Evaluating the 3D EnKF - VAR Hybrid Data Assimilation in GSI for Surface and Upper Level Analyses

ZHENG QI WANG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

Graduate Program in
EARTH AND SPACE SCIENCE

York University
Toronto, Ontario

August 2018

Abstract

This study examines the 3 – dimensional analysis produced using the Hybrid Ensemble Kalman Filter (EnKF) – Variational (VAR) Data Assimilation in the Gridpoint Statistical Interpolation (GSI) System. The data assimilation ingests the 1 – hour forecast High-Resolution Rapid Refresh (HRRR) and The Global Ensemble Forecast System, as the background and ensemble member set, respectively. Also, the conventional and satellite radiance observations are assimilated. The analysis covers a CONUS domain and has a 3 km horizontal resolution with 50 vertical native levels. The experiments focus on the advantages of using the flow – dependent background error in the hybrid scheme to dynamically characterize the model background error based on the flow of the day. From the case study results, the hybrid scheme has a higher accuracy in 2m temperature and 10m winds speed than the background and the 3D VAR scheme, especially in regions of weather systems such as frontal boundaries and low – pressure centers. Statistical comparisons of the surface analysis indicated the hybrid scheme outperformed the background and 3D VAR, but is unable to surpass the results from the Real – Time Mesoscale Analysis (RTMA). Also, the impact of the flow – dependent background error covariance in the hybrid scheme was compared with the terrain – following background error covariance in the RTMA. Upper level analysis comparison suggests the hybrid has a lower RMSE than the background and the 3D VAR for the lower and mid atmosphere but have similar results for the upper atmosphere. A brief sensitivity test on the vertical localization showed little impact on the upper level analysis. Lastly, the benefit of assimilating satellite radiance observation and the performance of the enhanced radiance bias correction in GSI was examined.

Acknowledgement

I would like to thank my supervisor Dr. Yongsheng Chen for the guidance and support in helping me to gain the experience in data assimilation. A special gratitude to Dr. Peter Taylor, who has taken me as a research assistance since high school and given me the opportunity to explore the research field in atmosphere science. Special thanks towards my other committee member Dr. Mark Gordon and Dr. Steven Wang for taking their time to give valuable feedback. Also thank you to The Weather Network for the financial support and feedback. I would like to express my appreciation to my girlfriend Garra, my mom, Kim and Dad, Mike for their support and the motivation talks during the entire journey. Lastly, it wouldn't have been an exciting two years without my friends and colleagues Brandon, Tim and Stefan.

Table of Contents

List of Tables

List of Figures

1 Introduction

As seen in The National Weather Services (NWS), The Weather Network (TWN) and the atmospheric community, there is an increasing demand for high-resolution and frequent updated surface and upper-level analyses. Within the forecasting and modeling community, the analyses provide a tool for situational awareness, nowcasting, forecast verification, model calibration and bias correction. While TWN has implemented a surface analysis technique to support its clients and users with point-based current weather information, other applications include climate-related studies through reanalysis systems, the energy industry and supporting air quality studies.

Current surface analysis, such as The Real-Time Mesoscale Analysis (RTMA) provides a 2.5 km spatial resolution representation of near-surface weather conditions, such as 2m temperature, 2m specific humidity, 2m dew point temperature, 10m wind speed/gust and surface visibility. This data assimilation system ingests near surface observation, model background and utilizes the static terrain-following background error covariance within 2D variational approach to produce gridded surface analysis (De Pondeca et al. 2011). RTMA delivers a more accurate analysis than the ones provided by the operational hourly NWP models, such as High-Resolution Rapid Refresh (HRRR) and Rapid Refresh (RAP). Other analysis products include the 3D RTMA and Rapid Updating Analysis (RUA) that is currently being developed by The Environmental Modeling Center (EMC) at The National Center of Environmental Predictions (NCEP) and The Global System Division (GSD) in The National Oceanic & Atmospheric Administration (NOAA). By extending the 2D analysis into a 3D analysis, it provides a 3D

representation of the current atmosphere through analysis and diagnostic variables such as temperature, wind, moisture, hydrometeors, clouds. The 3D variational data assimilation approach provides the capability to ingest upper-level observations such as radiosondes, satellite data, radar and aircraft measurements.

Additional comparisons had been made by Ancell et al. 2014, who studied the analysis results from their in-house 4km and 12km resolution RTMA and 2D Ensemble Kalman Filter (EnKF) data assimilation systems. Although the RTMA and EnKF produce analysis for various variables, the comparison is limited only to examine surface wind and temperature. The findings show that EnKF can produce a finer-scaled spatial structured surface analysis than RTMA for a mountainous region. This is consistent for variables such as 10m UV wind components and 2m temperature at both 4km and 12km resolution. A statistical comparison is conducted of both approaches at 4km and 12km resolutions over a two – month period. Results show that at 4km resolution, EnKF produces a more accurate wind speed analysis than RTMA for 90% of the same period. Similarly, this is also the case at 12km resolution, for 95% of the time. However, the temperature analysis generated by RTMA tend to outperform EnKF for both resolution specifications.

The surface analyses that were mentioned above were produced from either the variational or ensemble data assimilation. In both schemes, the objective is to provide the best analysis based on the influence between the background and the assimilated observations. The weighting between the background and observation is determined by the background error covariance and the observation error covariance. In the variational scheme, the background and the observation

terms form the cost function to solve the analysis control variables through minimization. In particular, the background error covariance used in the variational scheme is static and has near homogeneous and isotropic influences on the increments (Lorenc et al. 2000). However, a physically well – represented background error covariance can vary, depending on the flow of the day (Wang et al. 2008). Hence, the flow-dependent background error covariance in the ensemble scheme provides a necessary alternative that has isentropic influences that allow the increments to follow the structure of the flow of the day. For example, the influence of temperature increment will align with the structure of weather systems such as a frontal boundary or a low-pressure system. The flow-dependent background error covariance is estimated using the Monte Carlo method, which samples from the atmospheric probability density function (pdf) through a set of ensemble members (Hamill and Snyder 2000; Hamill 2001). In addition, Kalman Filter updates the analysis and the error covariance by assimilating the observations (Chen and Snyder 2006). Therefore, the flow-dependent error covariance in EnKF allows for the more appropriate weighting between the observation and the background based on the flow of the day (Houtekamer et al. 2005; Whitaker et al. 2008). However, the smaller set of ensembles lacks the ability to fully represent the pdf which results in the background covariance estimates to suffer from sampling error (Hamill and Snyder 2000), while larger sets of ensembles become too computationally expensive and unpractical for operational use. Overall, both the variational and ensemble data assimilation scheme has its advantages and disadvantages.

The objective of this study is to examine the benefit of combining the variational and ensemble scheme into an EnKF-VAR Hybrid data assimilation system for surface and upper-

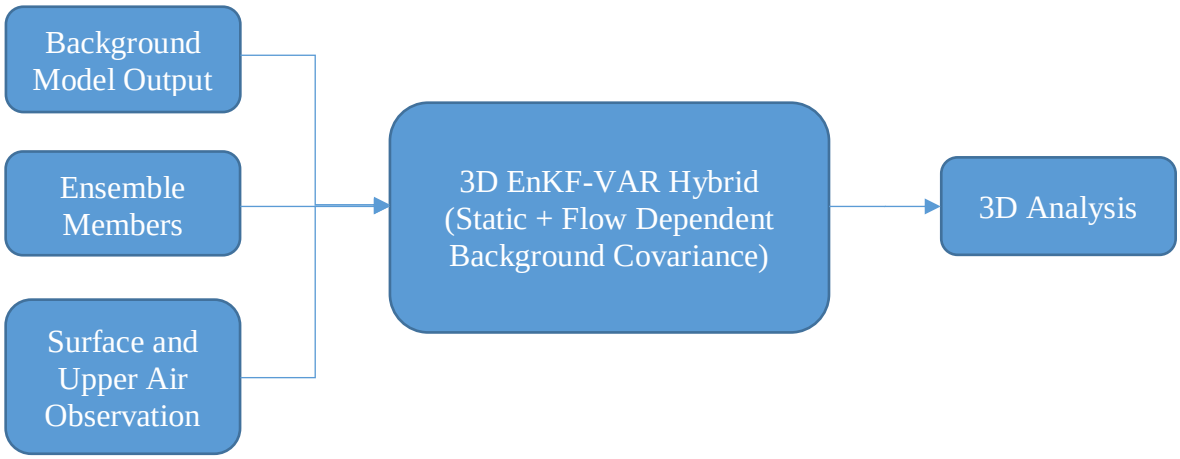
level analysis. In particular, the flow – dependent background error covariance from the ensemble scheme is incorporated into the static background error covariance matrix and the minimization cost function from the variational scheme (Wang 2010). The amalgamation of the two background error covariance matrices in the hybrid scheme ameliorates the sampling error issue with smaller ensemble members in used EnKF (Wang 2010). Studies by Hamill and Snyder (2000), Wang et al. (2013) and Wang et al. (2007a) suggests the hybrid data assimilation provides a more accurate analysis than solely using variational or EnKF scheme, especially when the size of ensemble is small and the model error is large (Wang et al. 2007a, 2008a). As a result, the reduction of ensemble size reduces the necessary computational resource, while improving the accuracy of the analysis through the combination of static and flow – dependent background error covariance. Lastly, the cross – variable covariance defined in the ensemble component within the hybrid scheme enable the temperature, wind and moisture increments to establish cross-variable relationship (Wang et al. 2008b). Meanwhile, the 2D VAR in RTMA only have uni-variable covariance, meaning that there is no correlation among the temperature, wind and moisture increments (De Pondeca et al. 2011). Although the hybrid scheme is widely used within operational NWP community (Buehner 2005; Wang et al. 2008b; Buehner et al. 2010b), it has not been implemented for operational analysis purposes. A brief summary of the background used in the 3D EnKF-VAR Hybrid, 3D EnKF, 3D VAR and 2D VAR RTMA is shown in Table 1.1. In addition, Figure 1.1.1 (a) –(d) illustrates the inputs, outputs and the algorithms for 3D EnKF-VAR Hybrid, 3D EnKF, 3D VAR and 2D VAR RTMA data assimilation schemes.

Background Error Covariance used in the Data Assimilation Schemes

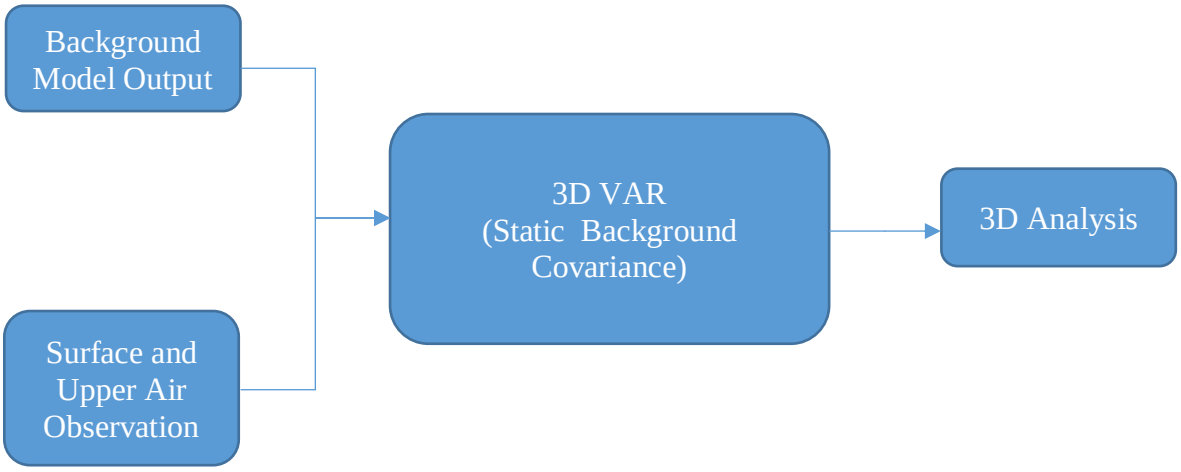
Data Assimilation Scheme	Background Error Covariance	Cross Variable Correlation
3D EnKF-VAR Hybrid	Combination of Static and Flow – Dependent	Yes
3D EnKF	Flow Dependent	Yes
3D VAR	Static	
2D VAR RTMA	Terrain – Following	No

Table 2.1.1: *Summary of the background used in the 3D EnKF-VAR Hybrid, 3D EnKF, 3D VAR and 2D VAR RTMA*

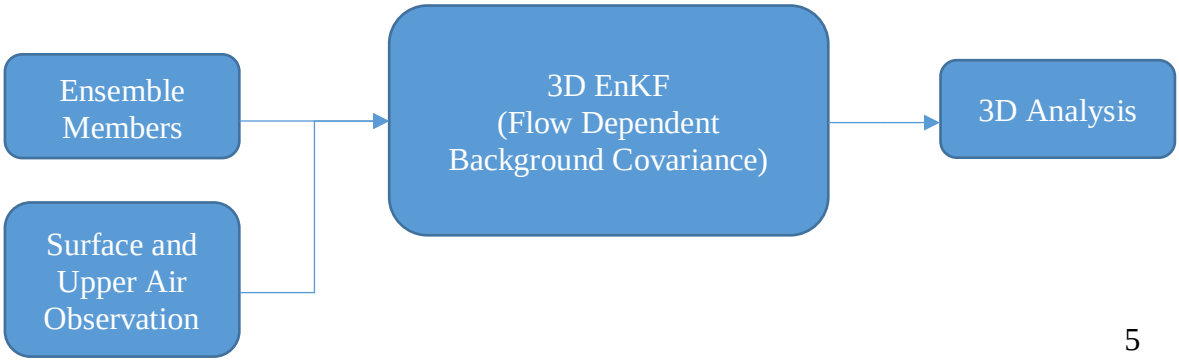
(a) 3D EnKF-VAR Hybrid



(b) 3D VAR



(c) 3D EnKF



(d) 2D VAR RTMA

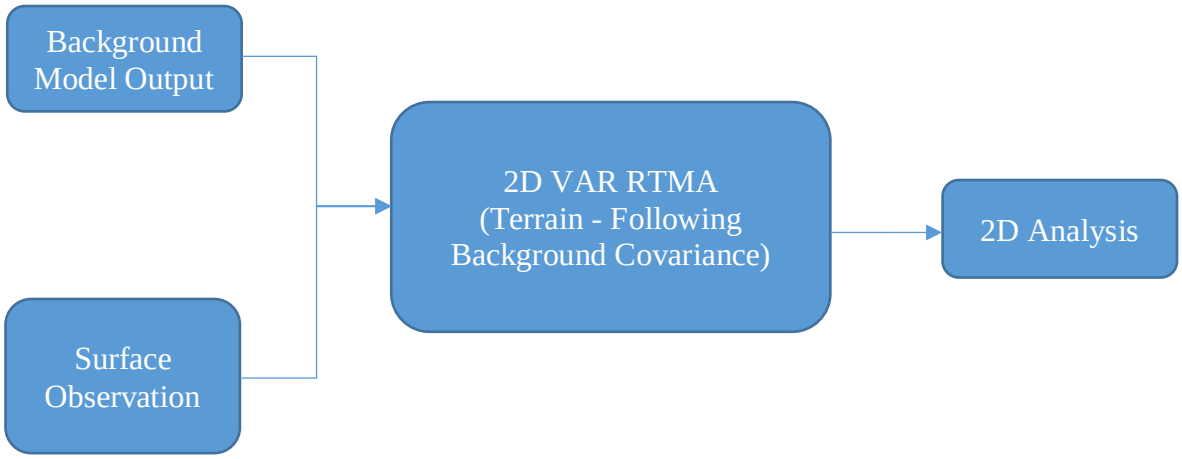


Figure 2.1.1*: Illustrates the inputs, outputs and the algorithms for (a) 3D EnKF-VAR Hybrid, (b) 3D EnKF, (c) 3D VAR and (d) 2D VAR RTMA data assimilation schemes.*

This study will focus on the performance of the 3D analysis produced by using 3D Hybrid Data Assimilation, which aims to investigate four components. The first component examines the effects of flow-dependent background error covariance to fit the surface increment based on current weather systems, such as low-pressure systems and frontal boundaries. The accuracy of the hybrid surface analysis will be compared against the surface analysis from 3D VAR and the existing operational 2D VAR RMTA surface analysis with terrain – following background error covariance. In particular, the impact between the flow-dependent and the terrain background error covariance on surface analysis will be examined. The second component discusses the potential benefit of extending the analysis from 2D to 3D and studies the influence of assimilating upper-level observations on the surface analysis. Specifically, one hopes the assimilation of upper-level observations such as radiosonde, radar data, aircraft measurements can provide an accurate representation of the planetary boundary layer and refine the surface

analysis. In addition, comparison of the upper-level analysis between 3D Hybrid and the 3D VAR scheme will be conducted. The third component consists of a series of sensitivity tests by varying the vertical localization of the ensemble covariance. One can argue that the physical characteristics of near-surface variables are highly localized. Therefore, an improvement of surface analysis could be made by experimenting with different length of vertical localization. Also, any effects on the upper-level analysis will be discussed. Lastly, the fourth component considers the improvement of the upper-level analysis from assimilating satellite radiances. It also touches on the effectiveness of the enhanced radiance bias correction on removing systematic radiance bias and improving the analysis. Overall, the objective of this study is to demonstrate that the hybrid data assimilation is a promising alternative to produce 3D analyses and to show the benefit of assimilation upper-level observation, including satellite radiance, aircraft, radiosonde and radar observations.

The 3D hybrid data assimilation will be conducted using the Gridpoint Statistical Interpolation (GSI) framework. It will ingest the High – Resolution Rapid Refresh model (HRRR) as the background, Global Ensemble Forecast System (GEFS) for ensemble members and the conventional and satellite observations provided by Global Data Assimilation System (GDAS). The 6 – hourly updated 3D analysis covers the CONUS domain with 50 native vertical levels with variables including temperature, UV wind components, specific humidity and surface pressure.

In this study, Section 2: Theoretical Background explains the theoretical background of the variational, EnKF and Hybrid scheme. It also describes the mathematical framework of incorporating ensemble covariance into the variational framework by introducing extended variables. This section also includes an overview of ensemble covariance localization and the enhanced radiance bias correction procedure in GSI. Section 3: Experimental Design discusses data used in the experiments and the configuration of the data assimilation. It also reviews the pre and post process procedures. Next, Section 4: Results and Discussion examines the results and performance of surface analysis produced using hybrid data assimilation. Lastly, Section 5: Conclusion and Future Work summaries this study with suggestions for further improvement in the future

2 Theoretical Background

2.1 Variational Data Assimilation Framework

The primary goal of data assimilation is to provide the best approximation of the current atmospheric state based on the most current available observations (in forms a vector) $\boldsymbol{y}^0$ and background model state (a concatenated vector of all the control variables) $\boldsymbol{x}$. As discussed by Wang (2010) and Shao et al (2016), the 3D Variational Data Assimilation in GSI consists of the minimization of a cost function to solve for the analysis control variables, shown in Eq. (2.1.1).

$$J(\boldsymbol{x}_1') = \frac{1}{2}(\boldsymbol{x}_1')^T \boldsymbol{B}_1^{-1}(\boldsymbol{x}_1') + \frac{1}{2}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}_1'\right)^T \boldsymbol{R}^{-1}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}_1'\right) \quad (2.1.1)$$

In this iterative process for the GSI regional analysis, $\boldsymbol{x}_1'$ represents the increment of the control variables such as stream function, velocity potential, virtual temperature, surface pressure and pseudo-relative humidity. In addition, the weighting of the first and second terms in Eq. (2.1.1) are balanced between the static background error covariance $\boldsymbol{B}_1$ and the observation error covariance $\boldsymbol{R}$. Also, the linearized observation operator $\boldsymbol{H}$ is used to translate the control variables from the model space into the observational space, as denoted by $\boldsymbol{H}\boldsymbol{x}_1'$. Lastly, the innovation vector $\boldsymbol{y}^{0'}$ is defined as the difference between the $\boldsymbol{y}^0$and $\boldsymbol{H}\boldsymbol{x}_1'$.

GSI was originally developed as a 3D variational data assimilation system, where the minimization is preconditioned upon the full static background error covariance, as discussed in Wang (2010). Buehner (2005) reviewed similar precondition technique on the static background

error covariance within the operational 3D Var data assimilation system in the Canadian Meteorological Centre. Due to the large size of the full background error covariance matrix (order of ~10^{14}), it is unpractical to use matrix inversion operations on $\boldsymbol{B}_1$ in the cost function minimization, as shown in Eq. (2.1.1). In other words, it would be too computationally expensive to follow the suggested cost function framework to compute the analysis. The rest of the section illustrates the mathematical framework that is implemented in GSI regional 3D-VAR to avoid the use of matrix inversion, as discussed in Wang (2010).

First consider the 3D VAR cost function in Eq. (2.1.1) when inverting the static background error covariance $\boldsymbol{B}_1$. The minimization of the cost function uses the preconditioned conjugate method, where one must set a precondition prior to minimization. In this case, Wang (2010) set it as:

$$\boldsymbol{z}_1' = \boldsymbol{B}_1^{-1}(\boldsymbol{x}_1') \qquad (2.1.2a)$$

Or,

$$\boldsymbol{x}_1' = (\boldsymbol{z}_1')\,\boldsymbol{B}_1 \qquad (2.1.2b)$$

The 3D Var cost function will become:

$$J(\boldsymbol{x}_1') = \frac{1}{2}(\boldsymbol{x}_1')^T\boldsymbol{z}_1' + \frac{1}{2}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}_1'\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}_1'\right) \quad (2.1.3a)$$

$$J(\boldsymbol{z}_1') = \frac{1}{2}(\boldsymbol{z}_1')^T\boldsymbol{B}_1\boldsymbol{z}_1' + \frac{1}{2}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}_1'\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}_1'\right) \quad (2.1.3b)$$

By taking the gradient with respect to $\boldsymbol{x}_1'$ and using the chain rule, Eq. (2.1.3a) will equate to:

$$\nabla_{\boldsymbol{x}_1'} J = \boldsymbol{z}_1' + \boldsymbol{H}^T \boldsymbol{R}^{-1}\left(\boldsymbol{H}\boldsymbol{x}_1' - \boldsymbol{y}^{0\prime}\right) \qquad (2.1.4)$$

Similarly, one can take the gradient of Eq. (2.1.3b) with respect to $\boldsymbol{z}_1'$:

$$\nabla_{\boldsymbol{z}_1'} J = \boldsymbol{B_1}\boldsymbol{z}_1' + \boldsymbol{B_1}\boldsymbol{H}^T \boldsymbol{R}^{-1}\left(\boldsymbol{H}\boldsymbol{x}_1' - \boldsymbol{y}^{0\prime}\right) = \boldsymbol{B_1}\nabla_{\boldsymbol{x}_1'} J \qquad (2.1.4)$$

Therefore, one can use the preconditioned minimization of the cost function to avoid inverting the background error covariance $\boldsymbol{B_1}$, as seen in Eq. (2.1.4)

2.2 Ensemble Kalman Filter Data Assimilation Framework

As previously mentioned, data assimilation consists of utilizing the observation of a state $\boldsymbol{y}$ and the prior model state $\boldsymbol{x}$ to obtain the best estimate of the true state $\boldsymbol{x}$ of a system at a given time t. In particular, Ensemble Kalman Filter (EnKF) follows the theoretical background behind Bayesian probability. Specifically, the true state of $\boldsymbol{x_t}$, at a given time t, can be estimated using the conditional probability density function (PDF), as stated in Chen and Snyder (2007). Houtekamer and Fuqing (2016) explained that EnKF produces a representation of the PDF using a series of states with perturbation, called ensembles members to compute the sample covariance error. The ensemble members are updated with new observations and propagate in time until the next time steps. The update equations to estimate the analysis are expressed in Eq. (2.2.2) and (2.2.3).

$$\overline{\boldsymbol{x^a}} = \overline{\boldsymbol{x^b}} + \boldsymbol{K}\left(\boldsymbol{y^o} - \overline{\boldsymbol{Hx^b}}\right) \tag{2.2.2}$$

$$\boldsymbol{K} = \boldsymbol{P^b H^T}\left(\boldsymbol{HP^b H^T} + \boldsymbol{R}\right)^{-1} \tag{2.2.3}$$

From Equation (2.2.2) and (2.2.3), $\overline{\boldsymbol{x^a}}$ is the analysis mean of the state vector, while $\overline{\boldsymbol{x^b}}$ refers to background state vector or the ensemble mean from the background model. The Kalman Gain matrix $\boldsymbol{K}$ gives the weighting between the observations and the background model, which is dependent on the observation error covariance $\boldsymbol{R}$ and the ensemble background error covariance $\boldsymbol{P}$. The observation vector $\boldsymbol{y^o}$ represents the measured quantity at each observation location and $\boldsymbol{H}$ represents the linear observation operator to translate the background state variables from model space into observation space.

In order to compute the Kalman Gain Matrix in Eq. (2.2.3), it is unnecessary and computationally expensive to use the full background error covariance matrix. A sample covariance matrix can rather be approximated using the set of ensembles and compute $\boldsymbol{P^b H^T}$ and $\boldsymbol{H P^b H^T}$, as shown in Eq. (2.2.4) and (2.2.5).

$$\boldsymbol{P^b H^T} = cov\left(\rho \circ \boldsymbol{x^b}, \boldsymbol{H x^b}\right) = \frac{1}{N_e - 1}\sum_{k=1}^{N_e}(\boldsymbol{x_k^b} - \overline{\boldsymbol{x^b}})(H\left(\boldsymbol{x_k^b}\right) - \overline{H(\boldsymbol{x^b})}) \quad (2.2.4)$$

$$\boldsymbol{H P^b H^T} = cov\left(\boldsymbol{H x^b}, \boldsymbol{H x^b}\right) = \frac{1}{N_e - 1}\sum_{k=1}^{N_e}\left(H\left(\boldsymbol{x_k^b}\right) - \overline{H(\boldsymbol{x^b})}\right)^2 \quad (2.2.5)$$

where,

$$\overline{\boldsymbol{x^b}} = \frac{1}{N_e}\sum_{k=1}^{N_e}\boldsymbol{x_k^b} \qquad (2.2.6)$$

$$\overline{H(\boldsymbol{x^b})} = \frac{1}{N_e}\sum_{k=1}^{N_e}H(\boldsymbol{x_k^b}) \qquad (2.2.7)$$

N_e represents the number of ensemble members.

As stated by Chen and Snyder (2007), perturbations that deviate from the analysis ensemble mean can be determined using ensemble square root filter. The perturbation for each ensemble member is updated for the k^{th} ensemble member is defined by:

$$\boldsymbol{x}_k^{\boldsymbol{a}} - \overline{\boldsymbol{x}^{\boldsymbol{a}}} = \boldsymbol{x}_k^{\boldsymbol{b}} - \overline{\boldsymbol{x}^{\boldsymbol{b}}} - \alpha \boldsymbol{K}\left(H\left(\boldsymbol{x}_k^{\boldsymbol{b}}\right) - \overline{H\left(\boldsymbol{x}_k^{\boldsymbol{b}}\right)}\right) \qquad (2.2.8)$$

$$\alpha = \left[1 + \sqrt{\frac{\sigma_o^2}{var\left(H(\boldsymbol{x}^{\boldsymbol{b}})\right) + \sigma_o^2}}\right] \qquad (2.2.9)$$

where the standard deviation of each observation type is denoted by σ_o. Also, $var\left(H\left(\boldsymbol{x}^{\boldsymbol{b}}\right)\right)$ refers to the variance of $H\left(\boldsymbol{x}^{\boldsymbol{b}}\right)$.

By combining the analysis ensemble mean and ensemble deviation components, it completes the EnKF update cycle. Since this experiment solely focused on obtaining ensemble analysis, it is unnecessary to propagate the analysis update forward in time for this study.

2.3 GSI 3D Hybrid Data Assimilation Framework

In the section, the minimization cost function of the 3D Ensemble – Variational Hybrid scheme (3D Hybrid) and its components will be briefly explained. The framework of the Regional 3D Hybrid in GSI to produce analysis was discussed in Wang et al (2007a), Wang et al (2007b), Wang (2010). The analysis increment in 3D Hybrid $\boldsymbol{x}'$ is defined by the combination of the two terms in Eq. (2.3.1). The first term considers the increment associated with the static background covariance from the variational scheme, $\boldsymbol{x}_1'$. The second term reflects on the analysis increment associated with the flow-dependent background covariance from the ensemble scheme.

$$\boldsymbol{x}' = \boldsymbol{x}_1' + \sum_{k=1}^{K} (\boldsymbol{a}_k \circ \boldsymbol{x}_k^e) \qquad (2.3.1)$$

Specifically, the second term in Eq. (2.3.1) depicts the linear combination of extended control variables $\boldsymbol{a}_k$ and ensemble perturbation $\boldsymbol{x}_k^e$. The ensemble perturbation is defined as:

$$\boldsymbol{x}_k^e = \frac{1}{\sqrt{K-1}} (\boldsymbol{x}_k - \overline{\boldsymbol{x}}) \qquad (2.3.2)$$

In Eq. (2.3.2), the ensemble forecast is represented by $\boldsymbol{x}_k$, while the mean ensemble forecast is depicted by $\overline{\boldsymbol{x}}$. The subscripts k in the variables $\boldsymbol{a}_k$ and $\boldsymbol{x}_k^e$ denotes the kth member of the ensemble and K is the ensemble size. In addition, the operator indicated by " ∘ " represents the Schur product, which is an element by element product between two vectors.

In order to obtain the analysis increment $\boldsymbol{x}'$, one must minimize the hybrid cost function, as shown in Eq. (2.3.3a) or (2.3.3b).

$$J(\boldsymbol{x}_1', \boldsymbol{a}) = \beta_1 J_1 + \beta_2 J_2 + J_o \qquad (2.3.3a)$$

In simple terms, the hybrid cost function in Eq. (2.3.3a) comprises three components. The first term states the background term that is related with the static covariance in the variational scheme. The second term represents the background term associated with the flow-dependent covariance in the ensemble scheme. Lastly, the third term is the observation term.

One can illustrate an detailed expression of the hybrid cost function in Eq. (2.3.3b).

$$J(\boldsymbol{x}_1', \boldsymbol{a}) = \beta_1 \frac{1}{2}(\boldsymbol{x}_1')^T \boldsymbol{B}_1^{-1}(\boldsymbol{x}_1') + \beta_2 \frac{1}{2}(\boldsymbol{a})^T \boldsymbol{A}^{-1}(\boldsymbol{a}) + \frac{1}{2}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}'\right)^T \boldsymbol{R}^{-1}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}'\right) \quad (2.3.3b)$$

In particular, $\boldsymbol{B}_1$ matrix is the static background covariance, which defines the spatial covariance of the analysis increment that is associated with the variational scheme, $\boldsymbol{x}_1'$. Next, the vector $\boldsymbol{a}$ is the concatenating vector of the extended control variables $\boldsymbol{a}_k$ for all of K numbers of ensemble members, as seen in Wang et al (2007b) and expressed in Eq. (2.3.4).

$$\boldsymbol{a} = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_k \\ \vdots \\ \boldsymbol{a}_K \end{pmatrix} \qquad (2.3.4)$$

In addition, the block-diagonal matrix $\boldsymbol{A}$ is used to constrain the concatenating extended control variable, $\boldsymbol{a}$. For instance, each block in the diagonal of $\boldsymbol{A}$ comprise of the same prescribed correlation matrix, $\boldsymbol{S}$ that constrains the spatial variation of $\boldsymbol{a}_k$.

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{S} & & & \boldsymbol{0} \\ & \ddots & & \\ & & \boldsymbol{S} & \\ & & & \ddots \\ \boldsymbol{0} & & & \boldsymbol{S} \end{pmatrix} \qquad (2.3.5)$$

In other words, the matrix $\boldsymbol{S}$ restricts any spatial correlation among the control variables to a limited radius distance. Hence, this is used as covariance localization. Note the ensemble covariance is not explicitly shown in Eq. (2.3.3b). However, the background error covariance $\boldsymbol{B}$ and the ensemble covariance $\boldsymbol{P}^e$ with localization constraints $\boldsymbol{S}$ are explicitly defined in Eq. (2.3.6), following the framework in Wang et al (2008a). Wang et al (2007b) and Wang et al (2008a) proved the equivalency of the hybrid cost function between implicitly defined ensemble covariance in Eq. (2.3.3b) and explicitly defines the ensemble covariance with localization in Eq. (2.3.6).

$$J(\boldsymbol{x}') = \frac{1}{2}\boldsymbol{x}'^{T}\left(\frac{1}{\beta_1}\boldsymbol{B_1} + \frac{1}{\beta_2}\boldsymbol{P}^{e} \circ \boldsymbol{S}\right)^{-1}\boldsymbol{x}' + \frac{1}{2}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}'\right)^{T}\boldsymbol{R^{-1}}\left(\boldsymbol{y}^{0'} - \boldsymbol{H}\boldsymbol{x}'\right) \quad (2.3.6)$$

Lastly, the weighting factor β_1 and β_2 balances the total background variances contribution between the static background covariance and the flow-dependent background covariance, respectively. According to Hamill and Snyder (2000), Wang et al. (2007a) and Wang et al (2008a), both weighting factors are constrained to conserve the total background error variance, as seen in Eq. (2.3.7).

$$\frac{1}{\beta_1} + \frac{1}{\beta_2} = 1 \qquad (2.3.7)$$

In the practical implementation of the hybrid cost function in Eq. (2.3.3b), the weighting factor between the static and flow-dependent background covariance is actually defined by $\frac{1}{\beta_1}$ and $\frac{1}{\beta_2}$ rather than β_1 and β_2. Section 2.4 provides further detail on this notion while demonstrating the mathematical framework that refrains from inverting the total background. For instance, the non-inverted total background covariance framework in Eq. (2.4.6) shows the weighting factor is expressed as $\frac{1}{\beta_1}$ and $\frac{1}{\beta_2}$.

2.4 Incorporating ensemble covariance from 3D VAR

As previously mentioned in Section 2.1, it is necessary to establish a precondition upon the full background error covariance in the variational cost function to avoid inverting a large dimensional matrix. This technique significantly reduces the computational cost to allow for the minimization process to be computationally feasible. On the other hand, the GSI's 3D Hybrid data assimilation scheme follows a similar precondition mathematical framework to incorporate ensemble covariance with variational background error covariance in the Hybrid cost function, as shown by Wang (2010). This can be done by extending the control variables and the background error covariance. In particular, the new control variables are defined by Eq. (2.4.1)

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{a} \end{pmatrix} \qquad (2.4.1)$$

Revisiting the hybrid analysis increment in Eq. (2.3.1) and digress further:

$$\boldsymbol{x}' = \boldsymbol{x}_1' + \sum_{k=1}^{K} (\boldsymbol{a}_k \circ \boldsymbol{x}_k^{\boldsymbol{e}}) = \ \boldsymbol{x}_1' + [diag(\boldsymbol{x}_1^{\boldsymbol{e}}) \ldots diag(\boldsymbol{x}_k^{\boldsymbol{e}})]\boldsymbol{a}. \qquad (2.4.1)$$

Note that the $diag$ operator reforms a vector into a diagonal matrix. By denoting:

$$\boldsymbol{D} = [diag(\boldsymbol{x}_1^{\boldsymbol{e}}) \ldots diag(\boldsymbol{x}_k^{\boldsymbol{e}})] \qquad (2.4.2)$$

Eq. (2.4.1) becomes:

$$\boldsymbol{x}' = \boldsymbol{x}_1' + \boldsymbol{D}\boldsymbol{a} \qquad (2.4.3)$$

In order to simplify further, consider the Eq (2.4.4), where $\boldsymbol{I}$ is an identity matrix:

$$\boldsymbol{C} = (\boldsymbol{I}, \boldsymbol{D}) \qquad (2.4.4)$$

Eq. (2.4.2) becomes:

$$\boldsymbol{x}' = (\boldsymbol{I}, \boldsymbol{D}) \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{a} \end{pmatrix} = \boldsymbol{C}\boldsymbol{x} \qquad (2.4.5)$$

By extending the background error covariance in the hybrid scheme, the first term in Eq. (2.3.3b) is preconditioned by the static background covariance $\boldsymbol{B}_1$ while the second term in Eq. (2.3.3b) is preconditioned with respect to $\boldsymbol{A}$ rather than the ensemble covariance. As previously mentioned in Section 2.3, the ensemble covariance is not explicitly defined in Eq. (2.3.3b). Overall, both $\boldsymbol{A}$ and $\boldsymbol{B}$ constrain the covariance of the extended variables in Eq. (2.4.1). Subsequently, two matrices can be joint in Eq. (2.4.6).

$$\boldsymbol{B} = \begin{pmatrix} \frac{1}{\beta_1}\boldsymbol{B_1} & 0 \\ 0 & \frac{1}{\beta_2}\boldsymbol{A} \end{pmatrix} \qquad (2.4.6)$$

Similar to the preconditioning used in Eq. (2.1.2a) of 3D Var scheme, the preconditioned conjugate gradient minimization can also be implemented in the hybrid scheme, using the newly defined $\boldsymbol{B}$ in Eq. (2.4.6).

$$\boldsymbol{z} = \boldsymbol{B}^{-1}(\boldsymbol{x}) = \begin{pmatrix} \beta_1 \boldsymbol{B}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \beta_2 \boldsymbol{A}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{a} \end{pmatrix} = \begin{pmatrix} \beta_1 \boldsymbol{B}_1^{-1} \boldsymbol{x}_1' \\ \beta_2 \boldsymbol{A}^{-1} \boldsymbol{a} \end{pmatrix} \qquad (2.4.7)$$

By following the same mathematical framework as the 3D Var scheme, one can demonstrate that $\nabla_{\boldsymbol{z}_1'} J = \boldsymbol{B_1} \nabla_{\boldsymbol{x}_1'} J$ also satisfies for the hybrid scheme. Eq. (2.4.8) and Eq. (2.4.9) shows the gradient of the hybrid cost function in Eq. (2.3.3b) with respect to $\boldsymbol{x}_1'$ and $\boldsymbol{a}$, respectively.

$$\nabla_{\boldsymbol{x}_1'} J = \beta_1 \boldsymbol{B}_1^{-1} \boldsymbol{x_1'} + \boldsymbol{H}^T \boldsymbol{R^{-1}} \left(\boldsymbol{H x'} - \boldsymbol{y^{0'}} \right) \qquad (2.4.8)$$

$$\nabla_{\boldsymbol{a}} J = \beta_2 \boldsymbol{A}^{-1} \boldsymbol{a} + \boldsymbol{D}^T \boldsymbol{H^T R^{-1}} \left(\boldsymbol{H x'} - \boldsymbol{y^{0'}} \right) \qquad (2.4.9)$$

One can combine Eq. (2.4.8) and (2.4.9) together:

$$\boldsymbol{\nabla_x J} = \begin{pmatrix} \boldsymbol{\nabla}_{\boldsymbol{x}_1'} \boldsymbol{J} \\ \boldsymbol{\nabla_a J} \end{pmatrix} = \begin{pmatrix} \beta_1 \boldsymbol{B}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \beta_2 \boldsymbol{A}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{a} \end{pmatrix} + \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{D} \end{pmatrix}^T \boldsymbol{H}^T \boldsymbol{R^{-1}} \left(\boldsymbol{H x'} - \boldsymbol{y^{0'}} \right) \qquad (2.4.10)$$

Using Eq. (2.4.1), (2.4.4) and (20), one can simplify Eq. (2.4.7) further:

$$\boldsymbol{\nabla_x J} = \begin{pmatrix} \boldsymbol{\nabla}_{\boldsymbol{x}_1'} \boldsymbol{J} \\ \boldsymbol{\nabla_a J} \end{pmatrix} = \boldsymbol{B}^{-1}(\boldsymbol{x}) + \boldsymbol{C}^T \boldsymbol{H}^T \boldsymbol{R}^{-1} \left(\boldsymbol{H C x} - \boldsymbol{y}^{O'} \right)$$

$$= \boldsymbol{z} + \boldsymbol{C^T H^T R^{-1}}\left(\boldsymbol{HCx} - \boldsymbol{y^{0'}}\right) \qquad (2.4.11)$$

Subsequently, one can also take the gradient of the hybrid cost function in Eq. (2.4.3b) with respect to $\boldsymbol{z}$. Note that $\boldsymbol{z}$ matrix comprises two components, $\beta_1 \boldsymbol{B_1^{-1}} \boldsymbol{x_1'}$ and $\beta_2 \boldsymbol{A^{-1}} \boldsymbol{a}$. Therefore, the gradient of the cost function can be equivalently determined by taking the gradient with respect of $\beta_1 \boldsymbol{B_1^{-1}} \boldsymbol{x_1'}$ and $\beta_2 \boldsymbol{A^{-1}} \boldsymbol{a}$, as shown in Eq. (2.4.12) and (2.4.13).

$$\nabla_{\beta_1 \boldsymbol{B_1^{-1}} \boldsymbol{x_1'}} J = \boldsymbol{x_1'} + \frac{1}{\beta_1} \boldsymbol{B_1 H^T R^{-1}}\left(\boldsymbol{Hx'} - \boldsymbol{y^{0'}}\right) \qquad (2.4.12)$$

$$\nabla_{\beta_1 \boldsymbol{A^{-1}} \boldsymbol{a}} J = \boldsymbol{a} + \frac{1}{\beta_2} \boldsymbol{A D^T H^T R^{-1}}\left(\boldsymbol{Hx'} - \boldsymbol{y^{0'}}\right) \qquad (2.4.13)$$

Next, Eq (2.4.12) and (2.4.13) can be combined to form Eq. (2.4.14):

$$\nabla_{\mathbf{z}} J = \begin{pmatrix} \nabla_{\beta_1 \boldsymbol{B_1^{-1}} \boldsymbol{x_1'}} J \\ \nabla_{\beta_1 \boldsymbol{A^{-1}} \boldsymbol{a}} J \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_1'} \\ \boldsymbol{a} \end{pmatrix} + \begin{pmatrix} \frac{1}{\beta_1} \boldsymbol{B_1} & 0 \\ 0 & \frac{1}{\beta_2} \boldsymbol{A} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{D} \end{pmatrix}^T \boldsymbol{H^T R^{-1}}\left(\boldsymbol{Hx'} - \boldsymbol{y^{0'}}\right) \quad (2.4.14)$$

Substitute Eq. (2.4.1), (2.4.4), (2.4.6) into Eq. (2.4.14), we obtain:

$$\nabla_{\mathbf{z}} J = \begin{pmatrix} \nabla_{\beta_1 \boldsymbol{B_1^{-1}} \boldsymbol{x_1'}} J \\ \nabla_{\beta_1 \boldsymbol{A^{-1}} \boldsymbol{a}} J \end{pmatrix} = \boldsymbol{x} + \boldsymbol{B C^T H^T R^{-1}}\left(\boldsymbol{Hx'} - \boldsymbol{y^{0'}}\right) \qquad (2.4.15)$$

Therefore, we have demonstrated the Eq. (2.4.15) and (2.4.16) for the hybrid scheme is true and is consistent with the 3D Var Scheme that was previously discussed. The significance of this

relationship indicates that the background error covariance within the hybrid and 3D VAR scheme do not need to be inverted to solve the cost function.

$$\nabla_{\mathbf{z}} J = \boldsymbol{B}\nabla_{\boldsymbol{x}} J. \qquad (2.4.16)$$

Overall, this section demonstrated the implementation of the ensemble covariance with the static covariance in the existing GIS 3D VAR framework. It also introduces preconditioned conjugate gradient to avoid inverting the background covariance within the minimization of the cost function. As a result, this would greatly decrease the computational cost in both schemes. One can learn more about iteratively solving for the analysis from the preconditioned cost function by using the precondition conjugate gradient method from Derber and Rosati (1989).

2.5 Localization

It is essential to consider covariance localization to eliminate spurious correlation within the model state that does not have any physical relationship. This occurs as the distance between the grid points in model state increases substantially large. For example, one can assume that a surface analysis increment such as 10m wind speed at two distant grid point locations will have no correlation. Therefore, the increment of an individual wind speed observation is constrained in the horizontal and vertical direction. As previously stated, the 3D Hybrid correlation function for localization that is applied to the ensemble covariance is implicitly defined by $\boldsymbol{A}$. Since the control variables span in the horizontal and vertical direction, Wang et al. (2013) discussed that the covariance localization consists of a horizontal and vertical component, $\boldsymbol{A}_h$ and $\boldsymbol{A}_v$. In particular, the horizontal covariance localization is translated from a model grid space into the spectral space, using the transformation matrix $\boldsymbol{L}$. Contrarily, the horizontal covariance localization can be converted back from spectral space into the model grid space using an inverse matrix $\boldsymbol{L}^{-1}$. Therefore, the horizontal covariance localization can be defined by Eq. (2.5.1).

$$\boldsymbol{A}_h = \boldsymbol{L}^{-1}\boldsymbol{A}_{hs}\boldsymbol{L} \qquad (2.5.1)$$

Where $\boldsymbol{A}_{hs}$ represents the horizontal covariance localization in spectral space. On the other hand, the vertical covariance localization $\boldsymbol{A}_v$ is characterized by a recursive filter transformation explained by Hayden and Purser (1995). This transformation consists for successive approximation in linearly sequential passes among all the vertical levels.

The correlation function ρ in Eq. (2.5.2) that was adopted from Gaspari and Cohn (1999) provides the weighting within the horizontal and vertical localization. The variable z represents the distance between the grid points in the model grid space and the observation location, while $2c$ is defined as the cutoff distance at which the correlation diminishes to zero. In GSI, the cutoff distance $2c$ of vertical covariance localization can be set in vertical grid point model levels or natural logarithm of the corresponding pressure levels. Further detail is provided in Section 4.2. Also, the cutoff distance of the horizontal localization is in units of kilometers. According to Wang (2010), the rate of convergence in the cost function minimization is not dependent on the length of the localization scale. Figure 1 illustrates the horizontal correlation weighting function with respect to the zonal and meridional direction in units of kilometers.

$$\rho = \begin{cases} -\frac{1}{4}\left(\frac{|z|}{c}\right)^5 + \frac{1}{2}\left(\frac{|z|}{c}\right)^4 + \frac{5}{8}\left(\frac{|z|}{c}\right)^3 - \frac{5}{3}\left(\frac{|z|}{c}\right)^2 + 1, & 0 \le |z| \le c \\ \frac{1}{12}\left(\frac{|z|}{c}\right)^5 - \frac{1}{2}\left(\frac{|z|}{c}\right)^4 + \frac{5}{8}\left(\frac{|z|}{c}\right)^3 + \frac{5}{8}\left(\frac{|z|}{c}\right)^2 - 5\left(\frac{|z|}{c}\right) + 4 - \frac{2}{3}\left(\frac{|z|}{c}\right), & c \le |z| \le 2c \\ 0, & 2c \le |z| \end{cases} \quad (2.5.2)$$

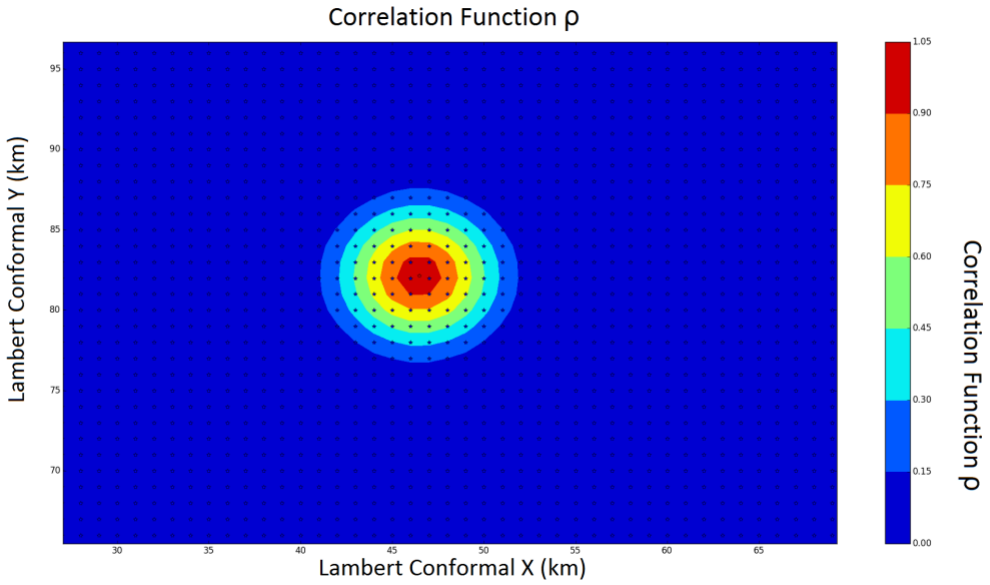


Figure 2.5.1*: The correlation function ρ used in the covariance localization, range from values of 0 to 1*

2.6 Enhanced Satellite Radiance Bias Correction in GSI

Bias correction for satellite radiance is essential to ensure the assimilated satellite observations are not deteriorating the analysis. As mentioned by Zhu et al. (2013) and Shao et al. (2016), radiance biases can originate from three sources, including the error caused from poor calibration of the satellite instruments, errors that are introduced from the radiative transfer model and errors from the background forecast model. The GSI enhanced radiance bias correction developed by Zhu et al. (2013) aims to correct biases from the first and second source of error and was used in this study. This variational scheme consists of an air-mass dependent component and a scan-angle component, which are implemented into the 3D Hybrid cost function minimization to update the predictor coefficients and compute the bias correction coefficients. Typically, this procedure is repeated iteratively at each analysis cycle for the bias correction coefficient to convergence into realistic values that can represent the radiance bias. Depending on the quality of the initialized bias correction coefficient, this process could take weeks to months of data assimilation cycles.

There are several advantages of using the Enhanced Radiance Bias Correction (ERBC) as opposed to the original Radiance Bias Correction (RBC) in GSI. Firstly, ERBC combines the two-component procedures into GSI, whereas the scan-angle component in RBC is computed using a package outside of GSI. Secondly, the modified preconditioning that is applied to bias correction coefficients in ERBC has a faster convergence rate of the minimization process. Lastly, ERBC is able to recognize new or missing radiance data and generate corresponding predictors on-the-fly. As a result, it removes the need to provide pre-initialized predictor values

that were originally required in RBC. As the RBC becomes obsolete in the current release packages in GSI, ERBC evolves to be a widely used radiance bias correction technique within the GSI community. For further detail on the mathematics framework on ERBC, refer to Zhu et al. (2013) and Shao et al. (2016).

3 Experiment Design

3.1 Data

The 3D Hybrid data assimilation ingests 3 data components, including model background, ensemble perturbation, and observations.

Firstly, the High-Resolution Rapid Refresh model is used as the model background, which has a 3km horizontal resolution with 50 native levels. Among the various models that are available, the horizontal resolution of HRRR is comparable with the 2.5 km horizontal resolution RTMA analysis. It offers hourly outputs of analysis and forecasts up to 18 hours. For this data assimilation study to be practically adopted in operational use, the HRRR's 1-hour forecast will be used to give a lead-time for the data assimilation to be completed reasonably close to the valid time. Due to the high volume of data, National Centers of Environmental Prediction (NCEP) stores the hourly HRRR model output in three separate grib2 files with different vertical level configurations. Namely, the model output is stored in native level (wrfnat), pressure level (wrfprs), and surface level (wrfsfc). These files are preprocessed to produce the initialized model background file that is compatible with GSI, which will be further explained in Section 3.3.

Secondly, the ensemble perturbations are retrieved from Global Ensemble Forecasting System (GEFS) in T574 grid configuration. It has a 33-35 km horizontal resolution with 64 pressure levels and 80 ensemble members, Zhou et al. (2017). The GEFS provides 6 hourly forecasts up to 8 days with the current grid configuration and an additional 8 days at different grid configuration with coarser horizontal resolution. The data are stored in Sigma files and can be downloaded from NCEP. Due to the limited available computing resource, only the 6-hour

ensemble forecast is used in the experiments. Since the HRRR model background and the GEFS ensemble data have different horizontal resolutions, GSI downscales the GEFS grid to be consistent with the HRRR's 3km horizontal resolution grid.

Part of this study aims to produce the optimal analysis increments by utilizing a large amount of high-quality observations. It is comprised of the surface, upper air and satellite observations. A majority of them are available from the Global Data Assimilation System (GDAS). The list of observations that can be assimilated in GSI is summarized in Table 3.1, where the observations that are used in the experiments are denoted with "*".These observations are stored in bufr file format and can be modified and manually updated with additional data by using utility packages within GSI. The observation bufr files are categorized based on the measurement instruments, such as conventional, satellite and radar datasets.

Observations Accepted in GSI	
Conventional Observation*	Precipitation rate observations from TMI
satellite winds observations*	SBUV/2 ozone observations from satellite NOAA-16, 17, 18, 19
AMSU-A 1b radiance (brightness temperatures) from satellites NOAA-15, 16, 17,18, 19 and METOP-A/B*	HIRS4 1b radiance observation from satellite NOAA-18, 19 and METOP-A/B
AMSU-B 1b radiance (brightness temperatures) from satellites NOAA-15, 16,17*	HIRS3 1b radiance observations from satellite NOAA-16, 17
Radar radial velocity Level 2.5 data	HIRS2 1b radiance from satellite NOAA-14
Precipitation rate observations from SSM/I	MSU observation from satellite NOAA 14
Microwave Humidity Sounder observation from NOAA-18, 19 and METOP-A/B*	GOES sounder radiance (sndrd1, sndrd2, sndrd3 sndrd4) from GOES-11, 12, 13, 14, 15.
AMSU-A and AIRS radiances from satellite AQUA	SSMI observation from satellite f13, f14, f15
SSMIS radiances from satellite f16	NEXRAD Level 2 radial velocity*
GOES sounder radiance from GOES-11, 12	GOES imager radiance from GOE-11, 12
Ozone Monitoring Instrument (OMI) observation NASA Aura	Infrared Atmospheric Sounding Interfero-meter sounder observations from METOP-A/B
The Global Ozone Monitoring Experiment (GOME) ozone observation from METOP-A/B	Aura MLS stratospheric ozone data from Aura

Table 3.2.1: *Types of observation that can be assimilated in GSI. The observations are that assimilated in the experiments are defined by an asterisk symbol "*".*

The conventional observation bufr dataset consists of temperature, moisture, pressure, and/or wind speed/direction measurements from surface METAR stations, radiosonde, aircraft,

ships, buoys. The conventional dataset also includes Vertical Azimuth Display (VAD) from NEXRAD radars. In order to illustrate the immense amount of observation data that are available, Table 3.2 shows the number of measurements for each conventional observation type that is assimilated to produce the analysis on May 5, 2018 at 00z.

Conventional Observation Types	Number of Observations
Radiosonde – Temperature, Moisture, Pressure, Wind components (u,v)	78
Aircraft – Temperature	15
Surface Marine (Ships, Buoys) - Temperature, Moisture, Pressure	2163
Surface Land METARS - Temperature, Moisture, Pressure	6964
NEXRAD Vertical Azimuth Display – Wind Components (u,v,z)	140
Aircraft – Wind components (u,v)	17
Surface Marine (Ships, Buoys) - Wind components (u,v)	2132
Surface Land METARS - Wind components (u,v)	6824

Table 3.1.1: *Number of measurement for each conventional observation type that are assimilated to produce the analysis on May 5, 2018 at 00z. Note that radiosonde data are generally available at 00z and 12z. The values shown the table reflect the numbers of observation* that passed quality control in GSI

Satellite Data offers a large amount of information that can be used in data assimilation to initialize numerical weather predictor (NWP) models. Unlike conventional observations, observations from a single satellite instrument can cover vast portions of the NWP domain. In particular, satellite data can potentially provide observational coverage in regions of scarce conventional observations, such as mountainous areas and over large bodies of water. For this

reason, numerous operational NWP system initialize their model by assimilating satellite observation, such as European Centre for Medium-Range Weather Forecasts (ECMWF), Global Forecasting System (GFS), North American Mesoscale Forecast System (NAM) and Rapid Refresh Model. Although this study focuses on using data assimilation to provide a 3D analysis rather than on initializing NWP models, the analysis can benefit from assimilating satellite data. The satellite data that are used in this study are summarized in Table 3.3, in which the channels that are used for each instrument and satellite vehicle are consistent with the configuration of the NCEP's NAM, GFS and RAP model to ensure the measurement used in the data assimilation are from well-collaborated instruments.

Satellite Instrument	On-Board Satellite Vehicle	Numbers of Channels	Orbit	Description
Microwave Humidity Sounder (MHS)	NOAA-18 NOAA-19 MetOp-A MetOp-B	5	Polar	– Provide vertical profile of temperature and humidity. – Channel 1 retrieves surface water vapour and temperature and Channel 2 detects only surface water vapour. – Channel 3-5 retrieve water vapour in the upper atmosphere.
Advanced Microwave Sounding Unit-A (AMSU-A)	NOAA-15 NOAA-18 NOOA-19 MetOp-A MetOp-B	15	Polar	– Measures outgoing radiances from the Earth surface to 3 hPa of the atmosphere. – Resolution at Nadir: 48 km – Channel 1 – 4: Retrieves Water Vapours – Channel 5 – 8: Retrieves Tropospheric Temperature – Channel 9 -14: Retrieves Stratospheric Temperature – Channel 15: Retrieves Cloud Top
High-Resolution Infrared Raditation Sounder 4 (HIRS4)	NOAA-19 MetOp-A MetOp-B	20	Polar	– Retrieves vertical profiles of Temperature and humidity from measured radiances – Resolution at Nadir: 10 km

Table 3.1.2*: A summary of the radiance satellite instruments and its measured data that are assimilated in this study. John et al. (2012), Karbou et al. (2005). The MHS, AMSU-A and HIRS4 instruments are employed on various polar-orbiting satellite vehicles. The* channel and resolution vary for each the instrument.

3.2 Configuration

Several configurations for the analysis produced by the 3D Hybrid data assimilation in GSI are considered. The analysis has a CONUS domain shown in Figure 3.2.1with 50 native vertical levels, covering over the southern 49 states of the U.S.A, the southern part of Canada and northern regions of Mexico. It also has 3km horizontal resolution and 50 native vertical levels. Since the GEFS ensemble forecast is only available every 6 hours, an analysis is also produced in 6 hours intervals (00z, 06z, 12z and 18z). Furthermore, the analysis includes updated variables such as temperature, wind speed, specific humidity and pressure.

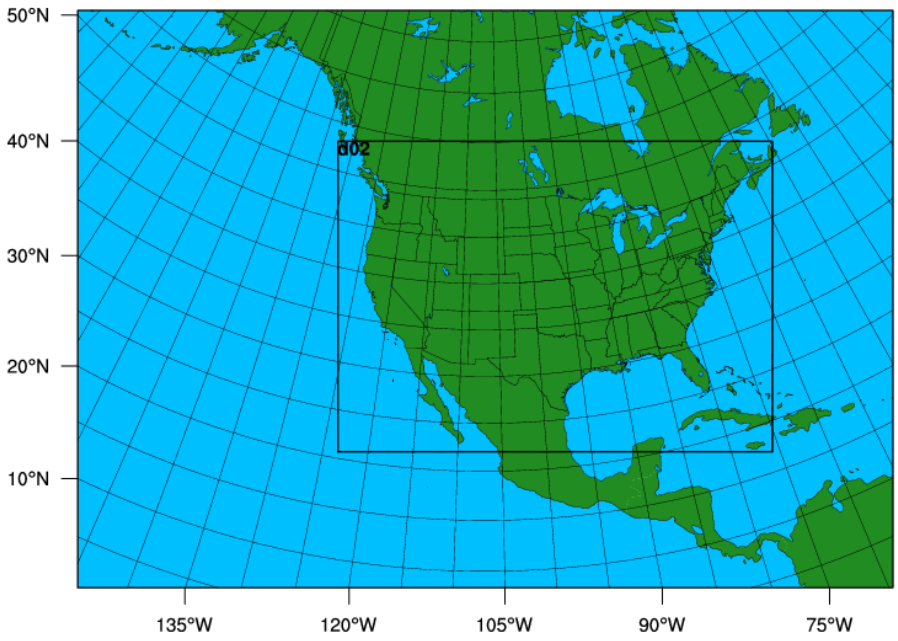


***Figure 3.2.1**: Illustration of the CONUS domain*

GSI has numerous parameters in the namelist to allow users to conveniently make the necessary changes that cater to their research objectives without modifying the source code. For instance, the hybrid data assimilation in GSI gives the option to adjust the weighting total

background covariance between the static and ensemble background error covariance, as explained in Section 2.3. For all the experiments in this study, the weightings for both schemes are balanced to be equal, by setting $\frac{1}{\beta_1}$ and $\frac{1}{\beta_2}$ to 0.5. One must enable GSI to update the 2m potential temperature and moisture increments on the analysis by modifying anavinfo file. The vertical profile observation errors of Temperature, uv wind component and relativity humidity are shown in Figure 3.2.2 (a) – (c), respectively. Similarly, the surface observation errors shown in Table 3.4 from the RTMA configuration in GSI are used in this study to ensure a fair comparison between the RTMA and hybrid analysis. It is important to note that the set of observation errors used in this study are smaller than the set used to initialize numerical weather models. This allows the analysis to be fitted closer to the observations, thus increasing its accuracy. Lastly, the total number of minimization iteration is set to 20 for all experiments. Refer to Table 3.6 for a summary of all experiments with its respective configuration for this study.

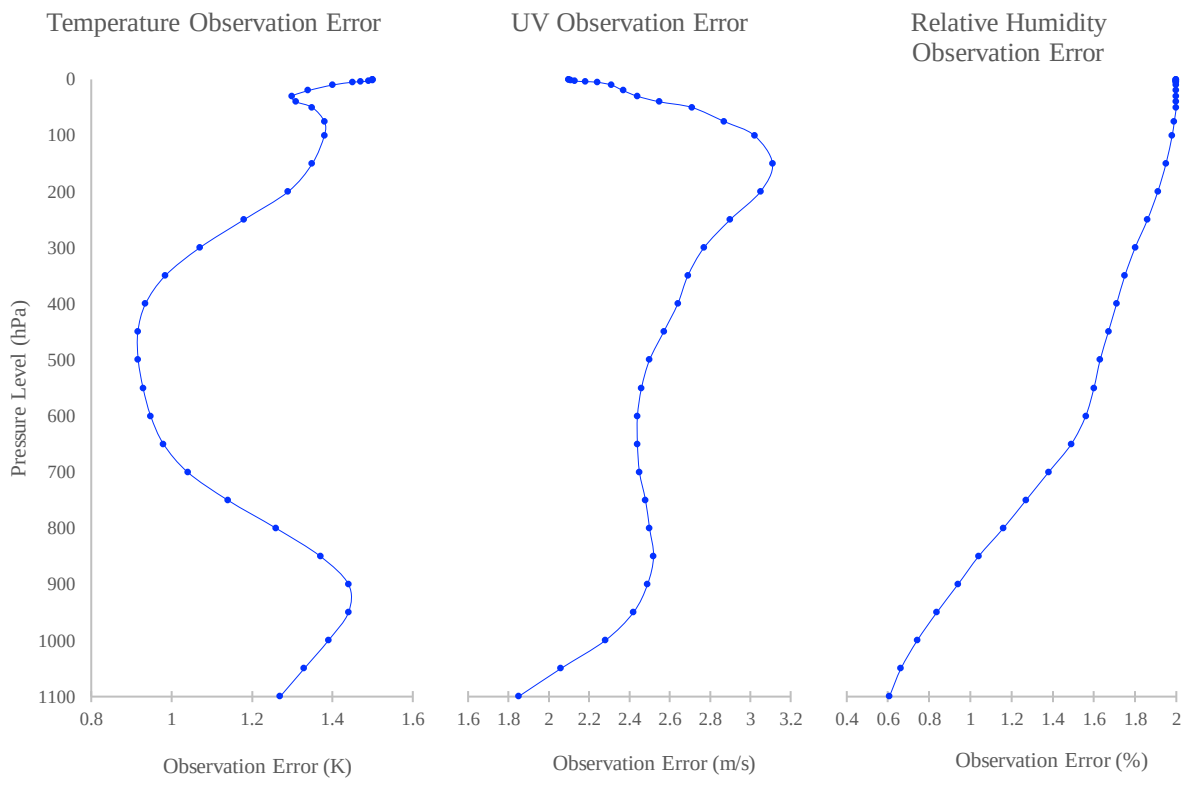


Figure 3.2.2*: Vertical profile of (a) temperature, (b) uv wind component and (c) relative humidity observation error used to characterize the observation error covariance* ***R****.*

Surface Observation Errors

Observation Type	Temperature (K)	Relative Humidity (%)	UV Wind Component (m/s)	Pressure (mb)
Surface Marine – Obs Type: 180	1	5.912		0.538
Surface METAR – Obs Type: 181	1	5.912		0.538
Surface Marine – Obs Type: 183	1.2	5.912		
Surface METAR – Obs Type: 187	1	5.912		0.538
Surface Marine – Obs Type: 280			2.628	
Surface METAR – Obs Type: 281			1.587	
Surface Marine – Obs Type: 284			1.079	
Surface METAR – Obs Type: 287			1.587	

Table 3.2.1*: Surface observation error, including(a) temperature, (b) uv wind component and (c) relative humidity used in characterizing the observation error covariance* ***R****.*

This study focuses on examining the results of the 3D Hybrid analysis with the effects of changing vertical localization and assimilating satellite radiances. The control run experiment was conducted for a period between May 5, 2018 at 00z to May 18, 2018 at 00z. It has a vertical and horizontal localization of 3 grid units and 100 km. Since the goal for this study is to produce analysis that captured localized features such as contrasting terrain and flow-dependent characteristics, the control run experiments (Hyb_ctrl) have relatively low vertical and horizontal localization of 3 grid units and 100 km, respectively. One can hypothesize that the low localizations can constrain the physically correlated features within the localized domain, as discussed in Section 2.5. Next, three of experiments with varying vertical localization were conducted to study its effects on surface and upper-level analysis increments. The experiments ran for a period between May 9, 2018 at 00z and May 13, 2018 at 00z, with the vertical localization set to 6, 9 and 12 grid units, which are abbreviated by Hyb_v6_h100, Hyb_v9_h100, Hyb_v12_h100, respectively. Finally, a separate part of the study focuses on the potential benefits of assimilating satellite radiances for a period between May 5, 2018 at 00z to May 18, 2018 at 00z. This experiment is denoted by Hyb_sat and uses a vertical and horizontal localization of 3 grid units and 100 km, respectively.

3D Hybrid - No Satellite Data are assimilated

Experiment Run	$\frac{1}{\beta_1}$	$\frac{1}{\beta_2}$	Total number of Minimization Iterations	Vertical Localization (Grid points)	Horizontal Localization (km)	Start Date	End Date
Hyb_ctrl	0.5	0.5	20	3	100	May 5, 2018 at 00z	May 18, 2018 at 00z
Hyb_v6_h100				6	100	May 9, 2018 at 00z	May 13, 2018 at 00z
Hyb_v9_h100				9	100		
Hyb_v12_h100				12	100		

3D Hybrid - Satellite Data are assimilated

Experiment Run	$\frac{1}{\beta_1}$	$\frac{1}{\beta_2}$	Total number of Minimization Iterations	Vertical Localization (Grid points)	Horizontal Localization (km)	Start Date	End Date
Hyb_sat	0.5	0.5	20	3	100	May 5, 2018 at 00z	May 18, 2018 at 00z

Table 3.2.2: *Summary of all experiments with its corresponding configurations. The main components of this study examined the potential benefit of assimilating surface and upper level observations on surface and upper level analysis. Also, investigate the effect of vertical and horizontal localizations on the analysis. Lastly, an experiment was done to study the benefits of assimilating satellite radiances. Overall, the Total number of Minimization Iterations, Vertical & Horizontal Localization,* $\frac{1}{\beta_1}$, $\frac{1}{\beta_2}$ *and experiential periods for each experiment are shown in this table.*

3.3 Background Preprocessing

Typically, GSI can ingest the background model output that is taken from the previous data assimilation cycle or directly from external model outputs. Since our experiments ingest from the external HRRR model output, a series of preprocessing procedures are necessary to ensure that the external model file format and data can be accepted in GSI. This process consists of utilizing Weather Research Forecast model (WRF) and WRF Preprocessing System (WPS) to convert the HRRR model Grib2 files into ARW (Advanced Research WRF) NETCDF files. As the initial step, WPS program is used to ungrib the wrfnat, wrfprs, wrfsfc HRRR Grib2 files and temporarily store the dataset into intermediate files. Next, the model domain and the terrestrial data are interpolated and linked to the intermediate files to generate WPS NETCDF file. Finally, WPS NETCDF file is ingested into WRF initialization code real.exe to convert into ARW NETCDF files.

3.4 Post Processing

The analysis produced by GSI does not explicitly provide some surface variables such as 2m temperature and 10m wind speed and direction. Therefore, post-processing is needed to compute the two surface variables from the state variables within GSI. Specifically, GSI has the option to update 2m potential temperature analysis θ and surface pressure analysis P. Consequently, the 2m temperature can be calculated by using the Poisson's Equation in Eq. 3.4.1.

$$T = \theta \left(\frac{P}{P_0}\right)^{\frac{R}{Cp}} \qquad (3.4.1)$$

Assume the reference pressure P_0 is 1000.00 hPa, the gas constant for dry air R is $287.04\ J * K^{-1} * kg^{-1}$ and the specific heat capacity Cp is $1004.67 * K^{-1} * kg^{-1}$.

As previously mentioned, GSI does not update the 10m wind speed and direction analysis. However, it does provide an updated 3D U and V wind component analysis fields. One can adopt a similar surface layer parameterization scheme from the WRF to compute the 10m winds. In order to be consistent with the WRF configuration for the operational HRRR model, Mellor–Yamada–Nakanishi–Niino (MYNN) scheme was considered. Based on the height of the first and second native level of the HRRR model, the MYNN scheme parametrizes the 10m U and V components to be equivalent to the first native level of the 3D U and V wind component analysis fields.

4 Results and Discussions

4.1 3D Hybrid Analysis (Control Run)

The 3D Hybrid Analysis Control Run experiment, denoted by Hyb_ctrl on Table 3.5, focuses on examining the benefits of the 3D Hybrid Data Assimilation while comparing the results from 3D VAR and RTMA as a benchmark. These comparisons highlight the performance of the 3D Hybrid, 3D VAR and RTMA using statistical and spatial comparisons. Specifically, the surface and the upper level analysis will be explored in detail in Section 4.1.1 and 4.1.2.

4.1.1 Surface Analysis Result

4.1.1.1 Case Study: Low Pressure Center off Lake Michigan

As several weather systems pass through the analysis domain throughout the study, it gives the opportunity to spatially examine the performance of the 3D Hybrid surface analysis. In particular, the flow-dependent error covariance employed in the 3D Hybrid scheme should produce a more accurate surface analysis in regions of the weather compared to the analysis produced by using solely the variational scheme or the background HRRR model analysis. In the first case study, a Texas Low Pressure Center converged with a Colorado Low Pressure Center over Wisconsin and The Great Lakes on May 9th, 2018 at 18z. The instability of the weather system is enhanced by frontal boundaries within the region, as shown by NOAA's surface analysis and radar image in Figure 4.1.1. From the 3D Hybrid analysis, the 2m temperature and 10m wind speed posterior (analysis) for the same time is portrayed in the contour in Figure 4.1.2a and 4.1.2b, respectively. The color scale of the contour is represented by the left colorbar, where the values are in units of Kelvins and m/s. The scatterplot depicts the analysis errors that were verified against the measurement of each observation station. The values of the analysis error are illustrated by the right colorbar, in units of Kelvin and m/s. Light color shading in the scatterplot suggests low errors. As the color shading deviates towards red or blue, it indicates the verification of the analysis has a strong positive/negative bias. In addition, the blue and red lines represent cold and warm fronts, while "H" and "L" symbolize high and low-pressure centers. From the 2m temperature contours, there is cold air mass situating over Wisconsin and Minnesota. On the other hand, a warm air mass was positioned over eastern parts of Illinois and Missouri. The temperature gradient that is dividing the warm and cold air mass can be seen by the color contour and warm/cold frontal boundaries in Figure 4.1.2a. As a result, the 2m

temperature analysis produced by the 3D Hybrid scheme is able to locate advancing air masses and can potentially depict certain features of weather systems. From the 10m wind speed contour, regions of relatively high wind are located off North and South Dakota and Nebraska, before channeling eastwards around the low-pressure center along the southern tip of Lake Michigan. Although the wind direction is not shown, this feature signifies one of the characteristics of a typical counter-clockwise rotation flow around a low-pressure center at the near surface in the Northern Hemisphere.

One can compute the increment of the 3D Hybrid analysis to examine the impact of using hybrid data assimilation in comparison with the background HRRR model. In other words, the increment is equivalent to the difference between posterior (analysis) and the prior (background model). The contour in Figure 4.1.3a and 4.1.3b displays the increment of 2m temperature and 10m wind speed, where the values are represented by the left colorbar. The scatterplot illustrates the error improvement between the posterior and prior, verified at each observation station. This can be determined by taking the difference between absolute error of the posterior and absolute error of the prior, as shown in Eq. (4.1.1). If the error improvement at an observation station equates to a negative value, it indicates that the absolute error decreased after data assimilation and had a positive impact on the accuracy of the updated analysis. On the other hand, a positive value suggests that the absolute error increases after data assimilation, hence have a negative impact on the accuracy of the updated analysis.

$$Error\ Improvement = abs\left(Error_{posterior}\right) - abs\left(Error_{prior}\right) \quad (4.1.1)$$

By comparing the radar image in Figure 4.1.1 with the 2m temperature increment contour in Figure 4.1.3a, one can recognize the hybrid data assimilation decreased the surface temperature by 2 – 3 K along the southern portion of the warm front situated over Illinois and Indiana. However, it increased the temperature by 1 – 3 K along the northern portion of the warm front, just east of Lake Michigan. Temperature adjustment also occurred for the regions with precipitation covering Minnesota and the low-pressure center just over Chicago. In addition, the blue shading of the scatterplots within these regions suggest the absolute error decreased by 1 – 3 K after the data assimilation. High increments with a decrease in absolute error can also be found along the cold front spanning from the Rockies in British Colombia through US-Canadian board and western Ontario. A similar comparison for 10m wind speed shows the posterior has decreased the wind speed by 1 – 3 m/s from the prior just over the low-pressure center and in the vicinity with precipitation. The wind increments are relatively smaller along the cold front in BC through western Ontario. However, a higher increment of magnitude 2 – 4 m/s was seen along the cold front in BC as it advanced southeastwards in the 00z and 06z analysis for May 10th, 2018. In all cases, the data assimilation decreased the absolute error by 1 – 4 m/s within the region of the weather system. However, there are regions with a significant increment that did not experience any weather systems throughout the study, such as the state of Colorado in the Rocky Mountains and western Virginia in the Appalachian Mountains. Since NWP such as the HRRR model lack the ability to resolve the surface variables in regions of high – contrasting terrain, data assimilation provides more accurate analysis for those regions by adjusting the surface analysis closer to the measurement from the nearby observations. Also, the flow-dependent error covariance from hybrid data assimilation is able to spatially characterize the analysis increment over the mountainous areas, based on the flow of the day.

In order to further demonstrate the effect of flow-dependent background error covariance from a hybrid data assimilation scheme, one can compare the analysis produced using the hybrid scheme against the analysis produced using the variational scheme. The contours in Figure 4.1.4a and 4.1.4b display the difference for 2m temperature and 10m wind speed posterior between the hybrid scheme and variational scheme. The values in the contour are represented by the left colorbar. The scatterplot depicts the absolute error difference between the posterior from the hybrid and variational scheme, shown in Eq. 4.1.2. If the scatterplot indicates a negative value by the blue color shading, it implies that the absolute error for the posterior from the hybrid scheme is less than the variational scheme. On the other hand, positive values shown in red color shading, suggesting that the absolute error for the hybrid scheme is greater than the variational scheme.

$$Error\ Difference = abs\left(Error_{Hybrid}\right) - abs(Error_{VAR}) \qquad (4.1.2)$$

In Figure 4.1.4a, there are considerable differences in the 2m temperature analysis produced between the hybrid and variational scheme, which can be found along the warm frontal boundaries south of Lake Michigan and the cold front situated near the western part of the US-Canadian border. Notable 1 – 3 K differences are located in areas of precipitation covering Lake Michigan, Lake Superior, Wisconsin, Minnesota and along British Colombia and Alberta. Figure 4.1.4b shows similar results, where there are respectable differences in 10m wind speed analysis between the hybrid and variational schemes within the weather system at southern tip Lake Michigan and along the cold front near the U.S-Canada Border. The magnitude of the difference

for these regions varies from 0.5 to 1.5 m/s. Consequently, values in the 2m temperature and 10m wind speed Hybrid-VAR differences for these regions have a strong correlation with the regions of high increment in Figure 4.1.3a and 4.1.3b. In addition, the generally blue shading of scatterplot within these regions in Figure 4.1.4a and 4.1.4b shows that the analysis produced using the hybrid scheme has a lower absolute error than the variational scheme. These findings strengthen the notion that the flow-dependent background error covariance in the hybrid scheme is able to capture the passage of the weather system and produce a more accurate 2m temperature and 10m wind speed analysis.

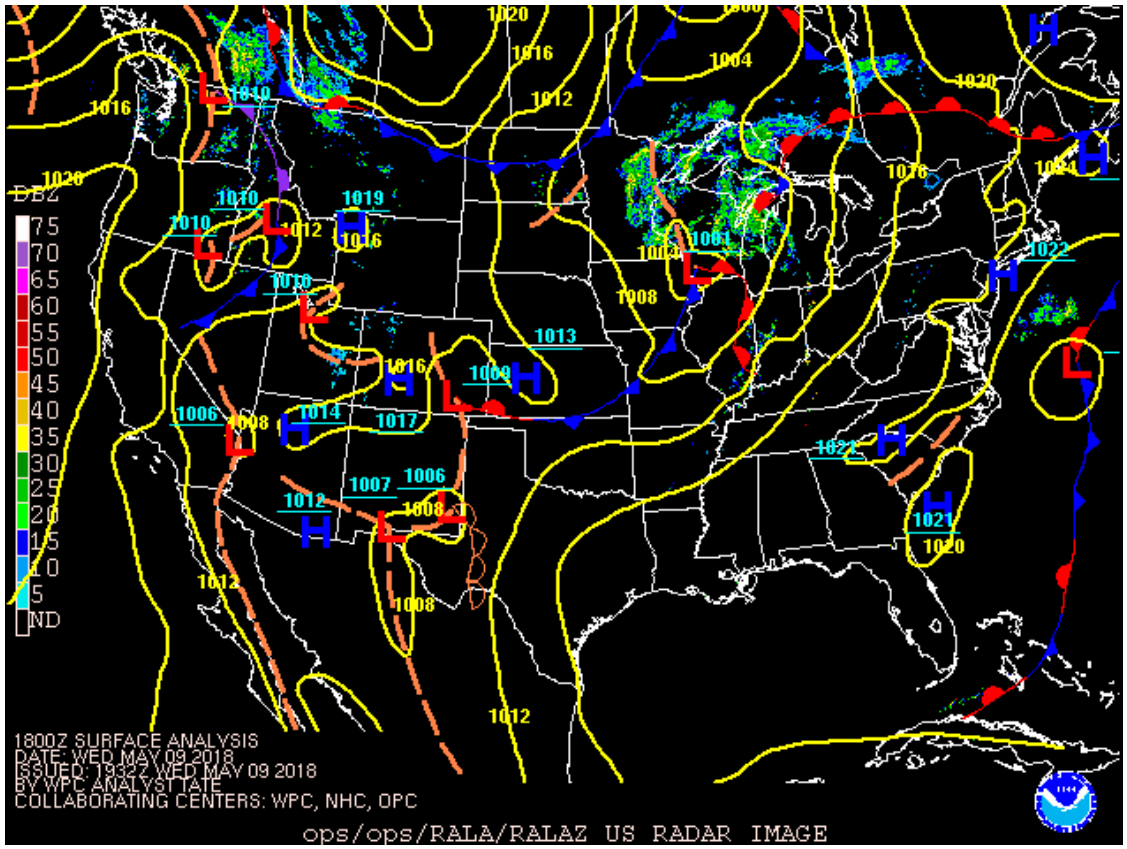


Figure 4.1.1*: Surface analysis with overlaying composite radar image on May 9th, 2018 at 18z. This figure was produced by NOAA.*

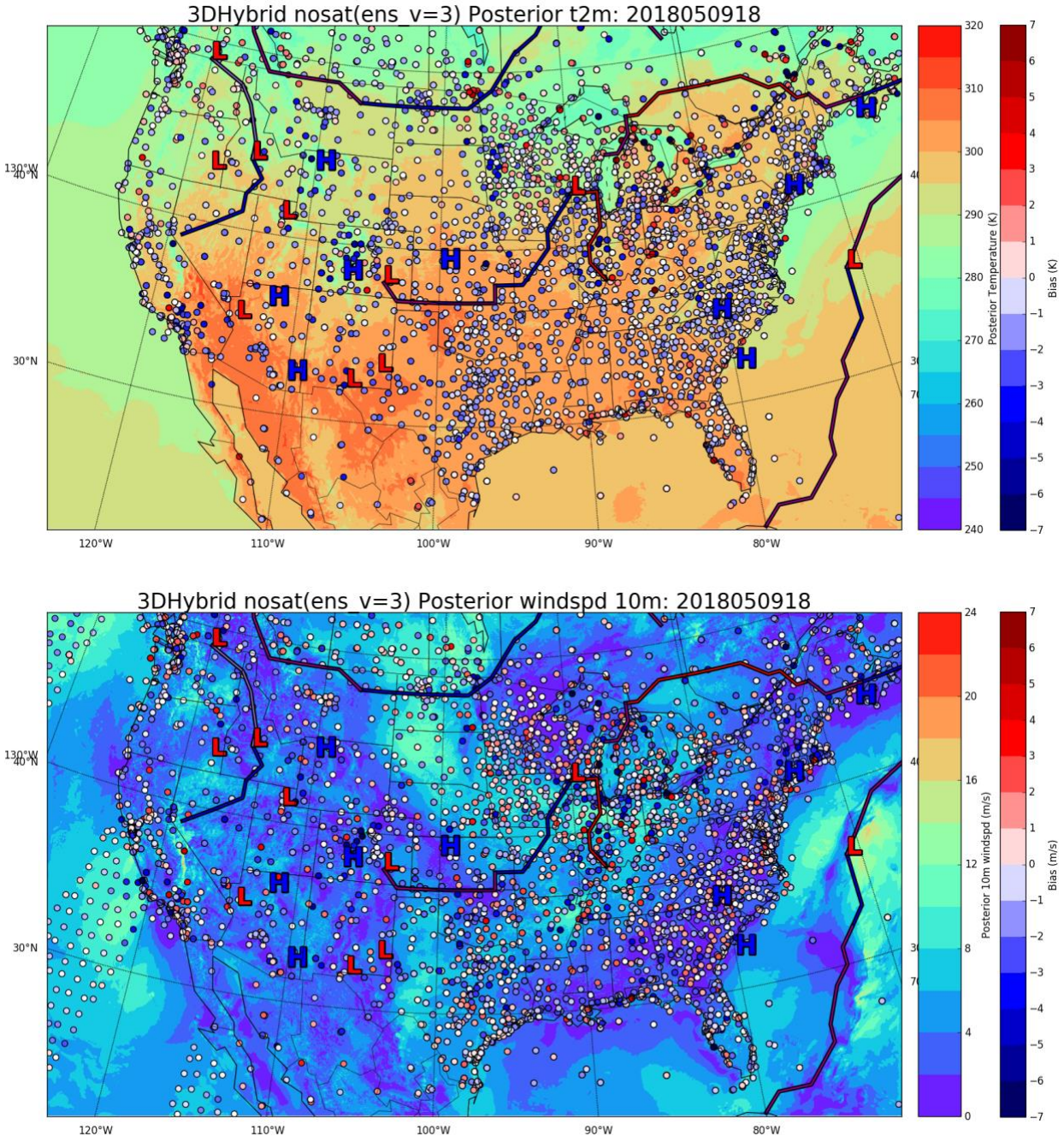


Figure 4.1.2: *The 2m temperature (a) and 10m wind speed (b) posterior for May 9th, 2018 at 18z, represented by the contours. The shading of the scatterplot depicts the analysis error at an individual observation station. The High and Low pressure center is represented by the H (blue) and L (red) symbols. The cold, warm and occluding fronts are portrayed by the blue, red and purple lines, respectively.*

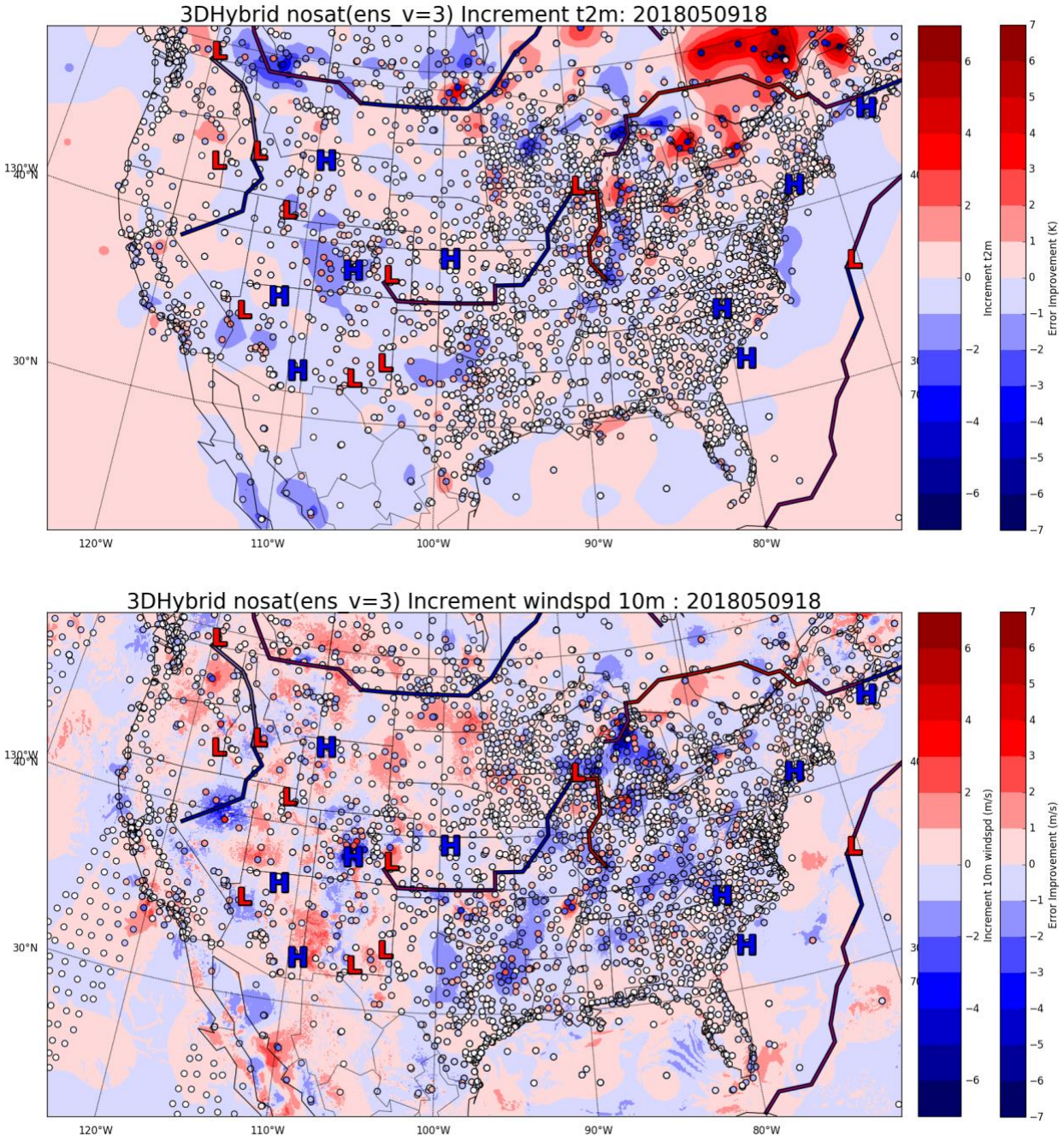


Figure 4.1.3*: The difference between the posterior and prior for 2m temperature (a) and 10m wind speed (b) on May 9th, 2018 at 18z, represented by the contour. The scatterplot displays the error improvement after assimilation (the difference between the absolute error of the posterior and the prior at the individual observation stations). Positive impact on the 3D Hybrid is denoted by negative (blue) error improvement values. Whereas, the negative impact is denoted by positive (red) improvement values. The High and Low-pressure center is represented by the H (blue) and L (red) symbols. The cold, warm and occluding fronts are portrayed by the blue, red and purple lines, respectively.*

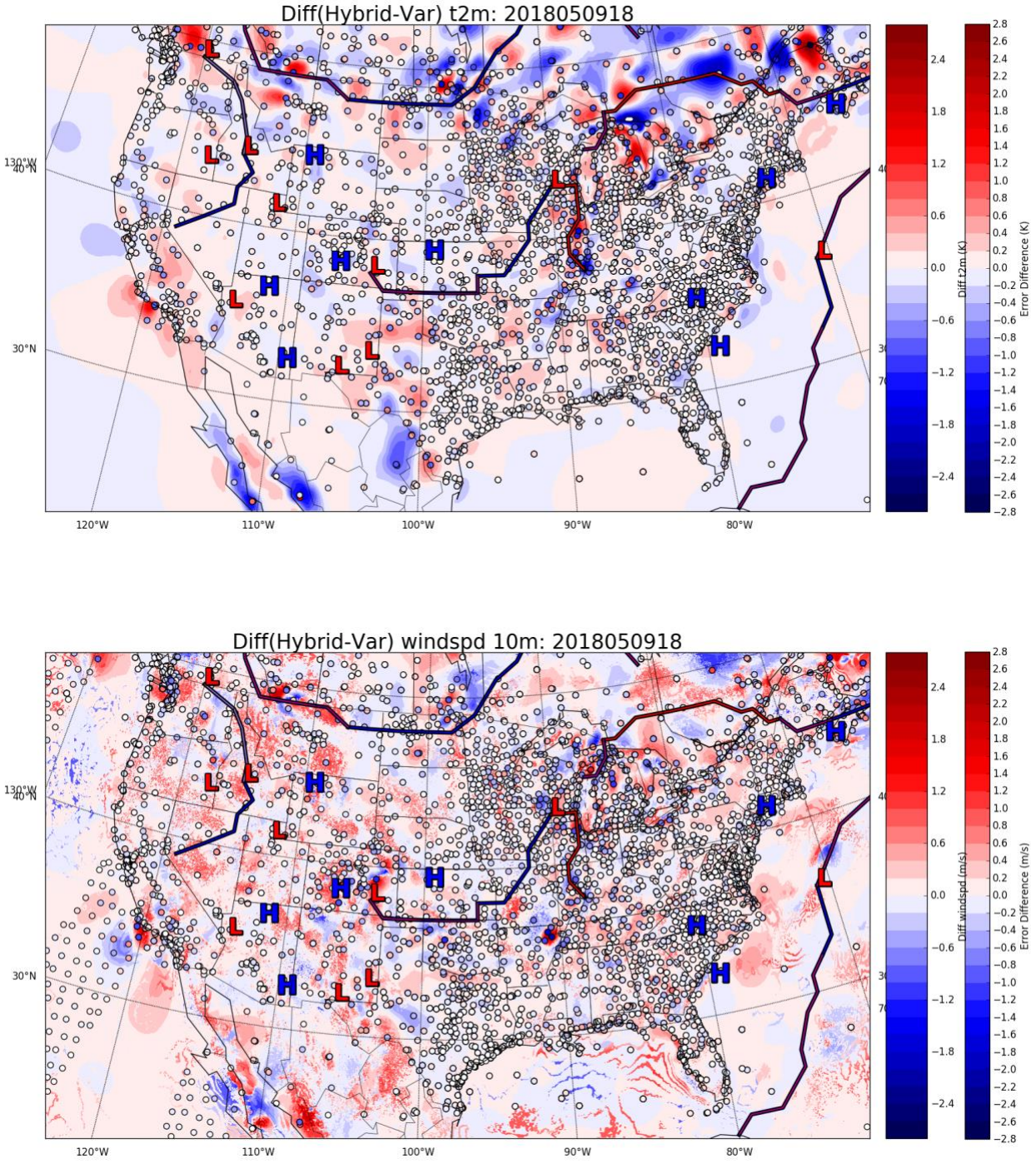


Figure 4.1.4*: The contour represents the difference between the hybrid and variational data assimilation analysis for 2m temperature (a) and 10m wind speed (b) on May 9th, 2017 at 18z.. The scatterplot depicts the absolute error difference between the hybrid and variational scheme, verified at the individual observation stations. Negative (Positive) values indicate the absolute error from the hybrid analysis is smaller (greater) than the variational analysis. The High and Low-pressure center is represented by the H (blue) and L (red) symbols. The cold, warm and occluding fronts are portrayed by the blue, red and purple lines, respectively.*

4.1.1.2 Case Study: Stationary Front Across US

In the second case study, a stationary front span crossed the U.S from Arizona through Missouri and the Northern Atlantic Coast of US on May 12, 2018 at 12z, as shown by the NOAA's surface analysis – radar image in Figure 4.1.5. In the recent days, pools of moisture from the Pacific were channeled across the country to the Atlantic coast by the stationary front. Consequently, developed precipitation was seen along the northern side of the front. The 2m temperature posterior in Figure 4.1.6a demonstrated the distinct temperature gradient that aligns with the position of the stationary front in Figure 4.1.5. Specifically, the temperature differences were as large as ~15℃ within the boundary that separated the cold and warm air masses. The scatterplot in the figure indicates that the posterior generally underestimates temperature compared with the observations.

Similar to the comparison in the first case study, the increment for 2m temperature and 10m wind speed for the second case study are shown in Figure 4.1.7a and 4.1.7b. The 2m temperature increment shows the hybrid data assimilation adjusted the temperature analysis by a magnitude of 2 – 5 ℃ in numerous regions along the stationary frontal boundary and the low-pressure with precipitation system in Utah – Colorado and Ohio – Pennsylvania. The large values in the increment within these regions suggests that the hybrid scheme is able to spatially adjust the position of frontal boundaries based on the assimilated observations and flow of the day information from the ensemble members. Similarly, the 10m wind speed increment reveals distinguishable increment of 1 – 4 m/s within low pressure center that is over Utah – Colorado. There are also other regions with a noticeable increment that did not experience any significant weather systems, which occurs quite frequently during the study. This can be attributed by

assimilating the surface wind observations using the hybrid scheme to adjust the wind analysis to capture highly localized surface wind features that are dependent on the surround terrains and frictional properties. In other words, the flow-dependent error covariance is able to characterize the surface wind flow over varying terrain based on the flow of the day, as previously discussed.

The predominantly blue shading in the scatterplot in Figure 4.1.7a and 4.1.7b indicates that the absolute error for 2m temperature and 10m wind speed have improved by 2 – 5 and 1 – 4 m/s within specific regions along the stationary front. The areas with an absence of weather systems have smaller absolute error improvement. There are also isolated points where the observation stations have verified that the absolute error has increased after the data assimilation. In particular, the number points with an increased absolute error for surface wind in Figure 4.1.7b and Figure 4.1.3b are greater than the amount for surface temperature in Figure 4.1.7a and Figure 4.1.3a. This can be attributed by the wide time window to accept observation in the data assimilation experiment. For instance, the most recent surface observations that are measured within -1.5 to +1.5 hours from the analysis time stamp are assimilated in this study. Although surface temperature fluctuation is typically minimal within this timescale, surface wind can change drastically depending on the local terrain and local-scale wind flow. Therefore, one can improve the analysis by determining a balance between the appropriate observation time window and assimilate the maximum amount of high-quality observations. For example, the time window of $\pm$12 minutes from the analysis time stamp in RTMA allows to assimilate measurements from ~14000 observation stations, according to De Pondeca et al. (2011).

The difference in 2m temperature analysis between the hybrid and variational scheme can be seen in Figure 4.1.8a, while a similar comparison is done for 10m wind speed shown in Figure 4.1.8b. The largest magnitude in temperature and wind speed differences are positioned along the stationary front. For instance, the Hybrid-VAR differences for the 2m temperature and wind speed analysis within low pressure system over Colorado – Utah varies from -0.6 to 1.5 ℃ and -1 to 1 m/s. Higher values for 2m temperature can be seen along a portion of the front, spanning from Northern Texas to Kansas and Northern Missouri. Within these regions, the primarily blue colour shading of the scatterplot in Figure 4.18a and 4.18b indicates that analysis absolute errors from the hybrid scheme are less than analysis absolute errors from the variational scheme. These findings demonstrated that hybrid data assimilation has a better ability to characterize the spatial position of the weather system such as a stationary front than the variational data assimilation.

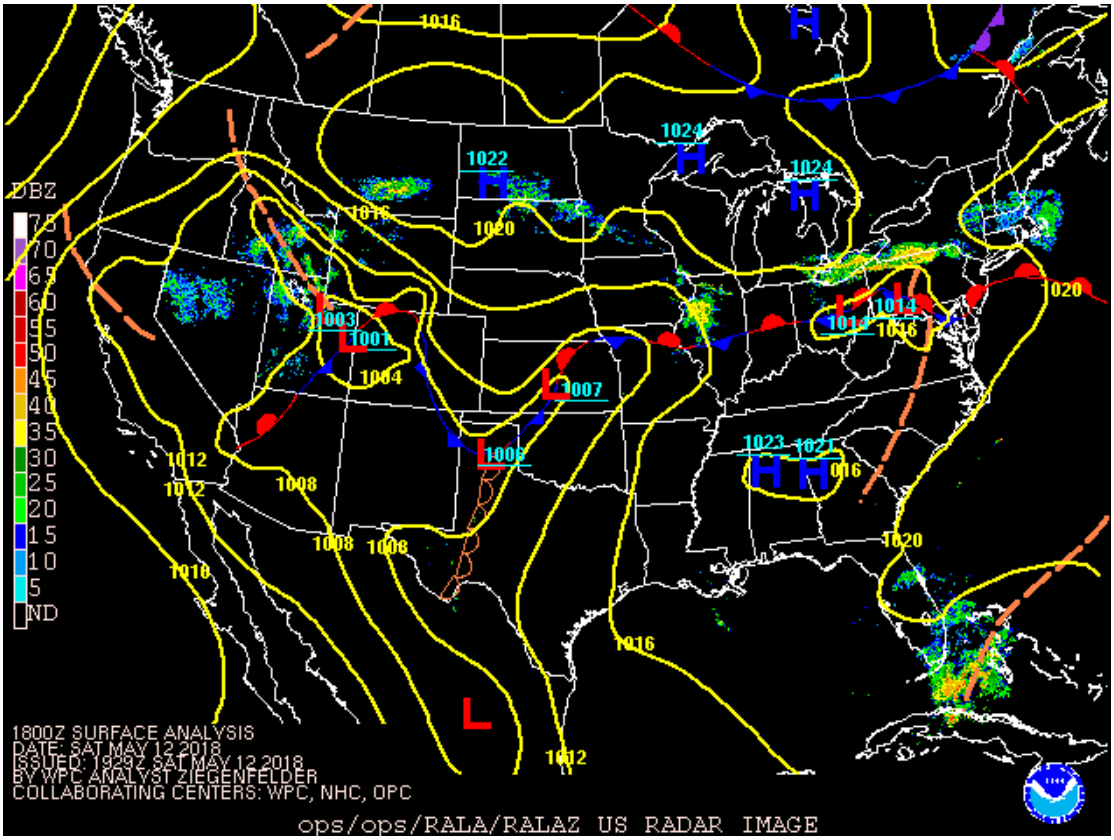


Figure 4.1.5: *Surface analysis with overlaying composite radar image on May 12th, 2018 at 18z. This figure was produced by NOAA.*

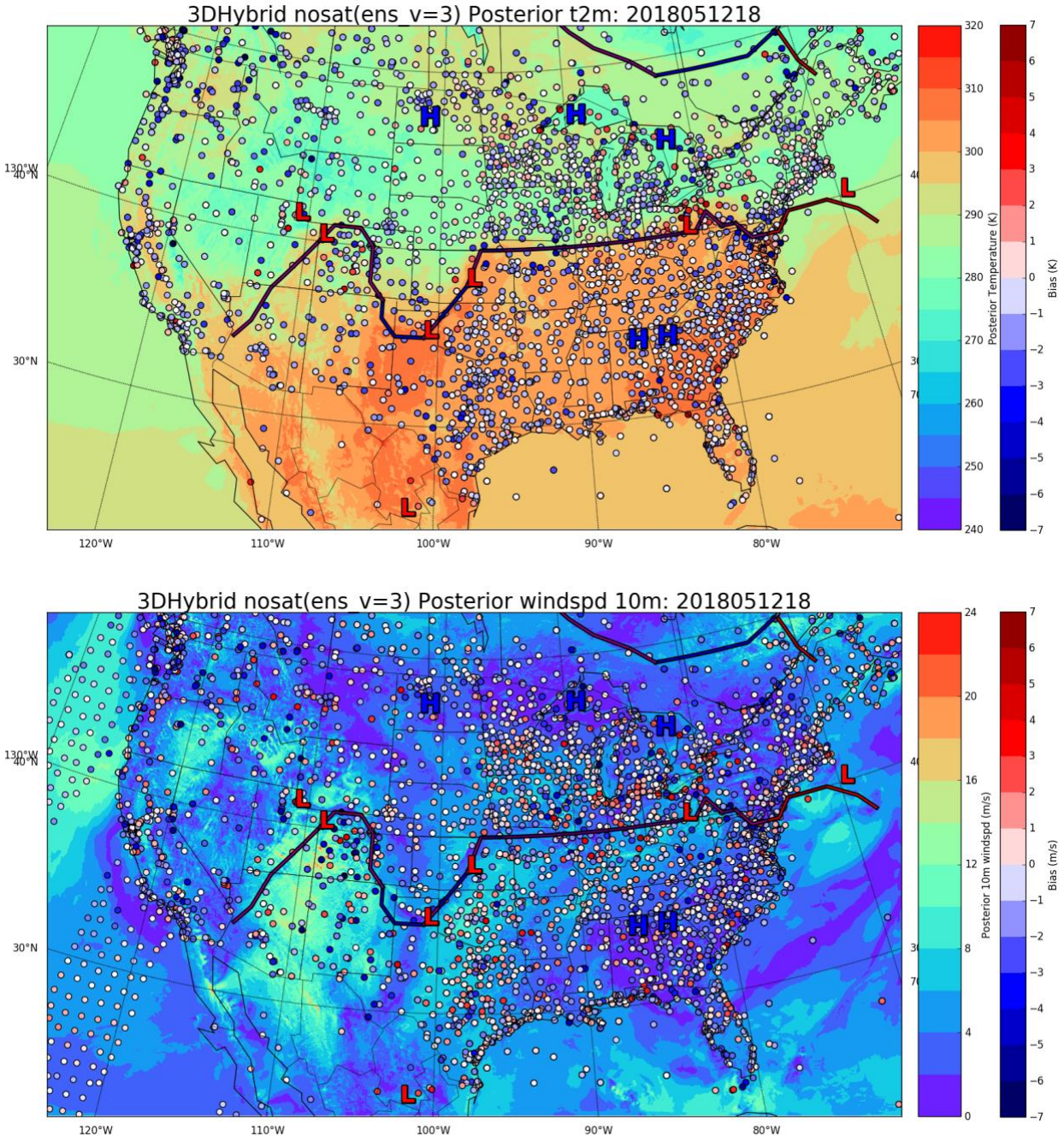


Figure 4.1.6: *The 2m temperature (a) and 10m wind speed (b) posterior for May 12th, 2018 at 18z, represented by the contours. The shading of the scatterplot depicts the analysis error at an individual observation station. The High and Low pressure center is represented by the H (blue) and L (red) symbols. The cold, warm and occluding fronts are portrayed by the blue, red and purple lines, respectively.*

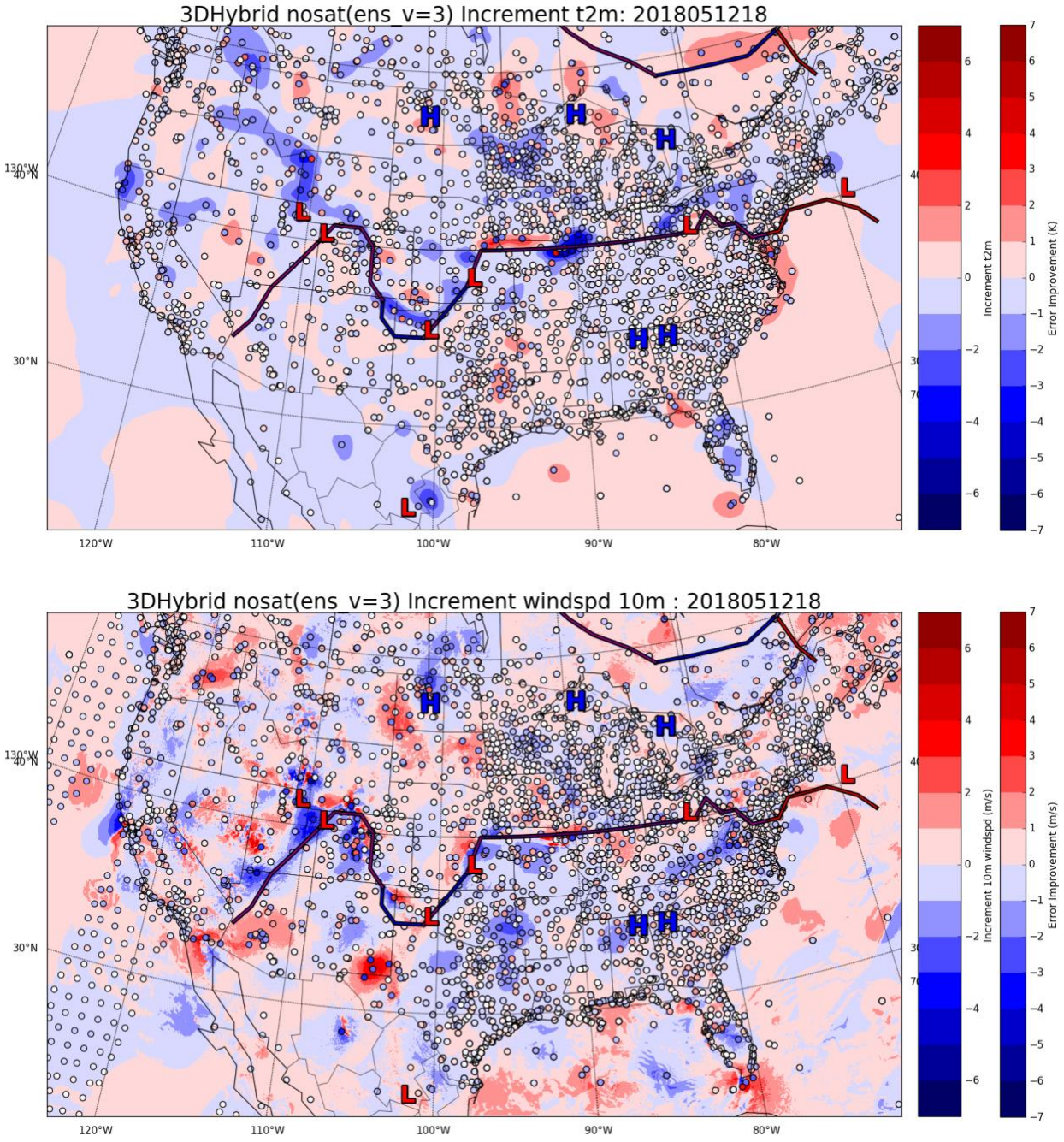


Figure 4.1.7*: The difference between the posterior and prior for 2m temperature (a) and 10m wind speed (b) on May 12th, 2018 at 18z, represented by the contour. The scatterplot displays the error improvement after assimilation (the difference between the absolute error of the posterior and the prior at the individual observation stations). Positive impact on the 3D Hybrid is denoted by negative (blue) error improvement values. Whereas, the negative impact is denoted by positive (red) improvement values. The High and Low-pressure center is represented by the H (blue) and L (red) symbols. The cold, warm and occluding fronts are portrayed by the blue, red and purple lines, respectively.*

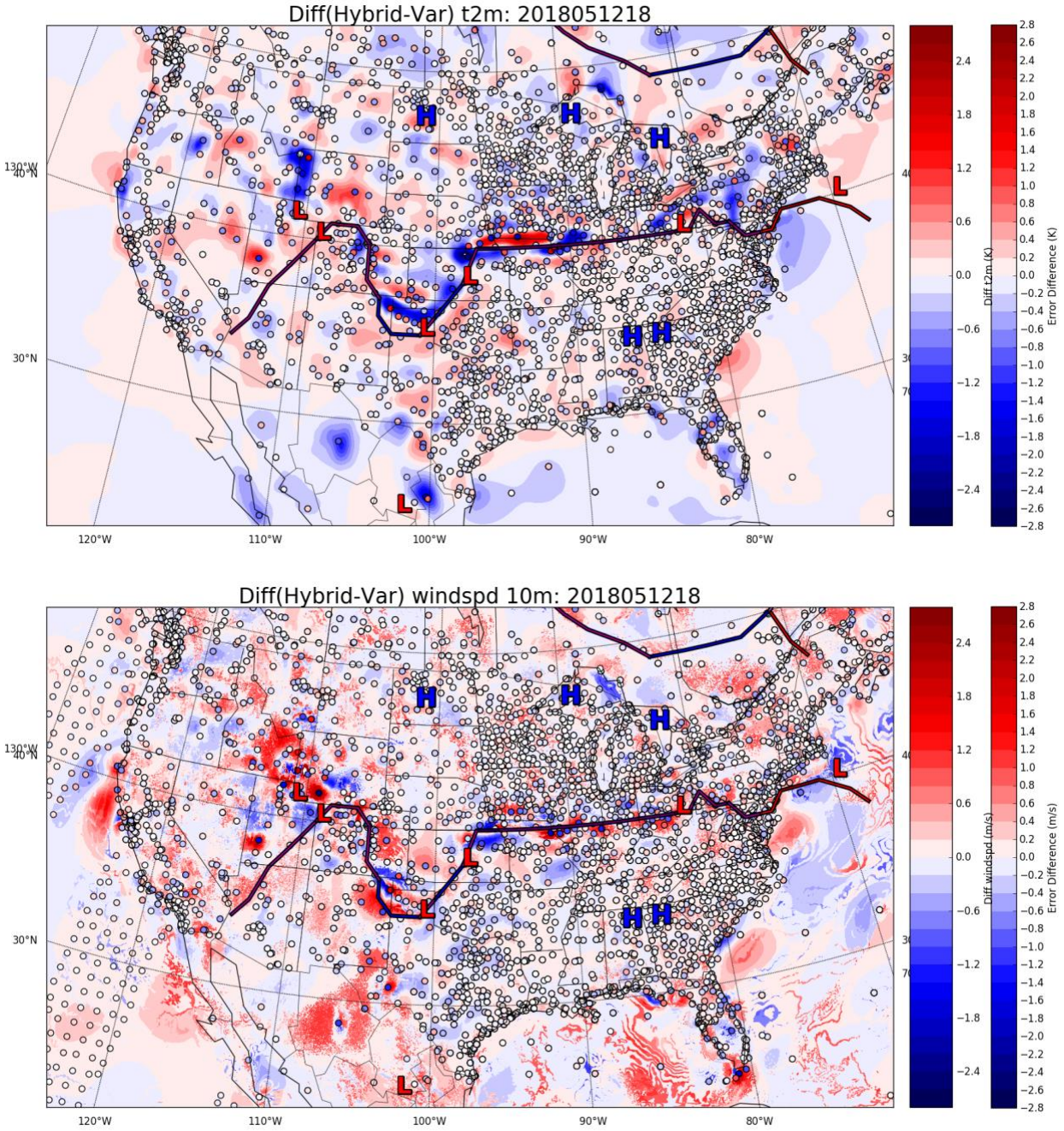
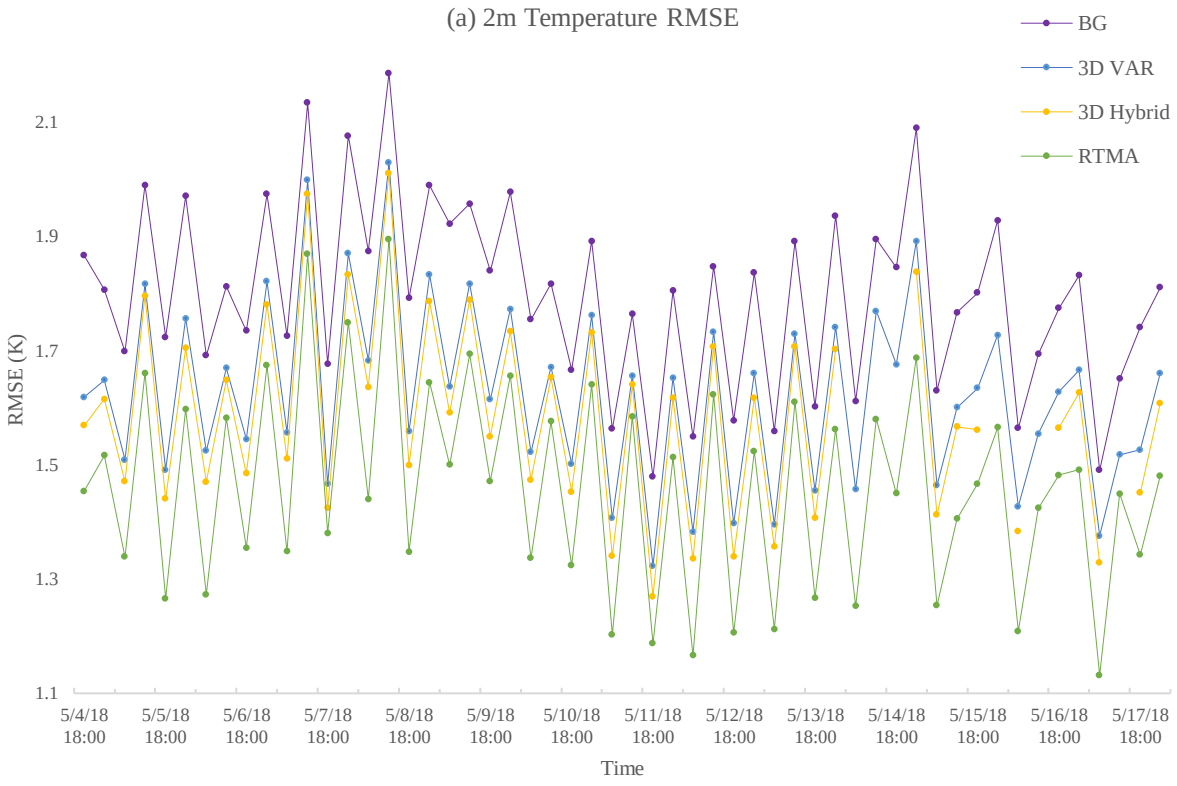


Figure 4.1.8*: The contour represents the difference between the hybrid and variational data assimilation analysis for 2m temperature (a) and 10m Wind Speed (b) on May 12th, 2017 at 18z. The scatterplot depicts the absolute error difference between the hybrid and variational scheme, verified at the individual observation stations. Negative (Positive) values indicate the absolute error from the hybrid analysis is smaller (greater) than the variational analysis. The High and Low-pressure center is represented by the H (blue) and L (red) symbols. The cold, warm and occluding fronts are portrayed by the blue, red and purple lines, respectively.*

4.1.1.3 Statistical Analysis

One can evaluate the performance of the 3D Hybrid analysis by computing a 6-hourly time series of the root mean square error (RMSE) for the periods between May 5th, 2018 at 00z to May 18th, 2018 at 00z. Similarly, the RMSE of background model (prior), 3D VAR and RTMA are also calculated and used as a benchmark. The RMSE comparisons for 2m temperature, 10m wind speed and 2m specific humidity are shown in Figure 4.1.3.1a, 4.1.3.1b, 4.1.9c. The 2m temperature RMSE comparison indicates that the 3D hybrid data assimilation analysis outperformed the background and the 3D VAR. However, the RTMA 2m temperature analysis has a significantly lower RMSE than the other three analyses. Similarly, the 10m wind speed RMSE comparison illustrates the RMSE for 3D Hybrid analysis is lower than the background and 3D VAR. Although the RTMA 10m wind analysis has a generally lower RMSE than the 3D Hybrid, the 3D hybrid analysis marginally outperformed the RTMA for 28% of the time. Lastly, the 2m specific humidity RMSE comparison depicts 3D Hybrid analysis outperformed the background, but its RMSE is only marginally lower than 3D VAR. Nevertheless, RTMA has lower RMSE than 3D Hybrid and 3D VAR by ~0.1 g/kg. Overall, the 3D Hybrid analysis demonstrated the ability to produce surface analyses that are more accurate than using the 3D variational scheme, but unable to surpass the accuracy of the RMTA.


(a) 2m Temperature RMSE
BG
3D VAR
3D Hybrid
RTMA
RMSE (K)
2.1
1.9
1.7
1.5
1.3
1.1
5/4/18 18:00
5/5/18 18:00
5/6/18 18:00
5/7/18 18:00
5/8/18 18:00
5/9/18 18:00
5/10/18 18:00
5/11/18 18:00
5/12/18 18:00
5/13/18 18:00
5/14/18 18:00
5/15/18 18:00
5/16/18 18:00
5/17/18 18:00
Time


(b) 10m Wind Speed RMSE
BG
3D VAR
3D Hybrid
RTMA
RMSE (m/s)
2.3
2.1
1.9
1.7
1.5
1.3
1.1
5/4/18 18:00
5/5/18 18:00
5/6/18 18:00
5/7/18 18:00
5/8/18 18:00
5/9/18 18:00
5/10/18 18:00
5/11/18 18:00
5/12/18 18:00
5/13/18 18:00
5/14/18 18:00
5/15/18 18:00
5/16/18 18:00
5/17/18 18:00
Time

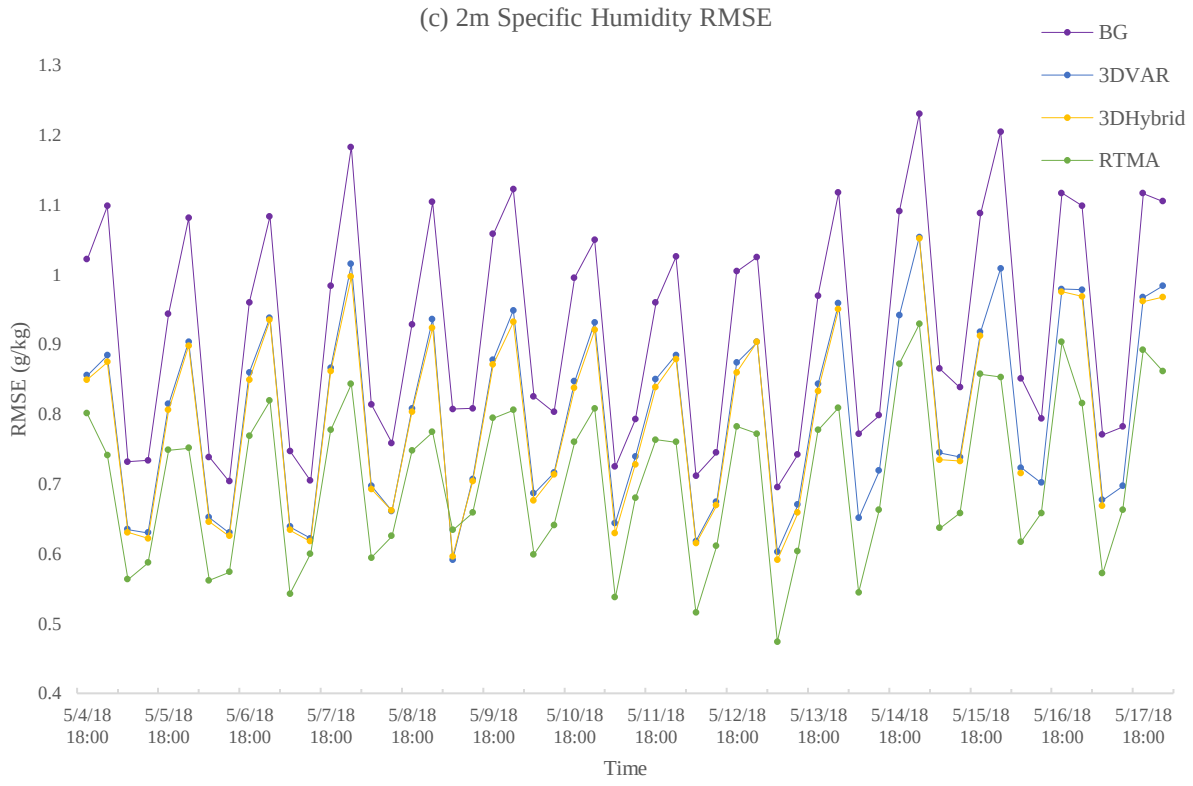


Figure 4.1.9*: 6 – hourly RMSE comparison between the background (purple line), 3D VAR (blue line), 3D Hybrid (orange line) and RTMA (green line) for 2m temperature (a), 10m wind speed (b) and 2m specific humidity (c). This study was conducted for the periods between May 5th, 2018 at 00z to May 18th, 2018 at 00z.*

In order to further examine the statistical performance of the surface analysis from the 3D Hybrid, 3D VAR and RTMA, a time and domain averaged RMSE is computed for the periods between May 5th, 2018 at 00z to May 18th, 2018 at 00z. Similar to the RTMA comparison in De Pondeca et al. (2011), the improvement percentage between the RMSE of prior and the posterior is obtained in this study by Eq. 4.1.3.1.

$$Improv\ (\%) = \frac{RMSE_{prior} - RMSE_{Posterior}}{RMSE_{Posterior}} * 100 \qquad Eq\ 4.1.3.1$$

The comparison includes 2m temperature, 10m wind speed and 2m specific humidity, as shown in Table 4.1. (a)-(c). From the 2m temperature comparison, the hybrid scheme has an improvement percentage of 13.46%, while the variational scheme has a lower improvement of 10.51 %. Meanwhile, the 10m wind speed comparison shows the hybrid improvement percentage of 19.67%, exceeding over the variational scheme of 14.61%. Lastly, the improvement percentage of the hybrid scheme marginally surpass the variational scheme by ~1% for the 2m specific humidity. The improvement percentage for RMTA cannot be computed because the prior (background) model data was not available for this study. However, a 15 days study conducted by De Pondeca et al. (2011) stated the RMTA's improvement percentage for 2m temperature, 10m wind speed and 2m specific humidity is 45%, 16% and 34% respectively. Although, one should be cautious comparing the results of this study with the conclusion from De Pondeca et al. (2011), as the two experiments were conducted for different time periods and background models (prior). Nevertheless, this comparison further demonstrated the Hybrid scheme produce a more accurate analysis than the variational scheme but is outperformed by RTMA.

2m Temperature RMSE (a)

Analysis	Prior (K)	Posterior (K)	Improvement (%)
Hybrid	1.81	1.59	13.46
VAR	1.81	1.64	10.51
RTMA	x	1.36	x

10m Wind Speed RMSE (b)

Analysis	Prior (m/s)	Posterior (m/s)	Improvement (%)
Hybrid	1.83	1.53	19.67
VAR	1.83	1.60	14.61
RTMA	x	1.44	x

2m Specific Humidity RMSE (c)

Analysis	Prior (g/kg)	Posterior (g/kg)	Improvement (%)
Hybrid	0.94	0.81	16.56
VAR	0.94	0.82	15.49
RTMA	x	0.56	x

Table 4.1.1*: A time and domain averaged RMSE comparison between the Hybrid, Variational schemes and RTMA for 2m Temperature (a), 10m Wind speed (b) and 2m Specific Humidity (c). The comparison includes results for the periods between May 5th, 2018 at 00z to May 18th, 2018 at 00z. The RMSE of the prior (background) and the posterior (analysis) with the improvement percentage (Eq. 4.1.3.1) are shown for each scheme.*

The terrain-following error covariance employed in the RTMA is a contributing factor that allows the 2D variational scheme to produce an accurate surface analysis. Since surface fields such as surface temperature, moisture and pressure exhibit strong dependency on the local terrain, one can incorporate terrain information to constrain the analysis increment spatially. As discussed by De Pondeca et al. (2011), the local terrain gradient is projected onto the autocovariance sub-matrices within the background error covariance matrix. As a result, analysis increment from a single observation measurement follows an anisotropic spatial distribution, following the local terrain rather than the isotropic analysis increment in a typical variational data assimilation system. On the other hand, the flow-dependent error covariance within the hybrid data assimilation scheme is able to characterize the local terrain through a weaker

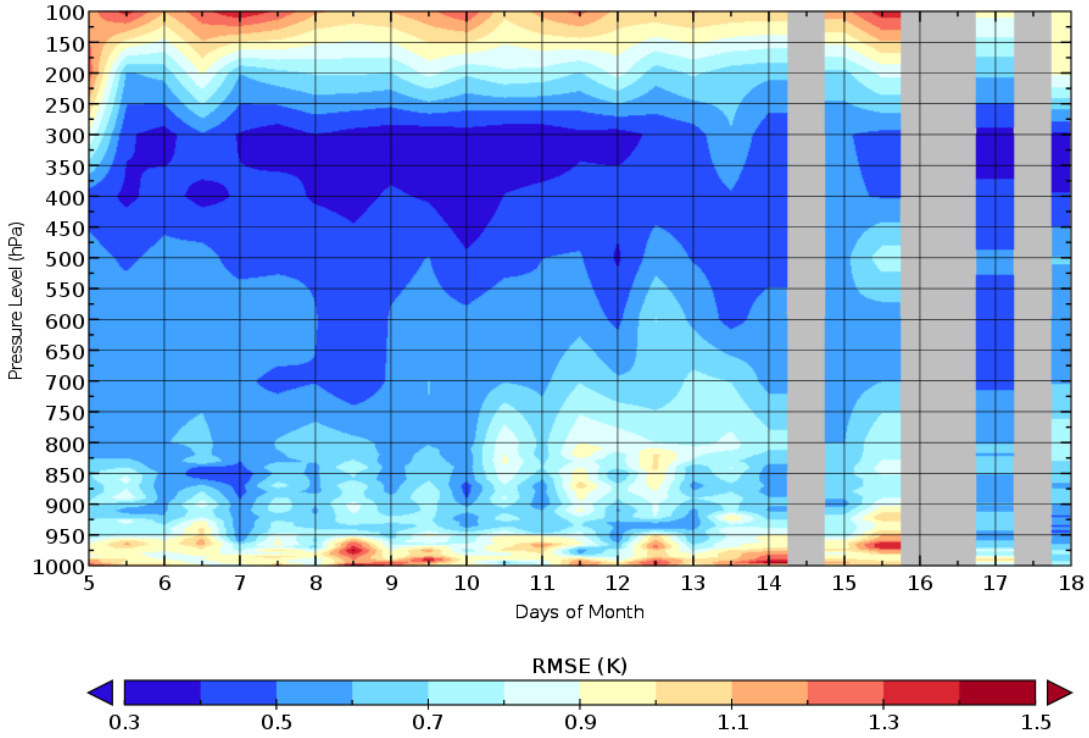
constraint than the terrain-following error covariance. As a result, the improvement percentage of 2m temperature and 2m specific humidity from the result RTMA in De Pondeca et al. (2011) is significantly higher than the results from the Hybrid scheme in this study, shown in Table 4.1. However, the improvement percentage of 10m wind speed from the Hybrid scheme is slightly higher than the published RMTA results. According to De Pondeca et al. (2011), the anisotropic terrain-following constraint on 10m wind speed is very weak to consider errors in wind circulations over mountains, rather than only around the mountains. Similarly, the flow-dependent error covariance in the hybrid scheme represents the errors of wind flow over terrain from the general circulation of the near-surface atmosphere. Although the comparison examines the results from different time periods, the findings suggest the weaker terrain following constraint in RTMA and the flow-dependent error covariance produce similar statistics on the accuracy of surface wind speed.

4.1.2 Upper Level Analysis Result

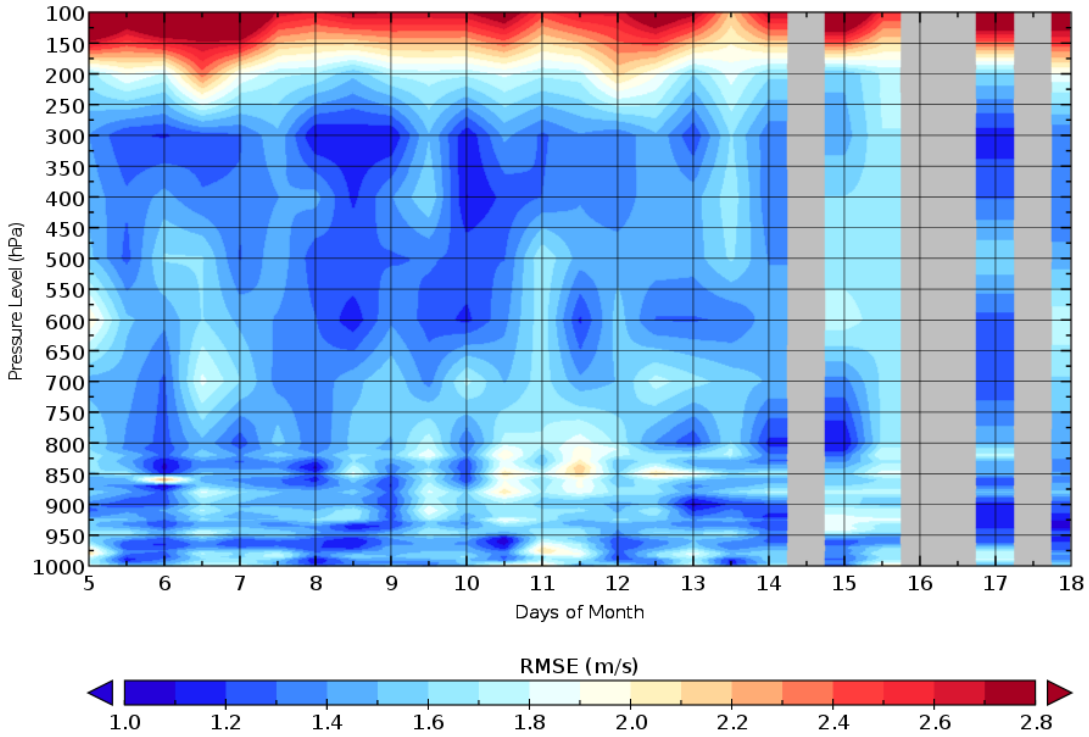
This section focuses on the statistical performance of the upper level analysis for the hybrid scheme during the period between May 5th, 2018 at 00z to May 18th, 2018 at 00z. The comparison consists of 12 – hourly vertical profiles of the RMSE verified with radiosonde observations over the CONUS domain for temperature, wind speed and specific humidity, as shown in Figure 4.1.10 (a) – (c). The RMSE vertical profile for temperature in Figure 4.1.10a reveals the RMSE at near-surface ranges from ~ 0.8 – 1.4 Kelvins, while the RMSE generally decreases to ~ 0.4 – 1.1K approaching the top of the planetary boundary layer (PBL) at 850 hPa pressure level. The RMSE continues to decreases within the mid-atmosphere at 500 hPa to ~ 0.4 – 0.9 K, before increasing to a peak of ~1.4 K above the tropopause at 200 hPa. The overall trend of the finding is consistent with results from the temperature RMSE comparison conducted for a period during the winter season by Wang et al. (2013). However, their RMSE values between 900 to 1000 hPa are slightly higher than the ones in this study, which can be caused by the different background model and the observation errors used in the hybrid system. Also, the relatively high RMSE values within the PBL can be contributed by high spatial and vertical variability in temperature induced by the strong thermal advection that typically occurs below 850 hPa. On the other hand, the high RMSE values between 250hPa to 100 hPa layer stems from the errors of background model, in which NWP commonly have difficulty to characterize the model's top boundary conditions that is within the tropopause. Unlike the temperature RMSE comparison, the wind speed RMSE vertical profile in Figure 4.1.10a exhibits the RMSE for the pressure levels below the tropopause is less dependent on the pressure level, where the RMSE at a given height varied from ~ 1.1 – 2.1 m/s. However, the maximum RMSE of 2.7 – 2.8 m/s are

seen within the tropopause, which is related to boundary condition in the background model. Contrarily, the results from Wang et al. (2013), suggests a decrease in wind speed RMSE with respect to height from 2.2 to 2 m/s within the 1000 hPa – 850 hPa layer, before increasing to 3.4 m/s at 100 hPa. Lastly, Figure 4.1.2.1c shows the specific humidity RMSE generally decreases with height. The RMSE spreads between ~ 0.4 – 1.0 g/kg below 700 hPa and values decreased to ~ 0.2 – 0.4 g/kg and 0.1 – 0.2 g/kg at 500 hPa and 300 hPa, respectively. The high RMSE in the lower atmosphere and lower values in upper troposphere can be justified by the high concentration of moisture generally situating below 700 hPa, while air loft tends to be drier in the most situations without convention weather systems. In addition, the specific humidity RMSE maxima between 950 – 750 hPa layer during May 11 at 00z to May 13 at 00z coincide with the temperature and wind speed RMSE maxima within the same period and atmospheric layer. Spatial comparison for upper level analysis in future work would provide further details to explain this situation.

(a) Temperature RMSE



(b) Wind Speed RMSE



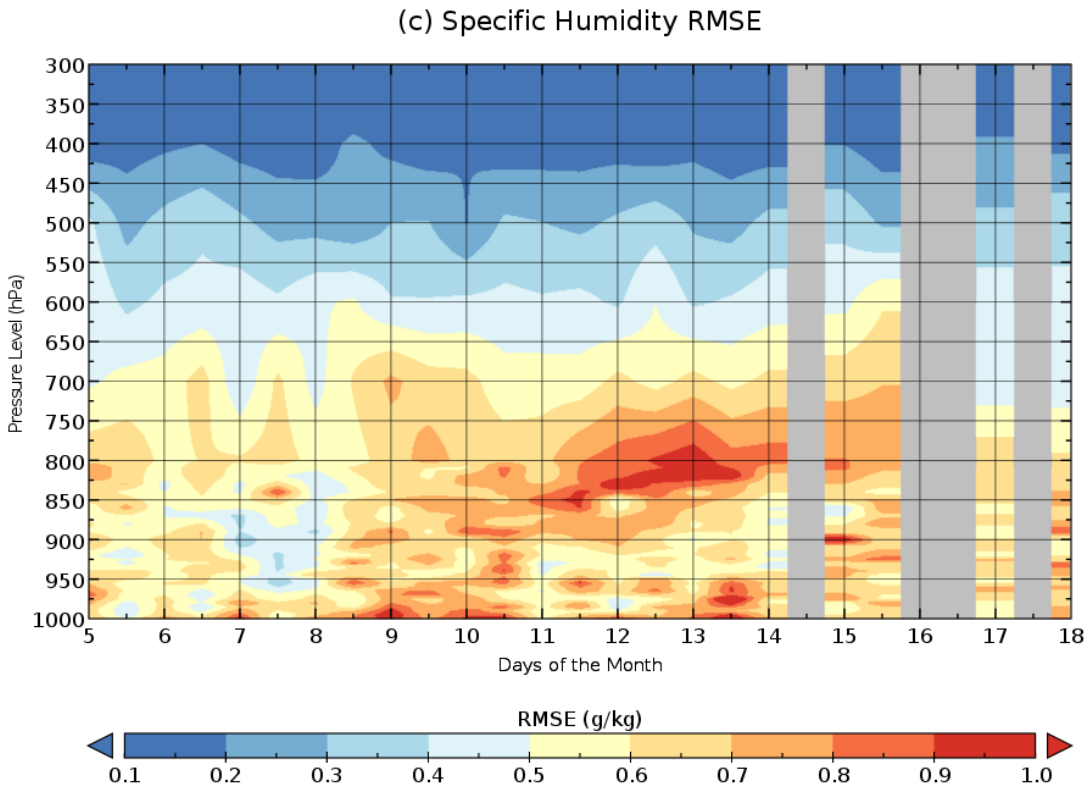


Figure 4.1.10*: The statistical performance of upper level analysis using the hybrid scheme for (a) Temperature, (b) wind speed, (c) specific humidity (d). The contours represent the RMSE vertical profile for the 12 – hourly comparison spanning from 00z on May 5th to 00z on May 18th, 2018.*

Similar statistical comparisons of the upper level analysis were conducted for the results using the variational scheme and the background model. These comparisons were used as a benchmark to evaluate the statistical performance of the hybrid scheme. The 12 – hourly RMSE time series for the hybrid scheme, variational scheme and the model background are examined at the pressure levels of 1000 hPa, 925 hPa, 850 hPa, 700 hPa, 500hPa and 200 hPa. The results for temperature, wind speed and humidity are plotted in Figure 4.1.11, 4.1.12 and 4.1.13, respectively. The temperature RMSE comparisons at 1000 hPa and 925 hPa suggest the hybrid scheme has a slightly lower RMSE than the variational scheme with the biggest difference of 0.14 – 0.16 K. The differences in RMSE are smaller for 850 hPa, with the values ranging from ~

0.05 – 0.08 during May 5th at 00z to May 15th at 12z, while slightly higher difference of ~ 0.16 – 0.23 K on May 17 and 18 at 00z. Furthermore, the RMSE values for hybrid and variational scheme are similar at 700 hPa and 500 hPa, while the RMSE for the hybrid scheme is slightly higher than the variational scheme by ~ 0.02 – 0.04 K at 200 hPa. Comparable results found in the wind RMSE comparison, where the hybrid scheme marginally outperformed the variational scheme with slight lower RMSE at 1000 hPa, 925 hPa, and 850 hPa. The most substantial differences in RMSE between the two schemes at 1000 hPa and 925 hPa are ~ 0.16 m/s and 0.18 m/s, respectively, while there is little difference in RMSE values at 700 hPa and 500hPa. However, the RMSE in the variational scheme is marginally lower than the hybrid scheme by ~0.04 m/s – 0.12 m/s at 200hPa. Lastly, the specific humidity RMSE comparison shows variational scheme incrementally outperformed the hybrid scheme at 1000 hPa and 925 hPa, while the two schemes have similar RMSE values at 850 hPa through 200 hPa.

The study by Wang et al. (2013) conducted a similar comparison to examine the statistical performance of the upper level analysis using hybrid and variational data assimilation. In their findings, the temperature RMSE for the hybrid scheme is lower than the variational scheme by approximately 0.1 K between 1000 hPa and 800 hPa. The difference in the RMSE between the two scheme decreases to ~ 0.05 K above 700 hPa, but the hybrid scheme consistently outperforms the variational scheme for all pressure levels. Similarly, the wind speed RMSE for the hybrid scheme is consistently lower than the variational scheme by ~ 0.2 m/s for all pressure levels. As a result, there is a discrepancy in the result between Wang et al. (2013) and this study. This could be caused by the different configured horizontal and vertical localizations in the two studies, which constrains the influence of the analysis increment of each

assimilated observation in the horizontal and vertical direction. For instance, the hybrid system in Wang et al. (2013) has set its horizontal and vertical localization to 1600 km and greater than 30 grid unit, respectively. On the other hand, the horizontal and vertical localization for our experiment is 100 km and 3 grid units. The small localization setting for this experiment preserved the localized features of the surface analysis. However, a more elongated localization is more suitable for upper level analysis because the features within the mid and upper troposphere are relatively uniform. Consequently, the upper level increment from an upper level observation is representative of a larger domain in the model space.

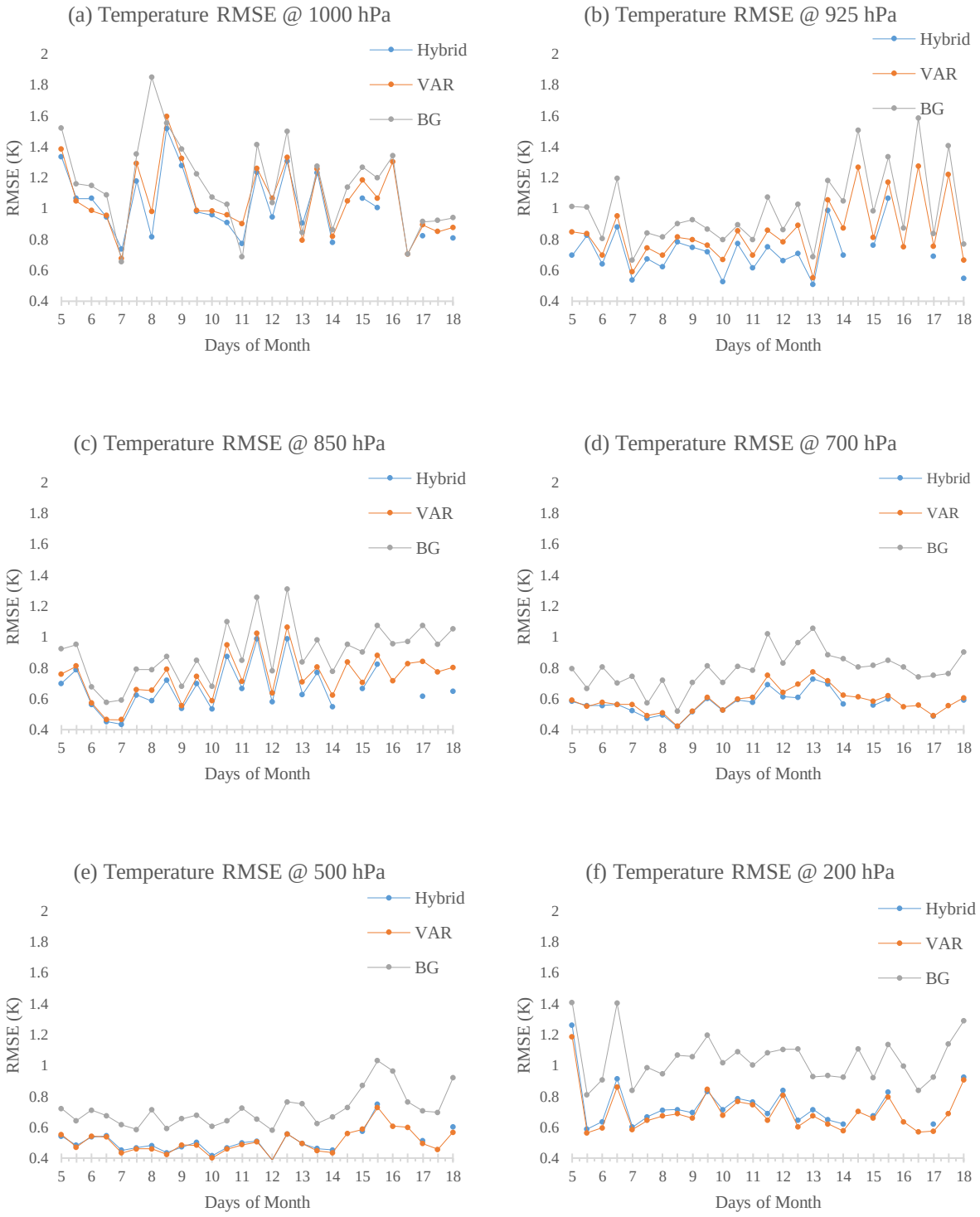


Figure 4.1.11: *12 – hourly time series are comparing temperature RMSE from the hybrid scheme (Blue), variational scheme (Orange) and the model background (Grey) for the period during May 5th, 2018 at 00z to May 18th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a) , 925 hPa (b) , 850 hPa (c) , 700 hPa (d), 500hPa (e) and 200 hPa (f)*

***Figure 4.1.12**: 12 – hourly time series are comparing wind speed RMSE from the hybrid scheme (Blue), variational scheme (Orange) and the model background (Grey) for the period during May 5th, 2018 at 00z to May 18th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a) , 925 hPa (b) , 850 hPa (c) , 700 hPa (d), 500hPa (e) and 200 hPa (f).*

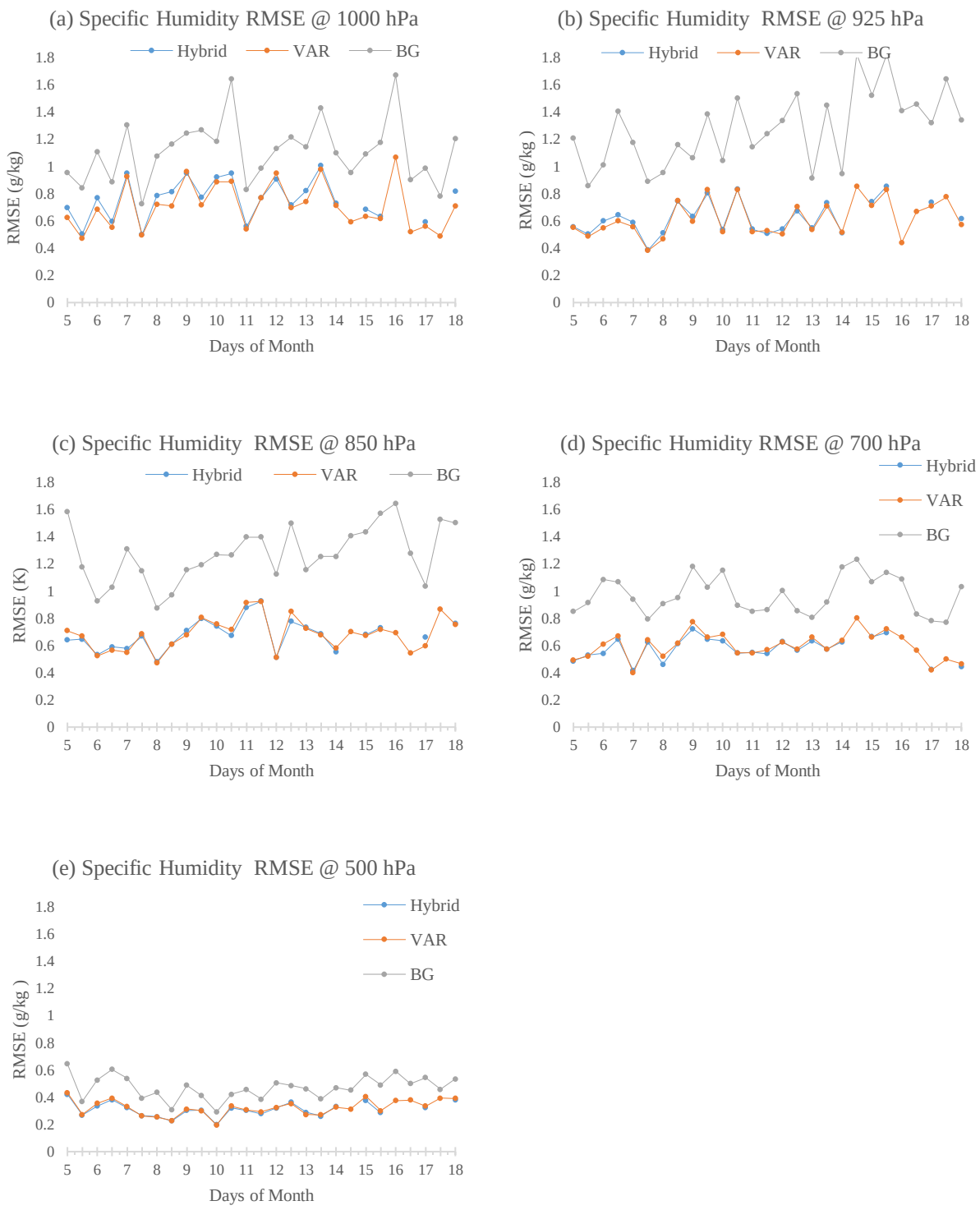


Figure 4.1.13*: 12 – hourly time series are comparing specific humidity RMSE from the hybrid scheme (Blue), variational scheme (Orange) and the model background (Grey) for the period during May 5th, 2018 at 00z to May 18th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a) , 925 hPa (b) , 850 hPa (c) , 700 hPa (d), 500hPa (e).*

4.2 The Effects of Vertical localization

One can conduct sensitivity tests on the vertical localization to examine the changes in the performance of surface and upper level analysis. In particular, four hybrid data assimilation experiments are conducted with vertical localization set to 3, 6, 9 and 12 vertical grid units, which are abbreviated by Hyb_v3_h100, Hyb_v6_h100, Hyb_v9_h100 and Hyb_v12_h100, respectively. Figure 4.2.1 illustrates the height above ground at each of the 50 model levels to help show the relationship between the vertical localization in grid units and the vertical distance. For example, 3 grid units localization for an observation measured near the 46^{th} model level is equivalent to the vertical distance between 46^{th} – 49^{th} model level or 43^{rd} – 46^{th} model level, which is ~ 1453m and 1159 m, respectively. It's worthwhile noting that vertical distances between the lower model levels are significantly smaller than distances at the higher model levels. For instance, the vertical distances between the 1^{st} – 4^{th} model level is ~100m. Due to the limiting computing resource, the four experiments were running during a 4-day period between May 9^{th} at 00z to May 13^{th} at 00z and moisture observations are not assimilated.

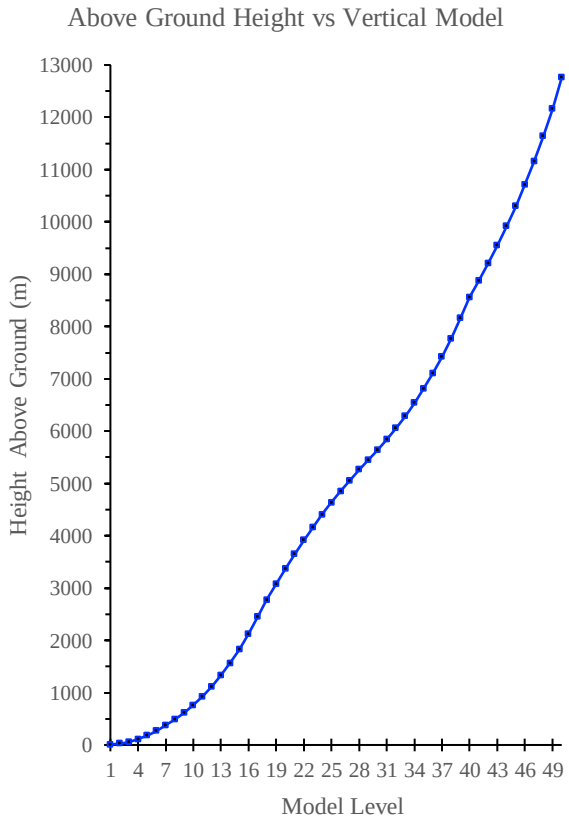


Figure 4.2.1*: Depicts the height above ground for each of the 50 model models.*

Firstly, the statistical performance of the surface analysis is examined for the hybrid data assimilation experiment with varying vertical localization settings. The time and domain averaged RMSE for the 2m temperature and 10m wind speed are shown in Table 4.2. The comparison also includes the improvement percentage between the analysis and the background, as described in Eq. 4.1.3.1. From the results, the 2m temperature and 10m wind speed RMSE marginally increases with for the experiment with high values of vertical localization. When the vertical localization increased from 3 to 12 grid units, the 2m temperature RMSE risen from 1.56850 K with 14.801 % improvement to 1.56882 K with 14.777% improvement. Similarly, the wind speed RMSE increased by ~ 0.00067 m/s and the improvement percentage decreased from

21.470 % to 21.416. Overall, statistical performance of the surface analysis did not improve by increasing vertical localization and expand the influence of increments from assimilating upper level observations.

2m Temperature RMSE (a)

Analysis	Posterior (K)	Improvement (%)
VAR	1.59904	12.608
Hyb_v3_h100	1.56850	14.801
Hyb_v6_h100	1.56871	14.786
Hyb_v9_h100	1.56875	14.782
Hyb_v12_h100	1.56882	14.777

10m Wind Speed RMSE (b)

Analysis	Posterior (m/s)	Improvement (%)
VAR	1.60021	15.978
Hyb_v3_h100	1.52786	21.470
Hyb_v6_h100	1.52808	21.452
Hyb_v9_h100	1.52836	21.429
Hyb_v12_h100	1.52853	21.416

Table 4.2.1*: A time and domain averaged RMSE comparison between the Hybrid scheme with vertical localization set to 3, 6, 9 and 12 grid units and the Variational schemes for 2m Temperature (a), 10m Wind speed (b). The comparison includes results for the periods between May 9th, 2018 at 00z to May 13th, 2018 at 00z. The RMSE of the posterior (analysis) and the improvement percentage (Eq. 4.1.3.1) are shown for each scheme.*

Further statistical comparisons were conducted to examine the performance of upper level analysis from adjusting the vertical localization. In particular, the 12 – hourly temperature and wind speed RMSE time series comparisons in Figure 4.2.2 and Figure 4.2.3, include the results at 1000 hPa, 925 hPa, 850 hPa, 500 hPa and 200 hPa for the hybrid scheme with vertical localization of 3, 6, 9 and 12 grid units. The results from the variational scheme are shown in the plot as a benchmark. The temperature and wind speed comparisons at all the pressure levels show a minimal difference in RMSE between hybrid scheme with localization varying from 3 to 12 grid units. This finding suggests there is an insignificant impact on the statistical performance

of upper analysis by modifying the vertical localization from 3 to 12 grid units with the horizontal localization set to 100 km. However, a conclusion cannot be made solely based on these results. An experiment using a hybrid ETKF -3D VAR data assimilation conduct by Wang et al. (2008a) shows significant improvement in RMSE for temperature and zonal and meridional wind speed by setting the optimal horizontal localization and weighting between static and ensemble covariance. For instance, their lowest RMSE for winds was achieved by placing 20% and 80% weighting on the static and ensemble covariance and a horizontal localization scale of 1414 km. This approach had an overall 20.7 % improvement over the 3D VAR analysis. Other studies such as Pan et al. (2014) utilized GSI's feature to vary the horizontal and vertical localization scale with height. For example, the horizontal localization at the surface is 700 km, but gradually increase up to 1050 km at the top of the model. Similarly, the vertical localization scale also increased with height, but it's also dependent on the variable type. This gives motivation for determining the most favorable combination of total covariance weighing, vertical and horizontal localization setting to produce the lowest RMSE for surface and upper level analysis in future work.

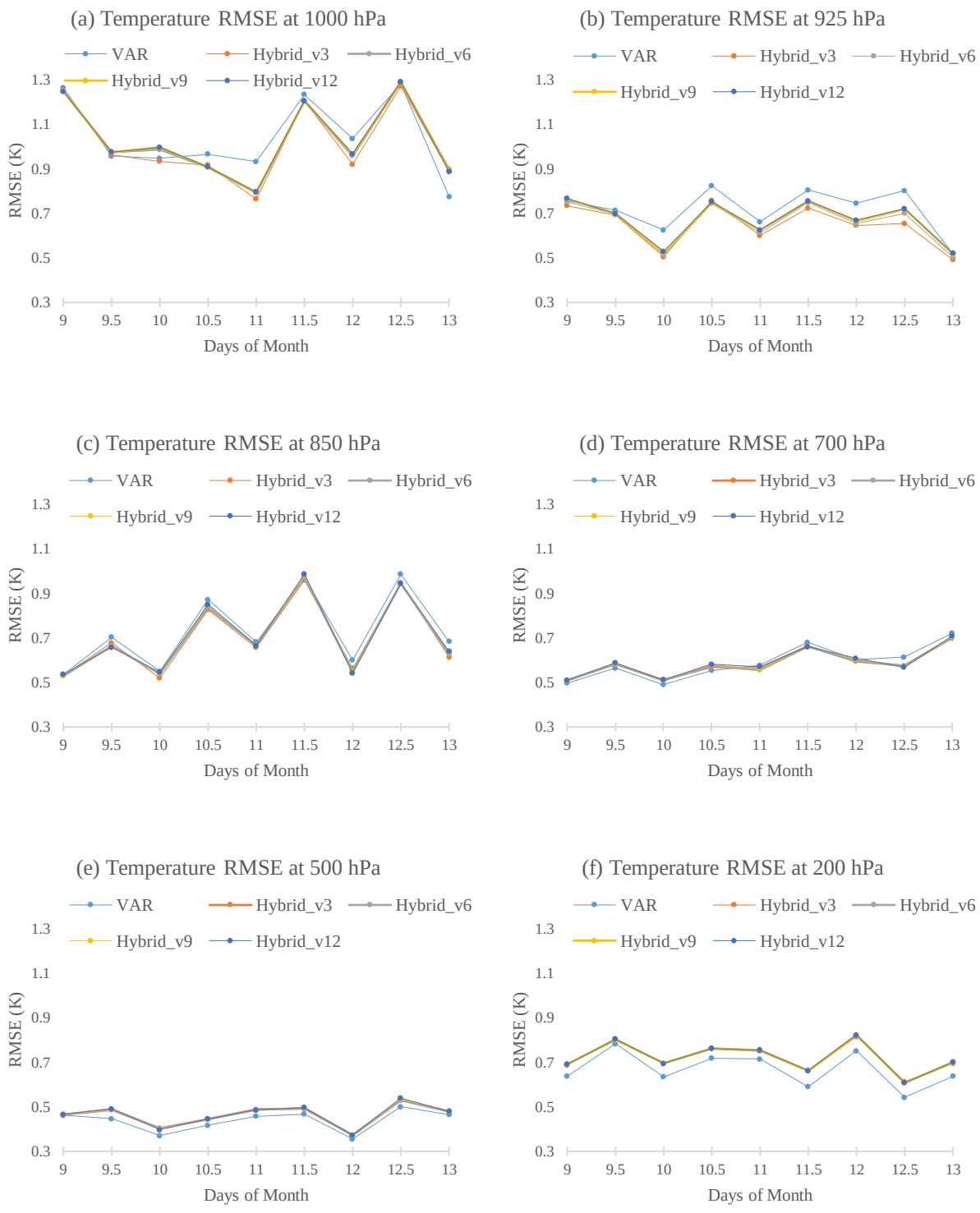


***Figure 4.2.2**: 12 – hourly time series are comparing temperature RMSE between the Hybrid scheme with vertical localization set to 3, 6, 9 and 12 grid units and the Variational schemes for the period during May 9th, 2018 at 00z to May 13th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a), 925 hPa (b) , 850 hPa (c) , 700 hPa (d), 500hPa (e) and 200 hPa (f).*

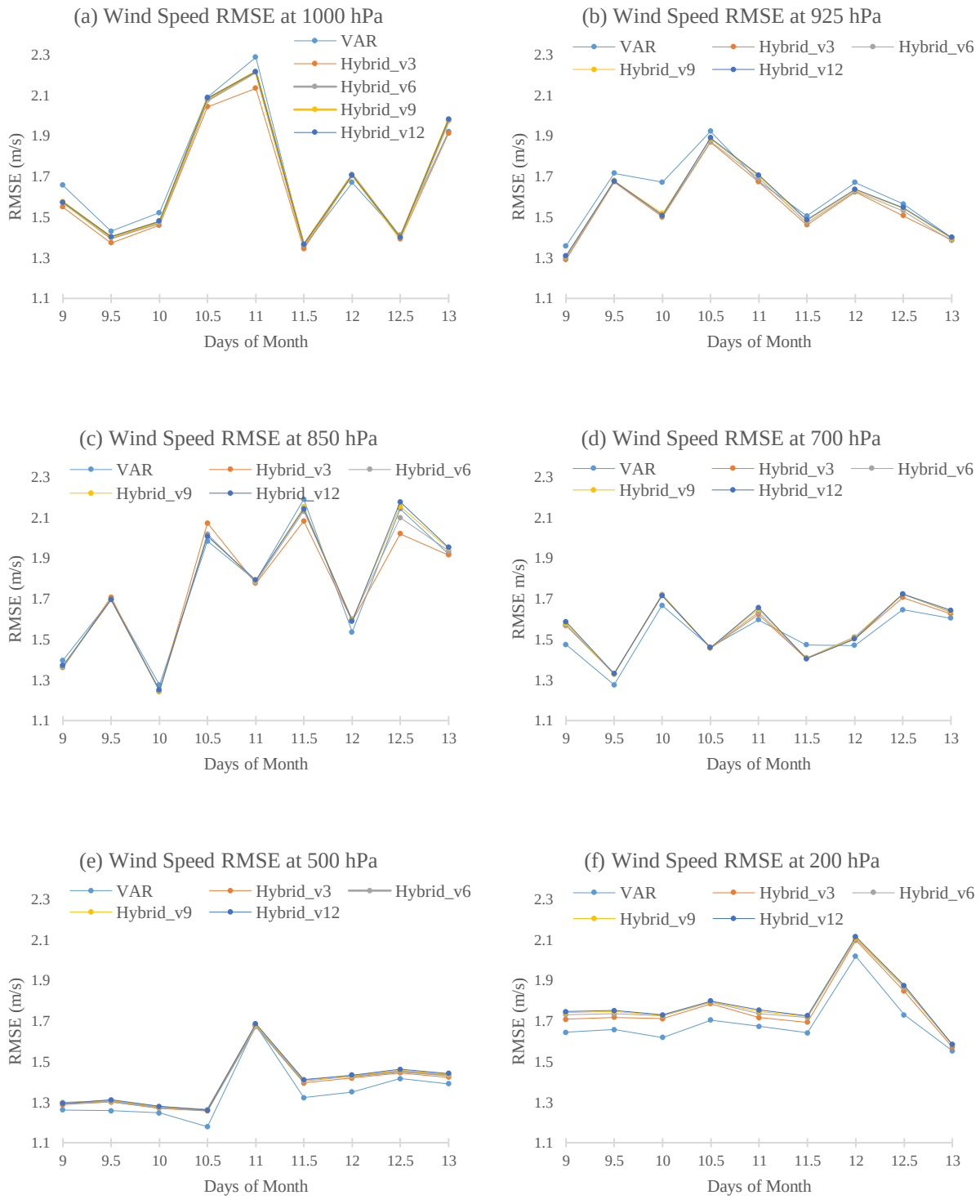


Figure 4.2.3: *12 – hourly time series are comparing wind speed RMSE between the Hybrid scheme with vertical localization set to 3, 6, 9 and 12 grid units and the Variational schemes for the period during May 9th, 2018 at 00z to May 13th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a), 925 hPa (b) , 850 hPa (c) , 700 hPa (d), 500hPa (e) and 200 hPa (f).*

4.3 The Assimilating Satellite Data

The following discussions address the impact of assimilating satellite data on improving the accuracy of the analysis. Unlike the radiosonde observations, the satellite data provides a broader data coverage with frequent measurement updates. Table 4.3 depicts the number of radiance observations that are ingested into the hybrid data assimilation after a 60 km data thinning for each instrument at 18z on May 7th, 2018.

Instrument	Number of Radiance Observations
AMSUA NOAA-15	0
AMSUA NOAA-18	15
AMSUA NOAA-19	0
AMSUA MetOp-A	17431
AMSUA MetOp-B	26031
MHS NOAA-19	0
MHS MetOp-A	7735
MHS MetOp-B	10745

Table 4.3.1*: The number of radiance observations measured from various instrument on the NOAA – 15, 18, 19 and MetOp – A, B satellite vehicles. These observations were assimilated for the analysis at 18z on May 7th, 2018. A 60 km data thinning is applied to reduce the density of observations.*

A Hybrid data assimilation experiment, ingesting satellite radiance and conventional observations was conducted for the period during 00z on May 5th to May 18th, 2018. The analysis cycles within the first 5 days are used to iteratively refine the radiance bias correction coefficient effectively ameliorate the radiance biases from the satellite observations. Typically, the radiance observations have systematic biases that can deteriorate the accuracy of the analysis. Therefore, bias correction on the radiance observations must be applied before they are assimilated, as discussed in Section 2.6. In order to demonstrate the impact of radiance bias correction, one can

compare brightness temperature mean bias error (MBE) for the bias-corrected prior (Guess) and posterior (analysis) against non-bias corrected prior and posterior. Figure 4.3.1a illustrates this comparison for channel 8 of Advanced Microwave Sounding Unit – A on the MetOps – A vehicle. The MBE for prior and posterior with bias correction are noticeably smaller than the values for prior and posterior without bias correction. The total bias correction applied to this channel is comparable with the monthly averaged value for -0.832 K that was published by NCEP. Furthermore, Figure 4.3.2a shows the bias correction improves the RMSE of prior and posterior from 0.5 – 0.9 K to 0.2 – 0.6 K and from 0.6 – 1.4 K to 0.18 – 0.5 K, respectively. Therefore, the results indicate the RMSE for both the prior and posterior were systematically reduced by applying radiance bias correction. However, the bias correction was not as effective for other channels and instruments, as seen in the comparison for channel 2 of Microwave Humidity Sounder on MetOp – B in Figure 4.3.1b. In particular, it had little impact on reducing the biases with periods when the MBE rather increased for prior and posterior with bias correction. The RMSE comparison in Figure 4.3.2b, further suggests the bias correction did not reduce but increased the RMSE in some cases. Overall, bias correction consistently improved the RMSE values for 20 out of 52 channels that were assimilated.

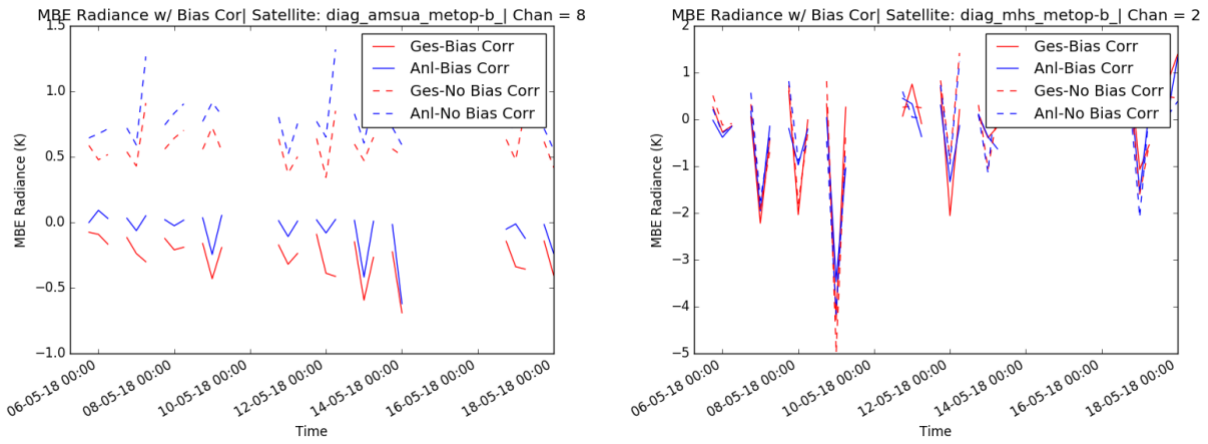


Figure 4.3.1*: Mean Bias Error (MBE) comparison between bias corrected prior and posterior (solid lines) against the non-bias corrected prior and posterior (dashed lines). The period of this comparison spans from 00z on May 5th, 2018 to 00z on May 18th, 2018. The results for channel 8 of AMSU-A on MetOp – B is shown on the left plot, while channel 2 of MHS on MetOp – B is shown on the right plot.*

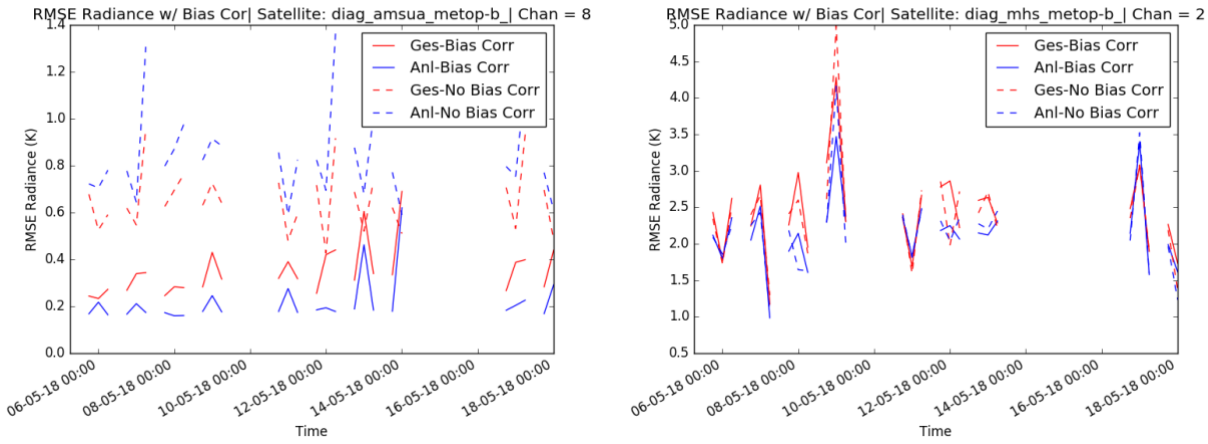


Figure 4.3.2: *Root Mean Square Error (RMSE) comparison between bias corrected prior and posterior (solid lines) against the non-bias corrected prior and posterior (dashed lines). The period of this comparison spans from 00z on May 5th, 2018 to 00z on May 18th, 2018. The results for channel 8 of AMSU-A on MetOp – B is shown on the left plot, while channel 2 of MHS on MetOp – B is shown on the right plot.*

As discussed above, the results indicate the radiance bias correction was not able to remove the symmetric bias and reduce the RMSE value for the significant sets of the channels. However, this procedure is essential to avoid introducing additional error into the analysis. A key component in the process is utilizing the radiance bias correction coefficients that are calibrated with a particular set of satellite observations and the background model. One must iteratively tune the set of coefficients through a series of analysis cycles to perform the optimal bias correction. Among the 32 channels with poor bias correction, the MBE comparison for channel 2 of MHS in Figure 4.3.1b, suggests the bias did not decrease for the prior and posterior with bias correction throughout the study. In other words, the radiance bias correction coefficient for this channel has not yet converged to a representative set of values. The satellite data assimilation in Zhu et al. (2014) mentioned their spin-up period to obtain optimal coefficients took the first 8 days of the analysis cycles. However, the length of the spin-up period can vary from weeks to months, depending on the initialized coefficient values, as stated by Zhu et al. (2014) and Shao et al. (2016). The set of initialized coefficients used in our experiment was taken from Global Data Assimilation System (GDAS) that is used to initialize the Global Forecast System (GFS). As a result, the initialized coefficients were not calibrated for the set of satellite observations and the HRRR model background that were used in our experiment. Consequently, a more extended spin-up period could be necessary for our experiment to allow the radiance bias correction coefficients to converge and effectively remove the systematic bias from all channels there were assimilated. Due to the limited available computing resources, this study was not able to include a longer period of spin-up and examine the improvement of the other 32 channels. Nevertheless, the satellite data assimilation experiment has demonstrated the effectiveness of the enhanced radiance bias correction on reducing systematic bias. Positive results were seen in large sets of

channels and further improvement can be made by introducing well-fitted initialized radiance bias correction coefficients and also having a longer spin-up period.

From the bias – corrected channels, one can examine the difference (OmB) between the observed brightness temperature and the simulated brightness temperature model output for the prior and the posterior. As background information on satellite data assimilation, the innovation is computed in the observational space, in which The Community Radiance Transfer Model (CRTM) is used as the radiance observation operator to convert model field such as temperature, moisture and ozone profiles into simulated brightness temperature, as stated by Shao et al. (2016). Figure 4.3.4a and 4.3.4b illustrates the brightness temperature OmB comparison for the prior and posterior from channel 8 of the AMUS – A MetOp – B satellite at 18z on May 10th, 2018. The scatter plot indicates the quality-controlled observation point with a 60 km data thinning, while the color shadings within the scatterplot represent brightness temperature OmB. The prior results in Figure 4.3.3a show the simulated model underestimates brightness temperature, compared to the bias – corrected observed brightness temperature for northern regions of the Midwest, while the opposite situation is true for areas south of Arizona and western parts of Mexico. On the other hand, OmB values generally decreased for these regions after the radiance observations are assimilated as seen by the posterior result in Figure 4.3.3b. The decrease in OmB from the prior to the posterior of this particular channel and instrument were found throughout the period of this experiment. This result is consistent with the reduced RMSE for the same channel in Figure 4.3.2a. Overall, the assimilation of bias correction radiance observation from channel 8 on AMSU – A MetOp – B satellite shows improvement on the accuracy of the analysis. A similar study was conducted for one of the channels with poor

bias correction. The OmB comparison between the prior and posterior from channel 2 of the MHS MetOp – B satellite are shown in Figure 4.3.4a and 4.3.4b. The results show relatively high OmB values both prior and posterior with no improvement in reducing OmB and RMSE after assimilating the radiance observation. In general, one can clearly distinguish the bias – corrected channels that improve the analysis by examining the MBE, RMSE and spatial OmB comparison and effectively assimilate radiance observations without introducing additional errors.

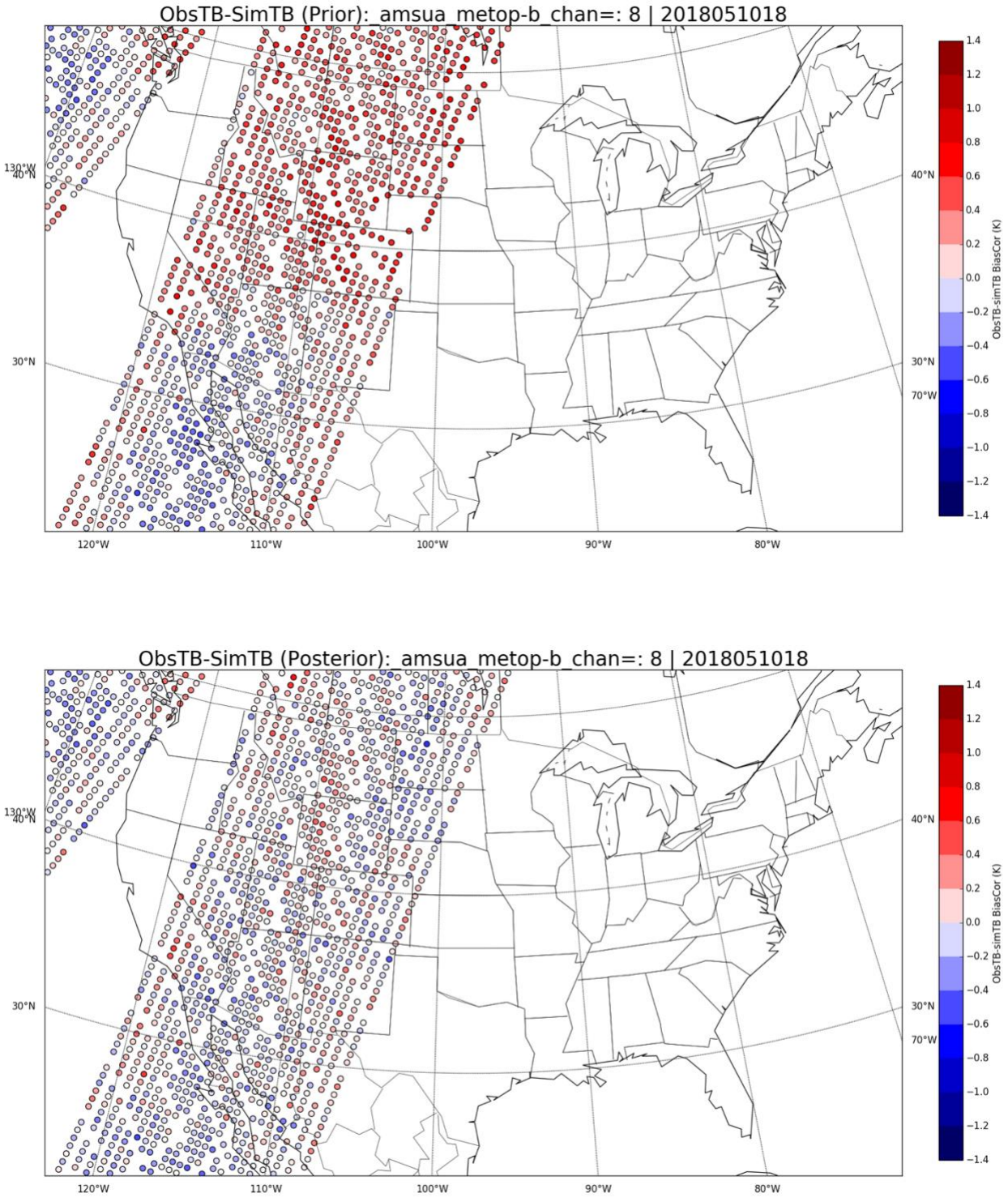


Figure 4.3.3: *The difference (OmB) between the bias-corrected observed brightness temperature and the simulated brightness temperature model output of the prior (Top) and the posterior (Bottom) for channel 8 of AMSU-A on MetOp – B at 18z on May 10th, 2018.*

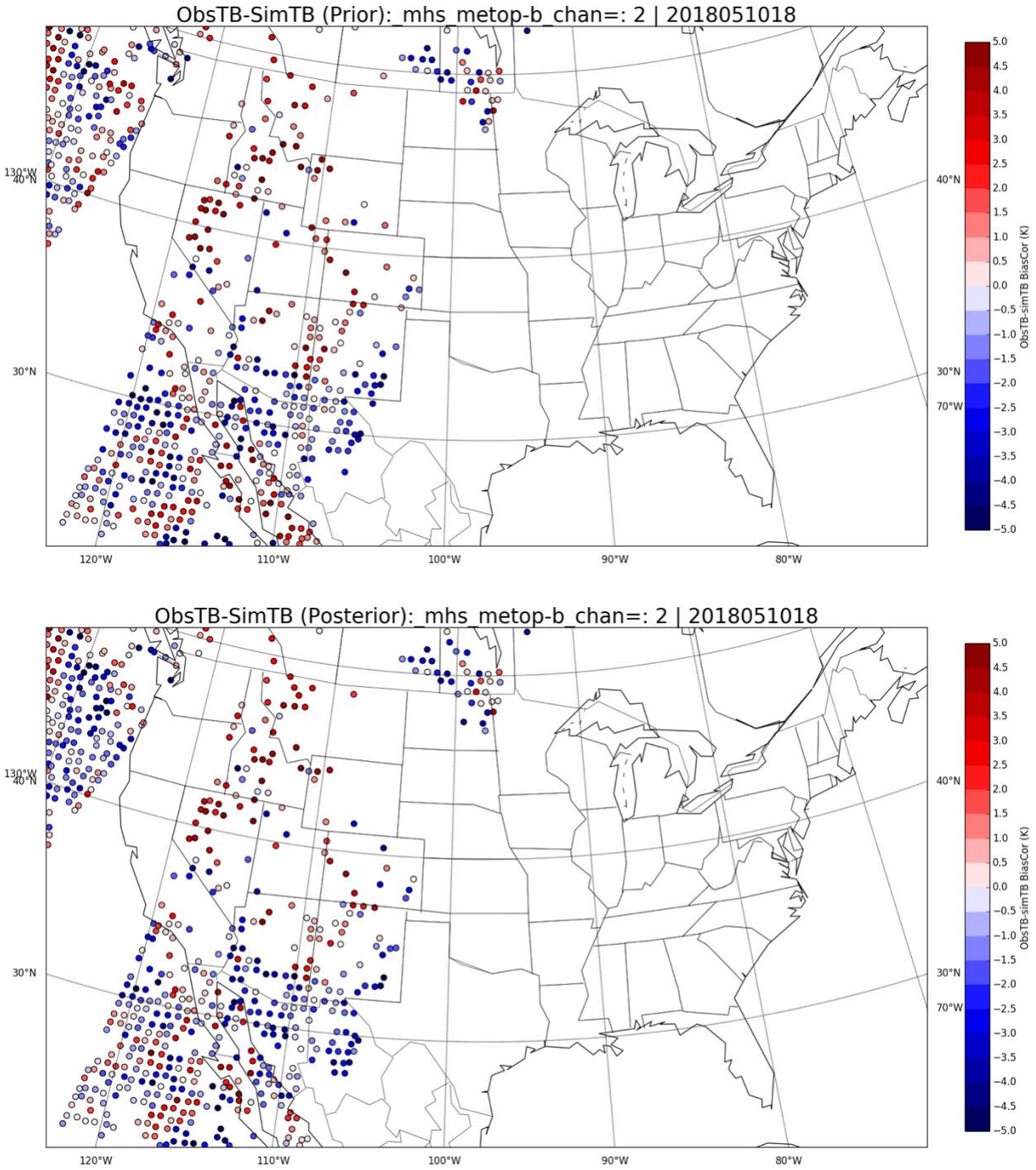


Figure 4.3.4: *The difference (OmB) between the bias corrected observed brightness temperature and the simulated brightness temperature model output of the prior (Top) and the posterior (Bottom) for channel 2 of MHS on MetOp – B at 18z on May 10th, 2018.*

In order to further examine the impact of assimilating satellite observation, the temperature and wind speed RMSE at 1000 hPa, 850 hPa, 700 hPa, 500 hPa and 200 hPa pressure level was plotted for the period during May 5th, 2018 at 00z to May 18th, 2018 at 00z in Figure 4.3.5 and 4.3.6. The comparison includes the analysis RMSE from assimilating conventional and satellite observation using the hybrid scheme (Hybrid_Sat), assimilating only the conventional observations using the hybrid scheme (Hybrid_noSat) and variational scheme (VAR). As expected, the temperature and wind speed RMSE are higher when satellite observations are assimilated, especially above 850 hPa. In particular, the temperature and wind speed RMSE at 200 hPa for Hybrid_Sat can be ~0.36 K and ~0.9 m/s higher than Hybrid_noSat. From previous discussions, radiance bias correction is essential to remove systematic bias within satellite observation and prevent channel with poor radiance bias correction to deteriorate the analysis. Although the results have demonstrated the bias correction used in this experiment was successfully able to remove biases from numerous channels, the bias correction coefficients were not able to represent the biases for a significant number of channels given the amount of time for spin-up. Therefore, an improvement on the analysis is expected with a longer spin up period to allow the bias correction coefficient to converge and successfully remove the biases before assimilating the radiance observations.

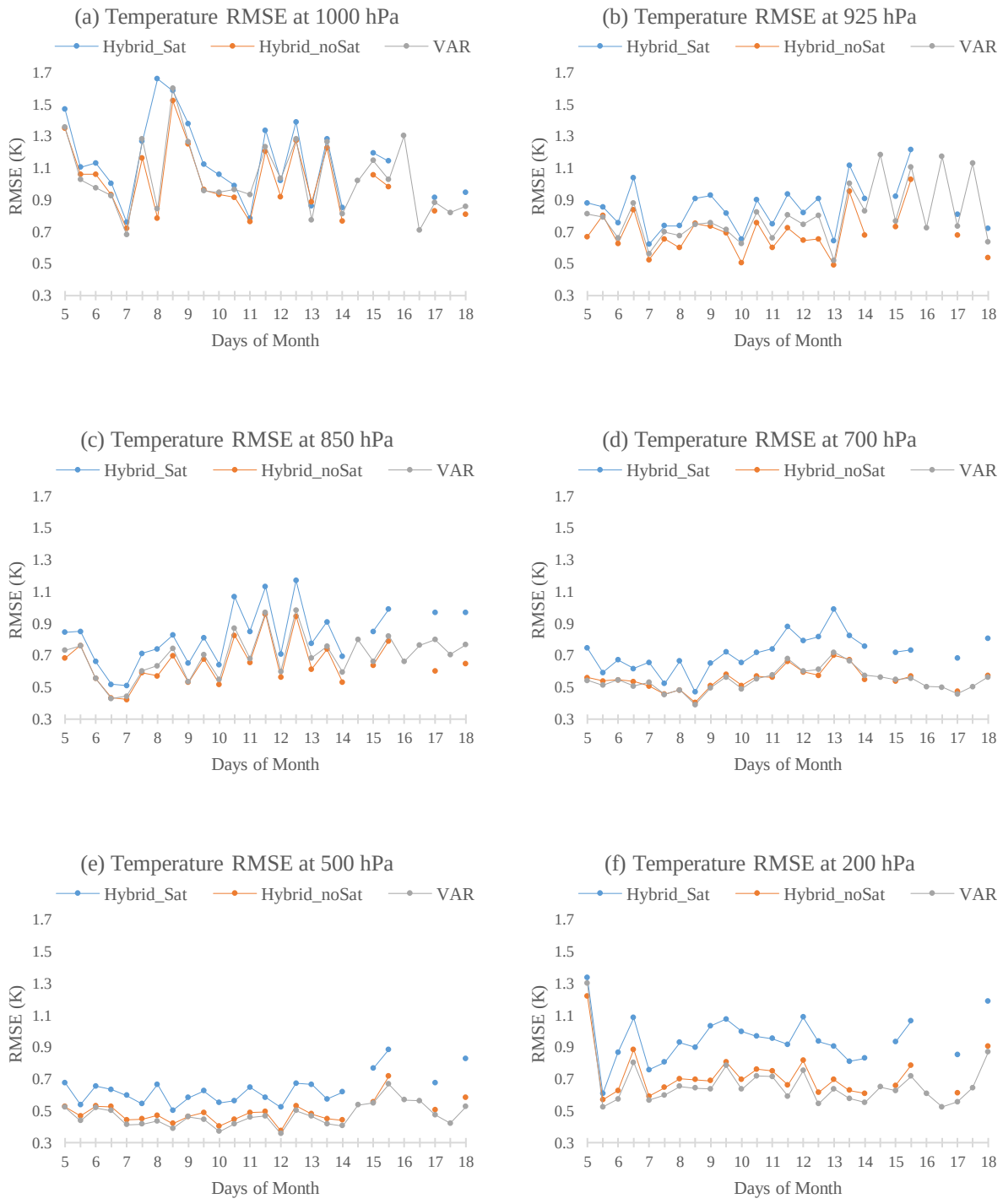


Figure 4.3.5: *12 – hourly time series are comparing temperature RMSE between the analysis RMSE from assimilating conventional and satellite observation using the hybrid scheme (Hybrid_Sat), assimilating only the convectional observations using the hybrid scheme (Hybrid_noSat) and variational scheme (VAR) for the period during May 5th, 2018 at 00z to May 18th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a), 925 hPa (b) , 850 hPa (c) , 700 hPa (d), 500hPa (e) and 200 hPa (f).*

Figure 4.3.6: *Figure 4.3.6: 12 – hourly time series comparing wind speed RMSE between the analysis RMSE from assimilating conventional and satellite observation using the hybrid scheme (Hybrid_Sat), assimilating only the conventional observations using the hybrid scheme (Hybrid_noSat) and variational scheme (VAR) for the period during May 5th, 2018 at 00z to May 18th, 2018 at 00z. The RMSE time series are shown at the pressure levels of 1000 hPa (a), 925 hPa (b), 850 hPa (c), 700 hPa (d), 500hPa (e) and 200 hPa (f).*

4.4 Computational Cost

The computational cost among the 3D Hybrid, 3D VAR and 2D RTMA data assimilation schemes depends on the several factors including the size and dimension of the analysis' domain, amount of assimilated observations, size of the background model and ensemble members. In Table 4.4.1, the computational cost for each scheme is summarized in terms of CPU and Memory that were allocated, running time and CPU* running time for each analysis cycle. Overall, the 3D Hybrid scheme uses the most CPU and memory per analysis cycle, while the 2D RTMA scheme uses the least. The 3D Hybrid scheme required the most CPU and memory because it ingest and store the very large global ensemble dataset (80 GB). Therefore, smaller computation cost for the hybrid scheme can be achieved by having additional pre-process procedure to extract a CONUS subset data from the global ensemble dataset and ingested it into GSI. On the other hand, the RTMA required less CPU because the dimension of the analysis is reduced from 3D to 2D analysis. Consequently, the background error covariance matrix, control variable vector and observation vector is reduced. The drawback is that RTMA does not assimilate upper air observations, does not incorporate flow – dependent background error covariance and does not provide upper air analysis.

	CPU	Memory Allocated (GB)	Running Time (min)	CPU* Running Time
3D Hybrid	624	2808	~ 60	374400
3D VAR	432	576	~ 150 min	248832
2D RTMA	96	128	~13 mins	512

Table 4.4.1: *The computational cost in terms of allocated CPU, memory, running time and CPU * Running time per analysis cycle for each of the schemes.*

5 Conclusions and Future Work

A study on the performance of the analysis produced by 3D hybrid data assimilation was presented. The 6 – hourly update analysis includes temperature, wind speed, specific humidity and pressure. It has a CONUS domain with 3 km horizontal resolution and 50 vertical native levels. The hybrid data assimilation ingests the HRRR 1 – hour forecast as the background and assimilates land and marine surface, VAD, aircraft, radiosonde and satellite observations. The results of the hybrid analysis are compared against the background, variational analysis and existing operational RTMA analysis. The objective of this study aims to demonstrate the benefit of incorporating the flow-dependent error covariance in the hybrid scheme to produce a well-represented analysis.

In order to examine the surface analysis, an experiment was conducted for the period between May 5th, 2018 at 00z to May 18th, 2018 at 00z. From the study cases, the 2m temperature and 10m wind speed analyses were able to depict features of the weather system, such as advancing air masses and flows around low – pressure centers. In addition, the increment comparison shows error improvements of 1 – 5 K and 1 – 4 m/s between the analysis and the background in regions of weather systems such as frontal boundary, regions of precipitation and low - pressure centers. The high increment with error improvement along the weather system suggests the flow-dependent error covariance is able capture the flow of the day and spatially characterize the increment along weather systems. However, 10m wind speed increment illustrated a marginal increase in errors for some regions. Since surface wind speed and direction is highly variable in time, a shortened observation time window could be ameliorated these error

by allowing only the observations that are measured close to the analysis timestamp. In future work, the observation time window should be changed from $\pm$ 1.5 hours to $\pm$ 12 mins, as set in the configuration of the RTMA. Further findings indicated high increment with error improvement are seen in regions of contrasting terrain in the Rockies and Appalachian Mountain, which suggest the flow-dependent error covariance it able to capture temperature variability and surface wind features that are dependent on flow over terrains. Lastly, when comparing the accuracy of the 2m temperature and 10m wind speed analysis difference between the hybrid and variational scheme, the hybrid analysis has a smaller absolute error by 1 – 3 K and 0.5 - 1.5 m/s in the regions of weather systems and complex terrain.

From the surface analysis statistical comparison, the RMSE time series for 2m temperature, 10m wind speed and 2m specific humidity indicated the hybrid scheme had a lower RMSE than the background and the variational scheme but had a higher RMSE than the RTMA analysis. However, the hybrid 10m wind speed analysis marginally outperformed the RTMA for 28% of the time. Further comparison depicts the improvement percentage for 2m temperature in the hybrid is 13.46 %, which was greater than the variational schemes at 10.51 %. Meanwhile, the 10m wind speed comparison showed the hybrid improvement percentage at 19.67% exceeded the variational scheme at 14.61%. Also, the improvement percentage of the hybrid scheme marginally surpass the variational scheme by ~1% for the 2m specific humidity. The comparison of 10m wind speed in this study with the RTMA results by De Pondeca (2011) suggested the improvement percentage between the two experiments are similar. Therefore, the weak terrain – following constraint of the background error covariance in RTMA and the flow-dependent error covariance in the hybrid scheme produced similar statistics on the accuracy of

surface wind speed, assuming the observation error covariance used in the RTMA and hybrid data assimilation were the same. In future work, one can conduct a 2D VAR experiment with terrain – following background error covariance and uses the same background, observations and configuration as this study. This allows for a more direct comparison of effect from using the terrain-following and flow-dependent background error covariance in the RTMA and hybrid scheme.

The upper – level analysis statistical comparison showed the temperature and wind speed RMSE for the hybrid scheme is lower than the VAR and BG at 1000 hPa, 925 hPa and 850 hPa pressure levels. The two schemes have similar RMSE values 500hPa, while the hybrid scheme has a higher RMSE than VAR for the upper – level at 200 hPa. On the other hand, specific humidity RMSE comparison shows variational scheme incrementally outperformed the hybrid scheme at 1000 hPa and 925 hPa, while the two schemes have similar RMSE values at 850 hPa through 200 hPa.

Next, a series of sensitivity tests on the vertical localization were examined to study the change in the performance of surface and upper – level analysis. Four experiments with vertical localization set to 3, 6, 9 and 12 vertical grid units were conducted for a 4 – day period between May 9th at 00z to May 13th at 00z. The surface analysis results show the 2m temperature RMSE increased from 1.56850 K with 14.801 % improvement to 1.56882 K with 14.777% improvement when the vertical localization increased from 3 to 12 grid units. Meanwhile, the wind speed RMSE increased by ~ 0.00067 m/s and the improvement percentage decreased from 21.470 % to 21.416. As a result, there is minimal impact on the statistical performance of the

surface analysis by increasing the vertical localization from 3 to 12 grid units. Similarly, an insignificant impact was found for the upper – level analysis. The results for other literature recommended a larger horizontal and vertical localization to be appropriate for upper – level analysis, which can be implemented in future work.

Finally, the benefit of assimilating satellite radiance using hybrid data assimilation was considered. The performance of the radiance bias correction on removing the systematic bias from the satellite instruments was studied. The results indicated the bias correction procedure have successfully removed the biases and improved the RMSE from 20 out of 52 channels. However, the temperature and wind speed RMSE profiles are higher when satellite observations were assimilated, especially above 850 hPa. Improvement on the bias correction is essential to effectively assimilate satellite radiance without introducing additional errors. Therefore, further enhancement on the bias correction procedure can be made in future work by introducing well-fitted initialized radiance bias correction coefficients and also having a more extended spin-up period to allow the bias correction coefficient to converge.

6 References:

Ancell, Brian C., Clifford F. Mass, Kirby Cook, and Brad Colman. 2014. “Comparison of Surface Wind and Temperature Analyses from an Ensemble Kalman Filter and the NWS Real-Time Mesoscale Analysis System.” Weather and Forecasting 29(4): 1058–75. http://journals.ametsoc.org/doi/abs/10.1175/WAF-D-13-00139.1.

Buehner, M. (2005). Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Quarterly Journal of the Royal Meteorological Society*, *131*(607), 1013–1043. https://doi.org/10.1256/qj.04.15

Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L., & He, B. (2010). Intercomparison of Variational Data Assimilation and the Ensemble Kalman Filter for Global Deterministic NWP. Part II: One-Month Experiments with Real Observations. *Monthly Weather Review*, *138*(5), 1567–1586. https://doi.org/10.1175/2009MWR3158.1

Chen, Y., & Snyder, C. (2006). Assimilating Vortex Position with an Ensemble Kalman Filter. *Monthly Weather Review*, *135*(5), 1828–1845. https://doi.org/10.1175/MWR3351.1

De Pondeca, M. S. F. V., Manikin, G. S., DiMego, G., Benjamin, S. G., Parrish, D. F., Purser, R. J., … Vavra, J. (2011). The Real-Time Mesoscale Analysis at NOAA’s National Centers for Environmental Prediction: Current Status and Development. *Weather and Forecasting*, *26*(5), 593–612. https://doi.org/10.1175/WAF-D-10-05037.1

Derber, J. C., Wu, W. (1998). The use of TOVS level-1b radiances in the NCEP SSI analysis system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 689–724. https://doi.org/10.1002/qj.49712656315

Gaspari, G., & Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757. https://doi.org/10.1256/smsqj.55416

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hamill, T. M., & Snyder, C. (2000). A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. *Monthly Weather Review*, *128*(8), 2905–2919. https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2

Harris, B. A., & Kelly, G. (2001). Satellite Radiance-Bias Correction Scheme for data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 1453–1468.

Hayden, C. M., & Purser, R. J. (1995). Recursive Filter Objective Analysis of Meteorological Fields: Applications to NESDIS Operational Processing. *Journal of Applied Meteorology*. https://doi.org/10.1175/1520-0450-34.1.3

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., & Hansen, B. (2005). Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Monthly Weather Review*, *133*, 604–620. https://doi.org/10.1175/MWR-2864.1

Houtekamer, P. L., & Zhang, F. (2016). Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. *Monthly Weather Review*, *144*(12), 4489–4532. https://doi.org/10.1175/MWR-D-15-0440.1

John, V. O., Holl, G., Buehler, S. A., Candy, B., Saunders, R. W., & Parker, D. E. (2012). Understanding intersatellite biases of microwave humidity sounders using global simultaneous nadir overpasses. *Journal of Geophysical Research Atmospheres*, *117*(2). https://doi.org/10.1029/2011JD016349

Karbou, F., Aires, F., Prigent, C., & Eymard, L. (2005). Potential of advanced microwave sounding unit-A (AMSU-A) and AMSU-B measurements for atmospheric temperature and humidity profiling over land. *Journal of Geophysical Research Atmospheres*, *110*(7), 1–16. https://doi.org/10.1029/2004JD005318

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L., Barker, D. M., Bray, J. R., Clayton, A. M., Dalby, T. , Li, D. , Payne, T. J. and Saunders, F. W. (2000), The Met. Office global three-dimensional variational data assimilation scheme. Q.J.R. Meteorol. Soc., 126: 2991-3012. doi:10.1002/qj.49712657002

Pan, Y., Zhu, K., Xue, M., Wang, X., Hu, M., Benjamin, S. G., … Whitaker, J. S. (2014). A GSI-Based Coupled EnSRF–En3DVar Hybrid Data Assimilation System for the Operational Rapid Refresh Model: Tests at a Reduced Resolution. *Monthly Weather Review*, *142*(10), 3756–3780. https://doi.org/10.1175/MWR-D-13-00242.1

Shao, H., J. Derber, X.-Y. Huang, M. Hu, K. Newman, D. Stark, M. Lueken, C. Zhou, L. Nance, Y.-H. Kuo, B. Brown, 2016: Bridging Research to Operations Transitions: Status and Plans of Community GSI. Bulletin of the American Meteorological Society, doi:10.1175/BAMS-D-13-00245.1, in press

Wang, X., Snyder, C., & Hamill, T. M. (2007a). On the Theoretical Equivalence of Differently Proposed Ensemble–3DVAR Hybrid Analysis Schemes. *Monthly Weather Review*, *135*(1), 222–227. https://doi.org/10.1175/MWR3282.1

Wang, X., Hamill, T. M., Whitaker, J. S., & Bishop, C. H. (2007b). A Comparison of Hybrid Ensemble Transform Kalman Filter–Optimum Interpolation and Ensemble Square Root Filter Analysis Schemes. *Monthly Weather Review*, *135*(3), 1055–1076. https://doi.org/10.1175/MWR3307.1

Wang, X., Barker, D. M., Snyder, C., & Hamill, T. M. (2008a). A Hybrid ETKF–3DVAR Data Assimilation Scheme for the WRF Model. Part I: Observing System Simulation Experiment. *Monthly Weather Review*, *136*(12), 5116–5131. https://doi.org/10.1175/2008MWR2444.1

Wang, X., Barker, D. M., Snyder, C., & Hamill, T. M. (2008b). A Hybrid ETKF–3DVAR Data Assimilation Scheme for the WRF Model. Part II: Real Observation Experiments. *Monthly Weather Review*, *136*(12), 5132–5147. https://doi.org/10.1175/2008MWR2445.1

Wang, X. (2010). Incorporating Ensemble Covariance in the Gridpoint Statistical Interpolation Variational Minimization: A Mathematical Framework. *Monthly Weather Review*, *138*(7), 2990–2995. https://doi.org/10.1175/2010MWR3245.1

Wang, X., Parrish, D., Kleist, D., & Whitaker, J. (2013). GSI 3DVar-Based Ensemble–Variational Hybrid Data Assimilation for NCEP Global Forecast System: Single-Resolution Experiments. *Monthly Weather Review*, *141*(11), 4098–4117. https://doi.org/10.1175/MWR-D-12-00141.1

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y., & Toth, Z. (2008). Ensemble Data Assimilation with the NCEP Global Forecast System. *Monthly Weather Review*, *136*(2), 463–482. https://doi.org/10.1175/2007MWR2018.1

Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., & Wobus, R. (2017). Performance of the New NCEP Global Ensemble Forecast System in a Parallel Experiment. *Weather and Forecasting*, WAF-D-17-0023.1. https://doi.org/10.1175/WAF-D-17-0023.1

Zhu, Y., Derber, J., Collard, A., Dee, D., Treadon, R., Gayno, G., & Jung, J. A. (2014). Enhanced radiance bias correction in the National Centers for Environmental Prediction's Gridpoint Statistical Interpolation data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *140*(682), 1479–1492. https://doi.org/10.1002/qj.2233

7 Appendix: Table of Acronyms

Acronym	Meaning
EnKF	Ensemble Kalman Filter – *Data assimilation technique*
VAR	Variational – *Data assimilation technique*
GSI	Gridpoint Statistical Interpolation System – *The data assimilation system that was used to run the variational and hybrid data assimilation experiments*
WRF	Weather Research and Forecasting Model
HRRR	High Resolution Rapid Refresh NWP Model – *Ingested as the background in the experiments*
GEFS	Global Ensemble Forecast System NWP Model– *Ingested as the ensemble members data set in the experiments*
GDAS	Global Data Assimilation System – *Observations from the system are taken to be ingested in the experiments*
RAP	Rapid Refresh NWP Model
RUC	Rapid Update Cycle NWP Model
RTMA	Real – Time Mesoscale Analysis
ERBC	Enhanced Radiance Bias Correction
EBC	Radiance Bias Correction
EMC	Environmental Modeling Center
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
GSD	Global Systems Division
NWS	National Weather Services
TWN	The Weather Network
RMSE	Root Mean Square Error
MBE	Mean Bias Error

***Table 4.4.1**: The list of Acronyms that were used in the thesis*