The Pennsylvania State University

The Graduate School

IMPROVING THE ANALYSIS AND FORECAST OF TROPICAL MESOSCALE CONVECTIVE SYSTEMS THROUGH ADVANCING THE ENSEMBLE DATA ASSIMILATION OF GEOSTATIONARY SATELLITE INFRARED RADIANCE OBSERVATIONS

A Dissertation in

Meteorology and Atmospheric Science

by

Man Yau Chan

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

October 2022

The dissertation of Man Yau Chan was reviewed and approved[1] by the following:

Xingchao Chen
Assistant Professor of Meteorology
Dissertation Advisor
Chair of Committee

David J. Stensrud
Department Head and Professor of Meteorology

Eugene E. Clothiaux
Professor of Meteorology

Helen Greatrex
Assistant Professor of Geography

Jeffrey L. Anderson
Data Assimilation Research Section Head
National Center for Atmospheric Research
Special Member

[1] Signatures are on file in the Graduate School

ABSTRACT

The advent of modern geostationary satellite infrared (IR) imagery has ushered in a new era in atmospheric observations. In recent years, tremendous progress has been made towards improving the forecasts of tropical cyclones and mid-latitude weather by the ensemble data assimilation (DA) of these IR observations. However, there is comparatively little work on assimilating these observations to improve the analysis and forecast of tropcial mesoscale convective systems (MCSs). Because these systems are the primary source of rainfall over the Tropics and influence global weather and climate, it is essential to accurately predict these systems. As such, the ultimate goal of my dissertation research is to improve the analysis and forecast of tropical MCSs through assimilating geostationary infrared radiance (GeoIR) observations.

My dissertation research has two aspects: 1) demonstrating the impacts of GeoIR ensemble DA on the analysis and forecast of tropical MCSs, and 2) devising a new computationally efficient ensemble DA algorithm to better exploit GeoIR observations. To demonstrate aspect 1, I started with assimilating real GeoIR water vapor channel brightness temperature (WV-BT) observations from the Himawari-8 geostationary satellite in a tropical squall line case over the Sumatra-Malaysia-Borneo region. The ensemble DA was executed through the state-of-the-art Pennsylvania State University Ensemble Kalman Filter (PSU-EnKF) system. Compared to a control experiment where only in-situ and satellite-derived atmospheric motion vectors (AMVs) were assimilated, the inclusion of WV-BT observations on top of the control experiment's observations improved the analysis and forecast of the tropical squall line. Specifically, the analysed squall line outflow positions and the analysis and forecast of the squall line's clouds were improved. These improvements scale with the frequency of GeoIR DA.

Since no MCS-resolving tropical reanalysis product currently exist over the

Maritime Continent and GeoIR DA has demonstrated positive impacts in my tropical squall case, I proceeded to create a high-resolution tropical MCS reanalysis (TMeCSR). This was done through combining an ensemble of MCS-resolving model runs (9-km horizontal grid spacing), in-situ observations, AMV observations and WV-BT observations through the PSU-EnKF. To further enhance the TMeCSR, I also introduced large-scale information from the European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis version 5 (ERA5). The TMeCSR is available during June, July and August of 2017, and spans a region covering most of the tropical Indian Ocean, most of continental Asia, the Maritime Continent, and the West Pacific.

Comparisons of TMeCSR and ERA5 against independent satellite retrievals indicate that TMeCSR's cloud and multiscale rain fields are better than those of ERA5. Furthermore, TMeCSR better captured the diurnal variability of rainfall and the statistical characteristics of MCSs. Forecasts initialized from TMeCSR also have more accurate rain and clouds than those initialized from ERA5. The TMeCSR and ERA5 forecasts have similar performances with respect to sounding and surface observations. These results indicate that TMeCSR is a promising MCS-resolving dataset for tropical MCS studies.

The second aspect of my dissertation research is to create a new computationally efficient ensemble DA algorithm: the bi-Gaussian ensemble Kalman filter (BGEnKF). The BGEnKF is motivated by the differences in the atmospheric dynamics, and thus the statistics, of clear atmospheric columns and cloudy atmospheric columns. Unlike the EnKF, which does not distinguish the differences between these two types of columns, the BGEnKF explicitly treats clear atmospheric columns separately from cloudy atmospheric columns. Furthermore, unlike earlier formulations in the literature, my formulation is computationally efficient and does not require laborous derivations and programming concerning stochastic subspaces.

To examine the advantages of my BGEnKF over the EnKF, I implemented the BGEnKF into the PSU-EnKF system and performed observing systems simulation experiments (OSSEs) using a case of tropical convection over the equatorial Indian Ocean. This case occurred during the onset of the October 2011 Madden-Julian Oscillation event. Only synthetic infrared window channel brightness temperatures (Window-BT) from the Meteorological Satellite 7 were assimilated. My results indicate that the BGEnKF outperformed the EnKF in this semi-idealized setting. These performance advantages were found in the horizontal wind vector components, temperature, specific humidity and WV-BT fields.

My OSSE tests with the BGEnKF are among the first to test a Gaussian mixture model EnKF (GMM-EnKF) with a realistic weather model. The encouraging results motivate further testing and development of the BGEnKF algorithm. If the BGEnKF consistently outperforms the EnKF at assimilating GeoIR observations, the BGEnKF might replace the EnKF in operational weather forecasting systems in the future.

Contents

List of Figures

Preface

The material in this dissertation is a combination of my first-authored published works and my exploration of ensemble DA theory. These published works are based on studies that I did during my tenure as a PhD candidate at the Department of Meteorology and Atmospheric Science, the Pennsylvania State University. I am the primary investigator in all of these published works because I 1) conceptualized and performed the experiments, 2) analyzed the results, and 3) drafted the manuscripts.

The list of my first-authored published works used in this dissertation is as follows.

- Chapter 4 is based on my 2021 publication in the Advances in Atmospheric Science [AAS; Chan and Chen (2021)].
- Chapter 5 is based on my accepted 2022 publication in the Journal of Advances in Modelling Earth Systems [JAMES; **CITATION COMING SOON**].
- Chapters 7 and 8 are based on a combination of my 2020 publication in the Monthly Weather Review [MWR; Chan et al. (2020a)] and a manuscript that is currently under review in JAMES. A preprint of the latter manuscript is available on the Earth and Space Science Open Access Archive [ESSOAR; Chan et al. (2022)].

Acknowledgements

This dissertation would be impossible without support from various sources. In particular, I would like to thank my advisor, Xingchao (XC) Chen, for his mentorship, advice, collaboration, and for the freedom to explore various aspects of ensemble DA. Without XC, I might not have gone as far in testing and developing GeoIR DA and the BGEnKF.

I am also grateful to Fuqing for taking me on as a PhD student back in 2017 and for providing various opportunities. Fuqing introduced me to the world of ensemble DA and GeoIR DA. The work on the BGEnKF was made possible by Fuqing introducing me to Jeffrey L. Anderson (Jeff). Fuqing's jovial personality, wisdom and keen insight are dearly missed.

Jeff's support and advice are also invaluable for my dissertation, particularly regarding the BGEnKF. Without his input during the algorithm development phases, I might have stopped at a far less satisfying version of the BGEnKF. I am eternally grateful for his insights and experiences on ensemble DA algorithm development, and for the opportunity to collaborate (and continue to collaborate) with him.

I would also like to thank my committee and the Department of Meteorology and Atmospheric Science for supporting my dissertation work and for their inputs. Their support and input are invaluable for my career development, science and for ensuring that I do not get tangled up in paperwork. Special kudos

to Karen Corl for her help with paperwork relating to my time at Penn State.

Most importantly, I would like to thank my husband, Christopher Hartman for his love, companionship, support and help, particularly when graduate school is stressing me out.

Finally, I would like to thank the following for funding my research and for the computational resources. This research is funded by the U.S. Department of Energy Office of Science Biological and Environmental Research, as part of the Regional and Global Model Analysis program area through the Water Cycle and Climate Extremes Modelling (WACCEM) scientific focus area. I would also like to thank Office of Naval Research (ONR) Grant N00014-18-1-2517, the Graduate School Fellowship from The Pennsylvania State University Graduate School, the National Science Foundation (Award 1712290), and the Advanced Study Program at the National Center for Atmospheric Research (NCAR) for funding my work. NCAR is sponsored by the National Science Foundation and any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. The experiments and analysis in this dissertation were done on the Stampede2 supercomputer at the Texas Advanced Computing Center and the Cori supercomputer at the National Energy Research Scientific Computing Center (NERSC).

Chapter 1

Introduction

1.1 Background

Tropical mesoscale convective systems (MCSs) are important. They are not only responsible for more than 60% of the total precipitation in the Tropics (Mohr et al., 1999; Houze, 2004; Nesbitt et al., 2006; Liu, 2011; Roca and Fiolleau, 2020; Chen et al., 2022), but also influence global weather through convective momentum transport Houze (1973); Zhang and McFarlane (1995); Badlan et al. (2017) and latent heating effects (Madden and Julian, 1971, 1972; Held and Hou, 1980; Hoskins and Karoly, 1981; Weickmann, 1983; Satoh, 1994; Wheeler and Kiladis, 1999; Johnson and Ciesielski, 2013). Furthermore, recent studies have demonstrated that tropical MCSs play important roles in the onset of monsoons and in the monsoonal overturning circulation (Chen et al., 2018b, 2021a,b). These monsoonal roles establish the significance of tropical MCSs in the regional and global climate (i.e., beyond weather timescales). It is thus important to study and to improve forecasts of tropical MCSs.

Numerical weather prediction (NWP) models and observations are important tools for studying and forecasting tropical MCSs. However, both tools have their drawbacks: observational information tends to be incomplete and

model information is error-prone. Observational information is incomplete because majority of the tropical MCSs occur over the ocean (Huang et al., 2018). Even if the sparse in-situ observations (Ingleby, 2017) were combined with frequently available (>8 scans / day) and dense [>1 pixel / (50 km)^2] satellite observations, the resulting observation data does not have sufficient information about the 3D characteristics of tropical MCSs. On the other hand, NWP models can only imperfectly represent atmospheric processes, are susceptible to errors in the initial and boundary conditions, and tend to amplify both sources of error. Improving the methods to address these drawbacks can thus accelerate tropical MCS research.

Ensemble data assimilation (DA) methods mitigate the drawbacks of both information sources through combining them. Specifically, ensemble DA combines current observations with an ensemble of NWP forecasts (henceforth, the forecast ensemble) from an earlier time point. This combination process is done by converting the observations into an ensemble of adjustments for the forecast ensemble. The conversion is based on the statistical relationships[1] linking the observed quantities to all model quantities (Kalnay, 2003; Fletcher, 2017a). If done appropriately, the model quantity adjustments will reduce the errors in the model information. As such, ensemble DA is an indispensible component of operational NWP systems (Geer et al., 2018).

The usefulness of ensemble DA for studying and forecasting tropical MCSs depends, in part, on the spatiotemporal availability of observations assimilated (Liu and Rabier, 2002; Benjamin et al., 2004; Ying et al., 2018; Ying and Zhang, 2018). One of the most frequently, densely and widely available tropical MCS observation sources is geostationary satellite infrared radiance (GeoIR) observations [>1 pixel/ hr / (10 km)^2]. Unfortunately, in operational NWP systems that cover the Tropics, only 1 GeoIR observation is assimilated at every ∼125 km [*e.g.*, ECMWF (2016) and Geer et al. (2018)]. This observation density

[1]These statistical relationships estimated from the forecast ensemble.

is too low to adequately sample the small convective cells small convective cells (< 100 km across) that comprise tropical MCSs. As such, the assimilation of denser GeoIR observations can potentially improve the accuracy of tropical MCSs resolved by the datasets produced by ensemble DA (henceforth, the analyses).

1.2 Research motivations

At the time of writing, most of the research on assimilating GeoIR observations into convection-permitting models was done in the context of mid-latitude weather systems and tropical cyclones. Research in these contexts have found that the addition of IR observations improved the analyzed thermodynamic and cloud fields (Vukicevic et al., 2004, 2006; Otkin, 2010, 2012; Zhang et al., 2016, 2018; Honda et al., 2018a; Minamide and Zhang, 2019; Sawada et al., 2019; Jones et al., 2020; Zhang et al., 2021b,a; Hartman et al., 2021). In contrast, relatively little work has been done in the context of tropical MCSs. Nonetheless, the pattern of improvement seen in the context of convection-permitting mid-latitude severe weather and tropical cyclone prediction should also be achievable in the context of tropical MCSs.

Ying and Zhang (2018) was among the first to examine the potential impacts of assimilating GeoIR observations on tropical MCSs. Using ensembles of gray-zone resolution simulations, Ying and Zhang (2018) performed idealized observing system simulation experiments (OSSEs) over the tropical Indian Ocean during the active phase of the October 2011 Madden-Julian Oscillation (MJO; Madden and Julian (1971, 1972)) event. Gray-zone resolution simulations are simulations that use horizontal grid spacing fine enough to resolve mesoscale convective systems, but too coarse to resolve individual convective updrafts (Wang et al., 2015; Chen et al., 2018c). Ying and Zhang (2018)

found that the inclusion of GeoIR observations can improve the analyzed cloud hydrometeor fields. Chan et al. (2020b) subsequently performed real data DA experiments for a similar time period and domain and found that the assimilation of real GeoIR observations can potentially improve cloud field analyses and predictions across a range of spatial scales, as well as short-term rainfall predictions. These improvements seen in Ying and Zhang (2018) and Chan et al. (2020b) are consistent with the improvements seen in the context of mid-latitude weather and tropical cyclone convection-permitting numerical weather prediction.

However, several important questions remain unanswered regarding the assimilation of GeoIR observations in the context of tropical MCSs. First, how does the assimilation of real-world GeoIR observations improve the thermodynamic and dynamic fields of tropical MCSs? Second, how does the accuracy of the analyses and predictions of tropical MCSs vary with the frequency of GeoIR DA? Furthermore, can GeoIR observations be used to construct a high resolution tropical MCS reanalysis dataset for the tropical MCS research and forecast community? Addressing these questions can potentially inform the future assimilation of GeoIR observations in the Tropics and improve the analyses and forecasts of tropical MCSs.

The appropriateness of the ensemble DA algorithm employed to assimilate GeoIR observations is another important research question. If the algorithm is inappropriate, changing the algorithm can potentially improve the accuracy of analyzed tropical MCSs.

A common assumption in popular ensemble DA algorithms is that the forecast ensemble is drawn from a Gaussian distribution (henceforth, Gaussian ensemble DA methods). This assumption is problematic in situations where the forecast ensemble is uncertain about the absence/presence of clouds (*i.e.*, some ensemble members are clear and others are cloudy; henceforth, mixed

forecast ensemble). Such situations frequently occur. Because the dynamical, thermodynamical and radiative characteristics of clear atmospheric columns are different from those of cloudy atmospheric columns, the statistical relationships connecting model quantities can vary depending on the absence/presence of clouds. In other words, the statistical relationships are nonlienar functions of the model variables. Since Gaussian distributions only permit linear statistical relationships, Gaussian ensemble DA methods are supoptimal for handling mixed ensemble forecasts. It is thus necessary to develop ensemble DA methods that can handle mixed forecast ensembles.

The ultimate goal of my PhD research is to advance the ensemble DA of GeoIR observations for tropical MCSs. There are two aspects to my research. The first aspect is to demonstrate the impacts of assimilating moderate amounts of GeoIR observations [$\sim$1 observation/(27 km)^2] for analyses and forecasts of tropical MCSs. These demonstrations are done using a popular Gaussian ensemble DA method: the ensemble Kalman filter (EnKF). The second aspect is to extend the EnKF to handle mixed forecast ensembles. Specifically, I have constructed a computationally efficient bi-Gaussian extension of the EnKF (the BGEnKF) and tested it with a toy model and a realistic weather model under idealized conditions.

1.3 Dissertation overview

This dissertation is broken into two parts. The first part focuses on the demonstrative aspect of my PhD research. Chapter 2 covers general concepts in ensemble DA, the ensemble DA workflow, how ensemble DA fits into operational forecasting, as well as my notation. The specific form of the EnKF (the ensemble square-root filter, or the EnSRF) I use will be discussed in Chapter 3. In Chapter 4, I will demonstrate the potential benefits of assimilating

moderately thinned GeoIR observations via the EnSRF using a tropical squall line case study over the Maritime Continent. Galvanized by these findings, I then constructed a three-month tropical MCS reanalysis (TMeCSR) product by combining in-situ observations, satellite-derived wind observations, and GeoIR observations, with an MCS-resolving weather model. The construction and performance of the TMeCSR are discussed in Chapter 5.

The BGEnKF aspect of my research is addressed in the second part of this dissertation. I will first show empirical evidence that clear and cloudy forecast members should be treated separately in Chapter 6. The BGEnKF will then be discussed in Chapter 7. Following that, the advantage of using the BGEnKF over the EnSRF will be demonstrated in Chapter 8 via observing system simulation experiments (OSSEs) of a case of tropical convection over the Indian Ocean. This dissertation will then conclude in Chapter 9.

Chapter 2

General DA concepts and notation

2.1 Overview

Before delving into the details of the EnKF, the BGEnKF, and GeoIR DA, some general concepts in DA and my notation should be discussed. This chapter will cover the essence of DA (Bayesian inference), the typical workflow of ensemble DA systems, and define the joint state vectors. The last item is heavily used throughout this dissertation.

2.2 DA and Bayesian inference

DA is a class of Bayesian inference methods that attempts to infer the current state of the atmosphere by combining prior information (usually from forecasts initialized at a previous time) with observation information. This combination accounts for uncertainties in either information and in the inferred state through using probability density functions (pdf). Supposing vector $\boldsymbol{\psi}$ represents an arbitrary state of the atmosphere and vector $\mathbf{y^o}$ contains the observations, DA combines the prior and observation information via Bayes'

rule

$$p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) = \frac{p\left(\boldsymbol{\psi}\right) p\left(\mathbf{y^o}|\boldsymbol{\psi}\right)}{p\left(\mathbf{y^o}\right)}. \tag{2.1}$$

Here, the posterior pdf $p\left(\boldsymbol{\psi}|\mathbf{y^o}\right)$ represents the inferred information and its uncertainty, and the prior pdf $p\left(\boldsymbol{\psi}\right)$ represents the prior information and its uncertainty. The observation information is introduced through the observation likelihood $p\left(\mathbf{y^o}|\boldsymbol{\psi}\right)$ (*i.e.*, the probability of obtaining $\mathbf{y^o}$ for every $\boldsymbol{\psi}$). Finally, the marginal $p\left(\mathbf{y^o}\right)$ normalizes the posterior pdf.

Bayes' rule computes the probability of every $\boldsymbol{\psi}$, given the available prior and observation information. For example, consider a $\boldsymbol{\psi}$ that has a cloud over State College, PA. Suppose forecasts initialized from a previous time suggest a high probability of cloudiness, but precise satellite observations are clear. Since the considered $\boldsymbol{\psi}$ agrees with the forecasts, $p\left(\boldsymbol{\psi}\right)$ is a large probability. However, $p\left(\mathbf{y^o}|\boldsymbol{\psi}\right)$ is small because the considered $\boldsymbol{\psi}$ disagrees with the precise observations. Bayes' rule thus computes a small posterior probability to the considered $\boldsymbol{\psi}$ [specifically, $p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) < p\left(\boldsymbol{\psi}\right)$]. To construct the posterior pdf for every $\boldsymbol{\psi}$, this computation can be naïvely repeated for every $\boldsymbol{\psi}$ in $\mathbb{R}^{N_\psi}$ space (N_ψ is the number of elements in $\boldsymbol{\psi}$).

In practical DA, evaluations of Bayes' rule [Eq. (2.1)] for every $\boldsymbol{\psi}$ in $\mathbb{R}^{N_\psi}$ space (henceforth, naïve evaluations of Bayes' rule) is undesirable for several reasons.

1. There are infinite possible $\boldsymbol{\psi}$'s, meaning that it is costly to evaluate Bayes' rule for all possible $\boldsymbol{\psi}$.

2. $\boldsymbol{\psi}$ usually contains $\sim 10^9$ elements, meaning that $p\left(\boldsymbol{\psi}|\mathbf{y^o}\right)$ and $p\left(\boldsymbol{\psi}\right)$ are small values. Rounding errors due to finite precision computations can thoroughly corrupt these pdf values.

These naïve evaluations are avoided in practical DA by assuming that $p\left(\boldsymbol{\psi}\right)$, $p\left(\mathbf{y^o}|\boldsymbol{\psi}\right)$, and $p\left(\boldsymbol{\psi}|\mathbf{y^o}\right)$ can be written in terms of specific families of proba-

bility distributions. For instance, Gaussian DA algorithms assume that $p(\boldsymbol{\psi})$ and $p(\mathbf{y^o}|\boldsymbol{\psi})$ are multivariate Gaussian pdfs[1], meaning that $p(\boldsymbol{\psi}|\mathbf{y^o})$ is also a multivariate Gaussian pdf[2]. Gaussian pdfs are entirely characterized by their mean vector and covariance matrix. Gaussian DA methods thus merely need to either analytically determine or construct an ensemble representation of the mean $\boldsymbol{\psi}$ and covariance matrix of $p(\boldsymbol{\psi}|\mathbf{y^o})$ (henceforth, the posterior mean and covariance matrix). There is thus no need for naïve evaluations of Bayes' rule in Gaussian DA.

Naïve evaluations are also unnecessary in many DA algorithms that assume $p(\boldsymbol{\psi})$ and/or $p(\mathbf{y^o}|\boldsymbol{\psi})$ are prescribed non-Gaussian pdfs. In the Gaussian mixture model (GMM) ensemble Kalman filter (GMM-EnKF; will be discussed later), it is assumed that 1) $p(\boldsymbol{\psi})$ is a GMM pdf and 2) $p(\mathbf{y^o}|\boldsymbol{\psi})$ is a Gaussian pdf. These assumptions result in a GMM $p(\boldsymbol{\psi}|\mathbf{y^o})$. GMM pdfs are entirely characterized by the number of Gaussian kernels and each Gaussian kernel's weights, mean state and covariance matrix. As such, GMM-EnKFs merely need to construct ensemble representations of these characteristic parameters, thus avoiding the naïve evaluations of Bayes' rule.

Even in DA methods that allow non-parameteric pdfs like particle filters [PFs; *e.g.*, van Leeuwen et al. (2019)], rank histogram filters [RHFs; Anderson (2010, 2019, 2020)], and the quantile conserving ensemble filter [QCEF; Anderson (2022)], there is no need to naïvely evaluate Bayes' rule. The algorithms described in the referenced literature only require Bayes' rule to be evaluated for a small number of times (for N_E forecast ensemble members, $\sim N_E$ evaluations of Bayes' rule).

[1] also known as multivariate normal distribution pdf

[2] For mathematical proof that the multiplication of two Gaussian pdfs result in a scaled Gaussian pdf, see Chapter 3.9.3.

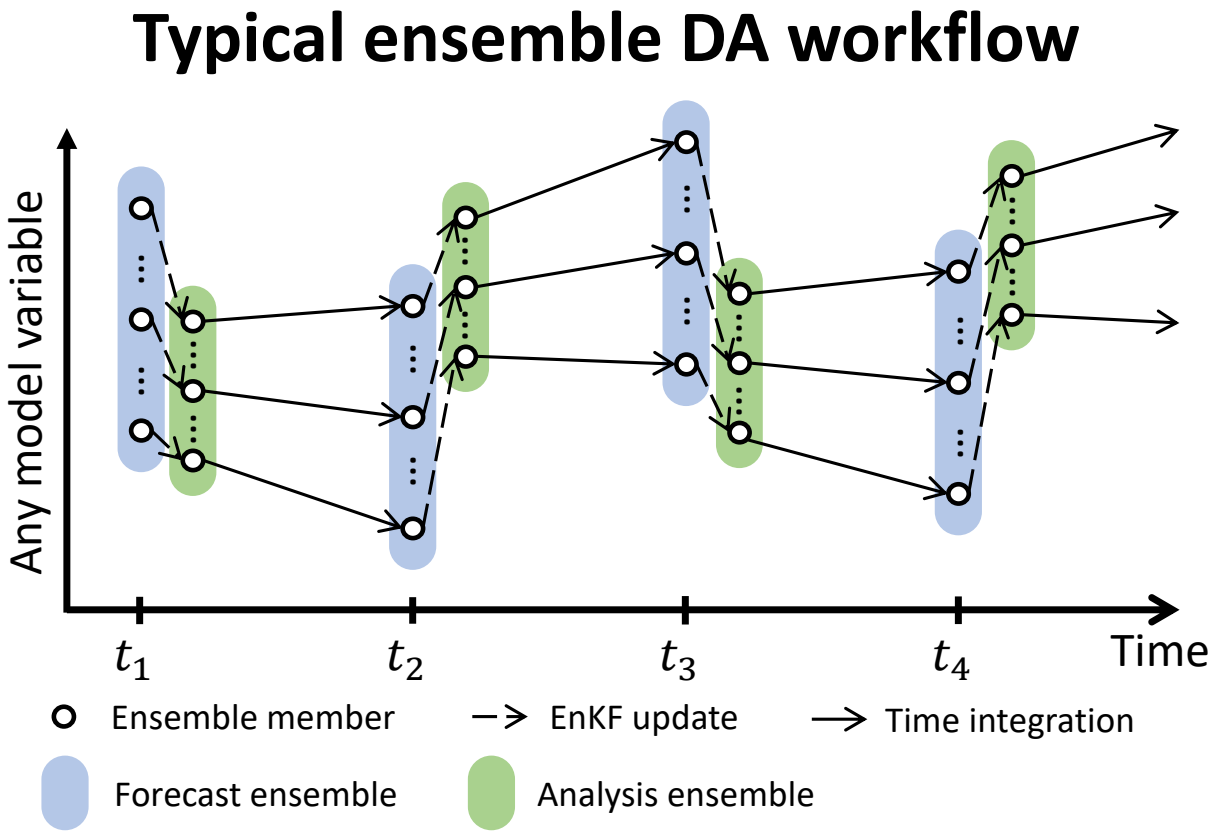


Figure 2.1: Typical workflow of an ensemble DA system.

2.3 Ensemble DA system workflow

Ensemble DA methods represent the prior pdf and posterior pdf using a Monte Carlo approach. Specifically, ensemble DA views the forecast ensemble as samples drawn from the prior distribution, and then constructs an analysis/posterior ensemble that follows the posterior distribution. The recipe to construct the analysis ensemble varies with the assumed forms of the prior pdf and observation likelihood. Later on in Chapters 3 and 7, I will discuss two ensemble DA algorithm in detail.

The typical workflow of an ensemble DA system is illustrated in Figure 2.1.

In general, ensemble DA systems have a temporal component because observations are valid at different points in time. Suppose we have an ensemble of forecasts at starting time t_1 (Figure 2.1; blue filled oval with white circles). Assimilating observations valid at t_1 into the forecast ensemble results in an analysis ensemble at t_1. This analysis ensemble is then integrated forward in time (via a forecast model; Figure 2.1) to produce a new forecast ensemble at time t_2. Observations valid at t_2 are then assimilated into this t_2 forecast ensemble to construct an analysis ensemble at t_2. The t_2 analysis ensemble is then integrated forward in time. This assimilate-then-integrate cycle can continue until all available observations are assimilated.

2.4 The joint state vector $\boldsymbol{\psi}$

It is necessary to explicitly define $\boldsymbol{\psi}$ before proceeding any further. My choice of atmospheric state vector $\boldsymbol{\psi}$ contains three components:

1. all forecast model variables (represented by the model state vector $\mathbf{x}$),
2. all observable quantities corresponding to $\mathbf{x}$ (represented by the observable state vector $\mathbf{y}$), and,
3. all auxiliary quantities needed to execute the DA procedure (represented by the auxiliary state vector $\boldsymbol{\xi}$).

As such,

$$\boldsymbol{\psi} \equiv \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \boldsymbol{\xi} \end{bmatrix}. \tag{2.2}$$

Suppose I am assimilating N_y observations in a situation where $\mathbf{x}$ contains N_x elements and $\boldsymbol{\xi}$ contains N_ξ elements. Then, $\mathbf{y}$ has N_y elements and $\boldsymbol{\psi}$ has $N_x + N_y + N_\xi \left(\equiv N_\psi\right)$ elements. Since $\boldsymbol{\psi}$ is constructed by concatenating or joining $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{\xi}$, I will sometimes refer to $\boldsymbol{\psi}$ as the joint state vector.

$\mathbf{y}$ can be constructed by applying the vector observation operator function $\mathscr{H}$ on $\mathbf{x}$. In other words,

$$\mathbf{y} \equiv \mathscr{H}(\mathbf{x}). \tag{2.3}$$

Furthermore, $\boldsymbol{\xi}$ can be constructed by applying a vector auxiliary operator function $\boldsymbol{\Xi}$ on $\mathbf{x}$, *i.e.*,

$$\boldsymbol{\xi} = \boldsymbol{\Xi}(\mathbf{x}). \tag{2.4}$$

As such, a forecast ensemble member's $\boldsymbol{\psi}$ (henceforth, $\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}}$) can be constructed from its forecast model state $\mathbf{x}_{\mathbf{n}}^{\mathbf{f}}$ via

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} \equiv \begin{bmatrix} \mathbf{x}_{\mathbf{n}}^{\mathbf{f}} \\ \mathscr{H}\left(\mathbf{x}_{\mathbf{n}}^{\mathbf{f}}\right) \\ \boldsymbol{\Xi}\left(\mathbf{x}_{\mathbf{n}}^{\mathbf{f}}\right) \end{bmatrix} \quad \forall\, n = 1, 2, \ldots, N_E \tag{2.5}$$

where N_E is the forecast ensemble size.

Finally, to extract the observable component from $\boldsymbol{\psi}$, we can define a linear operator $\mathbf{H}$ such that

$$\mathbf{y} = \mathbf{H}\,\boldsymbol{\psi}. \tag{2.6}$$

An appropriate definition of $\mathbf{H}$ is the following $N_y \times N_\psi$ matrix

$$\mathbf{H} \equiv \begin{bmatrix} \mathbf{0}_{\mathbf{N_y} \times \mathbf{N_x}} & \mathbf{I}_{\mathbf{N_y} \times \mathbf{N_y}} & \mathbf{0}_{\mathbf{N_y} \times \mathbf{N}_\xi} \end{bmatrix}. \tag{2.7}$$

Here, for arbitrary integers N_1 and N_2, $\mathbf{0}_{\mathbf{N_1} \times \mathbf{N_2}}$ is an $N_1 \times N_2$ matrix of zeros, and $\mathbf{I}_{\mathbf{N_1} \times \mathbf{N_1}}$ is an $N_1 \times N_1$ identity matrix.

Chapter 3

The ensemble square-root filter (EnSRF)

3.1 Overview

Ensemble Kalman filters (EnKF) are among the most popular ensemble DA methods in the geosciences (Keppenne et al., 2005; Reichle et al., 2009; Stammer et al., 2016; Edwards et al., 2015; Park and Xu, 2016; Helmert et al., 2018). A commonly used flavor of the EnKF is the ensemble square-root filter (EnSRF) proposed by Whitaker and Hamill (2002). Since all of the research in this dissertation utilizes the EnSRF, this chapter is dedicated to describing and explaining the EnSRF.

This chapter is structured in two layers: the conceptual layer and the mathematical layer. In the conceptual layer (the first layer), I will discuss the EnSRF at a conceptual level. The conceptual matters covered are:

1. the assumptions underlying the EnSRF,
2. a general outline of the EnSRF procedure,
3. the connection between the EnSRF and linear regression, and,

4. heuristic measures used to improve the EnSRF.

The mathematical (second) layer will cover:

1. a derivation of the posterior pdf and its parameters from first principles[1],
2. how an assumption in the EnSRF allows observations to be assimilated one at a time (*i.e.*, serial assimilation), and,
3. a construction of the EnSRF update procedure from the derived posterior pdf parameters.

For readability, some of the mathematical proofs and derivations have been consigned to this chapter's appendix (Chapter 3.9).

THE CONCEPTUAL LAYER

3.2 The EnSRF's assumptions

The EnSRF represents the combination of observation and forecast information through applying updates to an ensemble of forecasted atmospheric states $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}}\right\}$. These updates are derived from Bayesian inference (*i.e.*, Bayes' rule) assuming:

1. unbiased Gaussian forecast errors,
2. unbiased Gaussian observation errors, and,
3. all observations are made independently.

The forecast ensemble's assumed pdf is thus a Gaussian pdf that is completely characterized by two statistical parameters: the forecast mean state

[1] Note: this construction accounts for the rank deficiency of the prior covariance matrix

$\overline{\boldsymbol{\psi^f}}$, and the forecast covariance matrix $\boldsymbol{P^f_\psi}$. These are estimated via

$$\overline{\boldsymbol{\psi^f}} \equiv \frac{1}{N_E}\sum_{n=1}^{N_E} \boldsymbol{\psi^f_n} \quad \text{and} \quad \boldsymbol{P^f_\psi} \equiv \frac{1}{N_E - 1}\sum_{n=1}^{N_E} \boldsymbol{\psi^{f'}_n}\left(\boldsymbol{\psi^{f'}_n}\right)^\top, \tag{3.1}$$

where

$$\boldsymbol{\psi^{f'}_n} \equiv \boldsymbol{\psi^f_n} - \overline{\boldsymbol{\psi^f}} \quad \forall \quad n = 1, 2, \ldots, N_E. \tag{3.2}$$

The next two assumptions concern the observations. The unbiased Gaussian observation error assumption means that the likelihood function for a Gaussian pdf with mean $\mathbf{y^o}$ and covariance matrix $\boldsymbol{R}$.[2] Imposing the independent observation assumption means that 1) $\boldsymbol{R}$ is a diagonal matrix, and 2) observations can be assimilated one at a time (also known as serial assimilation)[3].

3.3 The EnSRF update procedure

As a result of the three assumptions, the EnSRF serially assimilates observations via the following procedure (Whitaker and Hamill, 2002)[4]:

1. Construct $\left\{\boldsymbol{\psi^f_1}, \boldsymbol{\psi^f_2}, \ldots, \boldsymbol{\psi^f_{N_E}}\right\}$ from the forecast ensemble of model state vectors $\left\{\mathbf{x^f_1}, \mathbf{x^f_2}, \ldots, \mathbf{x^f_{N_E}}\right\}$ by evaluating Eq. (2.5).
2. For $m = 1, 2, \ldots, N_y$,
   (a) Construct $\left\{\boldsymbol{\psi^a_1}, \boldsymbol{\psi^a_2}, \ldots, \boldsymbol{\psi^a_{N_E}}\right\}$ by assimilating the m-th observation into $\left\{\boldsymbol{\psi^f_1}, \boldsymbol{\psi^f_2}, \ldots, \boldsymbol{\psi^f_{N_E}}\right\}$ via the EnSRF update equation (see later).

[2] $\boldsymbol{R}$ is usually inputted by the user based on their knowledge of instrument errors.

[3] See Chapter 3.7 for an explanation on why serial assimilation is permitted for independent observations

[4] See Chapters 3.6 to 3.8 for the construction of the EnSRF procedure from the three assumptions.

(b) For $n = 1, 2, \ldots, N_E$, $\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} \leftarrow \boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n}}$.

3. Exit.

Note that step 2b means that we replace the forecast ensemble with the analysis ensemble.

The EnSRF update equation to assimilate the m-th observation y^o_m with observation error variance $\left(\sigma^o_m\right)^2$ into ensemble member n is:

$$\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n}} = \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} + \frac{\mathbf{c}_{\boldsymbol{\psi}, y^f_m}}{\left(\sigma^f_m\right)^2 + \left(\sigma^o_m\right)^2}\left[y^o_m - \overline{y^f_m} - \phi_m\left(y^f_{m,n} - \overline{y^f_m}\right)\right] \tag{3.3}$$

where $y^f_{m,n}$ is the m-th forecasted observation for member n, and is the $(N_x + m)$-th element of $\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}}$. Furthermore, $\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}$ is the forecast ensemble mean state, $\overline{y^f_m}$ is the forecasted ensemble mean observation, $\left(\sigma^f_m\right)^2$ is the ensemble variance forecasted observation, $\mathbf{c}_{\boldsymbol{\psi}, y^f_m}$ is an N_ψ-dimensional vector of forecast ensemble covariances linking $\boldsymbol{\psi}$ and y^f_m, and ϕ_m is the scalar factor proposed by Whitaker and Hamill (2002). These ensemble-related quantities are computed via:

$$\overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \equiv \frac{1}{N_E}\sum_{n=1}^{N_E} \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} \ , \quad \overline{y^f_m} \equiv \frac{1}{N_E}\sum_{n=1}^{N_E} y^f_{m,n} \ , \quad \left(\sigma^f_m\right)^2 \equiv \frac{1}{N_E - 1}\sum_{n=1}^{N_E}\left(y^f_{m,n} - \overline{y^f_m}\right)^2,$$

$$\phi_m \equiv \left(1 + \sqrt{\frac{\left(\sigma^o_m\right)^2}{\left(\sigma^f_m\right)^2 + \left(\sigma^o_m\right)^2}}\right)^{-1}, \text{ and,}$$

$$\mathbf{c}^{\mathbf{f}}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y^f_m} \equiv \frac{1}{N_E - 1}\sum_{n=1}^{N_E}\left(\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} - \overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)\left(y^f_{m,n} - \overline{y^f_m}\right).$$

Note that $\frac{\mathbf{c}_{\boldsymbol{\psi},y_m^f}}{\left(\sigma_m^f\right)^2+\left(\sigma_m^o\right)^2}$ is often called the Kalman gain matrix.

3.4 Visual illustration of the EnSRF's key features

To get the gist of how Eq. (3.3) updates the forecast ensemble, I plotted a bivariate demonstration of the EnSRF in Fig. 3.1. The EnSRF has several key features. First, the EnSRF pushes the prior ensemble members (blue capital letters) towards the observed value (red dashed line), and stops partway (black capital letters). Note that if the observation's error variance was lower (width of the observation likelihood function; *i.e.*, the horizontal width of red curve), the EnSRF will push the ensemble members closer to the observed value.

The second key feature is that the horizontal distance between any pair of ensemble members in Fig. 3.1 is decreased by the EnSRF. In other words, the posterior ensemble (black capital letters) has smaller ensemble variance in the observed quantity than that of the prior ensemble (blue capital letters).

Another key feature is that the EnSRF updates other quantities in $\boldsymbol{\psi}$ following linear regression. This can be seen in Fig. 3.1: the EnSRF pushed the ensemble members from the prior positions (blue capital letters) to their posterior positions (black capital letters) on diagonal trajectories (black thin lines) that are parallel to a regression line (solid blue diagonal line). Said regression line is constructed by treating prior ensemble (blue capital letters) as the data points and the observed quantity as the predictor, and is estimated via least squares.[5] [6]

[5] See Chapter 3.8 for the connection between the EnSRF and linear regression.

[6] Note that the linear regression coefficients are not explicitly computed in the EnSRF (*i.e.*,

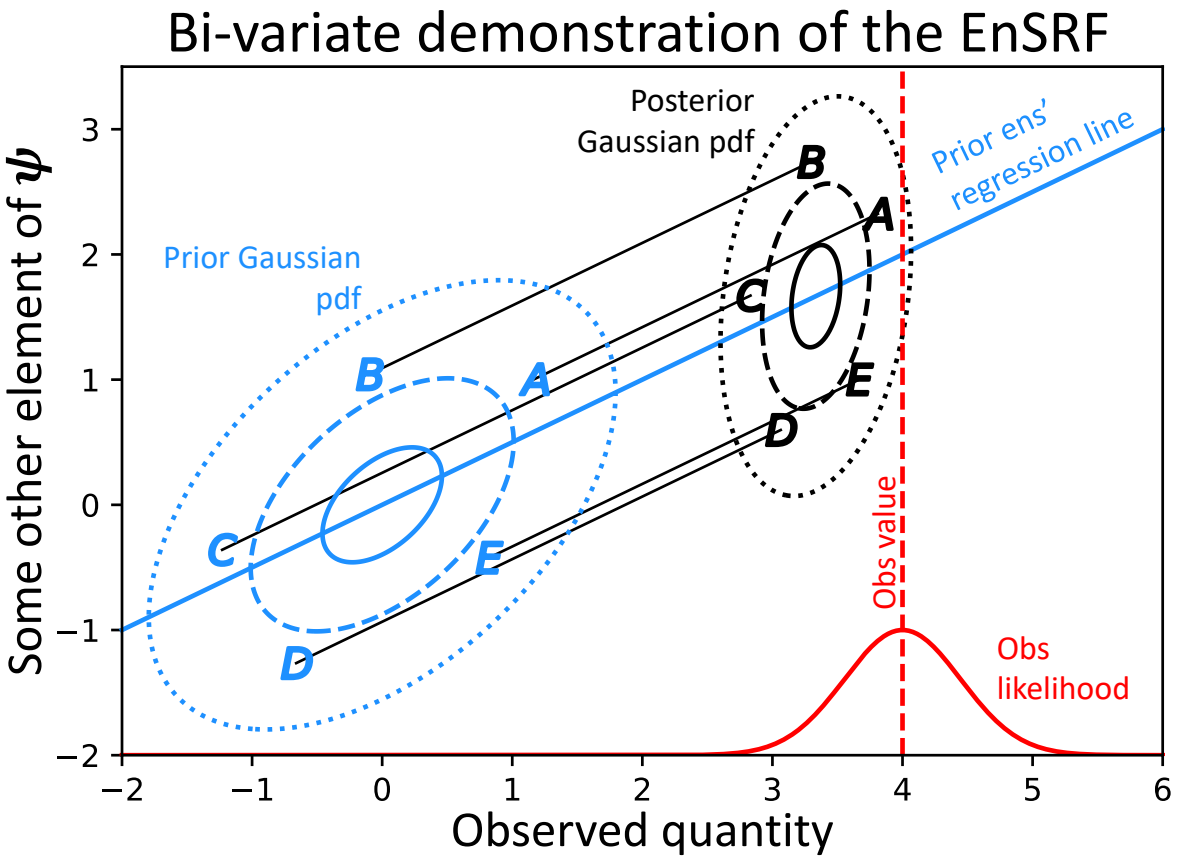


Figure 3.1: A bivariate illustration of how the EnSRF updates the prior ensemble through linear least squares relationships. The blue oval contours represent the prior Gaussian pdf, the blue capital letters (A, B, C, D & E) indicate 5 prior ensemble members, and the blue straight line is the least-squares linear relationship for the prior ensemble. Note that the relationship's predictor is the observed quantity. The red curve indicates the observation likelihood function and the vertical dashed red line indicates the observation value. The black oval contours represent the posterior Gaussian pdf the black capital letters (A, B, C, D & E) indicate 5 posterior ensemble members. The thin black lines show how the prior members are shifted to their posterior positions in a fashion parallel to the regression line.

---

the regression is done implicitly). However, the EnSRF's identical twin, the two-step ensemble adjustment Kalman filter [EAKF; Anderson (2003)] does explicitly compute the regression coefficient.

Finally, the EnSRF also reduces the ensemble's uncertainty of other quantities in $\boldsymbol{\psi}$. This effect is subtly illustrated in Fig. 3.1: the vertical displacements between any pair of ensemble members is slightly reduced by the EnSRF. In other words, the EnSRF can not only adjust unobserved quantities, but also reduce their uncertainties.

3.5 Heuristic strategies and modifications

3.5.1 Distance-dependent weakening of the EnSRF update (aka, localization)

Sampling errors can lead to erroneous estimates of the linear relationships implicitly used by the EnSRF. This issue is particularly egregious when the "true" or population linear relationship is weak[7]. Such sample-estimated relationships typically have small signal-to-noise relationships, meaning that the estimated relationships are suspect. For instance, if the Pearson correlation of the "true" linear relationship is +0.05, a 50-member ensemble's estimate of the relationship is likely severely contaminated by sampling noise. Left untreated, these problematic noisy relationships can result in EnSRF updates that undesirably degrade the ensemble.

In general, the strength of linear relationships (*i.e.*, the Pearson correlation coefficient) between any two geographical locations tend to weaken with their intervening distance. This means that the relationships used to update locations far from the observation site are likely weak, and therefore likely egregiously prone to sampling errors.

Another issue with the EnSRF is the potential inappropriateness of using linear regressions. Nonlinear relationships are likely when considering locations

[7] *i.e.*, the Pearson correlation coefficient is weak

that are far from the observation site. This can be inferred by considering Taylor expansions of some 3D atmospheric field as a function of Cartesian coordinates. Supposing the expansion is executed at the observation site, the linear expansion is only appropriate for sufficiently small distances from the observation site. As such, it is likely inappropriate to assume that the observed quantity and some faraway quantity in $\boldsymbol{\psi}$ to be linear.

Since the relationships between the observation site and faraway locations tend to be egregiously contaminated by sampling noise and/or nonlinear, the EnSRF's updates to these faraway locations are likely to degrade the ensemble's accuracy. The solution is thus to nullify these problematic long-distance updates. Specifically, updates to locations beyond a threshold distance (radius of influence, or ROI) from the observation site are disabled. This solution is called "localization" and is accomplished by multiplying a vector of "localization" factors (an N_ψ-dimensional vector $\boldsymbol{\rho_m}$) to the analysis increment. The localized EnSRF update equation can be written as

$$\boldsymbol{\psi_n^a} = \boldsymbol{\psi_n^f} + \boldsymbol{\rho_m} \circ \left\{ \mathbf{c}_{\boldsymbol{\psi}, y_m^f} \left[\frac{y_m^o - \overline{y^f} - \phi_m \left(y_{m,n}^f - \overline{y^f} \right)}{\left(\sigma_m^f \right)^2 + \left(\sigma_m^o \right)^2} \right] \right\} \tag{3.4}$$

where $\circ$ represents element-wise multiplication[8]. The elements of $\boldsymbol{\rho_m}$ that correspond to locations that are more than 1 ROI from the observation site are zero, and the remaining elements can be between 0 and 1 (inclusive).

Since atmospheric variables tend to vary smoothly in space, the localized EnSRF update should avoid introducing discontinuities.[9] As such, within the localized update zone, the localization factors are set up to vary smoothly from zero (at the zone's boundaries) to unity (at the center of the zone). This is typ-

[8] also known as a Hadamard product or a Schur product

[9] Severe model imbalances or a large amount of spurious gravity waves are likely to result from introducing discontinuities into the ensemble.

ically achieved by using localization factors that follow the Gaspari-Cohn fifth-order polynomial. This results in localization factors that strongly resemble a Gaussian function as a function of distance, except that the factors smoothly fall to zero at a distance of 1 ROI. As such, the EnSRF update is confined to a localized area without introducing problematic discontinuities.

Note that localization comes with a number of issues. Some of the more interesting ones are listed below.

1. Localization has been reported to induce model imbalances (Greybush et al., 2011). Nonetheless, this is a small price to pay to remove the deleterious long-distance updates.

2. ROI values are typically manually tuned. Since ROIs can vary depending on the phenomena resolved by the forecast model, observation density and the ensemble size (*e.g.*, Ying et al. (2018)), the optimal ROI values can vary from case-to-case. Unfortunately, manual optimization of ROIs is computationally expensive and labor intensive. Research into adaptive methods to estimate ROIs (Anderson, 2012; Lei and Anderson, 2014) is currently underway.

3. With localization, changing the order in which observations are assimilated can change the resulting posterior ensemble (Kotsuki et al., 2017; Hartman et al., 2021).

3.5.2 Dispersion inflation measures: RTPP and ABEI

The envelope of possible states represented by the forecast ensemble (*i.e.*, the forecast ensemble's variance/dispersion) should ideally contain the actual state of the atmosphere. Otherwise, the prior guess is inappropriate. Forecast ensembles with insufficient dispersion can arise due to 1) the EnSRF's tendency to remove too much ensemble dispersion in the previous DA cycle, 2)

unaccounted model errors, and/or 3) situations where the forecast ensemble is cloudless but the actual atmosphere is cloudy. The third item is a result of the fact that clear forecast ensembles have much smaller ensemble variances than cloudy or mixed ensembles. If under-dispersion is left untreated, the EnSRF's corrective ability is limited.

In this dissertation, two heuristic measures are used to maintain appropriate levels of ensemble dispersion. The first is the relaxation to prior perturbations (RTPP) proposed by Zhang et al. (2004). In RTPP, after assimilating all observations, the resulting ensemble members are modified via

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} \leftarrow \overline{\boldsymbol{\psi}^{\boldsymbol{a}}} + (1-\alpha)\left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} - \overline{\boldsymbol{\psi}^{\boldsymbol{a}}}\right) + \alpha\left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} - \overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right) \quad \forall \quad n = 1, 2, \cdots, N_E \tag{3.5}$$

where α is a user-specified value between 0 and 1 (inclusive) and is called the relaxation coefficient (or, occasionally, the mixing coefficient). RTPP essentially modifies the posterior members by mixing each posterior member's deviation from the posterior mean (posterior perturbations) with that of its prior form (prior perturbations).

The second measure to inflate ensemble dispersion is the adaptive background error inflation scheme [ABEI; Minamide and Zhang (2019)]. The ABEI is designed to treat situations where forecast ensemble is cloudless but the actual atmosphere is cloudy. Over geographical locations where such situations occur, the ABEI specifies an empirically-derived inflation factor. The ABEI then constructs a field of dispersion inflation factors over the entire forecast model's domain by spatially smoothing the previously specified inflation factors. The smoothing process is similar to that of Anderson (2007).

3.5.3 Adaptive observation error inflation (AOEI)

In situations where the forecast ensemble disagrees with infrared radiance observations on the presence/absence of clouds, the EnSRF update can result in model imbalances. This is because the forecast minus observation difference (henceforth, the innovation) tends to be large in such situations, thus resulting in large updates that can lead to model imbalances. The observation error variance $\left(\sigma_m^o\right)^2$ is can be inflated in such situations to reduce the update's magnitude.[10] This sort of heuristic measure is employed in most studies that assimilate all-sky infrared radiance observation via an EnKF.

The observation error inflation method used in this dissertation is the Adaptive Observation Error Inflation scheme [AOEI; Minamide and Zhang (2017)]. The AOEI asserts that the squared innovation for any infrared radiance observation d^2 should be smaller than the sum of the forecast and observation error variances. Mathematically,

$$d^2 \leq \left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2. \tag{3.6}$$

In situations where the above inequality is violated, the AOEI assumes that the violation arises from the model's inability to adequately resolve the observed feature (*i.e.*, representation errors). Since representation errors can be treated as a component of the observation error (Janjić et al., 2021), the AOEI inflates $\left(\sigma_m^o\right)^2$ until the above inequality is restored. Specifically, the AOEI modifies $\left(\sigma_m^o\right)^2$ via

$$\left(\sigma_m^o\right)^2 \leftarrow \max\left\{\left(\sigma_m^o\right)^2,\ d^2 - \left(\sigma_m^f\right)^2\right\}. \tag{3.7}$$

[10] To see why, notice that the $\left(\sigma_m^o\right)^2$ is in the denominator of update introduced by the EnSRF in Eq. (3.4).

THE MATHEMATICAL LAYER

3.6 Derivation of the EnSRF posterior pdf from first principles

3.6.1 Issue: $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is usually a singular matrix

We will now construct the EnSRF's posterior pdf though combining Bayes' rule with the Gaussian forecast error and Gaussian observation error assumptions. These assumptions are often naïvely interpreted to imply

$$p(\boldsymbol{\psi}) \equiv \frac{1}{\sqrt{(2\pi)^{N_\psi}\,\det\left(\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\right)}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\psi}-\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)^{\top} \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}{}^{-1}\left(\boldsymbol{\psi}-\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)\right\}, \tag{3.8}$$

and,

$$p\left(\mathbf{y^o}|\boldsymbol{\psi}\right) \equiv \frac{1}{\sqrt{(2\pi)^{N_y}\,\det\left(\boldsymbol{R}\right)}} \exp\left\{-\frac{1}{2}\left(\mathbf{H}\boldsymbol{\psi}-\mathbf{y^o}\right)^{\top} \boldsymbol{R}^{-1}\left(\mathbf{H}\boldsymbol{\psi}-\mathbf{y^o}\right)\right\}, \tag{3.9}$$

where $\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}$ and $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ are the sample mean vector and sample covariance matrix of the forecast ensemble $\left\{\boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{1}}, \boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{2}}, \ldots, \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{N_E}}\right\}$. Specifically,

$$\overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \equiv \frac{1}{N_E}\sum_{n=1}^{N_E} \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} \quad \text{and} \quad \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}} \equiv \frac{1}{N_E-1}\sum_{n=1}^{N_E} \boldsymbol{\psi}^{\boldsymbol{f}'}_{\boldsymbol{n}}\left(\boldsymbol{\psi}^{\boldsymbol{f}'}_{\boldsymbol{n}}\right)^{\top}, \tag{3.10}$$

where

$$\boldsymbol{\psi}^{\boldsymbol{f}'}_{\boldsymbol{n}} \equiv \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} - \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \quad \forall \;\; n = 1, 2, \ldots, N_E. \tag{3.11}$$

Finally, $\boldsymbol{R}$ is an $N_y \times N_y$ observation error covariance matrix.

The Gaussian prior pdf in Eq. (3.8) is usually invalid because $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is usually

singular. This singularity has two causes:

1. Some elements of $\boldsymbol{\psi}$ have purely linear and deterministic relationships with other elements of $\boldsymbol{\psi}$ (proven in Chapter 3.9.1).
2. $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ is computed from a small forecast ensemble (proven in Chapter 3.9.2).

Both conditions are usually present in practical DA. This is because $N_E \sim 50$ and $N_\psi \sim 10^9$ for practical DA ($N_E \ll N_\psi + 1$), and we often assimilate linear observations. A common example of the latter would be having zonal wind velocities in $\mathbf{x}$ and then assimilating zonal wind vector observations (*e.g.*, rawinsonde observations or satellite-derived motion vectors). As such, $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ is usually singular in practical DA, and Eq. (3.8) is thus usually invalid.

3.6.2 Solution: derive the posterior pdf a subspace where $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ is invertible

Since $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ is usually singular, the strategy to derive the posterior pdf has two steps:

1. In the subspace where $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ is invertible (henceforth, the $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace), derive the posterior pdf and its characteristic parameters.
2. Transform the posterior pdf's parameters from the $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace to the space that $\boldsymbol{\psi}$ lives in (henceforth, the full space).

Note that this strategy is not new – the ensemble transform Kalman filter [ETKF; Bishop et al. (2001)] is essentially based on this strategy.

The transformations used to convert between the $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace and the full subspace are as follows. Let the rank of $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ be N_s, meaning that the $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace has N_s-dimensions. The $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace can be accessed via the transformation

$$\mathbf{s} = \mathcal{S}^{\mathsf{T}} \left(\boldsymbol{\psi} - \boldsymbol{\psi}^{\boldsymbol{f}} \right) \tag{3.12}$$

where $\mathcal{S}$ is a $N_\psi \times N_s$ matrix whose columns form a set of orthonormal vectors that spans the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace. We will derive the EnSRF in terms of $\mathbf{s}$.

To go from the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace to the full space, we can use the "reverse" transformation

$$\boldsymbol{\psi}^* = \mathcal{S}\,\mathbf{s} + \overline{\boldsymbol{\psi}^{\boldsymbol{f}}}. \tag{3.13}$$

This Eq. (3.13) will be used to transform the derived EnSRF equations from the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace to the full space.[11]

Note that I did not explicitly write out the elements of $\mathcal{S}$. This is because all references to $\mathcal{S}$ vanishes after the EnSRF equations have been transformed to full space (will see that later on). All I require is that some form of $\mathcal{S}$ exists[12].

3.6.3 The posterior pdf and its parameters in the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace

We will now derive the posterior pdf and its parameters. Start with writing the Gaussian prior pdf [$p(\mathbf{s})$] in the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace:

$$p(\mathbf{s}) = \frac{1}{\sqrt{(2\pi)^{N_s}\det\left(\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{s}}\right)}} \exp\left\{-\frac{1}{2}\left(\mathbf{s}-\overline{\mathbf{s}^{\mathbf{f}}}\right)^{\top} \boldsymbol{P}^{\boldsymbol{f}\,-1}_{\boldsymbol{s}} \left(\mathbf{s}-\overline{\mathbf{s}^{\mathbf{f}}}\right)\right\}. \tag{3.14}$$

[11] Note that Eq. (3.13) does not strictly reverse the transformation into the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace [Eq. (3.12)]. For any $\boldsymbol{\psi}$ with a component in the nullspace of $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$, transforming to the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace [Eq. (3.13)] and then back to full space [Eq. (3.12)] results in losing the nullspace component. However, for any $\boldsymbol{\psi}$ that has no component in the nullspace of $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ Eqs. (3.12) and (3.13) reverse each other.

[12] $\mathcal{S}$ is guaranteed to exist for $N_s \geq 1$. This is because for any $N_s \geq 1$, there is at least one set of N_s orthonormal vectors that spans an N_s-dimensional subspace.

Here,

$$\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} \equiv \frac{1}{N_E} \sum_{n=1}^{N_E} \left(\mathbf{s}_{\mathbf{n}}^{\mathbf{f}} - \overline{\mathbf{s}^{\mathbf{f}}} \right) \left(\mathbf{s}_{\mathbf{n}}^{\mathbf{f}} - \overline{\mathbf{s}^{\mathbf{f}}} \right)^{\top}. \tag{3.15}$$

Note that $\overline{\mathbf{s}^{\mathbf{f}}}$ can be shown to be an N_s-dimensional zero vector.

The Gaussian observation likelihood can be written as

$$p\left(\mathbf{s}|\mathbf{y}^{\mathbf{o}}\right) \equiv \frac{1}{\sqrt{(2\pi)^{N_y} \det\left(\boldsymbol{R}\right)}} \times \exp\left\{ -\frac{1}{2} \left[\mathbf{H}\boldsymbol{\psi}^{*} - \mathbf{y}^{\mathbf{o}} \right]^{\top} \boldsymbol{R}^{-1} \left[\mathbf{H}\boldsymbol{\psi}^{*} - \mathbf{y}^{\mathbf{o}} \right] \right\}$$

where $\boldsymbol{\psi}^{*}$ is result of transforming $\mathbf{s}$ to full space. Said transformation is defined earlier in Eq. (3.13):

$$\boldsymbol{\psi}^{*} = \mathcal{S}\,\mathbf{s} + \overline{\boldsymbol{\psi}^{\boldsymbol{f}}}.$$

As such, the observation likelihood can be written as

$$p\left(\mathbf{s}|\mathbf{y}^{\mathbf{o}}\right) \equiv \frac{1}{\sqrt{(2\pi)^{N_y} \det\left(\boldsymbol{R}\right)}} \times \exp\left\{ -\frac{1}{2} \left[\mathbf{H}\mathcal{S}\,\mathbf{s} - \left(\mathbf{y}^{\mathbf{o}} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \right) \right]^{\top} \boldsymbol{R}^{-1} \left[\mathbf{H}\mathcal{S}\,\mathbf{s} - \left(\mathbf{y}^{\mathbf{o}} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \right) \right] \right\} \tag{3.16}$$

To proceed, we need Bayes' rule in the $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace:

$$p\left(\mathbf{s}|\mathbf{y}^{\mathbf{o}}\right) = \frac{p\left(\mathbf{s}\right) p\left(\mathbf{y}^{\mathbf{o}}|\mathbf{s}\right)}{p\left(\mathbf{y}^{\mathbf{o}}\right)}. \tag{3.17}$$

Since $p\left(\mathbf{s}\right)$ [Eq. (3.14)] and $p\left(\mathbf{y}^{\mathbf{o}}|\mathbf{s}\right)$ [Eq. (3.16)] are Gaussian pdfs, $p\left(\mathbf{s}\right) p\left(\mathbf{y}^{\mathbf{o}}|\mathbf{s}\right)$ results the following rescaled Gaussian pdf [*e.g.*, Anderson and Anderson (1999),

Fletcher (2017b)]:

$$p(\mathbf{s})\, p\left(\mathbf{y^o}|\mathbf{s}\right) = \frac{\alpha_s}{\sqrt{(2\pi)^{N_s} \det\left(\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}\right)}} \exp\left\{-\frac{1}{2}\left(\mathbf{s}-\overline{\mathbf{s^a}}\right)^{\top} \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}-1}\left(\mathbf{s}-\overline{\mathbf{s^a}}\right)\right\} \tag{3.18}$$

where

$$\overline{\mathbf{s^a}} \equiv \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} \mathcal{S}^{\top} \mathbf{H}^{\top}\left(\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\mathcal{S}^{\top}\mathbf{H}^{\top}+\boldsymbol{R}\right)^{-1}\left[\mathbf{y^o}-\mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right],$$

$$\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}} \equiv \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} - \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} \mathcal{S}^{\top} \mathbf{H}^{\top}\left(\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\mathcal{S}^{\top}\mathbf{H}^{\top}+\boldsymbol{R}\right)^{-1}\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}, \text{ and,}$$

$$\alpha_s = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{y^o}-\mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)^{\top}\left(\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\mathbf{H}^{\top}\mathcal{S}^{\top}+\boldsymbol{R}\right)^{-1}\left(\mathbf{y^o}-\mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)\right\}}{\sqrt{(2\pi)^{N_y}\det\left(\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\mathbf{H}^{\top}\mathcal{S}^{\top}+\boldsymbol{R}\right)}}.$$

See Chapter 3.9.3 for a derivation of Eq. (3.18). Note that the $\overline{\mathbf{s^f}}$ terms were removed because $\overline{\mathbf{s^f}}$ is a zero vector.

To obtain the posterior pdf $p(\mathbf{s}|\mathbf{y^o})$, consider that

$$p\left(\mathbf{y^o}\right) = \int_{\mathbb{R}^{N_S}} p\left(\mathbf{y^o} \cap \mathbf{s}\right) \mathrm{d}^{N_s}\mathbf{s} = \int_{\mathbb{R}^{N_S}} p\left(\mathbf{s}\right) p\left(\mathbf{y^o}|\mathbf{s}\right) \mathrm{d}^{N_s}\mathbf{s}. \tag{3.19}$$

Since $p(\mathbf{s})p(\mathbf{y^o}|\mathbf{s})$ is a Gaussian pdf times α_s, then $p(\mathbf{y^o}) = \alpha_s$. As such, by Bayes' rule [Eq. (3.17)], the posterior pdf is

$$p\left(\mathbf{s}|\mathbf{y^o}\right) = \frac{1}{\sqrt{(2\pi)^{N_s} \det\left(\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}\right)}} \exp\left\{-\frac{1}{2}\left(\mathbf{s}-\overline{\mathbf{s^a}}\right)^{\top} \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}-1}\left(\mathbf{s}-\overline{\mathbf{s^a}}\right)\right\} \tag{3.20}$$

with posterior parameters

$$\overline{\mathbf{s^a}} \equiv \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} \mathcal{S}^{\top} \mathbf{H}^{\top}\left(\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\mathcal{S}^{\top}\mathbf{H}^{\top}+\boldsymbol{R}\right)^{-1}\left[\mathbf{y^o}-\mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right], \text{ and,} \tag{3.21}$$

$$\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}} \equiv \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} - \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} \mathcal{S}^{\top} \mathbf{H}^{\top}\left(\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\mathcal{S}^{\top}\mathbf{H}^{\top}+\boldsymbol{R}\right)^{-1}\mathbf{H}\mathcal{S}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}. \tag{3.22}$$

3.6.4 Conversion of $\overline{\mathbf{s}^{\mathbf{a}}}$ and $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$ to full space

We will now execute the second step in Chapter 3.6.2: the transformation of the $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$ subspace posterior parameters ($\overline{\mathbf{s}^{\mathbf{a}}}$ and $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$) to their full space forms ($\overline{\boldsymbol{\psi}^{\boldsymbol{a}}}$ and $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}}$). $\overline{\mathbf{s}^{\mathbf{a}}}$ can be transformed to $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}}$ through Eq. (3.13). In other words,

$$\overline{\boldsymbol{\psi}^{\boldsymbol{a}}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \mathcal{S}\,\overline{\mathbf{s}^{\mathbf{a}}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top}\mathbf{H}^{\top}\left(\mathbf{H}\,\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top}\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\left[\mathbf{y}^{\mathbf{o}} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right]. \quad (3.23)$$

To simplify Eq. (3.23), notice that

$$\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top} = \mathcal{S}\left[\frac{1}{N_E}\sum_{n=1}^{N_E-1}\mathbf{s}_{\mathbf{n}}^{\mathbf{f}}\left(\mathbf{s}_{\mathbf{n}}^{\mathbf{f}}\right)^{\top}\right]\mathcal{S}^{\top} = \frac{1}{N_E}\sum_{n=1}^{N_E-1}\mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\mathbf{f}}\left(\mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\mathbf{f}}\right)^{\top} = \frac{1}{N_E}\sum_{n=1}^{N_E-1}\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}'}\left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}'}\right)^{\top}$$

$$\therefore \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top} = \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} \quad (3.24)$$

Substituting Eq. (3.24) into Eq. (3.23) thus gives

$$\overline{\boldsymbol{\psi}^{\boldsymbol{a}}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}\,\mathbf{H}^{\top}\left(\mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}\,\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\left[\mathbf{y}^{\mathbf{o}} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right]$$

We can also convert $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$ to its full space counterpart ($\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}}$) via:

$$\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}} = \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}\,\mathcal{S}^{\top}. \quad (3.25)$$

This expression is proven in Chapter 3.9.4. Combining Eq. (3.25) with the expression for $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$ [Eq. (3.15)] and Eq. (3.24) gives

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}} &= \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}\,\mathcal{S}^{\top} = \mathcal{S}\left[\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}} - \boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top}\mathbf{H}^{\top}\left(\mathbf{H}\,\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top}\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\mathbf{H}\,\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\right]\mathcal{S}^{\top} \\
&= \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top} - \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top}\mathbf{H}^{\top}\left(\mathbf{H}\,\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top}\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\mathbf{H}\,\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{f}}\,\mathcal{S}^{\top} \\
&= \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} - \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}\,\mathbf{H}^{\top}\left(\mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}\,\mathbf{H}^{\top} + \boldsymbol{R}\right)\mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}.
\end{aligned}$$

3.6.5 Summary

In summary, the posterior pdf is a Gaussian pdf characterized by the posterior mean $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}}$ and posterior covariance matrix $\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi}}$. The expressions for $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}}$ and $\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi}}$ are also known as the Kalman filter (KF) equations. For ease of reference, the KF equations are repeated here:

$$\overline{\boldsymbol{\psi}^{\boldsymbol{a}}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\mathbf{H}^{\top}\left(\mathbf{H}\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\left[\mathbf{y}^{\mathbf{o}} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right], \text{ and,} \tag{3.26}$$

$$\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi}} = \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}} - \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\mathbf{H}^{\top}\left(\mathbf{H}\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\mathbf{H}^{\top} + \boldsymbol{R}\right)\mathbf{H}\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}. \tag{3.27}$$

Some notes:

1. As promised in Chapter 3.6.2, all references to $\mathcal{S}$ vanish when we convert the EnSRF update equations from the $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ subspace to full space. As such, there is no need to write out any explicit formulation for $\mathcal{S}$ – we only require that $\mathcal{S}$ to exist for the derivation to work.

2. If $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ turns out to be non-singular (*i.e.*, invertible), the derivation of the KF equations presented here also applies. The only difference is that $\mathcal{S}$ becomes an $N_{\psi} \times N_{\psi}$ identity matrix.

3. The full space KF equations [Eqs. (3.26) and (3.27)] here are identical to those derived under the assumption that $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is invertible. However, my derivation is more rigorous because I do not assume invertible $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$.

3.7 Independent observations can be assimilated serially

If all observations are made independently (*i.e.*, assumption 3 in Chapter 3.2), they can be assimilated serially. To see why, notice that for independent

observations, the observation likelihood $p(\mathbf{y^o}|\mathbf{s})$ can be rewritten as

$$p(\mathbf{y^o}|\mathbf{s}) = p\left(y_1^o|\mathbf{s}\right) p\left(y_2^o|\mathbf{s}\right) \ldots p\left(y_{N_y}^o|\mathbf{s}\right)$$

and the marginal $p(\mathbf{y^o})$ can be rewritten as

$$p(\mathbf{y^o}) = p\left(y_1^o\right) p\left(y_2^o\right) \ldots p\left(y_{N_y}^o\right).$$

The $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace Bayes's rule then becomes

$$p\left(\mathbf{s}|\mathbf{y^o}\right) = \frac{p(\mathbf{s})\, p(\mathbf{y^o}|\mathbf{s})}{p(\mathbf{y^o})} = \frac{p(\mathbf{s})\, p\left(y_1^o|\mathbf{s}\right) p\left(y_2^o|\mathbf{s}\right) \ldots p\left(y_{N_y}^o|\mathbf{s}\right)}{p\left(y_1^o\right) p\left(y_2^o\right) \ldots p\left(y_{N_y}^o\right)}$$
$$= \frac{p(\mathbf{s})\, p\left(y_1^o|\mathbf{s}\right)}{p\left(y_1^o\right)} * \frac{p\left(y_2^o|\mathbf{s}\right)}{p\left(y_2^o\right)} * \frac{p\left(y_3^o|\mathbf{s}\right)}{p\left(y_3^o\right)} * \cdots * \frac{p\left(y_{N_y}^o|\mathbf{s}\right)}{p\left(y_{N_y}^o\right)}.$$

$p(\mathbf{s}|\mathbf{y^o})$ can thus be evaluated through the recurrence relation

$$p\left(\mathbf{s}|y_1^o, y_2^o, \ldots, y_{m+1}^o\right) = p\left(\mathbf{s}|y_1^o, y_2^o, \ldots, y_m^o\right) \frac{p\left(y_{m+1}^o|\mathbf{s}\right)}{p\left(y_{m+1}^o\right)} \tag{3.28}$$

which starts from

$$p\left(\mathbf{s}|y_1^o\right) = \frac{p(\mathbf{s})\, p\left(y_1^o|\mathbf{s}\right)}{p\left(y_1^o\right)}.$$

If the recurrence relation is evaluated $(N_y - 1)$ times, we will obtain $p(\mathbf{s}|\mathbf{y^o})$.

Notice that starting from $p\left(\mathbf{s}|y_1^o\right)$ is equivalent to starting by assimilating the first observation y_1^o into the forecast ensemble. The resulting ensemble is termed the $m = 1$ ensemble. Evaluating the Eq. (3.7) to obtain $p\left(\mathbf{s}|y_1^o, y_2^o\right)$ is then equivalent to assimilating y_2^o into the $m = 1$ ensemble (the resulting ensemble is termed the $m = 2$ ensemble). We can then assimilate the third ob-

servation y_3^o into the $m = 2$ ensemble to represent constructing $p\left(\mathbf{s}|y_1^o, y_2^o, y_3^o\right)$. The pattern can be repeated until all N_y observations are assimilated, thus obtaining $p\left(\mathbf{s}|y_1^o, \ldots, y_{N_y}^o\right)$ $(\equiv p(\mathbf{s}|\mathbf{y^o}))$. As such, because the independent observation assumption allows Bayes' rule to be evaluated recursively, we can assimilate the observations serially.

3.8 Construction of the EnSRF update algorithm

3.8.1 Single observation KF equations

We will now construct the EnSRF algorithm from the KF equations. Since the EnSRF explicitly assumes independent observations (Chapter 3.2), it will assimilate observations serially (Chapter 3.7). It is thus sufficient to construct the EnSRF's update procedure for assimilating one observation from the single observation form of the KF equations. The KF equations [Eqs. (3.26) and (3.27)] to assimilate an observation y_m^o into $\left\{\boldsymbol{\psi_1^f}, \boldsymbol{\psi_2^f}, \ldots, \boldsymbol{\psi_{N_E}^f}\right\}$ are:

$$\overline{\boldsymbol{\psi^a}} = \overline{\boldsymbol{\psi^f}} + \frac{\mathbf{c}_{\boldsymbol{\psi^f}, y_m^f}}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right], \quad \text{and,} \tag{3.29}$$

$$\boldsymbol{P_\psi^a} = \boldsymbol{P_\psi^f} - \mathbf{c}_{\boldsymbol{\psi^f}, y_m^f} \frac{1}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left(\mathbf{c}_{\boldsymbol{\psi^f}, y_m^f}\right)^{\mathrm{T}} \tag{3.30}$$

where the various quantities are as defined in Chapter 3.3.

3.8.2 Heuristic constraints to construct the EnSRF

Before constructing the EnSRF update procedure, note that for any specified posterior mean and covariance, there are infinite possible posterior ensem-

bles (Tippett et al., 2003)[13]. Heuristic constraints are thus needed to select a specific posterior ensemble and construct the EnSRF update procedure. The heuristic constraints I will use are essentially a rehash of the two-step EAKF thinking introduced by Anderson (2003).[14]

Heuristic constraint 1: the posterior ensemble is constructed via linear regression

The first heuristic constraint I will impose is that the construction of the posterior ensemble uses best-fit linear relationships between the observed quantity and the other elements of $\boldsymbol{\psi}$. Specifically, the regression treats the observed quantity y_m as the predictor, all elements in $\boldsymbol{\psi}$ as predicted quantities, and the prior ensemble $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}}\right\}$ as the data points used for regression. The resulting least-squares linear relationships can be shown to be

$$\boldsymbol{\psi}_{\mathbf{fit}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{\beta}_{\boldsymbol{m}}\left(y_m - \overline{y_m^f}\right) \tag{3.31}$$

where $\boldsymbol{\psi}_{\mathbf{fit}}$ is the linear model's predicted $\boldsymbol{\psi}$ for a given value of the observed quantity y_m (y_m is the $(N_x + m)$-th element in $\boldsymbol{\psi}$), and $\boldsymbol{\beta}$ is an N_ψ-dimensional vector of linear regression coefficients. Furthermore,

$$\boldsymbol{\beta}_{\boldsymbol{m}} \equiv \frac{\mathbf{C}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}}{\left(\sigma_m^f\right)^2}. \tag{3.32}$$

[13]The evidence is that we can always convert any ensemble of N_E randomly drawn N_ψ-dimensional white noise vectors into an ensemble with covariance $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}}$. Note that these white noise vectors must be drawn such that they span an N_E-dimensional vector space. Since there are infinite possible random draws, then there are infinite possible posterior ensembles that follow any given $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}}$ and $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}}$.

[14]Side-note: Whitaker and Hamill (2002) originally constructed the EnSRF by converting the stochastic Monte Carlo EnKF (Evensen, 1994; Burgers et al., 1998; Houtekamer and Mitchell, 1998) into a deterministic formulation. The main idea of Whitaker and Hamill (2002) was to replace the random draws representing $\boldsymbol{R}$ with a square-root modification factor.

As such, each forecast ensemble member can be written as

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{\beta}_{\boldsymbol{m}} \left(y_{m,n}^{f} - \overline{y_{m}^{f}} \right) + \mathbf{r_n} \quad \forall \quad n = 1, 2, \ldots, N_E \tag{3.33}$$

where $\mathbf{r_n}$ is the residual of $\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}}$ that is unaccounted by the linear regression. Note because $\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}$ is a point in Eq. (3.31), thus $\overline{\mathbf{r_n}}$ is an N_ψ-dimensional vector of zeros.

The heuristic constraint then is that we seek posterior ensemble members of the form

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{\beta}_{\boldsymbol{m}} \left(y_{m,n}^{a} - \overline{y_{m}^{f}} \right) + \mathbf{r_n} \quad \forall \quad n = 1, 2, \ldots, N_E \tag{3.34}$$

where $\left\{ y_{m,1}^{a}, y_{m,2}^{a}, \ldots, y_{m,N_E}^{a} \right\}$ are to be determined. In other words, the posterior members are constructed by linear regression. Some algebraic manipulation of Eq. (3.34) yields:

$$\begin{aligned} \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} &= \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{\beta}_{\boldsymbol{m}} \left(y_{m,n}^{a} - y_{m,n}^{f} + y_{m,n}^{f} - \overline{y_{m}^{f}} \right) + \mathbf{r_n} \\ &= \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{\beta}_{\boldsymbol{m}} \left(y_{m,n}^{f} - \overline{y_{m}^{f}} \right) + \mathbf{r_n} + \boldsymbol{\beta}_{\boldsymbol{m}} \left(y_{m,n}^{a} - y_{m,n}^{f} \right). \end{aligned}$$

Substituting in Eq. (3.33) yields an expression to construct the posterior members through linear regression:

$$\therefore \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} = \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} + \boldsymbol{\beta}_{\boldsymbol{m}} \left(y_{m,n}^{a} - y_{m,n}^{f} \right) \quad \forall \quad n = 1, 2, \ldots, N_E. \tag{3.35}$$

Heuristic constraint 2: $y^{a'}_{m,n} = \alpha_m y^{f'}_{m,n}$

The second desired heuristic constraint is

$$y^{a'}_{m,n} = \alpha_m y^{f'}_{m,n} \qquad \forall \quad n = 1, 2, \ldots, N_E \tag{3.36}$$

where

$$y^{a'}_{m,n} \equiv y^{a}_{m,n} - \overline{y^{a}_{m}}\,, \qquad y^{f'}_{m,n} \equiv y^{f}_{m,n} - \overline{y^{f}_{m}}\,, \tag{3.37}$$

$$\overline{y^{a}_{m}} \equiv \frac{1}{N_E}\sum_{n=1}^{N_E} y^{a}_{m,n} \qquad \overline{y^{f}_{m}} \equiv \frac{1}{N_E}\sum_{n=1}^{N_E} y^{f}_{m,n}\,,$$

and α_m is a positive constant to be determined. To use this second constraint, substitute Eq. (3.37) and then Eq. (3.36) into the first constraint's posterior member update equation [Eq. (3.35)]:

$$\begin{aligned}\boldsymbol{\psi^a_n} &= \boldsymbol{\psi^f_n} + \boldsymbol{\beta_m}\left(y^{a}_{m,n} - y^{f}_{m,n}\right) = \boldsymbol{\psi^f_n} + \boldsymbol{\beta_m}\left(\overline{y^{a}_{m}} + y^{a'}_{m,n} - \left(\overline{y^{f}_{m}} + y^{f'}_{m,n}\right)\right) \\ &= \boldsymbol{\psi^f_n} + \boldsymbol{\beta_m}\left(\overline{y^{a}_{m}} + \alpha_m y^{f'}_{m,n} - \left(\overline{y^{f}_{m}} + y^{f'}_{m,n}\right)\right).\end{aligned}$$

As such, the posterior members that obey the two heuristic constraints have the following form:

$$\boxed{\boldsymbol{\psi^a_n} = \boldsymbol{\psi^f_n} + \boldsymbol{\beta_m}\left[\overline{y^{a}_{m}} - \overline{y^{f}_{m}} + (\alpha_m - 1)\, y^{f'}_{m,n}\right] \qquad \forall \quad n = 1, 2, \ldots, N_E.} \tag{3.38}$$

The EnSRF update equations

All that is left is to construct a set of $\left\{y^{a}_{m,1}, y^{a}_{m,2}, \ldots, y^{a}_{m,N_E}\right\}$ such that the heuristically-constrained posterior members [Eq. (3.38)] obey the single observation KF equations [Eqs. (3.29) and (3.30)]. We can obtain $\overline{y^{a}_{m}}$ by equating the ensemble mean of the heuristically-constrained posterior members [Eq.

(3.38)] with the mean state KF equation [Eq. (3.29)]. In other words,

$$\overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \boldsymbol{\beta}_{\boldsymbol{m}}\left(\overline{y_m^a} - \overline{y_m^f} + (\alpha_m - 1)\,\overline{y_{m,n}^{f'}}\right) = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \frac{\mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right]$$

where the left-hand side comes from Eq. (3.35) and the right-hand side comes from Eq. (3.29). Since $\overline{y_{m,n}^{f'}} = 0$, substituting in the definition of $\boldsymbol{\beta}_{\boldsymbol{m}}$ [Eq. (3.32)] and doing some manipulation gives

$$\begin{aligned}
&\overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \frac{\mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}}{\left(\sigma_m^f\right)^2}\left(\overline{y_m^a} - \overline{y_m^f}\right) = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} + \frac{\mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right] \\
&\Leftrightarrow \frac{1}{\left(\sigma_m^f\right)^2}\left(\overline{y_m^a} - \overline{y_m^f}\right) = \frac{1}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right]
\end{aligned}$$

$$\boxed{\therefore \overline{y_m^a} = \overline{y_m^f} + \frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right].} \tag{3.39}$$

The next step is to determine a set of $\left\{y_{m,1}^{a'}, y_{m,1}^{a'}, \ldots, y_{m,N_E}^{a'}\right\}$ that causes in the heuristically-constrained posterior members to satisfy the KF equation to update the covariance matrix [Eq. (3.30)]. By the second heuristic constraint [Eq. (3.36)], this procedure simplifies to seeking a α_m such that said KF equation is satisfied. Supposing

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}'} \equiv \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} - \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \quad \text{and} \quad \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}'} \equiv \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} - \overline{\boldsymbol{\psi}^{\boldsymbol{a}}},$$

we can derive the following expression for $\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}'}$ by combining Eqs. (3.29), (3.38), and (3.39):

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}'} = \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}'} + \boldsymbol{\beta}_{\boldsymbol{m}}\,(\alpha_m - 1)\,y_{m,n}^{f'}. \tag{3.40}$$

An expression α_m can then be derived by combining Eq. (3.40) with the definition of sample covariance matrices, and then equating to the KF covariance update equation [Eq. (3.30)]. In other words,

$$\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}} \equiv \frac{\sum_{n=1}^{N_E} \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}\prime} \left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}\prime}\right)^{\top}}{N_E - 1} = \frac{\sum_{n=1}^{N_E} \left[\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}\prime} + \boldsymbol{\beta_m} \left(\alpha_m - 1\right) y_{m,n}^{f\prime}\right] \left[\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}\prime} + \boldsymbol{\beta_m} \left(\alpha_m - 1\right) y_{m,n}^{f\prime}\right]^{\top}}{N_E - 1}$$
$$= \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} + \left(\alpha_m - 1\right)^2 \left(\sigma_m^f\right)^2 \boldsymbol{\beta_m} \boldsymbol{\beta_m}^{\top} + \left(\alpha_m - 1\right) \left(\boldsymbol{\beta_m} \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top} + \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \boldsymbol{\beta_m}^{\top}\right)$$

Substituting in the definition of $\boldsymbol{\beta_m}$ [Eq. (3.32)] gives

$$\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}} = \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} + \frac{\left(\alpha_m - 1\right)^2}{\left(\sigma_m^f\right)^2} \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top} + \frac{\left(\alpha_m - 1\right)}{\left(\sigma_m^f\right)^2} \left(\mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top} + \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top}\right)$$

$$\therefore \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}} = \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} + \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \left(\frac{\left(\alpha_m - 1\right)^2 + 2\left(\alpha_m - 1\right)}{\left(\sigma_m^f\right)^2}\right) \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top}. \tag{3.41}$$

Equating with the KF covariance update equation [Eq. (3.30)] gives

$$\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} - \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \frac{1}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2} \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top} = \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} + \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f} \left(\frac{\left(\alpha_m - 1\right)^2 + 2\left(\alpha_m - 1\right)}{\left(\sigma_m^f\right)^2}\right) \mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}^{\top}.$$

This suggests

$$\frac{-1}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2} = \frac{\left(\alpha_m - 1\right)^2 + 2\left(\alpha_m - 1\right)}{\left(\sigma_m^f\right)^2}.$$

With some algebraic manipulation, we obtain

$$\alpha_m^2 - \left(1 - \frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\right) = 0.$$

Since the expression in the parenthesis is always positive, the roots of the quadratic α_m equation are real. These roots are

$$\alpha_m = \pm\sqrt{1 - \frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} = \pm\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}}$$

To select one of the roots, consider that the negative root would cause $\|y_n^{a'} - y_n^{f'}\|$ to be bigger than that of the positive root. The negative root would thus cause larger updates to the prior ensemble of $\boldsymbol{\psi}$ vectors than the positive root. Since larger $\boldsymbol{\psi}$ updates are more likely to cause model imbalance, we select the positive root. Hence,

$$\boxed{\alpha_m = +\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}}.} \tag{3.42}$$

Combining our desired linear regression form of the posterior members [Eq.

(3.38)] with the derived expression for $\overline{y_m^a}$ and α_m thus gives

$$\boldsymbol{\psi_n^a} = \boldsymbol{\psi_n^f} + \boldsymbol{\beta_m}\left[\overline{y_m^a} - \overline{y_m^f} + (\alpha_m - 1)\, y_{m,n}^{f'}\right]$$

$$= \boldsymbol{\psi_n^f} + \boldsymbol{\beta_m}\left[\overline{y_m^f} + \frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right] - \overline{y_m^f} + \left(\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} - 1\right) y_{m,n}^{f'}\right]$$

$$= \boldsymbol{\psi_n^f} + \boldsymbol{\beta_m}\left[\frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right] + \left(\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} - 1\right) y_{m,n}^{f'}\right]$$

Considering the definition of $\boldsymbol{\beta_m}$ [Eq. (3.32)] gives

$$\boldsymbol{\psi_n^a} = \boldsymbol{\psi_n^f} + \frac{\mathbf{c}_{\boldsymbol{\psi^f}, y_m^f}}{\left(\sigma_m^f\right)^2}\left[\frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f}\right] + \left(\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} - 1\right) y_{m,n}^{f'}\right]$$

$$= \boldsymbol{\psi_n^f} + \frac{\mathbf{c}_{\boldsymbol{\psi^f}, y_m^f}}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left[y_m^o - \overline{y_m^f} + \frac{\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} - 1}{\frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}}\, y_{m,n}^{f'}\right]$$

Notice that

$$\frac{\sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} - 1}{\frac{\left(\sigma_m^f\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} = \frac{-\left(1 - \sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}}\right)}{1 - \frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}} = \frac{-1}{1 + \sqrt{\frac{\left(\sigma_m^o\right)^2}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}}} = -\phi_m$$

where ϕ_m is as defined in Chapter 3.3.

We have thus derived the EnSRF update equation [Eq. (3.3)] stated in Chapter 3.3:

$$\boxed{\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} = \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} + \frac{\mathbf{c}_{\boldsymbol{\psi}^{\boldsymbol{f}}, y_m^f}}{\left(\sigma_m^f\right)^2 + \left(\sigma_m^o\right)^2}\left(y_m^o - \overline{y_m^f} - \phi_m y_{m,n}^{f'}\right).}$$

The procedure to use the EnSRF update equation to serially assimilate N_y observations is described in Chapter 3.3.

3.9 Appendix

3.9.1 Proof: $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is singular for $N_E < N_\psi + 1$

To determine the rank of $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$, start by considering the forecast ensemble's perturbation $\boldsymbol{\psi}$ vectors. The n-th such vector is

$$\begin{aligned}
\boldsymbol{\psi}^{\boldsymbol{f}'}_{\boldsymbol{n}} &\equiv \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} - \frac{1}{N_E}\sum_{n=1}^{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} \\
&= \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{1}} + \cdots + \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n-1}} + \frac{N_E-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} + \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n+1}} + \cdots + \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{N_E}} \\
&= \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{1}} + \cdots + \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n-1}} + \left(1+\frac{-1}{N_E}\right)\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n}} + \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n+1}} + \cdots + \frac{-1}{N_E}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{N_E}}
\end{aligned}$$

As such,

$$\boldsymbol{\psi}^{\boldsymbol{f}'}_{\boldsymbol{n}} = \begin{bmatrix}\boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{1}} & \boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{2}} & \cdots & \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{N_E}}\end{bmatrix}\begin{bmatrix}\nu \\ \vdots \\ \nu \\ 1+\nu \\ \nu \\ \vdots \\ \nu\end{bmatrix}$$

where the column vector has N_E elements,

$$\nu = -1/N_E, \tag{3.43}$$

and the $1-\nu$ value in the column vector occurs in the n-th element.

Supposing

$$\boldsymbol{\Psi}^{\boldsymbol{f}} \equiv \begin{bmatrix}\boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{1}} & \boldsymbol{\psi}^{\boldsymbol{f}}_{\mathbf{2}} & \cdots & \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{N_E}}\end{bmatrix},$$

we can thus write

$$\left[\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}'} \quad \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}'} \quad \cdots \quad \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}'}\right] = \boldsymbol{\Psi}^{\boldsymbol{f}}\,\mathbf{V} \tag{3.44}$$

where $\mathbf{V}$ is the following $N_E \times N_E$ matrix

$$\mathbf{V} \equiv \begin{bmatrix} 1+\nu & \nu & \cdots & \nu & \nu \\ \nu & 1+\nu & \cdots & \nu & \nu \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \nu & \nu & \cdots & 1+\nu & \nu \\ \nu & \nu & \cdots & \nu & 1+\nu \end{bmatrix}. \tag{3.45}$$

To figure out the rank of $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$, rewrite the sample-based definition of $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ in terms of $\boldsymbol{\Psi}^{\boldsymbol{f}}$. Thus,

$$\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} = \frac{1}{N_E-1}\sum_{n=1}^{N_E} \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}'}\left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}'}\right)^{\top} = \frac{1}{N_E-1}\left[\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}'} \quad \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}'} \quad \cdots \quad \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}'}\right]\left[\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}'} \quad \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}'} \quad \cdots \quad \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}'}\right]^{\top}$$

$$\therefore \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}} = \boldsymbol{\Psi}^{\boldsymbol{f}}\,\mathbf{V}\,\mathbf{V}^{\top}{\boldsymbol{\Psi}^{\boldsymbol{f}}}^{\top}. \tag{3.46}$$

We can therefore infer that

$$\operatorname{rank}\left(\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}\right) \leq \min\left\{\operatorname{rank}\left(\boldsymbol{\Psi}^{\boldsymbol{f}}\right), \quad \operatorname{rank}(\mathbf{V}), \quad N_\psi\right\}. \tag{3.47}$$

Since $\boldsymbol{\Psi}^{\boldsymbol{f}}$ is an $N_\psi \times N_E$ matrix,

$$\operatorname{rank}\left(\boldsymbol{\Psi}^{\boldsymbol{f}}\right) \leq \min\left\{N_E, N_\psi\right\}.$$

Note that if $N_E > N_\psi$ and $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}}\right\}$ are a linearly independent set of vectors, then the rank of $\boldsymbol{\Psi}^{\boldsymbol{f}}$ is N_E. If this linear independence holds but $N_E < N_\psi$, then $\operatorname{rank}\left(\boldsymbol{\Psi}^{\boldsymbol{f}}\right) = N_\psi$.

It can be shown that the rank of $\mathbf{V}$ is $N_E - 1$. First, notice that the sum of the

first $N_E - 1$ rows of $\mathbf{V}$ gives the following row vector

$$\begin{bmatrix} 1+\nu+(N_E-2)\nu & 1+\nu+(N_E-2)\nu & \cdots & 1+\nu+(N_E-2)\nu & (N_E-1)\nu \end{bmatrix} = -\begin{bmatrix} \nu & \nu & \cdots & \nu & 1+\nu \end{bmatrix}.$$

The definition of ν [Eq. (3.43)] was used to simplify the above row vector. The above indicates that the last row of $\mathbf{V}$ is the negative sum (*i.e.*, a linear combination) of the first $N_E - 1$ rows. As such, the row space of $\mathbf{V}$ is the same as the following $(N_E - 1) \times N_E$ matrix

$$\mathbf{V}^* = \begin{bmatrix} 1+\nu & \nu & \cdots & \nu & \nu \\ \nu & 1+\nu & \cdots & \nu & \nu \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \nu & \nu & \cdots & 1+\nu & \nu \end{bmatrix}. \tag{3.48}$$

The rows of $\mathbf{V}^*$ are linearly independent. This can be proven by contradiction. Pretend that the last row can be expressed as some linear sum of the first $N_E - 2$. In other words,

$$\begin{bmatrix} \nu & \nu & \cdots & \nu & 1+\nu & \nu \end{bmatrix} = \sum_{r=1}^{N_E-2} a_r \left(\nu \mathbf{1}_{\mathbf{N_E}\times\mathbf{1}}{}^\top + \hat{\mathbf{e}}_{\mathbf{r}}^\top \right) \tag{3.49}$$

where $a_1, a_2, \ldots, a_{N_E-2}$ are scalar coefficients, $\mathbf{1}_{\mathbf{N_E}\times\mathbf{1}}$ is an N_E-dimensional vector of ones, and $\hat{\mathbf{e}}_{\mathbf{r}}$ is an N_E-dimensional vector where the r-th element is unity and all other elements are zeros.

The N_E-th element on both sides of Eq. (3.49) is

$$\nu = \sum_{r=1}^{N_E-2} a_r \, \nu. \tag{3.50}$$

There is no term relating to $\hat{\mathbf{e}}_{\mathbf{r}}$ here because the summation in Eq. (3.49) does not include $\hat{\mathbf{e}}_{\mathbf{N_E}}$.

Now examine the first element (*i.e.*, the leftmost element) on both sides of Eq. (3.49):

$$\nu = \sum_{r=1}^{N_E-2} a_r \nu + a_1. \tag{3.51}$$

Substituting Eq. (3.50) into Eq. (3.51) results in

$$\nu = \nu - 1 \tag{3.52}$$

which is clearly a contradiction. As such, the last row of $\mathbf{V}^*$ cannot be a linear combination of the other rows of $\mathbf{V}^*$. Similar proofs by contradiction can be executed on all rows of $\mathbf{V}^*$, thus indicating that the rows of $\mathbf{V}^*$ are linearly independent.

Since the row space of $\mathbf{V}$ is the same as the row space of $\mathbf{V}^*$, the row space of $\mathbf{V}$ has $N_E - 1$ dimensions. Furthermore, since $\mathbf{V}$ is symmetric, its column space also has $N_E - 1$ dimensions. In other words,

$$\text{rank}(\mathbf{V}) = N_E - 1. \tag{3.53}$$

Thus, the inequality in Eq. (3.47) becomes

$$\text{rank}\left(\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\right) \leq \min\left\{\text{rank}\left(\boldsymbol{\Psi}^{\boldsymbol{f}}\right), \quad N_E - 1, \quad N_\psi\right\}.$$

For the typical situation where $N_E < N_\psi$,

$$\text{rank}\left(\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\right) \leq \min\left\{\text{rank}\left(\boldsymbol{\Psi}^{\boldsymbol{f}}\right) \quad N_E - 1, \quad N_\psi\right\} \leq \left\{N_E, \quad N_E - 1, \quad N_\psi\right\}.$$
$$\therefore \text{rank}\left(\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\right) \leq N_E - 1.$$

The maximum rank of $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is thus guaranteed to be $N_E - 1$. As such, $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is singular for typical ensemble sizes.

3.9.2 Proof: $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is singular if a subset of $\boldsymbol{\psi}$ has deterministic linear relationships with other elements of $\boldsymbol{\psi}$

Suppose some elements of $\boldsymbol{\psi}$ are have deterministic linear relationships with other elements of $\boldsymbol{\psi}$. Let $\mathbf{s}$ represent the latter elements of $\boldsymbol{\psi}$ and let N_s represent the number of elements in $\mathbf{s}$ (*i.e.*, $N_s < N_\psi$). As such, one can construct any $\boldsymbol{\psi}$ from $\mathbf{s}$ via a $N_\psi \times N_s$ matrix $\mathbf{L}$, *i.e.*,

$$\boldsymbol{\psi} = \mathbf{L}\,\mathbf{s}. \tag{3.54}$$

Since $N_s < N_\psi$, $\mathbf{L}$ has a maximum rank of N_s.

To see why $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is singular in such cases, consider the definition of $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$

$$\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}} \equiv \overline{\left(\boldsymbol{\psi}^{\boldsymbol{f}} - \overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)\left(\boldsymbol{\psi}^{\boldsymbol{f}} - \overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)^{\top}} \tag{3.55}$$

where the overbars indicate an ensemble averages. For simplicity assume that $N_E \gg N_\psi$. Applying Eq (3.54) into Eq (3.55) gives

$$\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}} = \mathbf{L}\left\{\overline{\left(\mathbf{s}^{\mathbf{f}} - \overline{\mathbf{s}^{\mathbf{f}}}\right)\left(\mathbf{s}^{\mathbf{f}} - \overline{\mathbf{s}^{\mathbf{f}}}\right)^{\top}}\right\}\mathbf{L}^{\top} = \mathbf{L}\,\mathbf{P}_{\mathbf{s}}\,\mathbf{L}^{\top} \tag{3.56}$$

where

$$\mathbf{P}_{\mathbf{s}} \equiv \overline{\left(\mathbf{s}^{\mathbf{f}} - \overline{\mathbf{s}^{\mathbf{f}}}\right)\left(\mathbf{s}^{\mathbf{f}} - \overline{\mathbf{s}^{\mathbf{f}}}\right)^{\top}}.$$

Since $\mathbf{s}$ has N_s elements, $\mathbf{P}_{\mathbf{s}}$ has a maximum rank of N_s. Thus,

$$\operatorname{rank}\left(\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}\right) = \operatorname{rank}\left(\mathbf{L}\,\mathbf{P}_{\mathbf{s}}\,\mathbf{L}^{\top}\right) \leq \max\left\{\operatorname{rank}(\mathbf{L}), \operatorname{rank}(\mathbf{P}_{\mathbf{s}}), \operatorname{rank}\left(\mathbf{L}^{\top}\right)\right\}$$

Since the ranks of $\mathbf{P_s}$ and $\mathbf{L}$ are both N_s, and $N_s < N_\psi$, then, $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is rank deficient. This rank deficiency happens even if $N_E \gg N_\psi$. As such, $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi}}$ is singular if any element in $\boldsymbol{\psi}$ have deterministic linear relationships with the other elements in $\boldsymbol{\psi}$.

3.9.3 Proof: Multiplication of two Gaussian pdfs results in a scaled Gaussian pdf

Suppose we have two Gaussian pdfs for an N_p-dimensional state vector $\mathbf{p}$:

$$p_A(\mathbf{p}) \equiv \frac{1}{\sqrt{(2\pi)^{N_p}\det(A)}} \exp\left\{-\frac{1}{2}(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A})^\top \mathbf{A}^{-1}(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A})\right\} \tag{3.57}$$

and

$$p_B(\mathbf{p}) \equiv \frac{1}{\sqrt{(2\pi)^{N_\omega}\det(A)}} \exp\left\{-\frac{1}{2}(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B})^\top \mathbf{B}^{-1}(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B})\right\}. \tag{3.58}$$

$\boldsymbol{\mu}_\mathbf{A}$ is a constant N_p-dimensional vector, and $\mathbf{A}$ is some invertible $N_s \times N_s$ covariance matrix. Furthermore, $\boldsymbol{\Omega}$ is an $N_\omega \times N_p$ matrix where $N_\omega < N_p$, $\boldsymbol{\mu}_\mathbf{B}$ is a constant N_ω-dimensional vector, and $\mathbf{B}$ is an invertible $N_\omega \times N_\omega$ covariance matrix.

The goal of this section is to show that

$$p_A(\mathbf{p})\,p_B(\mathbf{p}) = \frac{\alpha}{\sqrt{(2\pi)^{N_p}\det(\mathbf{G})}} \exp\left\{-\frac{1}{2}(\mathbf{p}-\boldsymbol{\mu}_\mathbf{G})^\top \mathbf{G}^{-1}(\mathbf{p}-\boldsymbol{\mu}_\mathbf{G})\right\} \tag{3.59}$$

where

$$\mathbf{G} = \mathbf{A} - \mathbf{A}\boldsymbol{\Omega}^{\top}\left(\mathbf{B} + \boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}, \tag{3.60}$$

$$\boldsymbol{\mu}_{\mathbf{G}} = \boldsymbol{\mu}_{\mathbf{A}} + \mathbf{A}\boldsymbol{\Omega}^{\top}\left(\mathbf{B} + \boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\left(\boldsymbol{\mu}_{\mathbf{B}} - \boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}\right), \text{ and,} \tag{3.61}$$

$$\alpha = \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} - \boldsymbol{\mu}_{\mathbf{B}}\right)^{\top}\left(\mathbf{B} + \boldsymbol{\Omega}\,\mathbf{A}\,\boldsymbol{\Omega}^{\top}\right)^{-1}\left(\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} - \boldsymbol{\mu}_{\mathbf{B}}\right)\right\}}{\sqrt{(2\pi)^{N_\omega}\det\left(\mathbf{B} + \boldsymbol{\Omega}\,\mathbf{A}\,\boldsymbol{\Omega}^{\top}\right)}}. \tag{3.62}$$

Proving Eq. (3.59) confirms that the multiplication of two Gaussian pdfs gives a Gaussian pdf that is scaled by the scalar factor α.

To derive Eq. (3.59), begin with multiplying the two Gaussian pdfs:

$$\begin{aligned} &p_A(\mathbf{p})\,p_B(\mathbf{p}) \\ &= C\exp\left\{-\frac{1}{2}\left(\mathbf{p} - \boldsymbol{\mu}_{\mathbf{A}}\right)^{\top}\mathbf{A}^{-1}\left(\mathbf{p} - \boldsymbol{\mu}_{\mathbf{A}}\right) - \frac{1}{2}\left(\boldsymbol{\Omega}\mathbf{p} - \boldsymbol{\mu}_{\mathbf{B}}\right)^{\top}\mathbf{B}^{-1}\left(\boldsymbol{\Omega}\mathbf{p} - \boldsymbol{\mu}_{\mathbf{B}}\right)\right\} \\ &= C\exp\left\{\left(\mathbf{p} - \boldsymbol{\mu}_{\mathbf{A}}\right)^{\top}\mathbf{A}^{-1}\left(\mathbf{p} - \boldsymbol{\mu}_{\mathbf{A}}\right) + \left(\boldsymbol{\Omega}\mathbf{p} - \boldsymbol{\mu}_{\mathbf{B}}\right)^{\top}\mathbf{B}^{-1}\left(\boldsymbol{\Omega}\mathbf{p} - \boldsymbol{\mu}_{\mathbf{B}}\right)\right\}^{-1/2}. \end{aligned} \tag{3.63}$$

Here,

$$C \equiv \frac{1}{\sqrt{(2\pi)^{N_p}\det(\mathbf{A})}} * \frac{1}{\sqrt{(2\pi)^{N_\omega}\det(\mathbf{B})}}. \tag{3.64}$$

Because this derivation is ~~painfully~~ long, the remainder of the derivation is broken into parts. In the first part, we will derive the quadratic term in the exponent of Eq. (3.59). In the second part, we will derive the $\alpha/\sqrt{(2\pi)^{N_p}\det(\mathbf{G})}$ factor in Eq. (3.59).

Part 1: deriving the quadratic term in Eq. (3.59)

We will now concentrate on addressing the terms in the curly braces, *i.e.*,

$$\begin{aligned}
&\left(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A}\right)^\top \mathbf{A}^{-1}\left(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A}\right)+\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B}\right)^\top \mathbf{B}^{-1}\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B}\right)\\
&=\mathbf{p}^\top\mathbf{A}^{-1}\mathbf{p}-2\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}\mathbf{p}+\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\mathbf{p}^\top\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{p}-2\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{p}+\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\\
&=\mathbf{p}^\top\left(\mathbf{A}^{-1}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}\right)\mathbf{p}-2\left(\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}+\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}\right)\mathbf{p}+\left(\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\right)
\end{aligned}$$

$$\begin{aligned}
&\therefore\left(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A}\right)^\top \mathbf{A}^{-1}\left(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A}\right)+\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B}\right)^\top \mathbf{B}^{-1}\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B}\right)\\
&=\mathbf{p}^\top\left(\mathbf{A}^{-1}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}\right)\mathbf{p}-2\left(\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\right)^\top\mathbf{p}+\left(\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\right).
\end{aligned}\tag{3.65}$$

Notice: if we invert $\mathbf{G}^{-1}$ [Eq. (3.60)] and apply the Woodbury matrix identity, we get

$$\mathbf{G}^{-1}=\mathbf{A}^{-1}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}.\tag{3.66}$$

For convenience, let's also define

$$\mathbf{q}\equiv\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}.\tag{3.67}$$

Substituting Eqs. (3.66) and (3.67) into Eq. (3.65) gives

$$\begin{aligned}
&\left(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A}\right)^\top \mathbf{A}^{-1}\left(\mathbf{p}-\boldsymbol{\mu}_\mathbf{A}\right)+\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B}\right)^\top \mathbf{B}^{-1}\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_\mathbf{B}\right)\\
&=\mathbf{p}\left(\mathbf{A}^{-1}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\Omega}\right)\mathbf{p}-2\left(\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\Omega}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\right)^\top\mathbf{p}+\left(\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\right)\\
&=\mathbf{p}\mathbf{G}^{-1}\mathbf{p}-2\mathbf{q}^\top\mathbf{p}+\left(\boldsymbol{\mu}_\mathbf{A}^\top\mathbf{A}^{-1}\boldsymbol{\mu}_\mathbf{A}+\boldsymbol{\mu}_\mathbf{B}^\top\mathbf{B}^{-1}\boldsymbol{\mu}_\mathbf{B}\right).
\end{aligned}$$

To proceed, we will assume $\mathbf{G}^{-1}$ is invertible and use the multivariate form of completing the square:

$$\mathbf{p}^\top\mathbf{G}^{-1}\mathbf{p}-2\mathbf{q}^\top\mathbf{p}=\left(\mathbf{p}-\mathbf{G}\mathbf{q}\right)^\top\mathbf{G}^{-1}\left(\mathbf{p}-\mathbf{G}\mathbf{q}\right)-\mathbf{q}^\top\mathbf{G}\mathbf{q}.\tag{3.68}$$

Applying Eq. (3.68) gives

$$\begin{aligned}&\left(\mathbf{p}-\boldsymbol{\mu}_{\mathbf{A}}\right)^{\top}\mathbf{A}^{-1}\left(\mathbf{p}-\boldsymbol{\mu}_{\mathbf{A}}\right)+\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_{\mathbf{B}}\right)^{\top}\mathbf{B}^{-1}\left(\boldsymbol{\Omega}\mathbf{p}-\boldsymbol{\mu}_{\mathbf{B}}\right)\\&=\mathbf{p}\mathbf{G}^{-1^{\top}}\mathbf{p}-2\mathbf{b}^{\top}\mathbf{p}+\left(\boldsymbol{\mu}_{\mathbf{A}}^{\top}\mathbf{A}^{-1}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\right)\end{aligned}\tag{3.69}$$

Substituting Eq. (3.69) into Eq. (3.63) gives

$$p_A(\mathbf{p})\,p_B(\mathbf{p})\equiv CD\exp\left\{-\frac{1}{2}(\mathbf{p}-\mathbf{G}\mathbf{q})^{\top}\mathbf{G}^{-1}(\mathbf{p}-\mathbf{G}\mathbf{q})\right\}\tag{3.70}$$

where

$$D\equiv\exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\mu}_{\mathbf{A}}^{\top}\mathbf{A}^{-1}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\right)-\mathbf{q}^{\top}\mathbf{G}\mathbf{q}\right]\right\}.\tag{3.71}$$

If we can show that

$$CD=\frac{\alpha}{\sqrt{(2\pi)^{N_p}\det(\mathbf{G})}}\tag{3.72}$$

and

$$\mathbf{G}\mathbf{q}=\boldsymbol{\mu}_{\mathbf{G}}\equiv\boldsymbol{\mu}_{\mathbf{A}}+\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\mathbf{B}+\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\left(\boldsymbol{\mu}_{\mathbf{B}}-\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}\right)\tag{3.73}$$

then Eq. (3.70) is the same as Eq. (3.59) [*i.e.*, we have derived Eq. (3.59)].

We will now derive Eq. (3.73). By Eq. (3.67),

$$\mathbf{G}\mathbf{q}=\left(\mathbf{A}^{-1}+\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\right)^{-1}\left(\mathbf{A}^{-1}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\right)\tag{3.74}$$

$$=\left[\mathbf{A}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\right]\left(\mathbf{A}^{-1}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\right)\tag{3.75}$$

$$=\boldsymbol{\mu}_{\mathbf{A}}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\left[\mathbf{A}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\right]\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}$$

$$=\boldsymbol{\mu}_{\mathbf{A}}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\left[\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}}$$

$$=\boldsymbol{\mu}_{\mathbf{A}}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\left[\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}}$$

$$=\boldsymbol{\mu}_{\mathbf{A}}-\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\mathbf{A}\boldsymbol{\Omega}^{\top}\left[\mathbf{B}^{-1}-\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}}.$$

Therefore,

$$\begin{aligned}\mathbf{Gq} = \; & \boldsymbol{\mu}_\mathbf{A} - \mathbf{A}\boldsymbol{\Omega}^\top \left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^\top + \mathbf{B}\right)^{-1} \boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} \\ & + \mathbf{A}\boldsymbol{\Omega}^\top \left[\mathbf{B}^{-1} - \left(\mathbf{B} + \mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^\top\right)^{-1}\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_\mathbf{B} \qquad (3.76) \\ = \; & \boldsymbol{\mu}_\mathbf{A} - \mathbf{A}\boldsymbol{\Omega}^\top \left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^\top + \mathbf{B}\right)^{-1} \boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} + \mathbf{A}\boldsymbol{\Omega}^\top \left[\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^\top + \mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_\mathbf{B} \qquad (3.77)\end{aligned}$$

$$\therefore \mathbf{Gq} = \boldsymbol{\mu}_\mathbf{A} + \mathbf{A}\boldsymbol{\Omega}^\top \left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^\top + \mathbf{B}\right)^{-1} \left(\boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B}\right) \qquad (3.78)$$

To get from (3.74) to (3.75), the Woodbury matrix identity was applied on $\left(\mathbf{A}^{-1} + \boldsymbol{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\Omega}\right)^{-1}$. To get from (3.76) to (3.77), the Woodbury matrix identity was applied on $\left(\mathbf{B} + \mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^\top\right)^{-1}\mathbf{B}\right)^{-1}$. We have thus derived Eq. (3.73).

Part 2: Deriving $\alpha/\sqrt{(2\pi)^{N_p}\det(\mathbf{G})}$ in Eq. (3.59)

The next stage is to show that show Eq. (3.72):

$$CD = \frac{\alpha}{\sqrt{(2\pi)^{N_p}\det(\mathbf{G})}}$$

where α is defined in Eq. (3.62) to be

$$\alpha \equiv \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B}\right)^\top \left(\mathbf{B} + \boldsymbol{\Omega}\,\mathbf{A}\,\boldsymbol{\Omega}^\top\right)^{-1} \left(\boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B}\right)\right\}}{\sqrt{(2\pi)^{N_\omega}\det\left(\mathbf{B} + \boldsymbol{\Omega}\,\mathbf{A}\,\boldsymbol{\Omega}^\top\right)}}.$$

To show Eq. (3.72), we just have to show that

$$D = \exp\left\{-\frac{1}{2}\left(\boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B}\right)^\top \left(\mathbf{B} + \boldsymbol{\Omega}\,\mathbf{A}\,\boldsymbol{\Omega}^\top\right)^{-1} \left(\boldsymbol{\Omega}\boldsymbol{\mu}_\mathbf{A} - \boldsymbol{\mu}_\mathbf{B}\right)\right\} \qquad (3.79)$$

and

$$C = \frac{1}{\sqrt{(2\pi)^{N_\omega + N_p} \det(\mathbf{G}) \det\left(\mathbf{B} + \mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top\right)}}. \tag{3.80}$$

To show Eq. (3.79), we can start with Eq. (3.71), substitute in the identities of Eqs. (3.60) and (3.67).

$$D = \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\mu}_\mathbf{A}^\top \mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right) - \mathbf{q}^\top \mathbf{G} \mathbf{q}\right]\right\}$$

$$= \exp\left\{\begin{array}{l}\left(\boldsymbol{\mu}_\mathbf{A}^\top \mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right) \\ \quad -\left(\mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right)^\top \left(\mathbf{A}^{-1} + \mathbf{\Omega}^\top \mathbf{B}^{-1} \mathbf{\Omega}\right)^{-1} \left(\mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right)\end{array}\right\}^{-1/2}$$

Apply Woodbury matrix identity on $\left(\mathbf{A}^{-1} + \mathbf{\Omega}^\top \mathbf{B}^{-1} \mathbf{\Omega}\right)^{-1}$ gives

$$D = \exp\left\{\begin{array}{l}\left(\boldsymbol{\mu}_\mathbf{A}^\top \mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right) \\ \quad -\left(\mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right)^\top \left[\mathbf{A} - \mathbf{A} \mathbf{\Omega}^\top \left(\mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top + \mathbf{B}\right)^{-1} \mathbf{\Omega} \mathbf{A}\right] \\ \quad\quad * \left(\mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\right)\end{array}\right\}^{-1/2}.$$

Now, we expand the ~~annoyingly long~~ second term in the exponent

$$D = \exp\left\{\begin{array}{l}\boldsymbol{\mu}_\mathbf{A}^\top \mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B} \\ \quad -\boldsymbol{\mu}_\mathbf{A}^\top \mathbf{A}^{-1} \boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{A}^\top \mathbf{\Omega}^\top \left(\mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top + \mathbf{B}\right)^{-1} \mathbf{\Omega} \boldsymbol{\mu}_\mathbf{A} \\ \quad -\boldsymbol{\mu}_\mathbf{A}^\top \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B} + \boldsymbol{\mu}_\mathbf{A}^\top \mathbf{\Omega}^\top \left(\mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top + \mathbf{B}\right)^{-1} \mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B} \\ \quad -\boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \mathbf{\Omega} \boldsymbol{\mu}_\mathbf{A} + \boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top \left(\mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top + \mathbf{B}\right)^{-1} \mathbf{\Omega} \boldsymbol{\mu}_\mathbf{A} \\ \quad -\boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B} + \boldsymbol{\mu}_\mathbf{B}^\top \mathbf{B}^{-1} \mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top \left(\mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top + \mathbf{B}\right)^{-1} \mathbf{\Omega} \mathbf{A} \mathbf{\Omega}^\top \mathbf{B}^{-1} \boldsymbol{\mu}_\mathbf{B}\end{array}\right\}^{-1/2},$$

realize that the two $\mathbf{A}^{-1}\boldsymbol{\mu}_{\mathbf{A}}$ terms cancel out

$$D = \exp\left\{\begin{array}{l}\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\end{array}\right\}^{-1/2},$$

employ the Woodbury matrix identity on $\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}$ in the last term

$$D = \exp\left\{\begin{array}{l}\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}}\end{array}\right\}^{-1/2},$$

and manipulate the last term in the exponent to get

$$D = \exp\left\{\begin{array}{l}\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\left(\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)\left[\mathbf{B}^{-1}-\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}}\end{array}\right\}^{-1/2}.$$

Applying the Woodbury matrix identity on $\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}$ then results

in

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\left(\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)\left[\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right]^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2}.$$

A bit more manipulation on the last term in the exponent gives:

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\left(\boldsymbol{\mu}_{\mathbf{B}}^{\top}\right)\left[\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right]^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2}.$$

Chucking in the Woodbury identity for $\left[\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right]^{-1}$ gives:

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\left(\boldsymbol{\mu}_{\mathbf{B}}^{\top}\right)\left[\mathbf{B}^{-1}-\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2}.$$

Expanding the last term in the exponent gives

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\left[\mathbf{B}^{-1}-\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2},$$

and the two $\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}$ terms cancel out to give

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad -\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}+\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad +\boldsymbol{\mu}_{\mathbf{B}}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2}.$$

To proceed further, recognize that $\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}$ and $\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}$ are scalars. This means that

$$\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} = \left(\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\right)^{\top} = \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}},$$

and,

$$\begin{aligned}&\boldsymbol{\mu}_{\mathbf{B}}^{\top}\mathbf{B}^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} = \left\{\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}\right\}^{\top} \\ &= \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}.\end{aligned}$$

As such,

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -2\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}}+2\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{B}}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2}.$$

We then collect the two terms with the factor of 2:

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -2\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left[\mathbf{B}^{-1}-\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{B}}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2},$$

manipulate $\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\mathbf{B}^{-1}$ to get:

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -2\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left[\mathbf{B}^{-1}-\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}\right]\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{B}}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2},$$

and apply the Woodbury matrix identity on $\left(\mathbf{B}+\mathbf{B}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}\right)^{-1}\mathbf{B}\right)^{-1}$ to get

$$D = \exp\left\{\begin{array}{l} \boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}} \\ \quad -2\boldsymbol{\mu}_{\mathbf{A}}^{\top}\boldsymbol{\Omega}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}} \\ \quad +\boldsymbol{\mu}_{\mathbf{B}}^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\boldsymbol{\mu}_{\mathbf{B}} \end{array}\right\}^{-1/2}.$$

Completing the square for the exponent, and bringing the $-1/2$ power back into the exponential function then gives us Eq. (3.79):

$$\therefore D = \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_{\mathbf{B}}-\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}\right)^{\top}\left(\boldsymbol{\Omega}\mathbf{A}\boldsymbol{\Omega}^{\top}+\mathbf{B}\right)^{-1}\left(\boldsymbol{\mu}_{\mathbf{B}}-\boldsymbol{\Omega}\boldsymbol{\mu}_{\mathbf{A}}\right)\right\}.$$

All that is left to complete this derivation is to show Eq. (3.80). Starting

from Eq. (3.64),

$$C \equiv \frac{1}{\sqrt{(2\pi)^{N_p}\det(\mathbf{A})}} * \frac{1}{\sqrt{(2\pi)^{N_\omega}\det(\mathbf{B})}} = \left\{(2\pi)^{N_\omega+N_p}\det(\mathbf{B})\det(\mathbf{A})\right\}^{-1/2}$$
$$= \left\{(2\pi)^{N_\omega+N_p}\det(\mathbf{B})\det(\mathbf{A})\frac{\det(\mathbf{G})}{\det(\mathbf{G})}\right\}^{-1/2} = \left\{(2\pi)^{N_\omega+N_p}\det(\mathbf{G})\frac{\det(\mathbf{B})\det(\mathbf{A})}{\det(\mathbf{G})}\right\}^{-1/2}$$
$$= \left\{(2\pi)^{N_\omega+N_p}\det(\mathbf{G})\left[\det(\mathbf{B})\det(\mathbf{A})\det\left(\mathbf{G}^{-1}\right)\right]\right\}^{-1/2}$$

Now we chuck in Eq. (3.66) to get

$$C = \left\{(2\pi)^{N_\omega+N_p}\det(\mathbf{G})\left[\det(\mathbf{B})\det(\mathbf{A})\det\left(\mathbf{A}^{-1}+\mathbf{\Omega}^{\top}\mathbf{B}^{-1}\mathbf{\Omega}\right)\right]\right\}^{-1/2},$$

and employ the generalized matrix determinant lemma on the terms in the square brackets to obtain

$$C = \left\{(2\pi)^{N_\omega+N_p}\det(\mathbf{G})\det\left(\mathbf{B}+\mathbf{\Omega}\,\mathbf{A}\,\mathbf{\Omega}^{\top}\right)\right\}^{-1/2}$$
$$\therefore C = \frac{1}{\sqrt{(2\pi)^{N_\omega+N_p}\det(\mathbf{G})\det\left(\mathbf{B}+\mathbf{\Omega}\,\mathbf{A}\,\mathbf{\Omega}^{\top}\right)}}.$$

We have thus proven Eq. (3.80).

3.9.4 Derivation: $\mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}\,\mathcal{S}^{\top} = \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}}$

An appropriate identity is needed to convert $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$ to its full space counterpart ($\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{a}}$). To derive this identity, suppose that we have an ensemble of $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{f}}$ subspace vectors with a mean of $\overline{\mathbf{s}^{\mathbf{a}}}$ and a covariance of $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{a}}$. Let these vectors be denoted by $\left\{\mathbf{s}_{\mathbf{1}}^{\mathbf{a}}, \mathbf{s}_{\mathbf{2}}^{\mathbf{a}}, \ldots, \mathbf{s}_{\mathbf{N_E}}^{\mathbf{a}}\right\}$. These vectors can be transformed to full space through Eq. (3.13):

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}\,*} = \mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\mathbf{a}} + \overline{\boldsymbol{\psi}^{\boldsymbol{f}}} \qquad \forall \quad n = 1, 2, \ldots, N_E. \tag{3.81}$$

If we define the covariance matrix of $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{\alpha}*}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{\alpha}*}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{\alpha}*}\right\}$ as $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{\alpha}}$, we can derive an identity to convert $\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{\alpha}}$ to $\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{\alpha}}$. Specifically,

$$
\begin{aligned}
\boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{\alpha}} &\equiv \frac{1}{N_E-1}\sum_{n=1}^{N_E}\left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{\alpha}*}-\overline{\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{\alpha}*}}\right)\left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{\alpha}*}-\overline{\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{\alpha}*}}\right)^{\mathsf{T}} \\
&= \frac{1}{N_E-1}\sum_{n=1}^{N_E}\left(\mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}+\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}-\left(\mathcal{S}\,\overline{\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}}+\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)\right)\left(\mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}+\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}-\left(\mathcal{S}\,\overline{\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}}+\overline{\boldsymbol{\psi}^{\boldsymbol{f}}}\right)\right)^{\mathsf{T}} \\
&= \frac{1}{N_E-1}\sum_{n=1}^{N_E}\left(\mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}-\mathcal{S}\,\overline{\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}}\right)\left(\mathcal{S}\,\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}-\mathcal{S}\,\overline{\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}}\right)^{\mathsf{T}} \\
&= \mathcal{S}\left[\frac{1}{N_E-1}\sum_{n=1}^{N_E}\left(\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}-\overline{\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}}\right)\left(\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}-\overline{\mathbf{s}_{\mathbf{n}}^{\boldsymbol{\alpha}}}\right)^{\mathsf{T}}\right]\mathcal{S}^{\mathsf{T}}.
\end{aligned}
$$

$$
\boxed{\therefore \boldsymbol{P}_{\boldsymbol{\psi}}^{\boldsymbol{\alpha}} = \mathcal{S}\,\boldsymbol{P}_{\boldsymbol{s}}^{\boldsymbol{\alpha}}\,\mathcal{S}^{\mathsf{T}}.} \tag{3.82}
$$

Chapter 4

Real data GeoIR DA tests on a tropical squall line

4.1 Overview and goals

As discussed in Chapter 1, the corrective impact of assimilating real world GeoIR observations into tropical MCS datasets is a relatively unexplored area of research. In this chapter, I will explore this area using the case of a tropical squall line that swept into the Maritime Continent. Ensembles of gray-zone resolution Weather Research and Forecasting (WRF) model simulations and real-world conventional observations were used. I also assimilated IR observations from the Advanced Himawari Imager (AHI), which is on-board the Himawari-8 geostationary satellite (Bessho et al., 2016).

4.2 Materials and Methods

4.2.1 Setup of simulation ensembles

To generate WRF simulation ensembles of the squall line, I employed the Advanced Research WRF (Skamarock et al., 2008) model (version 3.8.1). The

simulation domain covers the area plotted in Figure 4.1a and has 560 × 450 horizontal grid points with 9-km grid spacing. 45 model levels were employed, with the bottommost 9 levels within the lowest 1-km of the model, and the model top pressure set to 20 hPa.

For the WRF model's physics and dynamics, I employed the updated Goddard shortwave scheme (Chou and Suarez, 1999) and the Global Circulation Model version of the Rapid Radiative Transfer Model (RRTMG) longwave radiation scheme (Iacono et al., 2008). Surface processes were generally simulated using the unified Noah land surface physics scheme (Chen and Dudhia, 2001), the only exception being that surface skin temperatures were diagnosed via the method of Zeng and Beljaars (2005). The Yonsei University (YSU) boundary layer scheme (Hong et al., 2006) was employed to handle subgrid-scale turbulent mixing. Finally, I utilized the scheme of Thompson et al. (2008) to parameterize cloud microphysical processes.

Note that while the 9-km horizontal grid spacing is too coarse to resolve individual convective updrafts, it is sufficient to resolve the basic dynamics of tropical MCSs. There are several reasons behind choosing this horizontal grid spacing. First, due to the rapid motion of the tropical squall line (Figure 4.1a), a large domain is required. Due to our limited computational resources, it is difficult to utilize smaller grid spacings in this large domain. Second, earlier work over the nearby Indian Ocean have shown that this 9-km grid spacing is sufficiently small to realistically replicate the precipitation, circulation, thermodynamic, and radiation features of organized mesoscale convection over tropical oceans (Wang et al., 2015). Finally, this 9-km grid spacing is widely employed in studies that examine the overturning and the physical mechanisms of organized tropical convection (Chen et al., 2018b,c,a; Ruppert et al., 2020; Ruppert and Chen, 2020; Chen et al., 2020, 2022).

I have also opted to avoid the use of cumulus parameterizations in this

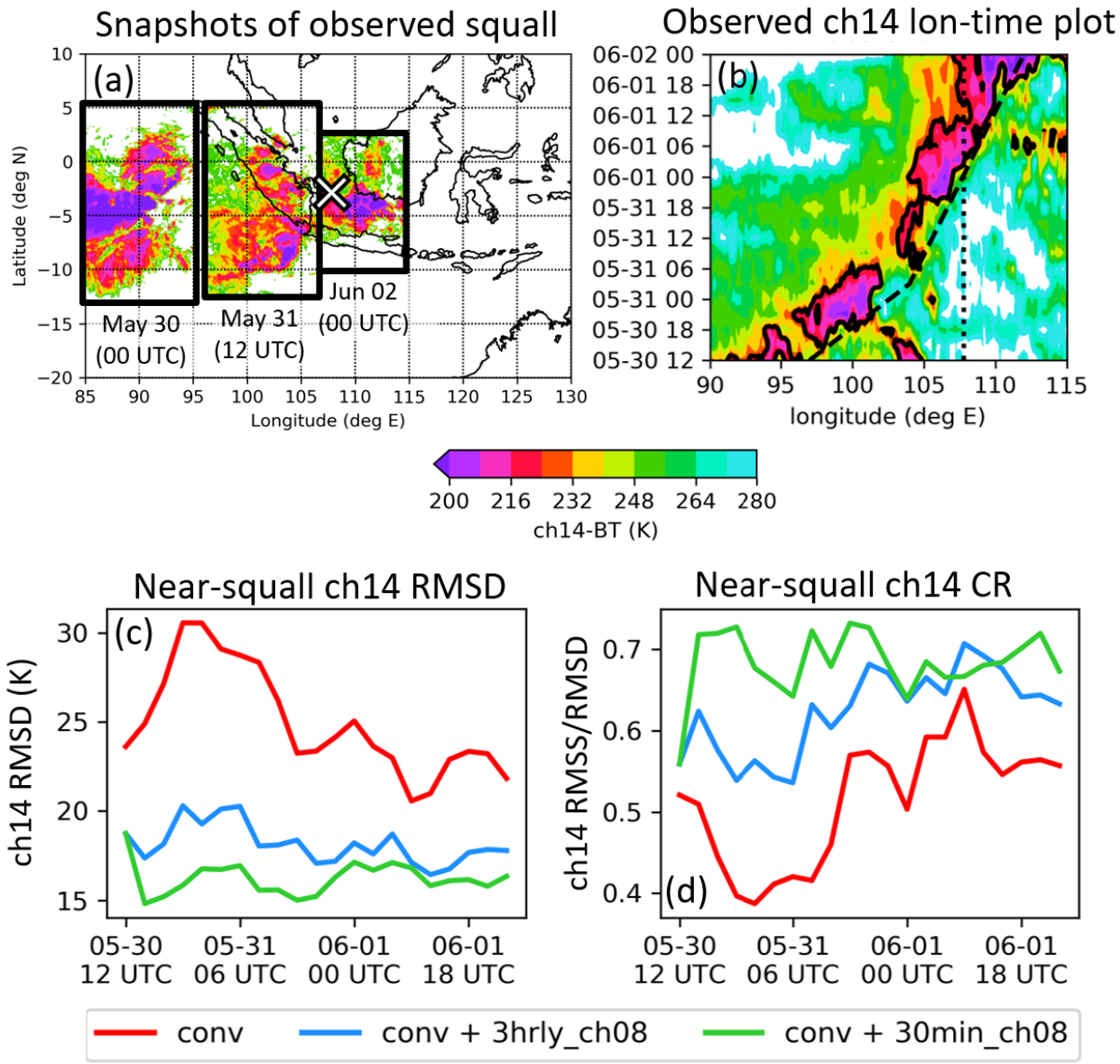


Figure 4.1: Plots showing (a) snapshots of the observed squall from AHI's channel 14, and (b) longitude-time diagram of the observed ch14-BT between 3.6 °S to 4.4 °S. The black-and-white cross in (a) indicates the location of the surface wind observations that will be used for validation later. The dashed black line in (b) indicates the estimated longitudes of the squall's leading edge at 4 °S, and the dotted black line indicates the position of the surface station used for validation. Also shown are (c) the ch14-BT RMSD of the analysis ensemble mean, as well as (d) the CR of the analysis ensemble. The RMSD and CR are computed in a squall-following region that is bounded by the 15 °S and 5 °N latitude circles. Furthermore, said squall-following region extends 15 degrees west of the leading edge, and 5 degrees east of the leading edge.

study. The reason is that earlier studies in the neighboring tropical Indian Ocean were able to obtain reasonable tropical MCSs without cumulus parameterizations (Chen et al., 2018b,c,a; Ruppert et al., 2020; Ruppert and Chen, 2020; Chen et al., 2020, 2022). These studies were able to do so because the 9-km grid spacing was enough to resolve mesoscale convective updrafts.

In this study, I initialized 50-member WRF ensemble simulations, starting on 30th May 2017 (0000 UTC). To construct the needed 50 sets of initial conditions, I utilized the European Centre for Medium-Range Weather Forecasts' (ECMWF's) fifth-generation global reanalysis data (ERA5; Hersbach et al. 2020) 10-member ensemble fields on 26 May (0000 UTC), 27 May (0000 UTC), 28 May (0000 UTC), 29 May (0000 UTC), and 30 May (0000 UTC), as well as the ERA5 reanalysis fields on 30 May (0000 UTC). To be precise, for each of the 5 dates, I generated 10 perturbations from the ERA5 10-member ensemble by subtracting the ensemble mean on said date from each of the 10-members. This procedure produced a set of 50 perturbations. I then added the ERA5 reanalysis fields on May 30 (0000 UTC) to the 50 perturbations to produce 50 sets of initial conditions. The ERA5 reanalysis fields were also used to generate the boundary conditions. Inother words, all 50-members were subjected to the same boundary conditions. The ensemble was then integrated forward for 12 hours to generate flow-dependent ensemble statistics on 30 May (1200 UTC). All data assimilation experiments will begin on 30 May (1200 UTC).

4.2.2 Observation sources

To examine the questions raised in section 1, I performed experiments where I assimilated non-IR conventional observations and satellite all-sky IR observations. The non-IR conventional observations include surface station, pilot, rawinsonde, radiosondes, and ship observations from the Global Telecommunication System, which are archived in the National Center for Atmospheric Research's Research Data Archive (NCAR RDA). The non-IR conventional ob-

servations also include satellite-derived atmospheric motion vectors from the Cooperative Institute for Meteorological Satellite Studies (CIMSS) Tropical Cyclone Archive. These non-IR conventional observations were assimilated every 3 hours in all experiments.

Aside from the non-IR conventional observations, I also assimilated AHI channel 8 (6.2 μm) infrared brightness temperature observations (ch08-BT). The AHI instrument is on board the Himawari-8 geostationary satellite, which is located above the Equator, at 140.7°E. Under clear-sky conditions, ch08-BT observations typically reflect the upper tropospheric water vapor content (Otkin 2012). These observations are generally available every 10 minutes.

4.2.3 Setup of data assimilation experiments

To investigate the important questions raised in section 1, I performed three DA experiments. The first experiment assimilated non-IR conventional observations every three hours (`conv` experiment). In the second experiment, I assimilated both ch08-BT and non-IR conventional observations every 3 hours (`conv+ch08_3hrly`). Finally, in the third experiment, I assimilated non-IR conventional observations every 3 hours, and ch08-BT observations every 30 minutes (`conv+ch08_30min`).

In this study, I employed the Pennsylvania State University WRF Ensemble Kalman Filter system (PSU-EnKF; Meng and Zhang (2007, 2008)) as our DA system. This version of the PSU-EnKF runs on the EnSRF (Whitaker and Hamill, 2002) discussed in Chapter 3, and is parallelized using the low-latency strategy proposed by Anderson and Collins (2007).

The PSU-EnKF algorithm employed here is a slightly modified version of the PSU-EnKF described in Chan et al. (2020b) [the ensemble square-root filter algorithm (EnSRF) in their section 2c]. The only modification to the algorithm is

that the mean of the forecasted observations is no longer calculated using the ensemble mean. Instead, the mean of the forecasted observations is calculated by first computing the simulated observations for all forecast members, and then taking the ensemble mean of the result. This modification was made to mitigate the artificial dry bias produced by the previous EnSRF implementation when infrared observations are assimilated. The cause of this artificial dry bias is described in Section 3b of Chan et al. (2020b).

Aside from that, I also performed 80% relaxation to prior perturbations to prevent filter divergence (Zhang et al., 2004). The variables updated by the PSU-EnKF include the three-dimensional winds, water vapor mixing ratio, liquid cloud mixing ratio, ice cloud mixing ratio, rain mixing ratio, snow mixing ratio, graupel mixing ratio, temperature, and pressure. All localizations are performed using the Gaspari-Cohn 1999 fifth-order polynomial (Gaspari and Cohn, 1999).

With regards to the assimilation of conventional observations, I assimilated the atmospheric motion vectors with a horizontal radius of influence (HROI) of 100 km, and no vertical localization. Sounding and aircraft observations were assimilated with a 700 km HROI and a 5 model layer vertical radius of influence (VROI). With regards to surface observations, Metéorologique Aviation Régulière (METAR) observations, surface synoptic observations (SYNOP) observations and ship observations were respectively assimilated with HROIs of 600 km, 300 km and 1400 km. All three surface observation types were assimilated with a 45 model layer VROI.

To assimilate the ch08-BT observations, I employed the Community Radiative Transfer Model (CRTM), release 2.3.0, to calculate the IR AHI brightness temperatures from the WRF ensemble. The ch08-BT observations were thinned to a spacing of 30 km, and then assimilated with a horizontal radius of influence of 100 km (Chan et al., 2020b). Furthermore, I employed the adaptive

observation error inflation scheme (AOEI) of Minamide and Zhang (2017), as well as the adaptive background error inflation scheme (ABEI) of Minamide and Zhang (2019), to assimilate the ch08-BT observations. No vertical localization was performed for the ch08-BT observations. Better results might be possible with more tuning, which can be explored in future studies.

Aside from these settings, no ch08-BT observations were rejected in this study. Observations with large innovations are often rejected because these observations might be erroneous (*e.g.*, Järvinen and Undén (1997); Cardinali et al. (2003)). In my setup, the AOEI inflates the observation error of ch08-BT observations with large innovations. The AOEI thus weakens the magnitude of the analysis increments coming from these large innovation ch08-BT observations. In other words, AOEI functions as a safety measure against potentially problematic ch08-BT observations. There was thus no need for me to reject any ch08-BT observations in this study.

It should also be noted that no bias correction for the ch08-BT observations were employed here because I found the ch08-BT biases are small in this tropical squall line case. As discussed in Chan et al. (2020b), the square of the bias is a component of the mean-squared difference (MSD) between the ensemble mean and observed ch08-BT. The `conv`, `conv+ch08_3hrly` and `conv+ch08_30min` experiments' squared prior ch08-BT biases are 0.49 K^2, 0.25 K^2, and 0.00 K^2, respectively. In contrast, the prior ch08-BT MSDs of the `conv`, `conv+ch08_3hrly` and `conv+ch08_30min` experiments are 68.9 K^2, 23.0 K^2, and 16.0 K^2, respectively. In other words, the squared prior ch08-BT biases only account for less than 1.1% of the prior ch08-BT MSD. Thus, I consider the ch08-BT biases to be small and did not employ bias correction when assimilating ch08-BT observations.

4.3 Results and discussion

4.3.1 Description of the observed tropical squall line

Before delving into the corrective impacts of assimilating GeoIR, I will describe the observed behavior of this study's tropical squall line case. Tropical squall lines are a frequently occurring type of tropical MCSs over the Maritime Continent (Lo and Orton, 2016; Chan et al., 2019; Sun et al., 2020). On 30th May 2017, the AHI observed a southeastward propagating MCS over the equatorial Indian Ocean (Figure 4.1a). By 1500 UTC on 30th May, the mesoscale convective system evolved into a southeastward propagating, bow-shaped squall line. At this time, the squall line was generally aligned in the northeast-southwest direction, with the northeastern tip of the squall line grazing the west coast of Sumatra (not shown here). For the next 12 hours, AHI observations indicate that the squall did not penetrate the mountain range that runs along most of the west coast of Sumatra.

Around 0900 UTC on 31st May, the northeastern half of the squall line penetrated the mountain range. Within the next 3 hours, the northeastern half and the southwestern half of the squall were split apart. For the remainder of the study, I will focus on the northeastern half because it passed over surface meteorological stations. The northeastern half developed a north-south alignment and propagated eastwards (Figure 4.1a), passing through Sumatra, and then the Straits of Malacca. This propagation behavior and morphology is rather typical of squalls in the Sumatra-Borneo region(Lo and Orton, 2016; Chan et al., 2019; Sun et al., 2020). By 0000 UTC on 2nd June, the squall line reached southwestern coast of Borneo (Figure 4.1a) and proceeded to dissipate.

It is worth noting here that this tropical squall line case coincides with a westerly wind burst. According to the ERA5, the observed squall line is collo-

cated with the advancing edge of a westerly wind burst (not shown). Furthermore, the Australian Bureau of Meteorology's (BoM's) Real-time Multivariate MJO index Wheeler and Hendon (2004) indicated that during the squall line's time frame, the MJO's active phase was advancing from the Indian Ocean into the Maritime Continent. As westerly wind bursts are often associated with the onset of the MJO's active phase (Zhang, 2005), the MJO index supports the association between the squall line and a westerly wind burst seen in the ERA5 data. Future work can examine the association between westerly wind bursts and tropical squall lines in the region.

4.3.2 Assimilating half-hourly ch08-BT improved analyzed clouds

The first question I will address in this study is whether adding all-sky IR observations into DA can improve the analyses and prediction of a tropical squall line. To that end, I will be comparing the `conv` and the `conv+ch08_30min` experiments. Furthermore, to focus on the squall line itself, the experiments will be compared in an area around the squall line's leading edge.

To determine the position of the leading edge, I computed the meridional average of the observed AHI channel 14 brightness temperatures (ch14-BT) for latitudes between 3.6 °S and 4.4 °S. The result is plotted as a longitude-time diagram in Figure 4.1b. Because the squall line's clouds show up as strong cold signals in ch14-BT, I can determine the position of the leading edge of the squall line from Figure 4.1b. The inferred positions of the leading edge are indicated by a dashed black line in Figure 4.1b. For the rest of this article, I will be comparing the experiments in the vicinity of these inferred positions.

I will begin with examining the impacts of including half-hourly ch08-BT observations, on top of 3-hourly conventional observations, on the cloud fields. Since ch14-BT was not assimilated and does reflect the presence and absence

of clouds, the root-mean-square difference (RMSD) between the analyzed and observed ch14-BT can be used to gauge the quality of the analyzed cloud fields. To be more precise, I will be comparing the ch14-BT RMSDs computed for a region surrounding the squall line's leading edge (Figure 4.1c). The region covers a rectangular area between 10 °S and 5 °S, as well as between 15 degrees west and 5 degrees east of the leading edge (Figure 4.1b). As seen in Figure 4.1c, `conv+ch08_30min`'s RMSDs are always substantially smaller than conv's RMSD. On average, `conv+ch08_30min`'s RMSD (16.2 K) is ∼35% smaller than `conv`'s RMSD (25.0 K). In other words, the assimilation of IR observations, on top of conventional observations, can improve analyses of the squall line's cloud field. This improvement is in line with earlier work (Ying and Zhang, 2018).

4.3.3 Assimilating half-hourly ch08-BT improved analyzed outflow position

Aside from improving the cloud field, the inclusion of half-hourly ch08-BT observations also improved the position of the squall line surface outflow. Surface outflows can generate converging surface winds ahead of the squall line, and thus encourage the formation of new cells ahead of the squall line [*e.g.*,Gamache and Houze (1982); Moncrieff and Liu (1999)]. Since the initiation of new cells ahead of the squall line is important for squall line propagation [*e.g*, Fovell et al. (2006)], improving the position of the analyzed surface outflow can improve the analyses and prediction of the squall line.

To see the improvement in surface outflow, consider that the outflow in this squall line is characterized by strong surface westerlies. Figure 4.2a and 3c show longitude-time diagrams of the meridionally-averaged analysis mean zonal wind at 300-m above sea level in the `conv` and `conv+ch08_30min` experiments. The thick black contours trace the observed 224 K ch14-BT contour, which indicates the position of the squall line. The meridional average was per-

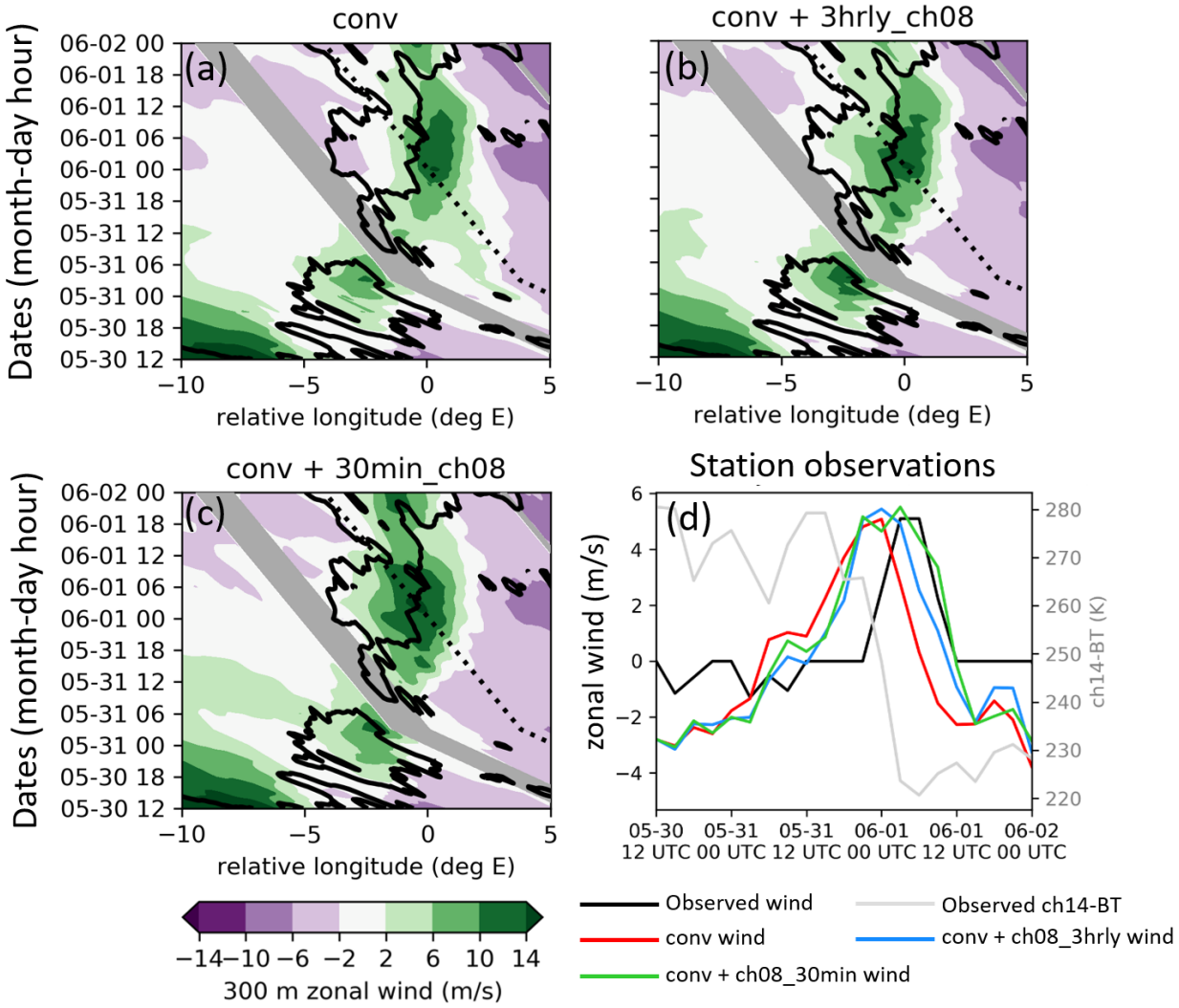


Figure 4.2: Squall-relative longitude-time diagrams of the analyzed zonal wind, averaged between 3.6 °S to 4.4 °S, at 300 m above sea level (a,b,c). The thick black contours in panels (a), (b), and (c) indicate the 224 K contour of the observed ch14-BT, averaged between the same latitudes. Also, the gray shading indicates the storm-relative longitudes of the mountain range along the west coast of Sumatra, where no 300-m zonal wind information is available in the model. Finally, panel (d) shows time series of the observed surface zonal wind, analyzed zonal wind, as well as the observed ch14-BT at a surface station located at 2.7 °S, 107.8 °E.

formed between 3.6 °S and 4.4 °S. Between 1200 UTC 31 May and 1800 UTC 1 June, conv's analyzed outflow is mostly dislocated east of the observed leading edge (Figure 4.2a; black contours). In contrast, the analyzed zone of strong westerlies in conv+ch08_30min overlaps substantially with the observed squall (Figure 4.2c). These longitude-time diagrams suggest that the half-hourly assimilation of ch08-BT observations likely improved the positions of the surface outflow.

To further confirm the improvement of the surface outflow position, I examined three-hourly unassimilated surface zonal wind observations from a surface station located at 2.7 °S, 107.8 °E (marked with white cross in Figure 4.1a). These surface observations were obtained from the Integrated Surface Database (ISD), which is maintained by the National Oceanic and Atmospheric Administration (NOAA). The time-series of the observed and analyzed zonal winds for this station is plotted in Figure 4.2d.

As can be inferred from Figure 4.2d, conv+ch08_30min's analyzed squall line surface outflow outperformed that of conv in several ways. The first way is in terms of the onset time of the squall line surface outflow at the station. The onset of the squall line surface outflow is indicated by strengthening westerlies. As can be seen in Figure 4.2d, the onset time seen in conv+ch08_30min's analyzed surface outflow (around 1200 UTC May 31) is slightly closer to the observed onset time (around 1800 UTC May 31) than that of conv (around 0600 UTC May 31). Aside from the difference in onset times, Figure 4.2d indicates that conv's squall line outflow leaves the station (around 0600 UTC June 1) 6 hours earlier than the observed squall line (around 1200 UTC June 1), whereas conv+ch08_30min's squall line outflow leaves the station at roughly the same time as the observed squall line. These differences in onset and exit times imply that conv's squall line outflow is displaced east of the actual outflow, and that this displacement error is largely reduced in conv+ch08_30min. This reduction in surface outflow displacement error is consistent with Figures 3a

and 3c.

It should also be noted here that the `conv+ch08_30min` also outperforms `conv` is in terms of the fit to the unassimilated zonal wind observations when the squall line is passing over the surface station (from 2100 UTC on 31 May to 1200 UTC on 1 June). This improvement is likely because the eastward displacement error is substantially smaller in `conv+ch08_30min` than in `conv`.

Interestingly, the improvements in surface outflow position introduced by spatiotemporally dense IR DA occurred in the absence of dense surface observations. Previous studies have suggested that dense surface observations can improve squall line surface outflows and the associated gust front positions (*e.g.*, Chen et al. (2016)). Our results thus suggest that spatiotemporally dense IR observations can be a potential substitute for dense surface observations when constraining errors associated with squall line surface outflows and gust front positions.

The surface outflow position improvements introduced by half-hourly IR DA can be plausibly explained by the combination of two complementary effects. The first effect comes from the cloud field improvements introduced by half-hourly ch08-BT DA. These cloud field improvements include improvements to the position of the analyzed squall line clouds, which in turn, leads to improved positioning of the squall line rainfall. Since the surface outflow is induced by rainfall, it is plausible for the improved rainfall position to improve the position of the surface outflow.

The second effect arises from the positive correlation between forecasted ch08-BT just behind the observed leading edge, and the ~950 hPa zonal wind around the leading edge of the surface outflow (henceforth, gust front). For instance, at 0000 UTC on 1 June, the forecasted ensemble mean gust front is at roughly 1 °E relative longitude (similar to the analyzed gust front in Figure

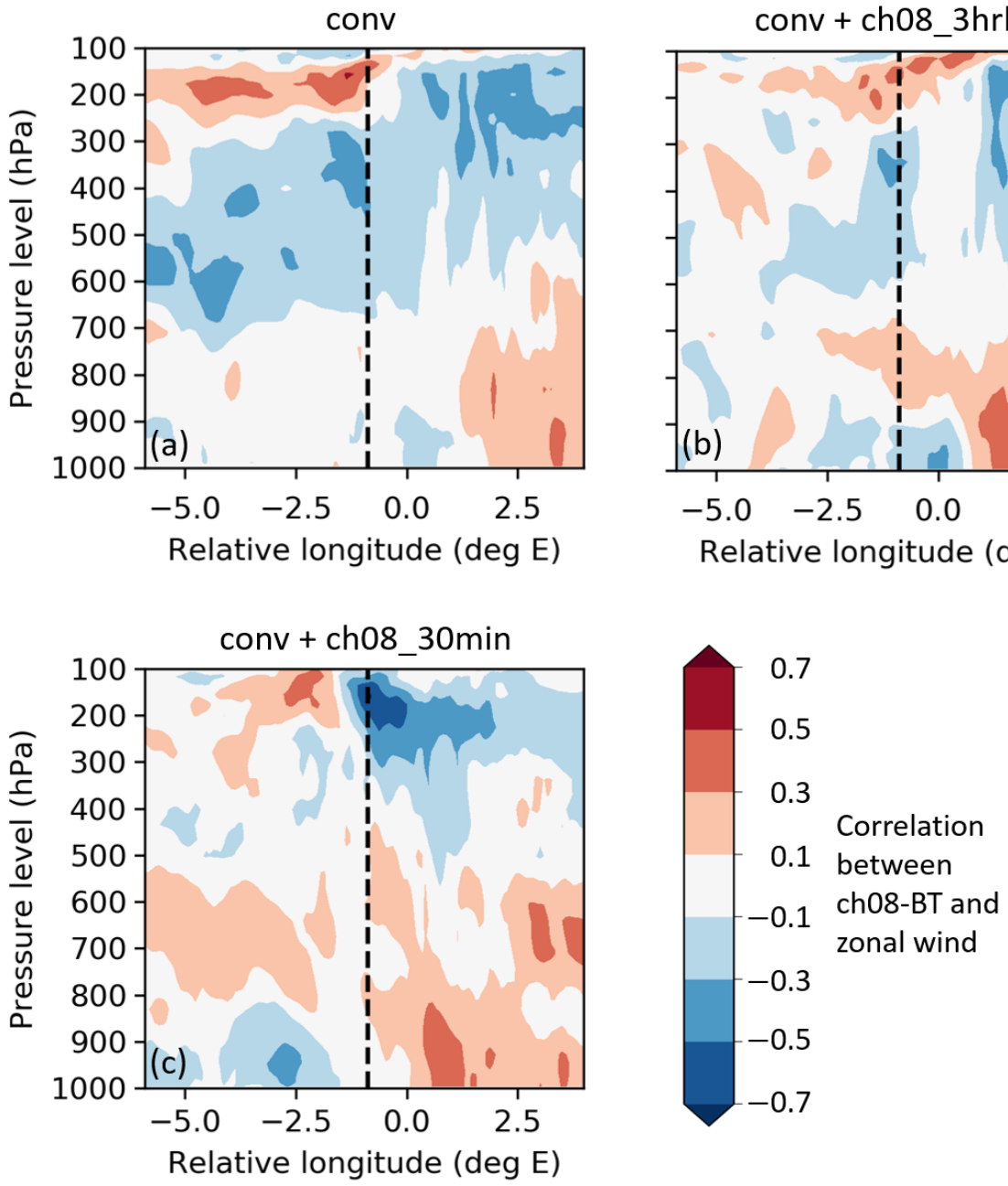


Figure 4.3: Correlations between ch08-BT and zonal wind on June 1st (00 UTC) for the `conv` experiment (a), the `conv+ch08_3hrly` experiment (b), and the `conv+ch08_30min` experiment (c). These correlations are plotted for a ch08 observation located 1 degree west from the squall's leading edge (vertical dashed lines). Note that the plotted values are the average of correlations across 10 zonal cross-sections between 3.6 °S to 4.4 °S.

4.2c). Likewise, Figure 4.3c indicates that the ∼950 hPa zonal wind at ∼1 °E relative longitude is positively correlated to forecasted ch08-BT just behind the leading edge (-1 °E relative longitude). In other words, said positive correlation is collocated with the gust front.

To see how this positive gust front zonal wind correlation allows the EnKF to shift the gust front, consider that the forecasted squall line clouds tend to be ahead (In other words, displaced eastwards) of the observed squall line clouds. Since the observed squall line's highest clouds tend to occur at around -1 °E relative longitude (not shown, but can be inferred from Figure 4.1b), this eastward displacement indicates that the forecasted clouds at -1 °E relative longitude should be lower than that of the observations. In other words, the forecasted ch08-BT is higher than that of the observations at -1 °E relative longitude. This results in a negative ch08-BT innovation at -1 °E relative longitude. Because of the positive correlation, this negative innovation results in a negative westerly (or easterly) analysis increment at the gust front. Since the wind in the gust front is mostly westerly, the analysis increment weakens the westerly at the gust front. This weakening means that the forecasted gust front is pushed westward. As such, the assimilation of ch08-BT observations near the observed leading edge improved the position of the gust front through the correlations between ch08-BT and the gust front's zonal wind.

4.3.4 Assimilating half-hourly ch08-BT improved deterministic forecasts of cloud fields

Aside from improving the analyzed cloud fields and surface outflow position, the half-hourly assimilation of IR observations also improved the short-term deterministic forecasts of the squall line. These deterministic forecasts were initiated from the EnKF analysis means. Figure 4.4 shows the meridionally-averaged, deterministically forecasted ch14-BT. Like Figure 4.1c, 3, and 4, the meridional average was done between 3.6 °S and 4.4 °S. As can be seen in

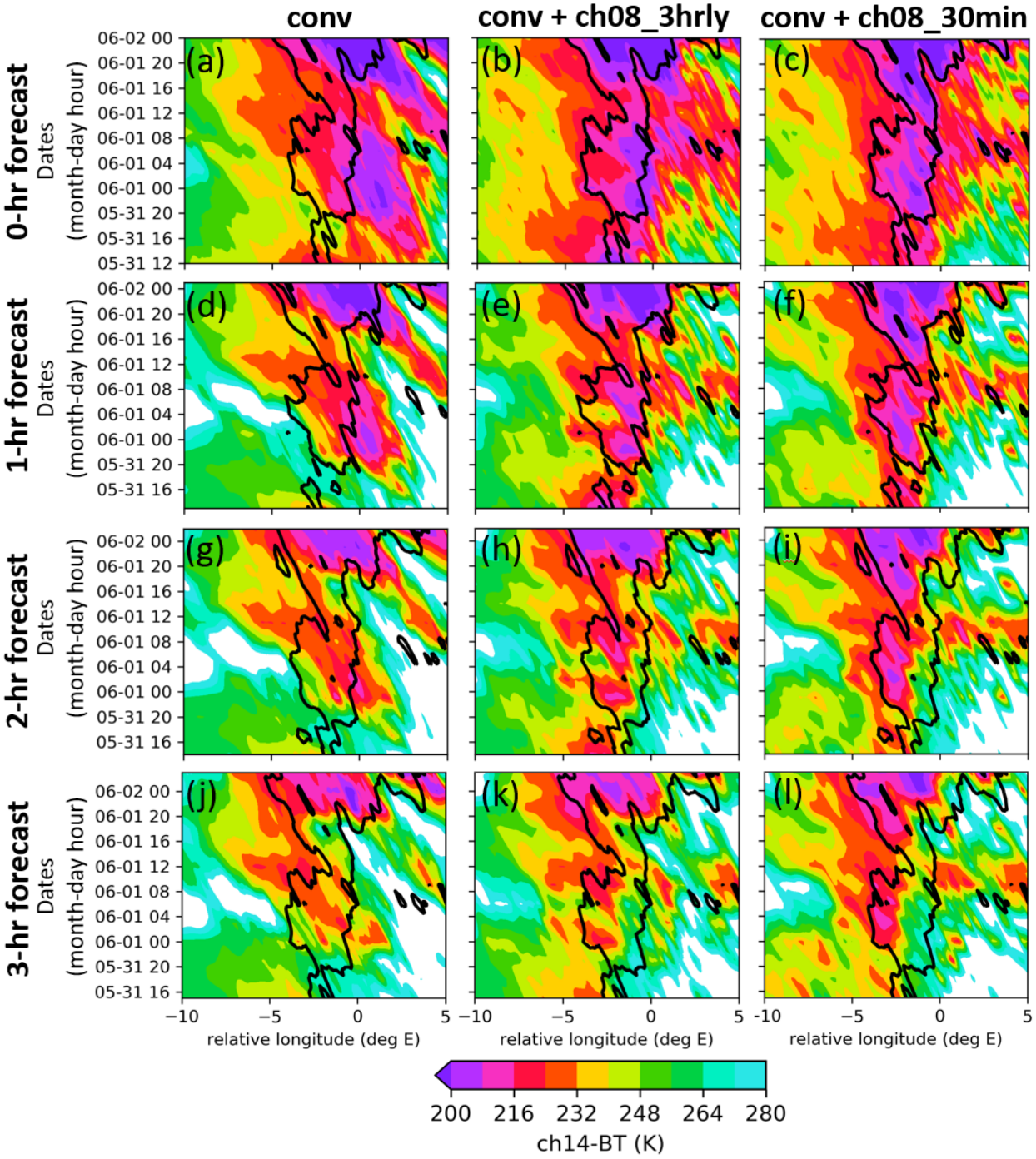


Figure 4.4: *See next page for caption*

Figure 4.4 *(previous page)*: Storm-relative longitude-time diagrams of deterministically forecasted ch14-BT (shading), as well as the 224 K contour of the observed ch14-BT (black contours). All plotted ch14-BT are averaged between 3.6 °S to 4.4 °S. Panels a, d, g & j show the deterministic forecasts from the conv experiment. Similar plots were also produced for the `conv+ch08_3hrly` (b, e, h & k) and `conv+ch08_30min` (c, f, i & l) experiments. The forecasted ch14-BT and the observed ch14-BT at the corresponding times are shown for the start times of the forecasts (a, b & c), at a lead time of 1 hour (d, e & f), at a lead time of 2 hours (g, h & i), and a lead time of 3 hours (j, k & l). Note that the deterministic forecasts from 13 initiation times are shown in each panel. The first deterministic forecasts were initiated on May 31 (12 UTC), and subsequent deterministic forecasts were initiated every 3 hours, up till and including June 2 (00 UTC).

Figure 4.4, for the first 3 hours of the deterministic forecasts, the positions of `conv+ch08_30min`'s deterministically forecasted clouds are closer to the observations than those of `conv`, especially for the forecasts initialized after 1200 UTC June 1. Furthermore, the squall-relative ch14-BT RMSD (calculated the same was as Figure 4.1c) of `conv+ch08_30min`'s deterministic forecasts are lower than those of the conv (not shown) for the first three hours. To be more precise, on average, `conv+ch08_30min`'s ch14-BT deterministic forecast RMSDs are about 20% smaller than those of `conv` from initiation to up to a lead time of 2 hours. For lead times of 3 hours and longer, the ch14-BT deterministic forecast RMSDs of both experiments are statistically indistinguishable. These improvements to the short-term deterministic forecasts of cloud fields are akin to those seen in Chan et al. (2020b).

4.3.5 Reducing the frequency of ch08-BT DA degraded analyses and deterministic forecasts

Another important question raised in section 1 is how the accuracy and prediction of the tropical squall line vary with the frequency of IR DA. To examine this question, I will be comparing the `conv+ch08_30min` and `conv+ch08_3hrly` experiments.

Reducing the frequency of ch08-BT DA generally degraded the analyses and deterministic forecasts of cloud fields (Figures 4.1 and 4.4). According to Figure 4.1c, decreasing ch08-BT DA frequency from half-hourly to three-hourly generally increased in the analyses' ch14-BT RMSD. On average, the RMSDs increased by 10%, from 16.2 K to 18.1 K. Furthermore, Figure 4.4 indicates that reduction in ch08-BT DA frequency degraded the deterministically forecasted positions of the squall line for the three hours of the deterministic forecast, and the degradations are especially noticeable for the forecasts initialized after 1200 UTC June 1 (Figures 4.4e and 4.4h). Taken together, these degradations imply that a higher frequency of data assimilation is preferable for short-term cloud field analyses and forecasts.

The analyzed positions of the squall line's outflow were also degraded when the frequency of ch08-BT DA was reduced from half-hourly to three-hourly. Between 31 May (1200 UTC) and 1 June (1800 UTC), `conv+ch08_3hrly`'s analyzed strong zonal wind zone does not overlap much with the observed squall (Figure 4.2b), whereas that of the `conv+ch08_30min` substantially overlaps the observed squall (Figure 4.2c). Furthermore, while the validation surface station data indicates the analyzed outflows of `conv+ch08_3hrly` and `conv+ch08_30min` arrived at the station at the same time (Figure 4.2d), the `conv+ch08_30min`'s analyzed outflow is clearly better than that of `conv+ch08_3hrly` when the squall line passed over the surface station during 0000 UTC – 1200UTC on June 1. In other words, high frequency data assimilation is preferable for improving the analyzed position of storm outflows.

This degradation in squall line outflow position is also consistent with the ensemble correlations between simulated ch08-BT and zonal wind (Figure 4.3b). While similar near-surface dipole correlation patterns exist in both the `conv+ch08_30min` and `conv+ch08_3hrly` experiments, the dipole pattern in `conv+ch08_3hrly` is dislocated 1.5 degrees east of the ch08-BT observation location. In other

words, conv+ch08_3hrly's capability in moving the squall line surface outflow position westwards is limited when compared to that of conv+ch08_30min.

It should be noted here that while decreasing the frequency of ch08-BT DA did dramatically degrade the quality of the analyzed cloud fields, analyzed outflow, and forecasted clouds, these quantities in conv+ch08_3hrly still outperformed those conv. In other words, even a moderate frequency of ch08-BT DA was able to improve these quantities beyond what is achieved by non-IR conventional observation DA.

4.4 Conclusions and areas for future work

In this study, I have delved into the hitherto unexplored dynamic and thermodynamic impacts of assimilating IR observations into the case of a tropical squall line. In line with earlier work [*e.g.*, Ying and Zhang (2018)], I confirmed that introducing ch08-BT observations into DA improved the analyzes and predictions of a tropical squall line's cloud field. Even though the ch08-BT observations are insensitive to lower-tropospheric zonal winds, the assimilation of these observations improved the position of the squall line's lower tropospheric outflow zone. This improvement was likely due to the improved cloud fields, as well as the near-surface dipole forecast ensemble correlations between ch08-BT and lower-tropospheric zonal winds. Finally, increasing the frequency of ch08-BT observations improved the analyzed cloud field, forecasted cloud field, and the analyzed outflow position.

Since this study is the first of its kind to specifically examine the improvements from introducing GeoIR observations into tropical MCS analyses and prediction, there are many areas for future work. These areas can be broken into three types: completeness, the assimilation of other observations and areas for future DA algorithm development.

The first area for completeness concerns the fact that I have only examined the impacts on one tropical squall line in the Maritime Continent. As such, future work can examine if similar improvements can be found in other tropical MCSs, even in other tropical areas. Second, comparisons against surface station temperature observations (not shown) indicate that the simulated squall line cold pools tend to be much stronger than observed. These overly strong simulated cold pools might be due to model errors in the surface parameterization or radiation schemes. Future work can thus examine tuning these schemes to produce more realistic cold pools. Aside from that, future work can also explore whether the improvements obtained here are retained when the model's horizontal grid spacing is reduced to a value that allows for the explicit resolution of convective updrafts (usually 1 km or less).

Future work can also explore the assimilation of other uncommonly used observations. For instance, it is possible that the inclusion of observations from additional water vapor channels can improve the analysis and prediction tropical MCSs. This improvement might be possible because different channels perceive different parts of the atmospheric column (Schmit et al., 2005). Aside from that, several studies have indicated that sub-hourly AMV data assimilation has the potential to improve forecasts of moist convection (Otsuka et al., 2015; Kunii et al., 2016). Right now, AMVs are generally assimilated every 3 6 hours. Thus, the impact of frequently assimilating infrared-derived atmospheric motion vector observations is also another avenue of future study.

Finally, future work can also delve into methods that address certain assumptions that are broken in IR DA. One important assumption in the EnKF is that the simulated observations have an approximately linear relationship with modelled atmospheric variables. However, the relationship between simulated IR observations and modelled atmospheric is often nonlinear [*e.g.*, Steward (2012)]. Other cutting-edge DA methods might be able handle this nonlinear-

ity. Such methods include particle filters [*e.g.*, van Leeuwen (2011); Penny and Miyoshi (2016); Poterjoy (2016)], iterative EnKF methods (Lorentzen and Naevdal, 2011), and Gaussian-mixture model methods [*e.g.*, Anderson and Anderson (1999); Sondergaard and Lermusiaux (2013b); Chan et al. (2020a)]. Aside from linearity, the EnKF also assumes that the forecast ensemble follows a Gaussian distribution. However, the displacements of cloud features among ensemble members can result in non-Gaussian properties, breaking this Gaussian assumption. As such, future work can explore methods that address displacement errors. Such methods include variants of the feature calibration and alignment technique (Hoffman et al., 1995; Hoffman and Grassotti, 1996; Brewster, 2003; Nehrkorn et al., 2003, 2015; Stratman et al., 2018) and the recently developed multiscale alignment data assimilation (Ying, 2019).

Chapter 5

A high resolution tropical MCS reanalysis (TMeCSR)

5.1 Overview and goals

In the previous chapter, I have demonstrated that assimilating GeoIR observations can improve the analyses and forecasts of a tropical MCS. This suggests that GeoIR DA can be used to create a high resolution tropical MCS reanalysis data product. Currently, no reanalysis product capable of explicitly resolving tropical MCSs exists. The creation of such a product is thus likely beneficial for the tropical MCS research community.

This chapter will discuss the creation and validation of a hitherto non-existent tropical MCS reanalysis (TMeCSR) product. The TMeCSR is created by refining the popular global ERA5 reanalysis (Hersbach et al., 2020) with high-resolution GeoIR brightness temperature (BT) observations and an MCS-resolving regional model (WRF). This dataset is available hourly for the three boreal summer months (June, July, August) of 2017 and covers a region spanning the majority of the Indian Ocean, tropical/subtropical continental Asia, the Maritime Continent, and the Western Pacific. The TMeCSR ensemble means, centermost

members and ensemble variances are publicly available on the Pennsylvania State University's Data Commons.[1]

Note that the three boreal summer months of 2017 were chosen because unusually widespread flooding occurred in continental Asia. More than 50 million people were affected and more than a thousand were killed by these floods. Since MCSs are responsible for the majority of rainfall in the region (Mohr et al., 1999; Houze, 2004; Nesbitt et al., 2006; Liu, 2011; Roca and Fiolleau, 2020), the TMeCSR would be useful for future studies of this unusually anomalous monsoon season.

Although limited in spatial and temporal coverage, TMeCSR captured more than 1200 MCS events, as estimated using the MCS tracking algorithm of Feng et al. (2021b). Furthermore, as I will show later, the MCSs that occurred during the TMeCSR period resemble the climatological statistics of MCSs in the region. Therefore, TMeCSR provides the tropical MCS community with a trove of data suitable for in-depth investigations of real-world MCSs.

5.2 Materials and methods

5.2.1 Setup of the 9-km WRF ensemble

The data assimilation method used to generate TMeCSR (described later) requires ensembles of forecast model realizations to assimilate observations. To explicitly resolve tropical MCSs, I employed the Advanced Research WRF model (Skamarock et al., 2008), version 3.8.1, with a 9-km horizonal grid spacing. The TMeCSR WRF domain covers the geographical area plotted in the panels of Figure 5.4 and has 800 × 640 horizontal grid points. Furthermore, the

[1] For access to the full prior and posterior ensemble fields, contact Dr. Xingchao Chen.

WRF model domain has 45 model levels, with the bottommost 9 levels within the lowest 1-km of the model, and a model top pressure of 20 hPa. The WRF physics schemes used here are same as in Chapter 4.2.1. Furthermore, no cumulus parameterizations are used (see Chapter 4.2.1).

The WRF ensemble is constructed by applying 50 perturbations to the ERA5-sourced initial conditions. These 50 perturbations originate from the ECMWF's 50-member ensemble fields from The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble [TIGGE; Swinbank et al. (2016)]. These 50-member TIGGE fields were also valid for 12 UTC on 19th May 2017.The 50-member WRF ensemble was then integrated freely forward for 12 hours to generate forecast ensemble statistics that are specific to the WRF model dynamics and instantaneous atmospheric conditions (often called "flow-dependent statistics"). For this initial 12-hour integration (or spin up) and for the entirety of the TMeCSR dataset, the WRF ensemble utilizes boundary conditions from ERA5. In other words, all ensemble members use identical boundary conditions. DA begins on 00 UTC on 20th May 2017.

5.2.2 Observations used

A variety of observations was utilized to generate TMeCSR. First, TMeCSR utilized observations from the World Meteorological Organization (WMO) Global Telecommunication System (GTS). GTS observations are available from the National Center for Atmospheric Research (NCAR) Research Data Archive (RDA). More specifically, I assimilated sounding observations, surface station observations, ship observations, aircraft observations, and satellite-retrieved atmospheric motion vector (AMV) observations from the GTS.

TMeCSR also assimilated water vapor channel brightness temperature data (WV-BT) from Meteosat-8's SEVIRI (channel 5) and Himawari-8's AHI (channel

8).[2] Both sensors have a central wavelength of 6.25μm. Furthermore, under clear-sky conditions in the tropics, these sensors are typically sensitive to the atmospheric layer between 500 hPa to 100 hPa (i.e., the middle to upper troposphere; CIMSS).

Both the SEVIRI and AHI WV-BT observations were thinned to a separation distance of ~27-km.[3] Since the half-way point between the two satellites' sub-satellite longitudes is 91.1°E, SEVIRI WV-BT observations were assimilated to the west of 91.1°E, and AHI WV-BT observations were assimilated to the east of 91.1°E.

5.2.3 TMeCSR Workflow

Figure 5.1(a) illustrates the workflow used to generate the TMeCSR dataset. TMeCSR begins at 00 UTC on 20th May 2017, after 12 hours of spin up from 12 UTC on 19th May 2017. On this date, the prepared WRF ensemble (blue bubble at 00 UTC on 20th May in Figure 5.1(a)) and relevant observations were ingested by the PSU-EnKF to produce an observation-constrained analysis ensemble (green bubble on the same date). The analysis ensemble is then integrated for 1 hour to produce a forecast ensemble at 01 UTC on 20th May 2017 (blue bubble at said date in Figure 5.1(a)). Observations valid at 01 UTC on 20th May 2017 were then assimilated to produce the analysis ensemble for this date (green bubble on 01 UTC 20th May in Figure 5.1(a)). This cycle of time integration and data assimilation repeats hourly up till, and including, 23 UTC on 19th August 2017.

[2]The AHI WV-BT data was obtained from the Japanese Aerospace Exploration Agency (JAXA) P-Tree system and the SEVIRI WV-BT data was obtained from the European Organisation for the Exploitation of Meteorological Satellite (EUMETSAT) Data Center. The latter was accessed through EUMETSAT's Earth Observation Portal (EOP).

[3]Note that no averaging was performed over multiple raw observations to produce the to-be-assimilated WV-BT observations (i.e., the super-observation strategy was not employed).

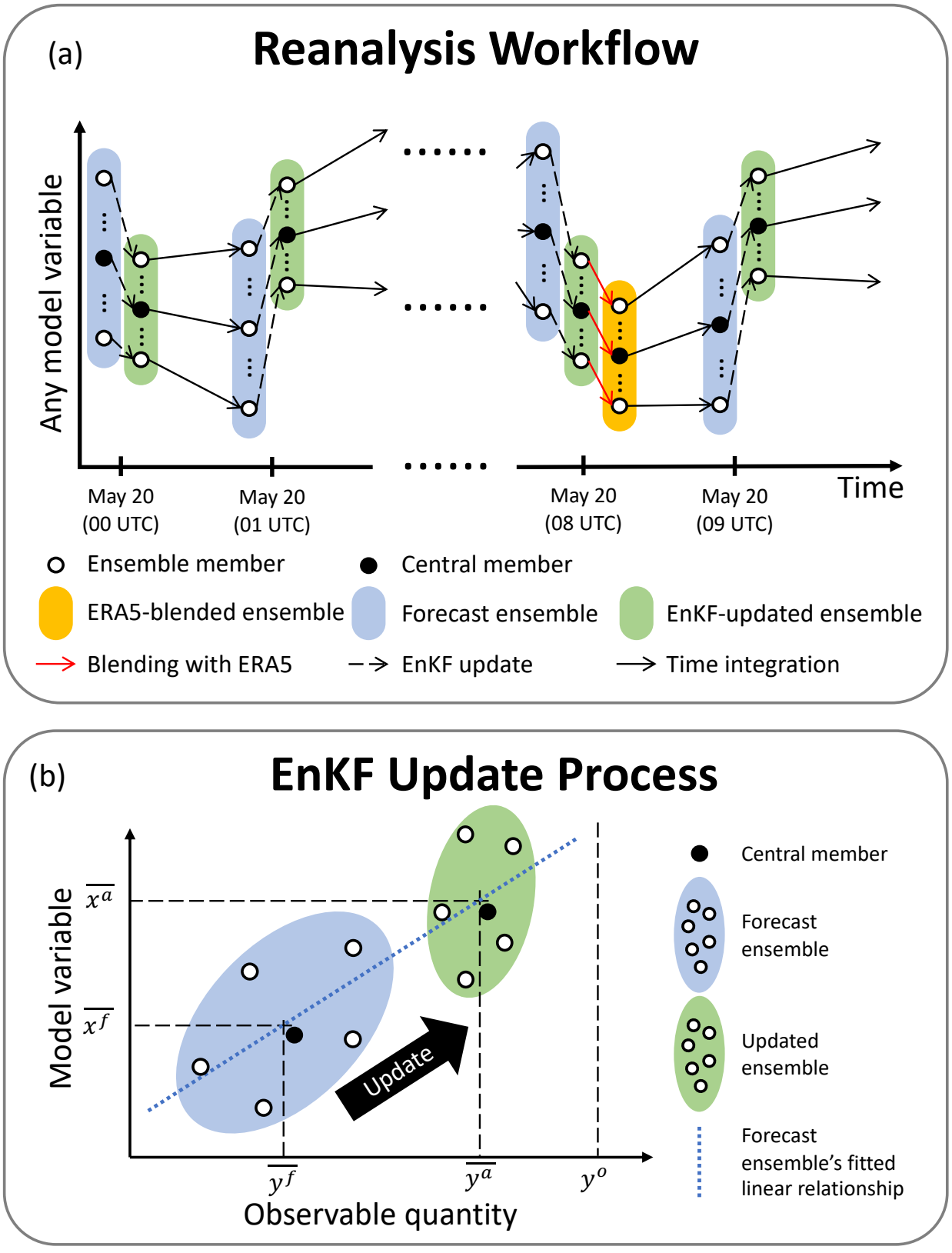


Figure 5.1: *See next page for caption*

Figure 5.1 *(previous page)*: Schematic diagrams representing (a) the overall workflow and (b) the EnKF update process used to generate the TMeCSR dataset. At the start of each hour (e.g., 00 UTC on 20th May), the TMeCSR assimilates observations into an ensemble of WRF forecasts using the PSU-EnKF. The resulting EnKF-updated ensemble is then integrated to the next hour using the WRF model. This hourly assimilation-integration cycle is performed throughout the three summer months of 2017. Additionally, every 12 hours (08 UTC and 20 UTC of each day) the large-scale information from the EnKF-updated ensemble is mixed with large-scale information from ERA5 prior to running the WRF integration (red arrows in panel (a)). The EnKF process updates the WRF forecast ensemble every hour by first linearly regressing the forecast ensemble's model variables against the forecasted observable quantities. The WRF ensemble is then shifted in phase space in accordance with the statistics of the WRF ensemble and prescribed observation errors (black block arrow in (b) with the text "Update" in white). The EnKF also contracts the spread in the ensemble to represent the greater confidence in the ensemble mean values after data assimilation.

Note that starting from 08 UTC on 20th May, I blended the large-scale information of the analysis ensemble mean with that of ERA5 (red arrows and orange bubble in Figure 5.1(a)). This blending is done every 12 hours (i.e., every day at 08 UTC and 12 UTC) throughout the time period of TMeCSR. The blending procedure will be described later.

5.2.4 PSU-EnKF setup

The PSU-EnKF system used in this study runs on the EnSRF algorithm described in Chapter 3.3, and parallelized using the high-latency strategy of Anderson and Collins (2007). See Chapters 3.5 and 4.2.3 for the heuristic measures, radii of influence and RTPP relaxation coefficients used in this study. Note that it is possible to produce better results by further tuning the various radii of influence (ROIs). Our tuning of the ROIs is limited as the tuning process involves rerunning TMeCSR.

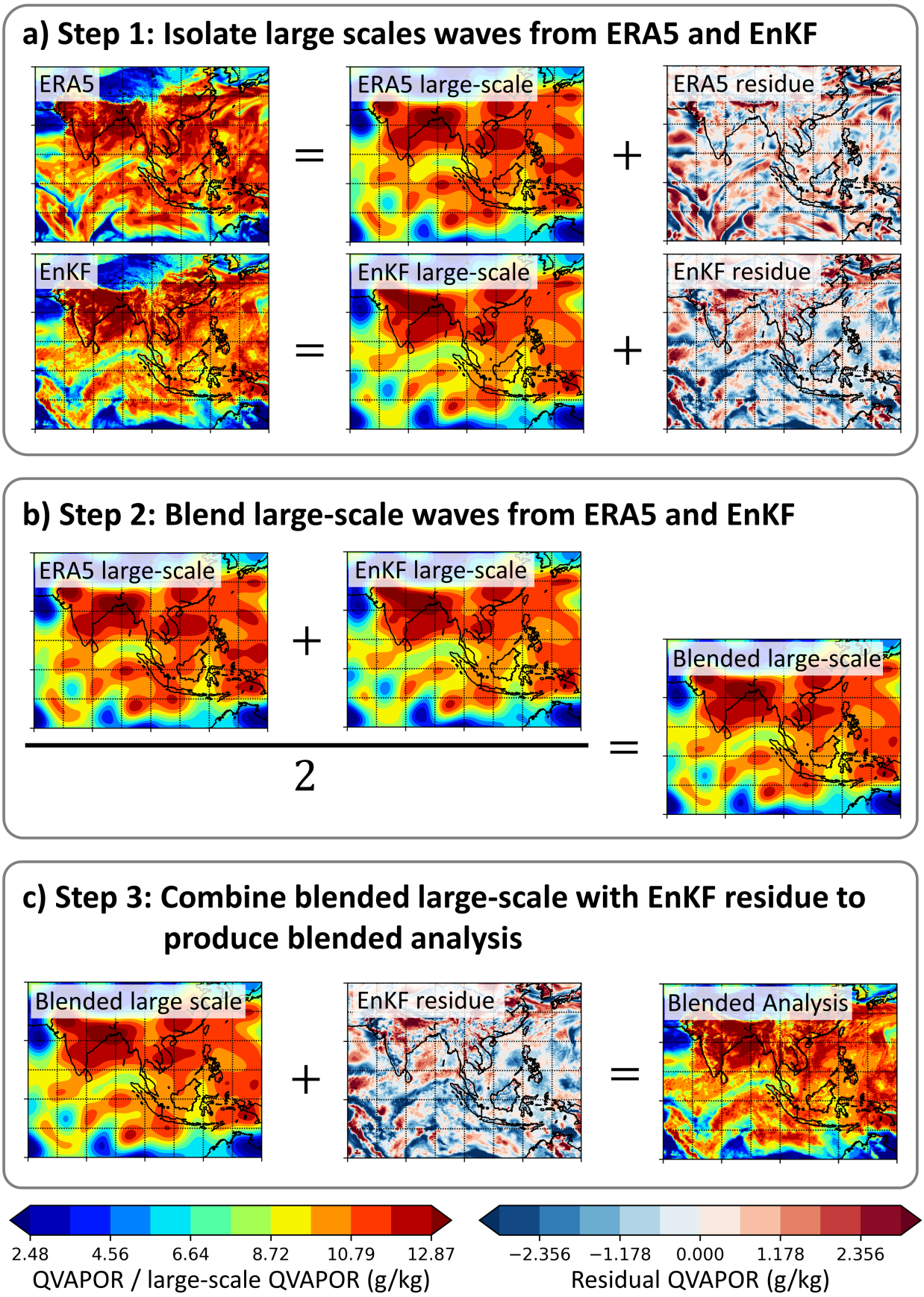


Figure 5.2: *See next page for caption*

Figure 5.2 *(previous page)*: A step-by-step illustration of the procedure used to blend the large-scale information from the EnKF-updated ensemble with the large-scale information from ERA5. In the first step (a), low-pass filtering is performed to separate information with horizontal wavelengths greater than 1000 km from information with horizontal wavelengths shorter than 1000 km. This filtering is performed on both the EnKF-updated ensemble mean and ERA5. In the second step (b), the two pieces of large-scale information is blended by taking the average of the two large-scale fields. Small-scale (horizontal wavelengths < 1000 km) information from the EnKF-updated ensemble mean is then introduced into the blended large-scale information to produce a blended analysis (c).

I also employed bias correction on the WV-BT observations. The bias correction method is the same as the simplest zeroth-order polynomial bias correction in Otkin et al. (2018), except for one difference. Instead of utilizing observation-forecast differences from a preceding DA cycle to generate the bias correction, I utilized observation-forecast differences valid at the current DA cycle. More specifically, WV-BT observations that were previously discarded by the thinning process (see earlier section) were utilized to estimate the domain-averaged WV-BT observation bias. This use of never-assimilated observations helps to maintain the independence between the observations to be assimilated and the forecasted observables.

5.2.5 Twice-a-day blending with large-scale information from ERA5

As mentioned earlier, the large-scale (> 1000 km) information of the TMeCSR ensemble mean field is blended with that of ERA5 every day at 08 UTC and 20 UTC. This is done to prevent the drift of the regional WRF model during the long-term integration.[4] The blending procedure is executed on the zonal winds (U), meridional winds (V), perturbation potential temperature (T), water vapor

[4]Specifically, without blending, the deterministic forecasts initialized from our reanalysis had larger error saturation values in the outgoing longwave radiation than those initialized from the ERA5.

mixing ratios (QVAPOR), perturbation pressure (P) and perturbation geopotential (PH) variable fields of the WRF model.

The blending process is as follows (see Figure 5.2 for an illustration). For each model layer and listed variable, a low-pass filter (1000-km wavelength threshold) is executed on the ensemble mean to isolate the ensemble mean's large-scale component. The same filter is used to isolate the ERA5's large-scale component. These two large-scale components are then averaged with equal weights (*i.e.*, 50-50 weights)[5] to construct the blended large-scale component. The ensemble mean's large-scale component is then replaced by the blended large-scale component.

A couple of notes on the blending process. I avoided blending the lowest 8 model levels to retain the high-resolution WRF-simulated atmospheric boundary layer. Furthmore, the blending was only performed at 08 UTC and 20 UTC. This is because ERA5 performs DA every 12-hours and assimilates 12 hours of observations (Hersbach et al., 2020) during each DA algorithm call. Blending at any time of day with ERA5 would thus introduce observation information from the future into the blended analysis. This will cause the observations and forecast ensemble to lose their statistical independence. To limit this effect, the blending is performed towards the end of the ERA5's observation windows (08 UTC and 20 UTC).

5.3 Performance of the TMeCSR

5.3.1 Cloud fields in the TMeCSR

I will first examine the performance of the TMeCSR in terms of the cloud fields. This is assessed by comparing the TMeCSR's and ERA5's outgoing long-

[5] Note that the 50-50 weights were used because no unassimilated observations were available for us to determine the relative accuracies of the ERA5 or the TMeCSR.

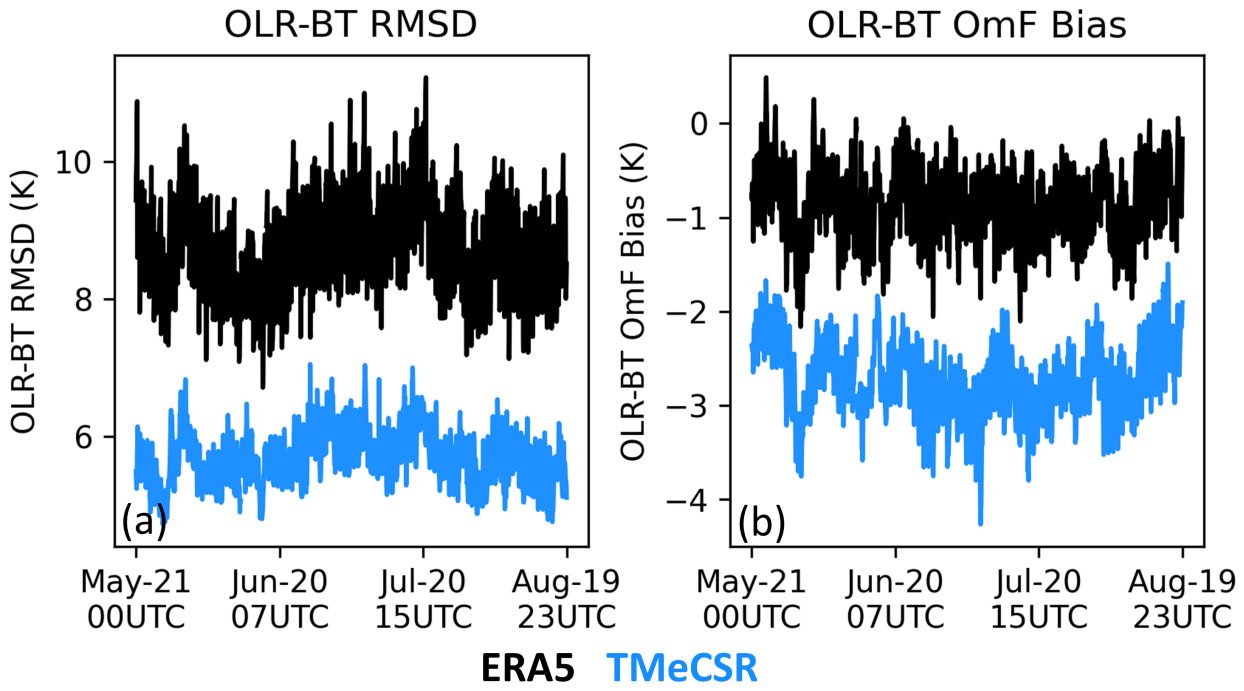


Figure 5.3: Cloud field performance statistics of ERA5 and TMeCSR as a function of date. These performance statistics are obtained by subtracting each dataset's forecasted OLR-BT data from the CERES OLR-BT data (i.e., observation minus forecast, or "OmF"). Both the root-mean-square of the differences (RMSD; a) and the average of the differences (ie, biases; b) are plotted.

wave radiation fields to the level 3 synoptic outgoing longwave radiation (OLR; SYN1deg) data (Doelling et al., 2013) from the Clouds and the Earth's Radiant Energy System (CERES) project (Wielicki et al., 1996). A qualitative cloud field assessment was also performed using SEVIRI channel 10 (10.8 μm central wavelength) and AHI channel 14 (11.2 μm central wavelength) infrared brightness temperature observation. For ease of interpretation, the OLR values were converted to OLR brightness temperatures (OLR-BT) using the Stefan-Boltzmann law:

$$BT = \left(\frac{OLR}{5.67 \times 10^{-8}} \right)^{1/4} \tag{5.1}$$

where BT stands for OLR-BT, and the denominator is the Stefan-Boltzmann constant. This conversion is permitted because most of the Earth's blackbody radiation is in the longwave regime.

Figure 5.3 shows the performance statistics of the TMeCSR and ERA5 OLR-BTs as functions of time. TMeCSR's forecast ensemble mean OLR-BT root-mean-square-differences (RMSDs) from the CERES OLR-BT are typically ∼30% lower than the ERA5's (Figure 5.3(a)). In contrast, TMeCSR's OLR-BT observation minus forecast (OmF) biases (roughly -3 K) are larger than ERA5's OLR-BT OmF biases (roughly -1 K). Since biases are a component of RMSDs, the RMSD improvements suggest that the degradations in the biases are overwhelmed by the improvements of the cloud fields at spatial scales smaller than the WRF domain. The CERES OLR-BT data assessment thus indicates that the TMeCSR cloud fields are improved over the ERA5 cloud fields.

Note that the strong negative OmF OLR-BT biases of TMeCSR indicate that TMeCSR is under producing clouds. This under-production is likely due to limitations in the WRF model. Since a 9-km horizontal grid spacing is employed, TMeCSR has difficulty resolving trade cumuli. Furthermore, the currently available WRF bulk microphysics schemes tend to underestimate the spatial extent of the stratocumuli surrounding MCSs (*e.g.*, Planche et al. (2019)). Nonetheless, the fact that TMeCSR has ∼30% lower OLR-BT RMSDs than ERA5 suggests that TMeCSR has better resolved clouds than ERA5.

To get a clearer picture of how the cloud fields of TMeCSR are better than those of ERA5, I compared the observed Window-BT field against those of TMeCSR and ERA5. The ERA5 Window-BT field was estimated from their OLR fields using the approach of Yang and Slingo (2001). The TMeCSR Window-BT field was estimated by first computing the 50 WRF members's Window-BT fields via the Yang and Slingo (2001) method, and then taking the ensemble mean. Figure 5.4 shows the comparison at a random time (07 UTC on 23 June).

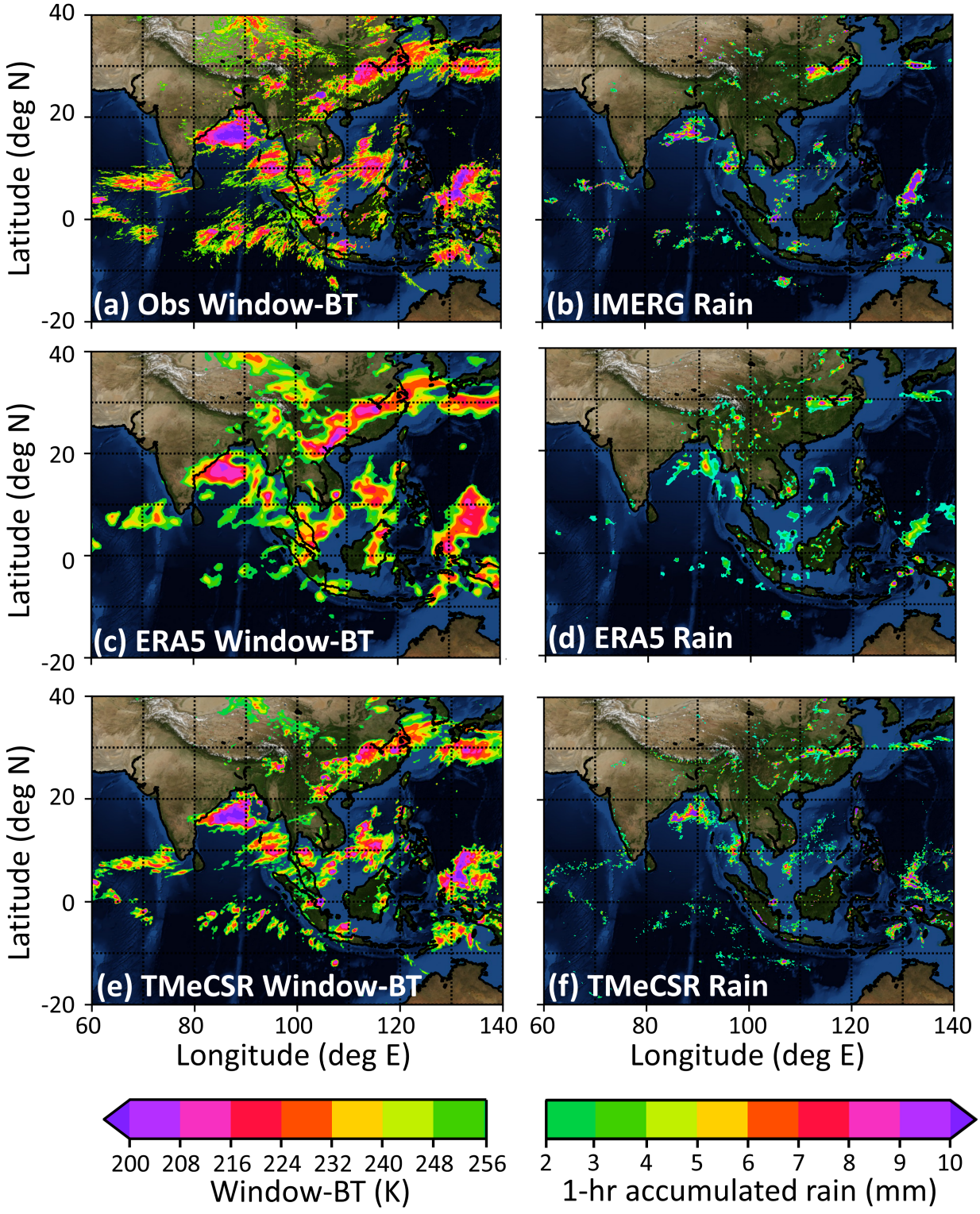


Figure 5.4: *See next page for caption*

Figure 5.4 *(previous page)*: Visual assessments of the cloud and rain information produced by TMeCSR and ERA5 against satellite-observed cloud and rain information on a typical date (07 UTC on 23rd June). The satellite-observed Window-BT data (a), ERA5-based Window-BT values (c), and TMeCSR-based Window-BT values (e) are compared for the visual cloud assessment. For the visual rain assessment, rain data from the IMERG (b) is compared against the ERA5 rain values (d) and the TMeCSR rain data (f).

A comparison of Figures 5.4(a), (c) and (e) reveals that TMeCSR has better cloud systems than ERA5. While the ERA5 MCS cloud systems were in the general location of the satellite observed MCSs, these ERA5 systems have overly broad stratiform regions (green regions in Figure 5.4(c)), insufficiently tall convective cores (purple and red regions in Figure 5.4(c)), and generally lack the observed fine structures (features with minor axis lengths smaller than 500-km). In contrast, TMeCSR's MCS cloud systems are better situated, avoid the overly broad stratiform regions, have more reasonable cloud top heights, and possess much of the observed fine-scale structures.

As a side note, the qualitative comparisons in the preceding paragraph do not consider features that are smaller than 100-km. Sub 100-km features are essentially invisible in Figure 5.4 since the plots were made over a large geographical area. In other words, ERA5 is not penalized for having a coarser data grid than TMeCSR. The noted qualitative differences are likely due to differences in the model dynamics and DA methods used in ERA5 and TMeCSR.

5.3.2 Rain fields in the TMeCSR

I also assessed whether the TMeCSR rainfall field are an improvement over ERA5's. This was done by comparing TMeCSR and ERA5 against the Global Precipitation Measurement (GPM) mission's Integrated Multi-Satellite Retrievals for GPM (IMERG) Version 6 Final Run (Huffman et al., 2020; Tan et al., 2019) rainfall data. Note that the TMeCSR rainfall fields are not updated by the PSU-

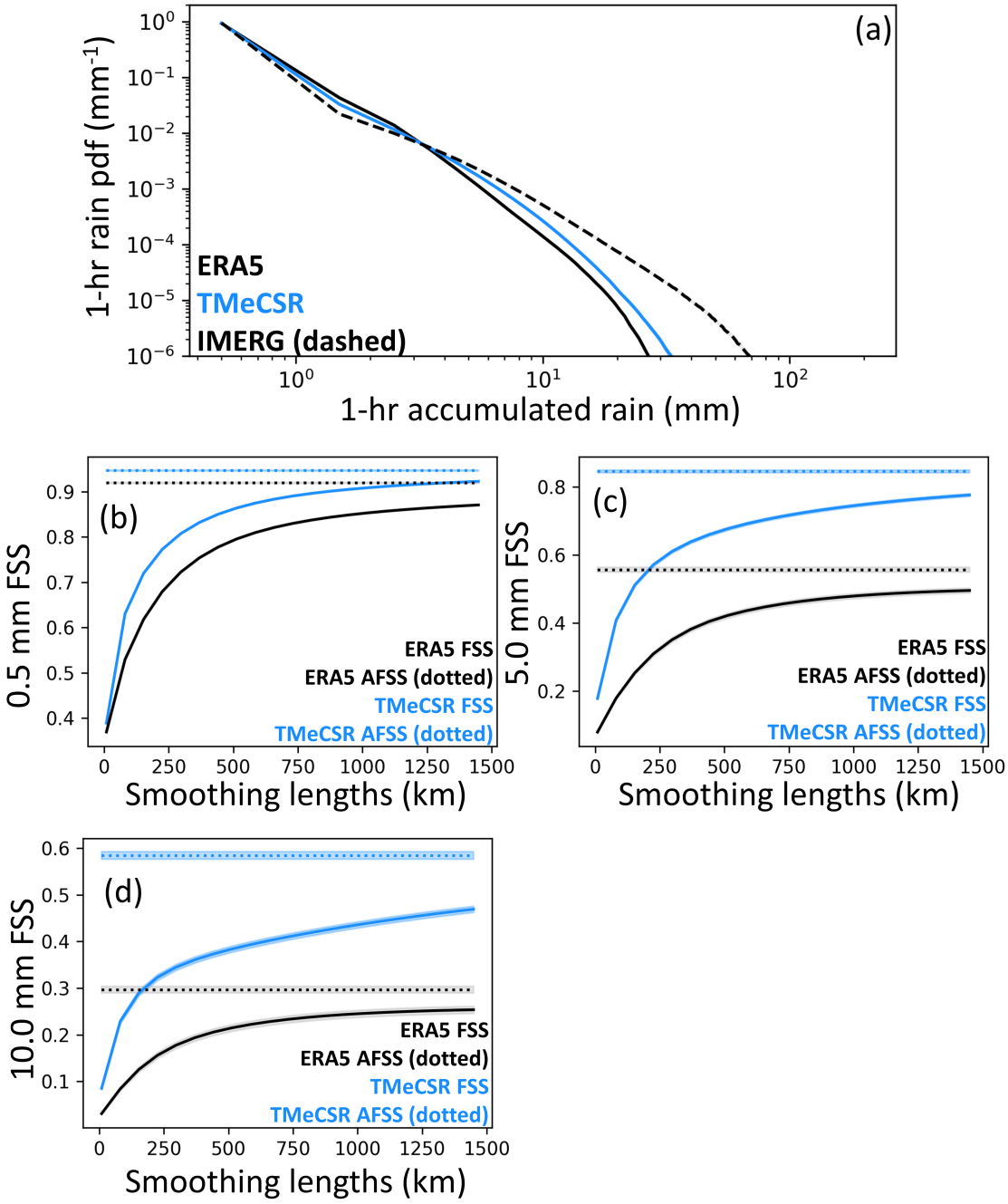


Figure 5.5: *See next page for caption*

Figure 5.5 *(previous page)*: Rain statistics of the IMERG, ERA5 and TMeCSR datasets. The pdfs of 1-hour accumulated rain are shown for all three datasets (a). The FSS's of the ERA5 and TMeCSR rain fields, with respect to the IMERG rain data, are shown as functions of smoothing lengths for 1-hour accumulated rain thresholds of 0.5-mm (b), 5.0-mm (c) and 10.0-mm (d). The AFSS values for both datasets, for each threshold, are also shown in the FSS plots. In all 4 panels, the half-width of the shadings indicates 2 times the standard error of the plotted quantity.

EnKF, meaning that these fields are solely due to model integration.

Figure 5.5(a) shows the time-averaged probability density functions (pdfs) of rainfall estimated from ERA5, TMeCSR, and the IMERG. ERA5 and TMeCSR were both interpolated to the IMERG's data grid prior to estimating the pdfs. Both ERA5 and TMeCSR overestimate the occurrence of 1-hour accumulated rain values under 3 mm, and underestimate the occurrence of 1-hour accumulated rain values above 3 mm. Nonetheless, the TMeCSR rain rate pdf is an improvement over the ERA5's because the former is closer to the observed rain rate pdf than the latter. This improvement likely arose from the use of an MCS-resolving forecast model in TMeCSR.

To assess how the TMeCSR 1-hour accumulated rain errors vary with spatial scale, I examined the scale-selective fractions skill scores (FSS) proposed by Roberts and Lean (2008) for 1-hour accumulated rain thresholds greater than 0.5 mm, 5 mm, and 10 mm (see Roberts and Lean (2008) for the FSS calculation algorithm). These three thresholds respectively correspond to light-to-heavy rain rates, moderate-to-heavy rain rates, and heavy rain rates.

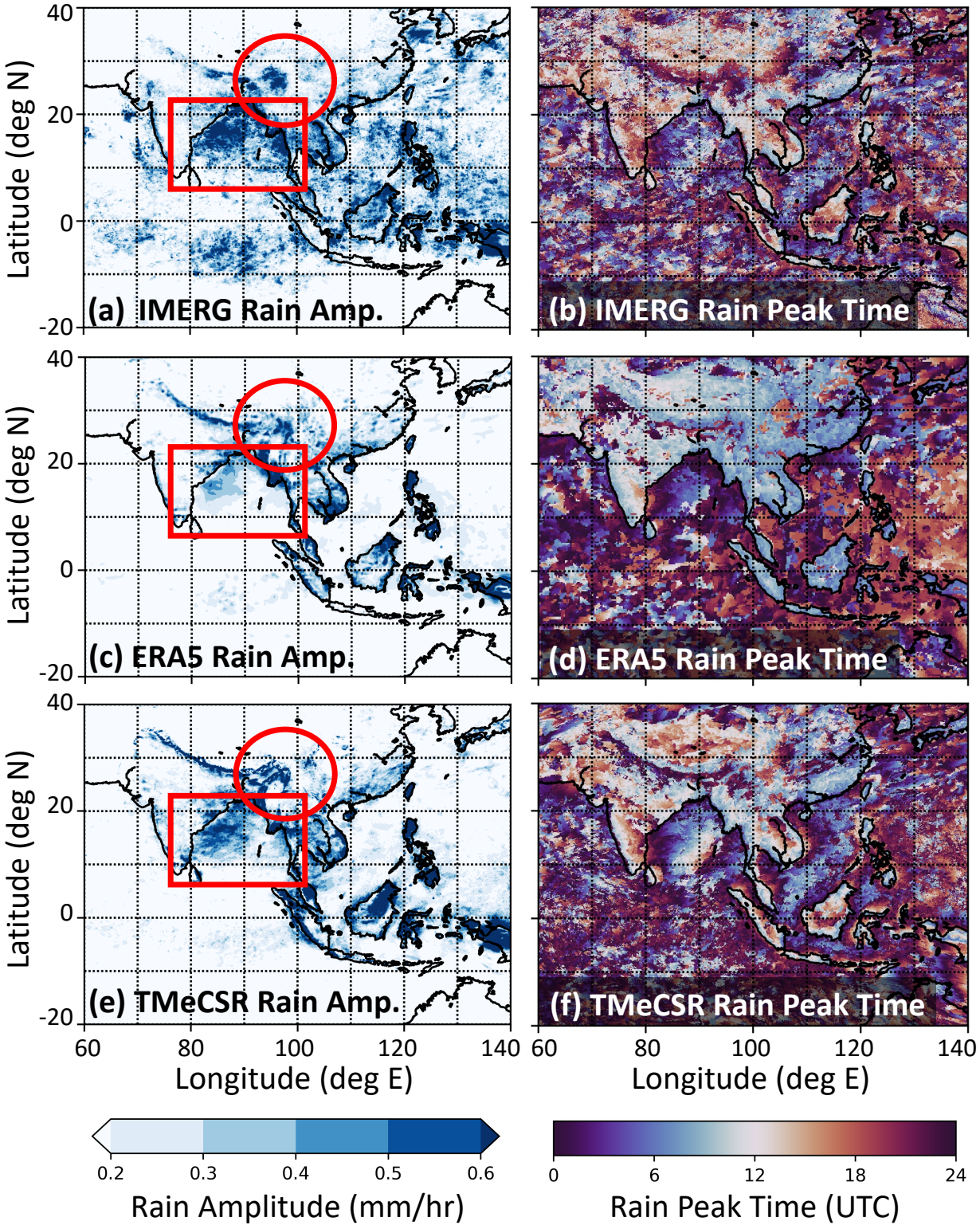


Figure 5.6: *See next page for caption*

Figure 5.6 *(previous page)*: Visual assessments of the diurnal rain amplitudes (a, c, and e) and diurnal rain peak hours (b, d and f) in TMeCSR (e and f) and ERA5 (c and d) against the diurnal rain amplitudes and diurnal rain peak hours observed by the IMERG (a and b). The red rectangles and red ovals in panels (a), (c), and (e) highlight areas where the TMeCSR's diurnal rain amplitudes better matched those observed than the ERA5.

The FSS data plotted in Figures 5.5(b), (c), and (d) indicate that the TMeCSR rain field is better than that of ERA5 in a variety of ways. First, TMeCSR's FSS values at infinite smoothing length (also known as the asymptotic FSS, or AFSS) are significantly closer to unity than those of ERA5 for all three thresholds. This implies that TMeCSR's rain occurrence biases are significantly lower than those in ERA5. These differences in occurrence biases are consistent with Figure 5.5(a): TMeCSR's pdf is closer to those of the IMERG than ERA5. Second, the TMeCSR's FSS values are always larger than the ERA5's for all examined thresholds and smoothing lengths. This across-the-board FSS difference implies that the TMeCSR rain field errors are smaller than those of ERA5 for all examined spatial scales. Finally, even without smoothing (0-km smoothing length), the TMeCSR's rain field errors are significantly smaller than that of ERA5.

To get a clearer picture of how the TMeCSR rainfall fields are improved over ERA5's, I plotted the rainfall fields on 07 UTC on 23 June 2017 in Figures 5.4(b), 5.4(d) and 5.4(f). Note that similar results are likely for other times (not shown). While ERA5 succeeded in capturing the general locations of the MCS-produced rainfall, the light rain areas (< 5 mm) on the edges of the MCSs are too large, and the moderate-to-heavy rain areas in the centers of the MCSs are too small and weak. These two issues are also consistent with the ERA5 pdf in Figure 5.5(a). In contrast, TMeCSR has more accurate rainfall locations as well as more realistic MCS rain features (Figure 5.4(f)).

Note that the qualitative comparisons in the previous paragraph did not consider rain features that are smaller than 100-km. These small features are ignored so that the ERA5 data is not unfairly penalized for having a coarser data grid than TMeCSR.

5.3.3 Rainfall diurnal cycles

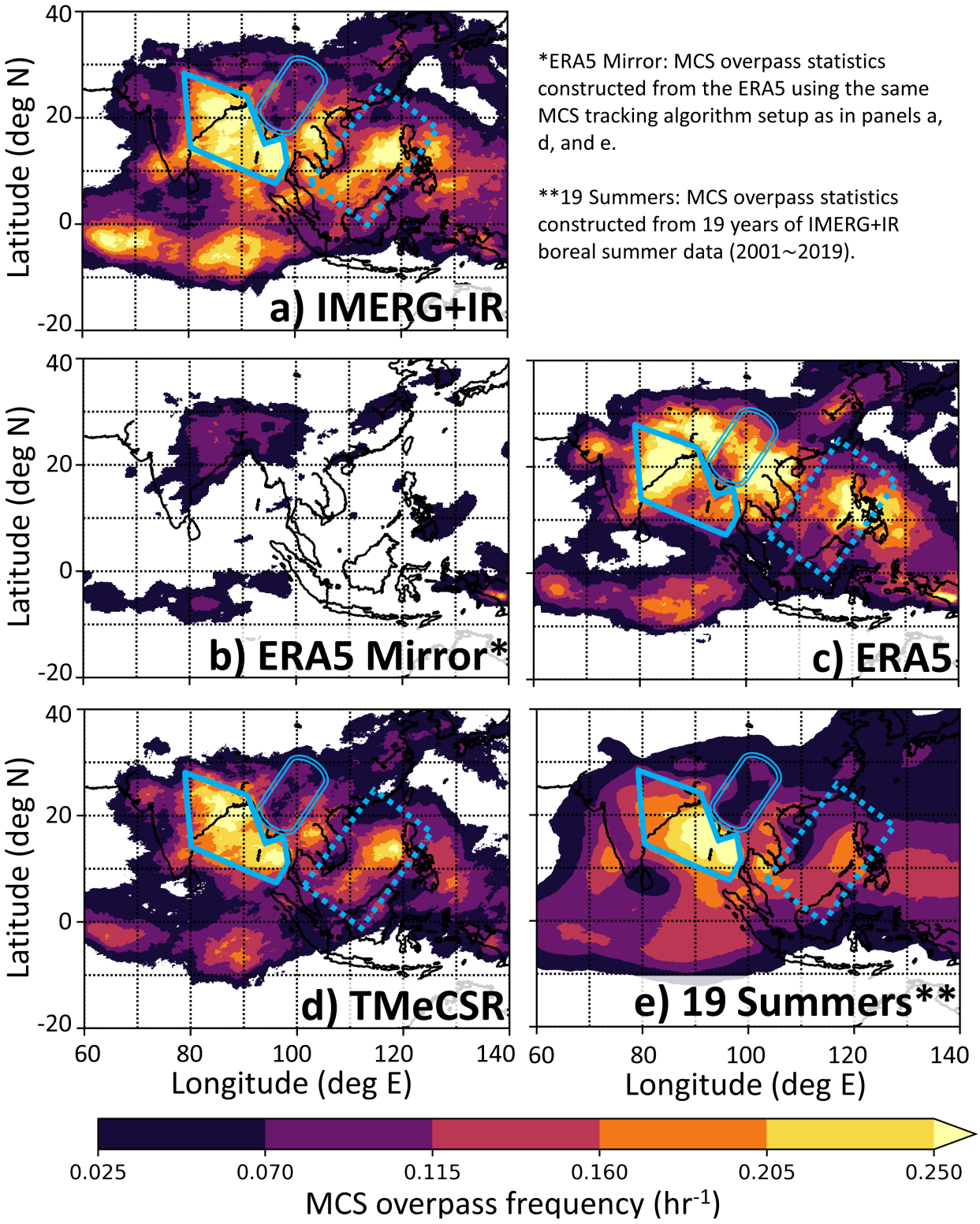
Since the rainfall diurnal cycles are one of the most fundamental modes of precipitation variability, the rainfall diurnal cycles resolved by TMeCSR were also assessed. Furthermore, the characterization of the rainfall diurnal cycle can help us understand not only the initiation and evolution mechanisms of tropical MCSs, but also the mechanisms that drive the regional and global climate (Chen et al. 2015; Aiguo Dai 2001). Accurately resolving rainfall diurnal cycles in reanalysis products would thus be a boon for research into these mechanisms.

I examined the diurnal rain amplitudes and diurnal rain peak times in the IMERG, ERA5 and TMeCSR (Figure 5.6). These two quantities were produced by first generating hourly rain composites over the entire domain using data during the TMeCSR period (June, July, and August 2017). The diurnal rain amplitude at any grid point or observation pixel is defined as half of the difference between the composites' diurnal rain rate minimum and diurnal rain rate maximum. The diurnal rain peak time at any grid point or observation pixel is defined as the hour at which the rain rate is maximized. These two quantities were examined because they characterize the diurnal rain cycle.

The diurnal rain amplitudes of both TMeCSR and ERA5 were examined first (Figure 5.6(a), (c) and (e)). On average, the amplitudes of TMeCSR (0.1700 ± 0.00030 mm/hr) and ERA5 (0.1403 ± 0.00050 mm/hr) are significantly smaller than that of the IMERG (0.2069 ± 0.00024 mm/hr). The TMeCSR average amplitude is nonetheless significantly closer to the IMERG's than ERA5's. Beside

the domain-averaged amplitudes, the TMeCSR amplitudes better resemble the IMERG's than ERA5's over the Bay of Bengal (red rectangles in Figure 5.6(a), (c) and (e)). Furthermore, compared to the ERA5, the TMeCSR better captured the observed circular region of low diurnal rainfall amplitude around the south-eastern edge of the Tibetan plateau (red ovals in Figure 5.6(a), (c) and (e)). It is otherwise difficult to distinguish differences in the performances of the TMeCSR and ERA5 diurnal rainfall amplitude. Overall, TMeCSR has better rain amplitudes than ERA5.

While it is difficult to visually distinguish the performances of the rain peak times of TMeCSR and ERA5 over the ocean, several important differences stand out over land (Figure 5.6(b), (d) and (f)). First, the TMeCSR rain peak times over land better resemble the observed dramatic variations than ERA5. These dramatic variations are likely associated with complex topography. Second, the TMeCSR peak rain times are noticeably closer to the IMERG than ERA5 over the Middle East, the southern half of the Indian peninsula, the Tibetan plateau, the Yellow River Basin, and the entirety of Southeast Asia (including the Borneo-Sumatra region). In other words, the TMeCSR rain peak time performances are generally better than ERA5 over land. This difference in peak time performance is likely due to ERA5's relatively coarser horizontal grid (Sato et al. 2009; Love et al. 2011) and its use of a cumulus parameterization scheme (Folkins et al. 2014; ECMWF 2016a). More importantly, from these rain peak time comparisons and the earlier rain amplitude comparisons, I can conclude that TMeCSR generally better captured the diurnal rain peak times than ERA5.



*ERA5 Mirror: MCS overpass statistics constructed from the ERA5 using the same MCS tracking algorithm setup as in panels a, d, and e.

**19 Summers: MCS overpass statistics constructed from 19 years of IMERG+IR boreal summer data (2001~2019).

Figure 5.7: *See next page for caption*

Figure 5.7 *(previous page)*: MCS frequencies from 1st July 2017 to 20th August 2017 for the IMERG+IR (a), ERA5 (b & c), TMeCSR (d) and TMeCSR centermost member (e). These frequencies are estimated on locations arranged on a 0.1° by 0.1° grid. At each location, the frequency is estimated by counting how often FLEXTRKR-identified MCS-associated clouds pass over the location. Note also that the ERA5 frequencies are calculated in two different ways: using the MCS identification criteria of Feng et al. (2021b) (b) and using the coarse resolution MCS identification criteria (c; see text).

5.3.4 Tropical MCS frequency, duration, size and rainfall

Since the purpose of TMeCSR is to provide many MCS cases for the tropical MCS and convective parameterization communities, it is important to determine if the MCSs in TMeCSR are more realistic than those of ERA5. To do so, I used the Flexible Object Tracker [FLEXTRKR; Feng et al. (2018, 2019, 2021b,a)] to identify tropical MCSs in the TMeCSR, ERA5 and observational data. MCSs in TMeCSR and ERA5 were identified and tracked by applying FLEXTRKR to their OLR-estimated Window-BT (Yang and Slingo, 2001) and rain rate fields. To identify and track observed MCSs, this algorithm was also applied to a combination of the IMERG rain rate product and a global geostationary satellite Window-BT data product (Janowiak et al., 2001).[6]. Because the NASA global brightness temperature product was not available for 20 days in June 2017, our MCS examination was performed from 1st July 2017 to 20th August 2017.

Before proceeding with the MCS validation, note that the FLEXTRKR MCS identification criteria used by Feng et al. (2021b) are grossly inappropriate for ERA5. This is because the resulting spatial distribution of MCS frequencies (Figure 5.7(b)) was highly unrealistic. The unrealistic MCS frequencies are likely because the coarse horizontal resolution of ERA5 prevents it from fully resolving the intense rain rates in the MCS strong convection zones (Bélair and

[6]The latter product is produced by the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) and archived at the National Aeronautics and Space Administration's (NASA's) Goddard Earth Sciences Data and Information Services Center (GES DISC)

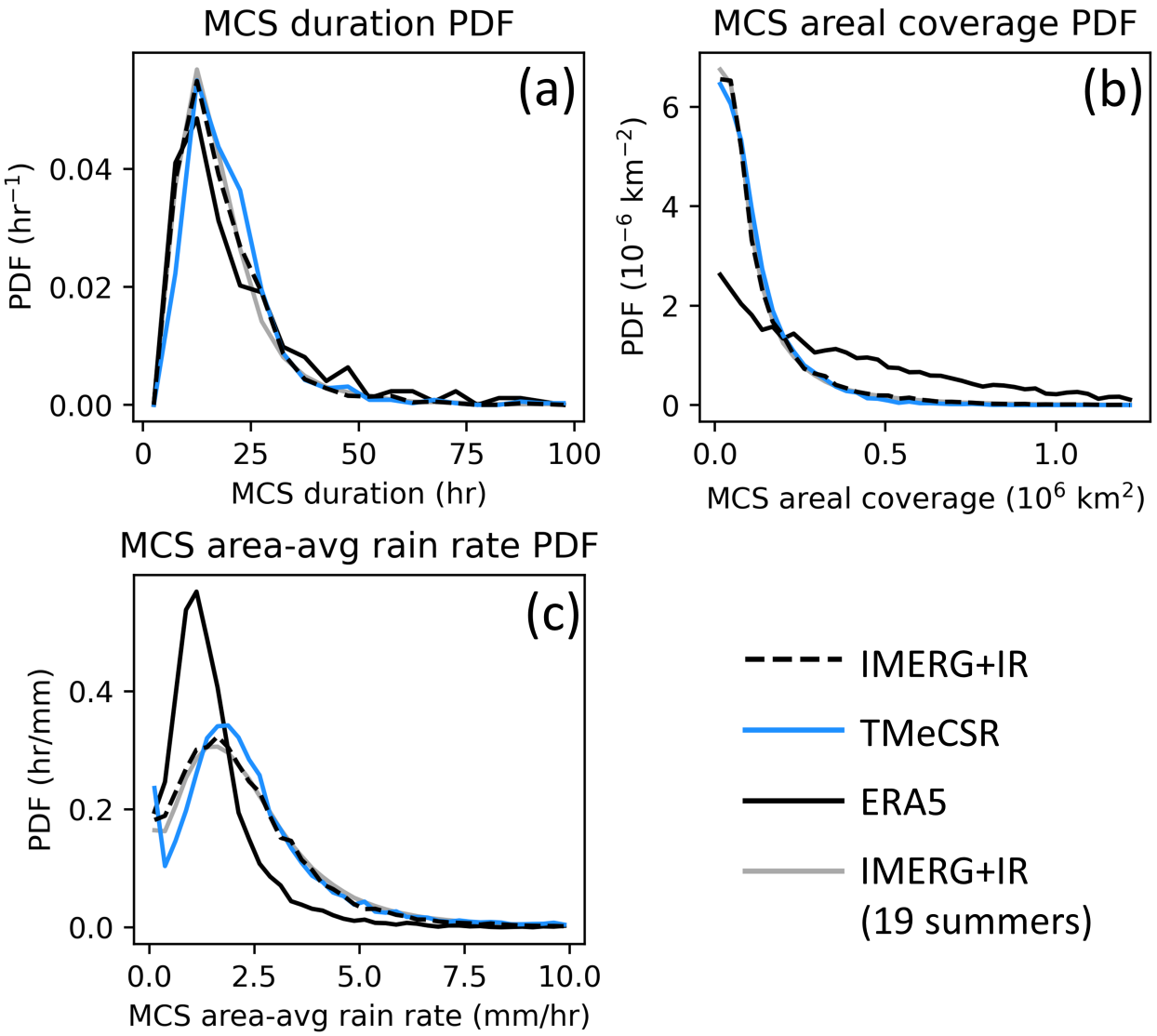


Figure 5.8: Probability density functions (pdfs) of (a) the MCS durations, (b) MCS sizes and (c) MCS area-averaged rain rates for the various datasets. Note that the gray solid curves are pdfs generated from MCSs identified in 19 summers (2001–2019) of IMERG+IR data.

Mailhot, 2001). As such, an alternative set of MCS identification criteria should be used for ERA5 that better matches the resolution of the dataset.

A more appropriate set of MCS identification criteria was suggested by Feng et al. (2021a) for evaluating climate simulations with 25-km grid spacings (henceforth, the coarse resolution criteria). The coarse resolution criteria differ from the Feng et al. (2021b) criteria in terms of the specified precipitation feature rain area threshold, area-averaged rain rate threshold, and rain rate skewness threshold [the exact values are available in Feng et al. (2021b) and Feng et al. (2021a)]. Since the resolution of ERA5 is comparable to the resolution of the climate simulations evaluated by Feng et al. (2021a), I adopted the coarse resolution criteria for tracking MCSs in ERA5. The spatial distribution of MCS frequencies resulting from applying the coarse resolution criteria on ERA5 [Figure 5.7(c)] is much more realistic than the earlier distribution [Figure 5.7(b)]. A similar approach was also used in Chen et al. (2021b) to generate a reasonably realistic spatial distribution of MCS frequency from ERA5. Given these results, in the subsequent discussions, I use the Feng et al. (2021b) criteria to identify MCSs in the IMERG+IR and the TMeCSR, and the coarse resolution criteria to identify MCSs in ERA5.

I will begin with evaluating the TMeCSR MCS frequency spatial patterns (Figure 5.7(d)), with a focus on notable features in the climatological spatial pattern of MCS frequencies (Figure 5.7(e)). This climatological spatial pattern is produced by applying FLEXTRKR on 19 years of June-July-August IMERG+IR data (2001 2019). Like the climatological pattern (Figure 5.7(e))), the IMERG+IR (Figure 5.7(a)) and TMeCSR (Figure 5.7(d)) have strong MCS frequencies (>0.205 per hour) over the Bay of Bengal (solid blue polygons in Figure 5.7). In contrast, ERA5 has low MCS frequencies in the middle of the Bay of Bengal (the dark spot in the blue polygon of Figure 5.7(c)). Near the Barail Range (blue oval with double lines), TMeCSR captured the dramatic low MCS frequency feature seen in both the IMERG+IR and the climatological data. This feature is nearly nonexistent in ERA5. Aside from that, near the Philippines, both the climatology and IMERG+IR indicate a hotspot of MCS frequency that runs from the Philippines to the South China Sea (blue dashed rect-

angle). TMeCSR captured a similar hotspot whereas the corresponding hotspot in ERA5 is more confined to the Philippines. These suggest that TMeCSR better captured the spatial pattern of MCS frequencies than ERA5.

Note that the TMeCSR MCS frequencies appear to be low-biased with respect to those of IMERG+IR over the South China Sea (blue dashed rectangles in Figure 5.7), the Bay of Bengal (solid blue polygons in Figure 5.7), and the equatorial Indian Ocean. There are several plausible causes for these biases: 1) FLEXTRKR is slightly overestimating the observed MCS frequencies, 2) TMeCSR's WRF setup tends to underproduce MCSs, and/or 3) the TMeCSR's WRF setup tends to produce overly small MCSs. Future work is necessary to uncover the source of these biases and mitigate them.

The characteristics of the MCSs resolved by the IMERG+IR, TMeCSR and ERA5 (Figure 5.8) were also examined. The lifetimes (Figure 5.8(a)), sizes (Figure 5.8(b)) and area-averaged rain rates (Figure 5.8(c)) of the observed MCSs have pdfs that are closer to those of TMeCSR than ERA5. These differences are the starkest for MCS lifetimes beyond 30 hours, MCS sizes beyond 0.25×10^6 km^2, and MCS area-averaged rain rates beyond 2.5 mm/hr. For these ranges, the TMeCSR pdfs are nearly indiscernible from those of IMERG+IR whereas the ERA5 pdfs deviate noticeably from the IMERG+IR pdfs. These results thus indicate that the MCS characteristics of TMeCSR are more realistic than those of ERA5. Since TMeCSR also produced more realistic spatial patterns of MCS frequencies, TMeCSR thus has more realistic MCSs than ERA5.

It is interesting to note that the properties of the observed MCSs in the TMeCSR period are quite comparable to the region's climatological summertime MCSs. As seen in Figure 5.8, the climatological pdfs of MCS properties (estimated from 19 years of June-July-August IMERG+IR data) are similar to those of IMERG+IR and TMeCSR. In other words, although TMeCSR covers a limited period, the MCSs that occurred during that period are quite represen-

tative of the region's summertime MCSs.

5.4 Conclusions

In this work, we have constructed a 9-km grid spacing ensemble-based tropical MCS reanalysis dataset through combining a cutting-edge ensemble data assimilation system (PSU-EnKF) with all-sky GeoIR observations, GTS observations, ERA5-based boundary conditions, and large-scale information from ERA5. The resulting TMeCSR dataset covers the majority of the Indian Ocean, tropical/subtropical continental Asia and the Maritime Continent, for the three summer months of 2017, and contains more than 1200 MCS events. By virtue of its spatial resolution, TMeCSR can explicitly resolve tropical MCSs. Aside from that, TMeCSR contains hourly analyses of the basic WRF output fields (e.g., 3D winds, potential temperature, hydrometeor mass mixing ratios), as well as instantaneous information about diabatic processes (e.g., longwave and shortwave radiation fluxes).

Validations against CERES OLR and IMERG rain data indicate that, compared to ERA5, TMeCSR better captured the tropical cloud and rainfall fields. Furthermore, TMeCSR outperformed ERA5 in terms of capturing the diurnal rainfall cycle throughout most of the domain. Using an MCS tracking algorithm, we found that TMeCSR produced more realistic MCSs than ERA5. Interestingly, the MCSs that occurred during the TMeCSR period have properties that are similar to the region's typical summertime MCSs. This suggests that TMeCSR is a useful dataset for MCS studies despite its limited time coverage. These encouraging results motivate the possibility of future studies using the TMeCSR dataset to investigate tropical MCSs or to inform convective parameterization.

Chapter 6

Forecast uncertainties about the presence/absence of clouds lead to mixture statistics

6.1 Overview

In the previous chapters, I have demonstrated that assimilating GeoIR observations via the EnSRF can improve the analyses and forecasts of tropical MCSs (Chapter 4), and can be used to produce a tropical MCS reanalysis (TMeCSR; Chapter 5; Chan and Chen (2021)). These results suggest that the assumptions in the EnKF (Chapter 3.2) are functional, at the very least. Nonetheless, as we will see in this chapter, the Gaussian forecast assumption is imperfect. Replacing this assumption with a more appropriate forecast error assumption could potentially improve the impacts of GeoIR DA.

The goal of this chapter is to demonstrate that the forecast statistics are mixed when the forecast ensemble is uncertain about the absence/presence of clouds (henceforth, mixed ensemble). This is because clear atmospheric columns and cloudy atmospheric columns follow different statistics (Harnisch

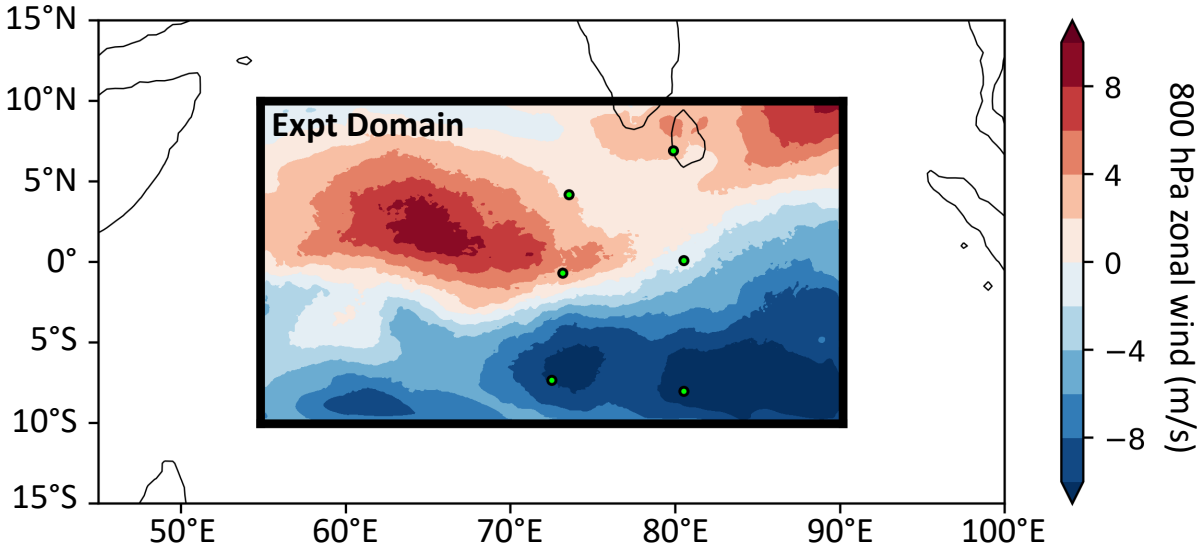


Figure 6.1: Study's domain over the Indian Ocean. The thin black lines indicate coastlines and the filled contours indicate the 800 hPa zonal wind field inside the study's domain. The locations of the DYNAMO sounding array are indicated in green-filled black circles.

et al., 2016; Minamide and Zhang, 2017, 2019). This point will be demonstrated using WRF ensembles from a case of tropical convection over the Indian Ocean (Wang et al., 2015; Chan et al., 2020b).[1]

6.2 Setup of WRF ensemble

An ensemble of WRF simulations for a case of tropical convection over the Indian Ocean will also be examined (1200 UTC on 15 October 2011). This case occurred during the onset of a Madden-Julian Oscillation (MJO) event (Madden and Julian, 1971, 1972), and was chosen for a variety of reasons. First, the general features of its tropical convection can be replicated in regional WRF

[1]Note that I have also demonstrated this point using the WRF ensemble generated in Chapter 4.2.1. See Chan et al. (2020a) for the figures and discussion.

models (Wang et al., 2015). Furthermore, this case was observed by the Dynamics of the MJO (DYNAMO) field campaign (Zhang et al., 2013; Zhang and Yoneyama, 2017) and has been explored in multiple simulation studies (Zhang et al., 2017; Ying and Zhang, 2017; Fu et al., 2017; Chen et al., 2018c; Chen and Zhang, 2019). This case has also been used in earlier studies to examine the potential impacts of assimilating IR-BTs on the analyses and forecasts of tropical convection (Ying and Zhang, 2018; Chan et al., 2020b). In other words, this October 2011 case over the equatorial Indian Ocean can be reasonably replicated in simulations and is reasonably established. This case is thus suitable for our investigation.

The setup of the WRF model for this case is identical to that of Chapter 4.2.1, except for the domain location, domain size and the chosen microphysics scheme. The domain is located over the tropical Indian Ocean (see Figure 6.1), has 432×243 horizontal grid points with 9-km horizontal grid spacing and 45 model levels. The bottommost 9 levels are within the lowest 1-km of the atmosphere and the pressure level at the top of the domain is set to 20 hPa. A 20-second time step is used to integrate the WRF model in time. The microphysics scheme used is the WRF double-moment 6-class scheme (WDM6; Lim and Hong (2010)).

The WRF ensemble is constructed by combining two datasets from the European Center for Medium-Range Forecasts (ECMWF): the ECMWF Reanalysis Version 5 (ERA5; Hersbach et al. (2020)) and the ECMWF's 50-member perturbed forecasts (Swinbank et al., 2016). The ERA5 dataset was download for every hour between 0000 UTC on 15 October to 1800 UTC on 18 October from the ECMWF's Climate Data Store (CDS). The ECMWF's perturbed forecasts were produced as part of The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; Swinbank et al. (2016)) and was downloaded for 0000 UTC on 15 October from the ECMWF's Meteorological Archival and Retrieval System (MARS).

The ERA5 and ECMWF's 50-member perturbed forecasts (TIGGE ensemble) were processed using the WRF Preprocessing System and WRF's real data processor (`real.exe`) to produce a set of 51 WRF initial conditions files. Note that the ERA5 was used to fill in the data missing from the TIGGE ensemble above 200 hPa. The 50 WRF initial conditions from the TIGGE ensemble were then re-centered on the ERA5 WRF initial condition file. The end result is a 51-member ensemble of WRF initial conditions, where member 51 is based entirely on the ERA5 (*i.e.*, the 51-st ensemble perturbation is zero). Afterwards, the 10th member is set aside[2], resulting in a 50 member esnemble of WRF initial conditions.

The 50 members are integrated forward for 12 hours to construct flow-dependent statistics. Note that the the lower and lateral boundary conditions used in this study are based entirely on the hourly ERA5 dataset (*i.e.*, the boundary conditions are unperturbed). While perturbed boundary conditions can increase the ensemble spread, the ensemble spread is usually reasonable even with unperturbed boundary conditions (not shown).

6.3 Differences between clear statistics and cloudy statistics

Before proceeding, it is necessary to briefly discuss how I identify an ensemble member's model column as clear (henceforth, clear member column) or cloudy (henceforth, cloudy member column). Since our WRF simulations used a 9-km horizontal grid spacing and trade cumuli typically have widths of ∼1-km, our simulations could not properly resolve trade cumuli. We thus

[2]This member will be used to construct the nature run for the WRF OSSE experiment later.

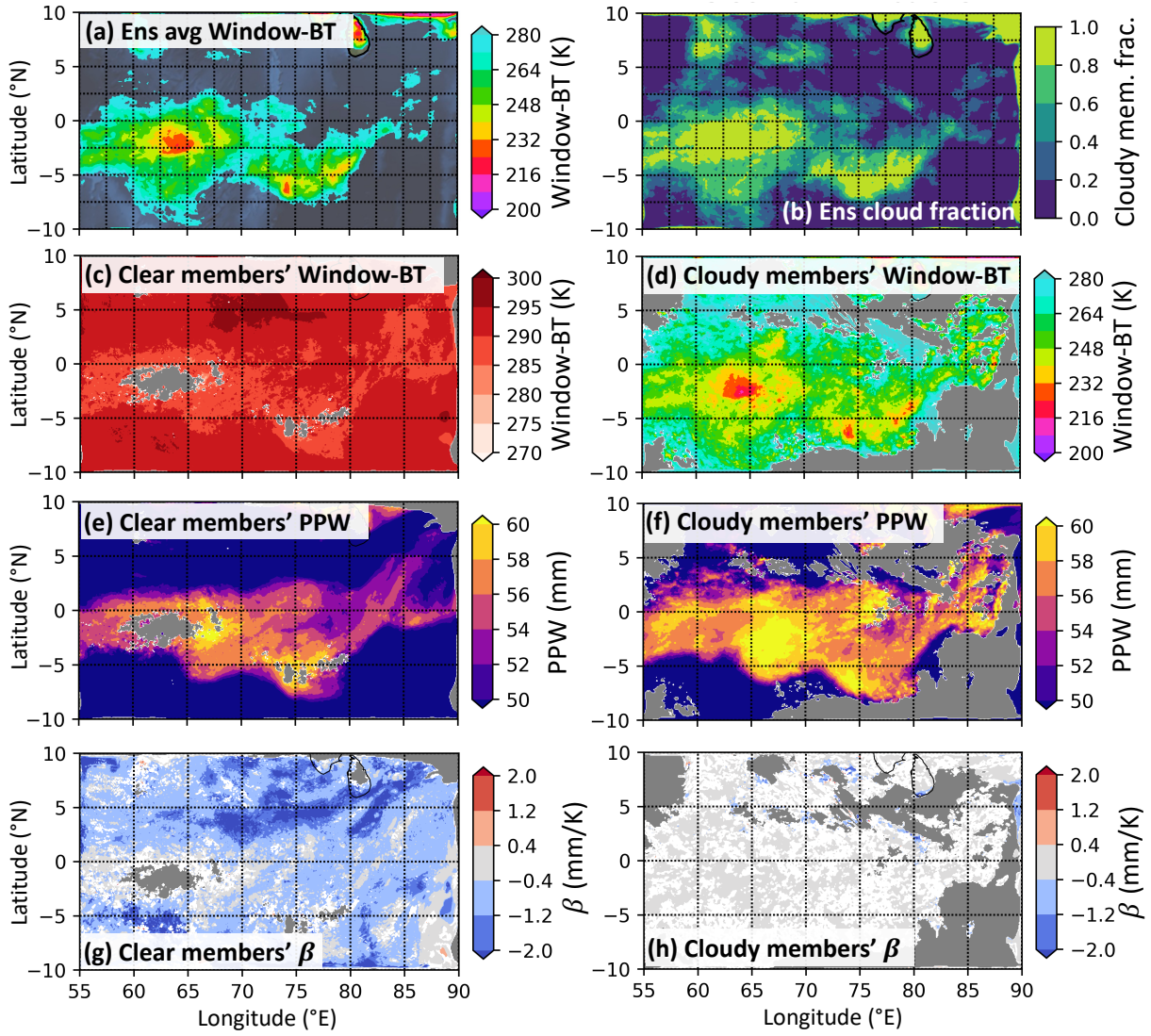


Figure 6.2: *See next page for caption*

considered columns with trade cumuli and entirely cloud-free columns as clear member columns, and the remaining members as cloudy member columns. Since trade cumuli do not typically grow above the melting layer (Johnson et al., 1999), cloud-less columns and trade cumuli do not possess frozen wa-

Figure 6.2 *(previous page)*: Latitude-longitude plots of various ensemble statistics at 1200 UTC on 15 October 2011 to illustrate the differences between clear and cloudy sky members at every model column. These quantities are generated using the 50-member ensemble described in section 6.2. The y-axes indicate latitude (degrees North), and the x-axes indicate longitude (degrees East). The plotted quantities are: the prior ensemble mean Window-BT (a), the fraction of cloudy member columns in the prior ensemble at every grid column (b), the mean Window-BTs of clear member columns (c), the mean Window-BT of cloudy member columns (d), the mean pseudo precipitable water (PPW) for clear member columns (e), the mean PPW for cloudy member columns (f), the linear regression coefficient between Window-BT and PPW (β) for clear member columns (g), and the β values for cloudy member columns (h). The gray shadings in panels c, e & g indicate locations where there are either less than 5 clear member columns, the clear member columns' Window-BT sample variance is zero, or the clear member columns' PPW sample variance is zero. The gray shadings in panels d, f & h indicate locations where there are either less than 5 cloudy member columns, the cloudy member columns' Window-BT sample variance is zero, or the cloudy member columns' PPW sample variance is zero. The white shadings in panels g indicate areas where the clear member columns' sample correlation between PPW and Window-BT is statistically insignificant, and likewise for the white shadings in panel h.

ter. It thus seems appropriate to use column-integrated ice mass content (ξ) to distinguish between clear and cloudy member columns. At model column (i, j), we compute ξ via:

$$\xi \equiv \int_0^{z_{top}} \rho(q_i + q_s + q_g) dz. \tag{6.1}$$

Here, z_{top} is the model top altitude and ρ represents air density. q_i, q_s and q_g are the mass mixing ratios of ice, snow and graupel, respectively. In this study, we will consider columns with $\xi \geq 1$ g/m^2 as cloudy, and columns with $\xi < 1$ g/m^2 as clear. The cloudy and clear Window-BT statistics do not vary noticeably for thresholds between 0.8-1.2 g/m^2. Future studies can refine the threshold value or seek better ways to separate clear and cloudy column members.

To illustrate the notion that clear and cloudy member columns can simultaneously exist in an ensemble, we plotted maps of the ensemble averaged Window-BT (Figure 6.2(a)) and the fraction of cloudy member columns in the ensemble (Figure 6.2(b)). These ensemble quantities are constructed from the spun-up 50-member WRF ensemble described in section 6.2. Though the ensemble captured the general appearance of the organized convective features seen in the nature run (Figures 6.1 and 6.2(a)), the ensemble was uncertain about the presence/absence of clouds over much of the domain (Figure 6.2(b)). This uncertainty is particularly noticeable over regions where the ensemble averaged Window-BT was between 248 K and 280 K.

Several differences between clear and cloudy member columns can be seen from Figure 6.2. First, the average Window-BT values of clear member columns are typically warmer than 280 K, whereas the average Window-BT values of cloudy member columns are cooler than 280 K (Figure 6.2(c & d)). This difference is well known. As such, the Window-BT ensemble statistics of an ensemble of clear and cloudy member columns (henceforth, mixed ensemble) will exhibit mixed statistics.

The clear and cloudy member columns also differ noticeably in terms of their humidity fields and the Kalman gain linking Window-BT innovations to humidity increments. For the ease of visualization, we examined through a column-integrated measure of humidity that is a linear function of the WRF model state: the pseudo precipitable water (PPW). The PPW is defined as

$$\text{PPW} \equiv \frac{g}{P_{\text{sfc}} - P_{\text{top}}} \int_0^1 q_v d\eta \tag{6.2}$$

where q_v refers to water vapor mass mixing ratio (QVAPOR), P_{sfc} and P_{top} refer to model surface pressure and model top pressure, and η refers to the WRF model's vertical coordinate. The PPW can be derived from the definition of

precipitable water by applying the hydrostatic approximation, the definition of WRF η levels, and by assuming that P_{sfc} and P_{top} are constants ($P_{\text{sfc}} \equiv 1000$ hPa, $P_{\text{top}} \equiv 20$ hPa).

We opted to use the linear PPW over precipitable water (PW) because PW is a nonlinear function of the model state. The Kalman gain linking PW to Window-BT within the same model column thus is not mathematically equivalent to taking a column-integral of the Kalman gain linking QVAPOR to Window-BT. In contrast, the Kalman gain linking PPW to Window-BT within the same model column is mathematically equivalent to taking a column-integral of the Kalman gain linking QVAPOR to Window-BT. Looking at PPW over PW thus allows us to get a more accurate sense of what the EnKF would do to QVAPOR within a model column.

Figure 6.2(c & d) indicate that the PPW of cloudy member columns is higher than that of clear member columns. This is because clouds require nearly saturated humidity to materialize. As such, when the ensemble is mixed, mixture statistics in the humidity fields are likely.

We also examined the component of the Kalman gains responsible for propagating Window-BT innovations to QVAPOR: the least squares linear regression coefficient linking Window-BT to QVAPOR (Anderson, 2003). For the ease of visualization, we looked at the coefficient linking Window-BT to PPW within the same column. This coefficient (β) is defined as

$$\beta \equiv \frac{\text{Cov}(\text{PPW}, \text{BT})}{\text{Var}(\text{BT})}. \tag{6.3}$$

$\text{Cov}(\text{PPW}, \text{BT})$ denotes the prior ensemble covariance between PPW and Window-BT within said model column, and $\text{Var}(\text{BT})$ denotes the prior ensemble variance of Window-BT within the same column. In the limit where Var(BT) is much smaller than the observation error, the Kalman gain turns into β.

As can be seen from Figure 6.2(e & f), the clear member columns' statistically significant β values are generally an order of magnitude larger than those of the cloudy member columns. This difference suggests that the statistical relationship between Window-BT and humidity can vary dramatically depending on the absence/presence of clouds.

6.4 Implication: EnSRF is sub-optimal for mixed ensembles

The forecast statistics of clear member columns are different from cloudy member columns. As such, whenever mixed ensembles occur, it seems appropriate to assume that there are two sets of linear relationships present that link observable quantities to model variables: one set for clear member columns and another set for cloudy member columns. Since the Gaussian forecast assumption (Chapter 3.2) only permits a single set linear relationships (see Chapter 3.8), the EnSRF cannot handle two sets of linear relationships. The EnSRF is thus sub-optimal to assimilate observations into mixed ensembles.

In the next chapter, I will introduce an extension of the EnSRF algorithm that can handle multiple sets of linear relationships.

Chapter 7

The bi-Gaussian ensemble Kalman filter (BGEnKF)

7.1 Introduction and overview

As discussed in Chapter 6, the Gaussian prior assumption is sub-optimal because this assumption cannot handle mixture statistics. Since Gaussian ensemble DA methods have been remarkably successful at assimilating GeoIR BTs (Otkin, 2012; Zhang et al., 2016; Minamide and Zhang, 2018; Honda et al., 2018b; Otkin and Potthast, 2019; Sawada et al., 2019; Okamoto et al., 2019; Zhang et al., 2019; Geer et al., 2019; Chan et al., 2020b; Jones et al., 2020; Chan and Chen, 2021), replacing the Gaussian prior assumption with one that can handle mixture statistics can potentially enhance these successes.

One approach to addressing such mixed statistics would be to handle the clear member columns separately from the cloudy member columns. This can be achieved by assuming that the prior ensemble is drawn from two Gaussian kernels: one Gaussian kernel for the clear member columns and another Gaussian kernel for the cloudy member columns (Chan et al., 2020a). In other words, the Gaussian prior assumption is replaced with the assumption that

the prior ensemble is drawn from a two-kernel Gaussian mixture model (GMM) distribution (henceforth, the bi-Gaussian prior assumption). The bi-Gaussian prior assumption is a weaker form of the Gaussian prior assumption since a Gaussian distribution is a special form of the bi-Gaussian distribution.

The goal of this chapter is to propose a computationally efficient and scalable bi-Gaussian extension of the EnKF (BGEnKF) to assimilate GeoIR observations. Similar to the chapter on the EnSRF (Chapter 3), this chapter is broken into two layers: the conceptual layer and the mathematical layer. The conceptual layer will cover:

1. the assumptions underlying my BGEnKF,

2. a general outline of the BGEnKF procedure,

3. a comparison between previous GMM extensions of the EnKF and my BGEnKF, and,

4. heuristic measures used to improve my BGEnKF.

The mathematical layer will cover:

1. a derivation of the bi-Gaussian posterior pdf from the BGEnKF assumptions, and,

2. the construction of my BGEnKF from the bi-Gaussian posterior pdf.

Note that throughout this chapter, I will formulate the BGEnKF in terms of a serial filter. The serial BGEnKF procedure is identical to that of Chapter 3.3, except that the EnKF update procedure is replaced by the BGEnKF procedure.

THE CONCEPTUAL LAYER

7.2 Assumptions underlying the BGEnKF

7.2.1 List of assumptions

Similar to the EnKF (Chapter 3), the BGEnKF assimilates observations into an ensemble of forecasted atmospheric states $\left\{\boldsymbol{\psi}_1^f, \boldsymbol{\psi}_2^f, \ldots, \boldsymbol{\psi}_{N_E}^f\right\}$ via Bayes' rule. The BGEnKF procedure can be derived by making the following assumptions.

1. Observation errors are unbiased and Gaussian.

2. Observations can be assimilated serially.[1]

3. When assimilating an observation, I assume that the creation process of any forecast member can approximated by the following procedure.

   (a) Flip a weighed coin to select between two Gaussian kernels (clear kernel and cloudy kernel).

   (b) Then, draw this member from the selected kernel.

4. The kernel that a forecast member is drawn from can be determined by checking whether the member's simulated observation column is clear or cloudy.

5. The forecast ensemble covariance matrix for every kernel is invertible[2].

[1] In other words, observation errors are assumed to be independent. See Chapter 3.7.

[2] This assumption is necessary because if both covariance matrices are singular, there is a considerable chance that the subspace of one covariance matrix is different from that of the other covariance matrix. Without a common subspace, the subspace trick used in Chapter 3.6.2 cannot be used to write the forecast pdf. As such, I decided to assume that the two forecast covariance matrices are invertible.

Side-note: the EnKF and BGEnKF make the same assumptions about observation errors (*i.e.*, independent Gaussian observation errors and Gaussian observation likelihoods).

7.2.2 The bi-Gaussian forecast pdf

The BGEnKF differs from the EnKF in terms of the assumptions made about the forecast ensemble. Assumption 3 implies that the forecast pdf is bi-Gaussian. Bi-Gaussian pdfs are defined as the weighted sum of two Gaussian kernels where the weights sum to unity. As such, a bi-Gaussian pdf is defined by the mean state, covariance matrix and weight assigned to each kernel (*i.e.*, there are six defining parameters).

To write out this forecast pdf, suppose quantities related to the clear kernel are denoted by the subscript clr and quantities related to the cloudy kernel are denoted by the subscript cld. Suppose further that g is a placeholder that can be replaced with "clr" or "cld", $\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}}$ is the mean state of kernel g, $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}$ is the covariance matrix of kernel g, and w^f_g is the weight of kernel g. Assuming that the kernel covariance matrices are invertible[3], forecast pdf can be written as (Alspach and Sorenson, 1972; Anderson and Anderson, 1999; Dovera and Della Rossa, 2011; Reich, 2012; Sondergaard and Lermusiaux, 2013a):

$$p(\boldsymbol{\psi}) = w^f_{\text{clr}}\,\mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\text{clr}}}, \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\text{clr}}\right) + w^f_{\text{cld}}\,\mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\text{cld}}}, \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\text{cld}}\right). \tag{7.1}$$

Here, I am using $\mathcal{G}(\cdot;\ \ \cdot,\cdot)$ to denote a Gaussian pdf. For a K-dimensional state vector $\boldsymbol{p}$ with some mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$, I define

$$\mathcal{G}(\boldsymbol{p};\ \ \boldsymbol{\mu},\boldsymbol{C}) \equiv \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{C})}} \exp\left\{-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu})^\top \boldsymbol{C}^{-\mathbf{1}} (\boldsymbol{p}-\boldsymbol{\mu})\right\}. \tag{7.2}$$

[3] (*i.e.*, assumption 5 in Chapter 7.2.1)

7.2.3 The resulting bi-Gaussian posterior pdf

Suppose an observation y^o with error variance σ^{o2} is assimilated. Combining the bi-Gaussian forecast pdf with the Gaussian observation likelihood through Bayes' rule produces the following bi-Gaussian posterior pdf (Alspach and Sorenson, 1972; Anderson and Anderson, 1999; Dovera and Della Rossa, 2011; Reich, 2012; Sondergaard and Lermusiaux, 2013a)[4]:

$$p\left(\boldsymbol{\psi}|y^o\right) = w^a_{\text{clr}}\,\mathcal{G}\left(\boldsymbol{\psi};\ \overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\text{clr}}}, \boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\text{clr}}\right) + w^a_{\text{cld}}\,\mathcal{G}\left(\boldsymbol{\psi};\ \overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\text{cld}}}, \boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\text{cld}}\right) \tag{7.3}$$

where $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{g}}}$ represents the posterior average state of cluster g, $\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\boldsymbol{g}}$ represents the posterior covariance matrix of cluster g, and w^a_g is the posterior weight of cluster g. $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{g}}}$ and $\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\boldsymbol{g}}$ are related to $\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}}$ and $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}$ via the Kalman filter (KF) equations (Alspach and Sorenson, 1972; Anderson and Anderson, 1999; Dovera and Della Rossa, 2011; Reich, 2012; Sondergaard and Lermusiaux, 2013a):[5]

$$\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{g}}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}} + \mathbf{K_g}\left[\mathbf{y^o} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}}\right],\ \text{ and, }\ \boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\boldsymbol{g}} = \left(\mathbf{I} - \mathbf{K_g}\,\mathbf{H}\right)\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}} \tag{7.4}$$

where

$$\mathbf{K_g} \equiv \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}\,\mathbf{H}^\top\left(\mathbf{H}\,\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}\,\mathbf{H}^\top + \sigma^{o2}\right)^{-1}. \tag{7.5}$$

The posterior weights are related to the forecast pdf's parameters via

$$w^a_g = \frac{w^f_g\,\alpha_g}{w^f_{\text{clr}}\,\alpha_{\text{clr}} + w^f_{\text{cld}}\,\alpha_{\text{cld}}},\ \text{ and, }\ \alpha_g = \mathcal{G}\left(y^o;\ \boldsymbol{H}\,\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}},\ \sigma^{o2} + \boldsymbol{H}\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{g}}\boldsymbol{H}^\top\right). \tag{7.6}$$

[4] This posterior pdf be derived later.

[5] See Chapter 3.6 for the derivation.

7.3 The BGEnKF update procedure

7.3.1 Outline of serial filtering procedure

The BGEnKF assimilates observations into the forecast ensemble through the following serial assimilation procedure:

1. Construct $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}}\right\}$ from the forecast ensemble of model state vectors $\left\{\mathbf{x}_{\mathbf{1}}^{\mathbf{f}}, \mathbf{x}_{\mathbf{2}}^{\mathbf{f}}, \ldots, \mathbf{x}_{\mathbf{N_E}}^{\mathbf{f}}\right\}$ by evaluating Eq. (2.5) (see Chapter 7.3.2).

2. For $m = 1, 2, \ldots, N_y$,

   (a) Select the m-th observation.

   (b) Split the ensemble into the two clusters (see Chapter 7.3.2).

   (c) Run through some heuristic checks to determine whether to use the EnKF or the BGEnKF (see Chapters 7.5.2 and 7.5.3).

   (d) If any heuristic check fails, put all ensemble members into one of the cluster.[6]

   (e) Estimate each kernel's mean state, covariance matrix and weight (see Chapter 7.3.3).

   (f) Construct $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{\alpha}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{\alpha}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{\alpha}}\right\}$ by assimilating the m-th observation into $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}}\right\}$ via the three-stage BGEnKF update procedure (see Chapter 7.3.4).

   (g)

   (h) For $n = 1, 2, \ldots, N_E$, $\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} \leftarrow \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{\alpha}}$.

3. Exit.

[6] This will cause the BGEnKF to become mathematically equivalent to the EnKF.

7.3.2 On sorting the forecast members into clusters

In step 2b of the procedure outline (Chapter 7.3.1), the BGEnKF uses assumption 4 in the earlier list of assumptions to split the ensemble into the two clusters (one cluster for the clear kernel, one cluster for the cloudy kernel). For simplicity, if a member's column-integrated ice mass content [defined in Eq. (6.1)] at the observation site exceeds 1 g/m^2, the member is sorted into the cloudy cluster. If otherwise, the member is sorted into the clear cluster.

Because the column-integrated ice mass content is now involved in the BGEnKF procedure, I will include this quantity at every observation site into the auxiliary variable component of $\boldsymbol{\psi}$ [see Eq. (2.2)]. The construction of $\left\{\boldsymbol{\psi}_{\mathbf{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\mathbf{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}}\right\}$ in the serial BGEnKF procedure thus involves evaluating each member's column-integrated ice mass content at every observation site [Eq. (2.5)].

7.3.3 On estimating the forecast pdf's parameters

To estimate the parameters defining the bi-Gaussian forecast pdf, suppose the set S_{clr} contains the indices[7] of forecast members drawn from the clear kernel, and the set S_{cld} contains the indices of forecast members drawn from the cloudy kernel. I can then compute the number of forecast members in cluster g (N_g^f) via

$$N_g^f \equiv \text{count}\left(S_g\right) \tag{7.7}$$

[7] These indices are the subscripts I use to denote my members. In other words, the index for $\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}}$ is n.

where count$(\cdot)$ counts the number of elements in S_g. The mean state $\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}}$, covariance matrix $\overline{\boldsymbol{P}_{\boldsymbol{g}}^{\boldsymbol{f}}}$ and weight w_g^f for kernel g can then be estimated via

$$\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}} \equiv \frac{1}{N_g^f} \sum_{n \in S_g} \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}}, \qquad \overline{\boldsymbol{P}_{\boldsymbol{g}}^{\boldsymbol{f}}} \equiv \frac{1}{N_g^f - 1} \sum_{n \in S_g} \left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} - \overline{\boldsymbol{\psi}_{\mathbf{g}}^{\boldsymbol{f}}} \right) \left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} - \overline{\boldsymbol{\psi}_{\mathbf{g}}^{\boldsymbol{f}}} \right)^{\mathsf{T}},$$
$$\text{and,} \quad w_g^f \equiv \frac{N_g^f}{N_{\text{clr}}^f + N_{\text{cld}}^f}. \tag{7.8}$$

7.3.4 The three-stage BGEnKF update procedure

The BGEnKF assimilates the m-th observation into $\left\{ \boldsymbol{\psi}_{\boldsymbol{1}}^{\boldsymbol{f}}, \boldsymbol{\psi}_{\boldsymbol{2}}^{\boldsymbol{f}}, \ldots, \boldsymbol{\psi}_{\boldsymbol{N_E}}^{\boldsymbol{f}} \right\}$ via a three-stage update procedure. These three stages are illustrated in Figure 7.1. In order of execution, these stages are: 1) the double EnKF stage, 2) the shrinking cluster member deletion stage, and 3) the expanding cluster member resampling stage. An outline of the three-stage procedure is available at the end of this section.

The double EnKF stage

The first stage [Figure 7.1(a)] is to represent the KF updates to the clusters' mean states and covariance matrices. Like the rest of this dissertation, I use the ensemble square root filter of Whitaker and Hamill (2002) (EnSRF) to update each cluster's members. The EnSRF update equation (Whitaker and Hamill, 2002) for members in cluster g (*i.e.*, $\left\{ \boldsymbol{\psi}_{\boldsymbol{n_g}}^{\boldsymbol{a}} | \forall\ n_g \in S_g \right\}$) is

$$\boldsymbol{\psi}_{\boldsymbol{n_g}}^{\boldsymbol{a}} = \boldsymbol{\psi}_{\boldsymbol{n_g}}^{\boldsymbol{f}} + \mathbf{K_g} \left(y^o - \boldsymbol{H} \overline{\boldsymbol{\psi}_{\boldsymbol{g}}} \right) - \phi_g \mathbf{K_g} \left(\boldsymbol{H} \boldsymbol{\psi}_{\boldsymbol{n_g}}^{\boldsymbol{f}} - \boldsymbol{H} \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}} \right) \quad \forall\ n_g \in S_g. \tag{7.9}$$

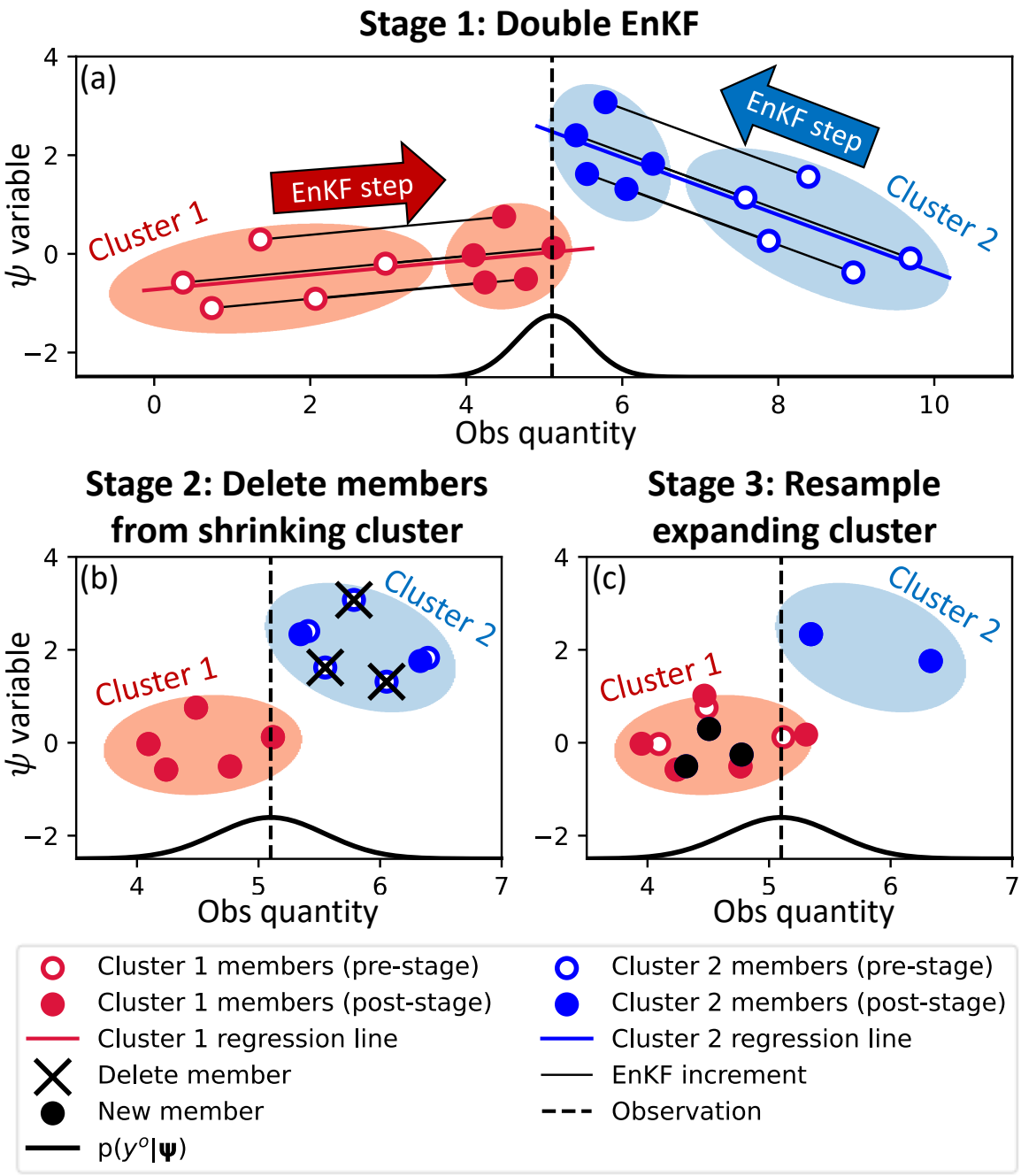


Figure 7.1: *See next page for caption*

Figure 7.1 *(previous page)*: A bivariate demonstration of the three-stage process of the BGEnKF algorithm. The light red ovals highlight cluster 1 members and the light blue ovals highlight cluster 2 members. Prior to running the BGEnKF update, the prior members have already been separated into two clusters. The BGEnKF's first stage is to employ the EnKF update equations on the two clusters separately (panel a). In the second stage (panel b), the BGEnKF identifies the shrinking cluster (the blue cluster 2 in this case), deletes an appropriate number of members from this cluster, and adjusts the remaining members to prevent the deletion from changing this cluster's mean. The BGEnKF's final stage (panel c) is to recreate the deleted members by resampling from the expanding cluster (cluster 1).

Here, $\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n_g}}$ represents a forecast ensemble member in cluster g. The Kalman gain matrix of cluster g ($\boldsymbol{K_g}$) can be computed via Eq. (7.5)). ϕ_g is the EnSRF's square-root modification factor (Whitaker and Hamill, 2002), which can be computed via

$$\phi_g \equiv \left\{ 1 + \sqrt{\frac{\sigma^{o2}}{\sigma^{o2} + \text{Var}\left(\boldsymbol{H}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}\right)}} \right\}^{-1} . \tag{7.10}$$

Note that the EnSRF-based cluster update equations can be replaced with those from the two-step ensemble adjustment Kalman filter (EAKF) of Anderson (2003). This is because the two filters have mathematically identical ensemble member update procedures.

The member deletion stage

In the second and third stages of the BGEnKF (Figure 7.1(b & c)), the number of ensemble members in each cluster (*i.e.*, cluster sizes) is updated to be consistent with the cluster's posterior weight [Eq. (7.6)]. The post-BGEnKF size of cluster g (N^a_g) can be determined by

$$N^a_g \equiv \text{round}\left(N_E * w^a_g\right) \tag{7.11}$$

where round$(\cdot)$ indicates rounding $\cdot$ to the nearest integer.

If the size of a cluster is reduced by the assimilation of y^o, members will be deleted from this cluster (Figure 7.1(b)). The number of members to be deleted N_{del} is defined as

$$N_{\text{del}} \equiv \begin{cases} N^f_{\text{clr}} - N^a_{\text{clr}} & \text{if } N^a_{\text{clr}} < N^f_{\text{clr}}, \\ \\ N^f_{\text{cld}} - N^a_{\text{cld}} & \text{if } N^a_{\text{cld}} < N^f_{\text{cld}}. \end{cases} \tag{7.12}$$

For simplicity, I will delete the members with the smallest N_{del} forecast-simulated observation perturbations. Since the deletion will cause the cluster's mean state to deviate from the theoretical mean state [Eq. (7.4)], I will recenter the remaining members around this theoretical value.

The resampling stage

If the size of one cluster is reduced by the assimilation of y^o, the other cluster compensates by increasing in size. The resampling procedure I use[8] is as follows. Supposing that the subscript "pre" denotes expanding cluster quantities before resampling, we can compute the pre-resampling perturbations $\left\{\boldsymbol{\psi}^{\boldsymbol{a}'}_{\boldsymbol{n}} | n \in S_{\text{pre}}\right\}$ via

$$\boldsymbol{\psi}^{\boldsymbol{a}'}_{\boldsymbol{n}} \equiv \boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n}} - \overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\mathbf{pre}}} \quad \forall \ n \in S_{\text{pre}} \tag{7.13}$$

where $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\mathbf{pre}}}$ is the expanding cluster's mean state and S_{pre} is the set of member indices in the expanding cluster before resampling. Furthermore, let $\boldsymbol{\Psi}_{\mathbf{pre}}$ is a matrix where each column contains a pre-resampling perturbation, and $\boldsymbol{\Psi}_{\mathbf{post}}$ is a matrix where each column contains a post-resampling perturbation. Supposing the pre-resampling cluster size is denoted by N_{pre} and the post-

[8] I proposed this resampling procedure in Chan et al. (2020a). This procedure is computationally efficient.

resampling cluster size is denoted by N_{post}, then $\mathbf{\Psi_{pre}}$ is an $N_\psi \times N_{\text{pre}}$ matrix and $\mathbf{\Psi_{post}}$ is an $N_\psi \times N_{\text{post}}$ matrix. If I denote the ℓ-th member index in S_{pre} as $n_{\text{pre},\ell}$, and likewise for the ℓ-th member index in S_{post}, we can explicitly write out $\mathbf{\Psi_{pre}}$ and $\mathbf{\Psi_{post}}$:

$$\begin{aligned} \mathbf{\Psi_{pre}} &\equiv \begin{bmatrix} \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},1}} & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},2}} & \cdots & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}}} \end{bmatrix}, \\ \mathbf{\Psi_{post}} &\equiv \begin{bmatrix} \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{post},1}} & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{post},2}} & \cdots & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{post},\boldsymbol{N}_{\mathbf{post}}}} \end{bmatrix}. \end{aligned} \tag{7.14}$$

My resampling strategy can be thus expressed as

$$\mathbf{\Psi_{post}} \equiv \mathbf{\Psi_{pre}} \boldsymbol{T} \tag{7.15}$$

where $\boldsymbol{T}$ is an $N_{\text{pre}} \times N_{\text{post}}$ matrix of resampling coefficients. The form of $\boldsymbol{T}$ I use is

$$\boldsymbol{T} \equiv \begin{bmatrix} k\boldsymbol{I}_{\boldsymbol{N}_{\mathbf{pre}} - \boldsymbol{N}^*_{\mathbf{new}}} & \mathbf{0}_{\left(\boldsymbol{N}_{\mathbf{pre}} - \boldsymbol{N}^*_{\mathbf{new}}\right) \times \boldsymbol{N}^*_{\mathbf{new}}} & \mathbf{0}_{\left(\boldsymbol{N}_{\mathbf{pre}} - \boldsymbol{N}^*_{\mathbf{new}}\right) \times \boldsymbol{N}_{\mathbf{new}}} \\ \mathbf{0}_{\boldsymbol{N}^*_{\mathbf{new}} \times \left(\boldsymbol{N}_{\mathbf{pre}} - \boldsymbol{N}^*_{\mathbf{new}}\right)} & \boldsymbol{I}_{\boldsymbol{N}^*_{\mathbf{new}}} & \boldsymbol{E} \end{bmatrix} \tag{7.16}$$

where N_{new} is the number of members being added and N^*_{new} is related to N_{new} and N_{pre}. Specifically,

$$N_{\text{new}} \equiv N_{\text{post}} - N_{\text{pre}}, \quad \text{and,} \quad N^*_{\text{new}} \equiv \begin{cases} N_{\text{new}} - 1 & \forall\ N_{\text{new}} \leq N_{\text{pre}} \\ N_{\text{pre}} & \text{otherwise} \end{cases}. \tag{7.17}$$

Furthermore, for arbitrary integers η and μ, $\boldsymbol{I_\eta}$ is an $\eta \times \eta$ identity matrix, $\mathbf{0}_{\boldsymbol{\eta}\times\boldsymbol{\mu}}$ is an $\eta \times \mu$ matrix of zeros. k is the following scalar inflation factor

$$k \equiv \sqrt{\frac{N_{\text{new}} + N_{\text{pre}} - 1}{N_{\text{pre}} - 1}} \quad \text{(note that } k \geq 1\text{)}. \tag{7.18}$$

Finally, the matrix $\boldsymbol{E}$ in Eq. (7.19) is an $N^*_{\text{new}} \times N_{\text{new}}$ matrix that I define as

$$\boldsymbol{E} \equiv \frac{k-1}{N_{\text{new}}} \mathbf{1}_{\boldsymbol{N^*_{\text{new}}}\times\boldsymbol{N_{\text{new}}}} + \boldsymbol{L_E}(\boldsymbol{L_W})^{-\mathbf{1}}\boldsymbol{W}. \tag{7.19}$$

Here, $\mathbf{1}_{\boldsymbol{N^*_{\text{new}}}\times\boldsymbol{N_{\text{new}}}}$ denotes an $N^*_{\text{new}} \times N_{\text{new}}$ matrix whose elements are all set to unity. Furthermore, $\boldsymbol{W}$ is an $N^*_{\text{new}} \times N_{\text{new}}$ matrix of the form

$$\boldsymbol{W} \equiv \left[\boldsymbol{I}_{\boldsymbol{N^*_{\text{new}}}} \quad \mathbf{0}_{\boldsymbol{N^*_{\text{new}}}\times\left(\boldsymbol{N_{\text{new}}}-\boldsymbol{N^*_{\text{new}}}\right)}\right] - \frac{1}{N_{\text{new}}} \mathbf{1}_{\boldsymbol{N^*_{\text{new}}}\times\boldsymbol{N_{\text{new}}}}. \tag{7.20}$$

Supposing that $\text{Chol}(\boldsymbol{S})$ denotes the Cholesky decomposition of an arbitrary symmetric matrix $\boldsymbol{S}$, I define

$$\boldsymbol{L_W} \equiv \text{Chol}\left(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}\right), \tag{7.21}$$

and

$$\boldsymbol{L_E} \equiv \text{Chol}\left(\frac{N_{\text{new}}}{N_{\text{pre}} - 1}\boldsymbol{I}_{\boldsymbol{N^*_{\text{new}}}} - \frac{(k-1)^2}{N_{\text{new}}}\mathbf{1}_{\boldsymbol{N^*_{\text{new}}}\times\boldsymbol{N^*_{\text{new}}}}\right). \tag{7.22}$$

This resampling procedure is guaranteed to preserve the mean and covariance of the expanding cluster.

Outline of the three-stage BGEnKF update procedure

Stage 1: Double EnKF [illustrated in Figure 7.1(a)]

1. Do g = clr, cld
   (a) For cluster g, compute the Kalman gain [$\boldsymbol{K_g}$; Eq. (7.5)] and square-root modification factor [ϕ_g; Eq. (7.10)].
   (b) Evaluate Eq. (7.9) for every ensemble member in cluster g.

Stage 2: Shrinking cluster member deletion [illustrated in Figure 7.1(b)]

1. Evaluate Eq. (7.11) to determine the targeted cluster sizes after assimilating the observation
2. If $N^a_{\text{clr}} < N^f_{\text{clr}}$, the clear cluster will be considered as the shrinking cluster.
3. If $N^a_{\text{cld}} < N^f_{\text{cld}}$, the cloudy cluster will be considered as the shrinking cluster.
4. If no shrinking cluster has been identified, terminate the current stage.
5. Compute N_{del} using Eq. (7.12).
6. Compute the current mean state of the shrinking cluster.
7. Delete the members with the smallest N_{del} forecast-simulated observation perturbations within the shrinking cluster.
8. Compute the mean state of the remaining members in the shrinking cluster.
9. Subtract the mean found in step 8 from the mean found in step 6.
10. Add the difference computed in step 9 to each of the remaining members in the shrinking cluster to recenter said members on the pre-deletion shrinking cluster mean state.

Stage 3: Resample expanding cluster members [illustrated in Figure 7.1(c)]

1. Evaluate Eq. (7.11) to determine the targeted cluster sizes after assimilating the observation
2. If $N^a_{\text{clr}} > N^f_{\text{clr}}$, the clear cluster will be considered as the expanding cluster.
3. If $N^a_{\text{cld}} > N^f_{\text{cld}}$, the cloudy cluster will be considered as the expanding cluster.
4. If no expanding cluster has been identified, terminate the current stage.
5. Compute N_{new} and N^*_{new} using Eq. (7.17).
6. Compute the expanding cluster's mean state vector.
7. Construct the expanding cluster's perturbation vectors via Eq. (7.13).
8. Construct matrix $\boldsymbol{W}$ by evaluating Eq. (7.20).
9. Construct $\boldsymbol{L_W}$ and $\boldsymbol{L_E}$ by evaluating Eqs. (7.21) and (7.22).
10. Construct $\boldsymbol{E}$ by evaluating Eq. (7.19).
11. Construct $\boldsymbol{T}$ by evaluating Eq. (7.16).
12. Evaluate Eq. (7.15) to resample the expanding cluster perturbations.
13. Add the expanding cluster's mean state (computed in step 6) to the resampled perturbations to construct the resampled expanding cluster ensemble members.

7.4 A discussion of various GMM-EnKFs (including my BGEnKF)

7.4.1 Overview

Various studies have proposed different methods to estimate the prior GMM PDF parameters and extended the Gaussian EnKFs to handle GMM prior PDFs (GMM-EnKFs). There are two main challenges with using GMM-EnKFs for numerical weather prediction: 1) estimating the forecast pdf's defining GMM parameters, and 2) representing the change from the forecasted kernel weights to the posterior kernel weights. In this section, I will elaborate on these challenges and how my BGEnKF circumvents these challenges.

7.4.2 On estimating the forecast pdf's GMM parameters

GMM pdfs are defined by four parameters (Alspach and Sorenson, 1972): 1) the number of Gaussian kernels, 2) the weight of each kernel, 3) the mean state vector of each kernel, and 4) the covariance matrix of each kernel. These parameters must be specified/estimated from the forecast ensemble to perform any form of GMM Bayesian inference. GMM-EnKF formulations (including mine) either use a heuristic approach or an objective approach to specify/estimate these parameters.

The GMM-EnKFs of Anderson and Anderson (1999) and Bengtsson et al. (2003) use heuristic approaches to estimate the forcast pdf's GMM parameters. Anderson and Anderson (1999) efficiently estimated the parameters by heuristically assuming that the number of Gaussian kernels is the same as the number of ensemble members and that the center of each kernel corresponds to an ensemble member. Furthermore, the covariance of each kernel is a rescaled version of the entire ensemble's covariance. However, these assumptions are inappropriate for infrared DA because only two kernels should

suffice to approximate the clear-sky and cloudy members, and each kernel has a different covariance. Bengtsson et al. (2003) then proposed a GMM EnKF that constructed G kernels by grouping ensemble members in the vicinity of G randomly selected members, in high-dimensional state space. This method is very susceptible to sampling noise (Sondergaard and Lermusiaux, 2013a).

More objective approaches to estimate the forecast pdf's GMM parameters were suggested by Dovera and Della Rossa (2011) and Sondergaard and Lermusiaux (2013a). Dovera and Della Rossa (2011) started by assuming a small number of kernels and then fitted the kernel parameters to the forecast ensemble using an expectation maximization (EM) algorithm. However, in spaces with more than 100 dimensions, the EM algorithm can frequently run into values that are smaller than the computer's floating point precision. Furthermore, EM algorithms can be expensive to employ in high-dimensional spaces.

Sondergaard and Lermusiaux (2013a) proposed an objective approach to fit all 4 forecast pdf GMM parameters. Essentially, the EM kernel estimation algorithm is executed multiple times, each time with different numbers of kernels, to determine different possible GMM forecast pdfs. The GMM forecast pdf that best fits the ensemble (determined using the Bayesian information criterion) is then selected. Since EM algorithms are expensive and susceptible in high-dimensional spaces, Sondergaard and Lermusiaux (2013a) also proposed executing the EM algorithm in a stochastic subspace based on the forecast model's governing equations (Sapsis and Lermusiaux, 2009, 2012; Ueckermann et al., 2013). Unfortunately, utilizing this stochastic subspace requires deriving and coding a set of dynamically orthogonal field equations from the model equations. This procedure is harder and more labor-intensive than constructing model adjoint operators for four-dimensional variational data assimilation. Simpler, more efficient and more appropriate methods of estimating the prior GMM PDF parameters are thus necessary in order to use GMM ensemble DA to practically assimilate satellite infrared observations.

My BGEnKF utilizes a heuristic approach to determine the forecast pdf's GMM parameters. Specifically, I assume that there are only 2 kernels and every member can be assigned to one of them based on whether the member's simulated observed column is clear/cloudy (assumption 4 in Chapter 7.2.1). These assumptions circumvent the need to use EM-based kernel fitting, thus avoiding the issues with computational complexity (Dovera and Della Rossa, 2011), finite numerical precision (Dovera and Della Rossa, 2011), programming difficulty (Sondergaard and Lermusiaux, 2013a) and difficulty in maintainence (Sondergaard and Lermusiaux, 2013a).

7.4.3 On representing the updates to the kernel weights

Another challenge of GMM ensemble DA is the difficulty in updating the ensemble in a way consistent with the posterior parameters. The traditional approach is to directly compute the posterior parameters first, and then draw random samples based on these posterior parameters (Anderson and Anderson, 1999; Dovera and Della Rossa, 2011; Sondergaard and Lermusiaux, 2013a). Standard drawing methods require the square-roots of the posterior covariance matrices (Rasmussen and Williams, 2005), which are computationally expensive to determine in a high-dimensional space. Even with localization $N_\psi \sim 10^4$, standard matrix square-root computations have a computational complexity of $\sim N_\psi^3 = 10^{12}$. While Sondergaard and Lermusiaux (2013a,b) performed the sampling in a low-dimensional subspace, their method is difficult to employ for complex models (see previous paragraph).

A second approach to update the ensemble is to first adjust the number of members in each kernel's cluster until the number of members in each cluster is consistent with the posterior weights, and then perform perturbed observation EnKF updates on the ensemble members in each cluster (Bengtsson et al., 2003). The cluster sizes are adjusted by duplicating (deleting) random mem-

bers in clusters where the weight is increased (decreased) by DA. Since this second approach relies on perturbing the observations to prevent the emergence of identical posterior ensemble members, a deterministic EnKF cannot be employed with this approach.[9]

My BGEnKF does not require using the square-root of the posterior kernel covariance matrices. All that is done is to delete members from the shrinking cluster, and then resample members from the expanding cluster using a matrix multiplication. This matrix multiplication effectively resamples the expanding cluster in the subspace spanned by this cluster's pre-resampling perturbations. The most expensive part of this multiplication is multiplying the submatrix $\boldsymbol{E}$ to $\boldsymbol{\Psi_{pre}}$. For $N^*_{\text{new}} \sim 10$ and $N_\psi \sim 10^4$, this submatrix multiplcation procedure has a computational complexity of $\sim N_\psi \left(N^*_{\text{new}}\right)^2 = 10^6$. This computational complexity is 6 orders of magnitude smaller than the approaches involving the square root of covariance matrices ($\sim 10^{12}$). Furthermore, unlike the GMM method of Bengtsson et al. (2003) my BGEnKF is entirely deterministic.[10].

7.5 Heuristic measures used with my BGEnKF

7.5.1 Localization

The BGEnKF is likely more susceptible to sampling noise than the EnKF because the sample size used to estimate each cluster's mean state and Kalman gain are smaller than the sample size used to estimate the mean state and covariance matrix of the entire ensemble. As such, I employ two heuristic measures that are similar to those of Chan et al. (2020a). First, I spatially localize the BGEnKF analysis increment using the Gaspari-Cohn fifth order polynomial [GC99; Gaspari and Cohn (1999)]. If $\boldsymbol{\rho}$ represents a vector of GC99 localization

[9] See Whitaker and Hamill (2002) for a discussion on the shortcomings of stochastic EnKFs.

[10] Note that if a stochastic form of my BGEnKF is desired, just replace the EnSRF procedure with the Monte Carlo EnKF of Burgers et al. (1998)

factors, I construct the localized updated extended state vector for member n via

$$\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} \leftarrow \boldsymbol{\rho} \circ \left(\boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{a}} - \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} \right) + \boldsymbol{\psi}_{\boldsymbol{n}}^{\boldsymbol{f}} \tag{7.23}$$

where $\circ$ represents element-wise multiplication. In the cases where either $w_{\text{clr}}^{f} = 1$ or $w_{\text{cld}}^{f} = 1$ (*i.e.,* the bi-Gaussian prior p.d.f. turns Gaussian), this localization method is identical to Kalman gain localization [*e.g.*, Anderson et al. (2009), Meng and Zhang (2008), Whitaker et al. (2008), Houtekamer and Zhang (2016)].

Note that this localization method [Eq. (7.23)] localizes the impacts of replacing clear members with cloudy members (or *vice versa*). As an example, suppose the BGEnKF replaces a cloudy forecast member with a clear analysis member. The localization process [Eq. (7.23)] first computes the difference between the cloudy forecast member and the clear analysis member (*i.e.*, the member's change due to the BGEnKF). This difference is then localized and applied to the cloudy forecast member. The resulting member follows the clear analysis member at the observation site and becomes increasingly like the cloudy forecast member with increasing distance from the observation site. Future work can examine other approaches to localize the impacts of deleting and replacing ensemble members.

7.5.2 Handling overly small clusters

The second heuristic sampling error mitigation measure is to switch from using the BGEnKF to using the EnKF whenever the pre-resampling expanding cluster is too small ($N_{\text{pre}} < 0.8N_E$), or whenever any cluster is too small (less than $0.1N_E$). A similar heuristic measure is used in Chan et al. (2020a).

7.5.3 Mitigating unphysical weight updates

Another issue specific to the BGEnKF is its occasional tendency to generate unphysical weight updates. Specifically, the BGEnKF occasionally expands the clear cluster when a cloudy observation is assimilated, and *vice versa*. This is because the BGEnKF does not explicitly consider whether an observation is clear or cloudy when assimilating it.

The BGEnKF is automatically switched to the EnKF whenever an unphysical weight update is detected. To do so, I first identify the whether the observation to be assimilated is definitively clear or cloudy. For instance, in the case of Window-BT values over tropical ocean, observation values warmer than 290 K are definitively clear, and observation values cooler than 280 K are definitively cloudy. If the observation is definitively clear, but the cloudy cluster is expanded by the BGEnKF, or *vice versa*, the BGEnKF will switch over to the EnKF.

THE MATHEMATICAL LAYER

7.6 Derivation of the BGEnKF posterior pdf

I will now derive the posterior pdf [Eq. (7.3)] resulting from assimilating observations with Gaussian errors (*i.e.*, Gaussian observation likelihood) into a bi-Gaussian forecast pdf [Eq. (7.1)]. For generality, the posterior pdf will first be derived for a forecast pdf with an arbitrary number of Gaussian kernels (N_k). I will then write out the posterior pdf for the bi-Gaussian forecast pdf.

7.6.1 Derivation for an arbitrary GMM forecast pdf

A GMM forecast pdf with N_k kernels can be written as

$$p(\boldsymbol{\psi}) \equiv \sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\boldsymbol{\psi};\ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}},\ \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\right) \tag{7.24}$$

where w_g^f are scalar weights such that

$$\sum_{g=1}^{N_k} w_g^f = 1, \quad \text{and,} \quad w_g^f \in [0,\, 1] \quad \forall \quad g = 1, 2, \ldots, N_k.$$

Futhermore, $\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}}$ is the mean state of kernel g and $\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}$ is the forecast covariance matrix of kernel g. Note that the forecast covariance matrix of every kernel is assumed to be invertible[11].

[11] I will explain why this assumption is necessary in the **APPENDIX**

For Gaussian observation errors, the observation likelihood is

$$p\left(\mathbf{y^o}|\boldsymbol{\psi}\right) \equiv \mathcal{G}\left(\mathbf{H}\,\boldsymbol{\psi};\ \ \mathbf{y^o},\, \boldsymbol{R}\right).$$

Bayes' rule thus informs us that the posterior pdf is

$$p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) \equiv \frac{p\left(\boldsymbol{\psi}\right) p\left(\mathbf{y^o}|\boldsymbol{\psi}\right)}{p\left(\mathbf{y^o}\right)} = \frac{\left[\sum_{g=1}^{N_k} w_g^f\, \mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\right)\right] * \mathcal{G}\left(\mathbf{H}\,\boldsymbol{\psi};\ \ \mathbf{y^o},\, \boldsymbol{R}\right)}{p\left(\mathbf{y^o}\right)}$$
$$= \frac{1}{p\left(\mathbf{y^o}\right)} \sum_{g=1}^{N_k} w_g^f \left[\mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\right) * \mathcal{G}\left(\mathbf{H}\,\boldsymbol{\psi};\ \ \mathbf{y^o},\, \boldsymbol{R}\right)\right].$$

The multiplication of two Gaussian kernels yields a scaled Gaussian kernel (proven in Chapter 3.9.3). By Eq. (3.59),

$$p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) = \frac{1}{p\left(\mathbf{y^o}\right)} \sum_{g=1}^{N_k} w_g^f\, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^{\top} + \boldsymbol{R}\right) \mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}}\right) \quad (7.25)$$

where

$$\begin{aligned} \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}} &= \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}} + \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^{\top}\left(\mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\left[\mathbf{y^o} - \mathbf{H}\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}}\right], \quad \text{and,} \\ \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}} &= \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}} - \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^{\top}\left(\mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^{\top} + \boldsymbol{R}\right)^{-1}\mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}} \end{aligned} \quad (7.26)$$

See Chapter 3.9.3 for the derivation of the expressions in Eq. (7.26).

To proceed, consider that $p\left(\mathbf{y^o}\right)$ serves to normalize Bayes' rule. Thus,

$$p\left(byo\right) = \int_{\mathbb{R}^{N_\psi}} p\left(\mathbf{y^o} \cap \boldsymbol{\psi}\right) \mathrm{d}^{N_\psi}\boldsymbol{\psi} = \int_{\mathbb{R}^{N_\psi}} p\left(\boldsymbol{\psi}\right) p\left(\mathbf{y^o}|\boldsymbol{\psi}\right) \mathrm{d}^{N_\psi}\boldsymbol{\psi}.$$

$$\therefore p\left(\mathbf{y^o}\right) = \int_{\mathbb{R}^{N_\psi}} \left\{ \sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right) \mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}} \right) \right\} \mathrm{d}^{N_\psi}\boldsymbol{\psi}.$$

$$= \sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right) \int_{\mathbb{R}^{N_\psi}} \mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}} \right) \mathrm{d}^{N_\psi}\boldsymbol{\psi}.$$

Since the integral of $\mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}} \right)$ is unity,

$$p\left(\mathbf{y^o}\right) = \sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right). \tag{7.27}$$

Substituting Eq. (7.27) into Eq. (7.25) yields the following GMM posterior pdf:

$$p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) = \frac{\sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right) \mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}} \right)}{\sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right)}.$$

Thus,

$$\boxed{p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) = \sum_{g=1}^{N_k} w_g^a \mathcal{G}\left(\boldsymbol{\psi};\ \ \overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{a}}},\, \boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{a}} \right)} \tag{7.28}$$

where

$$w_g^a \equiv \frac{w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right)}{\sum_{g=1}^{N_k} w_g^f \, \mathcal{G}\left(\mathbf{H}\,\overline{\boldsymbol{\psi}_{\boldsymbol{g}}^{\boldsymbol{f}}};\ \ \mathbf{y^o},\, \mathbf{H}\,\boldsymbol{P}_{\boldsymbol{\psi},\boldsymbol{g}}^{\boldsymbol{f}}\,\mathbf{H}^\top + \boldsymbol{R} \right)}. \tag{7.29}$$

The posterior pdf [Eq. (7.28)] is clearly a GMM pdf. Note that $\sum_{g=1}^{N_k} w_g^a = 1$.

7.6.2 The BGEnKF posterior pdf

The posterior pdf for a bi-Gaussian forecast pdf can be obtained by setting $N_k = 2$, defining the first Gaussian kernel ($g = 1$) as the clear kernel, and defining the second Gaussian kernel ($g = 2$) as the cloudy kernel. Eq. (7.28) then becomes

$$p\left(\boldsymbol{\psi}|\mathbf{y^o}\right) = w^a_{\text{clr}}\,\mathcal{G}\left(\boldsymbol{\psi};\ \overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\text{clr}}}, \boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\text{clr}}\right) + w^a_{\text{cld}}\,\mathcal{G}\left(\boldsymbol{\psi};\ \overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\text{cld}}}, \boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\text{cld}}\right)$$

where

$$\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{g}}} = \overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}} + \mathbf{K_g}\left[\mathbf{y^o} - \mathbf{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}}\right],\quad \boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\boldsymbol{g}} = \left(\mathbf{I} - \mathbf{K_g}\,\mathbf{H}\right)\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}},$$

$$\mathbf{K_g} \equiv \boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}\,\mathbf{H}^\top\left(\mathbf{H}\,\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}\,\mathbf{H}^\top + \boldsymbol{R}\right)^{-1},$$

$$w^a_g = \frac{w^f_g\,\alpha_g}{w^f_{\text{clr}}\,\alpha_{\text{clr}} + w^f_{\text{cld}}\,\alpha_{\text{cld}}},\ \text{ and, }\ \alpha_g = \mathcal{G}\left(\mathbf{y^o};\ \mathbf{H}\,\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}},\ \boldsymbol{R} + \mathbf{H}\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{g}}\mathbf{H}^\top\right).$$

To obtain the expressions for assimilating one observation, replace $\mathbf{y^o}$ with y^o, replace $\boldsymbol{R}$ with ${\sigma^o}^2$, and use an appropriate definition[12] of $\mathbf{H}$.

7.7 Construction of the BGEnKF procedure

7.7.1 Overview

The BGEnKF represents converting the prior bi-Gaussian pdf to the posterior bi-Gaussian pdf through updating the members in each cluster and adjusting the number of members in each cluster (henceforth, the cluster size). The former is represented by executions the EnSRF update procedure within each cluster (*i.e.*, double EnSRF executions), and the latter is represented by delet-

[12] Specifically, $\mathbf{H}$ becomes a row vector with N_ψ elements where all elements but one are zeroes. The only non-zero element is unity.

ing members from one cluster and adding members to the other cluster. In this section, I will construct and justify the BGEnKF update procedure.

7.7.2 Ensemble adjustments to reflect the posterior kernels: the double EnSRF stage

As shown in Chapter 7.6, $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{g}}}$ and $\boldsymbol{P}^{\boldsymbol{a}}_{\boldsymbol{\psi},\boldsymbol{g}}$ are related to $\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}}$ and $\boldsymbol{P}^{\boldsymbol{f}}_{\boldsymbol{\psi},\boldsymbol{g}}$ through the KF equations. This suggests that cluster g's posterior members can be constructed by applying the EnSRF procedure on the cluster g's forecast members.[13] In other words, a posterior member in cluster g is related to its forecasted version through:[14]

$$\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n_g}} = \boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n_g}} + \mathbf{K_g}\left(y^o - \boldsymbol{H}\overline{\boldsymbol{\psi}_{\boldsymbol{g}}}\right) - \phi_g \mathbf{K_g}\left(\boldsymbol{H}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{n_g}} - \boldsymbol{H}\overline{\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}}\right) \quad \forall \ n_g \in S_g$$

where

$$\phi_g \equiv \left\{1 + \sqrt{\frac{\sigma^{o2}}{\sigma^{o2} + \mathrm{Var}\left(\boldsymbol{H}\boldsymbol{\psi}^{\boldsymbol{f}}_{\boldsymbol{g}}\right)}}\right\}^{-1}.$$

To be clear, the EnSRF is executed twice: once for the clear kernel, and once for the cloudy kernel. The clear kernel EnSRF uses the clear kernel's forecast statistics and the cloudy kernel EnSRF uses the cloudy kernel forecast statistics. No clear kernel information is used in the cloudy kernel EnSRF and no cloudy kernel information is used in the clear kernel EnSRF.

[13] See Chapter 3.8 for the construction of the EnSRF.
[14] This equation is a copy of Eq. (7.9)

7.7.3 Ensemble adjustments to reflect the posterior weights: the member deletion stage

Since the g-th forecasted weight is represented by the number of forecast members in cluster g (N_g^f), to represent the g-th posterior weight, the number of posterior members in the g-th cluster is

$$N_g^a \equiv \text{round}\left(N_E * w_g^a\right).$$

Since the posterior weights are often different from the forecasted weights (see Chapter 7.6.2), the BGEnKF will delete members from the cluster with reduced weight (*i.e.*, $w_g^f > w_g^a$ or $N_g^f > N_g^a$; henceforth, the shrinking cluster), and add members to the cluster with increased weight (*i.e.*, $w_g^f < w_g^a$ or $N_g^f < N_g^a$; henceforth, the expanding cluster).

The number of members to be deleted N_{del} from the shrinking cluster is as defined Eq. (7.12):

$$N_{\text{del}} \equiv \begin{cases} N_{\text{clr}}^f - N_{\text{clr}}^a & \text{if } N_{\text{clr}}^a < N_{\text{clr}}^f, \\ \\ N_{\text{cld}}^f - N_{\text{cld}}^a & \text{if } N_{\text{cld}}^a < N_{\text{cld}}^f. \end{cases}$$

The choice of which members to delete requires heuristic constraints. Since deleting members will shift the cluster mean state, I will select members that are closest to the cluster mean state minimize this shift. However, computing the Euclidean distance between each shrinking cluster's $\boldsymbol{\psi}$ and the cluster mean state is likely to cause an inter-process communications bottleneck. As such, I will delete the N_{del} members with the smallest forecast-simulated observation perturbations [*i.e.*, the N_{del} members with the smallest $\left(y_n^f - y^o\right)^2$ values].

After deleting the members, I recentered the shrinking cluster's members

around the theoretical shrinking cluster posterior mean state [Eq. (7.4)]. This is done by evaluating the following equations in order. First compute the shifted mean state $\overline{\boldsymbol{\psi}_{\mathbf{shifted}}}$ by

$$\overline{\boldsymbol{\psi}_{\mathbf{shifted}}} \equiv \frac{1}{N^{a}_{\text{shrink}}} \sum_{n \in S^{a}_{\text{shrink}}} \boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n}}$$

where S^{a}_{shrink} contains the member indices for the remaining members in the shrinking cluster and

$$N^{a}_{\text{shrink}} \equiv \text{count}\left(S^{a}_{\text{shrink}}\right).$$

I then recentered the remaining shrinking cluster members by evaluating

$$\boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n}} \leftarrow \boldsymbol{\psi}^{\boldsymbol{a}}_{\boldsymbol{n}} - \overline{\boldsymbol{\psi}_{\mathbf{shifted}}} + \overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\mathbf{theory}}} \quad \forall \quad n \in S^{a}_{\text{shrink}}$$

where $\overline{\boldsymbol{\psi}^{\boldsymbol{a}}_{\text{theory}}}$ is the theoretical shrinking cluster posterior mean state [Eq. (7.4)].

Note that I have decided against trying to set the shrinking cluster's post-deletion covariance matrix to be equal to the theoretical shrinking cluster covariance matrix [Eq. (7.4)]. My decision is based on the fact that it is mathematically impossible for the shrinking cluster to have the same covariance matrices before and after deleting members (Chan et al., 2020a)[15]. Future work can investigate whether there is merit in attempting to adjust the remaining cluster members such that they represent some aspects of the theoretical shrinking cluster covariance matrix. Some possible aspects include the variances, the covariance linking the observed quantity to $\boldsymbol{\psi}$, and the N^{a}_{shrink} leading singular vectors[16].

[15] Proof: the rank of the post-deletion covariance matrix is smaller than the rank of the theoretical covariance matrix. As such, there is no linear operator that can transform the post-deletion covariance matrix to the theoretical covariance matrix.

[16] Note that attempting to sampling the leading singular vectors might be computationally expensive because the computational complexity of singular vector decomposition (SVD) is

7.7.4 Ensemble adjustments to reflect the posterior weights: the cluster resampling stage

The central idea of my resampling strategy

If members are deleted from one cluster, the other cluster expands to ensure that the total number of ensemble members is unchanged. To increase the expanding cluster's size from N_{pre} to N_{post}, the expanding cluster's ensemble members are resampled. The expanding cluster's sample mean state and sample covariance matrix should not be altered by resampling.

The computationally efficient resampling strategy I proposed in Chan et al. (2020a) is to resample within the extended state subspace spanned by the expanding cluster's ensemble members (henceforth referred to as the subspace resampling strategy). This is the easiest to formulate in terms of the perturbations of the expanding cluster's members. Supposing that the subscript "pre" denotes expanding cluster quantities before resampling, we can compute the pre-resampling perturbations $\left\{\boldsymbol{\psi}^{\boldsymbol{\alpha}'}{}_{\boldsymbol{n}} | n \in S_{\text{pre}}\right\}$ via

$$\boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}} \equiv \boldsymbol{\psi}^{\boldsymbol{\alpha}}_{\boldsymbol{n}} - \overline{\boldsymbol{\psi}^{\boldsymbol{\alpha}}_{\mathbf{pre}}} \quad \forall \ n \in S_{\text{pre}} \tag{7.30}$$

where $\overline{\boldsymbol{\psi}^{\boldsymbol{\alpha}}_{\mathbf{pre}}}$ is the expanding cluster's mean state and S_{pre} is the set of member indices in the expanding cluster before resampling.

The central idea of the subspace resampling strategy is to construct a new set of perturbations via linear combinations of the pre-resampling perturbations. I will denote all post-resampling expanding cluster quantities with the subscript "post". Let S_{post} denote the set of member indices in the post-resampling expanding cluster. S_{post} thus includes the member indices in S_{pre}

$\sim N_\psi^3$. Even if the SVD is executed locally (*i.e.*, $N_\psi \sim 10^4$), the computational complexity is $\sim 10^{12}$.

and the indices of the members deleted in the deletion stage. If I represent the set of post-resampling perturbation vectors as $\left\{\boldsymbol{\psi}^{\boldsymbol{\alpha}^*}_{\boldsymbol{n}^*} | n^* \in S_{\text{post}}\right\}$, the strategy's central idea can then be mathematically expressed as

$$\boldsymbol{\psi}^{\boldsymbol{\alpha}^*}_{\boldsymbol{n}^*} \equiv \sum_{n \in S_{\text{pre}}} \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}} \, T_{n,n^*} \qquad \forall \;\; n^* \in S_{\text{post}}$$

where T_{n,n^*} is a to-be-determined scalar factor controlling how the n-th pre-resampling perturbation contributes to the n^*-th post-resampling perturbation. This linear combination idea can be more succinctly expressed as

$$\boldsymbol{\Psi}_{\mathbf{post}} \equiv \boldsymbol{\Psi}_{\mathbf{pre}} \boldsymbol{T}. \tag{7.31}$$

Here, $\boldsymbol{\Psi}_{\mathbf{pre}}$ is a matrix where each column contains a pre-resampling perturbation, and $\boldsymbol{\Psi}_{\mathbf{post}}$ is a matrix where each column contains a post-resampling perturbation. Supposing the pre-resampling cluster size is denoted by N_{pre} and the post-resampling cluster size is denoted by N_{post}, then $\boldsymbol{\Psi}_{\mathbf{pre}}$ is an $N_\psi \times N_{\text{pre}}$ matrix and $\boldsymbol{\Psi}_{\mathbf{post}}$ is an $N_\psi \times N_{\text{post}}$ matrix. If we denote the ℓ-th member index in S_{pre} as $n_{\text{pre},\ell}$, and likewise for the ℓ-th member index in S_{post}, we can explicitly write out $\boldsymbol{\Psi}_{\mathbf{pre}}$ and $\boldsymbol{\Psi}_{\mathbf{post}}$:

$$\begin{aligned}
\boldsymbol{\Psi}_{\mathbf{pre}} &\equiv \left[\boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},1}} \quad \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},2}} \quad \cdots \quad \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}}}\right], \\
\boldsymbol{\Psi}_{\mathbf{post}} &\equiv \left[\boldsymbol{\psi}^{\boldsymbol{\alpha}^*}_{\boldsymbol{n}_{\mathbf{post},1}} \quad \boldsymbol{\psi}^{\boldsymbol{\alpha}^*}_{\boldsymbol{n}_{\mathbf{post},2}} \quad \cdots \quad \boldsymbol{\psi}^{\boldsymbol{\alpha}^*}_{\boldsymbol{n}_{\mathbf{post},\boldsymbol{N}_{\mathbf{post}}}}\right].
\end{aligned} \tag{7.32}$$

Finally, $\boldsymbol{T}$ is an $N_{\text{pre}} \times N_{\text{post}}$ matrix containing all of the T_{n,n^*} values [*i.e.*, element (n, n^*) of $\boldsymbol{T}$ is equal to T_{n,n^*}].

Heuristic constraints and arguments to formulate $\boldsymbol{T}$

$\boldsymbol{T}$ should be constructed such that the post-resampling perturbations have a covariance matrix equal to that of pre-resampling perturbations (covariance conservation condition) and have a mean of zero (zero sum condition). The first condition implies

$$\frac{1}{N_{\text{pre}}+N_{\text{new}}-1}\boldsymbol{\Psi}_{\mathbf{post}}\boldsymbol{\Psi}_{\mathbf{post}}{}^{\top}=\frac{1}{N_{\text{pre}}-1}\boldsymbol{\Psi}_{\mathbf{pre}}\boldsymbol{\Psi}_{\mathbf{pre}}{}^{\top}$$
$$\therefore \boldsymbol{\Psi}_{\mathbf{pre}}\boldsymbol{T}\boldsymbol{T}^{\top}\boldsymbol{\Psi}_{\mathbf{pre}}{}^{\top}=\frac{N_{\text{pre}}+N_{\text{new}}-1}{N_{\text{pre}}-1}\boldsymbol{\Psi}_{\mathbf{pre}}\boldsymbol{\Psi}_{\mathbf{pre}}{}^{\top}.$$

The covariance conservation condition is satisfied if

$$\boldsymbol{T}\boldsymbol{T}^{\top}=k^{2}\boldsymbol{I}_{\boldsymbol{N}_{\mathbf{pre}}} \tag{7.33}$$

where the symbols are as defined in Chapter 7.3.4.

The zero sum condition can be expressed as

$$\boldsymbol{\Psi}_{\mathbf{post}}\mathbf{1}_{\boldsymbol{N}_{\mathbf{post}}\times\mathbf{1}}=\mathbf{0}_{\boldsymbol{N}_{\psi}\times\mathbf{1}}$$

where the symbols are as defined in Chapter 7.3.4. The zero sum condition implies

$$\boldsymbol{\Psi}_{\mathbf{pre}}\boldsymbol{T}\mathbf{1}_{\boldsymbol{N}_{\mathbf{post}}\times\mathbf{1}}=\mathbf{0}_{\boldsymbol{N}_{\psi}\times\mathbf{1}}. \tag{7.34}$$

To proceed, consider that the sum of the pre-resampling perturbations is zero. In other words,

$$\boldsymbol{\Psi}_{\mathbf{pre}}c\,\mathbf{1}_{\boldsymbol{N}_{\mathbf{pre}}\times\mathbf{1}}=\mathbf{0}_{\boldsymbol{N}_{\psi}\times\mathbf{1}} \tag{7.35}$$

where c is any real number. Comparing Eq. (7.34) and Eq. (7.35) suggests

$$\boldsymbol{T}\mathbf{1}_{\boldsymbol{N}_{\mathbf{post}}\times\mathbf{1}}=c\,\mathbf{1}_{\boldsymbol{N}_{\mathbf{pre}}\times\mathbf{1}}.$$

In other words, if the elements in every row of $\boldsymbol{T}$ sum to c, the zero-sum property is fulfilled. For convenience, I have set $c = k$. Thus,

$$\boldsymbol{T}\,\mathbf{1}_{\boldsymbol{N_{post}}\times\mathbf{1}} = c\,\mathbf{1}_{\boldsymbol{N_{pre}}\times\mathbf{1}}. \tag{7.36}$$

While the covariance conservation condition [Eq. (7.33)] and the zero sum condition [Eq. (7.36)] provide constraints on the properties of $\boldsymbol{T}$, additional conditions are needed to completely constrain $\boldsymbol{T}$.[17] Heuristic arguments can be used to constrain our choice of $\boldsymbol{T}$. Firstly, it is desired that the pre-resampling expanding cluster perturbations are preserved as much as possible. However, as shown in appendix A of Chan et al. (2020a), it is impossible to preserve all the perturbations for all possible values of N_{new}. Thus, a second property is considered: if there are some perturbations that are difficult to preserve, I should at least preserve the directions of the perturbations.

A formulation of $\boldsymbol{T}$ and its general properties

One formulation of $\boldsymbol{T}$ that satisfies these properties is:

$$\boldsymbol{T} \equiv \begin{bmatrix} k\boldsymbol{I}_{\boldsymbol{N_{pre}}-\boldsymbol{N^*_{new}}} & \mathbf{0}_{\left(\boldsymbol{N_{pre}}-\boldsymbol{N^*_{new}}\right)\times\boldsymbol{N^*_{new}}} & \mathbf{0}_{\left(\boldsymbol{N_{pre}}-\boldsymbol{N^*_{new}}\right)\times\boldsymbol{N_{new}}} \\ \mathbf{0}_{\boldsymbol{N^*_{new}}\times\left(\boldsymbol{N_{pre}}-\boldsymbol{N^*_{new}}\right)} & \boldsymbol{I}_{\boldsymbol{N^*_{new}}} & \boldsymbol{E} \end{bmatrix}$$

where the various symbols are as defined in Chapter 7.3.4. Since $N^*_{\text{new}} < N_{\text{new}}$, $\boldsymbol{E}$ is a rectangular matrix with more columns than rows. I will construct matrix

[17] To see why additional conditions are needed, consider that $N_{\text{pre}}N_{\text{post}}$ constraints are needed to uniquely specify every element of $\boldsymbol{T}$. However, Eq. (7.33) only provides N^2_{pre} constraints and Eq. (7.36) only provides N_{pre} constraints. This means the covariance conservation and zero sum conditions only provide $N^2_{\text{pre}} + N_{\text{pre}}$ constraints. Since $N_{\text{post}} > N_{\text{pre}}$, if $N_{\text{pre}} > 2$ and $N_{\text{post}} > 3$, the total number of constraints is needed than those provided by the covariance conservation and zero sum conditions.

$\boldsymbol{E}$ shortly.

Note that whenever $N_{\text{new}} > N_{\text{pre}}$, the $k\boldsymbol{I}_{\boldsymbol{N}_{\text{pre}}-\boldsymbol{N}^*_{\text{new}}}$ component vanishes from $\boldsymbol{T}$. Furthermore, whenever $N_{\text{new}} = 1$, the $\boldsymbol{I}_{\boldsymbol{N}^*_{\text{new}}}$ and $\boldsymbol{E}$ components vanish from $\boldsymbol{T}$.

This formulation of $\boldsymbol{T}$ can be shown to satisfy the two desired properties. Performing the resampling by evaluating Eq. (7.31) yields

$$\boldsymbol{\Psi}_{\mathbf{post}} = \begin{bmatrix} \boldsymbol{\Psi}_{\mathbf{infl}} & \boldsymbol{\Psi}_{\mathbf{copy}} & \boldsymbol{\Psi}_{\mathbf{comb}} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{\Psi}_{\mathbf{infl}} &= \begin{bmatrix} k\boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},1}} & k\boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},2}} & \cdots & k\boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}-\boldsymbol{N}^*_{\mathbf{new}}}} \end{bmatrix} \\
\boldsymbol{\Psi}_{\mathbf{copy}} &= \begin{bmatrix} \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}-\boldsymbol{N}^*_{\mathbf{new}}+1}} & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}-\boldsymbol{N}^*_{\mathbf{new}}+2}} & \cdots & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}}} \end{bmatrix} \\
\boldsymbol{\Psi}_{\mathbf{comb}} &= \begin{bmatrix} \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}-\boldsymbol{N}^*_{\mathbf{new}}+1}} & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}-\boldsymbol{N}^*_{\mathbf{new}}+2}} & \cdots & \boldsymbol{\psi}^{\boldsymbol{\alpha}'}_{\boldsymbol{n}_{\mathbf{pre},\boldsymbol{N}_{\mathbf{pre}}}} \end{bmatrix} \boldsymbol{E}
\end{aligned}$$

The appearance of $\boldsymbol{\Psi}_{\mathbf{infl}}$ and $\boldsymbol{\Psi}_{\mathbf{copy}}$ in $\boldsymbol{\Psi}_{\mathbf{post}}$ indicates that all of the pre-resampling expanding cluster perturbation directions are preserved in the post-resampling expanding cluster. Out of N_{pre} of the post-resampling expanding cluster perturbations, $N_{\text{pre}} - N^*_{\text{new}}$ of them are inflated versions of pre-resampling expanding cluster perturbations ($\boldsymbol{\Psi}_{\mathbf{infl}}$), and the remaining N^*_{new} perturbations are preserved exactly by the transfer ($\boldsymbol{\Psi}_{\mathbf{copy}}$). The $\boldsymbol{\Psi}_{\mathbf{comb}}$ indicates that N_{new} perturbations were constructed by a linear combination of the N^*_{new} preserved perturbations mentioned earlier.

The choice of which perturbations to inflate and preserve is arbitrary. For simplicity, I have opted to inflate the members with the smallest $N_{\text{pre}} - N^*_{\text{new}}$ member indices.

Construction of the linear combination matrix $\boldsymbol{E}$

The matrix $\boldsymbol{E}$ can be constructed by considering that $\boldsymbol{T}$ satisfies the covariance conservation and zero sum conditions. Substituting the formulated $\boldsymbol{T}$ [Eq. (7.31)] into the covariance conservation condition [Eq. (7.33)] and doing some algebra yields

$$\boldsymbol{E}\,\boldsymbol{E}^{\top} = \frac{N_{\text{new}}}{N_{\text{pre}} - 1}\boldsymbol{I}_{\boldsymbol{N}^*_{\textbf{new}}}. \tag{7.37}$$

Furthermore, substituting the formulated $\boldsymbol{T}$ [Eq. (7.31)] into the zero sum condition [Eq. (7.36)] and doing some algebra yields

$$\boldsymbol{E}\,\boldsymbol{1}_{\boldsymbol{N}_{\textbf{new}}\times\boldsymbol{1}} = (k-1)\,\boldsymbol{1}_{\boldsymbol{N}^*_{\textbf{new}}\times\boldsymbol{1}}. \tag{7.38}$$

For $N_{\text{new}} \geq 2$, there is at least one $\boldsymbol{E}$ that satisfies Eq. (7.37) and Eq. (7.38).

To proceed, it is convenient to define

$$\boldsymbol{E}' \equiv \boldsymbol{E} - \overline{\boldsymbol{E}}$$

where

$$\overline{\boldsymbol{E}} \equiv \frac{k-1}{N_{\text{new}}}\boldsymbol{1}_{\boldsymbol{N}_{\textbf{new}}\times\boldsymbol{N}^*_{\textbf{new}}}. \tag{7.39}$$

As such, I can construct $\boldsymbol{E}$ if I have some means to construct $\boldsymbol{E}'$. To construct $\boldsymbol{E}'$, I will use Eq. (7.38) to determine something about $\boldsymbol{E}'$.

$$\boldsymbol{E}\,\boldsymbol{1}_{\boldsymbol{N}_{\textbf{new}}} = \left(\overline{\boldsymbol{E}} + \boldsymbol{E}'\right)\boldsymbol{1}_{\boldsymbol{N}_{\textbf{new}}} = (k-1)\,\boldsymbol{1}_{\boldsymbol{N}^*_{\textbf{pre}}\times\boldsymbol{1}} + \boldsymbol{E}'\,\boldsymbol{1}_{\boldsymbol{N}_{\textbf{new}}}.$$

Comparing the above expression with Eq. (7.38) yields

$$\boldsymbol{E}'\,\boldsymbol{1}_{\boldsymbol{N}_{\textbf{new}}\times\boldsymbol{1}} = \boldsymbol{0}_{\boldsymbol{N}^*_{\textbf{new}}\times\boldsymbol{1}}. \tag{7.40}$$

A second piece of information about $\boldsymbol{E}'$ can be obtained from Eq. (7.37):

$$\boldsymbol{E}\,\boldsymbol{E}^{\top} = \left(\overline{\boldsymbol{E}} + \boldsymbol{E}'\right)\left(\overline{\boldsymbol{E}} + \boldsymbol{E}'\right)^{\top} = \overline{\boldsymbol{E}}\,\overline{\boldsymbol{E}}^{\top} + \overline{\boldsymbol{E}}\,\boldsymbol{E}'^{\top} + \boldsymbol{E}'\,\overline{\boldsymbol{E}}^{\top} + \boldsymbol{E}'\,\boldsymbol{E}'^{\top}.$$

Considering Eq. (7.39) and Eq. (7.40), it can be shown that $\boldsymbol{E}'\,\overline{\boldsymbol{E}}^{\top}$ and $\overline{\boldsymbol{E}}\,\boldsymbol{E}'^{\top}$ matrices of zeros. Combining the above expression with Eq. (7.37) yield

$$\boldsymbol{E}\,\boldsymbol{E}^{\top} = \overline{\boldsymbol{E}}\,\overline{\boldsymbol{E}}^{\top} + \boldsymbol{E}'\,\boldsymbol{E}'^{\top} = \frac{N_{\text{new}}}{N_{\text{pre}} - 1}\boldsymbol{I}_{\boldsymbol{N}^*_{\text{new}}}.$$

As such, I have two useful conditions to construct $\boldsymbol{E}'$ with:

$$\boldsymbol{E}'\,\boldsymbol{E}'^{\top} = \frac{N_{\text{new}}}{N_{\text{pre}} - 1}\boldsymbol{I}_{\boldsymbol{N}^*_{\text{new}}} - \frac{(k-1)^2}{N_{\text{pre}}}\boldsymbol{1}_{\boldsymbol{N}^*_{\text{new}} \times \boldsymbol{N}^*_{\text{new}}} \tag{7.41}$$

and Eq. (7.40).

If I view each column of $\boldsymbol{E}'$ as a data sample, Eq. (7.40) and Eq. (7.41) can be viewed as equations related to the mean and covariance of the data samples. While I can construct $\boldsymbol{E}'$ via random draws from a normal distribution and apply appropriate linear transforms to construct $\boldsymbol{E}'$[18], I prefer using a deterministic approach for the ease of coding the BGEnKF. As such, I have opted to use

$$\boldsymbol{E}' \equiv \boldsymbol{L}_{\boldsymbol{E}}\left(\boldsymbol{L}_{\boldsymbol{W}}\right)^{-1}\boldsymbol{W} \tag{7.42}$$

where the various symbols are defined in Chapter 7.3.4. The resulting $\boldsymbol{E}'$ satisfies Eq. (7.40) and Eq. (7.41).

In summary, $\boldsymbol{E}$ is constructed by adding $\overline{\boldsymbol{E}}$ [Eq. (7.39)] to $\boldsymbol{E}'$ [Eq. (7.42)].

[18] This is random draw approach is used in Chan et al. (2020a).

7.7.5 The reason behind performing the EnSRF updates before cluster size adjustments

I opted to execute the EnSRF before the cluster size adjustments because the post-deletion shrinking cluster covariance matrix is more susceptible to sampling noise than that before the deletion. Furthermore, resampling the expanding cluster will not improve the expanding cluster covariance matrix as the resampling procedure is designed to preserve this matrix. As such, performing the EnSRF after cluster size adjustments is more susceptible to sampling errors than performing the EnSRF before the cluster size adjustments. I thus decided against performing the EnSRF after the cluster size adjustments.

Chapter 8

The BGEnKF VS the EnKF: idealized tests using WRF OSSEs

8.1 Overview

In the previous chapter, I discussed a computationally efficient BGEnKF algorithm to handle mixtures of clear and cloudy ensemble members. This chapter sets out to demonstrate the benefits of handling these mixtures via idealized tests using Observing System Simulation Experiments (OSSEs) involving a realistic weather model (WRF). Specifically, I will demonstrate that the BGEnKF generally outperforms the EnKF at assimilating synthetic infrared window channel brightness temperature (Window-BT) observations from a geostationary satellite.

The case used for these OSSEs is that of tropical convection over the equatorial Indian Ocean during the onset of the October 2011 Madden-Julian Oscillation [MJO; Madden and Julian (1971), Madden and Julian (1972)]. This case was chosen for several reasons. First, the general features of its tropical convection can be replicated in regional WRF models (Wang et al., 2015). Furthermore, this case was observed by the Dynamics of the MJO (DYNAMO)

field campaign (Zhang et al., 2013; Zhang and Yoneyama, 2017) and has been explored in multiple simulation studies (Zhang et al., 2017; Ying and Zhang, 2017; Fu et al., 2017; Chen et al., 2018c; Chen and Zhang, 2019). This case has also been used in earlier studies to examine the potential impacts of assimilating IR-BTs on the analyses and forecasts of tropical convection (Ying and Zhang, 2018; Chan et al., 2020b). In other words, this October 2011 case over the equatorial Indian Ocean can be reasonably replicated in simulations and is reasonably established. This case is thus suitable for our investigation.

8.2 Materials and methods

8.2.1 Description of October 2011 tropical convection case

I chose to use a three-day period during the onset of the October 2011 MJO event (15 October 2011 to 18 October 2011). My study domain is indicated by the red box in Figure 8.1(a), and is identical to that of Chan et al. (2020b). This domain is mostly over ocean, with a small amount of land in the north (southern tip of India and Sri Lanka).

Over this three-day period, two persistent regions of enhanced convection (henceforth, "convective regions") were observed in the 4-km Global IR Dataset of Janowiak et al. (2001). Said dataset is currently jointly produced and maintained by the National Centers for Environmental Prediction (NCEP), Climate Prediction Center (CPC) and the National Weather Service (NWS). To visualize these convective regions, I first converted the 3D longitude-latitude-time array of window channel brightness temperatures (Window-BT) from the 4-km Global IR Dataset (henceforth, MERG dataset) into a 2D longitude-time array by averaging between 10 °S and 10 °N. The 2D longitude-time array of Window-BT values were then converted into a 2D longitude-time array of

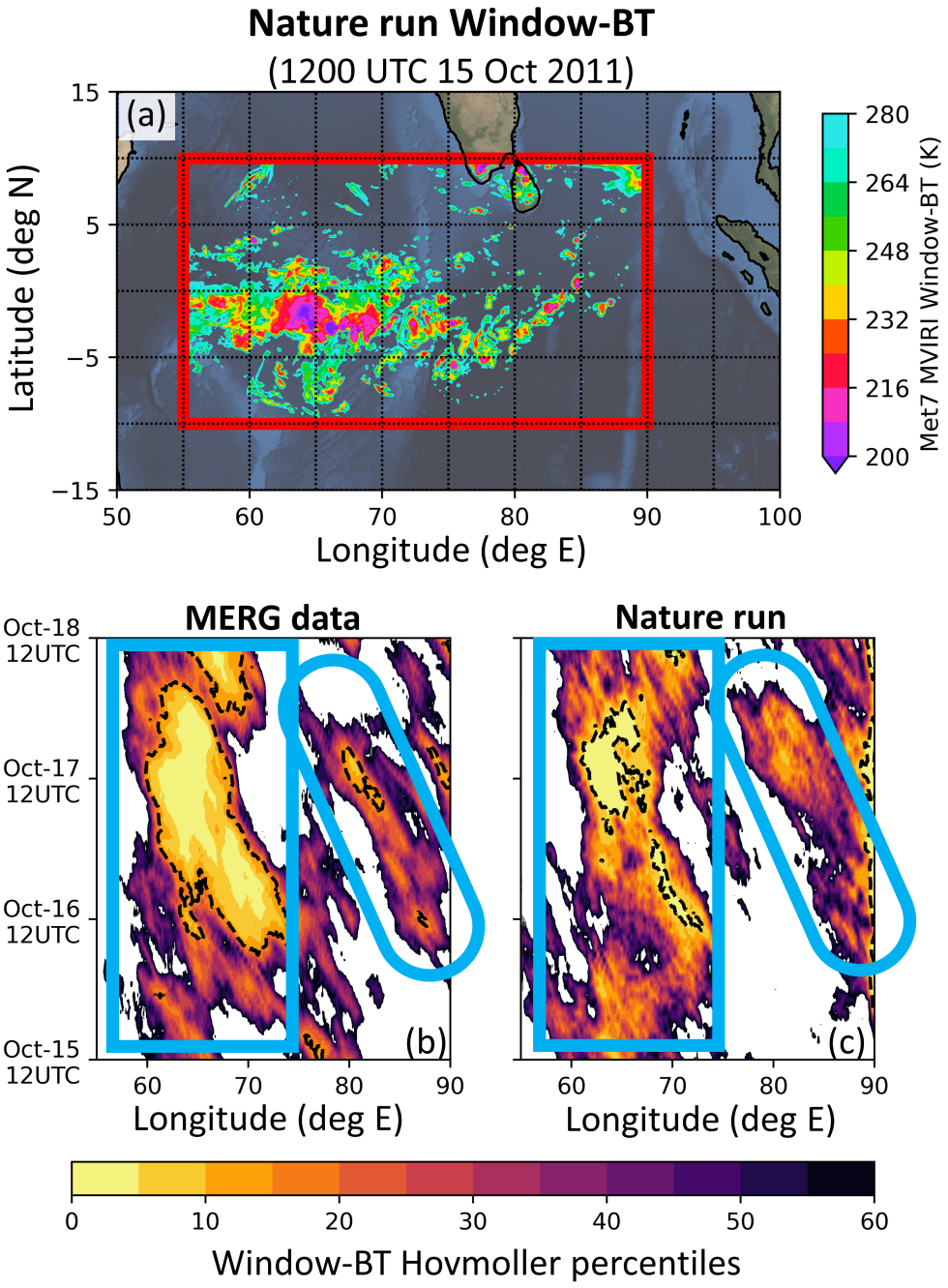


Figure 8.1: *See next page for caption*

Figure 8.1 *(previous page)*: (a) Plot of my OSSE domain overlaid with the nature run's simulated Window-BT field at 1200 UTC on 15th October 2011. The red box in panel (a) indicates my study domain. Also shown are longitude-time diagrams for the MERG dataset (b) and nature run (c). In panels (b) and (c), the shadings indicate Window-BT Hovmoller percentile values. These Window-BT Hovmoller percentile values are constructed by first averaging Window-BT values between between 10°S and 10°N at every hmy to produce a time-longitude array of latitudinally-averaged Window-BT values. Said arrays are then converted into percentiles before producing the longitude-time percentile values. Note that the dashed black contmys in (b) and (c) indicate areas where the time-longitude arrays of latitudinally-averaged Window-BT values are below 260 K. The features highlighted in the blue rectangle and the blue oval are discussed in the text.

percentiles and plotted the percentile data in Figure 8.1(b). This percentile processing is necessary to identify these two convective regions in the nature run later.

The two observed convective regions are indicated by a blue rectangle and a blue oval in Figure 8.1(b). The first system (blue rectangle) occurred between 60 °E and 75 °E and persisted beyond the three-day period. Westward propagation was observed in some of the clouds in this first system, most notably between 1200 UTC on 16 October and 0000 UTC on 18 October. The second system (blue oval) appeared on the eastern edge of the study domain at 1200 UTC on 16th October and exhibited a westward propagation that is similar to that of the first system. I will later assess my OSSE's nature run simulation by checking the nature run against these two convective regions.

8.2.2 Setup of WRF ensemble and nature run

The setup of the WRF model and the construction of the 51-member ensemble initial conditions are as described in Chapter 6.2. One of the fifty members generated from TIGGE will be used to construct the nature run. Specifically, I desire a nature run that is roughly one ensemble standard deviation from

the experiments' ensembles. To select an appropriate initial condition file for such a nature run, I first integrate the 51 members forward for 12 hours (from 0000 UTC to 1200 UTC on 15 October 2011). This integration is performed to generate flow-dependent ensemble statistics that are consistent with the WRF model. After the 12-hour integration, I compute the following perturbation length metric (D^2) for each of the 51 ensemble members

$$D^2(n) \equiv \frac{1}{N_S\, N_i\, N_j} \sum_{v \in S} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \left(\frac{\Lambda(i,j,v,n) - \langle \Lambda(i,j,v) \rangle_n}{\sigma_{i,j,v}} \right)^2 . \tag{8.1}$$

$\Lambda(i,j,v,n)$ here is the value of a WRF-derived field v at horizontal index location (i,j) for ensemble member n. Furthermore, $\langle \Lambda(i,j,v) \rangle_n$ is the ensemble average of $\Lambda(i,j,v,n)$, and $\sigma_{i,j,v}$ is the ensemble standard deviation of $\Lambda(i,j,v,n)$. This means that the expression in the parentheses of Eq. (8.1) is the spread-normalized displacement of ensemble member n from the ensemble mean at location (i,j) for variable field v. The set S contains three 2D variables (precipitable water, column mass, and mass-integrated kinetic energy) and N_s is the size of the set S (*i.e.*, $N_S = 3$). Furthermore, N_i ($\equiv 432$) is the number of east-west grid points and N_j ($\equiv 243$) is the number of north-south grid points. The metric in Eq. (8.1) can thus be interpreted as being proportional to the spread-normalized Euclidean length of the n-th ensemble perturbation. As such, a D^2 value of unity means that the ensemble member is generally displaced from the ensemble mean by 1 standard deviation.

I define our nature run member to be the member whose D^2 value is closest to unity at 1200 UTC on 15 October. As a result, the nature run is based on member 10 of the TIGGE ensemble. The remaining 50 WRF members will be used for my cycling OSSE DA experiments.

8.2.3 Sanity check of nature run

Before proceeding, I checked the nature run by comparing it against the MERG data. Figure 8.1(b & c) shows longitude-time diagrams of the Window-BT percentiles from the MERG dataset and my nature run. The construction of these percentiles is explained in Chapter 8.2.1 and in the caption of Figure 8.1.

I have opted to display the Window-BT percentiles over the Window-BT values because the WRF model tends to under produce clouds (*i.e.*, when compared to satellite observations, the nature run Window-BTs are warm biased). This is illustrated by the dashed contmys in Figure 8.1(b & c), which highlights areas where the latitudinally-averaged values of Window-BT were cooler than 260 K. These areas are substantially larger in the MERG data than in the nature run, meaning that the nature run under produced clouds. When I plotted the latitudinally-averaged Window-BT data from MERG and the nature run using the same Window-BT color range, it was difficult to visually identify persistent convective features in both datasets because of said cloud biases in the nature run (not shown). Since converting the Window-BT values to percentile values weakens the visual interference from the cloud biases, I opted to display the Window-BT percentiles over the Window-BT values.

Figure 8.1(c) indicates that the nature run also exhibits the two persistent convective features observed in the MERG dataset (see Chapter 8.2.1). These persistent convective features are indicated by the blue rectangle and blue oval in Figure 8.1(c). Not only did the nature run's two persistent convective features occur in locations and times similar to those of the MERG dataset (Figure 8.1(b)), these nature run features also exhibited westward propagation patterns similar to those of the MERG dataset. As such, the nature run simulation reasonably replicated the anomalous convective behavior of the real atmosphere between 15 October to 18 October 2011. This nature run was thus used for my OSSE tests.

8.2.4 Setup of DA experiments to test the BGEnKF

To test the BGEnKF, three 50-member ensemble experiments were conducted. All three experiments started at 1200 UTC on 15 October and terminated at 1200 UTC on 18 October, with hourly DA cycling (73 cycles in total).

In the first experiment, no observations were assimilated, meaning that the ensemble was effectively allowed to integrate freely (henceforth, NoDA experiment). The NoDA experiment serves as a baseline for comparing the performance of the EnKF and BGEnKF. In most of the discussions about the OSSE experiments later, the NoDA will mainly be used to construct normalized measures of errors in other two DA experiment, and to measure imbalances induced by DA.

The other two experiments are the EnKF and BGEnKF experiments. As the names suggest, the only difference between the EnKF and BGEnKF experiments is in the DA algorithm employed. The EnKF experiment will assimilate observations using the PSU-EnKF's (Meng and Zhang, 2007, 2008) default EnKF algorithm. In contrast to the EnKF experiment, the BGEnKF experiment will assimilate observations using the newly implemented BGEnKF algorithm (see Chapter 7 for the details).

As a first approach to testing the BGEnKF, only synthetic Window-BT observations will be assimilated in both DA experiments. These synthetic observations are constructed by first running the Community Radiative Transfer Model (CRTM) release 2.3.0 on the nature run (see Chapters 8.2.2 and 8.2.3). The nature run's Window-BT values were then thinned to a horizontal spacing of 27-km (∼11,500 observations per DA cycle). White noise with a standard deviation of 3 K was then added to the thinned nature run Window-BT

values to simulate instrument noise, thus constructing the synthetic observations. Future work can investigate if our findings can be extended to situations where an entire suite of operationally-assimilated observations and observations from different infrared channels are assimilated.

The heuristic strategies described in Chapter 3.5 are employed to assimilate the synthetic Window-BT observations. Specficially, a 100-km horizontal radius of influence (Houtekamer and Mitchell, 2001; Greybush et al., 2011; Houtekamer and Zhang, 2016) and no vertical localization was employed to assimilate the Window-BT observaitons. RTPP with an 80% weight on the forecast perturbations was also used.

Aside from these common heuristic strategies, I also restricted the BGEnKF/EnKF from updating the domain-averaged specific humidity (QVAPOR) using Window-BT observations. This measure was necessary because both the BGEnKF and the EnKF experienced filter divergence that are likely related to DA-induced dry biases. This divergence occurred after 48 hours of cycling. To test postulated cause of the filter divergence, I replaced the 3D posterior mean QVAPOR field ($\overline{q_v^a}$) with the following modified mean QVAPOR field ($\overline{q_v^*}$):

$$\overline{q_v^*(i,j,k)} \equiv \overline{q^a(i,j,k)} - \frac{1}{N_i\, N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \left\{ \overline{q_v^a(i,j,k)} - \overline{q_v^f(i,j,k)} \right\}. \tag{8.2}$$

Here, (i,j,k) refer to the west-east, south-north and bottom-top indices of the 3D QVAPOR fields and $\overline{q_v^f}$ refers to the 3D prior mean QVAPOR field. Employing Eq. (8.2) effectively nullifies the BGEnKF/EnKF's analysis increment to the domain-averaged QVAPOR. Upon employing Eq. (8.2), the filter divergence disappeared from the BGEnKF and EnKF experiments. This disappearance supports the notion that the DA-induced dry biases are the likely cause of the filter divergence. As such, this bias-nullifying measure will be used in both the BGEnKF and EnKF experiments discussed in this publication. Though beyond

the scope of my current study, the origin of this deleterious DA-induced dry bias warrants further investigation.

8.2.5 Execution wall-time: BGEnKF VS EnKF

Before proceeding, a comparison of the execution wall-times of the BGEnKF and the EnKF is warranted. The BGEnKF algorithm took ∼30 seconds to assimilate ∼11,500 observations using 228 Intel Knight's Landing computer cores [distributed across 7 computational nodes on the National Energy Research Scientific Computing Center (NERSC) Cori supercomputer; each core has a clock rate of 1.4 GHz]. Assimilating the same observations via an EnKF algorithm took ∼20 seconds of wall-time. For a fair comparison, this EnKF algorithm used the exact same code structure and computing resources, but with the cluster transfer and auxiliary variable update steps disabled. In other words, the BGEnKF used ∼10 seconds more wall-time than the EnKF.

This ∼10-second difference should be assessed in the context of the wall-time for the entire PSU-EnKF executable. The other components of the PSU-EnKF took ∼100 seconds to execute. As such, the BGEnKF only added ∼10% wall-time to the entire PSU-EnKF executable. The BGEnKF algorithm is thus likely affordable for research and operational groups that are already running serially-assimilating EnKFs [*e.g.*, Anderson et al. (2009)].

8.3 Results and discussion

8.3.1 Note on validation metrics

In the discussions to follow, for the ease of visualizing the performances of the EnKF and BGEnKF experiments as functions of time and model level, I will be showing plots of normalized root-mean-square errors (nRMSEs) and

normalized biases as functions of time and model level. Both quantities are normalized using the root-mean-square errors (RMSEs) of the NoDA experiment. The EnKF experiment's nRMSE at model level k and date t is defined as

$$\text{EnKF nRMSE}(k,t) \equiv \frac{\text{EnKF RMSE}(k,t)}{\text{NoDA RMSE}(k,t)} \tag{8.3}$$

and likewise for that of the BGEnKF and NoDA experiments (the NoDA's nRMSE values are always 1). Note that if a filter results in nRMSEs > 1.0, the assimilation of Window-BT via said filter degraded the ensemble with respect to the NoDA experiment (the converse is true for nRMSEs < 1.0). I also define the normalized bias of the EnKF experiment to be

$$\text{EnKF normalized bias}(k,t) \equiv \frac{\text{EnKF bias}(k,t)}{\text{NoDA RMSE}(k,t)}, \tag{8.4}$$

and likewise for the BGEnKF and NoDA experiments. Note that all biases here are computed by subtracting the nature run fields from the forecast ensemble mean fields.

The nRMSEs and normalized biases were examined for six variable fields: the zonal wind velocity component field (U), the meridional wind velocity component field (V), the temperature field (T), the QVAPOR field (Q), the Window-BT field, and the upper tropospheric infrared water vapor channel field (WV-BT; central wavelength of 6.2 microns). The nRMSEs are plotted in Figures 8.2 and 8.4(a & b) and the normalized biases are plotted in Figures 8.3 and 8.4(c & d). All quantities are computed using forecast ensembles. Two common features can be inferred from these nRMSE and normalized bias plots: 1) the BGEnKF experiment has generally better or similar nRMSEs and milder or similar normalized biases than the EnKF experiment, and 2) the spatiotemporal variations in the nRMSEs and normalized biases of the BGEnKF and EnKF experiments are quite similar.

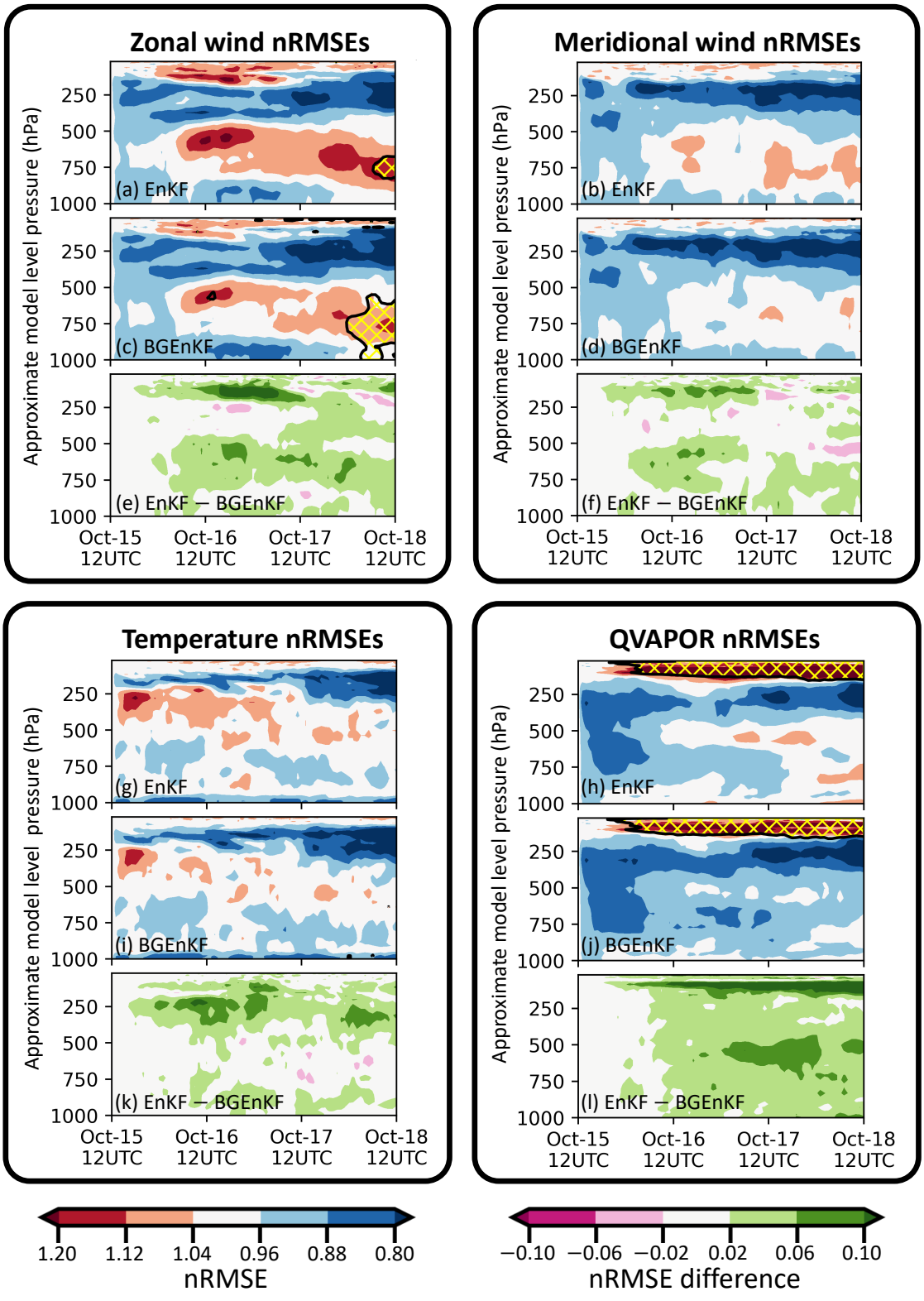


Figure 8.2: *See next page for caption*

Figure 8.2 *(previous page)*: Plots of various prior ensemble statistics as functions of time and model level. For ease of interpretation, the model levels are displayed in terms of their approximate pressure levels (estimated using the definition of eta levels in WRF and assuming a surface pressure of 1000 hPa). The shadings indicate the NoDA-normalized RMSEs [nRMSEs; defined in Eq. (8.3)] for the EnKF (a, b, g & h) and BGEnKF (c, d, i & j) experiments, as well as the nRMSE differences between the EnKF and BGEnKF experiments (e, f, k & l). The nRMSEs and nRMSE differences are shown for the U field (a, c & e), V field (b, d & f), T field (g, i & k), and Q field (h, j & l). The areas outlined with a black contmy and filled with yellow hatching have consistency ratios (spread/error) less than 0.75. Note that the statistics displayed here are based on forecast ensembles.

8.3.2 On differences in the BGEnKF's and the EnKF's performances during DA cycling

I will begin with describing and discussing the first common feature: the nRMSEs and normalized biases of the BGEnKF experiment tend to be either better than or comparable to those of the EnKF experiment (Figures 8.2 to 8.4). The BGEnKF's lower nRMSEs are apparent from the nRMSE difference plots of the U, V, T and Q fields – subtracting the BGEnKF's nRMSEs from the EnKF's nRMSEs generally results in positive values (green shadings in Figure 8.2(e, f, k & l). The BGEnKF experiment also has better WV-BT nRMSEs than the EnKF experiment (Figure 8.4(b)). In the case of the normalized biases, the BGEnKF experiment has noticeably milder biases than the EnKF experiment in terms of the 100 hPa U field (Figure 8.3(c & e)), the 400–100 hPa T field (Figure 8.3(i & k)), the Window-BT field (Figure 8.4(e)), and WV-BT field (Figure 8.4(f)). Otherwise, the BGEnKF and EnKF experiments have similar bias values. These results suggest that the BGEnKF is more suitable for assimilating all-sky Window-BT than the EnKF.

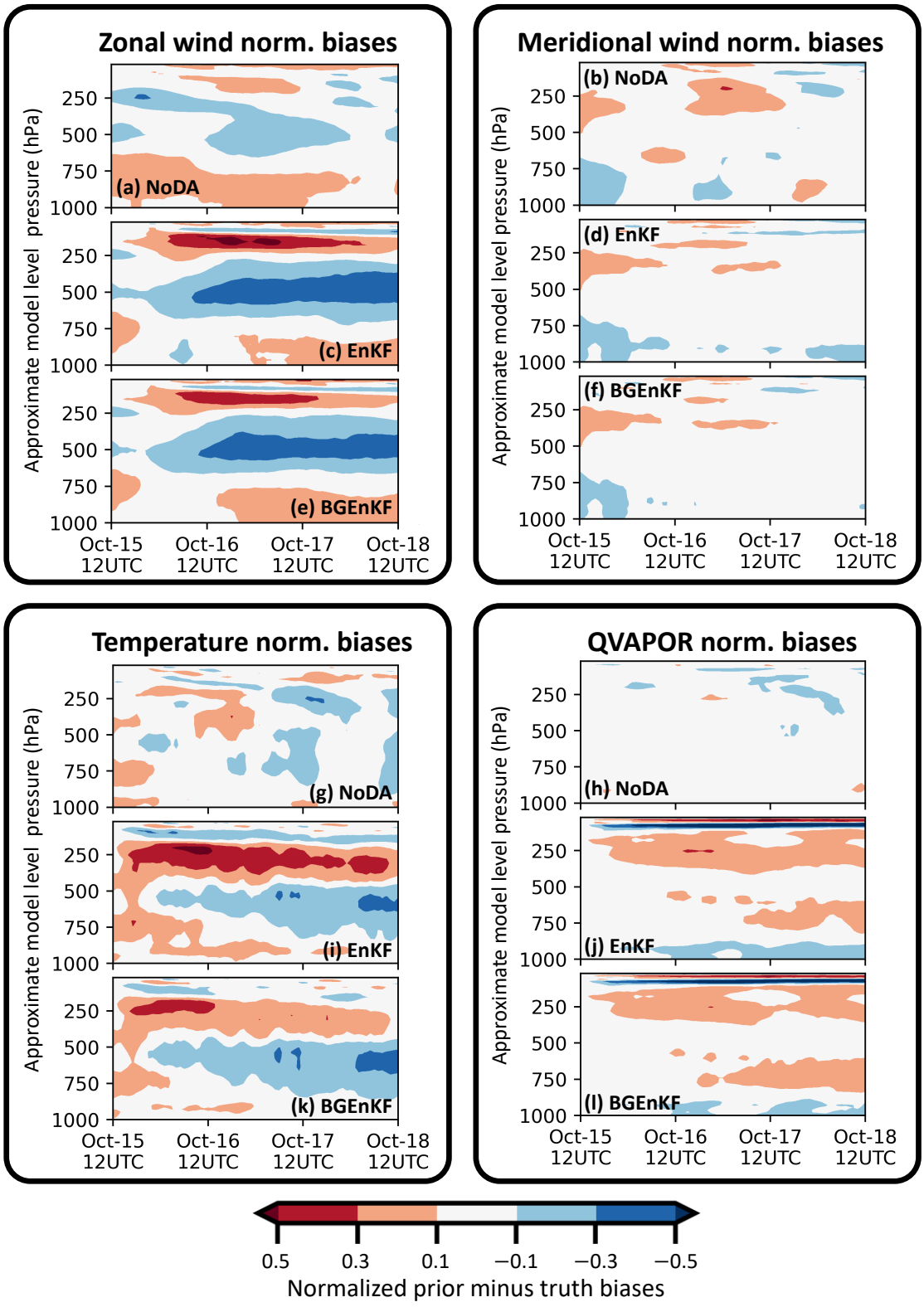


Figure 8.3: *See next page for caption*

Figure 8.3 *(previous page)*: Plots of various prior ensemble normalized biases as functions of time and model level. These normalized biases are displayed for the U field (a, c & e), V field (b, d & f), T field (g, i & k), and Q field (h, j & l), for the NoDA (a, b, g & h), EnKF (c, d, i & j) and BGEnKF (e, f, k & l) experiments. Similar to Figure 8.2, the model levels are displayed in terms of approximate pressure levels. See the Eq. (8.4) for the definition of the normalized biases.

The BGEnKF's performance advantages over the EnKF can be separated into two types. In the first type, the assimilation of Window-BT observations via the EnKF results in improvements relative to the NoDA experiment (*e.g.*, EnKF nRMSEs < NoDA nRMSEs), and replacing the EnKF with the BGEnKF caused greater improvements (henceforth termed the "greater improvement type of performance advantage"). According to Figures 8.2 and 8.4, this type of performance advantage occurs in multiple places: 1) the 800 hPa to 1000 hPa U field nRMSEs during the first 56 cycles, 2) the 100 hPa to 500 hPa U field nRMSEs during the last 36 DA cycles, 3) the near surface and ∼250 hPa V field nRMSEs from 0000 UTC on 16th October to 0000 UTC on 17th October, 4) between 100 hPa to 300 hPa in the T field nRMSEs for most cycles, 5) between 250 to 600 hPa in the Q field nRMSEs for most cycles, and in the WV-BT nRMSEs for most DA cycles after 0000 UTC on 16th October. The BGEnKF experiment also has smaller biases in the Window-BT and WV-BT fields than the EnKF experiment (Figure 8.4). This type of performance advantage over the EnKF is likely due to the BGEnKF's ability to handle mixture statistics. As such, the BGEnKF might be more suitable for assimilating Window-BT observations than the EnKF because the BGEnKF can create greater improvements than the EnKF.

The BGEnKF experiment's second type of performance advantage over the EnKF experiment is where Window-BT DA via the EnKF degraded the ensemble relative to the NoDA experiment (*e.g.*, EnKF nRMSEs > NoDA nRMSEs), and replacing the EnKF with the BGEnKF resulted in milder degradation (henceforth, the "milder degradation type of performance advantage"). In terms of nRMSEs

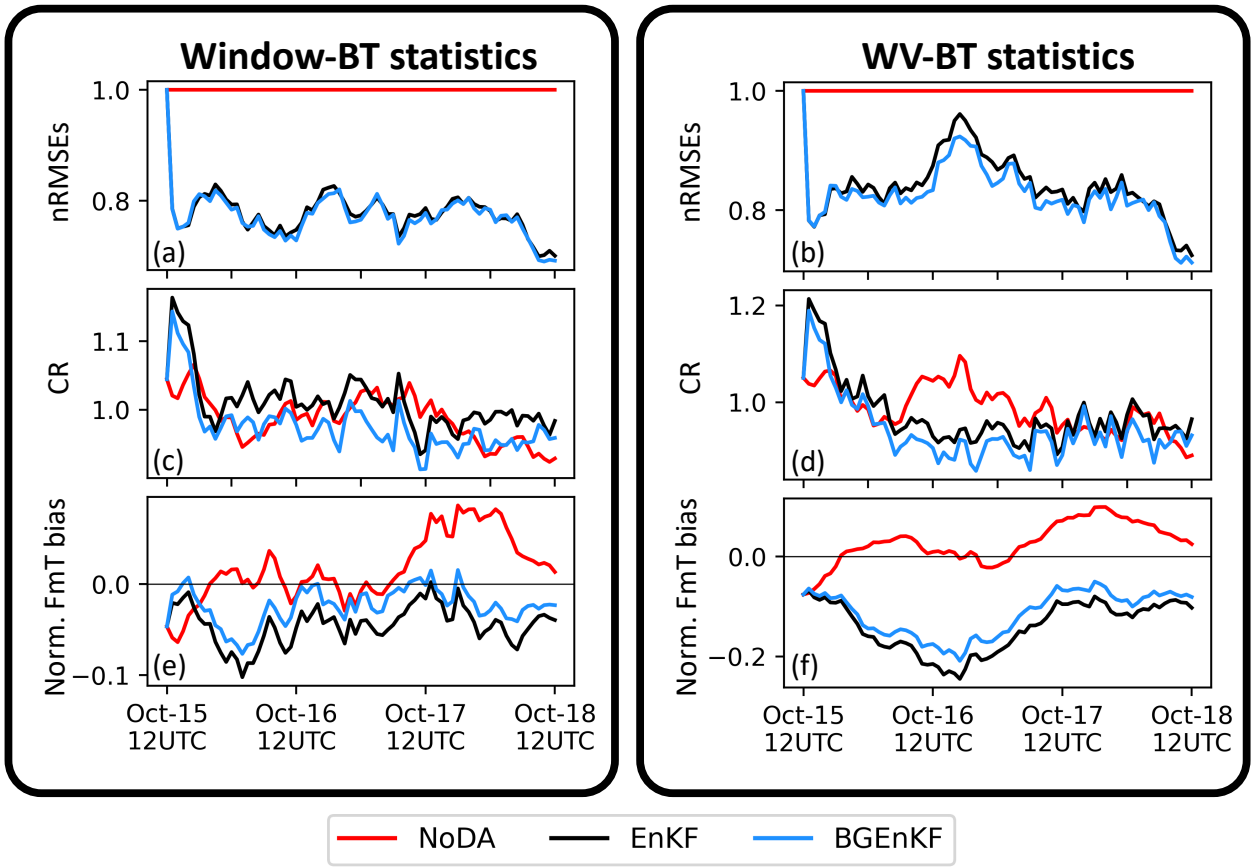


Figure 8.4: Time-series showing the performance statistics of the three experiments' prior ensembles in terms of Window-BT (a, c & e) and WV-BT (b, d & f). The definitions of nRMSEs (a & b) and normalized prior minus truth (Norm. FmT bias; e & f) are the same as in Figures 5 to 8. Like Figures 5 and 6, the consistency ratio (CR; c & d) here is defined as the ratio of spread to error.

(Figure 8.2), such situations are noticeable at the 100 hPa tropopause level and 500–700 hPa levels for the U and V fields , at the 200–500 hPa model levels for the T field, and at the 100 hPa level for the Q field. Such situations are also noticeable in the normalized biases of the ∼100 hPa U field and the 100–400 hPa T field (Figure 8.3). The likely origin of these DA-induced degradations will be discussed in section 8.3.3.

This milder degradation type of performance advantage might originate from a combination two factors. As in the case of the greater improvement type of performance advantage, because the BGEnKF is better at handling mixture statistics than the EnKF, the BGEnKF should generally result in milder performance degradations than the EnKF in situations where the prior ensemble is mixed. Aside from that, because the BGEnKF tends to result in smaller consistency ratios (CRs; spread-to-error ratio; partially illustrated in this paper), degrading analysis increments from the BGEnKF are likely smaller than those of the EnKF. Some evidence of this can be seen in Figure 8.2(a & c) – compared to the EnKF, the BGEnKF have larger areas where the U field CRs are less than 0.75. This too results in the BGEnKF having milder performance degradation than the EnKF. Since both factors can lead to the same type of performance advantage, it is difficult to gauge the relative contributions of these factors towards said performance advantage. These relative contributions deserve investigation in future work.

The BGEnKF tends to result in smaller CRs than the EnKF because the BGEnKF can outright convert all clear member columns to cloudy member columns, or *vice versa*. Since clear and cloudy member columns are very different, having both types of columns present at the same time boosts the ensemble spread. If all clear member columns are converted to cloudy member columns, or *vice versa*, large perturbations relative to the ensemble mean are thus replaced with smaller perturbations. This replacement results in reduced ensemble dispersion. Since the EnKF lacks this mechanism of ensemble spread removal, the BGEnKF tends to remove more ensemble spread than the EnKF. The BGEnKF can thus result in smaller CRs than the EnKF. Future work can investigate if a larger RTPP coefficient would be more appropriate for the BGEnKF.

Note that while the BGEnKF generally had lower nRMSEs than the EnKF, there are occasional transient situations where the opposite happens. For in-

stance, at around 0000 UTC on 17th October the BGEnKF's U nRMSEs were slightly higher than the EnKF at 250 hPa (Figure 8.2(e)). Another example would be T nRMSEs at around 1200 UTC on 17th October (Figure 8.2). Nonetheless, if I integrated the forecast ensembles' nRMSEs with respect to pressure at every cycle, the resulting mass-weighted nRMSEs of the BGEnKF experiment will be lower than that of the EnKF experiment.

Before proceeding, note that I have also examined day-long deterministic forecasts initialized from the analysis means of the EnKF and BGEnKF experiments (not shown here). The BGEnKF experiment's RMSE performance advantage over the EnKF experiment persists for up to 9 hours of lead time in terms of the U, V and T fields. In terms of the 500–800 hPa Q field RMSEs, the BGEnKF experiment's RMSE advantage over the EnKF experiment persisted throughout the 24 hours of integration. These results are as expected since the BGEnKF experiment has lower RMSEs than the EnKF experiment during DA cycling.

8.3.3 On the similar patterns observed in the performances of the BGEnKF and EnKF experiments

Another common feature in Figures 8.2 to 8.4 is that the spatiotemporal patterns in nRMSEs and normalized biases of the EnKF and BGEnKF experiments are similar. For instance, relative to the NoDA experiment (nRMSEs = 1), Window-BT DA with either algorithm tends to degrade the 500–800 hPa U nRMSEs, and improve the 100–500 hPa U nRMSEs (Figure 8.2(a & c)). The similarities in these spatiotemporal patterns suggest that there is a strong similarity in way the EnKF and BGEnKF algorithms are employed in the experiments.

The similarities in the spatiotemporal patterns of the BGEnKF's and EnKF's nRMSEs and normalized biases are likely caused by the frequent activation of the BGEnKF's single kernel mode. The single kernel mode is triggered when the ensemble failed to clear any of the heuristic checks described in Chapter

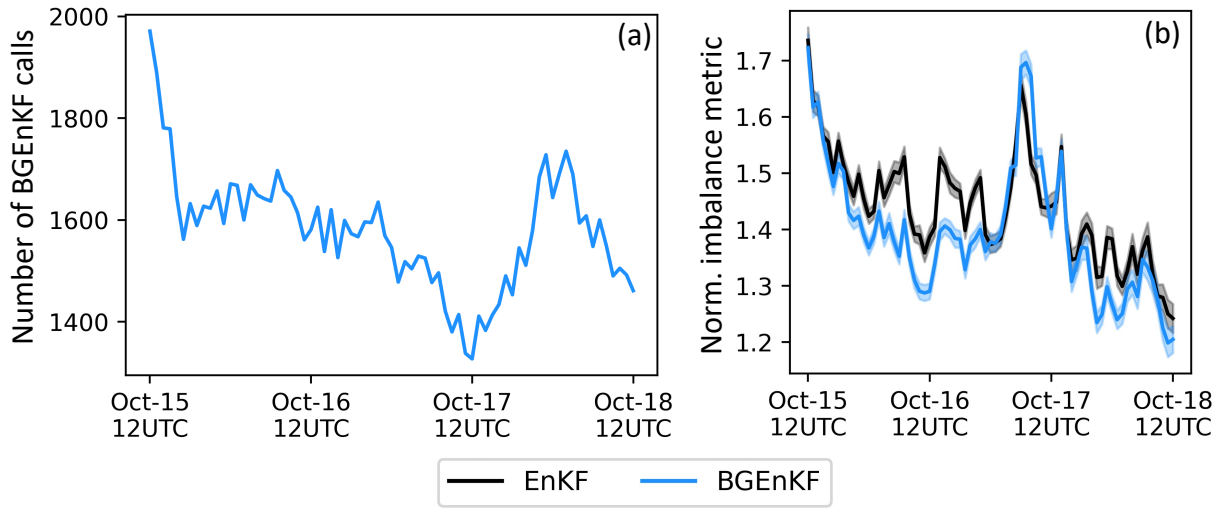


Figure 8.5: Plots showing the frequencies at which the two kernel BGEnKF update procedure is called in the BGEnKF experiment (a), and the normalized imbalance metric statistics for both the BGEnKF and EnKF experiments (b). For reference, 11502 IR observations are assimilated at each DA cycle. The normalized imbalance metric is defined in the text. The solid curves in (b) indicate the ensemble average of every member's normalized imbalance metric and the half-width of the shadings in (b) indicate twice the standard error of the members normalized imbalance metric.

7.5. As can be seen in Figure 8.5(a), the two kernel mode of the BGEnKF algorithm is only called to assimilate ∼10% of the Window-BT observations. The remaining ∼90% of observations assimilated in the BGEnKF experiments are assimilated using the single kernel mode. Since the single kernel mode of the BGEnKF is mathematically identical to the EnKF algorithm, the single kernel mode's high frequency of activation (∼90%) explains the strong similarities in nRMSE and normalized biases of the two DA experiments.

Note that the BGEnKF experiment noticeably outperformed the EnKF experiment despite the fact that the two kernel mode of the BGEnKF is infrequently activated. For instance, according to Figure 8.2(h, j & l), for the 24 cycles on 17th October and between 500 hPa to 700 hPa, the BGEnKF experiment had 0.06–0.1 less Q nRMSEs than the EnKF experiment. Since the EnKF experi-

ment had Q nRMSEs of ∼1 then, the BGEnKF was able to introduce a ∼6–10% improvement over the EnKF. These are considerable improvements since the BGEnKF was only called on ∼10% of the Window-BT observations.

It is also interesting to discuss the origins of the nRMSE and normalized bias degradations seen in the EnKF and BGEnKF experiments relative to the NoDA experiment (*e.g.*, EnKF nRMSEs > 1). Such degradations are observed in the nRMSEs of U, V and T fields. Since these two experiments have worse-than-NoDA nRMSEs and biases at similar model levels and times (Figures 8.2 and 8.3), these two experiments likely have similar sources of degradation. Considering that the EnKF is effectively employed on ∼90% of the Window-BT observations in the BGEnKF experiment, these common degradations are likely caused by the EnKF algorithm.

There are three plausible ways that the EnKF can contribute towards the worse-than-NoDA nRMSEs: 1) non-Gaussianity in the forecast ensemble, 2) sampling errors, and 3) biases introduced by the assimilation of Window-BT. The first factor essentially originates from the fact that clear extended state vectors have different statistics from cloudy extended state vectors. Sampling errors can also introduce errors into the analysis, particularly over regions where the ensemble correlations are weak. This factor is likely present in my experiments because no vertical localization is currently. Finally, since biases are a component of RMSEs (*e.g.*, Ying and Zhang (2017), Ying and Zhang (2018), and Chan et al. (2020b)), biases introduced by Window-BT DA can contribute towards worse-than-NoDA RMSEs. While the contribution of biases to worse-than-NoDA RMSEs can be easily inferred (see the next paragraphs), the first two factors cannot be easily teased apart.

To understand the contribution of biases to the occurrence of worse-than-NoDA RMSEs (*i.e.*, nRMSEs > 1), I computed the following fraction as a function

of model level and time ($f_{\text{bias}}(k,t)$). For the EnKF experiment, I defined

$$\text{EnKF's } f_{\text{bias}}(k,t) \equiv \sqrt{\frac{[\text{EnKF's biases}(k,t)]^2 - [\text{NoDA's biases}(k,t)]^2}{[\text{EnKF's RMSEs}(k,t)]^2 - [\text{NoDA's RMSEs}(k,t)]^2}}$$

and likewise for the BGEnKF experiment. f_{bias} can be interpreted as the fractional contribution of biases to the worse-than-NoDA RMSE performance.

I found that for about 25–45% of the worse-than-NoDA situations (nRMSEs > 1) in the U and T fields, majority of the nRMSE degradation (*i.e.*, $f_{\text{bias}} \geq 0.6$) can be explained by the the introduction of biases. Mathematically,

$$p(f_{\text{bias}} > 0.6 \,|\text{nRMSE} > 1) \in (0.25, 0.45).$$

As such, though DA-induced biases are important contributors towards the worse-than-NoDA RMSEs of either DA filters, the net contribution coming from other factors is also important. Future work should examine separating and quantifying the relative importance of these three factors towards the worse-than-NoDA RMSEs.

8.3.4 On the origin of biases in the EnKF and BGEnKF experiments

The U, T, Q, Window-BT and WV-BT biases introduced by Window-BT DA are also interesting to examine. Since the Q analysis increments were subject to bias removal (see last paragraph of Chapter 8.2.4), I will discuss the Q biases separately from the other biases. The introduced U and T biases are likely caused by persistently cold forecast minus truth (FmT) Window-BT biases in both the BGEnKF and EnKF experiments (Figure 8.4(e)). Furthermore, the cold biased FmT Window-BT values indicate an over preponderance of clouds. Since WV-BT values are cooler in cloudy situations than clear situations, the cold WV-

BT FmT biases are likely also due to the over-cloudiness of the forecast ensemble.

To understand the origin of the cold FmT Window-BT biases, I examined the analysis ensembles' Window-BT biases. Upon running the CRTM on the analysis ensembles of the EnKF and BGEnKF experiments, I found that the normalized analysis minus truth (AmT) Window-BT normalized biases were around -0.25 (not shown). These bias values are a factor of 5 larger than the forecast's FmT normalized biases of around -0.05 (Figure 8.4(e)). Since Window-BT values are cooler in the cloudy situations than clear situations, the assimilation of Window-BT observations resulted in overly cloudy analysis ensembles. Though the time-integration of these analysis ensembles dramatically reduced the over cloudiness (the normalized biases went from -0.25 to -0.05), some over cloudiness likely remained. As such, the U, T, Window-BT and WV-BT biases are likely caused by the EnKF and BGEnKF experiments introducing too many clouds into the analysis ensemble.

To understand why the EnKF and BGEnKF experiments introduced too many clouds, first consider that the two experiments have similar Window-BT biases and that the single kernel mode of the BGEnKF is called for ∼90% of the assimilated observations (Figures 8.4(e) and 8.5(a)). These suggest that the EnKF algorithm is a likely culprit behind the over-introduction of clouds.

I hypothesize that the EnKF-induced over-cloudiness results from the EnKF's inability to handle clear and cloudy column members separately and the strong sensitivity of Window-BTs to hydrometeors. When both clear and cloudy column members are present in the forecast ensemble, the EnKF's forecast mean state will contain some amount of clouds. Suppose that the correlations between Window-BT and hydrometeor mixing ratios are negative. If Window-BT observations with either small or negative innovations are assimilated, the clouds in the EnKF's mean state will either be unaffected (for small innova-

tions) or be increased (for negative innovations). Since the EnKF will also reduce the size of the ensemble members' perturbations, the ensemble thus contracts around a cloudy mean state. The result is that clear column forecast members gain some amount of clouds, even in situations where the innovation is close to zero. Since Window-BTs are very sensitive to the presence of clouds, running the CRTM on such members will generate cold cloudy Window-BT values. While it is possible for the EnKF to convert cloudy column members to clear column members, in my experience, this requires a large positive innovation due to the highly nonlinear relationship between Window-BT and hydrometeors. Thus, from the perspective of the Window-BT fields, the EnKF preferentially generates clouds over removing clouds (*i.e.*, the EnKF update procedure is asymmetric). This hypothesized mechanism of EnKF-induced over-cloudiness warrants future investigation.

Thus far, I have explained the likely origins of biases in all but the Q field. Since the analysis increment cannot modify the Q biases [see Eq. (8.2)], these biases are induced during the forecast step of the DA procedure. I can rule out the evaporation of DA-induced spurious clouds as an important source because the hydrometeor biases injected by increment are an order of magnitude smaller than the Q bias growth during integration (not shown). Other processes are likely causing the Q biases. Some possibilities include enhanced upward transport of Q from the surface, and enhanced latent fluxes from the ocean surface. The exact origin of these Q biases can be investigated in future work.

8.3.5 On dynamical imbalances

It is also important to check if the BGEnKF introduces more dynamical imbalances into the ensemble than the EnKF. To measure dynamical imbalance, I compute the root-mean-square of the second time derivative of surface pressure during the time integration phase of each DA cycle. These derivatives are

computed via centered differencing (Press and Flannery, 2010) on three consecutive snapshots of the surface pressure field. Said snapshots are spaced 30-minutes apart. This imbalance metric is often used to measure fast-moving gravity waves induced by dynamical imbalances and is computed for every ensemble member (Houtekamer and Mitchell, 2005; Temperton and Williamson, 1981).

Similar to nRMSEs and normalized biases, I will normalize the imbalance metric of each experiment against the imbalance metric of the NoDA experiment. A normalized imbalance metric value of 1 indicates that a normal amount of fast-moving gravity waves is present. A value greater than 1 indicates that a higher than normal amount of fast-moving gravity waves is present, thus indicating that the model state in question is likely unbalanced.

As can be seen from the time-series of normalized imbalance in Figure 8.5(b), the BGEnKF experiment generally has either statistically indistinguishable or milder imbalances than the EnKF experiment. The only exception to this trend happens between 0000 UTC to 1200 UTC on 17th October. The BGEnKF is thus likely more appropriate than the EnKF at assimilating Window-BT observations since the BGEnKF results in similar or more balanced model states than the EnKF.

8.4 Conclusions

In this study, I compare the BGEnKF against the EnKF using perfect model OSSEs with a realistic weather model (WRF) for a case of tropical convection. These OSSEs are executed using the state-of-the-art PSU-EnKF system. My results indicate that the BGEnKF outperforms the EnKF at assimilating synthetic Window-BT observations. I observe this performance advantage in terms of

the RMSEs and biases of the U, V, T, Q, Window-BT and WV-BT fields. This performance advantage is likely due to the BGEnKF's ability to handle mixtures of clear and cloudy column members. These performance advantages are achieved even though the BGEnKF s only activated for ∼10% of the assimilated Window-BT observations. As such, these promising results motivate future work into the BGEnKF using real data.

There are several large areas of future research for the BGEnKF. The first large area concerns refining the BGEnKF algorithm. Future work can, for instance, seek less heuristic approaches to sort the ensemble into clusters in a computationally efficient manner. One option is to combine clustering algorithms [*e.g.*, k-means (Forgy, 1965; Lloyd, 1982), support-vector machines (Cortes and Vapnik, 1995) and expectation maximization (Sondergaard and Lermusiaux, 2013a)] with dimension reduction methods [*e.g.*, Sondergaard and Lermusiaux (2013a), Reddy et al. (2020), Albarakati et al. (2021)]. Since cluster sizes, and thus sampling errors, can vary in each iteration of the serial BGEnKF loop, future work can investigate using adaptive or empirical localization methods (Anderson, 2012; Anderson and Lei, 2013; Lei and Anderson, 2014) to improve the BGEnKF's performance. Future work can also examine more sophisticated methods to regulate when the BGEnKF switches over to the EnKF (*e.g.*, using the Shapiro-Wilk test for normality).

Another area of future work is to hybridize the BGEnKF with other DA algorithms. Hybridization with kernel filters (Anderson and Anderson, 1999; Hoteit et al., 2008; Stordal et al., 2011; Hoteit et al., 2012; Liu et al., 2016; Stordal and Karlsen, 2017; Kotsuki et al., 2022) can be achieved by assigning the clear cluster's covariance to clear member kernels and likewise for the cloudy member kernels. Existing ensemble-variational hybrid DA algorithms (Hamill and Snyder, 2000; Lorenc, 2003; Buehner, 2005; Wang et al., 2007) can also be hybridized with the BGEnKF. For instance, the BGEnKF can replace the EnKF component of such methods. Hybridization with DA methods that employ

transport methods to update ensemble members (Reich, 2012; van Leeuwen, 2011; Marzouk et al., 2017; Hu and van Leeuwen, 2021; Evensen Geir et al., 2022) is also possible. This can provide a different method to shift members between clusters, as opposed to the current deletion-resampling method. Finally, the BGEnKF can be potentially hybridized with ensemble DA methods that allow non-parametric prior distributions. Such methods include particle filters (van Leeuwen, 2009; Poterjoy, 2016; Vetra-Carvalho et al., 2018; Poterjoy et al., 2019; van Leeuwen et al., 2019), the quantile conserving ensemble filter (Anderson, 2022), and the rank histogram filter (Anderson, 2010, 2019, 2020).

Since I have only tested the BGEnKF in a perfect model WRF OSSE using Window-BT observations, future work can test the BGEnKF in increasingly realistic scenarios, with other observation types, and/or in other Earth systems. For instance, since radar reflectivity observations are sensitive to the presence and absence of precipitation, the BGEnKF can potentially be better at assimilating such observations. The performance of the BGEnKF can also be compared with other popular DA algorithms in tests that assimilate the operational suite of atmospheric in-situ and remote observations. Imperfect model OSSEs and real data tests can also be done. The BGEnKF can also be tested in other Earth system components.

This study is among the first to demonstrate the potential of the BGEnKF with a high-order weather model. My BGEnKF is computationally efficient, scalable with parallelization, and likely straightforward to implement in existing serial EnKF DA systems. These algorithmic properties and this study's promising results motivate future research into developing, testing and applying the BGEnKF, or similar GMM-EnKFs, for Earth systems DA.

Chapter 9

Concluding remarks

The ultimate goal of my PhD research is to advance the ensemble DA of GeoIR observations for tropical MCSs. The work presented in this dissertation contributes to this goal by 1) demonstrating that assimilating GeoIR observations via an existing DA algorithm (the EnKF) can improve the analysis and forecasts of tropical MCSs, and 2) devising with a more appropriate DA algorithm (the BGEnKF) to assimilate GeoIR observations. Nonetheless, as pointed out at the end of Chapters 4, 5 and 8, there is much work that can be done in the future to advance the DA of GeoIR observations. I will highlight the more important ones here.

With regards to assimilating GeoIR observations with the EnKF for tropical MCSs, there is currently little published literature on the impacts of adding mildly thinned GeoIR radiance observations on top of operationally assimilated observations (Global Positioning System Radio Occultation, microwave radiances observed by low earth orbit satellites, *etc.*). My work thus far only combined GeoIR radiance observations with conventional in-situ observations. The next natural step would be to test if the assimilating mildly thinned GeoIR observations with all operationally assimilated observations can benefit the analysis and forecast of tropical MCSs. Ideally, these tests should be done using an operational EnKF system and done over a seasons or so. The resulting

test outcomes would accelerate the adoption of GeoIR radiances into operational DA over the Tropics.

Compared to applications of GeoIR observation, the BGEnKF algorithm described here has much largers areas for future work. The BGEnKF here is the first known GMM-EnKF algorithm that 1) uses a small number of kernels and 2) has demonstrated advantages over the EnKF using a realistic weather model (WRF). There are many algorithmic details that can be further refined (*e.g.*, tuning the localization radii, identifying ensemble clusters, heuristic strategies to prevent unphysical weight updates). Furthermore, the BGEnKF has only been tested with synthetic Window-BT observations in a tropical MCS setting. The BGEnKF can thus be tested with many other cloud-sensitive observations (*e.g.*, real/synthetic WV-BT observations, radar reflectivity observations, and lidar observations) and different Earth system models (*e.g.*, land surface models). It would also be interesting to compare the BGEnKF with other non-Gaussian ensemble DA algorithms (*e.g.*, particle filters and rank histogram filters).

Bibliography

Albarakati, A., and Coauthors, 2021: Model and data reduction for data assimilation: Particle filters employing projected forecasts and data with application to a shallow water model. *Computers and Mathematics with Applications*, doi:10.1016/j.camwa.2021.05.026.

Alspach, D. L., and H. W. Sorenson, 1972: Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, doi:10.1109/TAC.1972.1100034.

Anderson, J. L., 2003: A Local Least Squares Framework for Ensemble Filtering. *Monthly Weather Review*, **131 (4)**, 634–642, doi:10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2.

Anderson, J. L., 2007: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **59 (2)**, doi:10.1111/j.1600-0870.2006.00216.x.

Anderson, J. L., 2010: A Non-Gaussian Ensemble Filter Update for Data Assimilation. *Monthly Weather Review*, **138 (11)**, 4186–4198, doi:10.1175/2010MWR3253.1, URL http://journals.ametsoc.org/doi/10.1175/2010MWR3253.1.

Anderson, J. L., 2012: Localization and sampling error correction in ensemble

Kalman filter data assimilation. *Monthly Weather Review*, **140 (7)**, doi:10.1175/MWR-D-11-00013.1.

Anderson, J. L., 2019: A nonlinear rank regression method for ensemble Kalman filter data assimilation. doi:10.1175/MWR-D-18-0448.1.

Anderson, J. L., 2020: A marginal adjustment rank histogram filter for non-Gaussian ensemble data assimilation. *Monthly Weather Review*, **148 (8)**, doi:10.1175/MWR-D-19-0307.1.

Anderson, J. L., 2022: A Quantile-Conserving Ensemble Filter Framework. Part I: Updating an Observed Variable. *Monthly Weather Review*, **150 (5)**, doi:10.1175/mwr-d-21-0229.1.

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, doi:10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2.

Anderson, J. L., and N. Collins, 2007: Scalable Implementations of Ensemble Filter Algorithms for Data Assimilation. *Journal of Atmospheric and Oceanic Technology*, **24 (8)**, 1452–1463, doi:10.1175/JTECH2049.1, URL https://journals.ametsoc.org/view/journals/atot/24/8/jtech2049_1.xml.

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The data assimilation research testbed a community facility. *Bulletin of the American Meteorological Society*, **90 (9)**, 1283–1296, doi:10.1175/2009BAMS2618.1, URL https://journals.ametsoc.org/doi/10.1175/2009BAMS2618.1.

Anderson, J. L., and L. Lei, 2013: Empirical localization of observation impact in ensemble Kalman filters. *Monthly Weather Review*, **141 (11)**, doi:10.1175/MWR-D-12-00330.1.

Badlan, R. L., T. P. Lane, M. W. Moncrieff, and C. Jakob, 2017: Insights into convective momentum transport and its parametrization from idealized simulations of organized convection. *Quarterly Journal of the Royal Meteorological Society*, **143 (708)**, doi:10.1002/qj.3118.

Bélair, S., and J. Mailhot, 2001: Impact of horizontal resolution on the numerical simulation of a midlatitude squall line: Implicit versus explicit condensation. *Monthly Weather Review*, **129 (9)**, doi:10.1175/1520-0493(2001)129<2362:IOHROT>2.0.CO;2.

Bengtsson, T., C. Snyder, and D. Nychka, 2003: Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research D: Atmospheres*, doi:10.1029/2002jd002900.

Benjamin, S. G., and Coauthors, 2004: An hourly assimilation-forecast cycle: The RUC. *Monthly Weather Review*, **132 (2)**, 495–518, doi:10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2.

Bessho, K., and Coauthors, 2016: An introduction to Himawari-8/9 — Japan's new-generation geostationary meteorological satellites. *Journal of the Meteorological Society of Japan*, **94 (2)**, 151–183, doi:10.2151/jmsj.2016-009.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, **129 (3)**, 420–436, doi:10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2001)129%3C0420:ASWTET%3E2.0.CO;2https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0420_aswtet_2.0.co_2.xml.

Brewster, K. A., 2003: Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part I: Method description and simulation testing. *Monthly Weather Review*, **131 (3)**, 480–492, doi:10.1175/1520-0493(2003)131<0480:PCDAAA>2.0.CO;2.

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)**, doi:10.1256/qj.04.15.

Burgers, G., P. Jan van Leeuwen, G. Evensen, P. J. Van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, **126 (6)**, 1719–1724, doi:10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(1998)126%3C1719:ASITEK%3E2.0.CO;2.

Cardinali, C., L. Isaksen, and E. Andersson, 2003: Use and impact of automated aircraft data in a global 4DVAR data assimilation system. *Monthly Weather Review*, **131 (8 PART 2)**, 1865–1877, doi:10.1175//2569.1, URL http://www.ecmwf.int/.].

Chan, M.-Y., J. L. Anderson, and X. Chen, 2020a: An efficient bi-Gaussian ensemble Kalman filter for satellite infrared radiance data assimilation. *Monthly Weather Review*, doi:10.1175/mwr-d-20-0142.1.

Chan, M.-Y., and X. Chen, 2021: Improving Analyses and Forecasts of a Tropical Squall Line using Upper Tropospheric Infrared Satellite Observations. *Advances in Atmospheric Sciences, Accepted Manuscript*, doi:10.1007/S00376-021-0449-8, URL http://www.iapjournals.ac.cn/aas/en/article/doi/10.1007/s00376-021-0449-8http://www.iapjournals.ac.cn/aas/en/article/doi/10.1007/s00376-021-0449-8?viewType=HTML.

Chan, M.-Y., X. Chen, and J. L. Anderson, 2022: The potential benefits of handling mixture statistics via a bi-Gaussian EnKF: tests with all-sky satellite infrared radiances. *Earth and Space Science Open Archive*, 40, doi:10.1002/essoar.10512183.1, URL https://doi.org/10.1002/essoar.10512183.1.

Chan, M.-Y., J. C. F. Lo, and T. Orton, 2019: The structure of tropical Sumatra squalls. *Weather*, **74 (5)**, 176–181, doi:10.1002/wea.3375.

Chan, M.-Y., F. Zhang, X. Chen, and L. R. Leung, 2020b: Potential Impacts of Assimilating All-sky Satellite Infrared Radiances on Convection-Permitting Analysis and Prediction of Tropical Convection. *Monthly Weather Review*, doi: 10.1175/mwr-d-19-0343.1.

Chen, F., and J. Dudhia, 2001: Coupling and advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Monthly Weather Review*, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Chen, X., L. R. Leung, Z. Feng, and F. Song, 2021a: Crucial Role of Mesoscale Convective Systems in the Vertical Mass, Water and Energy Transports of the South Asian Summer Monsoon. *Journal of Climate*, **-1 (aop)**, 1–46, doi:10.1175/JCLI-D-21-0124.1, URL https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-21-0124.1/JCLI-D-21-0124.1.xml.

Chen, X., L. R. Leung, Z. Feng, F. Song, and Q. Yang, 2021b: Mesoscale Convective Systems Dominate the Energetics of the South Asian Summer Monsoon Onset. *Geophysical Research Letters*, **48 (17)**, e2021GL094873, doi:10.1029/2021GL094873, URL https://onlinelibrary.wiley.com/doi/full/10.1029/2021GL094873https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL094873https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL094873.

Chen, X., L. R. Leung, Z. Feng, and Q. Yang, 2022: Precipitation-Moisture Coupling Over Tropical Oceans: Sequential Roles of Shallow, Deep, and Mesoscale Convective Systems. *Geophysical Research Letters*, **49 (7)**, doi:10.1029/2022GL097836, URL https://onlinelibrary.wiley.com/doi/10.1029/2022GL097836.

Chen, X., R. G. Nystrom, C. A. Davis, and C. M. Zarzycki, 2020: Dynamical Structures of Cross-Domain Forecast Error Covariance of a Simulated Tropical Cyclone in a Convection-Permitting Coupled Atmosphere-Ocean Model.

Monthly Weather Review, **149 (1)**, 41–63, doi:10.1175/mwr-d-20-0116.1, URL https://journals.ametsoc.org/view/journals/mwre/149/1/mwr-d-20-0116.1.xml.

Chen, X., O. Pauluis, and F. Zhang, 2018a: Regional simulation of Indian summer monsoon intraseasonal oscillations at gray-zone resolution. *Atmospheric Chemistry and Physics*, doi:10.5194/acp-18-1003-2018.

Chen, X., O. M. Pauluis, L. R. Leung, and F. Zhang, 2018b: Multiscale atmospheric overturning of the Indian summer monsoon as seen through isentropic analysis. *Journal of the Atmospheric Sciences*, doi:10.1175/JAS-D-18-0068.1.

Chen, X., O. M. Pauluis, and F. Zhang, 2018c: Atmospheric overturning across multiple scales of an MJO event during the CINDY/DYNAMO campaign. *Journal of the Atmospheric Sciences*, doi:10.1175/JAS-D-17-0060.1.

Chen, X., and F. Zhang, 2019: Relative Roles of Preconditioning Moistening and Global Circumnavigating Mode on the MJO Convective Initiation During DYNAMO. *Geophysical Research Letters*, doi:10.1029/2018GL080987.

Chen, X., K. Zhao, J. Sun, B. Zhou, and W. C. Lee, 2016: Assimilating surface observations in a four-dimensional variational Doppler radar data assimilation system to improve the analysis and forecast of a squall line case. *Advances in Atmospheric Sciences*, doi:10.1007/s00376-016-5290-0.

Chou, M.-D., and M. J. Suarez, 1999: A Solar Radiation Parameterization Atmospheric Studies. *Technical Report Series on Global Modeling and Data Assimilation*.

Cortes, C., and V. Vapnik, 1995: Support-Vector Networks. *Machine Learning*, **20 (3)**, doi:10.1023/A:1022627411411.

Doelling, D. R., and Coauthors, 2013: Geostationary Enhanced Temporal Interpolation for CERES Flux Products. *Journal of Atmospheric and Oceanic Tech-*

nology, **30 (6)**, 1072–1090, doi:10.1175/JTECH-D-12-00136.1, URL https://journals.ametsoc.org/view/journals/atot/30/6/jtech-d-12-00136_1.xml.

Dovera, L., and E. Della Rossa, 2011: Multimodal ensemble Kalman filtering using Gaussian mixture models. *Computational Geosciences*, **15 (2)**, 307–323, doi:10.1007/s10596-010-9205-3.

ECMWF, 2016: IFS Documentation CY41R2 - Part I: Observations | ECMWF. *IFS Documentation CY41R2*, ECMWF, chap. 1, URL https://www.ecmwf.int/en/elibrary/16646-ifs-documentation-cy41r2-part-i-observations.

Edwards, C. A., A. M. Moore, I. Hoteit, and B. D. Cornuelle, 2015: Regional ocean data assimilation. *Annual Review of Marine Science*, **7**, doi:10.1146/annurev-marine-010814-015821.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99 (C5)**, 10 143–10 162, doi:10.1029/94JC00572, URL https://doi.org/10.1029/94JC00572http://doi.wiley.com/10.1029/94JC00572.

Evensen Geir, Vossepoel Femke C., and van Leeuwen Peter Jan, 2022: Particle Flow for a Quasi-Geostrophic Model. *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*, Springer International Publishing, Cham, 199–206, doi:10.1007/978-3-030-96709-3{_}20, URL https://doi.org/10.1007/978-3-030-96709-3_20.

Feng, Z., R. A. Houze, L. R. Leung, F. Song, J. C. Hardin, J. Wang, W. I. Gustafson, and C. R. Homeyer, 2019: Spatiotemporal characteristics and large-scale environments of mesoscale convective systems east of the rocky mountains. *Journal of Climate*, **32 (21)**, doi:10.1175/JCLI-D-19-0137.1.

Feng, Z., L. R. Leung, R. A. Houze, S. Hagos, J. Hardin, Q. Yang, B. Han, and J. Fan, 2018: Structure and Evolution of Mesoscale Convective Systems: Sensitivity to Cloud Microphysics in Convection-Permitting Simulations Over the

United States. *Journal of Advances in Modeling Earth Systems*, **10 (7)**, doi:10.1029/2018MS001305.

Feng, Z., F. Song, K. Sakaguchi, and L. R. Leung, 2021a: Evaluation of mesoscale convective systems in climate simulations: Methodological development and results from MPAS-CAM over the United States. *Journal of Climate*, **34 (7)**, doi:10.1175/JCLI-D-20-0136.1.

Feng, Z., and Coauthors, 2021b: A Global High-Resolution Mesoscale Convective System Database Using Satellite-Derived Cloud Tops, Surface Precipitation, and Tracking. *Journal of Geophysical Research: Atmospheres*, **126 (8)**, doi:10.1029/2020JD034202.

Fletcher, S. J., 2017a: Applications of Data Assimilation in the Geosciences. *Data Assimilation for the Geosciences*, Elsevier, 887–916, doi:10.1016/b978-0-12-804444-5.00023-4.

Fletcher, S. J., 2017b: *Data Assimilation for the Geosciences: From Theory to Application*. Elsevier, Cambridge, MA, United States of America, 1–957 pp.

Forgy, E. W., 1965: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, **21 (3)**.

Fovell, R. G., G. L. Mullendore, and S. H. Kim, 2006: Discrete propagation in numerically simulated nocturnal squall lines. *Monthly Weather Review*, doi:10.1175/MWR3268.1.

Fu, J. X., W. Wang, T. Shinoda, H. L. Ren, and X. Jia, 2017: Toward Understanding the Diverse Impacts of Air-Sea Interactions on MJO Simulations. *Journal of Geophysical Research: Oceans*, **122 (11)**, doi:10.1002/2017JC013187.

Gamache, J. F., and R. A. Houze, 1982: Mesoscale air motions associated with a tropical squall line. *Monthly Weather Review*, doi:10.1175/1520-0493(1982)110<0118:MAMAWA>2.0.CO;2.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125 (554)**, 723–757, doi:10.1256/smsqj.55416, URL http://doi.wiley.com/10.1002/qj.49712555417.

Geer, A. J., S. Migliorini, and M. Matricardi, 2019: All-sky assimilation of infrared radiances sensitive to mid- and upper-tropospheric moisture and cloud. *Atmospheric Measurement Techniques Discussions*, doi:10.5194/amt-2019-9.

Geer, A. J., and Coauthors, 2018: All-sky satellite data assimilation at operational weather forecasting centres. *Quarterly Journal of the Royal Meteorological Society*, doi:10.1002/qj.3202.

Greybush, S. J., E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, 2011: Balance and ensemble Kalman filter localization techniques. *Monthly Weather Review*, **139 (2)**, 511–522, doi:10.1175/2010MWR3328.1.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter-3D variational analysis scheme. *Monthly Weather Review*, **128 (8 II)**, doi:10.1175/1520-0493(2000)128<2905:ahekfv>2.0.co;2.

Harnisch, F., M. Weissmann, and Periáñez, 2016: Error model for the assimilation of cloud-affected infrared satellite observations in an ensemble data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, doi:10.1002/qj.2776.

Hartman, C. M., X. Chen, E. E. Clothiaux, and M.-Y. Chan, 2021: Improving the Analysis and Forecast of Hurricane Dorian (2019) with Simultaneous Assimilation of GOES-16 All-Sky Infrared Brightness Temperatures and Tail Doppler Radar Radial Velocities. *Monthly Weather Review*, **149 (7)**, 2193–2212, doi:10.1175/MWR-D-20-0338.1, URL https://journals.ametsoc.org/view/journals/mwre/149/7/MWR-D-20-0338.1.xml.

Held, I. M., and A. Y. Hou, 1980: Nonlinear axially symmetric circulations in a nearly inviscid atmosphere. *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1980)037<0515:NASCIA>2.0.CO;2.

Helmert, J., and Coauthors, 2018: Review of snow data assimilation methods for hydrological, land surface, meteorological and climate models: Results from a COST harmosnow survey. doi:10.3390/geosciences8120489.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, doi:10.1002/qj.3803.

Hoffman, R. N., and C. Grassotti, 1996: A technique for assimilating SSM/I observations of marine atmospheric storms: Tests with ECMWF analyses. *Journal of Applied Meteorology*, **35 (8)**, 1177–1188, doi:10.1175/1520-0450(1996)035<1177:ATFASO>2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0450(1996)035%3C1177:ATFASO%3E2.0.CO;2.

Hoffman, R. N., Z. Liu, J.-F. Louis, and C. Grassoti, 1995: Distortion Representation of Forecast Errors. *Monthly Weather Review*, **123 (9)**, 2758–2770, doi:10.1175/1520-0493(1995)123<2758:drofe>2.0.co;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(1995)123%3C2758:DROFE%3E2.0.CO;2.

Honda, T., S. Kotsuki, G. Y. Lien, Y. Maejima, K. Okamoto, and T. Miyoshi, 2018a: Assimilation of Himawari-8 All-Sky Radiances Every 10 Minutes: Impact on Precipitation and Flood Risk Prediction. *Journal of Geophysical Research: Atmospheres*, **123 (2)**, 965–976, doi:10.1002/2017JD027096.

Honda, T., and Coauthors, 2018b: Assimilating all-sky Himawari-8 satellite infrared radiances: A case of Typhoon Soudelor (2015). *Monthly Weather Review*, **146 (1)**, 213–229, doi:10.1175/MWR-D-16-0357.1.

Hong, S. Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, doi:10.1175/MWR3199.1.

Hoskins, B. J., and D. J. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2.

Hoteit, I., X. Luo, and D.-T. Pham, 2012: Particle Kalman Filtering: A Nonlinear Bayesian Framework for Ensemble Kalman Filters. *Monthly Weather Review*, **140 (2)**, 528–542, doi:10.1175/2011MWR3640.1, URL https://journals.ametsoc.org/doi/10.1175/2011MWR3640.1.

Hoteit, I., D. T. Pham, G. Triantafyllou, and G. Korres, 2008: A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. *Monthly Weather Review*, **136 (1)**, doi:10.1175/2007MWR1927.1.

Houtekamer, P., and H. L. Mitchell, 2005: Ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, **131 (613)**, 3269–3289, doi:10.1256/qj.05.135.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, **126 (3)**, 796–811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **129 (1)**, 123–137, doi:10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **144 (12)**, 4489–4532, doi:10.1175/MWR-D-15-0440.1.

Houze, R. A., 1973: A Climatological Study of Vertical Transports by Cumulus-Scale Convection. *Journal of the Atmospheric Sciences*, **30 (6)**, doi:10.1175/1520-0469(1973)030<1112:acsovt>2.0.co;2.

Houze, R. A., 2004: Mesoscale convective systems. doi:10.1029/2004RG000150.

Hu, C. C., and P. J. van Leeuwen, 2021: A particle flow filter for high-dimensional system applications. *Quarterly Journal of the Royal Meteorological Society*, **147 (737)**, doi:10.1002/qj.4028.

Huang, X., C. Hu, X. Huang, Y. Chu, Y.-h. Tseng, G. J. Zhang, and Y. Lin, 2018: A long-term tropical mesoscale convective systems dataset based on a novel objective automatic tracking algorithm. *Climate Dynamics 2018 51:7*, **51 (7)**, 3145–3159, doi:10.1007/S00382-018-4071-0, URL https://link.springer.com/article/10.1007/s00382-018-4071-0.

Huffman, G. J., D. T. Bolvin, E. J. Nelkin, E. F. Stocker, and J. Tan, 2020: V06 IMERG Release Notes. Tech. rep., National Aeronautics and Space Administration. URL https://gpm.nasa.gov/sites/default/files/2020-10/IMERG_V06_release_notes_201006_0.pdf.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research Atmospheres*, doi:10.1029/2008JD009944.

Ingleby, B., 2017: An assessment of different radiosonde types 2015/2016. *ECMWF Tech. Memo.*, **(807)**, 69 pp.

Janjić, T., Y. Ruckstuhl, and P. L. Toint, 2021: A data assimilation algorithm for predicting rain. *Quarterly Journal of the Royal Meteorological Society*, doi:10.1002/qj.4004.

Janowiak, J. E., R. J. Joyce, and Y. Yarosh, 2001: A real-time global half-hourly pixel-resolution infrared dataset and its applications. *Bulletin of the American Meteorological Society*, **82 (2)**, doi:10.1175/1520-0477(2001)082<0205:ARTGHH>2.3.CO;2.

Järvinen, H., and P. Undén, 1997: Observation screening and background quality control in the ECMWF 3D-Var data assimilation system. Tech. rep., European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading. doi:10.21957/lyd3q81, URL https://www.ecmwf.int/node/10352.

Johnson, R. H., and P. E. Ciesielski, 2013: Structure and properties of madden-julian oscillations deduced from DYNAMO sounding arrays. *Journal of the Atmospheric Sciences*, doi:10.1175/JAS-D-13-065.1.

Johnson, R. H., T. M. Rickenbach, S. A. Rutledge, P. E. Ciesielski, and W. H. Schubert, 1999: Trimodal characteristics of Tropical convection. *Journal of Climate*, **12 (8 PART 1)**, doi:10.1175/1520-0442(1999)012<2397:tcotc>2.0.co;2.

Jones, T. A., and Coauthors, 2020: Assimilation of GOES-16 Radiances and Retrievals into the Warn-on-Forecast System. *Monthly Weather Review*, **148 (5)**, 1829–1859, doi:10.1175/MWR-D-19-0379.1, URL https://journals.ametsoc.org/view/journals/mwre/148/5/mwr-d-19-0379.1.xml.

Kalnay, E., 2003: *Atmospheric Modelling, Data Assimilation*, Vol. 129. Cambridge University Press, 2441–2442 pp., doi:10.1256/00359000360683511, URL http://doi.wiley.com/10.1256/00359000360683511.

Keppenne, C. L., M. M. Rienecker, N. P. Kurkowski, and D. A. Adamec, 2005: Ensemble Kalman filter assimilation of temperature and altimeter data with bias correction and application to seasonal prediction. *Nonlinear Processes in Geophysics*, **12 (4)**, doi:10.5194/npg-12-491-2005.

Kotsuki, S., S. J. Greybush, and T. Miyoshi, 2017: Can we optimize the assimilation order in the serial ensemble Kalman filter? A study with the Lorenz-96 model. *Monthly Weather Review*, **145 (12)**, doi:10.1175/MWR-D-17-0094.1.

Kotsuki, S., T. Miyoshi, K. Kondo, and R. Potthast, 2022: A Local Particle Filter and Its Gaussian Mixture Extension Implemented with Minor Modifications to the LETKF. *Geoscientific Model Development Discussions*, **2022**, 1–38, doi:10.5194/gmd-2022-69, URL https://gmd.copernicus.org/preprints/gmd-2022-69/.

Kunii, M., M. Otsuka, K. Shimoji, and H. Seko, 2016: Ensemble Data Assimilation and Forecast Experiments for the September 2015 Heavy Rainfall Event in Kanto and Tohoku Regions with Atmospheric Motion Vectors from Himawari-8. *SOLA*, **12 (0)**, 209–214, doi:10.2151/sola.2016-042, URL https://www.jstage.jst.go.jp/article/sola/12/0/12_2016-042/_article.

Lei, L., and J. L. Anderson, 2014: Comparisons of empirical localization techniques for serial ensemble kalman filters in a simple atmospheric general circulation model. *Monthly Weather Review*, **142 (2)**, doi:10.1175/MWR-D-13-00152.1.

Lim, K. S. S., and S. Y. Hong, 2010: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models. *Monthly Weather Review*, doi:10.1175/2009MWR2968.1.

Liu, B., B. Ait-El-Fquih, and I. Hoteit, 2016: Efficient kernel-based ensemble Gaussian mixture filtering. *Monthly Weather Review*, **144 (2)**, doi:10.1175/MWR-D-14-00292.1.

Liu, C., 2011: Rainfall contributions from precipitation systems with different sizes, convective intensities, and durations over the tropics and subtropics. *Journal of Hydrometeorology*, doi:10.1175/2010JHM1320.1.

Liu, Z.-Q., and F. Rabier, 2002: The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Quarterly Journal of the Royal Meteorological Society*, **128**, 1367–1386.

Lloyd, S. P., 1982: Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, **28 (2)**, doi:10.1109/TIT.1982.1056489.

Lo, J. C. F., and T. Orton, 2016: The general features of tropical Sumatra Squalls. *Weather*, doi:10.1002/wea.2748.

Lorenc, A. C., 2003: Modelling of error covariances by 4D-Var data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **129 (595 PART B)**, 3167–3182, doi:10.1256/qj.02.131.

Lorentzen, R. J., and G. Naevdal, 2011: An iterative ensemble kalman filter. *IEEE Transactions on Automatic Control*, doi:10.1109/TAC.2011.2154430.

Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 Day Oscillation in the Zonal Wind in the Tropical Pacific. *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1971)028<0702:doadoi>2.0.co;2.

Madden, R. A., and P. R. Julian, 1972: Description of Global-Scale Circulation Cells in the Tropics with a 40–50 Day Period. *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1972)029<1109:dogscc>2.0.co;2.

Marzouk, Y., T. Moselhy, M. Parno, and A. Spantini, 2017: Sampling via measure transport: An introduction. *Handbook of Uncertainty Quantification*, doi:10.1007/978-3-319-12385-1{_}23.

Meng, Z., and F. Zhang, 2007: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part II: Imperfect model experiments. *Monthly Weather Review*, **135 (4)**, 1403–1423, doi:10.1175/MWR3352.1.

Meng, Z., and F. Zhang, 2008: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: Comparison with 3DVAR in a real-data case study. *Monthly Weather Review*, doi:10.1175/2007MWR2106.1.

Minamide, M., and F. Zhang, 2017: Adaptive observation error inflation for assimilating all-Sky satellite radiance. *Monthly Weather Review*, **145 (3)**, 1063–1081, doi:10.1175/MWR-D-16-0257.1.

Minamide, M., and F. Zhang, 2018: Assimilation of all-sky infrared radiances from Himawari-8 and impacts of moisture and hydrometer initialization on convection-permitting tropical cyclone prediction. *Monthly Weather Review*, **146 (10)**, 3241–3258, doi:10.1175/MWR-D-17-0367.1.

Minamide, M., and F. Zhang, 2019: An adaptive background error inflation method for assimilating all-sky radiances. *Quarterly Journal of the Royal Meteorological Society*, **145 (719)**, 805–823, doi:10.1002/qj.3466.

Mohr, K. I., J. S. Famiglietti, and E. J. Zipser, 1999: The contribution to tropical rainfall with respect to convective system type, size, and intensity estimated from the 85-GHz ice-scattering signature. *Journal of Applied Meteorology*, doi:10.1175/1520-0450(1999)038<0596:TCTTRW>2.0.CO;2.

Moncrieff, M. W., and C. Liu, 1999: Convection initiation by density currents: Role of convergence, shear, and dynamical organization. *Monthly Weather Review*, doi:10.1175/1520-0493(1999)127<2455:CIBDCR>2.0.CO;2.

Nehrkorn, T., R. N. Hoffman, C. Grassotti, and J.-F. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Quarterly Journal of the Royal Meteorological Society*, **129 (587)**, 195–218, doi:10.1256/qj.02.18, URL http://doi.wiley.com/10.1256/qj.02.18.

Nehrkorn, T., B. K. Woods, R. N. Hoffman, and T. Auligné, 2015: Correcting for Position Errors in Variational Data Assimilation. *Monthly Weather Review*, **143 (4)**, 1368–1381, doi:10.1175/MWR-D-14-00127.1.

Nesbitt, S. W., R. Cifelli, and S. A. Rutledge, 2006: Storm morphology and rainfall characteristics of TRMM precipitation features. *Monthly Weather Review*, doi:10.1175/MWR3200.1.

Okamoto, K., Y. Sawada, and M. Kunii, 2019: Comparison of assimilating all-sky and clear-sky infrared radiances from Himawari-8 in a mesoscale system. *Quarterly Journal of the Royal Meteorological Society*, **145 (719)**, 745–766, doi:10.1002/qj.3463, URL https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3463https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3463https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3463.

Otkin, J. A., 2010: Clear and cloudy sky infrared brightness temperature assimilation using an ensemble Kalman filter. *Journal of Geophysical Research Atmospheres*, doi:10.1029/2009JD013759.

Otkin, J. A., 2012: Assessing the impact of the covariance localization radius when assimilating infrared brightness temperature observations using an ensemble kalman filter. *Monthly Weather Review*, doi:10.1175/MWR-D-11-00084.1.

Otkin, J. A., and R. Potthast, 2019: Assimilation of All-Sky seviri infrared brightness temperatures in a regional-scale ensemble data assimilation system. *Monthly Weather Review*, doi:10.1175/MWR-D-19-0133.1.

Otkin, J. A., R. Potthast, and A. S. Lawless, 2018: Nonlinear bias correction for satellite data assimilation using taylor series polynomials. *Monthly Weather Review*, doi:10.1175/MWR-D-17-0171.1.

Otsuka, M., M. Kunii, H. Seko, K. Shimoji, M. Hayashi, and K. Yamashita, 2015: Assimilation experiments of MTSAT rapid scan atmospheric motion vectors on a heavy rainfall event. *Journal of the Meteorological Society of Japan*, **93 (4)**, 459–475, doi:10.2151/jmsj.2015-030, URL https://www.jstage.jst.go.jp/article/jmsj/93/4/93_2015-030/_article.

Park, S. K., and L. Xu, 2016: *Data assimilation for atmospheric, oceanic and hydrologic applications (Vol. III)*. doi:10.1007/978-3-319-43415-5.

Penny, S. G., and T. Miyoshi, 2016: A local particle filter for high-dimensional geophysical systems. *Nonlinear Processes in Geophysics*, doi:10.5194/npg-23-391-2016.

Poterjoy, J., 2016: A localized particle filter for high-dimensional nonlinear systems. *Monthly Weather Review*, **144 (1)**, 59–76, doi:10.1175/MWR-D-15-0163.1, URL http://journals.ametsoc.org/doi/10.1175/MWR-D-15-0163.1.

Poterjoy, J., L. Wicker, and M. Buehner, 2019: Progress toward the application of a localized particle filter for numerical weather prediction. *Monthly Weather Review*, doi:10.1175/MWR-D-17-0344.1.

Press, W., and B. Flannery, 2010: *Numerical Recipes in Fortran 90*, Vol. 35.

Rasmussen, C. E., and C. K. I. Williams, 2005: Gaussian Identities. *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, chap. Appendix A, 200–201.

Reddy, G. T., M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, 2020: Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, **8**, doi:10.1109/ACCESS.2020.2980942.

Reich, S., 2012: A Gaussian-mixture ensemble transform filter. *Quarterly Journal of the Royal Meteorological Society*, **138 (662)**, doi:10.1002/qj.898.

Reichle, R. H., M. G. Bosilovich, W. T. Crow, R. D. Koster, S. V. Kumar, S. P. P. Mahanama, and B. F. Zaitchik, 2009: Recent Advances in Land Data Assimilation at the NASA Global Modeling and Assimilation Office. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, doi:10.1007/978-3-540-71056-1{_}21.

Roberts, N. M., and H. W. Lean, 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, **136 (1)**, 78–97, doi:10.1175/2007MWR2123.1, URL https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml.

Roca, R., and T. Fiolleau, 2020: Extreme precipitation in the tropics is closely associated with long-lived convective systems. *Communications Earth & Environment*, doi:10.1038/s43247-020-00015-4.

Ruppert, J. H., and X. Chen, 2020: Island Rainfall Enhancement in the Maritime Continent. *Geophysical Research Letters*, doi:10.1029/2019GL086545.

Ruppert, J. H., X. Chen, and F. Zhang, 2020: Convectively forced diurnal gravity waves in the maritime continent. *Journal of the Atmospheric Sciences*, doi:10.1175/JAS-D-19-0236.1.

Sapsis, T. P., and P. F. Lermusiaux, 2009: Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D: Nonlinear Phenomena*, doi:10.1016/j.physd.2009.09.017.

Sapsis, T. P., and P. F. Lermusiaux, 2012: Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. *Physica D: Nonlinear Phenomena*, doi:10.1016/j.physd.2011.10.001.

Satoh, M., 1994: Hadley Circulations in Radiative–Convective Equilibrium in an Axially Symmetric Atmosphere. *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1994)051<1947:hcirei>2.0.co;2.

Sawada, Y., K. Okamoto, M. Kunii, and T. Miyoshi, 2019: Assimilating Every-10-minute Himawari-8 Infrared Radiances to Improve Convective Predictability. *Journal of Geophysical Research: Atmospheres*, doi:10.1029/2018JD029643.

Schmit, T. J., M. M. Gunshor, W. P. Menzel, J. J. Gurka, J. Li, and A. S. Bachmeier, 2005: Introducing the next-generation advanced baseline imager on GOES-

R. *Bulletin of the American Meteorological Society*, **86 (8)**, 1079–1096, doi:10.1175/BAMS-86-8-1079, URL http://speclib.

Skamarock, W., and Coauthors, 2008: A Description of the Advanced Research WRF Version 3. NCAR Tech. Note NCAR/TN–468+STR, 113 pp. *NCAR TECHNICAL NOTE*, doi:10.5065/D68S4MVH.

Sondergaard, T., and P. F. Lermusiaux, 2013a: Data assimilation with gaussian mixture models using the dynamically orthogonal field equations. Part I: Theory and scheme. *Monthly Weather Review*, **141 (6)**, 1737–1760, doi:10.1175/MWR-D-11-00295.1, URL https://journals.ametsoc.org/view/journals/mwre/141/6/mwr-d-11-00295.1.xml.

Sondergaard, T., and P. F. Lermusiaux, 2013b: Data assimilation with gaussian mixture models using the dynamically orthogonal field equations. Part II: Applications. *Monthly Weather Review*, **141 (6)**, 1761–1785, doi:10.1175/MWR-D-11-00296.1.

Stammer, D., M. Balmaseda, P. Heimbach, A. Köhl, and A. Weaver, 2016: Ocean Data Assimilation in Support of Climate Applications: Status and Perspectives. *Annual Review of Marine Science*, **8**, doi:10.1146/annurev-marine-122414-034113.

Steward, J. L., 2012: Practical Optimization Algorithms in the Data Assimilation of Large-Scale Systems with Non-Linear and Non-Smooth Observation Operators. Ph.D. thesis, Florida State University, URL http://purl.flvc.org/fsu/fd/FSU_migr_etd-5203.

Stordal, A. S., and H. A. Karlsen, 2017: Large sample properties of the adaptive gaussian mixture filter. *Monthly Weather Review*, **145 (7)**, doi:10.1175/MWR-D-15-0372.1.

Stordal, A. S., H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallès, 2011: Bridging the ensemble Kalman filter and particle filters: The adaptive

Gaussian mixture filter. *Computational Geosciences*, **15 (2)**, doi:10.1007/s10596-010-9207-1.

Stratman, D. R., C. K. Potvin, and L. J. Wicker, 2018: Correcting Storm Displacement Errors in Ensembles Using the Feature Alignment Technique (FAT). *Monthly Weather Review*, **146 (7)**, 2125–2145, doi:10.1175/MWR-D-17-0357.1, URL https://journals.ametsoc.org/view/journals/mwre/146/7/mwr-d-17-0357.1.xml.

Sun, X., and Coauthors, 2020: A subjective and objective evaluation of model forecasts of sumatra squall events. *Weather and Forecasting*, doi:10.1175/WAF-D-19-0187.1.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, doi:10.1175/BAMS-D-13-00191.1.

Tan, J., G. J. Huffman, D. T. Bolvin, and E. J. Nelkin, 2019: IMERG V06: Changes to the Morphing Algorithm. *Journal of Atmospheric and Oceanic Technology*, **36 (12)**, 2471–2482, doi:10.1175/JTECH-D-19-0114.1, URL https://journals.ametsoc.org/view/journals/atot/36/12/jtech-d-19-0114.1.xml.

Temperton, C., and D. L. Williamson, 1981: Normal mode initialization for a multilevel grid-point model. Part I: Linear aspects. *Monthly Weather Review*, **109 (4)**, doi:10.1175/1520-0493(1981)109<0729:NMIFAM>2.0.CO;2.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Monthly Weather Review*, **136 (12)**, 5095–5115, doi:10.1175/2008MWR2387.1, URL http://journals.ametsoc.org/doi/10.1175/2008MWR2387.1.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble Square Root Filters. *Monthly Weather Review*,

131 (7), 1485–1490, doi:10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2003)131%3C1485:ESRF%3E2.0.CO;2.

Ueckermann, M. P., P. F. Lermusiaux, and T. P. Sapsis, 2013: Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *Journal of Computational Physics*, doi:10.1016/j.jcp.2012.08.041.

van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. doi:10.1175/2009MWR2835.1.

van Leeuwen, P. J., 2011: Efficient nonlinear data assimilation for oceanic models of intermediate complexity. *IEEE Workshop on Statistical Signal Processing Proceedings*, doi:10.1109/SSP.2011.5967700.

van Leeuwen, P. J., H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, 2019: Particle filters for high-dimensional geoscience applications: A review. John Wiley and Sons Ltd, 2335–2365 pp., doi:10.1002/qj.3551.

Vetra-Carvalho, S., P. J. van Leeuwen, L. Nerger, A. Barth, M. U. Altaf, P. Brasseur, P. Kirchgessner, and J. M. Beckers, 2018: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. *Tellus, Series A: Dynamic Meteorology and Oceanography*, doi:10.1080/16000870.2018.1445364.

Vukicevic, T., T. Greenwald, M. Zupanski, D. Zupanski, T. Vonder Haar, and A. S. Jones, 2004: Mesoscale cloud state estimation from visible and infrared satellite radiances. *Monthly Weather Review*, doi:10.1175/MWR2837.1.

Vukicevic, T., M. Sengupta, A. S. Jones, and T. Vonder Haar, 2006: Cloud-resolving satellite data assimilation: Information content of IR window observations and uncertainties in estimation. *Journal of the Atmospheric Sciences*, doi:10.1175/JAS3639.1.

Wang, S., A. H. Sobel, F. Zhang, Y. Qiang Sun, Y. Yue, and L. Zhou, 2015: Regional simulation of the october and november MJO events observed during the CINDY/DYNAMO field campaign at gray zone resolution. *Journal of Climate*, **28 (6)**, 2097–2119, doi:10.1175/JCLI-D-14-00294.1.

Wang, X., C. Snyder, and T. M. Hamill, 2007: On the theoretical equivalence of differently proposed ensemble - 3DVAR hybrid analysis schemes. *Monthly Weather Review*, **135 (1)**, doi:10.1175/MWR3282.1.

Weickmann, K. M., 1983: Intraseasonal circulation and outgoing longwave radiation modes during Northern Hemisphere winter. *Monthly Weather Review*, doi:10.1175/1520-0493(1983)111<1838:ICAOLR>2.0.CO;2.

Wheeler, M., and G. N. Kiladis, 1999: Convectively Coupled Equatorial Waves: Analysis of Clouds and Temperature in the Wavenumber-Frequency Domain. *Journal of the Atmospheric Sciences*, doi:10.1175/1520-0469(1999)056<0374:CCEWAO>2.0.CO;2.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, doi:10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, **130 (7)**, 1913–1924, doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2002)130%3C1913:EDAWPO%3E2.0.CO;2.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP global forecast system. *Monthly Weather Review*, **136 (2)**, doi:10.1175/2007MWR2018.1.

Wielicki, B. A., B. R. Barkstrom, E. F. Harrison, R. B. Lee, G. L. Smith, and J. E. Cooper, 1996: Clouds and the Earth's Radiant Energy System (CERES): An

Earth Observing System Experiment. *Bulletin of the American Meteorological Society*, **77 (5)**, doi:10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2.

Yang, G. Y., and J. Slingo, 2001: The diurnal cycle in the tropics. *Monthly Weather Review*, **129**, 784–801, doi:10.1175/1520-0493(2001)129<0784:TDCITT>2.0.CO;2, URL http://centaur.reading.ac.uk/24887/.

Ying, Y., 2019: A multiscale alignment method for ensemble filtering with displacement errors. *Monthly Weather Review*, **147 (12)**, 4553–4565, doi:10.1175/MWR-D-19-0170.1, URL www.ametsoc.org/PUBSReuseLicenses.

Ying, Y., and F. Zhang, 2017: Practical and intrinsic predictability of multiscale weather and convectively coupled equatorial waves during the active phase of an MJO. *Journal of the Atmospheric Sciences*, **74 (11)**, 3771–3785, doi:10.1175/JAS-D-17-0157.1.

Ying, Y., and F. Zhang, 2018: Potentials in improving predictability of multiscale tropical weather systems evaluated through ensemble assimilation of simulated satellite-based observations. *Journal of the Atmospheric Sciences*, **75 (5)**, 1675–1698, doi:10.1175/JAS-D-17-0245.1.

Ying, Y., F. Zhang, and J. L. Anderson, 2018: On the selection of localization radius in ensemble filtering for multiscale quasigeostrophic dynamics. *Monthly Weather Review*, **146 (2)**, 543–560, doi:10.1175/MWR-D-17-0336.1.

Zeng, X., and A. Beljaars, 2005: A prognostic scheme of sea surface skin temperature for modeling and data assimilation. *Geophysical Research Letters*, doi:10.1029/2005GL023030.

Zhang, C., 2005: Madden-Julian Oscillation. doi:10.1029/2004RG000158.

Zhang, C., J. Gottschalck, E. D. Maloney, M. W. Moncrieff, F. Vitart, D. E. Waliser, B. Wang, and M. C. Wheeler, 2013: Cracking the MJO nut. *Geophysical Research Letters*, doi:10.1002/grl.50244.

Zhang, C., and K. Yoneyama, 2017: CINDY/DYNAMO field campaign: Advancing our understanding of MJO initiation. *World Scientific Series on Asia-Pacific Weather and Climate*, Vol. Volume 9, doi:10.1142/9789813200913{_}0027.

Zhang, F., M. Minamide, and E. E. Clothiaux, 2016: Potential impacts of assimilating all-sky infrared satellite radiances from GOES-R on convection-permitting analysis and prediction of tropical cyclones. *Geophysical Research Letters*, **43 (6)**, 2954–2963, doi:10.1002/2016GL068468.

Zhang, F., M. Minamide, R. G. Nystrom, X. Chen, S.-J. Lin, and L. M. Harris, 2019: Improving Harvey Forecasts with Next-Generation Weather Satellites: Advanced Hurricane Analysis and Prediction with Assimilation of GOES-R All-Sky Radiances. *Bulletin of the American Meteorological Society*, **100 (7)**, 1217–1222, doi:10.1175/BAMS-D-18-0149.1, URL https://journals.ametsoc.org/view/journals/bams/100/7/bams-d-18-0149.1.xml.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Monthly Weather Review*, **132 (5)**, 1238–1253, doi:10.1175/1520-0493(2004)132<1238:IOIEAO>2.0.CO;2.

Zhang, F., S. Taraphdar, and S. Wang, 2017: The role of global circumnavigating mode in the MJO initiation and propagation. *Journal of Geophysical Research*, **122 (11)**, 5837–5856, doi:10.1002/2016JD025665.

Zhang, G., and N. a. McFarlane, 1995: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model. *Atmosphere-Ocean*, **33 (3)**, 407–446, doi:10.1080/07055900.1995.9649539.

Zhang, J., S. Hu, and W. Duan, 2021a: On the sensitive areas for targeted observations in ENSO forecasting. *Atmospheric and Oceanic Science Letters*, 100054, doi:10.1016/j.aosl.2021.100054, URL https://linkinghub.elsevier.com/retrieve/pii/S167428342100026X.

Zhang, Y., F. Zhang, and D. J. Stensrud, 2018: Assimilating all-sky infrared radiances from GOES-16 ABI using an ensemble Kalman filter for convection-allowing severe thunderstorms prediction. *Monthly Weather Review*, doi:10.1175/MWR-D-18-0062.1.

Zhang, Y., and Coauthors, 2021b: Ensemble-Based Assimilation of Satellite All-Sky Microwave Radiances Improves Intensity and Rainfall Predictions for Hurricane Harvey (2017). John Wiley and Sons Inc, doi:10.1029/2021GL096410.

VITA

Man-Yau (“Joseph”) Chan

Address: Department of Meteorology and Atmospheric Science
The Pennsylvania State University
University Park, PA 16802

Education: Bachelors in Science (Honors with Distinction), Physics
National University of Singapore, 2017

First-authored publications:

1. **Chan, M.-Y.**, Chen X. and Anderson J. (*under review*): The potential benefits of handling clear and cloudy ensemble members separately through an efficient bi-Gaussian EnKF. *Journal of Advances in Modelling Earth Systems*.

2. **Chan, M.-Y.**, Chen X. and Leung R. L. (2022, *accepted*): A High-Resolution Tropical Mesoscale Convective System Reanalysis (TMeCSR). *Journal of Advances in Modelling Earth Systems*.

3. **Chan, M.-Y.**, and Chen X. (2021): Improving the Analyses and Forecasts of a Tropical Squall Line Using Upper Tropospheric Infrared Satellite Observations. *Advances in Atmospheric Sciences*. doi: 10.1007/s00376-021-0449-8

4. **Chan, M.-Y.**, Anderson J. L. and Chen X. (2020): An Efficient Bi-Gaussian Ensemble Kalman Filter for Satellite Infrared Radiance Data Assimilation. *Monthly Weather Review.* doi: 10.1175/MWR-D-20-0142.1

5. **Chan, M.-Y.**, Zhang F., Chen X. and Leung R. L. (2020): Impacts of Assimilating All-sky Satellite Infrared Radiances on Convection-Permitting Analysis and Prediction of Tropical Convection. *Monthly Weather Review.* doi: 10.1175/MWR-D-19-0343.1

6. **Chan, M. Y.**, Lo, J. C. and Orton, T. (2019), The structure of tropical Sumatra squalls. *Weather.* doi: 10.1002/wea.3375