Honours Project in Physics

A Study of Northeast Monsoon Forecast Errors

Author:

Chan Man Yau (A0116571J)

Supervisors:

Professor Lim Hock

Dr. Sun Xiang Ming

Abstract

The forecast error covariance matrix, $\boldsymbol{B}$, is essential for data assimilation. However, $\boldsymbol{B}$ is too big to be directly computed and stored. To overcome this problem, it is common to diagonalize $\boldsymbol{B}$ through a two step process. The first step transforms the forecast errors, $\boldsymbol{x^b}$, into a set of errors that possess zero cross-covariances, resulting in a block diagonal $\boldsymbol{B}$ containing only spatial auto-covariances. The second step then removes the spatial auto-covariances, resulting in a diagonal $\boldsymbol{B}$. This project focused on the first step and utilized linear relationships to decorrelate the stream function (ψ) and velocity potential (ϕ) forecast errors for the Singapore Variable resolution model, during the Maritime Continent Northeast monsoon season in 2015. However, the linear regression coefficient matrix relating ϕ and ψ is too big to be computed and must thus be modelled.

We will show that the two commonly used regression coefficient matrix models in the Weather Research and Forecasting Data Assimilation system do not perform well in the region and season considered. The first model assumes that ψ and ϕ are disconnected across model levels and are independent of the horizontal positions of ψ and ϕ considered. The second model disposes of the first assumption, but keeps the second assumption. Examinations of the columns of $\boldsymbol{B}$ (estimated via the National Meteorology Centre method) revealed that cross-covariances depend on the horizontal positions of the ψ and ϕ, meaning that the linear regression coefficients depend heavily on the horizontal positions considered. This breaks down the common assumption of the two models. We also considered the efficacy of the models in removing cross-covariances, and found them to be unsurprisingly bad at cross-covariance removal.

Examinations of the columns of $\boldsymbol{B}$ also revealed that the values of forecast error variables surrounding a point can be taken as samples of the forecast error variables at that point (neighbourhood equivalence assumption). Following this assumption, the forecast model grid is broken into subdomains and the regression coefficients are calculated between pairs of subdomains. This new model was found to be much better at removing cross-covariances because its assumption holds.

Acknowledgements

This project would not be possible without the guidance and mentorship of Professor Lim Hock from the National University of Singapore, Faculty of Science, Department of Physics, and Dr. Sun Xiangming from the Meteorological Services Singapore, Centre for Climate Research Singapore, Weather Modelling and Prediction. Their immense patience and understanding as I floundered through the theory, created buggy routines and occasionally overtaxing servers, are nothing short of admirable. I would also like to thanks Dr Huang Xiang-Yu and the staff from the Centre for Climate Research Singapore for their encouragement during this project and for letting me use their Athena computing cluster.

Aside from that, I would like to thank Soong Yun Ting and Liu Xiao Rui for being there when I raved and frothed at the mouth (metaphorically) while debugging routines and working out theoretical matters. Thank you Ng Wei Jie for letting me bounce ideas around. I would also like to thank Dr. Siu Zhuo Bin and John Ouyang for proofreading this thesis.

Lastly and most importantly, I would like to thank my parents, whose patience and love helped me to get through this project in one piece.

Contents

Chapter 1

Introduction

1.1 Data assimilation background

Need for data assimilation To obtain a good numerical weather forecast, the initial conditions must accurately represent the true state of the pre-forecast atmosphere (Caron and Fillion, 2010). Unfortunately, it is difficult in practice to obtain good initial conditions. Observations alone are insufficient to generate a full set of initial conditions — there are typically far fewer observations than model grid points (Bouttier and Courtier, 2002).[1] To make matters worse, while educated guesses (*e.g.*, result of a previous forecast) can easily generate a full set of initial conditions, they are susceptible to error. The solution is to combine the observations and educated guesses to produce a set of initial conditions that has the greatest probability of being accurate (Lorenc, 1986; Bannister, 2008a). This combination process is known as data assimilation, or DA.

DA problem To be exact, data assimilation seeks to determine a model state (state vector, $\boldsymbol{X}$) that has the maximum probability density of being true, given a set of observations (observation vector, $\boldsymbol{Y}$), and in the light of an educated guess (background state, $\boldsymbol{X^b}$). This probability density can be expressed using Bayes' theorem (Lorenc, 1986; Kalnay, 2003):

$$p\left(\boldsymbol{X}|\boldsymbol{Y}\right) \propto p\left(\boldsymbol{Y}|\boldsymbol{X}\right) p\left(\boldsymbol{X}\right). \tag{1.1}$$

Contribution from $\boldsymbol{X^b}$ The prior probability density, $p\left(\boldsymbol{X}\right)$, represents the probability that $\boldsymbol{X}$ is an accurate representation of the atmosphere, before considering the evidence.

[1] Forecast models typically run on at least $O\left(10^6\right)$ grid points and $O\left(10\right)$ variables. *I.e.*, an unrealistic $O\left(10^7\right)$ observations are needed for initialization.

During this phase, $\boldsymbol{X^b}$ is typically taken to be the best guess at what the the true state $\boldsymbol{X^t}$ looks like. As such, $p(\boldsymbol{X})$ is typically modelled by

$$p(\boldsymbol{X}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{X}-\boldsymbol{X^b}\right)^\top \boldsymbol{B}^{-1}\left(\boldsymbol{X}-\boldsymbol{X^b}\right)\right\},$$

where $\boldsymbol{B}$ is the background error covariance matrix (Lorenc, 1986). $\boldsymbol{B}$ is defined as

$$\boldsymbol{B} \equiv \frac{1}{N_T}\sum_{t=1}^{N_T}\left(\boldsymbol{X^b}(t)-\boldsymbol{X^t}(t)\right)\left(\boldsymbol{X^b}(t)-\boldsymbol{X^t}(t)\right)^\top \tag{1.2}$$

N_T is the number of samples considered, $\boldsymbol{X^t}$ is the true state of the atmosphere and t is an index representing the sample number. For the ease of notation, let us define the background state error vector,

$$\boldsymbol{x^b}(t) \equiv \boldsymbol{X^b}(t)-\boldsymbol{X^t}(t). \tag{1.3}$$

This also means that

$$\boldsymbol{B} = \frac{1}{N_T}\sum_{t=1}^{N_T}\boldsymbol{x^b}(t)\left[\boldsymbol{x^b}(t)\right]^\top$$

Contribution from $\boldsymbol{Y}$ The other term of Eqn (1.1), $p(\boldsymbol{Y}|\boldsymbol{X})$, is typically modelled by a similar Gaussian distribution, except that the reference (or the "mean") is taken to be $\boldsymbol{X}$ (Lorenc, 1986). Since $\boldsymbol{X}$ exists in a different basis from $\boldsymbol{Y}$, an operator, H, that maps $\boldsymbol{X}$ to $\boldsymbol{Y}$ is needed. Taken together, the model is

$$p(\boldsymbol{Y}|\boldsymbol{X}) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{Y}-H(\boldsymbol{X})\right]^\top \boldsymbol{R}^{-1}\left[\boldsymbol{Y}-H(\boldsymbol{X})\right]\right\}.$$

The observation error covariance matrix, $\boldsymbol{R}$, is defined as (Bouttier and Courtier, 2002),

$$\boldsymbol{R} \equiv \frac{1}{N_T}\sum_{t=1}^{N_T}\left[\boldsymbol{Y}(t)-H\left(\boldsymbol{X^t}(t)\right)\right]\left[\boldsymbol{Y}(t)-H\left(\boldsymbol{X^t}(t)\right)\right]^\top.$$

Full DA problem Bringing these two probability densities into Eqn (1.1) yields:

$$p(\boldsymbol{X}|\boldsymbol{Y}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{X}-\boldsymbol{X^b}\right)^\top \boldsymbol{B}^{-1}\left(\boldsymbol{X}-\boldsymbol{X^b}\right)-\frac{1}{2}\left[\boldsymbol{Y}-H(\boldsymbol{X})\right]^\top \boldsymbol{R}^{-1}\left[\boldsymbol{Y}-H(\boldsymbol{X})\right]\right\}.$$

DA cost function It is immediately apparent that maximizing $P(\boldsymbol{X}|\boldsymbol{Y})$ is equivalent to minimizing

$$J(\boldsymbol{X}) = \left(\boldsymbol{X} - \boldsymbol{X^b}\right)^\top \boldsymbol{B}^{-1} \left(\boldsymbol{X} - \boldsymbol{X^b}\right) + \left[\boldsymbol{Y} - H(\boldsymbol{X})\right]^\top \boldsymbol{R}^{-1} \left[\boldsymbol{Y} - H(\boldsymbol{X})\right]. \tag{1.4}$$

$J(\boldsymbol{X})$ is known in the literature as the cost function (Kalnay, 2003; Bannister, 2008a). In short, data assimilation determines the best initial conditions ($\boldsymbol{X^a}$) by minimizing Eqn (1.4), given a set of observations and an educated guess.[2]

1.2 Background error covariance matrix

Roles of $\boldsymbol{B}$ in DA The background error covariance matrix ($\boldsymbol{B}$), as defined in Eqn (1.2), plays several important roles in data assimilation. First of all, $\boldsymbol{B}$ weighs the contributions from $\boldsymbol{Y}$ and $\boldsymbol{X^b}$ to $\boldsymbol{X^a}$, according to their variances. The greater the relative variance of one component, the smaller its contribution to $\boldsymbol{X^a}$. Aside from that, $\boldsymbol{B}$ is also responsible for spreading the difference in the observation and the background state throughout the model grid and across different variables. This means that $\boldsymbol{B}$ allows a single observation to affect the entire set of initial conditions. Lastly, when multiple observations are assimilated, $\boldsymbol{B}$ causes the observations to interact at locations/variables different from the observation location and variable (Bouttier and Courtier, 2002; Kalnay, 2003; Bannister, 2008a). Clearly, an accurate $\boldsymbol{B}$ is essential for obtaining good initial conditions via data assimilation.[3]

1.2.1 Overwhelming size and unknown true state

Difficulties in obtaining $\boldsymbol{B}$ Unfortunately, it is impossible to compute $\boldsymbol{B}$ directly from the definition in Eqn (1.2) due to two problems. First of all the definition in Eqn (1.2) requires the unobtainable $\boldsymbol{X^t}$. This is typically handled by the National Meteorology Center method (Parrish and Derber, 1992), which will be covered in Chapter 2. Furthermore, since $\boldsymbol{B}$ is an $N \times N$ matrix (N being the number of grid points times variables in the model), $\boldsymbol{B}$ typically contains $O(10^{14})$ elements. In other words, computations involving $\boldsymbol{B}$ are severely constrained by the impractical amount of computational memory required ($\sim O(10^2)$ terabytes). These mean direct computations of $\boldsymbol{B}$ are clearly impractical (Bannister, 2008b).

[2] A solution to Eqn (1.4) can be derived under the assumption that the interpolation operator H can be linearized for small perturbations in $\boldsymbol{x}$. See Appendix A.1 for the details.

[3] All three roles of $\boldsymbol{B}$ can be derived from Eqn (1.4). See Appendix A.2 for the details.

Diagonalizing $\boldsymbol{B}$ by eigenvectors Transforming $\boldsymbol{x^b}$ into a vector of mutually uncorrelated elements immediately diagonalizes $\boldsymbol{B}$ (Lanczos, 1957; Bannister, 2008b). Since there are only N non-zero elements in the diagonalized $\boldsymbol{B}$, $\boldsymbol{B}$ will not have computational memory problems. In principle, the ideal way of doing so is through the use of the eigenvectors of $\boldsymbol{B}$. When $\boldsymbol{B}$ is transformed to the basis of its eigenvectors, it is diagonal. The transformation operator for $\boldsymbol{x^b}$ then is simply an $N \times N$ matrix containing $\boldsymbol{B}$'s eigenvectors as its rows.

Replace eigenvector transform with control variable transform Unfortunately, it is difficult to compute the eigenvectors if $\boldsymbol{B}$ cannot be computed in the first place. As such, the eigenvector transform method is replaced with a two-step transformation process (Parrish and Derber, 1992; Wu et al., 2002; Bannister, 2008b). Suppose that $\boldsymbol{x^b}$ contains the background error of L variables. The first step transforms $\boldsymbol{x^b}$ into a vector, $\boldsymbol{x^{b'}}$, where the cross-variable covariances (cross-covariance for short) are zero. The second step then removes the auto-covariances (covariance of the same variable, but between any pair of model grid points) of $\boldsymbol{x^{b'}}$, resulting in the vector $\boldsymbol{x^{b''}}$. Since $\boldsymbol{x^{b''}}$ will have zero cross-covariances and zero-autocovariances, its covariance matrix,

$$\left\langle \boldsymbol{x^{b''}}\, \boldsymbol{x^{b''\top}} \right\rangle = \frac{1}{N_T} \sum_{t=1}^{N_T} \boldsymbol{x^{b''}}\left(t\right)\, \boldsymbol{x^{b''}}\left(t\right)^{\top},$$

will be diagonal. This two-step transformation is typically called the control variable transform (CVT)[4].

Removing cross-covariances with least squares linear regression residues A good choice of variables for $\boldsymbol{x^{b'}}$ is the residues from performing least square linear regressions between pairs of $\boldsymbol{x^b}$ variables (Daley, 1991). To see the transformation, suppose that $\boldsymbol{x^b}$ is arranged into blocks containing only one variable each, or,

$$\boldsymbol{x^b}^{\top} = \left[\boldsymbol{x^b}_1^{\top}, \boldsymbol{x^b}_2^{\top}, \cdots, \boldsymbol{x^b}_l^{\top}, \cdots, \boldsymbol{x^b}_{L_1}^{\top}, \boldsymbol{x^b}_L^{\top} \right],$$

and $\boldsymbol{x^{b'}}$ is set up in a similar fashion,

$$\boldsymbol{x^{b'}}^{\top} \equiv \left[\boldsymbol{x^{b'}}_1^{\top}, \boldsymbol{x^{b'}}_2^{\top}, \cdots, \boldsymbol{x^{b'}}_l^{\top}, \cdots, \boldsymbol{x^{b'}}_{L_1}^{\top}, \boldsymbol{x^{b'}}_L^{\top} \right].$$

[4] A detailed description of the CVT method and the underlying mathematics of the first transformation can be found in Appendix B

The first transformation step is defined as

$$\boldsymbol{x}^{\boldsymbol{b}'}{}_l(t) = \boldsymbol{x}^{\boldsymbol{b}}{}_l(t) - \sum_{\ell=1}^{l-1} \boldsymbol{\alpha}_{l,\ell} \boldsymbol{x}^{\boldsymbol{b}'}{}_\ell(t), \tag{1.5}$$

where,

$$\boldsymbol{\alpha}_{l,\ell} \equiv \left\langle \boldsymbol{x}^{\boldsymbol{b}}{}_l \left(\boldsymbol{x}^{\boldsymbol{b}'}{}_\ell\right)^\top \right\rangle \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_\ell \left(\boldsymbol{x}^{\boldsymbol{b}'}{}_\ell\right)^\top \right\rangle^{-1}, \quad \ell = 1,\, 2,\, \cdots,\, l-2,\, l-1. \tag{1.6}$$

In other words, $\boldsymbol{x}^{\boldsymbol{b}'}{}_l(t)$ is the residue from performing a least square linear regression of $\boldsymbol{x}^{\boldsymbol{b}}{}_l$ against $\boldsymbol{x}^{\boldsymbol{b}'}{}_1$, $\boldsymbol{x}^{\boldsymbol{b}}{}_2$, $\cdots$, $\boldsymbol{x}^{\boldsymbol{b}'}{}_{l-1}$, $\boldsymbol{x}^{\boldsymbol{b}'}{}_l$. This transformation guarantees that $\left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_l \left(\boldsymbol{x}^{\boldsymbol{b}'}{}_\ell\right)^\top \right\rangle$ is zero for all $l \neq \ell$.[5]

Regression coefficient matrices must be modelled The CVT transform alone does not solve the overwhelming size problem – for L variables, there are $(L^2 - L)/2$ regression coefficient matrices, each with dimensions of $N/L \times N/L$. Furthermore, the auto-covariance matrices must be invertible to determine $\boldsymbol{\alpha}_{l,\ell}$. Unfortunately, there are typically insufficient samples of $\boldsymbol{x}^{\boldsymbol{b}}$ to ensure that the $N/L \times N/L$ auto-covariance matrices are invertible (Bannister, 2008b). These severely restrict our ability to determine $\boldsymbol{\alpha}_{l,\ell}$. However, it is possible to model $\boldsymbol{\alpha}_{l,\ell}$ by making some assumptions about how the regression coefficients vary spatially in the model grid. This modelling is the subject of this project.

1.2.2 Specificity to region and season

Regional and seasonal dependence The roles of $\boldsymbol{B}$ described earlier clearly imply that $\boldsymbol{B}$ contains information about how background errors are organized and connected. That in turn depends on the meteorological features of the region considered. These features also exhibit seasonal variations. As such, $\boldsymbol{B}$ changes with the region and season of study. These two dependences must be considered in any study of $\boldsymbol{B}$ (Bouttier and Courtier, 2002; Kalnay, 2003; Bannister, 2008b).

Paucity of Maritime Continent $\boldsymbol{B}$ literature While there is a wealth of literature concerning $\boldsymbol{B}$ in the mid-latitudes and on the global scale, scant attention has been given to $\boldsymbol{B}$ at the Maritime Continent. However, there is ample motivation to study $\boldsymbol{B}$ in this region. First of all, there is socioeconomic value in such a study: Indonesia, Malaysia and Singapore have a combined population of 85 million (Badan Perencanaan

[5] See the proof for Eqn (B.1) in Appendix B for the origin of this zero property.

Pembangunan Nasional, 2013). Furthermore, two of the most commonly used regression coefficient models explicitly assume that the regression coefficients do not change with the horizontal position of any pair of points considered (Chen et al., 2013; Descombes et al., 2015). In other words, the models assume that the regression coefficients are horizontally invariant and isotropic. Given the complex topography of the Maritime Continent (Chang et al., 2016), these models may not hold, necessitating a study of $\boldsymbol{B}$ to test them and possibly formulate a new model. Unfortunately, to date, we have only been able to find a single publication concerning $\boldsymbol{B}$ in this specific region (Chen et al., 2013). There is clearly a need for a study of $\boldsymbol{B}$ in the Maritime Continent.

Paucity of Northeast monsoon $\boldsymbol{B}$ literature To make matters worse, the above-mentioned singular study was only performed during the Southwest monsoon season (Chen et al., 2013). The climate of equatorial Southeast Asia has two monsoon seasons (Chang et al., 2005; Fong and Ng, 2012): the Southwest monsoon (July to September), the Northeast monsoon (December to March)[6]. In other words, a study $\boldsymbol{B}$ in the Northeast monsoon season (NEM) of the Maritime Continent is warranted.

Northeast monsoon season The Northeast monsoon season (NEM) begins in December when the latitude of maximal solar heating is in the Maritime Continent-northern Australian region. This generates a pressure trough south of the Equator. The cooling of the northern Eurasian landmass prompts the formation of a surface, cold-core, high pressure system over Siberia (Ding, 1994). The pressure gradient drives the general near-surface northeasterly monsoon flow that characterizes the Northeast monsoon.

Cold surges Southward bursts of intense cold air have also been observed to emanate from this system when it moves towards the coastline of China and the western Pacific (Wang, 2006). These bursts, also known as cold surges, arrive at the Maritime Continent from the northeast, and amplify the near-surface northeasterlies (Fong and Ng, 2012) due to the orientation of the region's topography (Chang et al., 2005). The topography of the Maritime Continent also acts to channel the cold surge towards the Equator (Chang et al., 2005). These cold surges are typically moistened by their trajectories over the South China Sea, and are associated with enhanced convection in the Maritime Continent (see Figure 6 of Chang et al., 2005).

[6]The two monsoon seasons are separated by intermonsoon seasons.

1.2.3 Dependence on source of $\boldsymbol{X^b}$

$\boldsymbol{B}$ varies with source of $\boldsymbol{X^b}$ $\boldsymbol{B}$ is also dependent upon the source of educated guesses used for data assimilation. Consider the scenario where $\boldsymbol{X^b}$ comes from the region's climatology and another scenario where $\boldsymbol{X^b}$ is the outcome of a previous numerical forecast. Intuitively, the two $\boldsymbol{X^b}$ will exhibit different errors, and thus have different $\boldsymbol{B}$'s. Clearly, these dependences must be accounted for when studying $\boldsymbol{B}$.

SINGV $\boldsymbol{X^b}$ source Given this dependence, a source of $\boldsymbol{X^b}$ must be selected in this project. The Singapore Variable resolution model (SINGV) is a forecasting model developed collaboratively by the Centre for Climate Research Singapore (Meteorological Services Singapore) and the UK MetOffice to enhance the forecasting capabilities of Singapore. To further enhance Singapore's forecasting capabilities, a study of $\boldsymbol{B}$, using SINGV as its source, is needed.

1.3 Overview

Goal The goal of this project is to determine an appropriate model for the regression coefficient matrices involved in the CVT, over the Maritime Continent, during the Northeast monsoon season, for the Singapore Variable resolution model.

Wind field $\boldsymbol{B}$ As a first approach to this topic, we will only consider 2 forecast error variables: the velocity potential forecast error, ϕ, and the stream function forecast error, ψ, of the horizontal wind components . Their corresponding forecast variables are denoted by the corresponding capitalized Greek letters, Φ and Ψ. The two variables, and their errors, are related to the horizontal wind field (Bijlsma et al., 1986), and its error, via

$$\boldsymbol{U_H} = \boldsymbol{\nabla}_{\boldsymbol{H}}\Phi - \boldsymbol{\nabla}_{\boldsymbol{H}} \times (\Psi \hat{\boldsymbol{z}}), \tag{1.7}$$

$$\boldsymbol{u_H} = \boldsymbol{\nabla}_{\boldsymbol{H}}\phi - \boldsymbol{\nabla}_{\boldsymbol{H}} \times (\psi \hat{\boldsymbol{z}}), \tag{1.8}$$

where

$$\boldsymbol{\nabla}_{\boldsymbol{H}} \equiv \hat{\boldsymbol{x}}\frac{\partial}{\partial x} + \hat{\boldsymbol{y}}\frac{\partial}{\partial y}.$$

The gradient term and curl terms can be viewed as the diverging part of the wind (wind error) and rotating part of the wind (wind error) respectively.

CVT formulation As such, the state vector of forecast errors at time t is

$$\boldsymbol{x^b}_t = \left[\boldsymbol{\psi}_t^\top, \boldsymbol{\phi}_t^\top\right]^\top, \tag{1.9}$$

where $\boldsymbol{\psi}_t$ is a vector of $N/2$ elements containing the forecast errors of ψ at all positions on the forecast model grid at time t, and likewise for $\boldsymbol{\phi}$. The corresponding control variable transform for this study is

$$\begin{cases} \boldsymbol{\psi} = \boldsymbol{\psi'} \\ \boldsymbol{\phi'} = \boldsymbol{\phi} - \boldsymbol{\alpha}_{\phi,\psi}\boldsymbol{\psi'}. \end{cases} \tag{1.10}$$

Two commonly used $\boldsymbol{\alpha}_{\phi,\psi}$ models A total of three ways to model $\boldsymbol{\alpha}_{\phi,\psi}$ will be considered in this project. The first model assumes that ψ and ϕ are only related when they are on the same model level, and that their relationship is independent of the horizontal positions considered. The second model removes the same-level restriction, allowing for inter-model level relationships. However, the horizontal invariance assumption applies to the second model as well. The algorithms for these models are illustrated in Figures 1.1 and 1.2.

New $\boldsymbol{\alpha}_{\phi,\psi}$ model The third model will follow a completely different assumption: the evaluations of ϕ and ψ in the vicinity of a selected model point are equivalent to samples of the evaluations of ϕ and ψ at the said model point. The resulting third model can be thought of as a coarse-resolution version of $\boldsymbol{B}$, allowing both horizontal and vertical variations in the relationship between ψ and ϕ. The way which regression coefficients are computed and employed in this model are illustrated in Figures 1.3 and 1.4, respectively.

Overview – Setup and checks We will begin with explaining how $\boldsymbol{x^b}$ can be determined by using pairs of forecasts, via the National Meteorology Center (NMC) method, in Chapter 2. Following that, Chapter 2 will detail the process of SINGV data generation and subsequent processing to produce $\boldsymbol{x^b}$. The SINGV forecast outputs will then be checked to ensure that it is capturing the general characteristics of the Northeast monsoon in Chapter 3.

Overview – modelling $\boldsymbol{\alpha}_{\boldsymbol{\phi\psi}}$ Upon establishing the sanity of the SINGV model, we will then examine some of the forecast error covariance features estimated by the NMC method in Chapter 4[7]. The negative implications of the observed features for the two commonly used models (shown in Figures 1.1 and 1.2) will then be elaborated. In response, Chapter

[7]Essentially, we will be looking at columns of a $\boldsymbol{B}$ estimated by the NMC.

4 will formulate new way of modelling regression coefficient matrices (the third model illustrated in Figures 1.3 and 1.4). Chapter 5 will then compare the performance of two commonly used models against that of the new model, and show that the new model far outperforms the former two. We will then conclude with some comments and directions for future work in Chapter 6.

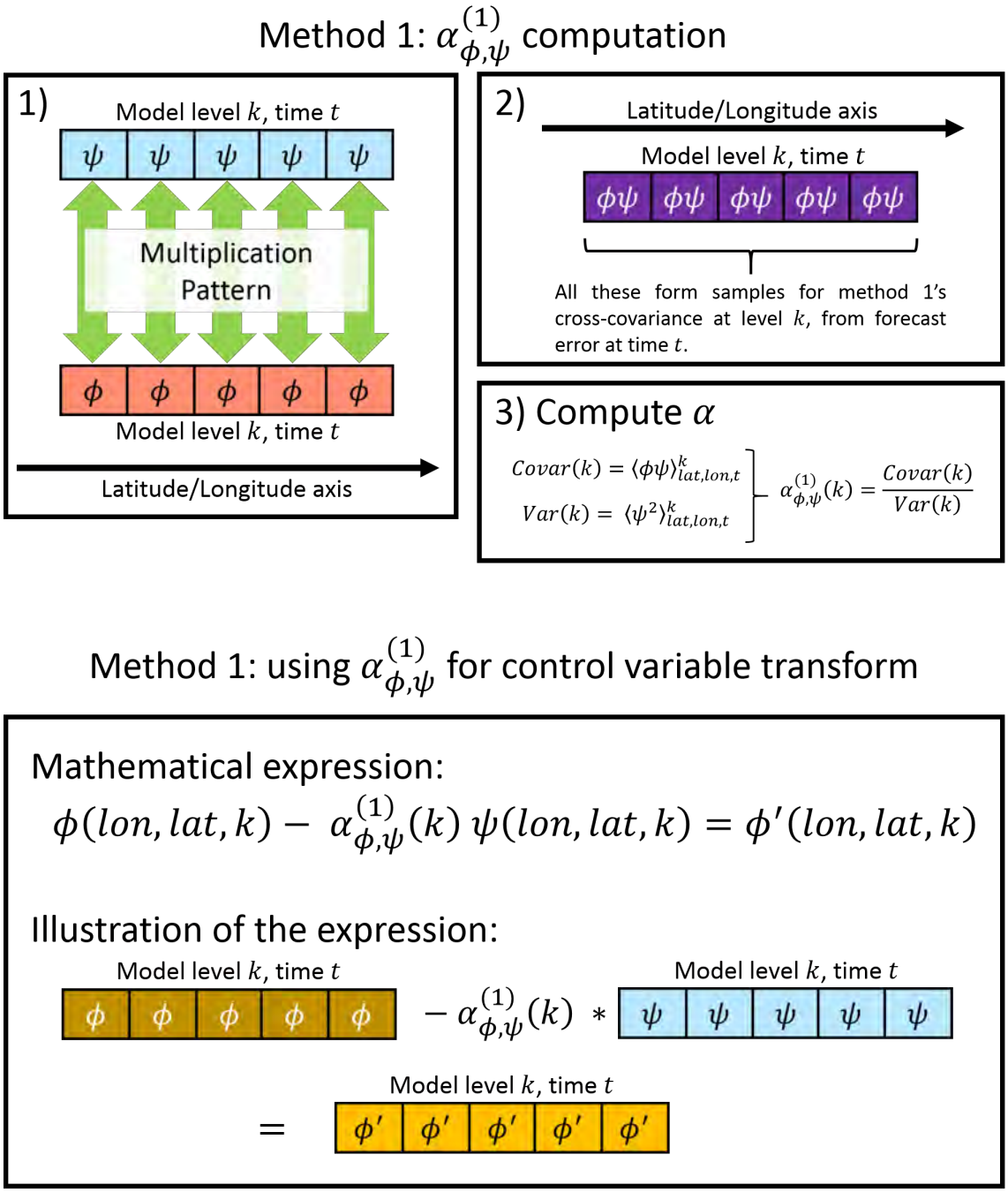
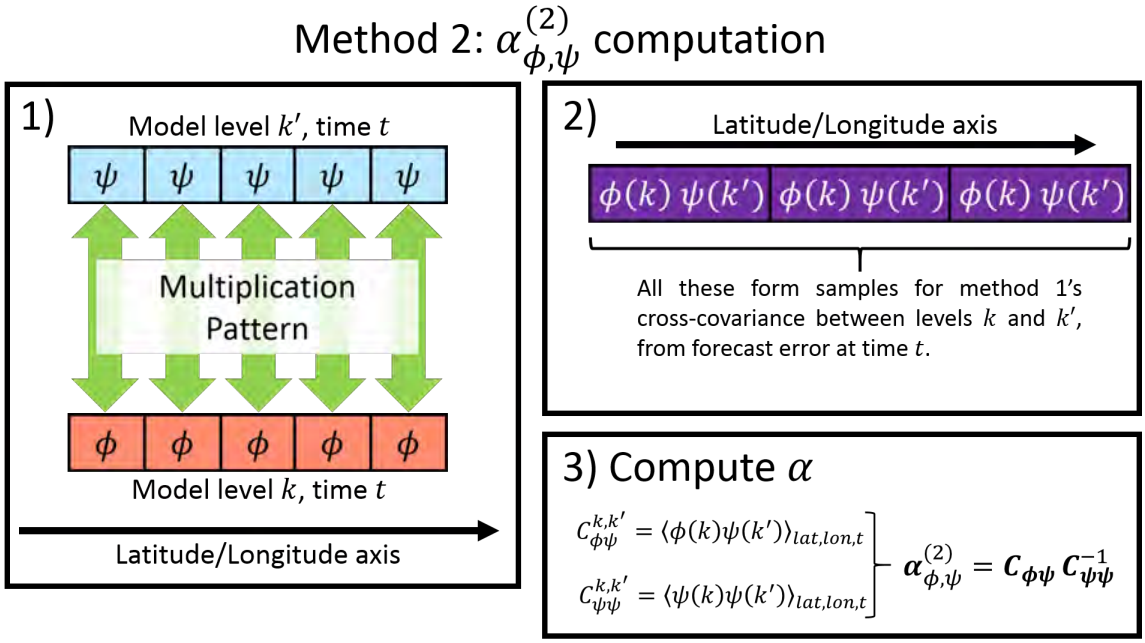


Figure 1.1: Illustration of how the first model computes regression coefficients (top) and how it is utilized for the control variable transform (bottom). Each colored box represents a model grid point with the indicated error variable indicated. The panels are to be read in sequence of the labelled numbers. This model assumes that ψ and ϕ are disconnected across model levels and their relationship on the same model level is invariant of horizontal position. In other words, there are only N_Z regression coefficients, one for each model level. Note that *lon*, *lat* and k refer to the longitude, latitude and model level of the model grid. The subscripts appended to the averages indicate the dimensions of averaging.



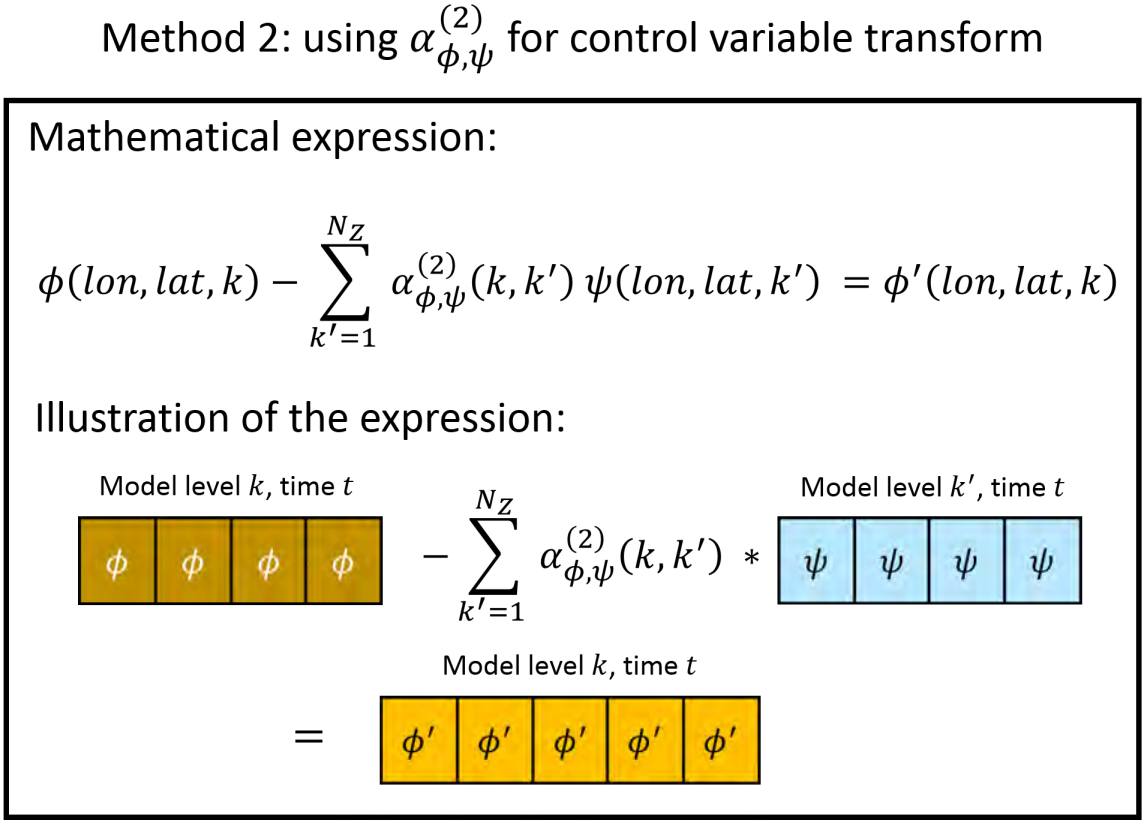


Figure 1.2: Illustration of how the second model computes regression coefficients (top) and how it is utilized for the control variable transform (bottom). Each colored box represents a model grid point with the indicated error variable indicated. The panels are to be read in sequence of the labelled numbers. This model assumes that the relationship between ψ and ϕ is invariant of horizontal position. In other words, there are only N_Z regression coefficients, one for each model level. Note that *lon*, *lat* and k refer to the longitude, latitude and model level of the model grid. The subscripts appended to the averages indicate the dimensions of averaging.

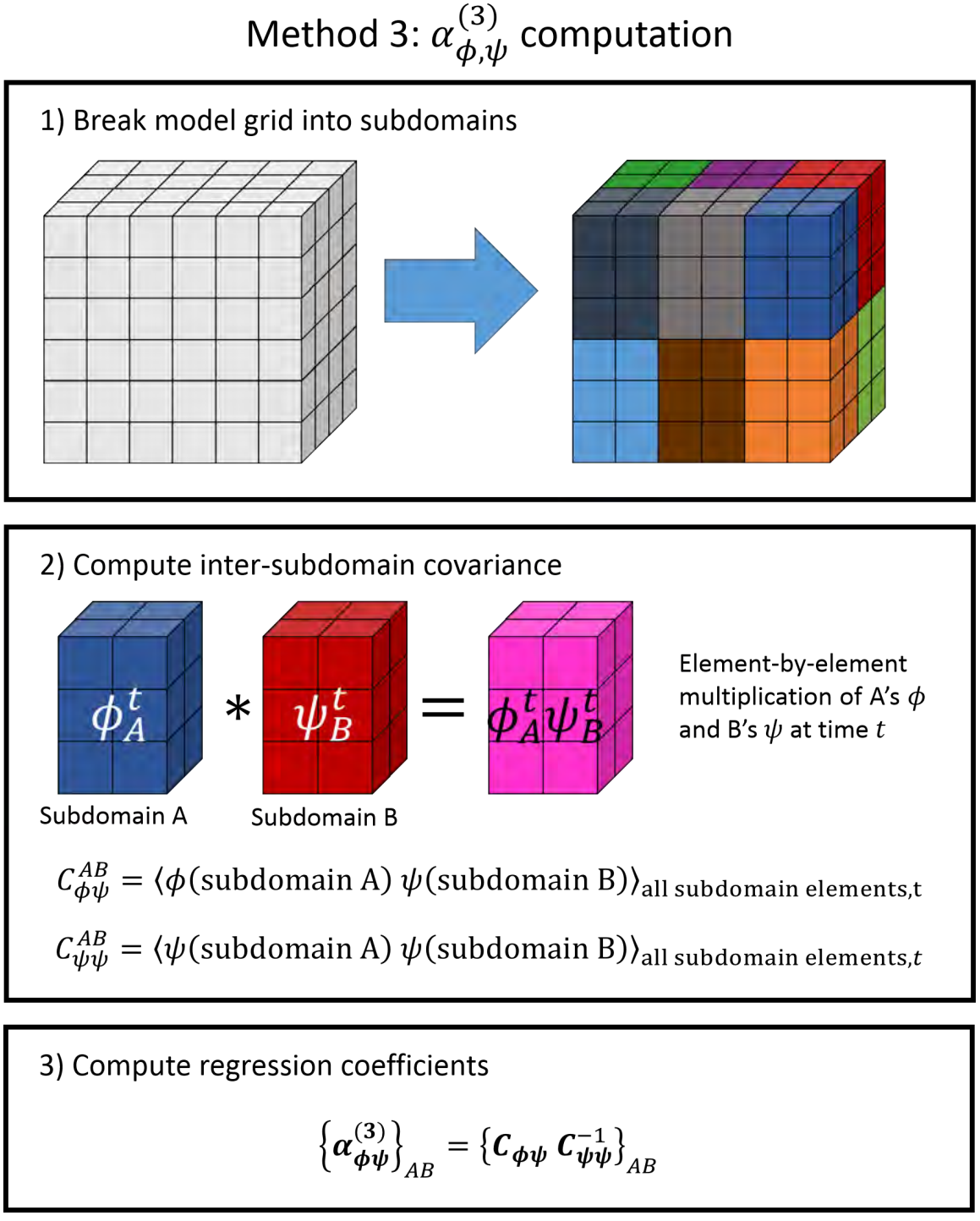


Figure 1.3: Illustration of how the third model computes regression coefficients. The panels are to be read in sequence of the labelled numbers. Each colored box represents a model grid point with the indicated error variable indicated. The panels are to be read in sequence of the labelled numbers. This model assumes that within a subdomain, all values of ψ and ϕ can be treated as samples of the center of the subdomain. The subscripts appended to the averages indicate the dimensions of averaging.

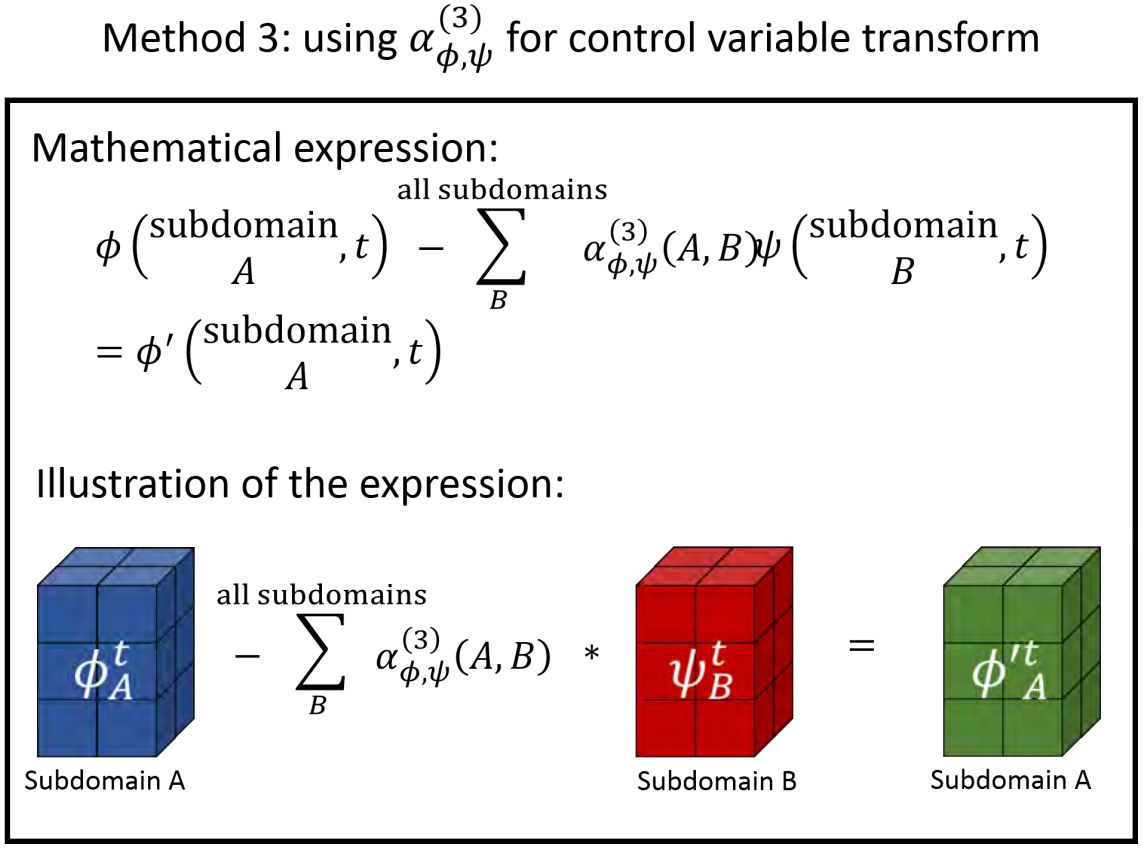


Figure 1.4: Illustration of how the third model is used in the control variable transform. The panels are to be read in sequence of the labelled numbers. Each colored box represents a model grid point with the indicated error variable indicated. The panels are to be read in sequence of the labelled numbers. This model assumes that within a subdomain, all values of ψ and ϕ can be treated as samples of the center of the subdomain.

Chapter 2

Materials and methods

Overview In this chapter, we will explain how $\boldsymbol{x^b}$ can be estimated using the National Meteorology Center method, in spite of the unknown $\boldsymbol{X^t}$. Following that, the setup of the Singapore Variable resolution model (SINGV) will be discussed. During that discussion, the exact manner which the National Meteorology Centre method is applied will also be elaborated. To obtain the velocity potential and stream functions, we utilized the GEN_BE package created by the National Corporation for Atmospheric Research and the UK Met Office.

2.1 Overcoming the X^t requirement with the NMC method

Purpose The National Meteorology Centre (NMC) method mitigates the requirement of the unobtainable $\boldsymbol{X^t}$ in any computation of $\boldsymbol{B}$ (Parrish and Derber, 1992). In essence, if two different $\boldsymbol{X^b}$ from the same source can be obtained for the same time, t, $\boldsymbol{B}$ can be approximated by

$$\boldsymbol{B} \approx \frac{1}{2N_T} \sum_{t=1}^{N_T} \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} - \boldsymbol{\mu} \right) \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} - \boldsymbol{\mu} \right)^\top, \tag{2.1}$$

where,

$$\boldsymbol{\mu} \equiv \frac{1}{N_T} \sum_{t=1}^{N_T} \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} \right).$$

Assumptions of NMC method The NMC method operates under under two assumptions. First of all, it is assumed that the auto-covariance of the two background errors are equivalent on average. In other words,

$$\sum_{t=1}^{N_T} \left(\boldsymbol{x^b}_{(1),t} \, \boldsymbol{x^b}^{\top}_{(1),t} \right) \approx \sum_{t=1}^{N_T} \left(\boldsymbol{x^b}_{(2),t} \, \boldsymbol{x^b}^{\top}_{(2),t} \right),$$

where $\boldsymbol{x^b}_{(1),t}$ and $\boldsymbol{x^b}_{(2),t}$ are the errors of $\boldsymbol{X^b}_{(1),t}$ and $\boldsymbol{X^b}_{(2),t}$ respectively. The second assumption is that $\boldsymbol{x^b}_{(1),t}$ and $\boldsymbol{x^b}_{(2),t}$ are mutually uncorrelated, or,

$$\sum_{t=1}^{N_T} \left(\boldsymbol{x^b}_{(1),t} \, \boldsymbol{x^b}^{\top}_{(2),t} \right) = \sum_{t=1}^{N_T} \left(\boldsymbol{x^b}_{(2),t} \, \boldsymbol{x^b}^{\top}_{(1),t} \right) = \boldsymbol{0}.$$

NMC derivation The NMC method in Eqn (2.1) can be shown to be consistent with the definition of $\boldsymbol{B}$ in Eqn (1.2) under these assumptions. Starting from Eqn (2.1), and drawing on the fact that $\boldsymbol{x^b}_{(1),t} \equiv \boldsymbol{X^b}_{(1),t} - \boldsymbol{X^t}_t$ and likewise for $\boldsymbol{x^b}_{(2),t}$,

$$\begin{aligned}
&\frac{1}{N_T} \sum_{t=1}^{N-1} \left\{ \boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} \right\} \left\{ \boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} \right\}^{\top} \\
&= \frac{1}{N_T} \sum_{t=1}^{N-1} \left\{ \left(\boldsymbol{x^b}_{(1),t} + \boldsymbol{X^t}_t - \boldsymbol{x^b}_{(2),t} - \boldsymbol{X^t}_t \right) \left(\boldsymbol{x^b}_{(1),t} + \boldsymbol{X^t}_t - \boldsymbol{x^b}_{(2),t} - \boldsymbol{X^t}_t \right)^{\top} \right\} \\
&= \frac{1}{N_T} \sum_{t=1}^{N-1} \left\{ \left(\boldsymbol{x^b}_{(1),t} - \boldsymbol{x^b}_{(2),t} \right) \left(\boldsymbol{x^b}_{(1),t} - \boldsymbol{x^b}_{(2),t} \right)^{\top} \right\} \\
&= \frac{1}{N_T} \sum_{t=1}^{N-1} \left\{ \boldsymbol{x^b}_{(1),t} \, \boldsymbol{x^b}^{\top}_{(1),t} - \boldsymbol{x^b}_{(1),t} \, \boldsymbol{x^b}^{\top}_{(2),t} - \boldsymbol{x^b}_{(2),t} \, \boldsymbol{x^b}^{\top}_{(1),t} + \boldsymbol{x^b}_{(2),t} \, \boldsymbol{x^b}^{\top}_{(2),t} \right\} \\
&\approx \frac{1}{N_T} \sum_{t=1}^{N-1} \left\{ \boldsymbol{x^b}_{(1),t} \, \boldsymbol{x^b}^{\top}_{(1),t} + \boldsymbol{x^b}_{(2),t} \, \boldsymbol{x^b}^{\top}_{(2),t} \right\} \approx 2 \sum_{t=1}^{N-1} \left\{ \boldsymbol{x^b}_{(1),t} \, \boldsymbol{x^b}^{\top}_{(1),t} \right\} \approx 2\boldsymbol{B}.
\end{aligned}$$

Modification due to bias There is a hidden pitfall in using the NMC method: the average of $\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t}$ may not be zero. As the general definition of a covariance matrix calls for mean removal before computing the covariance, the NMC method must be modified. The NMC method employed in this project is thus (Descombes et al., 2015):

$$\boldsymbol{B} \approx \frac{1}{2N_T} \sum_{t=1}^{N_T} \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} - \boldsymbol{\mu} \right) \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} - \boldsymbol{\mu} \right)^{\top},$$

where,

$$\boldsymbol{\mu} \equiv \frac{1}{N_T} \sum_{t=1}^{N_T} \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} \right).$$

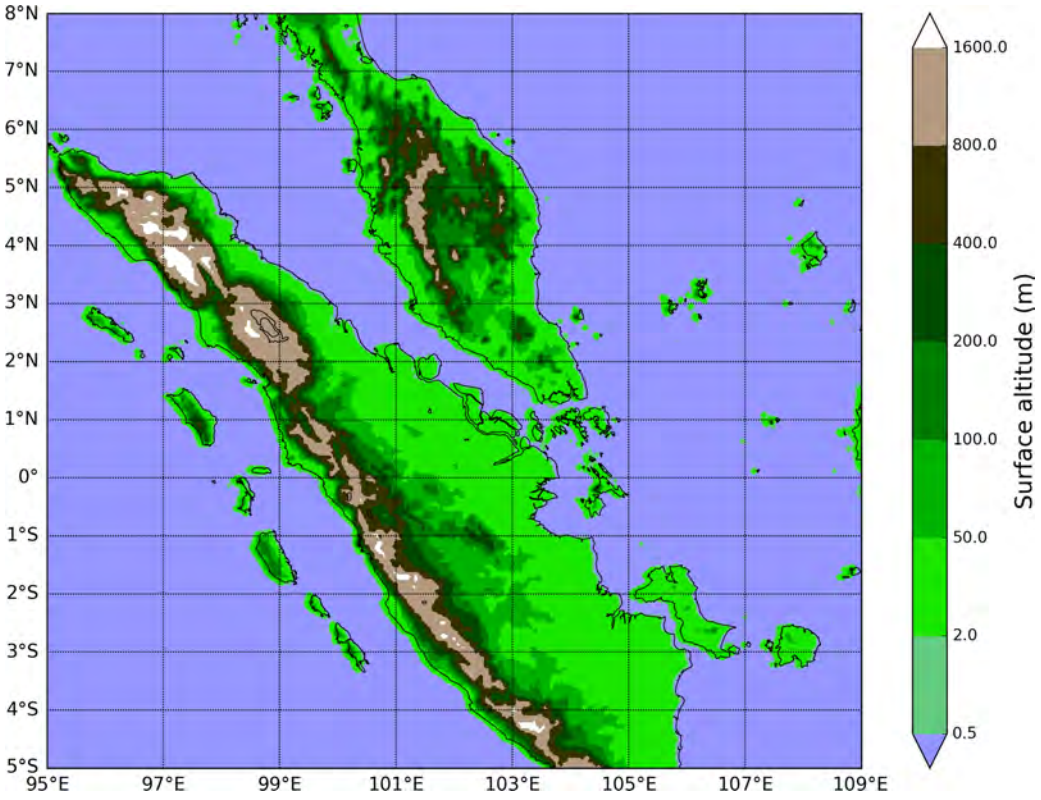


Figure 2.1: Plot of the surface heights of the SINGV simulation domain.

This also means that the background error at time t, $\boldsymbol{x^b}_t$, can be viewed as:

$$\boldsymbol{x^b}_t \approx \frac{1}{\sqrt{2}} \left(\boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} - \left\langle \boldsymbol{X^b}_{(1),t} - \boldsymbol{X^b}_{(2),t} \right\rangle \right).$$

Forecast perturbation-based NMC Since the source of $\boldsymbol{X^b}$ in this project is the Singapore Variable resolution model forecasts (SINGV), the NMC requires pairs of forecasts valid at the same time to form $\boldsymbol{B}$ (Parrish and Derber, 1992; Bannister, 2008b). If the time of data assimilation is T hours, then SINGV forecasts were initiated 12 hours and 6 hours before T, and their outputs at T are used as as the two $\boldsymbol{X^b}$ at time T.

2.2 SINGV model setup

Background and computational resource This study employed the Singapore Variable Resolution Model, version 2.1 (SINGV), which is modified from the UK MetOffice's (UKMO) Unified Model (UM). The SINGV v2.1 is produced by a multi-year collaboration project between the UK MetOffice and the Center for Climate Research Singapore to build a tropical convective-scale NWP for Singapore and the surrounding region. The model was run using the Center for Climate Research Singapore's High Performance Com-

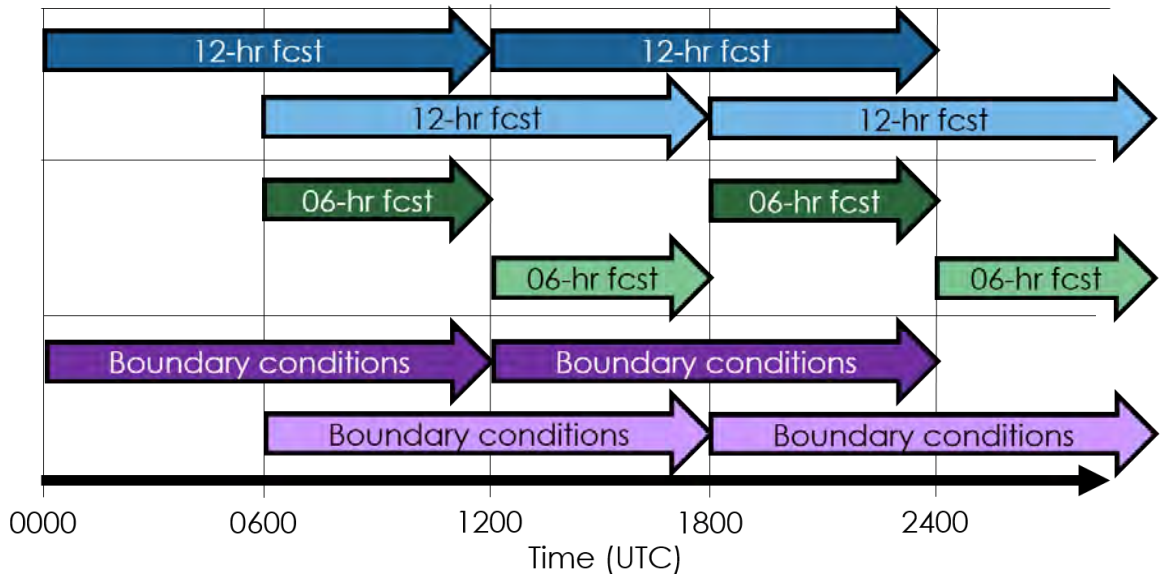


Figure 2.2: Illustration of how pairs of forecasts are set up, to be utilized by the NMC method. See text for the complementing description.

puting Cluster, Athena. Forecasts from the SINGV model were obtained over 15 days of the Northeast monsoon of 2015, from 5th December 2015, 0000 UTC, to 19th December 2015, 1800 UTC.

Model grid setup The simulation domain spans over a geographical region from 94.543°E to 109.2445°E and 5.74°S to 8.2325°N, with a horizontal spatial resolution of 4.5 km[1]. The model employed a total of 80 static, terrain-following, hybrid height model levels. The orography of the simulation domain has been plotted out in Figure 2.1.

Employing the NMC method As mentioned earlier, the NMC method can be used to sidestep the $\boldsymbol{X^t}$ requirement in computing $\boldsymbol{B}$. Pairs of $\boldsymbol{X^b}$ valid at the same times are needed for that approximation. In the case of SINGV, pairs of 12-hour and 6-hour forecasts were used. In other words, the NMC estimate of $\boldsymbol{x^b}_t$ is

$$\boldsymbol{x^b}_t \approx \frac{1}{\sqrt{2}} \left(\boldsymbol{X^b}_{\text{12-hr},t} - \boldsymbol{X^b}_{\text{06-hr},t} - \boldsymbol{\mu} \right), \quad \boldsymbol{\mu} \equiv \frac{1}{N_T} \sum_{t=1}^{N_T} \left(\boldsymbol{X^b}_{\text{12-hr},t} - \boldsymbol{X^b}_{\text{06-hr},t} \right) \tag{2.2}$$

We will use Figure 2.2 to illustrate the how the forecast pairs were set up. Suppose that a 12-hour forecast was initiated at 0000 UTC. Then to obtain the second forecast for the NMC, a 06-hour forecast was initiated at 0600 UTC. The initial conditions used for both forecasts were taken from the UM Global analysis data, released at 0000 UTC and 0600 UTC respectively. Both forecasts utilized the same set of hourly boundaries obtained

[1]Horizontal grid arrangement: 364 (longitudinal) by 346 (latitudinal).

from the 36-hour UM Global that was released at 0000 UTC. The pair of forecasts at 1200 UTC forms the first forecast pair for the NMC. This process is repeated at 0600 UTC, 1200 UTC, 1800 UTC, and so forth, every 6 hours. Note that this method of matching boundary conditions is known to enhance the accuracy of the NMC method (Berre et al., 2006; Bannister, 2008a).

Initial and boundary conditions The boundary and initial conditions for SINGV were taken from the UM Global 36-hour global forecasts. These global forecasts are performed daily at 00, 06, 12 and 18 hours UTC and are released to the Center for Climate Research Singapore upon completion. The SINGV system then performs a vertical linear and horizontal bilinear interpolation to pass the UM Global data into the aforementioned SINGV grid.

2.3 GEN_BE processing

2.3.1 Pre-GEN_BE processing

Need for pre-GEN_BE processing To use the GEN_BE package, the forecasts must be converted from the UK Met Office PP files into the Weather Research and Forecasting model NetCDF output files. Aside from the file format difference, the SINGV model uses a height-based, terrain following, vertical coordinate system, whereas the WRF model uses a pressure-based, terrain following, vertical coordinate system (η levels). The η level value of a point with pressure P^{WRF} with a model top pressure P^{top} and model surface pressure P^{sfc} is:

$$\eta = \frac{P^{\mathrm{WRF}} - P^{\mathrm{t}}}{P^{\mathrm{sfc}} - P^{\mathrm{t}}}$$

Hybrid height to η-level transform The WRF Preprocessing system is capable of transforming fields from hybrid height coordinates to η levels via vertical linear interpolation (metgrid step). To maximize interpolation accuracy, η levels that minimize the difference between η level pressure and hybrid height pressure were used. *I.e.*,

$$\frac{\partial}{\partial \eta_k}\left(P_{ijk}^{\mathrm{SINGV}} - P_{ijk}^{\mathrm{WRF}}\right)^2 = 0 \tag{2.3}$$

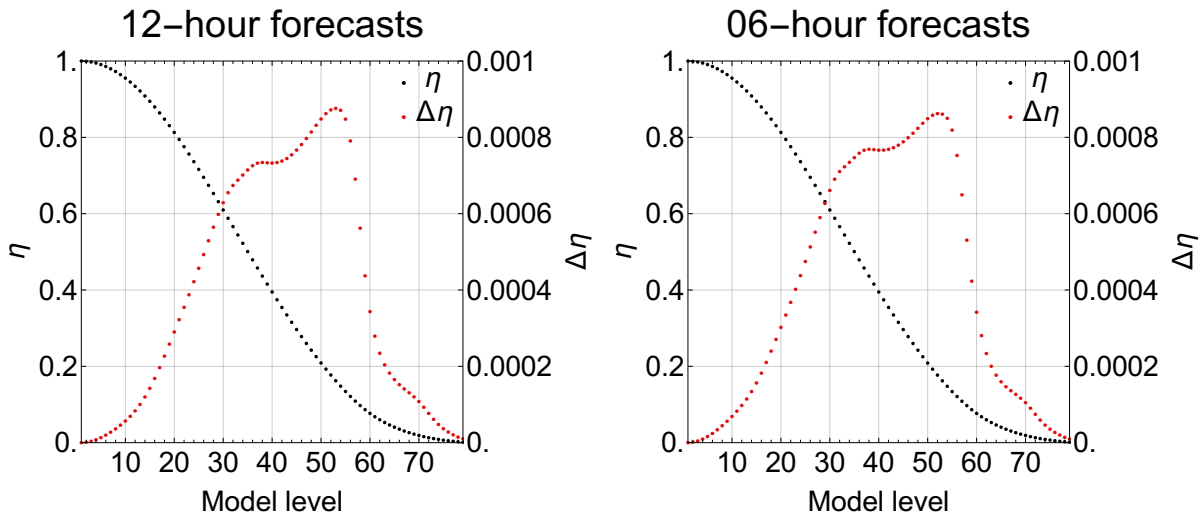


Figure 2.3: Plots of the average and standard deviation of η with respect to each model height, for the 12-hour forecast (left) and the 6-hour forecast (right).

Since $\eta = \dfrac{P^{\text{WRF}} - P^{\text{top}}}{P^{\text{sfc}} - P^{\text{top}}}$ implies $P^{\text{WRF}}_{ijk} = \eta_k \left(P^{\text{sfc}}_{ij} - P^{\text{top}}_{ij}\right) + P^{\text{top}}_{ij}$, the set of η values that solves Eqn (2.3) is thus found to be:

$$\eta_k = \frac{P^{\text{top}}_{ij} P^{\text{top}}_{ij} - P^{\text{top}}_{ij} P^{\text{SINGV}}_{ijk} - P^{\text{sfc}}_{ij} P^{\text{top}}_{ij} + P^{\text{sfc}}_{ij} P^{\text{SINGV}}_{ijk}}{P^{\text{top}}_{ij} P^{\text{top}}_{ij} - 2P^{\text{top}}_{ij} P^{\text{sfc}}_{ij} + P^{\text{sfc}}_{ij} P^{\text{sfc}}_{ij}} \tag{2.4}$$

η-levels are similar across 54 forecasts Surprisingly, despite the fact that Eqn (2.4) does not guarantee that the η values of each forecast set will be different, the η values ultimately turned out to be virtually identical. The average η value, and the standard deviation of η, across model levels, for both the 12-hour forecast (unstaggered) and the 6-hour forecast (staggered) are as shown in Figure 2.3.

ARW-WRF conversion After the η level and WPS procedure, the WRF model was run for 0 time steps to produce the SINGV data in the appropriate NetCDF format with the WRF model variables.

2.3.2 GEN_BE

Uses of GEN_BE The GEN_BE package was utilized on the WRF-processed forecast pairs to produce the velocity potential (Φ) and stream function (Ψ) for each pair. GEN_BE then follows Eqn (2.2) to produce the NMC-estimated forecast error.

Φ and Ψ determination As mentioned in the introduction, Φ and Ψ are defined for the horizontal wind field. In other words, for the k-th model level, Φ and Ψ are related to the horizontal wind field at level k, $\boldsymbol{U}_k$, by (Descombes et al., 2015)

$$\boldsymbol{\nabla_H} \cdot \boldsymbol{U}_k(x,y) = \nabla_H^2 \Phi(x,y,k) \ , \ \{\boldsymbol{\nabla_H} \times \boldsymbol{U}_k(x,y)\} \cdot \hat{\boldsymbol{z}} = \nabla_H^2 \Psi(x,y,k) \tag{2.5}$$

where

$$\boldsymbol{U}_k(x,y) \equiv U_{x,k}(x,y)\,\hat{\boldsymbol{x}} + U_{y,k}(x,y)\,\hat{\boldsymbol{y}} \ , \ \boldsymbol{\nabla_H} \equiv \hat{\boldsymbol{x}}\frac{\partial}{\partial x} + \hat{\boldsymbol{y}}\frac{\partial}{\partial y},$$

and $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ are the usual right-handed Cartesian unit vectors. The GEN_BE package solves Eqn (2.5) using a discrete sine transform, with zero boundary conditions. It then utilizes Eqn (2.2) to determine the forecast errors at time t, $\boldsymbol{x^b}_t$.

Removal surface level and boundaries The GEN_BE removes the first model level during its processing. Also, to prevent the boundary conditions of SINGV coming from the UM from interfering with the analysis, 18 points are cropped from both the north and south boundaries of the model grid and 17 points are cropped from both the east and west boundaries. This turns the model grid from a $364 \times 346 \times 80$ cuboid lattices of points (in longitude, latitude, vertical order) to $330 \times 310 \times 79$.

Chapter 3

Sanity checking the SINGV outputs

Purpose of chapter Before examining the features of the forecast error, it is important to ensure that the SINGV outputs display the general features of the Northeast monsoon. This acts as a sanity check to ensure that SINGV is behaving as expected. As such, we will now examine the synoptic features of the SINGV data set.

Overview This chapter will start with examining and confirming that the time-averaged thermodynamic and horizontal wind fields of the SINGV forecasts are as expected of the Northeast monsoon season. Afterwards, the forecasts will be examined for cold surges (important feature in the Northeast monsoon) as a further confirmation that the model is capturing the key Northeast monsoon features. We will then comment on an intriguing, short-lived, low-level, counter-clockwise circulation feature observed in the forecasts.

Hybrid height model levels will be used Before discussing the features of the SINGV dataset, note that the data in this section will be plotted based on the hybrid height model levels used in the SINGV model. However, the model level index will not be stated: instead, the rough atmospheric pressure level that corresponds to that model level will be stated. The correspondence between the model level and atmospheric pressure, and between model level and geopotential height, are as plotted in Figures 3.1 and 3.2.

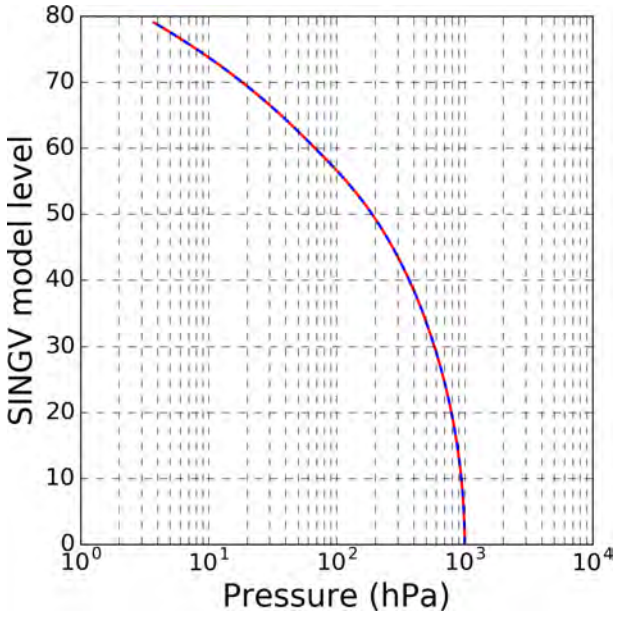


Figure 3.1: Plots of temporally and horizontally averaged pressure with respect to SINGV model levels. The solid red and sashed blue lines respectively indicate the averaged pressure profile of the 12-hour and 06-hour forecasts. Note that we have actually shaded one standard deviation on the left and right of pressure curve (see Figure 3.2 for example). However, the standard deviations are so small that they cannot be distinguished from the average curves. Also, both pressure profiles are virtually identical.

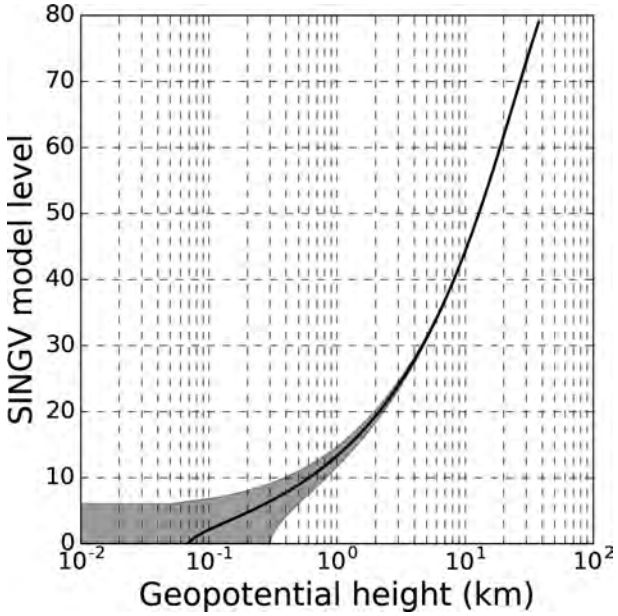


Figure 3.2: Plot of horizontally averaged geopotential height with respect to SINGV model levels. The shaded grey area indicates one standard deviation on either side of the averaged curve, per model level. Note that only one geopotential height curve is needed as the geopotential height of each model level remains the same across all forecasts in this study.

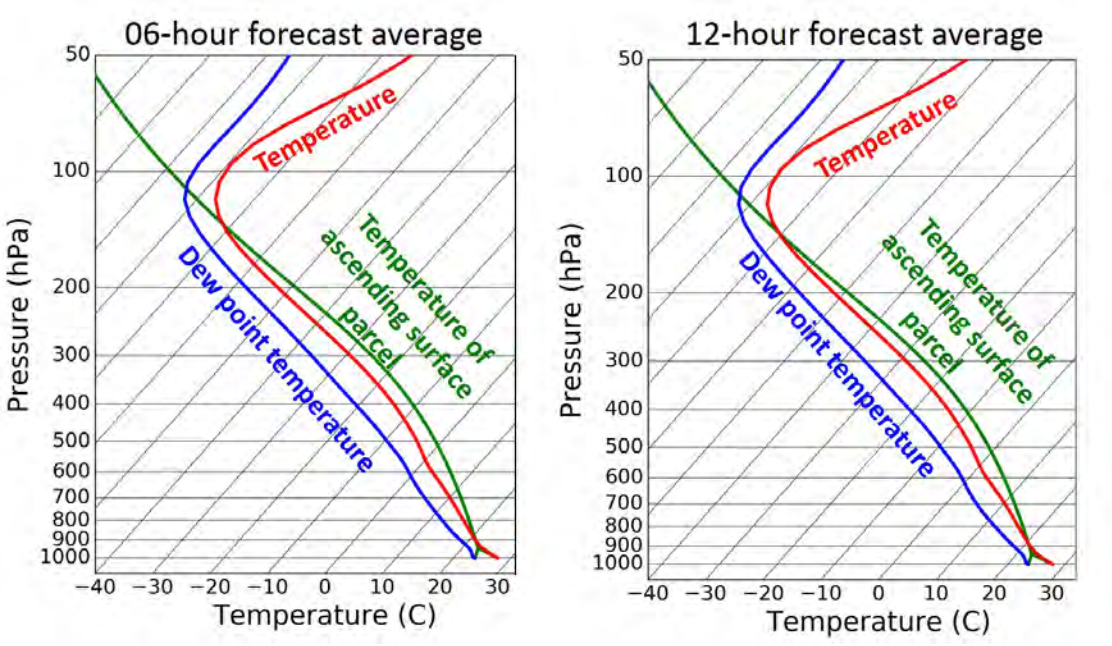


Figure 3.3: SkewT-logP plots of the two forecast sets (left: 06-hour forecasts, right: 12-hour forecasts), obtained by averaging pressure, specific humidity and temperature for every model level. The average is performed across the horizontal domain of each model level, across all forecasts with the same length. Temperature and dewpoint temperature profiles are as indicated by the solid red and blue lines respectively. The solid green curve indicates the profile of a surface air parcel as it ascends adiabatically. Note that the base of the green curve actually coincides with the base of the red curve. The green "leg" jutting out from the base of the blue curve traces the isotherm of the surface dew point temperature as surface parcel ascends. The place where the "leg" and the green curve meet is where the relative humidity of the surface parcel has reached 100%.

3.1 Time-averaged thermodynamic profile is as expected

Tropopause pressure level is correct SkewT-logP diagrams are useful for examining the likelihood of convection and the type of convection in the atmosphere. In Figure 3.3, skewT-logP[1] diagrams of both forecast sets are shown. The plotted profiles are constructed by taking the horizontal and temporal average of temperature, specific humidity and pressure for each model level. It is clear that the temperature minima occurs slightly above the 100 hPa level, suggesting that the tropopause resides roughly at that level. This is as expected for an equatorial region.

[1]These skewT-logP plots are produced with a modified version of the `SkewT` package in Python, released by Thomas Chubb, Department of Mathematical Sciences, Monash University, copyrighted 2014.

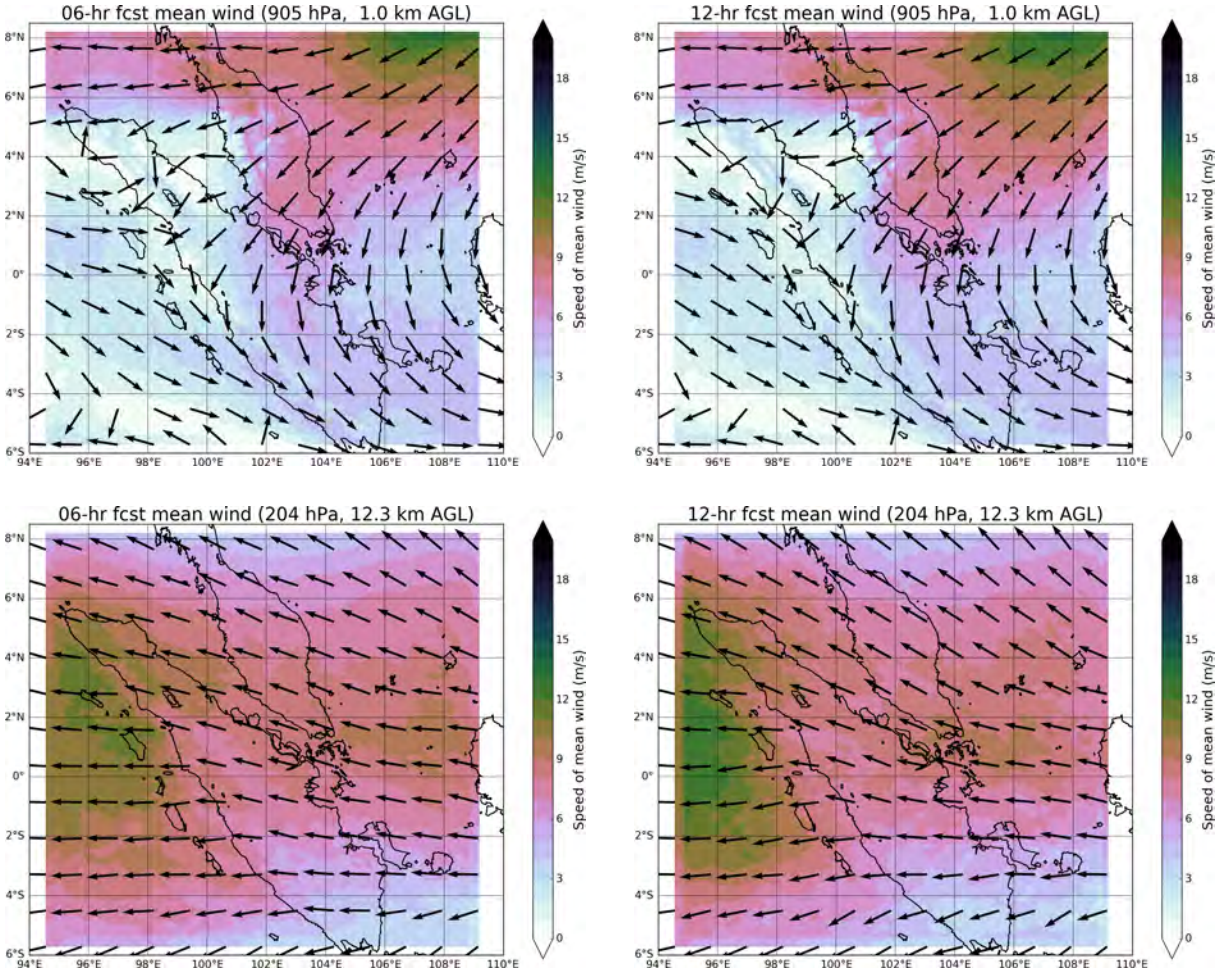


Figure 3.4: Contour plots of the time-averaged wind of the 06-hour forecasts (left column) and the 12-hour forecasts (right column), overlaid with arrows indicating the direction of wind flow, at the SINGV model levels with average pressures of 900 hPa (top) and 200 hPa (bottom). Note that the arrows only indicate the time-averaged wind direction, meaning that all arrows have the same lengths.

Atmosphere is unstable Figure 3.3 also shows that the atmosphere for both forecast sets are unstable and prime for deep convection. While the temperature of a surface air parcel undergoing dry adiabatic ascension is cooler than the environmental temperature for the bottom 100 hPa of the atmosphere, the difference is rather slight. In other words, only a small amount of energy is needed for the parcel to overcome the barrier to adiabatic ascension in the first 100 hPa of the atmosphere. Once the parcel passes the 900 hPa level (which is the level of free convection), the parcel ascends via a moist adiabat and remains continuously warmer than the atmosphere, allowing it to freely ascend, until it reaches the 130 hPa level. Clearly, the atmosphere is rather unstable. The closeness of this level to the tropopause indicates that the atmosphere is prime for deep convection. This is expected for the Maritime Continent during the Northeast monsoon (Chang et al., 2005).

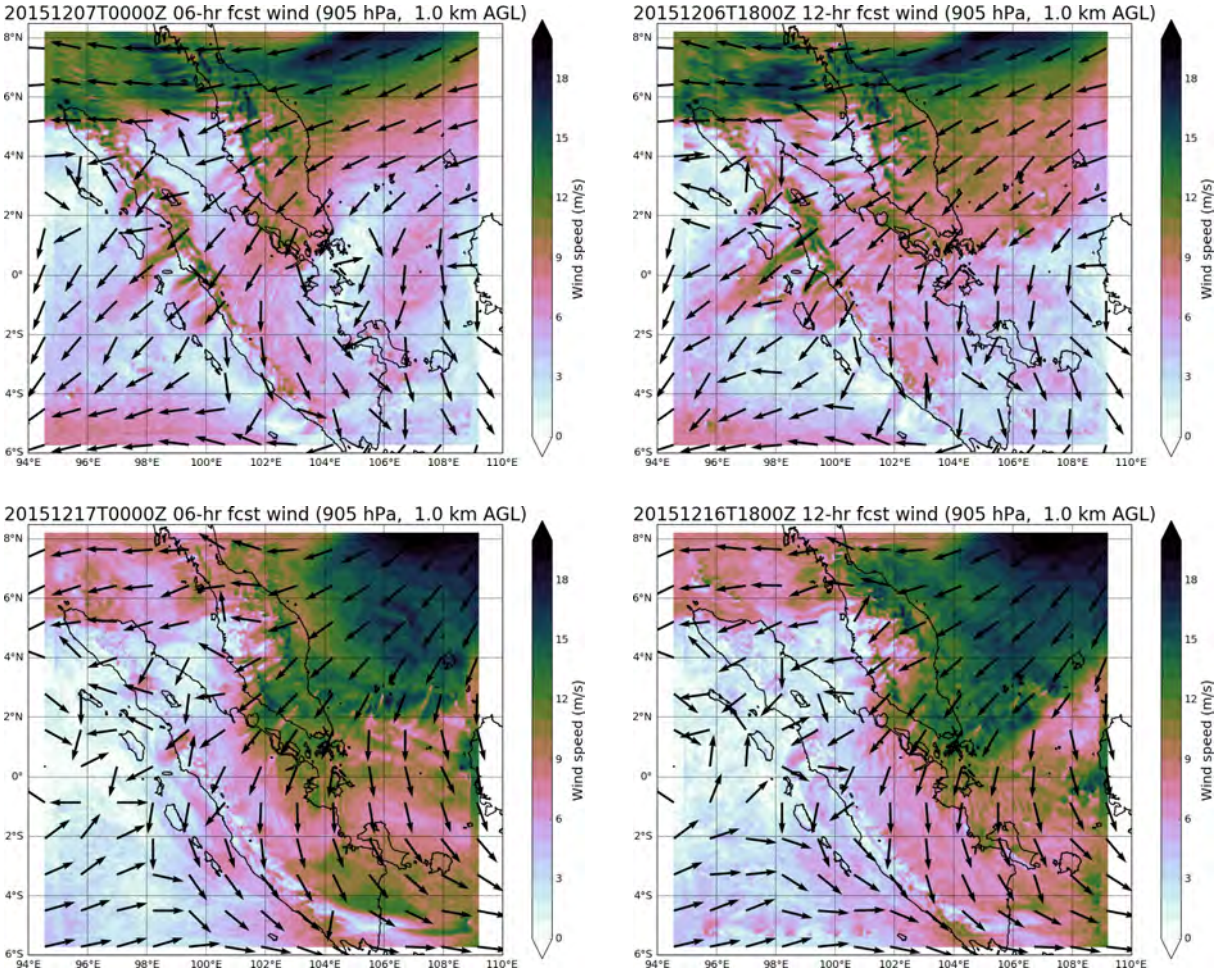


Figure 3.5: Same type of plots as Figure 3.4, but on the SINGV level corresponding to the 900 hPa. The data shown are snapshots of the cold surges at 0600 UTC, 7th December (top), and 0600 UTC, 17th December (bottom), observed in the 06-hour forecasts (left) and 12-hour forecasts (right).

3.2 900 hPa monsoon flow is as expected

Observed monsoon flow features Aside from that, on average, the 06-hour forecasts and the 12-hour forecast both show a low-level northeasterly wind entering the domain from the northeastern corner of the domain (Figure 3.4, 900 hPa plots). This wind flow is deflected leftwards as it approaches and crosses the Equator, resulting in a northwesterly low-level wind exiting the domain at the southeastern corner. This monsoon flow pattern is as expected from the monsoon's historical climatology (Chang et al., 2005; Fong and Ng, 2012).[2]

[2]The monsoon flow pattern is explained by the paragraph labelled "Northeast monsoon season" in the introduction.

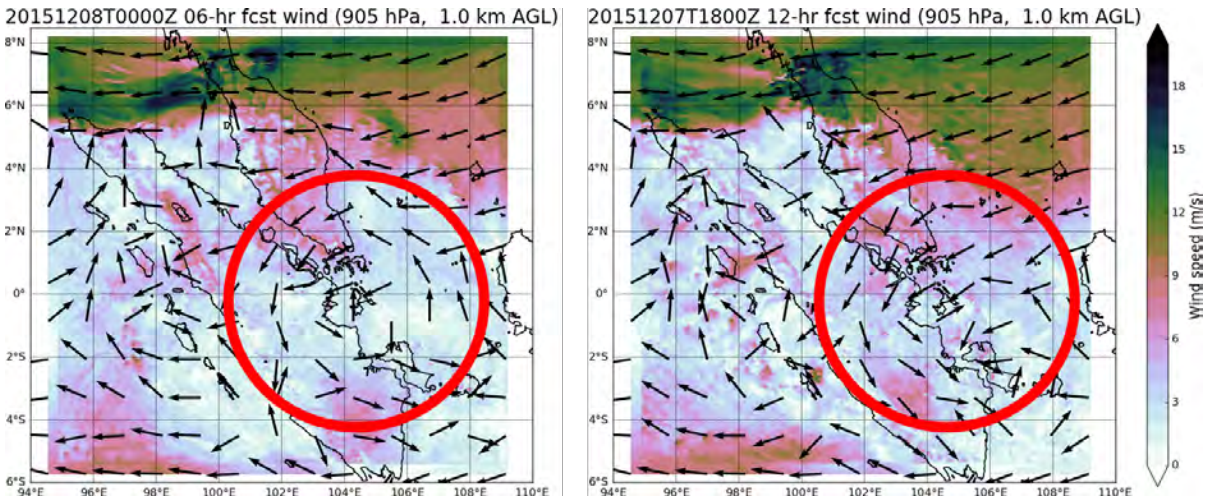


Figure 3.6: Same type of plots as Figure 3.4, but on the SINGV level corresponding to the 900 hPa. The data shows the Borneo vortex that appeared in the 06-hour forecast (left) and 12-hour forecast (right) valid on 0600 UTC, 8th December 2015. The vortex shows up as a closed, counterclockwise wind direction pattern (red circles) in both forecasts.

3.3 Cold surges manifested in the simulations

Cold surges observed Another important feature observed would be cold surges – bursts of low-level cold air that occasionally emanate from the Siberian surface high pressure system. The cold surge is a characteristic of the Northeast monsoon (Fong and Ng, 2012).[3] Two clear cold surges were observed in the forecasts. The cold surges appeared in the 06-hour and 12-hour forecasts valid from 5th December, 1200 UTC, to 9th December, 1200 UTC (Figure 3.5), and from 15th December, 0000 UTC, to the end of the simulations. Clearly, the SINGV forecasts are able to resolve this key feature of the Northeast monsoon season.

3.4 Interesting vortex features

Vortex-like feature observed We also observed a rather interesting large-scale counter-clockwise circulation (several hundred kilometres across) over equatorial South China Sea (Figure 3.6). The circulation shows up as a low-level closed, counter-clockwise, circulation loop on the forecasts, valid from 0000UTC, 8th December to 0600 UTC, 9th December (Figure 3.6). Interestingly, the Australian Bureau of Meteorology's Gradient Wind Analysis on the 9th of December 2015 also shows a similar circulation at the place where our

[3] A description and explanation of the cold surge can be found in the introduction, in the paragraph labelled "cold surges".

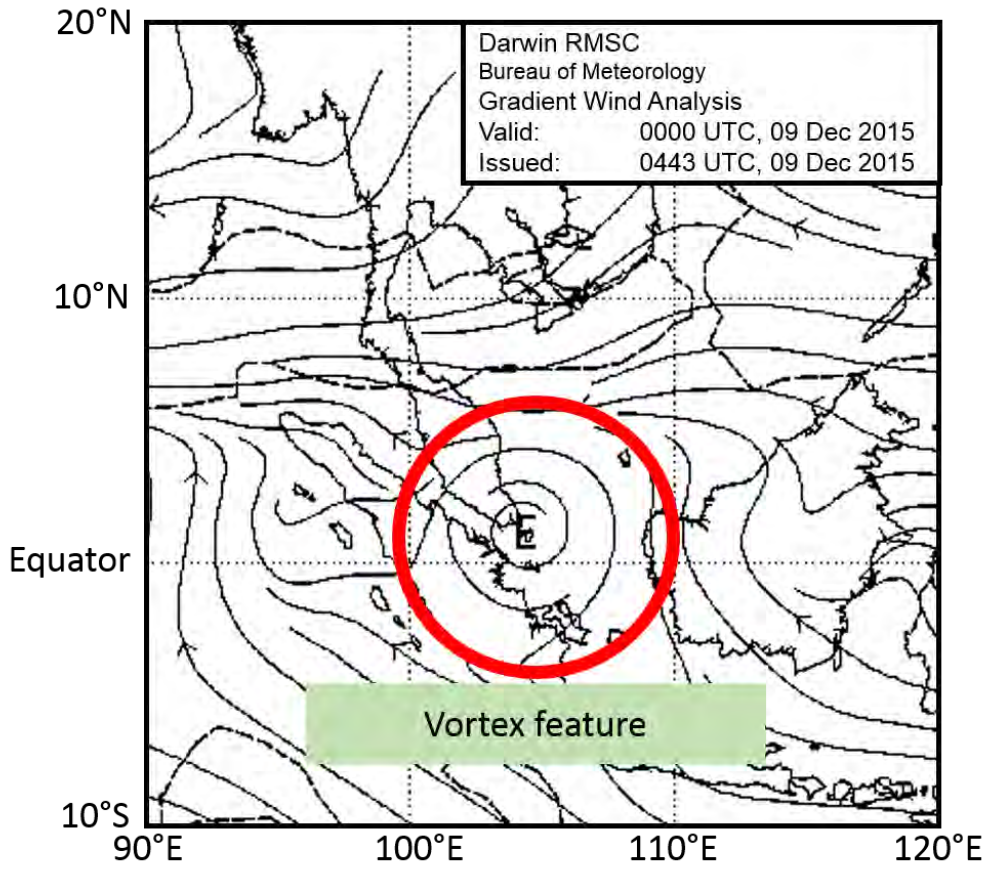


Figure 3.7: Streamlines plotted by the Australian Bureau of Meteorology (Regional Specialized Meteorological Center) on 9th December 2015, 0000 UTC. These streamlines show the wind flow at 1 km above the sea level. Note the vortex feature indicated in the red circle. This plot is modified from the analysis charts stored in the Australian Bureau of Meteorology Analysis Chart Archive (Australian Government, 2016).

circulation appears in the forecast (Figure 3.7). In other words, this circulation pattern may not be a spurious artefact of the model.

Uncertainty over identity of circulation At the time of writing, we are uncertain if the circulation is a Borneo vortex[4], which is a characteristic of the Northeast monsoon. While the scale and morphology of the circulation pattern matches, its position is too far west: Borneo vortices usually occur in the region between the western tip of Borneo and off the east coast of the Malay Peninsula (Chang et al., 2005).

Circulation unlikely to affect error covariances significantly Even if the circulation ultimately is not the characteristic Borneo vortex of the Northeast monsoon, the

[4]Low-level, converging, cyclonic vortices that tend to occur over the western part of Borneo (near the Equator, and east of the model domain), and are typically accompanied with deep convection.

short duration of this circulation means that it is unlikely to have much of an impact on the overall error covariance of the region. The circulation clearly vanished from the forecast after 0600 UTC, 10th December. In other words, the circulation only persisted for 12% of the 15 days considered. Furthermore, the circulation itself wandered from the southeastern part of the domain to the northeastern part of the domain, diluting the influence of any errors relating to the circulation over a large trajectory. As such, we can safely ignore the influence of the circulation when analysing the forecast error covariances.

3.5 Summary

In summary, we have confirmed that the SINGV forecasts are behaving as they should in the presence of the Northeast monsoon: the identified tropopause is in the correct pressure level, the monsoon flow is as expected and the characteristic cold surges are observed. The unusual circulation observed in the forecasts is also unlikely to have a great influence on the covariances of the forecast error. We can thus infer that the forecast error covariances estimated from these forecasts to representative of the Maritime Continent's Northeast monsoon.

Chapter 4

Error covariance features and regression coefficient models

Using NMC to examine error covariance features While it is impossible to determine $\boldsymbol{B}$ due to memory problems, it is possible to determine several columns of $\boldsymbol{B}$. In this chapter, we will utilize the NMC method to determine several columns of the $\boldsymbol{B}$ matrix. Each column of $\boldsymbol{B}$ contains the covariance between the forecast error of a particular variable, at a particular point on the model grid, against the forecast error of both variables, across the entire model grid. Such columns can be used to diagnose whether a regression coefficient model's assumptions holds.

Overview We will begin with explaining the physical meaning of the two error variables we are studying in this project. Following that, we will explain how the single-observation solution to the data assimilation cost function can be used to interpret physical meaning from columns of $\boldsymbol{B}$. Several columns of $\boldsymbol{B}$ will then be displayed and interpreted. Using these columns, we will overturn the horizontal invariance assumption utilized in two commonly used models of the control variable transform regression coefficients[1]. A new assumption will then be proposed from examining the columns of $\boldsymbol{B}$ and a new model will be formulated based on that[2].

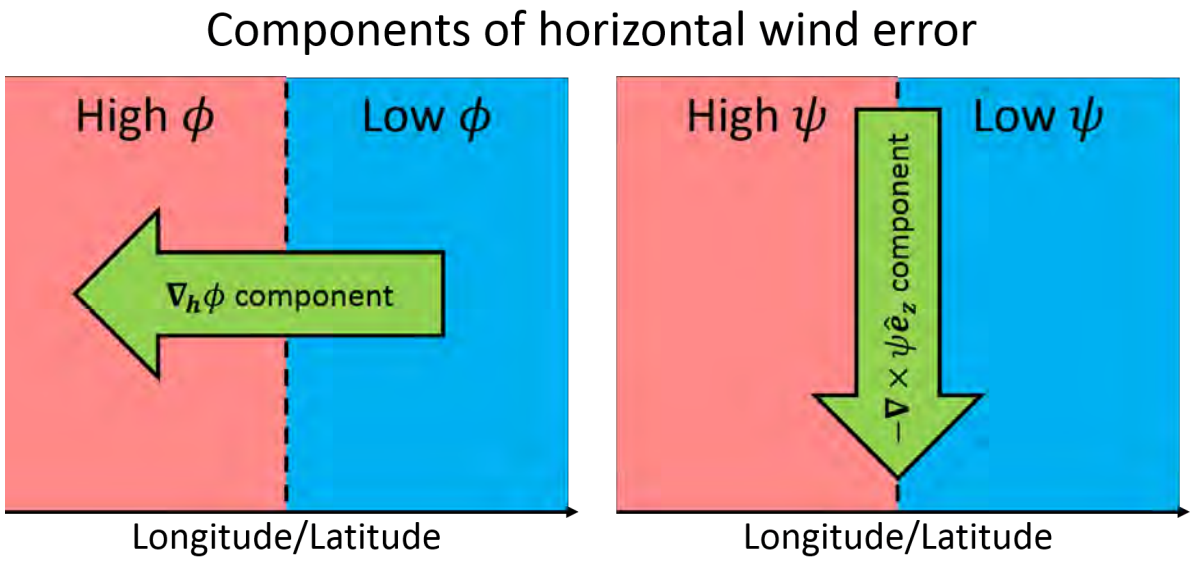


Figure 4.1: Diagrams showing the directions of the horizontal wind error vector field, induced by ϕ (left) and ψ (right). Note that the z-axis is pointing out of the page.

4.1 Physical meaning of ϕ and ψ

Meaning of ϕ As mentioned in the introduction, the error in the horizontal wind field can be written as

$$\boldsymbol{u}_k = \boldsymbol{\nabla}\phi\left(x,y,k\right) - \boldsymbol{\nabla}\times\Big(\psi\left(x,y,k\right)\hat{\boldsymbol{z}}\Big). \tag{4.1}$$

The general appearances of both parts of the wind field's forecast error can be directly inferred from the contour lines of ϕ and ψ. For the case of the divergent part of the wind ($\boldsymbol{\nabla}\phi$), consider the fact that the gradient of a scalar is perpendicular to the scalar's contour, and points in the direction of maximum increase (Figure 4.1, left panel). In other words, a positive hill of ϕ corresponds to a local flow convergence that is centered around the maximum and *vice versa* for a valley of ϕ (see Figure 4.2 for illustrations).

Meaning of ψ It is slightly more difficult to interprete ψ. Defining $\boldsymbol{u}_{\boldsymbol{\psi}} \equiv -\boldsymbol{\nabla}_{\boldsymbol{H}}\times(\psi\hat{\boldsymbol{z}})$, consider the following integral over an area bounded by a ψ contour line on a fixed model level:

$$\int_S \nabla_H^2\psi dA = \int_S \boldsymbol{\nabla}_{\boldsymbol{H}}\times\boldsymbol{u}_{\boldsymbol{\psi}}\cdot\hat{\boldsymbol{z}}dA.$$

[1]These are schematically illustrated in Figures 1.1 and 1.2.

[2]The computation of the regression coefficients under the new model is schematically illustrated in Figure 1.3. Figure 1.4 shows how this model is utilized in the control variable transform.

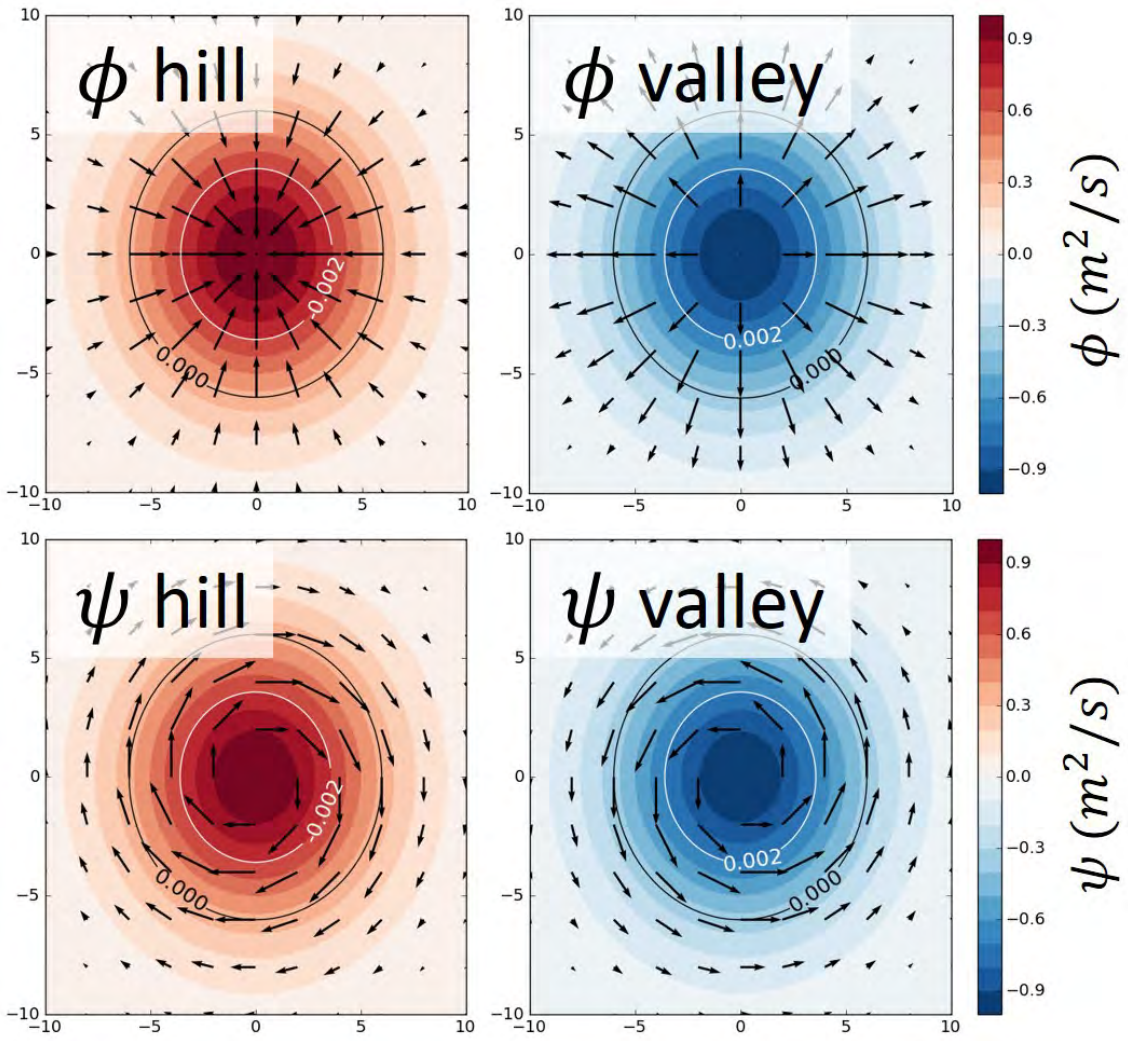


Figure 4.2: Contour plots showing examples of hills (left column) and valleys (right column) of ϕ (top row) and ψ (bottom row), overlaid with vector arrows indicating the respective components of wind implied by these two variables. The unfilled contours indicate horizontal divergence (top row) and the z-component of vorticity (bottom row).

Applying the 2D form of the divergence theorem and Green's theorem yields,

$$\oint_{\partial S} \boldsymbol{\nabla}_{\boldsymbol{H}} \psi \cdot d\boldsymbol{l}_{\perp} = \oint_{\partial S} \boldsymbol{u}_{\boldsymbol{\psi}} \cdot d\boldsymbol{l}_{\parallel}.$$

where $d\boldsymbol{l}_{\perp}$ is a vector pointing out of area S and perpendicular to the contour, and $d\boldsymbol{l}_{\parallel}$ is a length vector element along the contour, oriented such that S lies on the left of $d\boldsymbol{l}_{\parallel}$. Since the contour is on a fixed model level, then

$$d\boldsymbol{l}_{\parallel} \times \hat{\boldsymbol{z}} = d\boldsymbol{l}_{\perp}.$$

Applying this relation and considering $(\boldsymbol{\nabla_H}\psi)\cdot\left(d\boldsymbol{l}_\parallel\times\hat{\boldsymbol{z}}\right)=\left(d\boldsymbol{l}_\parallel\right)\cdot(\hat{\boldsymbol{z}}\times\boldsymbol{\nabla_H}\psi)$ yields,

$$\oint_{\partial S}\left(d\boldsymbol{l}_\parallel\right)\cdot(\hat{\boldsymbol{z}}\times\boldsymbol{\nabla_H}\psi)=\oint_{\partial S}\boldsymbol{u_\psi}\cdot d\boldsymbol{l}_\parallel,$$

or,

$$(\hat{\boldsymbol{z}}\times\boldsymbol{\nabla_H}\psi)\cdot d\boldsymbol{l}_\parallel=\boldsymbol{u_\psi}\cdot d\boldsymbol{l}_\parallel.$$

Since the gradient is horizontal and perpendicular to the contour line, $(\hat{\boldsymbol{z}}\times\boldsymbol{\nabla_H}\psi)$ yields a vector is either antiparallel or parallel to $d\boldsymbol{l}_\parallel$. The right panel of Figure 4.1 illustrates how $\boldsymbol{u_\psi}$ can be inferred from ψ fields using this idea. For the case of a ψ hill bounded by the contour line, $(\hat{\boldsymbol{z}}\times\boldsymbol{\nabla_H}\psi)$ is antiparallel to $d\boldsymbol{l}_\parallel$. In other words, cyclonic circulation will be represented by a ψ valley (see Figure 4.2 for illustrations).

4.2 Columns of $\boldsymbol{B}$

4.2.1 Interpreting columns of $\boldsymbol{B}$

Single-observation solution interprets $\boldsymbol{B}$ To understand the physical meaning of a column of $\boldsymbol{B}$, we need a way to relate the information in the column of error covariances to the result of data assimilation itself. Interpreting $\boldsymbol{B}$ through the lens of the single-observation solution to the data assimilation cost function in Eqn (1.4) is one intuitive way.

Single-observation solution conditions and statement Supposing that the observation corresponds to the n-th element on the model state vector, $\boldsymbol{X}$, then the ℓ-th element of the analysis state vector, $\boldsymbol{X^a}$, can be shown to be (Bannister, 2008a)

$$X^a_\ell - X^b_\ell = B_{\ell n}\frac{Y-X^b_n}{\sigma+B_{nn}}. \tag{4.2}$$

This is the single-observation solution[3], and it operates under three conditions. First of all, only one observation, Y, is assimilated. Secondly, this observation must be on the same location as one of the forecast model's grid points. Lastly, the observation must be an observation of one of the variables in the model. Henceforth, $\boldsymbol{X^a}-\boldsymbol{X^b}$ will be called the "analysis increment".

[3]The derivation of this solution is available in Appendix A.

An analysis increment is proportional to a column of $\boldsymbol{B}$ As can be seen from Eqn (4.2), the analysis increment is proportional to the n-th column of $\boldsymbol{B}$, with $\frac{Y-X_n^b}{\sigma+B_{nn}}$ as the proportionality constant. In other words, we can simply infer how $\boldsymbol{X^b}$ will be modified to form $\boldsymbol{X^a}$ from a column of $\boldsymbol{B}$. We will use this framework to determine the physical implications of the information contained in $\boldsymbol{B}$.

4.2.2 Computing columns of $\boldsymbol{B}$ from NMC

General form of $\boldsymbol{B}$ As mentioned earlier, we can determine several columns of $\boldsymbol{B}$ via the NMC method, without suffering from computer memory problems. We will explain this procedure in detail. For that, we need the form of $\boldsymbol{B}$. It is simply

$$\boldsymbol{x^b} \equiv \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\phi} \end{bmatrix} \Longrightarrow \boldsymbol{B} = \begin{bmatrix} \left\langle \boldsymbol{\psi}\boldsymbol{\psi}^\top \right\rangle_t & \left\langle \boldsymbol{\psi}\boldsymbol{\phi}^\top \right\rangle_t \\ \left\langle \boldsymbol{\phi}\boldsymbol{\psi}^\top \right\rangle_t & \left\langle \boldsymbol{\phi}\boldsymbol{\phi}^\top \right\rangle_t \end{bmatrix},$$

where $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ are vectors containing the respective forecast errors ($N/2$ elements each) and the averaging is done over time.

Meaning of a column of $\boldsymbol{B}$ The m-th column in the left half of $\boldsymbol{B}$ ($m < N/2$) contains 2 sets of information. It contains the covariance between ψ at the position corresponding to m-th element on the $\boldsymbol{\psi}$, versus ψ and ϕ everywhere on the model grid. Likewise, the m'-th column on the right half of $\boldsymbol{B}$ ($m' > N/2$) contains the covariance between ϕ at the $m' - N/2$ position on $\boldsymbol{\phi}$ vector, versus ψ and ϕ everywhere on the model grid.

NMC $\boldsymbol{B}$ column computation method From Eqn (2.2), we can compute the vectors of the NMC-estimated forecast errors at time index t via

$$\boldsymbol{\psi}_t \approx \frac{1}{\sqrt{2}} \left(\boldsymbol{\psi}_{\text{12-hr fcst},t} - \boldsymbol{\psi}_{\text{06-hr fcst},t} - \left\langle \boldsymbol{\psi}_{\text{12-hr fcst},t} - \boldsymbol{\psi}_{\text{06-hr fcst},t} \right\rangle_t \right),$$

and likewise for $\boldsymbol{\phi}$. As such, to compute the column of $\boldsymbol{B}$ corresponding to ψ at model index (i, j, k) (longitudinal, latitudinal, vertical), we simply use

$$\left\langle \boldsymbol{\psi}\, \psi\,(i,j,k) \right\rangle_t \approx \frac{1}{N_T} \sum_{t=1}^{N_T} \psi_t\,(i,j,k) * \boldsymbol{\psi}_t,$$

$$\left\langle \boldsymbol{\phi}\, \psi\,(i,j,k) \right\rangle_t \approx \frac{1}{N_T} \sum_{t=1}^{N_T} \psi_t\,(i,j,k) * \boldsymbol{\phi}_t.$$

A similar method can be utilized to compute the column of $\boldsymbol{B}$ corresponding to ϕ at (i, j, k). When interpreting the columns of $\boldsymbol{B}$, we will view the covariances on the left hand side at all displayed locations on the model grid.

4.3 Features observed in columns of $\boldsymbol{B}$

4.3.1 Deep convection

Recap: forecasts show deep convection tendencies With the single-observation solution framework to interpret columns of $\boldsymbol{B}$ and the NMC column estimation laid out, we are ready to examine the physical meaning implied by the columns of $\boldsymbol{B}$. We will begin with deep convection. As per the skewT-logP discussion in Chapter 3, the forecasts of the Northeast monsoon display a tendency for deep convection. As such, we would expect to see the forecast error covariances to reflect that.

Auto-covariances of ϕ at the surface level To confirm that, 9 columns of the $\left\langle \boldsymbol{\phi}\boldsymbol{\phi}^{\top} \right\rangle$ sub-matrix of $\boldsymbol{B}$ were computed. The 9 columns are such that their ϕ reference points are all located on the lowest model level (indicated by the 9 green dots in Figure 4.3). The covariances between ϕ at each of these 9 locations (green dots), and ϕ in a 450 km by 450 km square centred on each location, are plotted out in Figure 4.3.

Mesoscale surface convergence with positive ϕ observation increment Let us suppose that we are assimilating an observation of ϕ that is bigger than the corresponding background state's ϕ (*i.e.*, positive ϕ observation increment). Following the discussion on interpreting the columns of $\boldsymbol{B}$ through Eqn (4.2), Figure 4.3 implies that assimilating that observation at any green dot causes a hill of ϕ ($O\left(10^2\right)$ km diameter) to appear in the analysis increment, centred on the said green dot. Physically, this means that the assimilation process enhances mesoscale convergence ($O\left(10^2\right)$ km wide) on the surface level.

Deep atmosphere mesoscale divergence with positive ϕ observation increment Furthermore, when we consider the vertical structure of these ϕ-ϕ covariances, we notice that the positive observation increment at the surface triggers a ϕ valley aloft. This is demonstrated by Figure 4.4, which shows the vertical structure of the said covariances, from the three three green dots at 1°N in Figure 4.3. Clearly, assimilating the observation

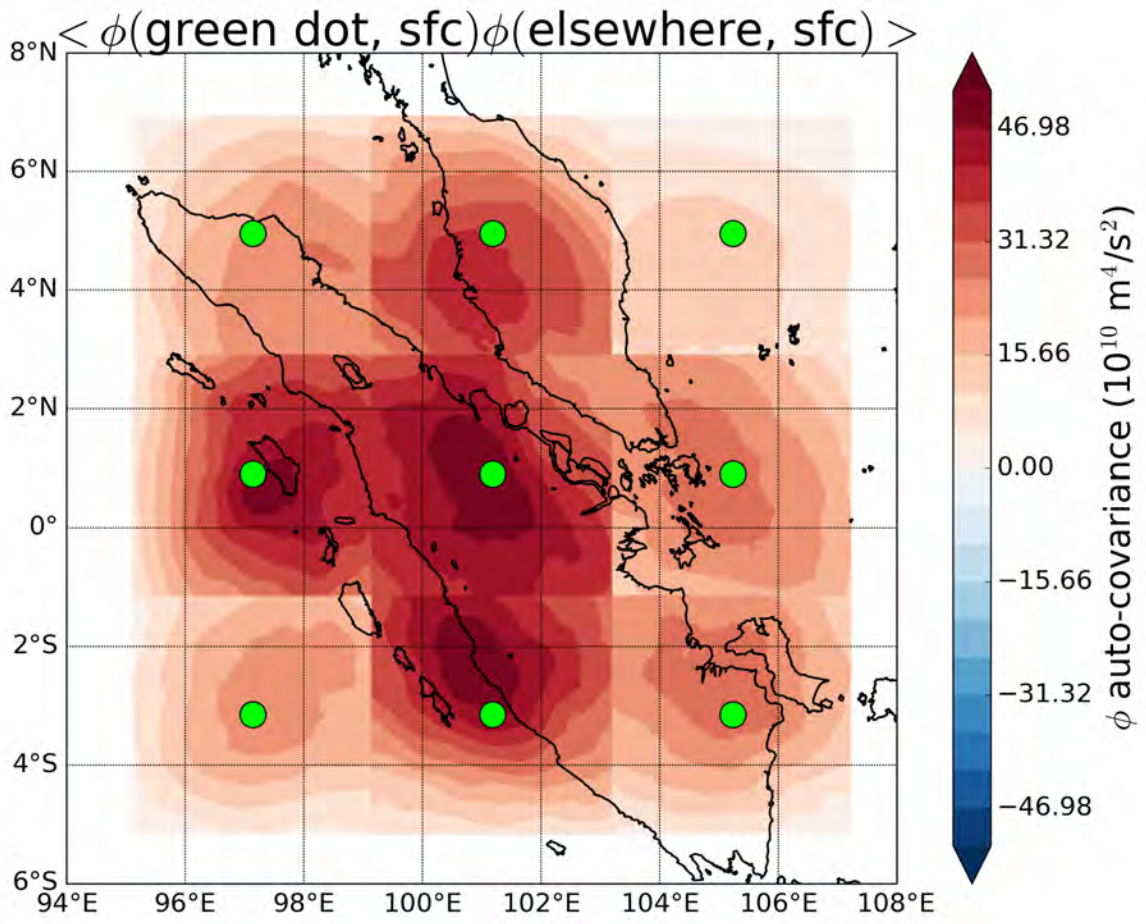


Figure 4.3: The contour plot shows the covariance between ϕ at the green dots and ϕ in squares centred on the green dots, on the surface level. The dimensions of each square is 450 km $\times$ 450 km, or roughly 4° $\times$ 4°. The landmasses in the domain are outlined in thick, solid black lines.

enhances divergence aloft ($O\,(10^2)$ km across). Furthermore, the valley feature occurs in the vicinity of the 50-th model level, which corresponds to the 110 hPa layer, according to Figure 3.1. In other words, the data assimilation triggers deep atmosphere mesoscale divergence.

Deep convection error feature Taken together, the mesoscale surface convergence and ∼100 hPa mesoscale divergence pattern implies that the error covariances are organized in a deep convection pattern. Furthermore, given the ubiquity of this pattern, it seems that using a positive ϕ observation increment anywhere on the surface enhances convection. In other words, the ϕ-ϕ covariances reflect the tendency for deep convection (as discussed in Chapter 3).

Foreshadow – Covariances ≈ variances near green dots An interesting observation can be made from Figures 4.3 and 4.4. Notice that the covariances tend to vary slowly

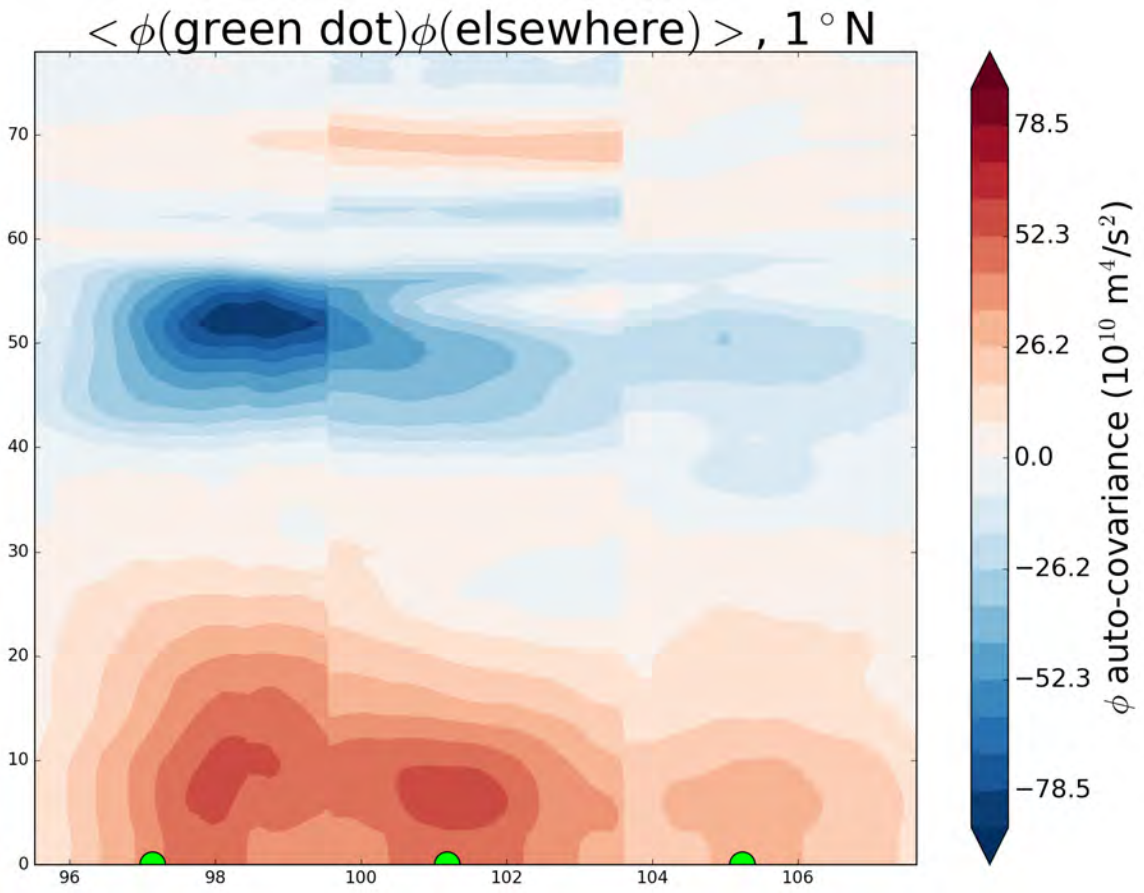


Figure 4.4: The contour plot shows the covariances of ϕ from the green dots along the 1°N latitude in Figure 4.3, and ϕ everywhere on a vertical slice of ϕ, along the 1°N latitude. This plot is separated into 3 vertical strips, each centred on a the green dot, spanning 4° longitude in the horizontal and all 79 model levels in the vertical. The horizontal axis is longitude and the vertical axis is model level.

immediately around the green dots. In other words, the covariances immediately around the green dots are roughly equal to the variance of ϕ at the green dots. This intriguing property will be used later when formulating our own regression coefficient model.

4.3.2 Vortex patterns

Surface circulations from ψ observation increment When we computed the covariance between ψ at the same 9 points, and ψ everywhere on the surface of the model, hills similar to those of ϕ-ϕ surface covariances appear (Figure 4.5). However, the physical meaning for the analysed wind field is completely different from that of ϕ. Assimilating a positive ψ observation increment generates positive ψ hills in the analysis increment. Following the interpretation in Figure 4.2, these ψ hills mean that this assimilation enhances clockwise (or anticyclonic) circulation surrounding the green dots.

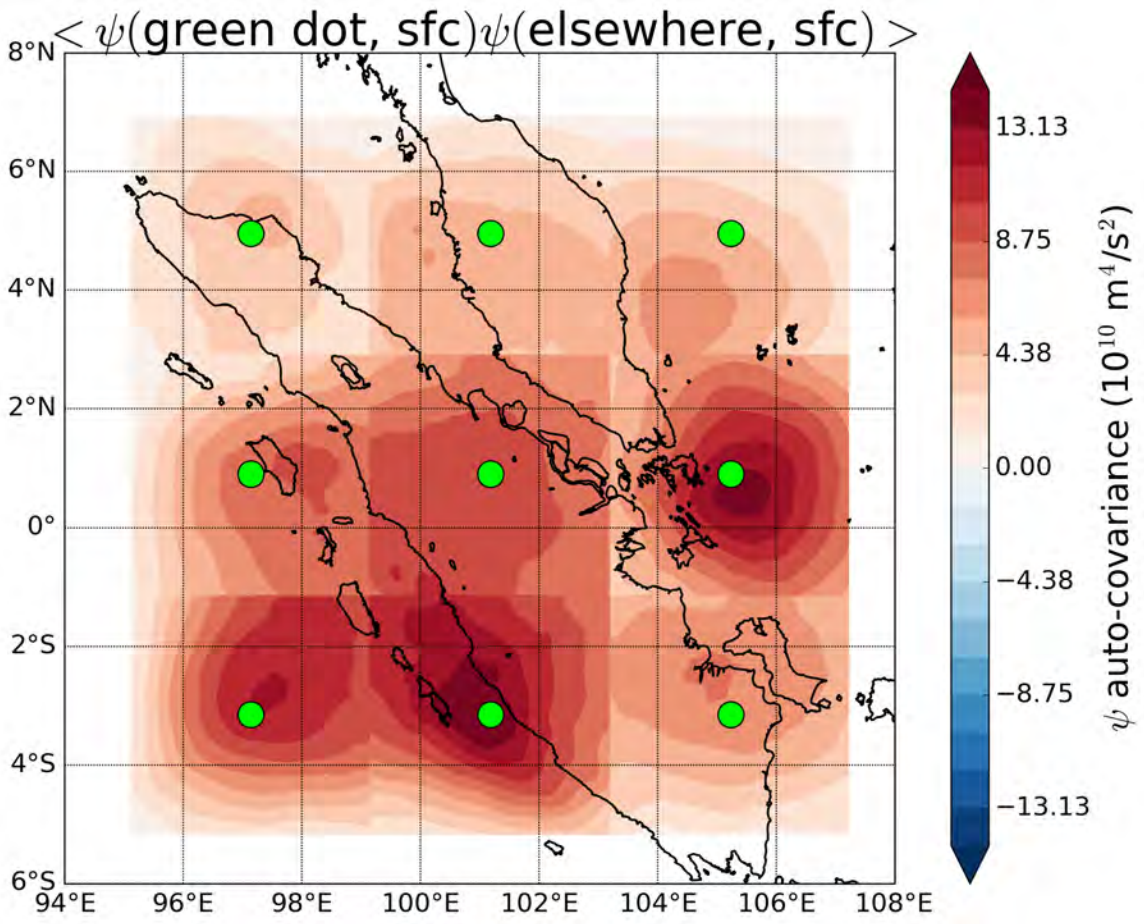


Figure 4.5: The contour plot shows the covariance between ψ at the green dots and ψ in squares centred on the green dots, on the surface level. The dimensions of each square is 450 km $\times$ 450 km, or roughly 4° $\times$ 4°. The landmasses in the domain are outlined in thick, solid black lines.

Foreshadow – Covariances $\approx$ variances near green dots At the time of writing, we are unable to determine the physical origin of these patterns for ψ, both on the surface (Figure 4.5) and in the vertical (Figure C.1 in Appendix C). However, like the case of ϕ-ϕ covariances, we notice that the covariances tend to be similar to variances immediately around the green dots. This pattern will be utilized later to formulate our own regression coefficient model.

4.3.3 Horizontally variant and anisotropic cross-covariances

ϕ-ψ covariance columns are horizontally variant and anisotropic We will now turn our attention to the cross-covariance between ϕ and ψ. The covariances between ϕ at the same 9 reference positions and ψ elsewhere in 450 km by 450 km squares centred on the reference positions are computed and plotted in Figure 4.6. It is immediately apparent that unlike the ϕ-ϕ and ψ-ψ columns of $\boldsymbol{B}$, these ϕ-ψ columns show vastly

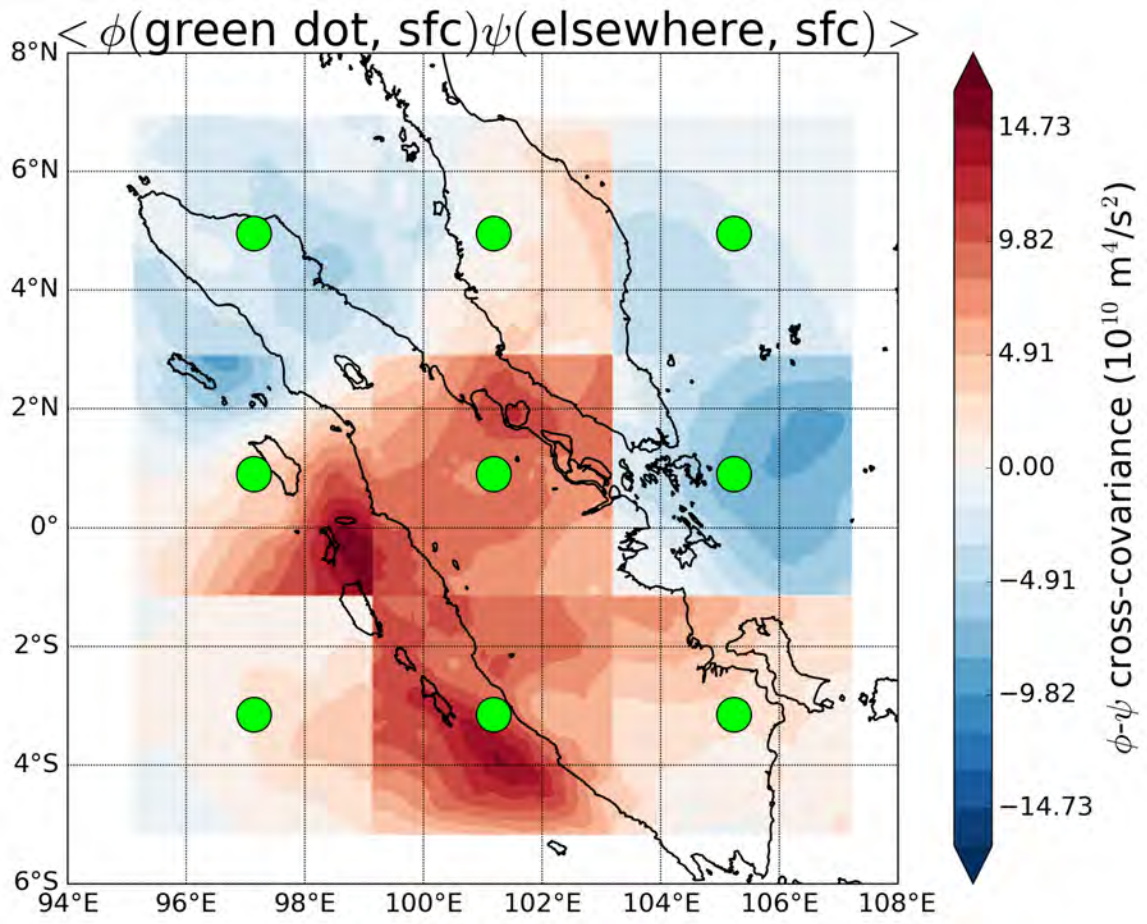


Figure 4.6: The contour plot shows the covariance between ϕ at the green dots and ψ in squares centred on the green dots, on the surface level. The dimensions of each square is 450 km × 450 km, or roughly 4° × 4°. The landmasses in the domain are outlined in thick, solid black lines.

different error patterns from column to column. Furthermore, the covariances do not vary isotropically around the green dots. In other words, the cross-covariance columns exhibit huge variations between horizontal locations and are severely anisotropic.[4]

Horizontal variations and anisotropy may be due to complicated topographic interactions The variations and anisotropy may be due to complicated topographic interactions. As an illustration, we will examine the cross-covariances connected to the green dot at 1°N, 97°E of Figure 4.6. Figure 4.7 display the cross-covariances between ϕ at the green dot, and everywhere else in the model grid, on the surface level.[5] When an assimilated observation of ϕ at the green dot (1°N, 97°E) increases ϕ at that location, the analysis increment will contain a ϕ hill centered on the green dot (see Figure 4.3) and a ψ dipole, with the negative patch to the northwest and the positive patch to the southeast

[4]Similar observations are also made when we consider the 9 columns corresponding to the covariance between ψ at the same 9 positions and ϕ everywhere on the surface (Figure C.2 in Appendix C).

[5]Essentially, an expanded version of the square in Figure 4.6 that is centred at 1°N, 97°E.

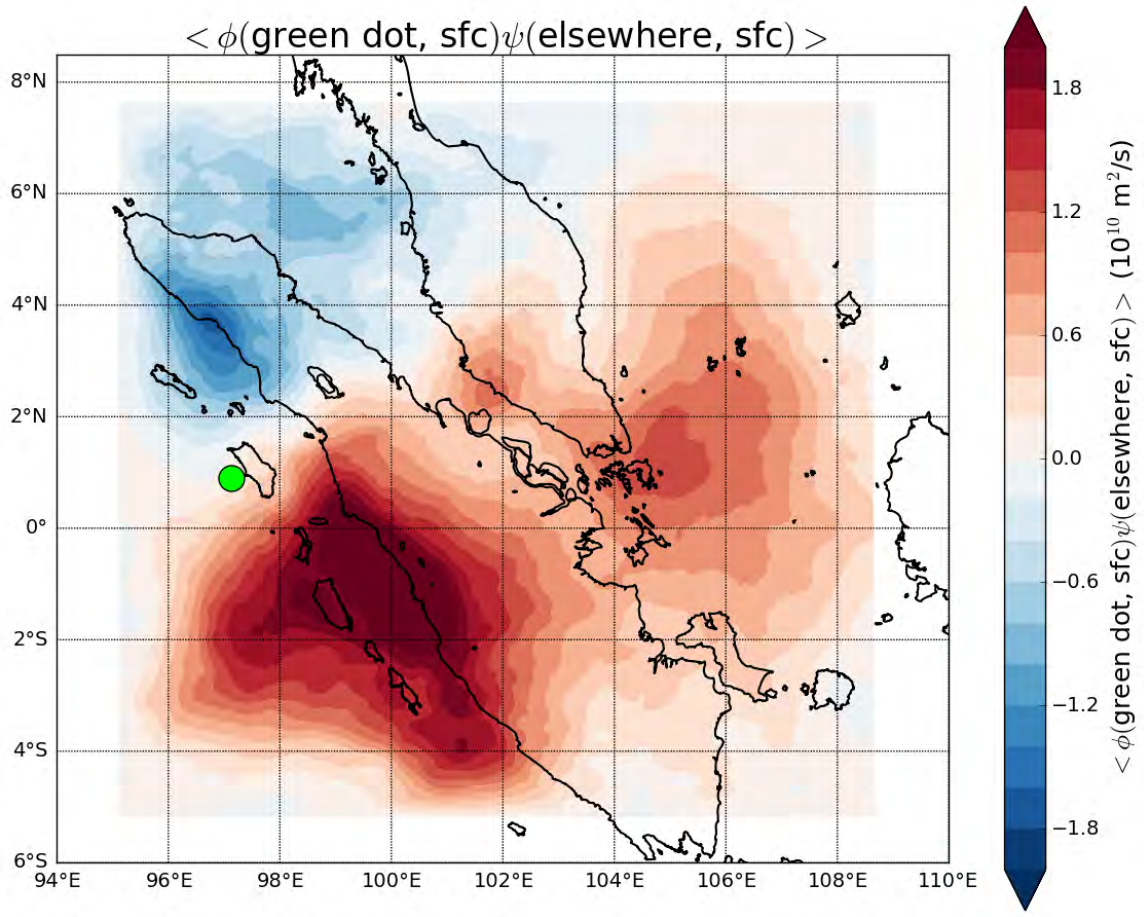


Figure 4.7: The contour plot shows the covariance between ϕ at the green dot, surface level, and ψ everywhere else on the surface level of the model grid. The landmasses in the domain are outlined in thick, solid black lines.

(Figure 4.7). This analysis increment can be interpreted as a southwesterly enhancement of the wind flowing through the green dot[6]. The southwesterly impinges on the island beside the green dot, resulting in surface convergence due to topographic obstruction, giving rise to the ϕ hill. Downstream of the green dot, this enhanced wind encounters the pass of the mountain range that runs along Sumatra (see Figure 2.1). As such, the wind partially diverges to the northwest and southeast and the remainder passes through the mountain pass. The divergence explains the intense ψ dipole feature on the windward side of the mountain range, while the jet resulting from the passage through the mountain pass explains the weaker dipole seen downstream of the pass. In the other locations of the domain, the topography is completely different, meaning that different interactions, and thus cross-covariance patterns, are to be expected.[7] In other words, topography may be part of the reason behind the dramatic horizontal variations and anisotropies observed.

[6]Refer to Figure 4.1 to see how the ψ part of the wind field can be interpreted from ψ contours

[7]Also, the atmospheric forcing can be different. In the eastern half of the domain, the error cross-covariances are likely influenced by the monsoon flow.

4.4 Regression coefficient models

4.4.1 Conventional regression coefficient models and their flawed assumptions

Flawed conventional regression coefficient models These horizontal variations and anisotropies are contrary to the assumptions of the two commonly used regression coefficient models (see Figures 1.1 and 1.2). The first model assumes that there are no horizontal variations in the regression coefficient and that there are zero covariances between variables on different model levels (Chen et al., 2013). The second model relaxes the second assumption, but keeps the assumption of horizontal invariance (Descombes et al., 2015). On top of that, these models assume that the relationship between variables on two model grid points are isotropic (Bannister, 2008b). This manifests in the lack of any horizontal directionality in these definitions of regression coefficients. However, as observed earlier, the cross-covariances are hardly horizontally invariant or isotropic. *I.e.*, the regression coefficients are likely to be horizontally variant and anisotropic! We can already foresee that these conventional models will have bad performance in our study.

Conventional models will still be used Despite the breakdown of the assumptions, these two models are still worth testing. First of all, the performance of these two models can be used to benchmark the performance of our own model (which will be introduced later). Furthermore, these two models result in very small matrices of α, and can thus be easily computed. Lastly, due to the limited number of forecast pairs, any computation of regression coefficient matrices must contend with the problem of undersampling. If there are insufficient samples of forecast errors, the forecast error auto-covariance matrices are definitely singular. In other words, inverting these matrices to determine the regression coefficients is out of the question. However, these two models circumvent this undersampling problem by considering all the points in each model level as a sample. This typically overcomes the problem of undersampling.

4.4.2 New regression coefficient matrix model

Requirements A better model than the two conventional models would be one that uses more realistic assumptions while preserving the benefit of computational ease and circumvents undersampling. Here, we will deduce a more suitable assumption from examining the earlier plots of the columns of $\boldsymbol{B}$.

Assumption: points near reference point are samples of a specified point The assumption that we have decided to use is: the valuations of ϕ and ψ in the immediate vicinity of a specified point are equivalent to samples of ϕ and ψ at the specified point. For the ease of writing, this assumption will be called the **neighbourhood equivalence assumption**. We will justify this assumption mathematically and from examining the earlier auto-covariance plots.

Mathematical justification of neighbourhood equivalence assumption Consider the Taylor expansion of ψ as a function of position and time. The expansion about a position $\boldsymbol{r}$, up to the linear order of $\boldsymbol{h}$, is simply,

$$\psi\left(\boldsymbol{r}+\boldsymbol{h},t\right)\approx\psi\left(\boldsymbol{r},t\right)+\boldsymbol{h}\cdot\boldsymbol{\nabla}\psi\left(\boldsymbol{r},t\right).$$

For the ease of discussion, the coordinates are Cartesian and $\boldsymbol{\nabla}$ is the 3D Cartesian gradient operator. We will also set the temporal mean of ψ to zero. Clearly, the temporal auto-covariance and variance of ψ in this small region are

$$\begin{aligned}\mathrm{Var}\left\{\psi\left(\boldsymbol{r}\right)\right\}&=\left\langle\psi\left(\boldsymbol{r}\right)^{2}\right\rangle,\\ \mathrm{Covar}\left\{\psi\left(\boldsymbol{r}\right),\psi\left(\boldsymbol{r}+\boldsymbol{h}\right)\right\}&\approx\left\langle\psi\left(\boldsymbol{r}\right)^{2}\right\rangle+\boldsymbol{h}\cdot\left\langle\psi\left(\boldsymbol{r}\right)\boldsymbol{\nabla}\psi\left(\boldsymbol{r}\right)\right\rangle.\end{aligned}$$

In other words, in the vicinity of $\boldsymbol{r}$, the valuation of the auto-covariance can be very close to the variance at $\boldsymbol{r}$! This implies that if the valuations of ψ on two nearby points, they can be treated as two samples of the same quantity.[8]

Justification from columns of $\boldsymbol{B}$: horizontal view The auto-covariance columns of $\boldsymbol{B}$ also support this notion. As mentioned in the earlier parts of this chapter, we noticed that the auto-covariances tend to be similar to the variances of the green dots immediately around the green dots. In fact, when we divide the auto-covariances of ϕ with the variance of ϕ, and likewise for ψ, in Figure 4.8, the resulting ratio tends to be within 0.2 of 1 within the 0.5° of the green dot. In other words, we can roughly consider ϕ and ψ within a $1^\circ\times1^\circ$ square centred at a point to be samples of ϕ and ψ at that point.

Justification from columns of $\boldsymbol{B}$: vertical view The same argument also applies in the vertical. From Figure 4.9, we infer that within roughly 5 model levels of a reference

[8]We acknowledge that there are scenarios where covariance of two variables can be equal to their variances, even though the two variables are clearly different. However, since we are dealing with auto-covariances, it seems intuitive to consider them the same quantity.

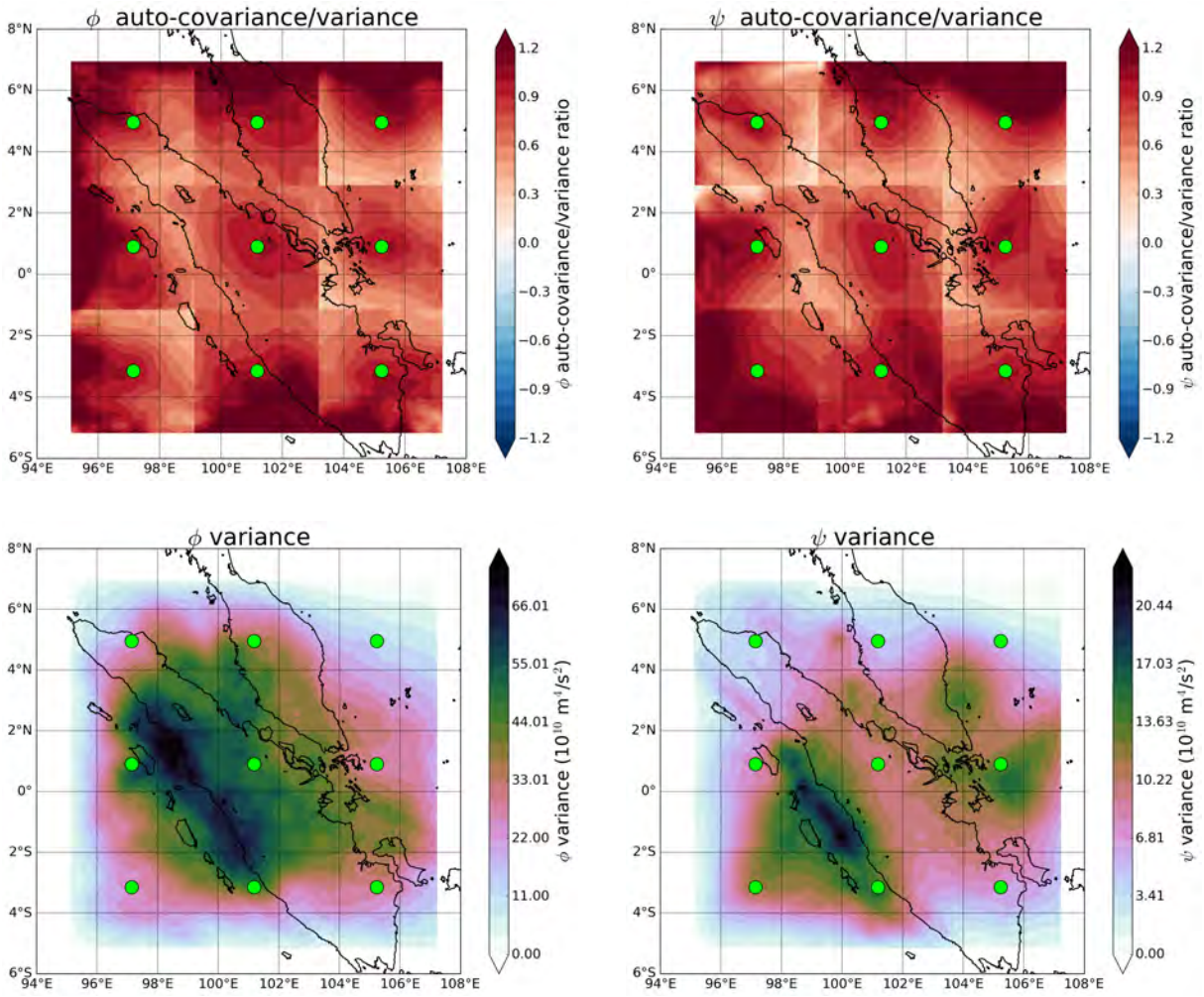


Figure 4.8: The horizontal contour plots show the ratio between the auto-covariances in Figures 4.3 (top left, for ϕ) and 4.5 (top right, for ϕ), against the local variances. The local variances have been plotted in the corresponding columns of the bottom row.

point, the neighbourhood equivalence assumption applies.[9]

Neighbourhood equivalence assumption allows affordable α computation The neighbourhood equivalence assumption causes the regression coefficient matrix computation to be affordable. Since the equivalence appears to hold for a set of points within a subdomains with sides 1°-by-1°-by-5 model levels, we can split the entire model grid into such subdomains. The regression coefficients only need to be calculated between each pair of such subdomains. This drastically reduces the size of the regression coefficient matrix, making it affordable to compute.

Neighbourhood equivalence assumption prevents undersampling Furthermore, the assumption prevents undersampling from happening. Under the assumption, all points

[9]However, due to computational speed limitations, we will consider groups of 10 model levels later.

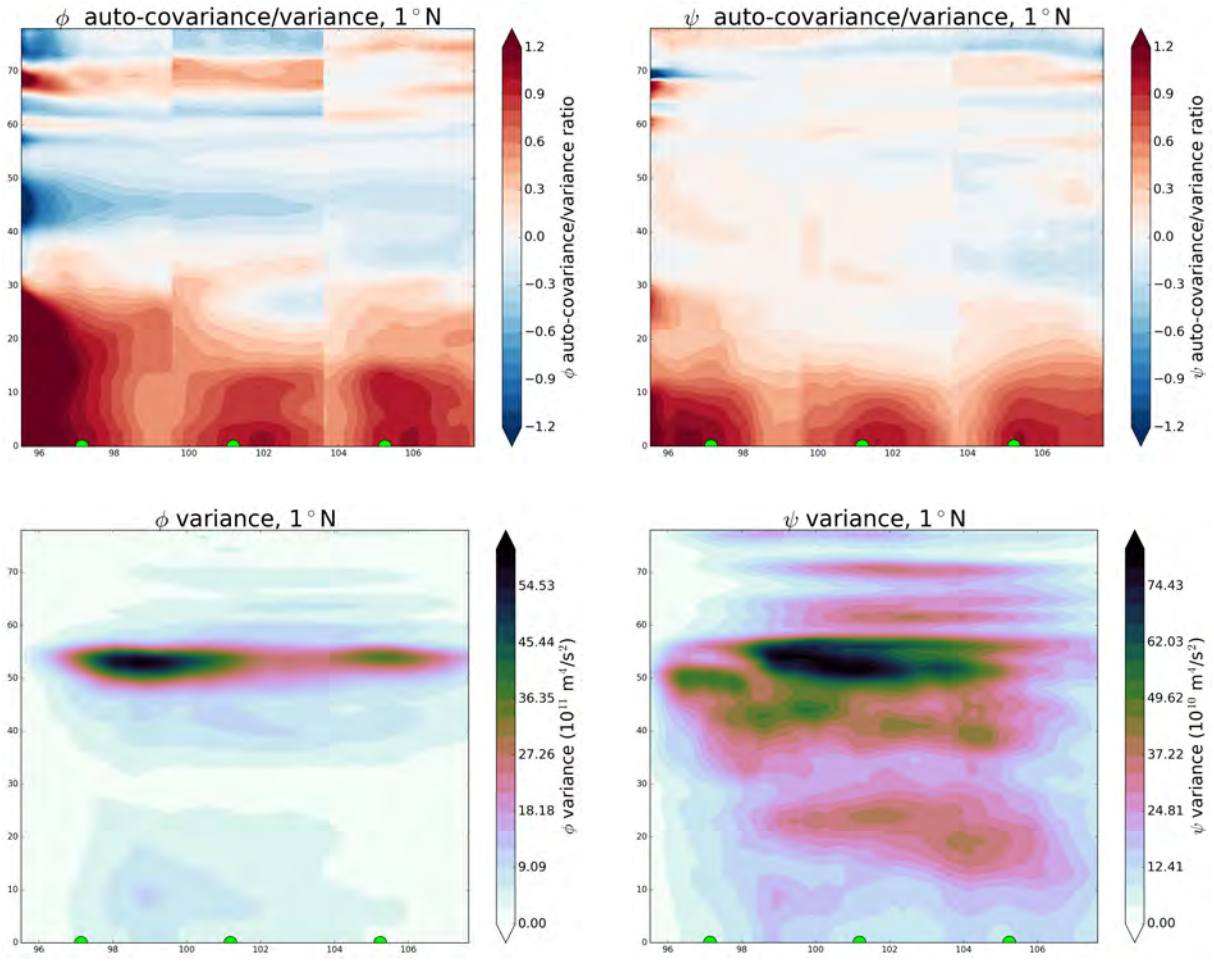


Figure 4.9: The vertical contour plots show the ratio between the auto-covariances in Figures 4.4 (top left, for ϕ) and C.1 (top right, for ϕ), against the local variances. The local variances have been plotted in the corresponding columns of the bottom row.

within a subdomain are samples. This dramatically increases the number of samples by a factor of the number of points in the subdomain. Furthermore, the size of the auto-covariance matrices needed to compute the regression coefficients is reduced by the same factor, reducing the number of samples needed to prevent singular matrices. Clearly, formulating a model of regression coefficients under this assumption is feasible.

Need for sampling pattern of new model However, the assumption alone does not tell us the exact way to compute the reduced auto-covariance matrices. Suppose that we want to determine the auto-covariance of ψ between subdomain A and subdomain B. By the definition of covariance, we need to multiply the mean-removed samples of ψ in both subdomains to compute the covariance.

Point-by-point multiplication sampling We must avoid multiplying each mean-removed sample in subdomain A to all the mean-removed samples in subdomain B (Figure

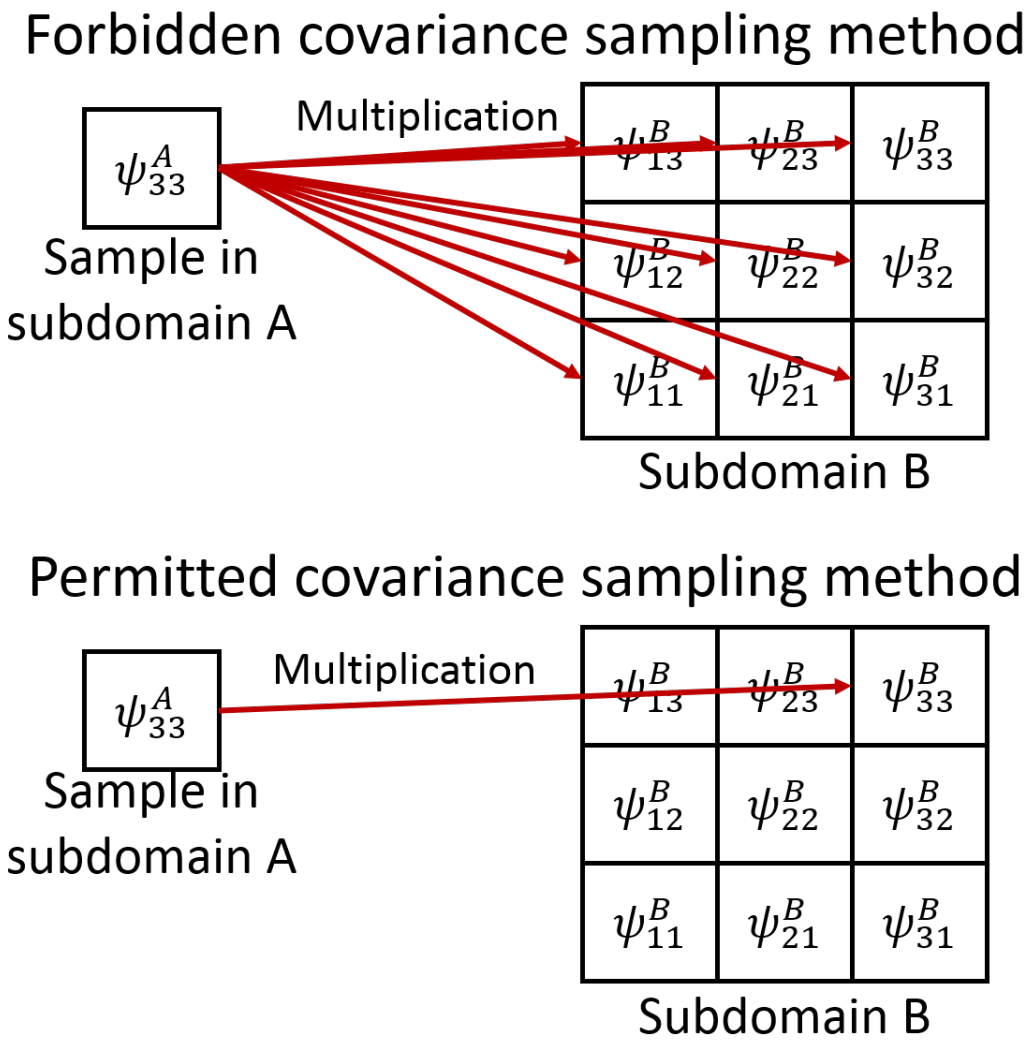


Figure 4.10: Two ways to use the samples in a pair of subdomains (A and B) to compute the multiples needed to generate auto-covariance of ψ between A and B.

4.10, top row). Mathematically, such an approach is exactly the same as taking the mean of all the values in A and B and use that as the sample. This approach kills off the increase in available samples. Instead, we should create the multiplication samples as per the second row in Figure 4.10. In other words, there is a one-to-one pairing of samples across the two subdomains. That will utilize all the samples, without causing undersampling.

Formulation of new model Following this sampling method, we can define the mathematical formulation of the new model. First, we will consider the (i, j, k)-th subdomain (counting from the bottommost-southmost-westmost subdomain) on the model grid to be the n-th subdomain. For the ease of computation, we will use subdomains that have the size of 1° by 1° by 10 model levels (to be precise, 20 by 20 by 10 model grid points). This

means that the model grid is split into a 3D array of $16 \times 15 \times 7$ subdomains. Thus, we can define the mapping from (i, j, k) to n,

$$n\left(i, j, k\right) = i + j * 16 + k * 16 * 15. \tag{4.3}$$

Then, we can define the full mathematical formulation of the model (see Figure 1.3 for an intuitive schematic illustration of the model).

$$\begin{aligned}
&\boldsymbol{\psi'_{n;t}}^{(3)} = \boldsymbol{\psi_{n;t}}^{(3)}, \quad \boldsymbol{\phi'_{n;t}}^{(3)} = \boldsymbol{\phi_{b;t}}^{(3)} - \sum_{i'=1}^{16} \sum_{j'=1}^{15} \sum_{k'=1}^{7} \alpha^{(3)}_{\phi,\psi'}\left(n, n'\right) \boldsymbol{\psi'_{n';t}}^{(3)} \\
&C^{(3)}_{\psi',\psi';\, n,n'} \equiv \sum_{t=1}^{N_T} \frac{\boldsymbol{\psi'_{n;t}}^{(3)} \cdot \boldsymbol{\psi'_{n';t}}^{(3)}}{20 * 20 * 10 * N_T}, \quad C^{(3)}_{\phi,\psi';\, n,n'} \equiv \sum_{t=1}^{N_T} \frac{\boldsymbol{\phi_{n;t}}^{(3)} \cdot \boldsymbol{\psi'_{n';t}}^{(3)}}{20 * 20 * 10 * N_T} \\
&\alpha^{(3)}_{\phi,\psi'}\left(n, n'\right) \equiv \left\{ \boldsymbol{C_{\phi,\psi'}}^{(3)} \left(\boldsymbol{C_{\psi',\psi'}}^{(3)} \right)^{-1} \right\}_{nn'}.
\end{aligned} \tag{4.4}$$

4.5 Summary

In summary, we have shown that the columns of the auto-covariance submatrices of $\boldsymbol{B}$ implicates deep convection and that the cross-covariance submatrices are not horizontally invariant and isotropic. Following this variance and anisotropic idea, we argued that the assumption behind two commonly used regression coefficient models (Figures 1.1 and 1.2) do not hold. A new assumption was then formulated from realizing that within the vicinity of a specified point, auto-covariances tend to be approximately equal to variances. A new model was then formulated based on this realization.

Chapter 5

Results from regression coefficient models

Overview The three regression coefficient models are now applied onto the outputs of the GEN_BE system. The first model assumes that the regression coefficients are horizontally invariant and isotropic, and that there is no relationship between different model levels (Figure 1.1). The second model also assumes horizontal invariance, but does not assume that error variables are disconnected across different model levels (Figure 1.2). The third method simply follows the neighbourhood equivalence assumption (Figures 1.3 and 1.4. The performance of the modelled regression coefficients are evaluated by their ability to remove cross-covariances after applying the first step of the CVT. Another result of interest would be the amount of $\boldsymbol{\phi}$ variance that the linear $\boldsymbol{\phi}$–$\boldsymbol{\psi}$ relationship can account for. If most of the variance of $\boldsymbol{\phi}$ can be explained by the $\boldsymbol{\phi}$–$\boldsymbol{\psi}$ relationship, we can consider $\boldsymbol{\phi'}$ to be completely decoupled from all other meteorological error variables in future works.

5.1 Performance

Basis of metric If a perfect set of regression coefficients can be produced, then the cross-covariance between variables in $\boldsymbol{x^{b'}}$ is zero[1]. We can utilize this to determine the performance of the modelled regression coefficient matrices.

[1]See the proof for Eqn (B.1) in Appendix B for the details.

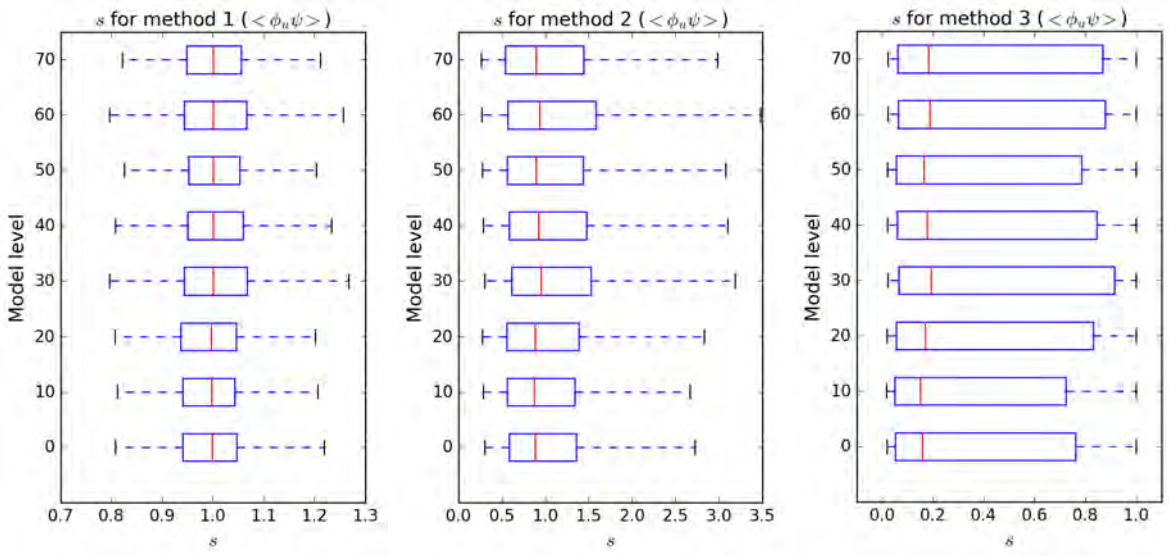


Figure 5.1: Box-and-whiskers plots of $s(\boldsymbol{r}, \boldsymbol{r'})$. These metrics were evaluated at the 72 $\boldsymbol{r'}$ locations specified in the text, for the first model (left), the second model (middle) and the third model (right). The boxes bound the interquartile range, the red vertical line indicates the median and the whiskers bound the 10th and 90th percentiles of the $9 * 160 * 150$ evaluations at each of the 8 level.

Description of s We used a metric to evaluate the performance of the regression coefficient models pairs of points in the model. The mathematical formulation is

$$s(\boldsymbol{r}, \boldsymbol{r'}) \equiv \left\| \sum_{t=1}^{N_T} \phi'(\boldsymbol{r}, t)\, \psi'(\boldsymbol{r'}, t) \right\| \Bigg/ \left\{ \sum_{t=1}^{N_T} \phi(\boldsymbol{r}, t)\, \psi(\boldsymbol{r'}, t) \right\}, \tag{5.1}$$

where $\boldsymbol{r}$ and $\boldsymbol{r'}$ are position vectors. This metric computes the ratio of the residual cross-covariance after the first step of CVT, against the original cross-covariance. The ideal value of s is thus zero. Naturally, the best model would be one that turns s to the smallest absolute value, when employed in the CVT.

s computer memory constraint Due to computational memory constraints, only 72 $\boldsymbol{r'}$ were sampled. Likewise, $\boldsymbol{r}$ was sampled every two grid points in the horizontal, and every grid point in the vertical. The 72 $\boldsymbol{r'}$ positions are split across 8 model levels[2], 9 per model level. The horizontal positions of each set of 9 points are the same as the green dots in Figure 4.8.

Plotting the s data The s valuations from these 72 test locations are grouped according to the 8 model levels of the test locations and then plotted in Figure 5.1. In other words, each box and whisker plot in Figure 5.1 contains $9 * 160 * 150$ evaluations of s.

[2]Model level indices are $k = 0, 10, 20, \cdots, 60, 70$.

Third model outperforms the first model According to Figure 5.1, 50% of the s evaluations for third model (below the median) have values less than 0.2. In other words, after applying the CVT, 50% of all tested points have cross-covariances with magnitudes that are less than 20% of the original cross-covariances. Furthermore, median to 90% percentile evaluations of s for the third model indicate that 40% of the new cross-covariances range from 20% to 100% of the original cross-covariances. In contrast, 40% of the s evaluations for the first model (10th percentile to median) indicate that the absolute value of the new cross-covariances are more than 80% of the original. To make matters worse, 40% of the s evaluations for the first model (median to 90th percentile) indicate that applying the first model resulted in larger cross-covariances! Clearly, the third model is far more adept at removing cross-covariances than the first model.

Third model outperforms the second model Similarly, the third model outperforms the second model. The 10th percentile to median evaluations of s for the second model indicates that applying the second model can only remove less than 80% of the cross-covariances for 40% of the points. In contrast, for the same range, the third model removes more than 80%. To make matters worse for the second model, the median to 90th percentile range of s evaluations indicate that applying the second model amplifies the cross-covariances! For the same percentile range, the third model is still able to constrain the absolute values of the new cross-covariances to be smaller than that of the original cross-covariances. The third model is clearly more suited to removing cross-covariances than the second model.

First and second model are unsuitable for our region The first model is clearly unsuitable for our region as it only modifies the cross-covariances by 20%, and it does amplify cross-covariances for 50% of the evaluated cases. While the second model is able to remove between more than 40% of the cross-covariances for 25% of the points tested (below the 25th percentile in Figure 5.1), the second model actually amplifies cross-covariances for roughly half the tested points (above the medians in Figure 5.1). Furthermore, 25% of the tested points (upwards of the 75th percentile) indicate that the second model actually amplifies the cross-covariances by more than 30%, and 10% of the tested points (upward of the 90th percentile) indicate that the second model actually increased the cross-covariance by more than 150%. These results mean that the first and second model are highly unsuitable for use in the CVT for our region of interest.

Violation of assumptions is the reason for first and second models' performance As per the discussion in the preceeding chapter, it is hardly surprising that the

performance of the first and second models is terrible in our region. In the preceeding chapter, we observed from Figure 4.6 that cross-covariances vary tremendously with the horizontal position considered and show great anisotropy around said position. As explained, this violates the horizontal invariance and isotropy assumption that underlies the first and second model. In other words, the bad performance by the two models are as expected.

5.2 Variance occupation

An additional boon for third model Of course, if the third model can also account for most of the variance of $\boldsymbol{\phi}$, future work utilizing this model can also neglect relationships between $\boldsymbol{\phi'}$ and other meteorological variables. To examine this, we defined the following fraction

$$f(\boldsymbol{r}) \equiv \left\{ \sum_{t=1}^{N_t} \phi(\boldsymbol{r}) \phi_b(\boldsymbol{r}) \right\} \Big/ \left\{ \sum_{t=1}^{N_t} \phi(\boldsymbol{r}) \phi(\boldsymbol{r}) \right\}, \tag{5.2}$$

where

$$\boldsymbol{\phi_b} \equiv \boldsymbol{\alpha}_{\phi,\psi} \boldsymbol{\psi'}.$$

In essence, if f is close to 1, then the linear relationship between ϕ and ψ is able to account for most of the variance in ϕ. The f evaluated for all three methods, across the model grid, are aggregated by model levels and displayed as box-and-whisker plots in Figure 5.2.

Third model explains most of ϕ variance Two observations can be made from Figure 5.2. First of all, from the horizontal scales of Figure 5.2, it is clear that utilizing the third model results in the greatest f valuations (which are very close to 1). In other words, when the third model is used, most of the variance in ϕ can be explained by ψ. In contrast, the other two methods only explain a small portion of the variance. As such, if the third model is implemented, subsequent studies can simply consider error variables apart from ψ to be largely unrelated to ϕ.

Fall-off in third model's f at upper level The second observation is that in the upper levels, there is a notable fall in the f valuations of the third model. This is most likely because the subdomains only cover the bottom 70 levels of the model and neglect the remaining 9 levels.

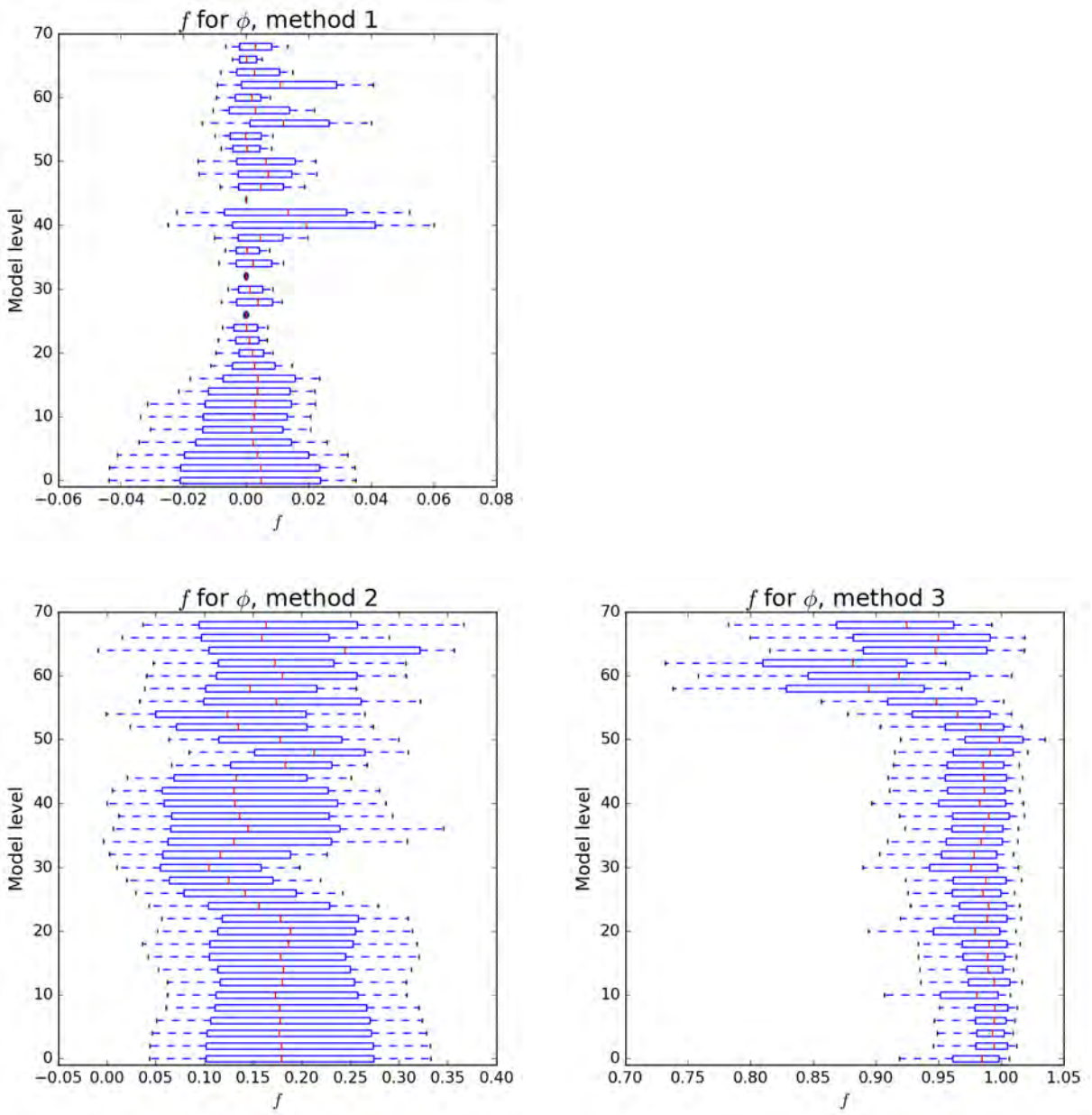


Figure 5.2: Box-and-whiskers diagrams of f obtained by the three methods of modelling regression coefficients. The boxes bound the interquartile range, the red vertical line indicates the median and the whiskers bound the 10th and 90th percentiles of the $9 * 160 * 150$ evaluations at each of the 8 level.

5.3 Summary

In summary, because the assumptions underlying the two commonly used models do not hold in reality, while the third model's assumption is based off the features observed in $\boldsymbol{B}$, the third model shows the greatest ability to remove cross-covariances. Furthermore, the third model has the additional benefit of accounting for most of $\boldsymbol{\phi}$'s variance, meaning that $\boldsymbol{\phi'}$ can be essentially considered to be decoupled from all other error variables introduced by future work.

Chapter 6

Conclusions and future work

Commonly used models are inadequate The goal of this project is to determine an appropriate model for the regression coefficient matrices involved in the CVT, over the Maritime Continent, during the Northeast monsoon season, using forecasts from the Singapore Variable resolution model. After sanity checking SINGV forecasts (5th December 2015 to 19th December 2015), we examined several columns of $\boldsymbol{B}$ and realized that the cross-covariance between $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ exhibits large variations depending on the horizontal location of the reference points. This means that any regression coefficient model based on the assumption that regression coefficients are invariant in the horizontal are unlikely to work well in the Maritime Continent, during the Northeast monsoon. Later tests confirmed their inadequate performance in this scenario.

Neighbourhood equivalence assumption and the third model is the best Examinations of the columns of $\boldsymbol{B}$ also revealed that there is a tendency for the auto-covariances between evaluations of an error variable at two nearby points to be approximately equal to their variances. A third model was formulated based on this assumption (neighbourhood equivalence assumption), and then tested. When utilized in a CVT, the third model removed most of the cross-covariances $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, making it an adequate model for our scenario, fulfilling our goal.

Third model's affordable computational cost It must be said here that the third model has a much higher computational cost than the other two. This is to be expected: the regression coefficient matrix generated by the third model is much bigger than the first and second model. However, this cost is still affordable! We were able generate the regression coefficients within 6 hours of computation in a laptop with an Intel i7 quad-core and 16 gigabytes of Random Access Memory, without using parallel computing

(only 1 core was employed). Furthermore, when we ran a CVT based on the third model, using the same laptop, without parallel programming, the first step of the CVT took only 3 minutes. Also, the algorithm that utilizes the third model for the CVT is actually an embarrassingly parallel algorithm (many repetitive matrix multiplications). It seems likely that if the model was rewritten for parallel computing, it would be affordable for data assimilation in operational forecasting.

Future work: increase number of subdomains in third model Even though our new model has much better performance than the other two, it is not optimal yet. In Chapter 4, we mentioned that the neighbourhood equivalence assumption appears to hold for 5 model levels. However, we utilized subdomains with 10 model levels each due to computational time limitations. Future work on this model can further enhance the performance of the model by using half the number of model levels per subdomain.

Future work: apply neighbourhood equivalence assumption to CVT's second step Aside from subdomain size optimization, another avenue of future work would be to apply the neighbourhood equivalence assumption to the second step of the CVT: the removal of auto-covariances in $\boldsymbol{x}^{\boldsymbol{b}'}$. We can safely say from Figure 4.5 that the assumption should hold fairly well when dealing with the auto-covariance of ψ as it is unchanged by the first step of the CVT. However, work is needed to determine if the assumption does hold for $\boldsymbol{\phi}'$. If that is true, it may be feasible to model the auto-covariance of $\boldsymbol{x}^{\boldsymbol{b}'}$ using the eigenvectors of the auto-covariance matrix of the subdomains. Future work can then test this new model against other pre-existing models of auto-covariances.

Future work: generality of neighbourhood equivalence assumption and third model A third avenue of future work would be to determine whether the neighbourhood equivalence assumption can be utilized on other commonly used variables in data assimilation (*e.g.*, air temperature and relative humidity). If so, then it may be possible to utilize the third model when removing their cross-covariances through the CVT.

Outlook for the future Advancements in computer technology have worked wonders for operational forecasting. In a mere three decades, the resolution of global circulation forecast models has improved from several hundreds of kilometres to the current less than 20 km resolution (the European Centre for Medium Range Weather Forecasts currently uses a 9 km horizontal resolution in their 36-hour global forecasts). In fact, at the time of writing, numerical weather prediction and limited area models with a horizontal resolution of less than 5 km are fairly common. It would not be surprising if sub-convective scale

(sub-kilometre scale) forecast models become common by the end of the next decade. However, it is difficult to expect the density of observation to match up to the resolution of forecast models. Clearly, $\boldsymbol{B}$'s capacity for spreading the influence of observations from their sparse locations throughout a model's initial conditions during data assimilation will prove to be important in this future. The art of data assimilation and studies on modelling $\boldsymbol{B}$ will remain invaluable for a long time.

Bibliography

Australian Government, Bureau of Meteorology. 2016. *Analysis chart archive*.

Badan Perencanaan Pembangunan Nasional. 2013. *Indonesian population projection*.

Bannister, R. N. 2008a. *A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances*, Quarterly Journal of the Royal Meteorological Society **134**, no. 637, 1951–1970, available at `arXiv:0801.1618v2`.

———. 2008b. *A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics*, Quarterly Journal of the Royal Meteorological Society **134**, no. 637, 1971–1996.

Berre, L., S. Ecaterina Stefanescu, and M. Belo Pereira. 2006. *The representation of the analysis effect in three error simulation techniques*, Tellus A **58**, no. 2, 196–209.

Bijlsma, S. J., L. M. Hafkenscheid, and Peter Lynch. 1986. *Computation of the streamfunction and velocity potential and reconstruction of the wind field*, Monthly Weather Review **114**, no. 8, 1547–1551.

Bouttier, F. and P. Courtier. 2002. *Data assimilation concepts and methods*, Meteorological training course lecture series.

Caron, J-F. and L. Fillion. 2010. *An examination of background error correlations between mass and rotational wind over precipitation regions*, Monthly Weather Review **138**, no. 2, 563–578.

Chang, C-P., P. a. Harr, and H-J. Chen. 2005. *Synoptic disturbances over the equatorial South China Sea and western Maritime Continent during boreal winter*, Monthly Weather Review **133**, no. 3, 489–503.

Chang, C-P., M-M. Lu, and H. Lim. 2016. *Monsoon convection in the Maritime Continent: interaction of large-scale motion and complex terrain*, Meteorological Monographs - Multiscale Convection-Coupled Systems in the Tropics: A Tribute to Dr. Michio Yanai **56**, 6.1–6.29.

Chen, Y., S. R. H. Rizvi, X. Y. Huang, J. Min, and X. Zhang. 2013. *Balance characteristics of multivariate background error covariances and their impact on analyses and forecasts in tropical and Arctic regions*, Meteorology and Atmospheric Physics **121**, no. 1-2, 79–98.

Daley, R. 1991. *Atmospheric data analysis*, Vol. 15, Cambridge University Press.

Descombes, G., T. Auligné, F. Vandenberghe, D. M. Barker, and J. Barré. 2015. *Generalized background error covariance matrix model (GEN-BE v2.0)*, Geoscientific Model Development **8**, no. 3, 669–696.

Ding, Y. 1994. *The summer monsoon in East Asia*, Monsoons over china, pp. 1–90.

Fong, M. and L. K. Ng. 2012. *The weather and climate of Singapore*, Meteorological Service Singapore, Singapore.

Kalnay, E. 2003. *Atmospheric modelling, data assimilation and predictability*, Vol. 129, Cambridge University Press.

Lanczos, C. 1957. *Applied analysis*, Dover Publications.

Lorenc, A. C. 1986. *Analysis methods for numerical weather prediction*, Quart J. R. Mrt. Soc. **112**, 1177–1194.

Parrish, D. F. and J. C. Derber. 1992. *The National Meteorological Center's Spectral Statistical-Interpolation Analysis System*, Vol. 120.

Wang, B. 2006. *The Asian monsoon*, Springer.

Wu, W-S., R. J. Purser, and D. F. Parrish. 2002. *Three-dimensional variational analysis with spatially inhomogeneous covariances*, Monthly Weather Review **130**, no. 12, 2905–2916.

Appendix A

The roles of the background error coviance matrix in DA

Overview Here, the roles of $\boldsymbol{B}$ in data assimilation (mentioned in Chapter 1) will be shown through the single-observation and double-observation assimilation solutions that minimize Eqn (1.4). As such, this chapter will be divided into two sections: the derivation of the solutions to Eqn (1.4) and the application of said solutions to demonstrate said roles of $\boldsymbol{B}$. Afterwards, this chapter will conclude with a brief summary. Note that the single-observation solution and several roles of $\boldsymbol{B}$ in data assimilation will be referenced in later chapters of this document to interpret columns of $\boldsymbol{B}$.

A.1 Solutions to Cost Function

A.1.1 Best linear unbiased estimate solution

Starting point – cost function As mentioned in Chapter 1, goal of data assimilation is to determine a model state that minimizes Eqn (1.4). In the field's parlance, this model state is called the analysis state, $\boldsymbol{X^a}$. $\boldsymbol{X^a}$ will be derived in this section. For convenience, the cost function is rewritten in tensor notation,

$$\begin{aligned} J(\boldsymbol{X}) = & X_i B_{ij}^{-1} X_j + X_i^b B_{ij}^{-1} X_j^b - X_i B_{ij}^{-1} X_j^b - X_i^b B_{ij}^{-1} X_j \\ & + Y_m R_{mn}^{-1} Y_n + H(\boldsymbol{X})_m R_{mn}^{-1} H(\boldsymbol{X})_n - Y_m R_{mn}^{-1} H(\boldsymbol{X})_n - H(\boldsymbol{X})_m R_{mn}^{-1} Y_n. \end{aligned}$$

Consider the gradient of the cost function:

$$\frac{\partial J(\boldsymbol{X})}{\partial x_l} = B^{-1}_{lj} X_j \; + \; X_i B^{-1}_{il} \; - \; B^{-1}_{lj} X^b_j \; - \; X^b_i B^{-1}_{il} \\ + \; \frac{\partial}{\partial x_l} \Bigg\{ H(\boldsymbol{X})_m \, R^{-1}_{mn} H(\boldsymbol{X})_n \; - \; Y_m R^{-1}_{mn} H(\boldsymbol{X})_n \; - \; H(\boldsymbol{X})_m \, R^{-1}_{mn} Y_n \Bigg\}.$$

Since $\boldsymbol{X} = \boldsymbol{X^a}$ minimizes the cost function, then,

$$\left.\frac{\partial J(\boldsymbol{X})}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} = 0 = B^{-1}_{lj} X^a_j \; + \; X^a_i B^{-1}_{il} \; - \; B^{-1}_{lj} X^b_j \; - \; X^b_i B^{-1}_{il} \\ + \; \left.\frac{\partial H(\boldsymbol{X})_m}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} R^{-1}_{mn} H(\boldsymbol{X^a})_n \; + \; H(\boldsymbol{X^a})_m \, R^{-1}_{mn} \left.\frac{\partial H(\boldsymbol{X})_n}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} \\ - \; Y_m R^{-1}_{mn} \left.\frac{\partial H(\boldsymbol{X})_n}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} \; - \; \left.\frac{\partial H(\boldsymbol{X})_m}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} R^{-1}_{mn} Y_n.$$

Symmetric $\boldsymbol{B}^{-1}$ and $\boldsymbol{R}^{-1}$ Before going any further, it is useful to note that $\boldsymbol{B}$ and $\boldsymbol{R}$ are, by definition, symmetric and positive definite matrices. In other words, $\boldsymbol{B}$ and $\boldsymbol{R}$ are invertible and their inverses are symmetric. Applying these ideas condenses the cost function differential:

$$\left.\frac{\partial J(\boldsymbol{X})}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} = 0 = 2\, B^{-1}_{lj} X^a_j \; - \; 2\, B^{-1}_{lj} X^b_j \; + \; 2 \left\{ \left.\frac{\partial H(\boldsymbol{X})_m}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} R^{-1}_{mn} H(\boldsymbol{X^a})_n \right\} \\ - \; 2 \left\{ Y_m R^{-1}_{mn} \left.\frac{\partial H(\boldsymbol{X})_n}{\partial x_l}\right|_{\boldsymbol{X}=\boldsymbol{X^a}} \right\}.$$

Gradient of H Defining

$$\boldsymbol{H} \equiv \boldsymbol{\nabla} H(\boldsymbol{X}) \Big|_{\boldsymbol{X}=\boldsymbol{X^a}} \tag{A.1}$$

simplifies the problem even further:

$$B^{-1}_{lj} X^a_j \; - \; B^{-1}_{lj} X^b_j \; + \; H_{lm} R^{-1}_{mn} H(\boldsymbol{X^a})_n \; - \; Y_m R^{-1}_{mn} H_{nl} \; = \; 0.$$

Note that the gradient operator in Eqn (A.1) is in model space. In other words, for a N-element state vector and a M-element observation vector, $\boldsymbol{H}$ is a $M \times N$ matrix. Thus,

$$B^{-1}_{lj} X^a_j \; - \; B^{-1}_{lj} X^b_j \; = \; + Y_m R^{-1}_{mn} H_{nl} \; - \; H_{lm} R^{-1}_{mn} H(\boldsymbol{X^a})_n \\ \Longrightarrow \boldsymbol{B}^{-1} \left(\boldsymbol{X^a} - \boldsymbol{X^b} \right) = \boldsymbol{H}^\top \boldsymbol{R}^{-1} \left\{ \boldsymbol{Y} - H(\boldsymbol{X^a}) \right\}$$

Tangent linear hypothesis To proceed any further, it is assumed that $\boldsymbol{X^a}$ and $\boldsymbol{X^b}$ are close enough for H to be linear. In other words, for $\boldsymbol{X^a} + \boldsymbol{\delta X} = \boldsymbol{X^b}$,

$$H\left(\boldsymbol{X^a} + \boldsymbol{\delta X} - \boldsymbol{\delta X}\right) \approx H\left(\boldsymbol{X^a} + \boldsymbol{\delta X}\right) - \boldsymbol{H}\boldsymbol{\delta X} = H\left(\boldsymbol{X^b}\right) + \boldsymbol{H}\boldsymbol{X^a} - \boldsymbol{H}\boldsymbol{X^b}. \quad \text{(A.2)}$$

This assumption is called the "tangent linear hypothesis" in literature. Thus,

$$\begin{aligned}
&\boldsymbol{B}^{-1}\left(\boldsymbol{X^a} - \boldsymbol{X^b}\right) = \boldsymbol{H}^\top \boldsymbol{R}^{-1}\left\{\boldsymbol{Y} - H\left(\boldsymbol{X^a} + \boldsymbol{\delta X} - \boldsymbol{\delta X}\right)\right\} \\
&\implies \boldsymbol{B}^{-1}\left(\boldsymbol{X^a} - \boldsymbol{X^b}\right) = \boldsymbol{H}^\top \boldsymbol{R}^{-1}\left\{\boldsymbol{Y} - H\left(\boldsymbol{X^b}\right) - \boldsymbol{H}\boldsymbol{X^a} + \boldsymbol{H}\boldsymbol{X^b}\right\} \\
&\implies \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)\left(\boldsymbol{X^a} - \boldsymbol{X^b}\right) = \boldsymbol{H}^\top \boldsymbol{R}^{-1}\left\{\boldsymbol{Y} - H\left(\boldsymbol{X^b}\right)\right\}
\end{aligned}$$

Assume invertibility of $\left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)$ To proceed any further, it is necessary to assume that $\left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)$ is invertible. Thus:

$$\left(\boldsymbol{X^a} - \boldsymbol{X^b}\right) = \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top \boldsymbol{R}^{-1}\left[\boldsymbol{Y} - H\left(\boldsymbol{X^b}\right)\right]$$

Tidying up In the literature, $\boldsymbol{K}$ is usually written as $\boldsymbol{B}\boldsymbol{H}^\top\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\top + \boldsymbol{R}\right)^{-1}$. For completeness, the two forms of $\boldsymbol{K}$ can be shown to be equivalent. Suppose that:

$$\boldsymbol{B}\boldsymbol{H}^\top\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\top + \boldsymbol{R}\right)^{-1} = \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top \boldsymbol{R}^{-1}$$

Then,

$$\begin{aligned}
\boldsymbol{B}\boldsymbol{H}^\top &= \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1} \boldsymbol{H}^\top \boldsymbol{R}^{-1}\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\top + \boldsymbol{R}\right) \\
&= \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1}\left(\boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{B} + \boldsymbol{I}\right)\boldsymbol{H}^\top \\
&= \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1}\left(\boldsymbol{H}^\top \boldsymbol{R}^{-1}\boldsymbol{H} + \boldsymbol{B}^{-1}\right)\boldsymbol{B}\boldsymbol{H}^\top \\
&= \boldsymbol{B}\boldsymbol{H}^\top
\end{aligned}$$

Thus, the two forms of $\boldsymbol{K}$ are consistent with each other. For consistency with the literature, the following form of the solution will be used for the rest of this document[1]:

$$\boldsymbol{X^a} = \boldsymbol{X^b} + \boldsymbol{B}\boldsymbol{H}^\top\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\top + \boldsymbol{R}\right)^{-1}\left[\boldsymbol{Y} - H\left(\boldsymbol{X^b}\right)\right]. \quad \text{(A.3)}$$

[1] An alternative derivation of the solution, based on minimizing the error of the analysis state, can be found in the appendix.

A.1.2 Single and double-observation BLUE solutions

Coincidence assumption The roles of $\boldsymbol{B}$ in data assimilation can be directly observed from considering the solution in Eqn (A.3) for the case where only one observation is assimilated, and for the case where two observations are assimilated. To simplify the problem, it will be assumed that the observation(s) is located at a model grid point(s), and concerns one (two) of the variables in $\boldsymbol{B}$.

H under coincidence assumption Under this collocation assumption, the operator H is effectively the matrix $\boldsymbol{H}$. Let the normalized basis vector for that model point and position in model basis be $\hat{\boldsymbol{X}}_{\boldsymbol{n}}$ (*i.e.*, the n-th element on $\boldsymbol{X}$) and likewise, the observation basis vector be $\hat{\boldsymbol{Y}}_{\boldsymbol{m}}$. As such, $\boldsymbol{H}$ for the single and double-observation cases are,

$$\boldsymbol{H} = 1\hat{\boldsymbol{Y}}\hat{\boldsymbol{X}}_{\boldsymbol{n}}{}^{\top}, \text{ and, } \boldsymbol{H} = 1\hat{\boldsymbol{Y}}_{\mathbf{1}}\hat{\boldsymbol{X}}_{\boldsymbol{n_1}}{}^{\top} + 1\hat{\boldsymbol{Y}}_{\mathbf{2}}\hat{\boldsymbol{X}}_{\boldsymbol{n_2}}{}^{\top}, \tag{A.4}$$

respectively.

Single-observation solution For the single-observation case, it is clear that $\boldsymbol{R}$ consists of a single element (σ^2). Applying the single-observation $\boldsymbol{R}$ and $\boldsymbol{H}$ into Eqn (A.3) yields the single-observation data assimilation equation

$$X_l^a - X_l^b = B_{ln}\frac{Y - X_n^b}{B_{nn} + \sigma^2}. \tag{A.5}$$

Note that subscripts refer to the corresponding elements in the relevant vector/matrix.

Double-observation solution For convenience, suppose that the two observations for the double-observation cases are uncorrelated. Then, $\boldsymbol{R} = \sigma_1^2\hat{\boldsymbol{Y}}_{\mathbf{1}}\hat{\boldsymbol{Y}}_{\mathbf{1}}{}^{\top} + \sigma_2^2\hat{\boldsymbol{Y}}_{\mathbf{2}}\hat{\boldsymbol{Y}}_{\mathbf{2}}{}^{\top}$. It can be shown that the corresponding BLUE solution is

$$\begin{aligned} X_l^a - X_l^b = &\frac{B_{ln_1}}{\gamma}\left(y_1 - X_{n_1}^b\right) + \frac{B_{ln_2}}{\gamma}\left(y_2 - X_{n_2}^b\right) \\ &+ \frac{B_{ln_1}B_{n_2n_2} - B_{n_1n_2}B_{ln_2}}{\gamma}\left(Y_1 - X_{n_1}^b\right) + \frac{B_{ln_2}B_{n_1n_1} - B_{ln_1}B_{n_1n_2}}{\gamma}\left(Yy_2 - X_{n_2}^b\right) \end{aligned} \tag{A.6}$$

where $\gamma \equiv B_{n_1n_1}B_{n_2n_2} + B_{n_1n_1}\sigma_2^2 + B_{n_2n_2}\sigma_1^2 + \sigma_1^2\sigma_2^2 - B_{n_1n_2}^2$.

Both solutions match It can also be shown that the double-observation solution can be reduced to the single-observation solution when either σ_1^2 or σ_2^2 is set to infinity. This makes sense as when there is infinite observation error, the observation's impact on data assimilation should be zero.

A.2 Roles of $\boldsymbol{B}$ in data assimilation

***B* weighs *Y* and $\boldsymbol{X^b}$ locally** The first role of $\boldsymbol{B}$ in data assimilation can be illustrated using Eqn (A.5). Expressing and rearranging Eqn (A.5) at the n-th element of the state vector gives,

$$X_n^a = \frac{B_{nn}}{B_{nn} + \sigma^2} Y + \frac{\sigma^2}{B_{nn} + \sigma^2} X_n^b. \tag{A.7}$$

Clearly, $\boldsymbol{B}$ serves to weigh the contributions of the observation and background state, according to the relative confidence for the two pieces of information.

***B* spreads $\boldsymbol{Y} - H\left(\boldsymbol{X^b}\right)$ globally** Aside from local weighing, $\boldsymbol{B}$ also spreads the difference between the observation and background state, throughout the model grid and across model variables. This can be inferred by considering Eqn (A.5) and the general form of $\boldsymbol{B}$ for L variables:

$$\boldsymbol{B} = \begin{bmatrix} \left\langle \boldsymbol{x^b}_1 \boldsymbol{x^b}_1^\top \right\rangle & \left\langle \boldsymbol{x^b}_1 \boldsymbol{x^b}_2^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_1 \boldsymbol{x^b}_l^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_1 \boldsymbol{x^b}_{L-1}^\top \right\rangle & \left\langle \boldsymbol{x^b}_1 \boldsymbol{x^b}_L^\top \right\rangle \\ \left\langle \boldsymbol{x^b}_2 \boldsymbol{x^b}_1^\top \right\rangle & \left\langle \boldsymbol{x^b}_2 \boldsymbol{x^b}_2^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_2 \boldsymbol{x^b}_l^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_2 \boldsymbol{x^b}_{L-1}^\top \right\rangle & \left\langle \boldsymbol{x^b}_2 \boldsymbol{x^b}_L^\top \right\rangle \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \left\langle \boldsymbol{x^b}_l \boldsymbol{x^b}_1^\top \right\rangle & \left\langle \boldsymbol{x^b}_l \boldsymbol{x^b}_2^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_l \boldsymbol{x^b}_l^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_l \boldsymbol{x^b}_{L-1}^\top \right\rangle & \left\langle \boldsymbol{x^b}_l \boldsymbol{x^b}_L^\top \right\rangle \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \left\langle \boldsymbol{x^b}_{L-1} \boldsymbol{x^b}_1^\top \right\rangle & \left\langle \boldsymbol{x^b}_{L-1} \boldsymbol{x^b}_2^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_{L-1} \boldsymbol{x^b}_l^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_{L-1} \boldsymbol{x^b}_{L-1}^\top \right\rangle & \left\langle \boldsymbol{x^b}_{L-1} \boldsymbol{x^b}_L^\top \right\rangle \\ \left\langle \boldsymbol{x^b}_L \boldsymbol{x^b}_1^\top \right\rangle & \left\langle \boldsymbol{x^b}_L \boldsymbol{x^b}_2^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_L \boldsymbol{x^b}_l^\top \right\rangle & \cdots & \left\langle \boldsymbol{x^b}_L \boldsymbol{x^b}_{L-1}^\top \right\rangle & \left\langle \boldsymbol{x^b}_L \boldsymbol{x^b}_L^\top \right\rangle \end{bmatrix}. \tag{A.8}$$

$\boldsymbol{X^a}$ is the sum of $\boldsymbol{X^b}$ and a multiple of a column of $\boldsymbol{B}$. Clearly, any column of $\boldsymbol{B}$ cuts through an auto-covariance matrix, and $L-1$ cross-covariance matrices. The auto-covariance matrix's column spreads the observation-background difference throughout the model grid, for the assimilated variable. At the same time, the columns of the $L-1$ cross-

covariance matrices spreads the observation-background difference to different variables, across the entire model grid. In other words, $\boldsymbol{B}$ spreads the influence of $\boldsymbol{Y} - H\left(\boldsymbol{X^b}\right)$ throughout the entire set of initial conditions.

$\boldsymbol{B}$ overlaps the influences of different observations When multiple observations are assimilated, $\boldsymbol{B}$ allows their respective influences to overlap, as though two single-observation solutions were used. This can be inferred from the first line of the double-observation solution, Eqn (A.6):

$$\frac{B_{ln_1}}{\gamma}\left(y_1 - X^b_{n_1}\right) + \frac{B_{ln_2}}{\gamma}\left(y_2 - X^b_{n_2}\right).$$

Notice that for a N-element state vector, for all $l \in [1, N]$, $l \in \mathbb{Z}$, these terms do not cancel out or vanish. As such, the two observation-background differences are included in every point on the grid and across all variables, weighed by the corresponding columns of $\boldsymbol{B}$. In other words, $\boldsymbol{B}$ causes the influence of observations to overlap in the model grid.

$\boldsymbol{B}$ causes observation interactions Aside from the overlapping effect, $\boldsymbol{B}$ also causes the observation-background differences to interaction at grid points and/or variables that differ from the observation location and/or variables. The second line of the double-observation BLUE solution in Eqn (A.6) alludes to this effect:

$$\frac{B_{ln_1}B_{n_2n_2} - B_{n_1n_2}B_{ln_2}}{\gamma}\left(y_1 - X^b_{n_1}\right) + \frac{B_{ln_2}B_{n_1n_1} - B_{ln_1}B_{n_1n_2}}{\gamma}\left(y_2 - X^b_{n_2}\right)$$

In other words, at points/variables different from the observations, an additional interaction between the two observation-background differences pop up. $\boldsymbol{B}$ supplies the interaction itself.

$\boldsymbol{B}$ limits observation interactions It is also interesting to note that the interacting term from one observation vanishes at the location/variable of the other observation. Specifically, $B_{ln_1}B_{n_2n_2} - B_{n_1n_2}B_{ln_2} = 0$ when $l = n_2$, and, $B_{ln_2}B_{n_1n_1} - B_{ln_1}B_{n_1n_2} = 0$ when $l = n_1$. This means that at the position/variable of one observation, the only influence from the other observation comes from the overlapping part. In other words, $\boldsymbol{B}$ limits observation interactions to places/variables away from observation points/variables.

A.3 Summary of derivation

In summary, a general solution to the cost function of Eqn (1.4) have been derived under the tangent linear hypothesis and the assumption that $\left(\boldsymbol{B}^{-1}+\boldsymbol{H}^{\top}\boldsymbol{R}^{-1}\boldsymbol{H}\right)$ is invertible. Two specific solutions (single-observation and double-observation) were then derived from the general solution, and then utilized to demonstrate that $\boldsymbol{B}$ has the following roles in data assimilation:

1. $\boldsymbol{B}$ weighs $\boldsymbol{Y}$ and $\boldsymbol{X^b}$ locally,
2. $\boldsymbol{B}$ spreads $\boldsymbol{Y}-H\left(\boldsymbol{X^b}\right)$ globally,
3. $\boldsymbol{B}$ overlaps influences of different observations,
4. $\boldsymbol{B}$ causes observation interactions, and,
5. $\boldsymbol{B}$ limits observation interactions.

In the later chapters, the first two roles of $\boldsymbol{B}$ and the single-observation solutions will be referenced when interpreting the results of this project's computations.

Appendix B

Details about the first step of CVT

Purpose of CVT As described in the introduction, the goal of the CVT method is to transform the inputted state vector of background error variables ($\boldsymbol{x^b}$) into a vector of control variables ($\boldsymbol{x^{b''}}$) possessing zero covariances. When this happens, $\boldsymbol{B}$ is effectively transformed into an easily managed diagonal matrix.

Diagonalisation inspiration The CVT model actually emulates the effect of applying $\boldsymbol{B}$'s eigenvector matrices to itself, *i.e.*, linear algebra's diagonalisation. Consider:

$$\boldsymbol{B} = \boldsymbol{V}\,\boldsymbol{S}\,\boldsymbol{V}^\top$$

where $\boldsymbol{V}$ and $\boldsymbol{S}$ are matrices of the same dimensions as $\boldsymbol{B}$, the columns of $\boldsymbol{V}$ contain the eigenvectors of $\boldsymbol{B}$ and $\boldsymbol{S}$ is a diagonal matrix of the eigenvalues. In other words, if one defines $\boldsymbol{x^{b''}} \equiv \boldsymbol{V}^\top \boldsymbol{x^b}$, *i.e.*, $\boldsymbol{V}\,\boldsymbol{x^{b''}} = \boldsymbol{x^b}$, then,

$$\left\langle \boldsymbol{x^{b''}}\,\boldsymbol{x^{b''\top}} \right\rangle = \left\langle \boldsymbol{V}^\top \boldsymbol{x^b}\,\boldsymbol{x^{b\top}}\,\boldsymbol{V} \right\rangle = \boldsymbol{V}^\top \left\langle \boldsymbol{x^b}\,\boldsymbol{x^{b\top}} \right\rangle \boldsymbol{V} = \boldsymbol{V}^\top \boldsymbol{V}\,\boldsymbol{S}\,\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{S}.$$

In other words, the best CVT operator is none other than $\boldsymbol{V}^\top$. However, given that $\boldsymbol{B}$ must be fully determined before any computation of $\boldsymbol{V}$, this optimal operator cannot be used.

Formulation of $\boldsymbol{U_p}$ Instead, the CVT method seeks to emulate the diagonalizing effect of $\boldsymbol{V}$ through the two stages described in the introduction. We will focus on the cross-

covariance removal stage, whose operator is $\boldsymbol{U_p}$[1]. The transformation is simply

$$\boldsymbol{x^b} = \boldsymbol{U_p}\boldsymbol{x^{b'}},$$

where $\boldsymbol{x^b}$ and $\boldsymbol{x^{b'}}$ are as defined in Chapter 1. It should be apparant that

$$\boldsymbol{U_p} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{\alpha}_{1,2} & \boldsymbol{I} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\alpha}_{1,L-1} & \boldsymbol{\alpha}_{2,L-1} & \cdots & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{\alpha}_{1,L} & \boldsymbol{\alpha}_{2,L} & \cdots & \boldsymbol{\alpha}_{L-1,L} & \boldsymbol{I} \end{bmatrix}.$$

This also means that

$$\boldsymbol{B} = \boldsymbol{U_p} \left\langle \boldsymbol{x^{b'}} \boldsymbol{x^{b'}}^\top \right\rangle \boldsymbol{U_p}^\top.$$

Decorrelated $\boldsymbol{x^{b'}}$ If $\boldsymbol{x^{b'}}$ has zero cross-covariances (over time), then $\left\langle \boldsymbol{x^{b'}} \boldsymbol{x^{b'}}^\top \right\rangle$ is a block diagonal matrix. We will prove this zero cross-covariance property,

$$\left\langle \boldsymbol{x^{b'}}_l \, \boldsymbol{x^{b'}}_m^\top \right\rangle = 0 \qquad \forall \; l \neq m. \tag{B.1}$$

First layer of proof for Eqn (B.1) The proof for Eqn (B.1) will be done inductively. First, consider the scenario where there are only 2 background error variables, then $\boldsymbol{x^{b'}}_1 = \boldsymbol{x^b}_1$, $\boldsymbol{x^{b'}}_2 = \boldsymbol{x^b}_2 - \boldsymbol{\alpha}_{2,1}\boldsymbol{x^{b'}}_1$, and, $\boldsymbol{\alpha}_{2,1} \equiv \left\langle \boldsymbol{x^b}_2 \, \boldsymbol{x^{b'}}_1^\top \right\rangle \left\langle \boldsymbol{x^{b'}}_1 \, \boldsymbol{x^{b'}}_1^\top \right\rangle^{-1}$. Thus,

$$\begin{aligned} \left\langle \boldsymbol{x^{b'}}_2 \, \boldsymbol{x^{b'}}_1^\top \right\rangle &= \left\langle \boldsymbol{x^b}_2 \, \boldsymbol{x^{b'}}_1^\top \right\rangle - \left\langle \boldsymbol{\alpha}_{2,1} \boldsymbol{x^{b'}}_1 \, \boldsymbol{x^{b'}}_1^\top \right\rangle \\ &= \left\langle \boldsymbol{x^b}_2 \, \boldsymbol{x^{b'}}_1^\top \right\rangle - \left\langle \boldsymbol{x^b}_2 \, \boldsymbol{x^{b'}}_1^\top \right\rangle \left\langle \boldsymbol{x^{b'}}_1 \, \boldsymbol{x^{b'}}_1^\top \right\rangle^{-1} \left\langle \boldsymbol{x^{b'}}_1 \, \boldsymbol{x^{b'}}_1^\top \right\rangle = \boldsymbol{0}. \end{aligned}$$

Since $\left\langle \boldsymbol{x^{b'}}_1 \, \boldsymbol{x^{b'}}_2^\top \right\rangle = \left\langle \boldsymbol{x^{b'}}_2 \, \boldsymbol{x^{b'}}_1^\top \right\rangle^\top$, then $\left\langle \boldsymbol{x^{b'}}_1 \, \boldsymbol{x^{b'}}_2^\top \right\rangle = \boldsymbol{0}$. As such, Eqn (B.1) holds for the 2 background error variable case.

Inductive proof for Eqn (B.1) Now, assume that Eqn (B.1) holds for a set of n transformed background error variables and an additional transformed background error is added. Under this supposition, for Eqn (B.1) to hold for $n+1$ transformed background error variables, we just need to show that

$$\left\langle \boldsymbol{x^{b'}}_{n+1} \, \boldsymbol{x^{b'}}_m^\top \right\rangle = 0 \quad \forall \; m < n+1.$$

[1] In the field, $\boldsymbol{U_p}$ is often called the physical balance operator.

By the definitions in Chapter 1,

$$\left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle = \left\langle \boldsymbol{x}^{\boldsymbol{b}}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \sum_{p=1}^{n} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle$$

$$= \left\langle \boldsymbol{x}^{\boldsymbol{b}}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \left(\sum_{p=1}^{m} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle + \boldsymbol{\alpha}_{n+1,m} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{m}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle + \sum_{p=m+1}^{n} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle \right)$$

$$= \left\langle \boldsymbol{x}^{\boldsymbol{b}}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \boldsymbol{\alpha}_{n+1,m} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{m}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \left(\sum_{p=1}^{m} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle + \sum_{p=m+1}^{n} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle \right)$$

$$= \left\langle \boldsymbol{x}^{\boldsymbol{b}}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \left\langle \boldsymbol{x}^{\boldsymbol{b}}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \left(\sum_{p=1}^{m} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle + \sum_{p=m+1}^{n} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle \right)$$

$$= -\sum_{p=1}^{m} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle - \sum_{p=m+1}^{n} \boldsymbol{\alpha}_{n+1,p} \left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle .$$

In the two remaining summations, it is clear that $p \leq n$. Since Eqn (B.1) is supposed to hold for the first n variables, then, $\left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{p}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle = \mathbf{0}$ for $p \leq n$. This clearly means that

$$\left\langle \boldsymbol{x}^{\boldsymbol{b}'}{}_{n+1}\, \boldsymbol{x}^{\boldsymbol{b}'\top}{}_{m} \right\rangle = \mathbf{0}. \tag{B.2}$$

Notice that Eqn (B.2) rests on the supposition that Eqn (B.1) holds for the first n transformed background error variables, where n is an unspecified integer. In other words, Eqn (B.1) should also hold for n transformed background error variables if it holds for $n-1$ transformed background error variables. If we repeat this backtracking continuously, we will eventually arrive at the 2 transformed background error variables scenario. Eqn (B.1) has been previously proven to hold for that particular scenario. In other words, by induction, Eqn (B.1) holds for any number of transformed background error variables.

Appendix C

Additional plots

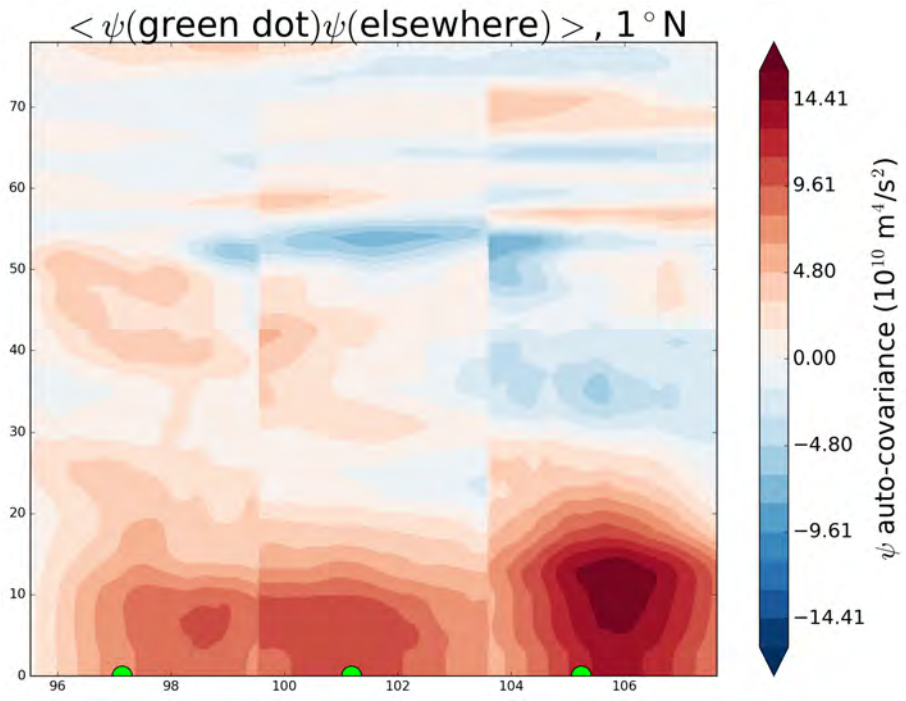


Figure C.1: The contour plot shows the covariance of ψ at the green dots on the surface, at 1°N, and ψ everywhere on a vertical slice of ϕ, along the 1°N latitude. This plot is separated into 3 vertical strips, each centred on a the green dot, spanning 2° longitude in the horizontal and all 79 model levels in the vertical. The horizontal axis is longitude and the vertical axis is model level. This plot is referenced to in the discussion near Figure 4.5

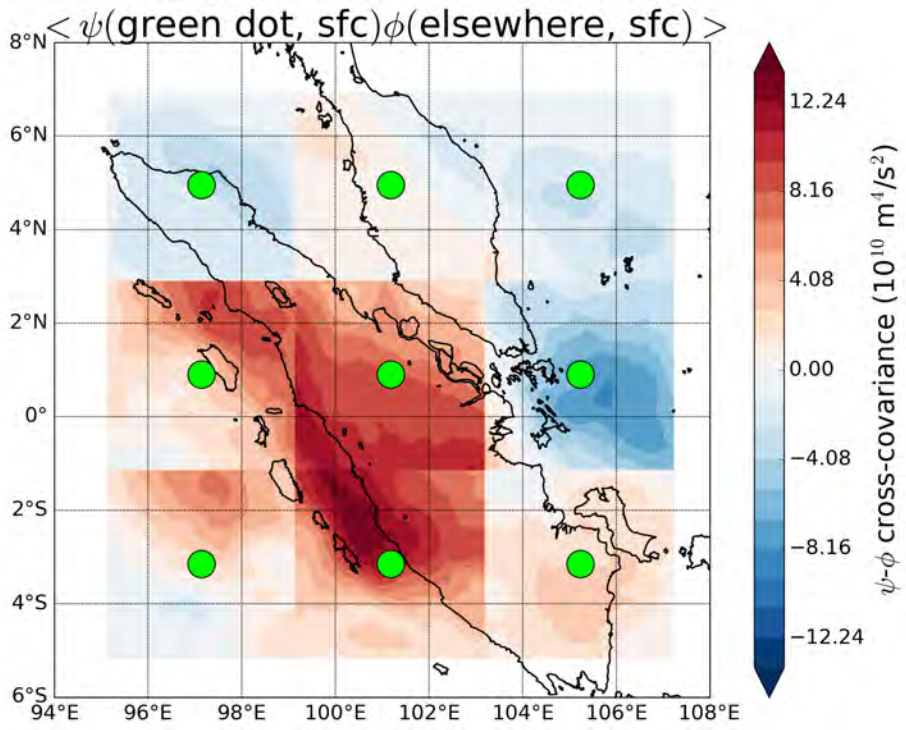


Figure C.2: The contour plot shows the covariance between ψ at the green dots and ϕ in squares centred on the green dots, on the surface level. The dimensions of each square is 450 km $\times$ 450 km, or roughly 4° $\times$ 4°. The landmasses in the domain are outlined in thick, solid black lines.

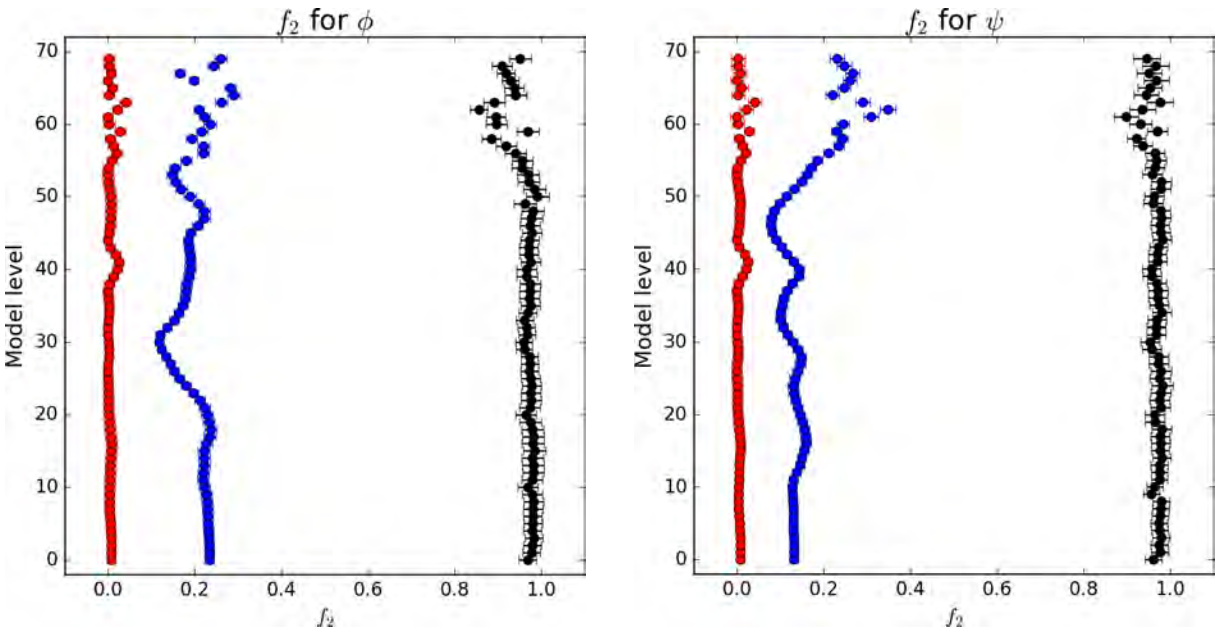


Figure C.3: Contribution fractions (Chen et al., 2013) for the three ways of modelling regression coefficients. The red data points are for the first model, the blue for the second model and the black for the third model.