 the three locations B1 through B3. Panel (a) shows the RMS of temperature increments
501 and panel (b) shows the RMS of water vapor increments. For all of the three locations,
502 RMS decreased until about 16 leading PCs were used. This number was close to that used
503 by Matricardi and Mcnally (2014) where 20 leading PCs were used for sounding channels
504 under clear-sky condition. This result indicated that most information contents in the
505 hyperspectral BTs could be captured by a smaller number of leading PCs. Also,
506 assimilating the leading PC scores instead of assimilating AIRS hyperspectral BTs could
507 provide significant computational savings with satisfactory accuracy.
508
509 **5. Summary and discussion**
510 Satellite-based hyperspectral observations such as those from AIRS and IASI have
511 thousands of infrared channels that contain information on atmospheric state with much
512 higher vertical resolution compared to observations from traditional sensors. However, the
513 large numbers of channels also lead to computational burden in retrieval and data
514 assimilation. Furthermore, most of the channels are highly correlated and the number of
515 pieces of independent information contained in the hyperspectral observations are usually
516 much smaller than the number of channels. Principal component analysis (PCA) was used
517 in this research to compress the observational information content contained in these
518 hyperspectral channels to a few leading principal components (PC). The corresponding PC
519 scores can then be assimilated into a PCA-based ensemble Kalman filter (EnKF) system.
520 In this proof-of-concept study using simulated observations, PCA was trained from
521 AIRS hyperspectral BTs simulated using the Community Radiative Transfer Model

(CRTM) and a large-domain convection-permitting simulation over the Indian Ocean that represents generic tropical ocean conditions. Brightness temperatures of 1670 AIRS channels were simulated over the domain, which showed large inter-channel correlations. Since PCA-based fast radiative transfer model is not used in this study, a subset of 199 channels were selected for subsequent analysis, which contains most of the variabilities in all AIRS channels, to balance information content and computational requirement. Then principal components were trained using 74,612 hyperspectral BTs samples.

AIRS hyperspectral BTs were simulated from the convection permitting ensemble simulations of Hurricane Harvey (2017) with CRTM. These hyperspectral BTs were converted to PC scores using the mean hyperspectral BTs of the training dataset and the PCs. The EnKF increments from assimilating AIRS hyperspectral BTs and from assimilating different numbers of leading PCs were compared. Result showed that assimilating about 10 to 20 leading PCs could yield increments that were nearly indistinguishable to that from assimilating hyperspectral measurements with 199 channels.

In this proof-of-concept study, we chose not to use PCA-based radiative transfer models so that the results shown in this work can be independent of their underlying assumptions and constraints. The drawback of this approach is that we had to use a subset of all AIRS channels which contains most of the information content. In a real PCA-based EnKF system, PCA-based radiative transfer model can be used to directly simulate PC scores at the full-spectral-resolution.

The current proof-of-concept study is based on simulated observations for both the training dataset and the reference truth for a tropical cyclone event; research is ongoing and/or planned to further explore the use of this approach for cycled real-data assimilation

545 under different atmospheric conditions including those over various land surfaces and/or
546 over higher latitudes.
547
548 **Acknowledgements**: This research is partially supported by NASA Grants
549 NNX16AD84G and NNX12AJ79G, ONR Grant N000140910526 and NOAA funding
550 under HFIP and NGGPS. The authors would like to thank Masashi Minamide, Scott Sieron
551 and Yue Ying for providing the WRF model output. Discussions with Eugene Clothiaux,
552 Scott Siron and Xianglei Huang are beneficial for this study. Computing are performed at
553 the Texas Advanced Computing Center.
554

6. **References**

Aires, F., W. B. Rossow, N. A. Scott, and A. Chédin, 2002: Remote sensing from the infrared atmospheric sounding interferometer instrument 1. Compression, denoising, and first-guess retrieval algorithms. *J. Geophys. Res.*, **107**, 4619, doi:10.1029/2001JD000955.

Antonelli, P., and Coauthors, 2004: A principal component noise filter for high spectral resolution infrared measurements. *J. Geophys. Res. Atmos.*, **109**, D23102, doi:10.1029/2004JD004862.

Aumann, H. H., and Coauthors, 2003: AIRS/AMSU/HSB on the aqua mission: design, science objectives, data products, and processing systems. *IEEE Trans. Geosci. Remote Sens.*, **41**, 253–264, doi:10.1109/TGRS.2002.808356.

Collard, A. D., and A. P. McNally, 2009: The assimilation of Infrared Atmospheric Sounding Interferometer radiances at ECMWF. *Q. J. R. Meteorol. Soc.*, **135**, 1044–1058, doi:10.1002/qj.410.

——, ——, F. I. Hilton, S. B. Healy, and N. C. Atkinson, 2010: The use of principal component analysis for the assimilation of high-resolution infrared sounder observations for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **136**, 2038–2050, doi:10.1002/qj.701.

Geer, A. J., and Coauthors, 2018: All-sky satellite data assimilation at operational weather forecasting centres. *Q. J. R. Meteorol. Soc.*, **144**, 1191–1217, doi:10.1002/qj.3202.

Goldberg, M. D., Yanni Qu, L. M. McMillin, W. Wolf, Lihang Zhou, and M. Divakarla, 2003: AIRS near-real-time products and algorithms in support of operational numerical weather prediction. *IEEE Trans. Geosci. Remote Sens.*, **41**, 379–389,

578 doi:10.1109/TGRS.2002.808307.
579 Guidard, V., N. Fourrié, P. Brousseau, and F. Rabier, 2011: Impact of IASI assimilation at
580 global and convective scales and challenges for the assimilation of cloudy scenes. *Q.*
581 *J. R. Meteorol. Soc.*, **137**, 1975–1987, doi:10.1002/qj.928.
582 Han, Y., P. van Delst, Q. Liu, F. Weng, B. Yan, R. Treadon, and J. Derber, 2006: JCSDA
583 Community Radiative Transfer Model (CRTM) - Version 1. *NOAA Tech. Rep.*,
584 **NESDIS 122**, 40.
585 Hannachi, A., I. T. Jolliffe, and D. B. Stephenson, 2007: Empirical orthogonal functions
586 and related techniques in atmospheric science: A review. *Int. J. Climatol.*, **27**, 1119–
587 1152, doi:10.1002/joc.1499.
588 Hilton, F., and Coauthors, 2012: Hyperspectral Earth Observation from IASI: Five Years
589 of Accomplishments. *Bull. Am. Meteorol. Soc.*, **93**, 347–370, doi:10.1175/BAMS-D-
590 11-00027.1.
591 Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF Single-Moment 6-Class Microphysics
592 Scheme (WSM6). *J. Korean Meteorol. Soc.*, **42**, 129–151.
593 Houtekamer, P. L., and F. Zhang, 2016: Review of the Ensemble Kalman Filter for
594 Atmospheric Data Assimilation. *Mon. Weather Rev.*, **144**, 4489–4532,
595 doi:10.1175/MWR-D-15-0440.1.
596 Huang, H.-L., and P. Antonelli, 2001: Application of Principal Component Analysis to
597 High-Resolution Infrared Measurement Compression and Retrieval. *J. Appl.*
598 *Meteorol.*, **40**, 365–388, doi:10.1175/1520-
599 0450(2001)040<0365:AOPCAT>2.0.CO;2.
600 ——, W. L. Smith, and H. M. Woolf, 1992: Vertical Resolution and Accuracy of

601 Atmospheric Infrared Sounding Spectrometers. *J. Appl. Meteorol.*, **31**, 265–274,
602 doi:10.1175/1520-0450(1992)031<0265:VRAAOA>2.0.CO;2.
603 Lim, K.-S. S., and S.-Y. Hong, 2010: Development of an Effective Double-Moment Cloud
604 Microphysics Scheme with Prognostic Cloud Condensation Nuclei (CCN) for
605 Weather and Climate Models. *Mon. Weather Rev.*, **138**, 1587–1612,
606 doi:10.1175/2009MWR2968.1.
607 Lin, H., S. S. Weygandt, A. H. N. Lim, M. Hu, J. M. Brown, and S. G. Benjamin, 2017:
608 Radiance Preprocessing for Assimilation in the Hourly Updating Rapid Refresh
609 Mesoscale Model: A Study Using AIRS Data. *Weather Forecast.*, **32**, 1781–1800,
610 doi:10.1175/WAF-D-17-0028.1.
611 Liu, X., W. L. Smith, D. K. Zhou, and A. Larar, 2006: Principal component-based radiative
612 transfer model for hyperspectral sensors: theoretical concept. *Appl. Opt.*, **45**, 201,
613 doi:10.1364/AO.45.000201.
614 Matricardi, M., 2010: A principal component based version of the RTTOV fast radiative
615 transfer model. *Q. J. R. Meteorol. Soc.*, **136**, 1823–1835, doi:10.1002/qj.680.
616 Matricardi, M., and A. P. McNally, 2014: The direct assimilation of principal components
617 of IASI spectra in the ECMWF 4D-Var. *Q. J. R. Meteorol. Soc.*, **140**, 573–582,
618 doi:10.1002/qj.2156.
619 McNally, A. P., P. D. Watts, J. A. Smith, R. Engelen, G. A. Kelly, J. N. Thépaut, and M.
620 Matricardi, 2006: The assimilation of AIRS radiance data at ECMWF. *Q. J. R.*
621 *Meteorol. Soc.*, **132**, 935–957, doi:10.1256/qj.04.171.
622 Minamide, M., 2018: On the predictability of tropical cyclones through all-sky infrared
623 satellite radiance assimilation. The Pennsylvania State University, 201 pp.

624 ——, and F. Zhang, 2018: Assimilation of All-Sky Infrared Radiances from Himawari-8
625 and Impacts of Moisture and Hydrometer Initialization on Convection-Permitting
626 Tropical Cyclone Prediction. *Mon. Weather Rev.*, **146**, 3241–3258,
627 doi:10.1175/MWR-D-17-0367.1.
628 ——, and ——, 2019: An adaptive background error inflation method for assimilating all-
629 sky radiances. *Q. J. R. Meteorol. Soc.*, **145**, 805–823, doi:10.1002/qj.3466.
630 Monahan, A. H., J. C. Fyfe, M. H. P. Ambaum, D. B. Stephenson, and G. R. North, 2009:
631 Empirical orthogonal functions: The medium is the message. *J. Clim.*, **22**, 6501–6514,
632 doi:10.1175/2009JCLI3062.1.
633 North, G. R., 1984: Empirical Orthogonal Functions and Normal Modes. *J. Atmos. Sci.*, **41**,
634 879–887, doi:10.1175/1520-0469(1984)041<0879:EOFANM>2.0.CO;2.
635 Schmit, T. J., M. M. Gunshor, W. P. Menzel, J. J. Gurka, J. Li, and A. S. Bachmeier, 2005:
636 Introducing the next-generation advanced baseline imager on GOES-R. *Bull. Am.*
637 *Meteorol. Soc.*, **86**, 1079–1096, doi:10.1175/BAMS-86-8-1079.
638 ——, P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017:
639 A closer look at the ABI on the goes-r series. *Bull. Am. Meteorol. Soc.*, **98**, 681–698,
640 doi:10.1175/BAMS-D-15-00230.1.
641 Susskind, J., C. D. Barnet, and J. M. Blaisdell, 2003: Retrieval of Atmospheric and Surface
642 Parameters From AIRS / AMSU / HSB Data in the Presence of Clouds. *IEEE Trans.*
643 *Geosci. Remote Sens.*, **41**, 390–409, doi:10.1109/TGRS.2002.808236.
644 Turner, D. D., R. O. Knuteson, H. E. Revercomb, C. Lo, and R. G. Dedecker, 2006: Noise
645 reduction of atmospheric emitted radiance interferometer (AERI) observations using
646 principal component analysis. *J. Atmos. Ocean. Technol.*, **23**, 1223–1238,

647 doi:10.1175/JTECH1906.1.

648 Wu, W., X. Liu, D. K. Zhou, A. M. Larar, Q. Yang, S. H. Kizer, and Q. Liu, 2017: The
649 application of PCRTM physical retrieval methodology for iasi cloudy scene analysis.
650 *IEEE Trans. Geosci. Remote Sens.*, **55**, 5042–5056,
651 doi:10.1109/TGRS.2017.2702006.

652 Xu, D., Z. Liu, X. Y. Huang, J. Min, and H. Wang, 2013: Impact of assimilating IASI
653 radiance observations on forecasts of two tropical cyclones. *Meteorol. Atmos. Phys.*,
654 **122**, 1–18, doi:10.1007/s00703-013-0276-2.

655 Ying, Y., and F. Zhang, 2017: Practical and Intrinsic Predictability of Multiscale Weather
656 and Convectively Coupled Equatorial Waves during the Active Phase of an MJO. *J.*
657 *Atmos. Sci.*, **74**, 3771–3785, doi:10.1175/JAS-D-17-0157.1.

658 Zhang, F., M. Minamide, and E. E. Clothiaux, 2016: Potential impacts of assimilating all-
659 sky infrared satellite radiances from GOES-R on convection-permitting analysis and
660 prediction of tropical cyclones. *Geophys. Res. Lett.*, **43**, 2954–2963,
661 doi:10.1002/2016GL068468.

662

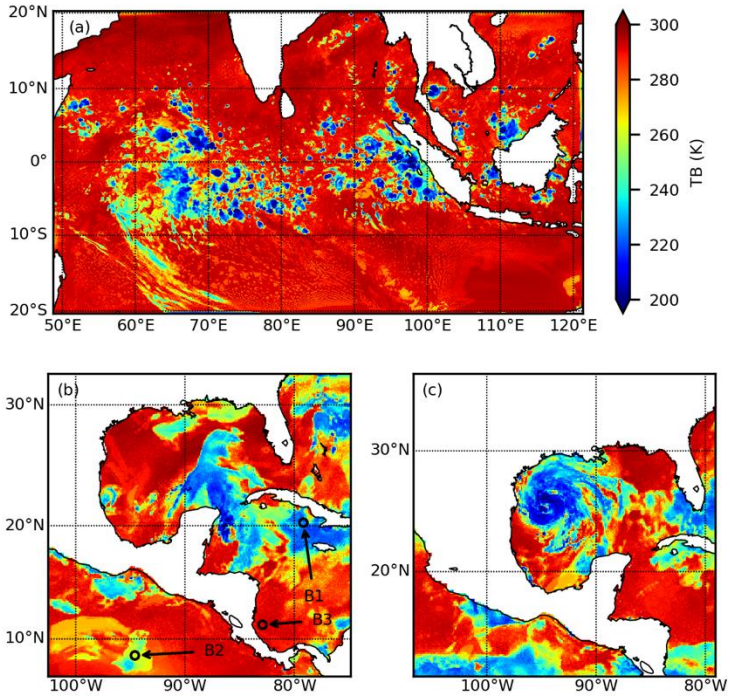
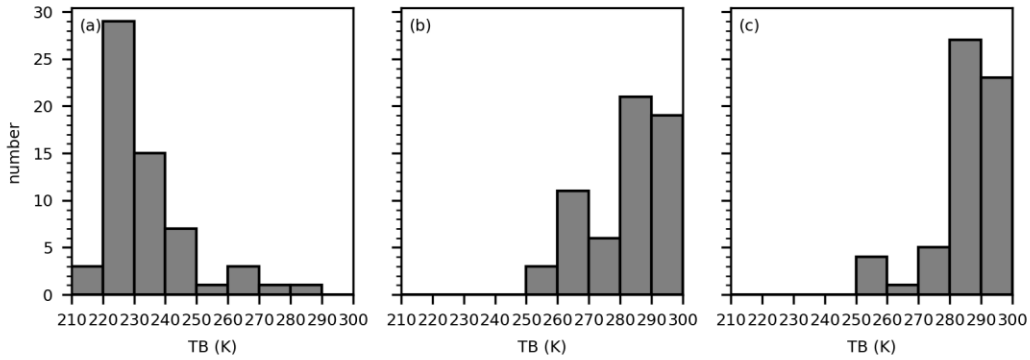


Figure 1: Simulated brightness temperatures for AIRS channel with wavelength 12.183 μm for (a) DYNAMO case at 0000 UTC 21 Oct. 2011, (b) early stage Harvey case at 1400 UTC 22 Aug. 2017, and (c) developed stage Harvey case at 0200 UTC 25 Aug. 2017. Small circles labeled with B1, B2, and B3 in panel (b) indicate high-cloud, mid-cloud, and low/no cloud location respectively.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-18-0454.1.

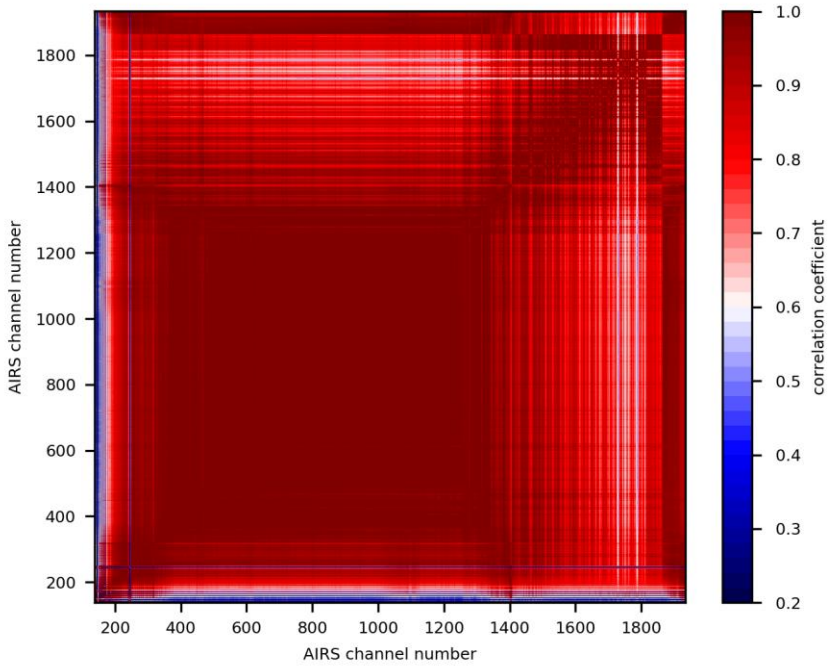


669

670 *Figure 2: Histograms of brightness temperatures of AIRS channel with wavelength 12.183*

671 *μm for all 60 ensemble members at locations labeled with (a) B1, (b) B2, and (c) B3 in*

672 *Figure 1(b).*

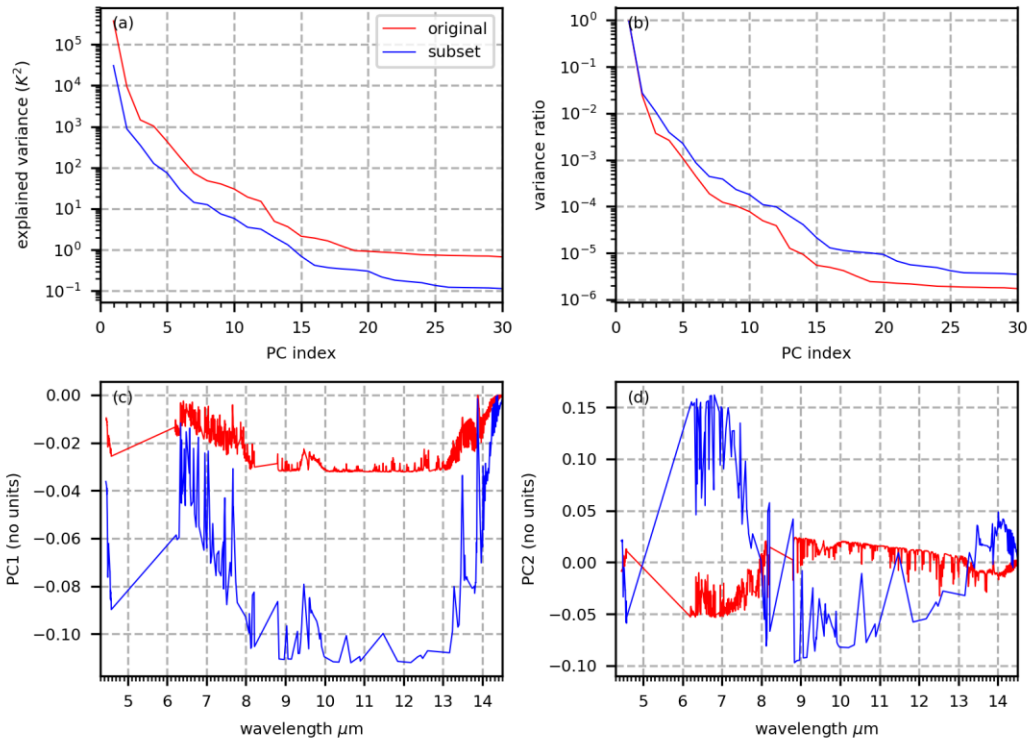


673

674 *Figure 3: Inter-channel correlations for simulated AIRS channels of DYNAMO case. The*

675 *minimum value of the color bar is set to 0.2. The actual minimum correlation value between*

676 *channels is -0.36.*



677

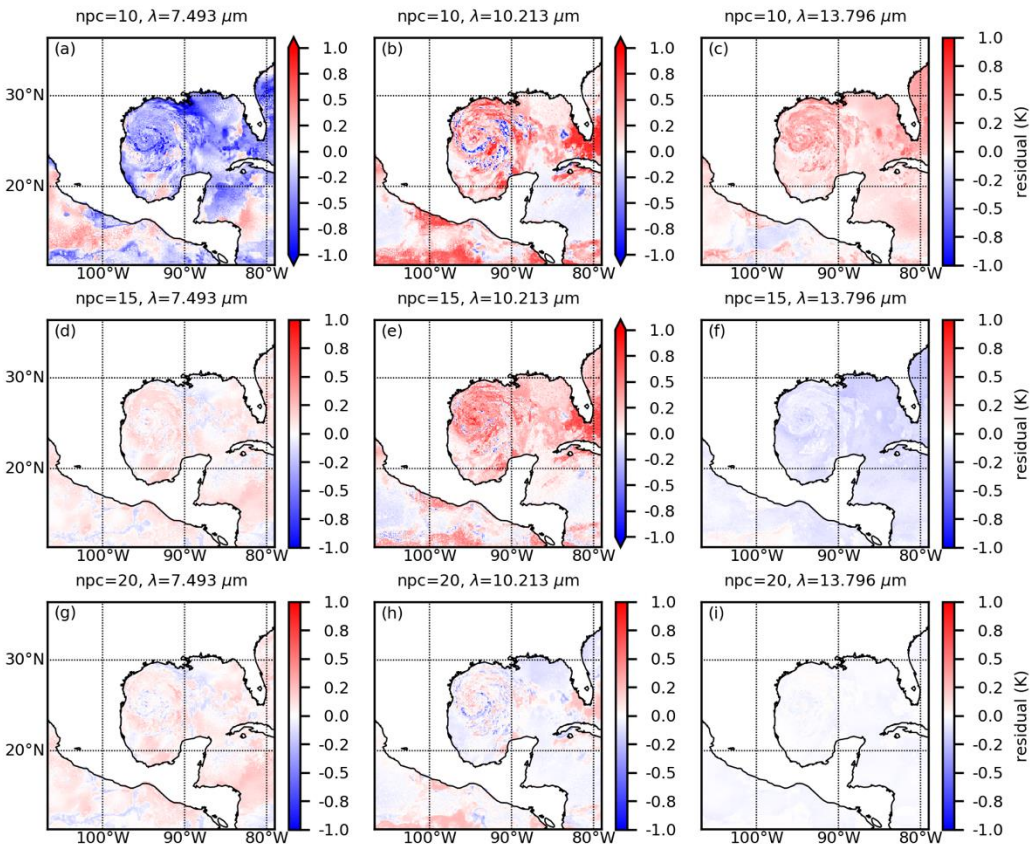
678

679 *Figure 4: (a) variance explained by each PC, (b) variance explained ratio for each PC, (c)*

680 *the 1st PC, and (d) the 2nd PC of the full-spectral-resolution BTs (red) and the subset BTs*

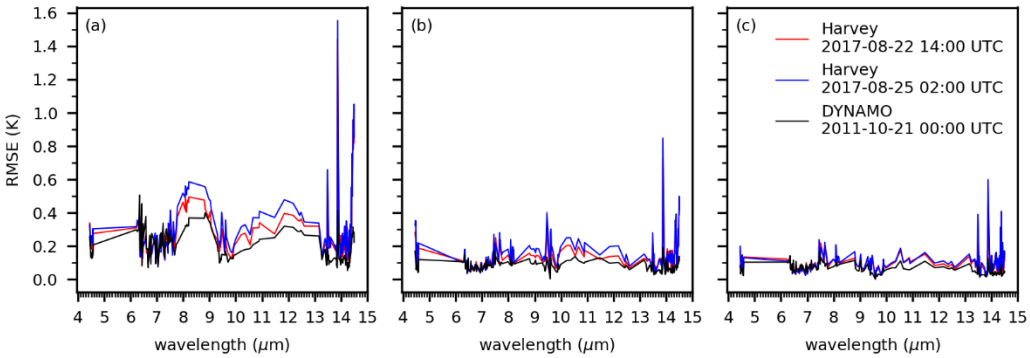
681 *(blue).*

682



683

684 *Figure 5: The reconstruction error when the first 10 (panels a-c), 15 (panels d-f), or 20*
685 *(panels g-i) leading PCs were used to reconstruct the hyperspectral BTs simulated from*
686 *the 1st member of case 3, i.e., the well-developed Harvey (2017) case, for a water vapor*
687 *channel with wavelength 7.493 μm (a, d, and g), a window channel with wavelength 10.213*
688 *μm (b, e, and h), and a temperature sounding channel with wavelength 13.796 μm (c, f, i).*



689

690 *Figure 6: Root mean square of reconstruction error over the entire domain of the*

691 *DYNAMO case (black), the 1st member of Harvey (2017) case at 1400 UTC 22 August 2017*

692 *(red), and at 0200 25 August 2017 (blue) for all the channels when using (a) 10, (b) 15,*

693 *and (c) 20 PCs.*

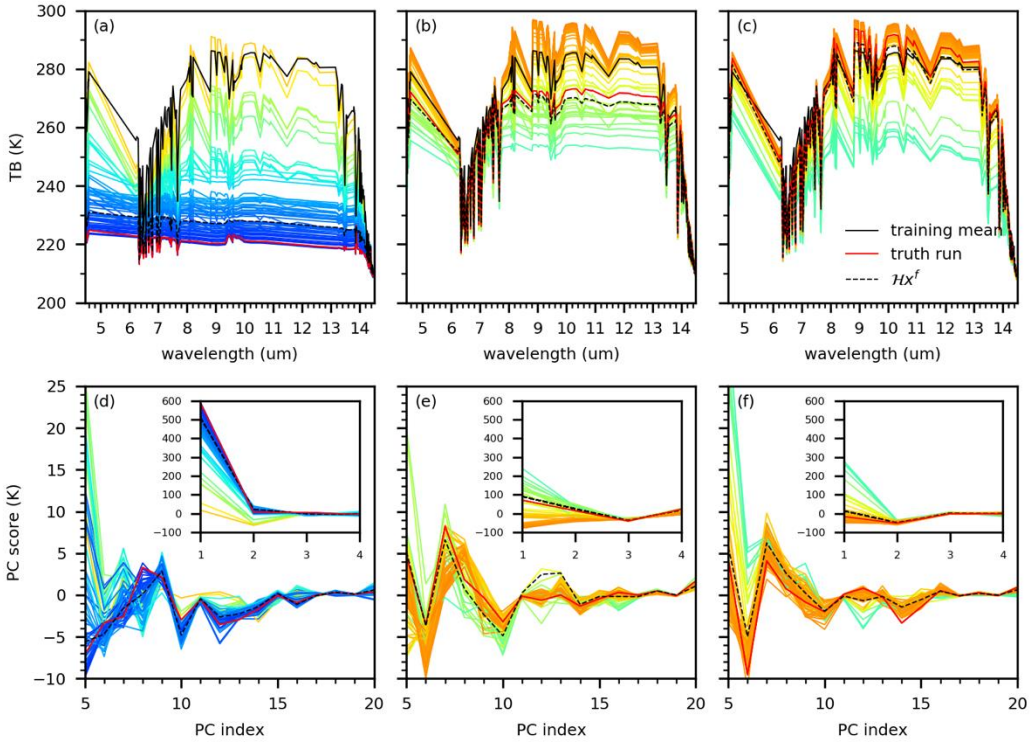
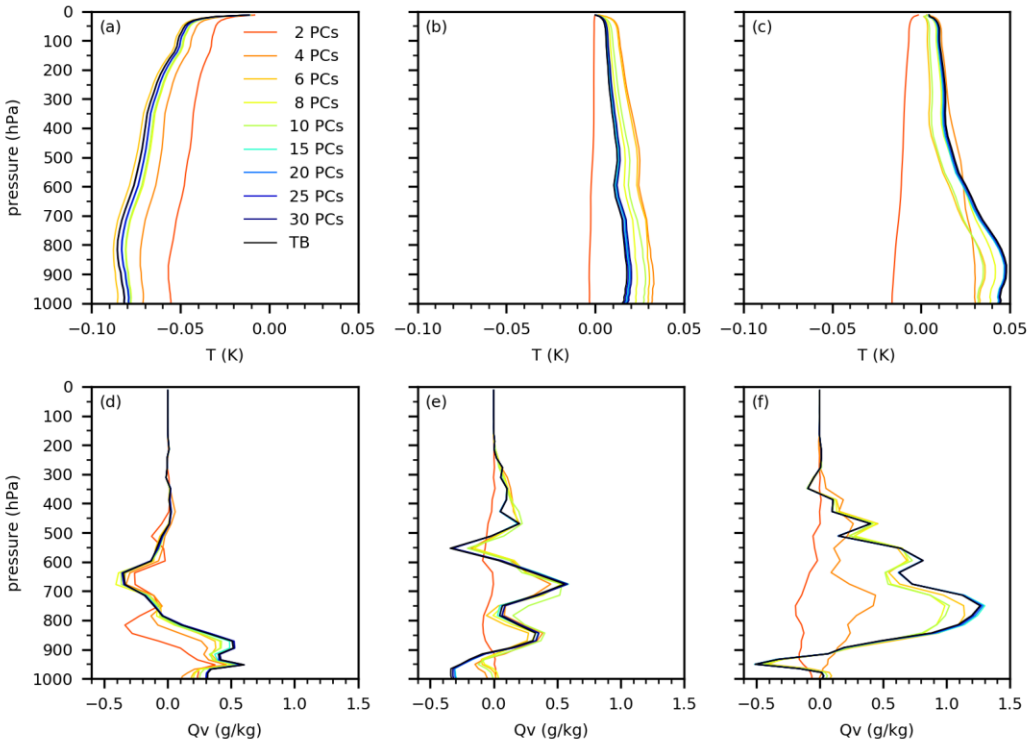


Figure 7: Simulated hyperspectral BTs of the 40th ensemble member (red lines), that of the other 59 ensemble members (other colored lines), hyperspectral BT calculated using ensemble mean atmospheric states as inputs to CRTM (black dashed lines), and the mean of hyperspectral BT of the training dataset (black lines) at location (a) B1, (b) B2, and (c) B3 labeled on Figure 1(b), and the 20 leading PC scores of the 40th ensemble member (red lines), that of the other 59 ensemble members (other colored lines), and that of the ensemble mean (black dashed lines) at location (d) B1, (e) B2, and (f) B3 labeled on Figure 1(b). The inner plots of panels (d-f) show PC scores corresponding to PC1~PC4 while the outer plots show that corresponding to PC5~PC20.



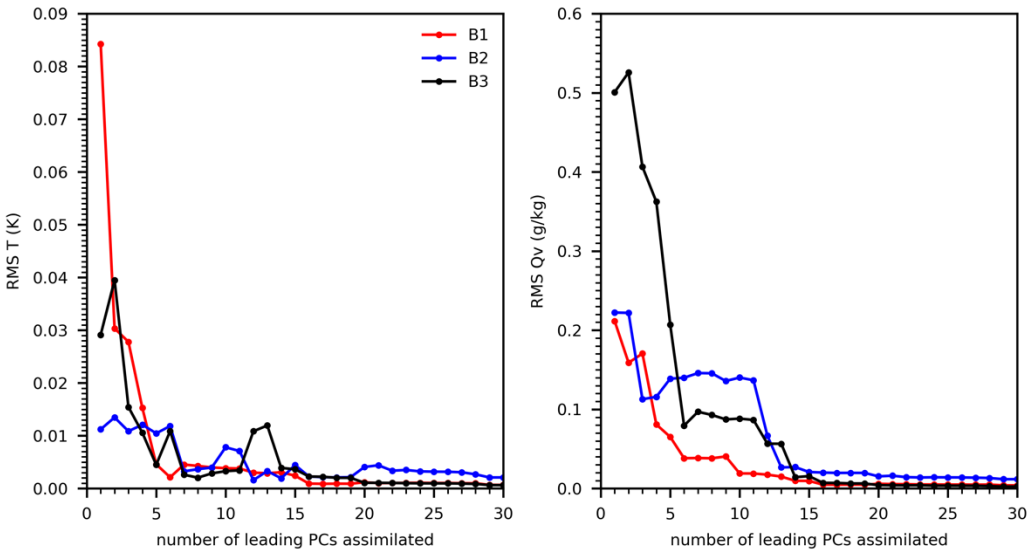
706

707

708 *Figure 8: EnKF increments on (a-c) temperature profiles and (d-f) from assimilating*

709 *hyperspectral BTs (black) and assimilating 2 to 30 leading PC scores (colored lines). The*

710 *legend in panel (a) shows the number of PC scores assimilated according to each color.*



711

712

713 *Figure 9: RMS of the difference between increments from assimilating AIRS brightness*

714 *temperature and PC scores calculated over all layers for (a) temperature profiles and (b)*

715 *water vapor profiles at the locations B1 (red), B2 (blue), and B3 (black) when different*

716 *number of PCs were assimilated.*

717