

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Partial analysis increments: theory and application to
optimisation of vertical localisation

verfasst von | submitted by

Martin Finn Parker

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 066 821

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Mathematik

Betreut von | Supervisor:

Ass.-Prof. Dr. Vladimir Kazeev B.Sc. M.Sc.

Acknowledgements

I have had the opportunity to learn from a range of inspiring individuals while working on this project. I wish to acknowledge the support of the following parties in producing my thesis.

The German Weather Service (DWD) for producing model data for the experiments described in the thesis and offering support in understanding their data assimilation system. In particular Annika Schomburg for producing and providing experimental data, and Hendrik Reich for sharing information on several aspects of the data assimilation system.

Theresa Diefenbach at the Ludwig-Maximilian-University (LMU) in Munich for providing initial code for the calculation of partial analysis increments and giving helpful thoughts, advice and guidance at all stages of the project.

The data assimilation group at the Department of Meteorology and Geophysics at the University of Vienna. In particular, Tobias Necker for helpful guidance on the first stages of the project, Philipp Griewank for guidance on later stages, and Martin Weissmann for providing ideas for the whole project, organising correspondence with the DWD and LMU, and proof-reading the final document.

My supervisor Vladimir Kazeev for guiding the whole project and providing support with mathematical proofs and notation.

Abstrakt auf Deutsch

Partial Analysis Increments (PAI) beziehen sich auf Spalten der Kalman-Gain, die mit ihren jeweiligen Beobachtungsabweichungen in einem Kalman-Filter multipliziert werden. Es wird eine neue, von Diefenbach et al. 2023 eingeführte Methode zur Verwendung von PAI mittels Annäherung von Analysen mit modifizierten Lokalisierungsfunktionen, *retrospektive Lokalisierung*, beschrieben. Es werden neue theoretische Ergebnisse zur Genauigkeit der retrospektiven Lokalisierung angegeben, die in führender Ordnung von Ensemblekorrelationen zwischen Analyseäquivalenten für die Beobachtungen abhängt. Die retrospektive Lokalisierung wird dann verwendet, um die vertikalen Lokalisierungsparameter für Satellitenradianzen in einem einmonatigen Vorhersageexperiment zu optimieren, bei dem ein nahezu betriebsbereiter Aufbau des regionalen numerischen Wettervorhersagesystems (NWP) des Deutschen Wetterdienstes (DWD) verwendet wird. Das Vorhersagesystem des DWD basiert auf dem ICOSahedral Non-hydrostatic Limited Area Model (ICON-LAM, Zängl et al. 2015) und dem Kilometerscale ENsemble Data Assimilation System (KENDA, Schraff et al. 2016). Analysen mit neuen Lokalisierungseinstellungen werden approximiert und mit Radiosondenwerten für Temperatur und relative Feuchte verglichen. Radiosonden werden nicht assimiliert, um ihre Unabhängigkeit als verifizierende Beobachtungen zu wahren. Für die drei Kanäle, die derzeit beim DWD verwendet werden, zeigen wir, dass die beste Strategie zur Minimierung von Fehlern bei Radiosonden zur Analysezeit darin besteht, sichtbare Strahlung im $0,6\ \mu\text{m}$ -Wellenlängenkanal nicht zu lokalisieren und infrarote Strahlung in den $6,2$ und $7,3\ \mu\text{m}$ -Kanälen hoch oben in der Atmosphäre zu lokalisieren. Bei allen Kanälen zeigte sich, dass eine vertikale Lokalisierungsfunktion mit mehreren Spitzenwerten von Vorteil sein könnte, insbesondere wenn der zweite Spitzenwert in der Nähe der Grenzschicht liegt. Es wird ein einfacher Kandidat für eine solche Funktion vorgeschlagen und gezeigt, dass sie niedrige Fehlerwerte ergibt. Insgesamt wurde 1 Monat als lang genug erachtet, um einige qualitative Trends in den Daten aufzuzeigen.

Abstract

Partial analysis increments (PAI) refer to columns of the Kalman gain multiplied by their respective observation deviations in a Kalman filter. A new method introduced by Diefenbach et al. 2023 for using PAI to approximate analyses with modified localisation functions, *retrospective localisation*, is described. Novel theoretical results are given on the accuracy of retrospective localisation, which depends to leading order on ensemble correlations between analysis equivalents for the observations. Retrospective localisation is then used to optimise vertical localisation parameters for satellite radiances in a one month forecast experiment using a near-operational setup of the German Weather Service (DWD) regional numerical weather prediction (regional NWP) system. The DWD forecast system is based on the the ICOSahedral Non-hydrostatic Limited Area Model (ICON-LAM, Zängl et al. 2015) and the Kilometer scale ENsemble Data Assimilation system (KENDA, Schraff et al. 2016) system. Analyses with new localisation settings are approximated and compared to radiosonde values for temperature and relative humidity. Radiosondes are not assimilated to preserve their independence as verifying observations. For the three channels currently in use at DWD, we show that to minimise error to radiosondes at analysis time, the best strategy is to not localise visible radiances in the $0.6\ \mu m$ wavelength channel, and to localise infrared radiances in the 6.2 and $7.3\ \mu m$ channels high up in the atmosphere. For all channels, there was evidence that a vertical localisation function with multiple peaks could be beneficial, particularly when the second peak is around the boundary layer. A simple candidate for such a function is proposed and shown to produce low error scores. Overall 1 month was considered long enough to show some qualitative trends in the data.

Contents

1	Introduction	6
2	Methods	9
2.1	Notation, introduction to data assimilation concepts and the LETKF	9
2.1.1	Sequential Data Assimilation Setup	9
2.1.2	The Local Ensemble Transform Kalman Filter (LETKF)	16
2.2	Satellite radiances and their assimilation	23
2.2.1	Radiative transfer	23
2.2.2	Visible $0.6\mu m$ and Infrared $6.2, 7.3\mu m$ Channels	23
2.2.3	Assimilating radiances	24
2.3	Partial analysis increments and novel results on their use in observation influence analysis . .	25
2.3.1	PAI definition and approximation with data assimilation outputs	25
2.3.2	Using PAI to Localise Retrospectively	27
2.4	Metrics and notation	36
2.5	Optimisation methods	39
2.5.1	Optimisation of the Gaspari Cohn function	39
2.5.2	Empirical Localisation Function (ELF) Derived from PAI	39
3	Data and model	44
3.1	Data	44
3.1.1	Model and assimilation setup	44
3.1.2	Data analysed	47
4	Illustrative example	49
4.1	Using PAI to optimise localisation - example at one location	49
4.1.1	Radiance assimilation at the test time and location	49
4.1.2	Using PAI to understand assimilation	51
4.1.3	Optimising the localisation parameters	53
5	Results	55
5.1	Behaviour of PAI vectors in whole sample	55
5.1.1	Averages of increments in the whole sample	55
5.1.2	Averages of benefit in the whole sample	56
5.2	Optimisation of Gaspari-Cohn localisation function parameters	57
5.2.1	Global optimisation of Gaspari-Cohn parameters	59
5.2.2	Reliability of results - subsampling	60
5.2.3	Optimising to different variables	61
5.2.4	Optimal parameters for clear/cloudy sky defined by observed radiances	61
5.3	Empirical localisation functions	63

5.3.1	Global coefficients	63
5.3.2	Clear/cloudy sky coefficients	64
5.4	Comparison of localisation strategies	65
5.4.1	Normalised sum of Gaspari-Cohn functions	66
5.4.2	Cost function for different localisation strategies	68
6	Discussion	71
6.1	Optimal localisation strategies	71
6.1.1	Comparison to other methods for choosing vertical radiance localisation	71
6.1.2	Comparison to DWD strategy	73
6.2	Limitations and potential for further study	73
7	Conclusion	75
A	Proofs of fundamental data assimilation results	78
B	Required theory from linear algebra	80
B.1	Elementwise/Schur/Hadamard product	80
B.2	Projection maps	80
B.3	The spectral norm	82
B.4	The Neumann series expansion	83
B.5	Eigenvalues and eigenvectors	83

Chapter 1

Introduction

Data assimilation is a technique used in atmospheric science to combine observational data and model output to estimate the current state of a dynamical system (Asch 2016). In operational weather forecasting centres, complex data assimilation algorithms make estimates of the current state of the atmosphere to be used as initial conditions for numerical forecasts. This is important due to the chaotic nature of atmospheric dynamics: small errors in initial conditions can lead to large forecast errors.

Data assimilation algorithms are often based on an optimal linear estimate, and this thesis explores a method for using this linearity to understand the impact of observations and parameters within the data assimilation system. For $x_b \in \mathbb{R}^n$ the model output (the ‘first guess’ for the state of the atmosphere), $d \in \mathbb{R}^m$ the error made when estimating observation values with the model, and $K \in \mathbb{R}^{n \times m}$ an optimal matrix, the estimate used for the current state of the model often takes the form:

$$x_b + Kd.$$

The vectors $K_{:,i}d_i \in \mathbb{R}^n$ are referred to as partial analysis increments (PAI), and can be used as a diagnostic for how observation i influences the estimate. It has been proposed that PAI can be used to quickly approximate the output of a data assimilation system under different settings so that many different parameter values can be evaluated and optimised (Diefenbach et al. 2023). In this thesis we explore the use of PAI vectors in understanding the effect of localisation, a technique used in data assimilation to account for issues relating to dimensionality and misrepresentation of errors in models and data assimilation algorithms. In particular, we aim to:

- Give a concise overview of the mathematical concepts behind the PAI method and the use of PAI to approximate altered localisation settings;
- Understand theoretically the error when using PAI to approximate localisation settings;
- Use PAI vectors from a near-operational data assimilation system to understand where radiance observations in different channels are beneficial to estimates for initial conditions for forecasts;
- Use PAI vectors to approximate different localisation settings for optimisation of localisation parameters and comparison of localisation strategies.

The thesis will focus on data assimilation with the Local Ensemble Transform Kalman Filter (LETKF; Hunt, Kostelich, and Szunyogh 2007) as used by the German Weather Service (Deutsche Wetterdienst, DWD). In this system, multiple forecasts are evaluated, each with different random perturbations of the initial conditions to represent the probability distribution of the final forecast (comparable to a Monte Carlo sample of the forecast distribution). Each forecast is referred to as an ensemble member, and the set of forecasts as an ensemble.

The LETKF is derived by taking the optimal value of the model state (the ‘analysis’) to be a linear combination of ensemble members. Ensembles are generally very small compared to the dimension of the model space, so ad hoc techniques are used to account for the errors induced by the small sample size. The most important of these techniques is ‘localisation’. In the DWD system this consists of applying the LETKF separately at each grid point in the model to increase the number of degrees of freedom - instead of one optimised weight per ensemble member, there is one optimal weight per gridpoint and ensemble member.

Localisation in the LETKF differs from other assimilation systems in that it is implemented through ‘local analysis’ - running the assimilation algorithm separately at different locations and only including the observations considered most influential at that point¹. In practise, local analysis is performed by replacing the inverse of the observation covariance matrix, R , with the Schur/Hadamard/Elementwise product of R^{-1} and some matrix ρ describing the localisation function evaluated at the specific grid point being considered. This can be interpreted as ‘inflating observation variance values where we want less influence of the specific observations on the assimilation’. Localisation in this sense also increases computational efficiency, as well as increasing the dimensionality of the assimilation. The main aims of localisation are to correct for sampling error and dimensionality issues relating to small ensembles, but localisation can also correct for other misrepresented errors in the assimilation system (Anderson 2016). In general, localisation is critical for effective ensemble Kalman filtering in atmospheric science (Houtekamer and Zhang 2016).

This thesis will focus on vertical localisation for cloud-affected satellite radiance observations, which have the potential to make significant contributions to forecast skill (McNally 2002). Incorporating radiance observations into operational weather forecasts poses many challenges, however. Satellite radiances observe the vertical integral of a complex structure function, so have no well-defined vertical location, and depend very nonlinearly on model variables through complex observation operators (Geiss et al. 2021). Assimilation affects optimal model states by adjusting variables (for example clouds) so that model equivalents of radiances match observations, but sometimes does this by adjusting for example the wrong clouds or clouds at the wrong level (Scheck, Weissmann, and Bach 2020).

Typical methods for localisation of satellite observations are based on taking a nominal height for the observation and letting localisation values decay away from this point. The localisation function is generally taken to be a 5th order polynomial similar to a Gaussian bell curve but with values going to zero outside of a certain range, referred to as a Gaspari-Cohn function (Gaspari and Cohn 1999). The Gaspari-Cohn function is completely described by the nominal location of the observation and width of the curve.

The nominal height is often taken to be the point where the observation operator is most sensitive to model variables so that only locations that are strongly correlated with the radiance observation assimilate radiances (Houtekamer et al. 2005, Hutt et al. 2020). This method has the further advantage that an ‘adaptive’ solution can be found: the nominal height can be changed over time and space as the sensitivity function changes. It is even possible to use the sensitivity function of radiances itself as a localisation function, eliminating the need for a Gaspari-Cohn function (Miyoshi and Sato 2007).

Localising based just on the direct relation between radiances and model variables could remove indirect positive impacts based on covariances between model variables, however. Other studies have tried to find localisation functions by considering areas where covariances with the satellite observation were high based on empirically-derived values (Fertig et al. 2007). Monte-Carlo methods have also been used to correct directly the sampling error, either to understand optimal localisation functions offline (Lei, Anderson, and Whitaker 2016, Lei et al. 2020a) or to create adaptive localisation functions based on assimilation outputs (Lei et al. 2020b). Simulated forecast and observation environments have also been used to calculate localisation functions that most reduce forecast error (Anderson and Lei 2013).

We take here a more operational approach, using outputs from a near-operational forecast and localising

¹ Localisation is more well known in algorithms that are run globally, where it consists of a modification of covariance matrix estimates.

where PAI vectors move towards verifying observations. This has the advantage that the many complexities of the operational system are shown in the data (if the observation is detrimental at a location due to other errors in the data assimilation system, our method should reflect this) but has the disadvantage that the results are specific to the data assimilation system being tested.

We start by addressing the first aim of introducing the PAI method in a detailed and formal way that is accessible to readers with a background in Mathematics and limited experience of data assimilation. Section 2.1 gives an introduction to the required concepts in data assimilation aimed at an audience from a mathematical background, and formalises some of the assumptions in data assimilation derivations so that theoretical results can be more easily described later. We then describe briefly satellite radiances and their assimilation in section 2.2, so that the reader has an understanding of the problems that we aim to address with the PAI method. Section 2.3 then introduces PAI vectors formally and states all of the approximations involved in their calculation and use in approximating analyses with altered vertical localisation functions.

To address the second aim of understanding theoretically where PAI vectors can be used to approximate new localisation functions, the second part of section 2.3 presents a leading order expression for the error in the localisation approximation. The result is based on several fairly unrestrictive assumptions, and has implications for computational methods.

Before tackling the third aim of evaluating the beneficial and detrimental impacts of radiance observations on initial conditions for weather forecasts, we describe the exact methods we will use for optimising the localisation functions, and the data set with which we will work. We propose some metrics, often based on the work of Diefenbach et al. 2023, to assess localisation strategies in section 2.4. We describe how we will use these metrics to optimise localisation strategies in section 2.5, including a novel derivation of a least squares method for optimising an idealised class of localisation functions. The data set, including the model used, is described in section 3.1, including some information on the verifying observations and assimilation. To illustrate the optimisation method that we will use, section 4.1 gives an example of the optimisation process for one location.

After this preparatory work, we will be able to describe with the PAI method where in the atmosphere radiance observations are useful. This is done first with some more exploratory plots in section 5.1, which also give an indication of some tendencies in the PAI vectors that could be responsible for beneficial and detrimental increments. Section 5.2 on optimisation of the Gaspari-Cohn parameters also gives intuition for where in the atmosphere observations are useful, considering different channels, some simple adaptive methods, and verifying to different variables.

The rest of the work addresses the fourth goal of optimising and critically assessing localisation strategies. Section 5.2 fulfils a second purpose of describing which Gaspari-Cohn functions are optimal on our data set, and for which conditions/verifying observations, while section 5.3 shows the results of the directly optimised idealised localisation function. Both of these methods suggest that a multi-peaked localisation function may be advantageous, so a simple method of producing such a function is given in section 5.4.1, before all localisation strategies are assessed using a cost function based on PAI in section 5.4.2.

Note on scope: The thesis aims to be accessible to an audience with a background in Mathematics and limited experience of data assimilation, while respecting convention in data assimilation notation. As such, extra detail is sometimes given for derivations that some readers may find unnecessary.

Chapter 2

Methods

2.1 Notation, introduction to data assimilation concepts and the LETKF

2.1.1 Sequential Data Assimilation Setup

The contents of this section can be found in most standard data assimilation text books. We refer in particular to Asch 2016 and Evensen, Vossepoel, and Leeuwen 2022.

Data assimilation can be defined as combining observations and model output in an optimal way to estimate the current state of a dynamical system (Asch 2016). We focus on *sequential data assimilation* in the Local Ensemble Transform Kalman Filter (Hunt, Kostelich, and Szunyogh 2007), as this is the form used in the experiments discussed later. Sequential data assimilation can be understood as combining observations and model output at the moment they are obtained, using this as initial conditions for the model, integrating the model forwards in time, and repeating (as opposed to combining observations from several points in time at once). The Local Ensemble Transform Kalman Filter is a computationally efficient data assimilation system based on Monte Carlo simulations.

Definitions

We first define concepts at a single point in time, then refer to the dynamical system in question later.

Definition 2.1.1 (Analysis, Model Space, Truth, Background distribution, Background Error (Asch 2016, p. 12-13)).

We aim to find the best estimate x_a of the mean x of a random variable X in some model space $\mathbb{R}^n$.

- x_a is referred to as the analysis;
- Refer to x as the truth;
- Refer to the distribution of X as the background distribution;
- Write $X = x + \epsilon_b$ with $\mathbb{E}(\epsilon_b) = 0$. Then ϵ_b is referred to as the background error.

In weather forecasting, X is the distribution of forecast outputs, and model space is a vector space of dimension approximately $n \approx 10^8$ consisting of values of (for example) temperature, wind speed and humidity at each point on a 3d grid. Inputs and outputs of forecast models are in this form.

In this thesis, the truth is a purely theoretical quantity and rarely considered. It is sometimes defined differently as a random variable (for example see Fortin et al. 2014), but for the derivations here we take it to refer to the deterministic mean.

Definition 2.1.2 (Background/First Guess (Asch 2016, p. 12-13)). *We observe background or first guess*

values x_b of X .

In weather forecasting the background x_b is the output of a forecast model.

Definition 2.1.3 (Observation, Observation operator, Model Equivalents (Asch 2016, p. 12-13)). *We take samples (observations) $y_o \in \mathbb{R}^m$ of Y , an observation random variable which is related to X . In particular, we model Y to have mean $h(x)$ for some function $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$.*

- *The function h is referred to as the observation operator;*
- *We will write linearized version of h as $\mathcal{H} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and represent this with a matrix $H \in \mathbb{R}^{m \times n}$ (corresponding to the gradient of $\mathcal{H}$)¹;*
- *Refer to $h(x_b) \in \mathbb{R}^m$ as the model equivalents of the observations;*
- *The vectors $y_o - h(x_b) \in \mathbb{R}^m$ are referred to as observation (first guess) departures;*
- *Write $Y = h(x) + \epsilon_o$. Then ϵ_o is referred to as the observation error.*

In weather forecasting, observations could be direct observations of model variables, for example temperature observations from surface stations or radiosondes, or indirect observations such as the satellite observations we are concerned with in this thesis. For direct observations, the observation operator h is generally just some kind of interpolation. For satellite radiances, h is a vertical integration of a complex sensitivity function depending on the model variables.

In data assimilation literature, the linearised observation operator can refer either to the full linearisation at a point x_0 : $\mathcal{H}(x) = h(x_0) + h'(x_0)x$ or just the gradient at the linearisation point: $\mathcal{H}(x) = h'(x_0)x$. The inclusion of the constant term does not affect any of the derivations, so is not given much attention in the rest of the thesis. We distinguish between the linear operator $\mathcal{H}$ and the matrix representation of its gradient H to avoid confusion relating to this distinction.

In the next definition, we make some strong assumptions on the distributions of X and Y . These assumptions will remain throughout the thesis as they are also used in operational weather forecasts. They are often tenuous.

Definition 2.1.4 (Background Covariance, Observation Covariance (Asch 2016, p. 16-17)). *Make the following assumptions. For $x \in \mathbb{R}^n$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ as before, and $B \in \mathbb{R}^{n \times n}$, $R \in \mathbb{R}^{m \times m}$ covariance matrices:*

$$X \sim \mathcal{N}(x, B), \tag{2.1}$$

$$Y \sim \mathcal{N}(h(x), R). \tag{2.2}$$

We refer to B as the background (error) covariance (matrix) and R as the observation (error) covariance (matrix).

The background error B is sometimes, but not always, modelled to be constant in time (Kalnay 2003, p. 176-177).

The background and observation covariance matrices are of particular importance in this thesis. The variances in B represent (Lahoz, Khatatov, and Menard 2010, p. 5):

- Errors in the preceeding initial conditions;
- Errors in the forward model $\mathcal{M}$.

The variances in R represent a combination of (Kalnay 2003, p. 199-200):

¹The notation $\mathcal{H}$ is not conventional in meteorology literature, and is intended to avoid confusion for readers from a mathematical background.

- *Representativeness error*: Observations are often considered as mean values over the given area they represent, but are actually subject to natural small scale variability;
- *Instrument error*: Natural errors made when taking observations;
- *Operator error*: The relation $\mathbb{E}(Y) = h(x)$ is often an approximation and thus a source of errors.

It is assumed that the background errors are independent of the model errors, i.e. that X is independent of Y .

Important assumption: In this thesis we will assume that R is diagonal. We make this assumption because:

- A diagonal R is used in many operational data assimilation systems, including those of the German Weather Service that will be considered later;
- For conventional observations (for example direct observations of temperature), this is a sensible assumption - the operator error should be very small for conventional observations and instruments are independent.

For less traditional observations such as the satellite radiances we will consider later, correlations between observations (the off-diagonal elements of R) could present more issues (Aravéquia et al. 2011) and are an active area of research.

The Kalman Filter

Under the assumptions defined in section 2.1.1, we can derive an optimal estimate for the analysis (our estimate of the mean of X). As in Asch 2016, p. 11, *optimal* can refer to either:

- Optimality in a *statistical* sense - that expected variance of the analysis is minimised, or;
- Optimality in a *variational* sense - that the value of some cost function can be minimised.

These can both be understood in a unified theory by taking a Bayesian setup (Evensen, Vossepoel, and Leeuwen 2022), so we consider here the optimum to be a maximum a posteriori estimate. We take a Gaussian prior distribution for x .

Although the exact expression for the analysis that we derive now is not used in the later chapters of this thesis, we give the full derivation here as it provides good intuition for later results and is a fundamental object in data assimilation. The result is derived with the linearized version of the observation operator, that is, the expectation of the observations is Hx , where x is the expectation of the background. Note how we have ignored here (and in subsequent derivations) the constant term in the linearized observation operator.

Theorem 2.1.1 (Evensen, Vossepoel, and Leeuwen 2022, p. 63-65). *Let $X \sim \mathcal{N}(x, B)$ and $Y \sim \mathcal{N}(Hx, R)$ for $H \in \mathbb{R}^{m \times n}$. Let x_b and y_o be observations of X and Y respectively. Then assuming a Gaussian prior distribution for x , the maximum a posteriori estimator for x is given by:*

$$x_a = x_b + K(y_o - Hx_b),$$

where

$$K = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}. \quad (2.3)$$

This expression can also be written as:

$$K = B H^T (H B H^T + R)^{-1}. \quad (2.4)$$

Proof. Let $f(x|x_b, y_o)$ refer to the posterior distribution of x . Then as the prior distribution is normal, we have:

$$f(x|x_b, y_o) \propto \exp \left[- (x_b - x)^T B^{-1} (x_b - x) - (y_o - Hx)^T R^{-1} (y_o - Hx) \right].$$

So maximising f over x is equivalent to minimising J , where:

$$J(x) = (x_b - x)^T B^{-1} (x_b - x) + (y_o - Hx)^T R^{-1} (y_o - Hx). \quad (2.5)$$

This is done by differentiating J . The full proof can be found in appendix A (theorem A.0.1). $\square$

The second expression (2.4) is more well known.

Definition 2.1.5 (Kalman Gain (Asch 2016, p. 17-18)). *When a Data Assimilation algorithm gives the optimal estimate:*

$$x_a = x_b + K(y_o - h(x_b)),$$

for some matrix $K \in \mathbb{R}^{n \times m}$, then K is referred to as the Kalman gain.

So in the current setup, the Kalman gain is $K = BH^T (HBH^T + R)^{-1}$.

It is important to understand the distribution of the error made when using x_a to estimate x :

Definition 2.1.6 (Analysis Error (Kalnay 2003, p. 155)). *For a data assimilation system where $x_a = x_b + K(y_o - h(x_b))$, refer to ϵ_a as the analysis error, where:*

$$\epsilon_a = X + K(Y - h(X)) - x.$$

It can be checked that, for linear h (that can be represented by the matrix H), $\mathbb{E}(\epsilon_a) = 0$. Looking at the variance in the linear case we have:

$$\mathbb{E}(\epsilon_a \epsilon_a^T) = \mathbb{E} \left[((I - KH)\epsilon_b + K\epsilon_o)((I - KH)\epsilon_b + K\epsilon_o)^T \right] \quad (2.6)$$

$$= (I - KH)\mathbb{E}(\epsilon_b \epsilon_b^T)(I - KH)^T + K\mathbb{E}(\epsilon_o \epsilon_o^T)K^T \quad (2.7)$$

$$= (I - KH)B(I - KH)^T + KRK^T. \quad (2.8)$$

The following result on the analysis covariance is important.

Proposition 2.1.1 (Kalnay 2003, p. 156-157). *Suppose we use a data assimilation algorithm with Kalman gain as in (2.4) and linear observation operator H . When all the relevant inverses are well defined, the covariance matrix for ϵ_a can be written both as:*

$$A := \mathbb{E}[\epsilon_a \epsilon_a^T] \quad (2.9)$$

$$= (B^{-1} + H^T R^{-1} H)^{-1}, \quad (2.10)$$

and:

$$A = (I_n - KH)B. \quad (2.11)$$

Proof. The proof consists of working through the different expressions for K and A , and is included in the appendix item A.0.1. $\square$

These formulations of A give us an intuition for the important result that **assimilation reduces spread**. Put more formally:

Proposition 2.1.2 (Derived by author but could be in other textbooks). *Suppose A and B are the analysis and background covariance matrices in a Kalman filter with Kalman gain as in (2.4). Then if $v \in \mathbb{R}^n$:*

$$v^T A v \leq v^T B v. \quad (2.12)$$

Note: This can be understood as: *assuming (2.1) and (2.2) hold, then all model variables, and all linear combinations of model variables, have variance in the analysis less than or equal to their variance in the background.*

Proof. Taking the second expression for A from proposition 2.1.1 and the second definition of the Kalman gain (2.4), we have:

$$\begin{aligned} A &= (I_n - KH)B \\ &= B - BH^T (HBH^T + R)^{-1} HB. \end{aligned}$$

The matrix $BH^T (HBH^T + R)^{-1} HB$ is positive semi-definite (as $(HBH^T + R)^{-1}$ is the inverse of a positive semi-definite matrix and we have multiplied by the same matrix on the left and right of this). Therefore:

$$\begin{aligned} v^T A v &= v^T B v - v^T BH^T (HBH^T + R)^{-1} HB v \\ &\leq v^T B v. \end{aligned}$$

□

Sequential Data Assimilation and Ensembles

The derivations in the previous section were carried out for a single point in time. In operational weather forecasts, we want information on a model state that varies in time. Introduce the following definition to capture this.

Definition 2.1.7 (Model Operator (Asch 2016, p. 16)). *In general, x (the mean of the background) evolves over time according to a dynamical system. To capture this, write $x^{(t_k)}$ to refer to the expected state x at time t_k , and write:*

$$x^{(t_k)} = \mathcal{M}_k(x^{(t_{k-1})}).$$

$(\mathcal{M}_k)_{k \geq 0}$ is referred to as the model operator.

Note: The above definition for the model operator does not describe probabilistically the behaviour of the background or any estimates we make for the mean. For now we only consider the model operator as a way of describing the propagation of the expected state. The background itself, X , is a description of the error of our model output as an estimate for the mean state, and the analysis, x_a , is a combination of model output and observations to be used as an estimate for the mean state.

In weather forecasting, $\mathcal{M}_k$ is a numerical integration of a set of model equations (conservation of mass, momentum and energy) and the evaluation of some physical equations describing atmospheric processes (for example radiation and cloud physics). It is nonlinear.

Use superscripts to specify at which point in time we have our observations and background values (for example $\{x_b^{(t_k)}\}_{k \in \mathbb{N}}$, $\{y_o^{(t_k)}\}_{k \in \mathbb{N}}$, $\{X^{(t_k)}\}_{k \in \mathbb{N}}$). The analysis is used as an initial condition for $\mathcal{M}_k$ and integrated forwards:

$$x_b^{(t_k)} = \mathcal{M}_k(x_a^{(t_{k-1})}).$$

Note that this already suggests that the assumption $X \sim \mathcal{N}(x, B)$ is tenuous.

To infer probabilistic information on how the distribution of X evolves in time, a set of perturbations can be stored and propagated with the model (Evensen 1994, Houtekamer and Mitchell 1998), as a Monte Carlo sample of the distribution of X . We use notation as in Hunt, Kostelich, and Szunyogh 2007 to address this as follows:

Definition 2.1.8 (Ensemble, Ensemble Mean (Kalnay 2003, p. 181-182)). *Let $x_{b,i}, i = 1, 2, \dots, N$ be a set of Monte Carlo samples from the distribution of X .*

- Refer to $\{x_{b,i}\}_{i=1}^N$ as an ensemble;
- Let $x_b = \bar{x}_b$ be the mean of the ensemble members $x_{b,i}$;
- Refer to the matrix X_b with columns $(X_b)_{:i} = x_{b,i} - x_b$ as the matrix of ensemble perturbations.

In weather forecasting, the forward operator is often computationally very expensive, so N is often between 20 and 100. This is very small compared to the size of model space (recall that model space has dimension approximately 10^8 in the experimental cases we will be concerned with).

An ensemble can be used to estimate covariances (Evensen 1994, Houtekamer and Mitchell 1998):

Definition 2.1.9 (Ensemble Covariance Estimate, P_b (Asch 2016, p. 98)). *Let X_b be a matrix of ensemble perturbations $x_{b,i} - x_b, i = 1, \dots, N$. Then the background ensemble covariance estimate is defined as*

$$P_b = \frac{1}{N-1} X_b X_b^T.$$

It is often considered advantageous to use P_b as the estimate for the background covariances as it is able to vary over time. In contrast, the empirically estimated background covariance B must be estimated before running a forecast and assimilating, so cannot depend on the state of the atmosphere (Kalnay 2003, p. 176-178).

We aim to produce a similar set of samples for the distribution of the analysis. We use the idea of ensemble covariance estimates to derive the analysis ensemble.

Definition 2.1.10 (Analysis Perturbations, Ensemble Analysis Error, P_a (Asch 2016, p. 158)). *Let $x_{a,i}$ be an ensemble drawn from distribution of the analysis $X + K(Y - HX)$, and let x_a be an estimator for the analysis mean.*

- Refer to the matrix X_a defined column-wise by $(X_a)_{:i} = x_{a,i} - x_a$ as matrix of analysis perturbations;
- Define the ensemble estimate of the analysis covariances as $P_a = \frac{1}{N-1} X_a X_a^T$ (as with the estimate of background covariances).

Later in the thesis x_a will be taken as the analysis mean, but this is not always the case.

We aim to have a method that satisfies (as columns of X_a should represent the distribution of ϵ_a):

$$\begin{aligned} \frac{1}{N-1} X_a X_a^T &\approx \mathbb{E}(\epsilon_a \epsilon_a^T) \\ &= (I - KH)B(I - KH)^T + KRK^T. \end{aligned}$$

As in (2.8). There are several ways to achieve this. One is the idea of square-root filters such as the LETKF discussed in the next section. Another, perhaps more easily interpretable idea is that of stochastic perturbations (Asch 2016, p. 157-160). Let $r_i \in \mathbb{R}^m, i = 1, \dots, N$ be a set of samples drawn from the distribution of

$Y = \mathbb{E}(Y)$, $\mathcal{N}(0, R)$, and let S be the matrix with columns equal to r_i . We have by independence that:

$$X_b S^T$$

is approximately 0. So then if $x_{a,i} = x_{b,i} + K(y_o + r_i - Hx_{b,i})$ and $x_a = x_b + K(y_o - Hx_b)$,

$$\frac{1}{N-1} X_a X_a^T = \frac{1}{N-1} ((I - KH)X_b + KS)((I - KH)X_b + KS)^T.$$

This is approximately equal to (by ignoring $X_b S^T$ terms):

$$(I - KH)P_b(I - KH)^T + K \frac{1}{N-1} S S^T K^T.$$

Assuming that B and P_b are equal (which is in general at least approximately true), we have that this is equal to:

$$(I - KH)B(I - KH)^T + K R K^T,$$

as desired. This gives us an example of the following definition.

Definition 2.1.11 (Ensemble Data Assimilation, Single Time Version (Terminology specific to this thesis to quickly introduce the ensemble method)). *Refer to a method of producing an optimal analysis ensemble $\{x_{a,i}\}_{i=1}^N$ from observations y_o and a background ensemble $\{x_{b,i}\}_{i=1}^N$ as an ensemble data assimilation algorithm.*

We are now ready to formally define sequential data assimilation.

Definition 2.1.12 (Sequential Ensemble Data Assimilation (Asch 2016, p. 156, 160, Ott et al. 2004)). *Suppose we have observations $y_o^{(t_k)}$ at given points in time $t_k, k = 0, 1, 2, \dots$, and an ensemble data assimilation algorithm for a single point in time. Then sequential data assimilation is applied inductively as follows. At time t_k :*

1. *Apply the ensemble data assimilation algorithm to $y_o^{(t_k)}$ and $\{x_{b,i}^{(t_k)}\}_{i=1}^N$ to get $\{x_{a,i}^{(t_k)}\}_{i=1}^N$.*
2. *Propagate each member forward in time individually: $x_{b,i}^{(t_{k+1})} = \mathcal{M}_{t_{k+1}}(x_{a,i}^{(t_k)})$.*

Sequential data assimilation is illustrated in one dimension (or for one coordinate of the multi-dimensional definition) in figure 2.1, taken from Evensen, Vossepoel, and Leeuwen 2022. Each line refers to the trajectory of the variable for one ensemble member. As we saw in proposition 2.1.2, the spread of the ensemble is reduced at each step: more ‘certainty’ from the observations is included in the analysis.

The definition of sequential data assimilation used here is purposely very broad. Various choices and modifications exist at both steps. For example, the mean x_a is often propagated forward on its own separately from the ensemble.

Localisation in Conventional Data Assimilation

As mentioned before, it is considered beneficial to substitute P_b for B in the formula for K to obtain the analysis, but this requires some adaptation of the matrix to deal with dimensionality issues. The main issues with using P_b as the background covariance are: (1) the ensemble size is generally far too small to estimate the covariance matrix accurately and (2) the ensemble size is generally far too small to produce an invertible estimate for the covariance, which is often necessary for assimilation algorithms.

Localisation is a method of modifying the ensemble-derived covariances to address these issues. In general, this can be implemented in two different ways (Asch 2016, p. 167-170):

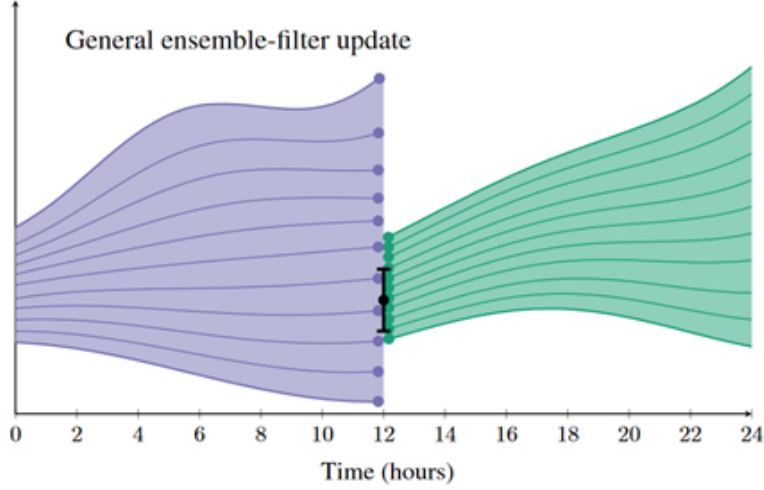


Figure 2.1: Taken from Evensen, Vossepoel, and Leeuwen 2022. Illustration of sequential ensemble data assimilation. The model is run for all ensemble members, assimilation is applied to the whole ensemble, and the process is repeated. Notice how the spread of the ensemble is reduced at the update step, reflecting proposition 2.1.2.

1. *Covariance localisation*: Taking pairs of variables that are expected to have approximately 0 covariance - for example variables that are physically far apart, and artificially reducing or setting to 0 the corresponding element of P_b ;
2. *Local analysis*: Computing separate analyses at different locations, and only including observations close to each area (as in some of the first ensemble data assimilation tests such as Houtekamer and Mitchell 1998).

2.1.2 The Local Ensemble Transform Kalman Filter (LETKF)

In the data assimilation system of the German Weather Service (DWD) on which the experiments in this paper are based, an efficient data assimilation algorithm called the *local ensemble transform Kalman filter* (LETKF) is used. It is motivated by the Kalman filter described in the previous section, and inherently incorporates the forecast ensemble into its formulation. The derivation here follows that of the original description in Hunt, Kostelich, and Szunyogh 2007, although the discussion of projections is a simplified version of arguments in Ott et al. 2004. The full description of the algorithm is given at the end of this section.

We recall for this section a cost function often used to derive Kalman filters. When deriving the Kalman gain in the proof of theorem 2.1.1 (given fully in appendix A, theorem A.0.1) we minimised an a posteriori distribution. For linear h represented by matrix H , this is equivalent to minimising the cost function:

$$J(x) = (x_b - x)^T B^{-1} (x_b - x) + (y_o - Hx)^T R^{-1} (y_o - Hx), \quad (2.13)$$

over x . The same argument works for nonlinear observation operator h , so we consider the full nonlinear cost function:

$$J(x) = (x_b - x)^T B^{-1} (x_b - x) + (y_o - h(x))^T R^{-1} (y_o - h(x)). \quad (2.14)$$

To see this, it may be helpful to note that J looks like the argument of the exponential in a multivariate normal distribution function. We will often work in this section with the full nonlinear operator.

Motivation of LETKF and Kalman Gain Equation

It has been observed in meteorological models that (at least locally) the subspace of the model space vectors that are unstable under the dynamical system is much smaller than the dimension of model space itself (Ott et al. 2004). It is therefore considered sensible to correct the background (locally) only on a subspace of the full model space. The span of the background ensemble deviations (the columns of X_b) is considered a good candidate for this subspace. The local ensemble transform Kalman filter (LETKF) is therefore based on looking for analysis vectors of the form:

$$x_a = x_b + X_b w, \quad (2.15)$$

for $w \in \mathbb{R}^N$. This is in line with other ensemble based Kalman filters (Asch 2016, p. 164). Indeed if we substitute the ensemble background covariance estimate P_b into (2.4), we get that:

$$\begin{aligned} x_a &= x_b + P_b H^T (H P_b H^T + R)^{-1} (y_o - h(x_b)) \\ &= x_b + \frac{1}{N-1} X_b \left[X_b^T (H P_b H^T + R)^{-1} (y_o - h(x_b)) \right], \end{aligned}$$

so the correction is also in the span of the columns of X_b in this case as well, and we are reassured that (2.15) is a sensible form for the analysis.

The LETKF is a method for choosing values for w , both for an analysis estimate and for ensemble members.

Note on inverting the ensemble background covariance estimate

Before deriving the LETKF, we adress some singularity issues.

Some care must be taken when substituting P_b into derivations as the small sample size means that the matrix is not in general invertible: it has rank at most $N-1$ for N the number of ensemble members and is therefore always singular. To address this, let $u_1, \dots, u_l$ be an orthonormal basis for the column space of X_b , and let:

$$Q = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_l \\ | & | & & | \end{pmatrix},$$

so that $Q Q^T$ represents the orthogonal projection of a vector in $\mathbb{R}^n$ onto the column space of X_b , a subspace of $\mathbb{R}^n$ (appendix B.2.2). Then we replace the inverse of P_b (which does not exist) with the Moore-Penrose pseudoinverse:

$$Q(Q^T P_b Q)^{-1} Q^T. \quad (2.16)$$

The inverse in (2.16) is strictly well defined (appendix B.2.3). In words, for an input vector v , the matrix (2.16) projects v into a coordinate system for the column space of X_b (the Q^T matrix), applies the inverse of P_b in this coordinate space (the $(Q^T P_b Q)^{-1}$ matrix), and then converts the coordinates back into model space vectors (the Q matrix).

Derivation of the LETKF Kalman gain

To derive the local ensemble transform Kalman filter, we would like to substitute our ensemble estimate of the background covariance into the cost function J given at the start of section 2.1.2. Using the pseudo-inverse (2.16) we can evaluate this as:

$$J(x) = (x_b - x)^T Q (Q^T P_b Q)^{-1} Q^T (x_b - x) + (y_o - h(x))^T R^{-1} (y_o - h(x)). \quad (2.17)$$

Note here we have used the full nonlinear observation operator h .

We now restrict to optimising over vectors of the form given by (2.15). We therefore write $x = x_b + X_b w$, and

substitute $x - x_b = X_b w$ into (2.17). The cost function (2.17) in terms of the w coordinates in (2.15) is then:

$$J(w) = (x_b - x)^T Q (Q^T P_b Q)^{-1} Q^T (x_b - x) + (y_o - h(x))^T R^{-1} (y_o - h(x)) \quad (2.18)$$

$$= (N - 1) w^T X_b^T Q (Q^T X_b X_b^T Q)^{-1} Q^T X_b w + (y_o - h(x_b + X_b w))^T R^{-1} (y_o - h(x_b + X_b w)) \quad (2.19)$$

$$= (N - 1) w_{||}^T w_{||} + (y_o - h(x_b + X_b w))^T R^{-1} (y_o - h(x_b + X_b w)). \quad (2.20)$$

where $w_{||}$ is the component of w orthogonal to the null space of X_b . A proof of the third equality can be found in appendix B in (proposition B.2.5).

If we alter the component of w in the null space of X_b , we do not change the value of the second term $(y_o - h(x_b + X_b w))^T R^{-1} (y_o - h(x_b + X_b w))$ in (2.20). So the minimiser of the cost function:

$$\tilde{J}(w) = (N - 1) w^T w + (y_o - h(x_b + X_b w))^T R^{-1} (y_o - h(x_b + X_b w)), \quad (2.21)$$

will also minimise the cost function (2.20). For this reason, we just consider the cost function (2.21) and omit the tilde in $\tilde{J}$ for ease.

As h is often a very complicated function, we need some more approximations to be able to minimise $J(w)$ in an efficient way. We use the following definition:

Definition 2.1.13 (Ensemble of model (observation) equivalents, background and analysis (Hunt, Kostelich, and Szunyogh 2007)). *Let $x_{b,i}$ be an ensemble in model space with x_b the ensemble mean. Let $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be the observation operator.*

- Define the ensemble of background model equivalents to be the set $\{y_{i,b}\}_{i=1}^N$, where $y_{i,b} = h(x_{i,b})$;
- Write the ensemble mean of $\{y_{i,b}\}_{i=1}^N$ over i as y_b ;
- Define the matrix of background model equivalent perturbations $Y_b \in \mathbb{R}^{m \times N}$ to be the matrix with columns $(Y_b)_{:i} = y_{i,b} - y_b$;
- The analysis ensemble of model equivalents, Y_a , is defined in the same way, replacing the background ensemble $\{x_{b,i}\}_{i=1}^N$ with the analysis ensemble $\{x_{a,i}\}_{i=1}^N$.

We now make a linearity assumption (Hunt, Kostelich, and Szunyogh 2007):

$$h(x_b + X_b w) = y_b + Y_b w. \quad (2.22)$$

This assumption is important, so we describe it in words here:

- Take a linear combination of the (model) ensemble perturbations, defined by coefficients w ;
- Add this to the ensemble mean;
- Apply the observation operator.

We are assuming this is the same as:

- Taking the same linear combination of the model equivalent ensemble perturbations, defined again by coefficients w ;
- Adding this to the model equivalent mean.

We can also interpret (2.22) as saying ' h is affine on the image of X_b ', an equivalent definition if h is sufficiently smooth. Indeed if:

$$\frac{\partial^2}{\partial w^2} h(x_b + X_b w) = 0 \text{ for all } w \in \mathbb{R}^N, \quad (2.23)$$

where the second derivative of a vector valued function is a 3-dimensional tensor, taking the Taylor expansion of $h(x_b + X_b w)$ gives (2.22). This shows that (2.22) is not far from saying that h is affine on the whole space.

Based on this assumption (2.22), we can re-write the cost function (2.21) as:

$$J(w) = (N - 1)w^T w + (y_o - y_b + Y_b w)^T R^{-1} (y_o - y_b + Y_b w). \quad (2.24)$$

This can be minimised over w to derive an optimal weight vector which we notate as w_a . The a in subscript refers to the fact that w_a is given by a specific formula (in this case it is the minimiser of (2.24)), and is not a variable argument of a function. The weight w_a is used for the ensemble mean estimate. The weights for the ensemble members will be derived later.

Theorem 2.1.2 (Hunt, Kostelich, and Szunyogh 2007). *The optimal weight vector w_a for the LETKF is given by:*

$$w_a = \tilde{P}_a Y_b^T R^{-1} (y_o - y_b).$$

Where

$$\tilde{P}_a = ((N - 1)I_N + Y_b^T R^{-1} Y_b)^{-1}. \quad (2.25)$$

Proof. Differentiate (2.24) and set to the derivative to 0. □

Definition 2.1.14 (Analysis error in ensemble space (Hunt, Kostelich, and Szunyogh 2007)). $\tilde{P}_a$ defined in (2.25) is referred to as the analysis error in ensemble space.

From theorem 2.1.2, we have an equation that is used for the analysis mean. This can be written as:

$$x_a = x_b + K(y_o - y_b)$$

where:

$$K := X_b \tilde{P}_a Y_b^T R^{-1}. \quad (2.26)$$

So again we have a Kalman gain matrix for the analysis.

Derivation of LETKF Analysis Ensemble

We again aim to find an analysis ensemble with a sensible sample covariance. The result will be that the LETKF analysis ensemble is taken as:

$$x_a = x_b + X_b \tilde{P}_a Y_b^T R^{-1} (y_o - y_b); \quad (2.27)$$

$$X_a = X_b W_a, \quad (2.28)$$

where $W_a \in \mathbb{R}^{N \times N}$ is defined to be:

$$W_a := [(N - 1)\tilde{P}_a]^{\frac{1}{2}}, \quad (2.29)$$

The square root refers to the unique positive semidefinite symmetric matrix W_a such that $W_a W_a^T = \tilde{P}_a$. The definition of $\tilde{P}_a$ is given in (2.25). Note how no stochastic sampling is required. We derive these equations in the following discussion.

First note that if we were to take an ensemble whose perturbations span the whole of the model space, optimising the raw LETKF cost function (2.21) without the linear approximation (2.22) over w is equivalent to optimising the original cost function (2.14) over x . We therefore look for analysis ensemble perturbations

X_a whose sample covariance $X_a X_a^T$ is comparable to the covariance of the analysis in the conventional data assimilation setup.

To avoid the need to sample more stochastic perturbations as in the previous section, we aim to produce the analysis ensemble linearly transforming the background ensemble perturbation matrix X_b .

The following proposition shows why $\tilde{P}_a$ is referred to as the analysis error in ensemble space and relates $\tilde{P}_a$ to the analysis covariance in conventional data assimilation. We assume a perfect system, so the span of the columns of X_b is now the full model space. The following result is stated in Hunt, Kostelich, and Szunyogh 2007 and we give a similar version assuming a large ensemble size so that $P_b = B$.

Proposition 2.1.3 (Hunt, Kostelich, and Szunyogh 2007). *Suppose $P_b = B$ is an exact estimate for the background covariance with $\text{rank}(P_b) = n$ and let A be the covariance matrix for ϵ_a in the Kalman filter from the previous section, that is the Kalman filter with Kalman gain matrix:*

$$K = P_b H^T (H P_b H^T + R)^{-1}.$$

Let $\tilde{P}_a$ be the analysis error in ensemble space. Then:

$$X_b^T A^{-1} X_b \approx (\tilde{P}_a)^{-1}$$

Proof. We use the expression in proposition 2.1.1 for A and replace B with P_b . Then:

$$X_b^T A^{-1} X_b = X_b^T (P_b^{-1} + H^T R^{-1} H) X_b \quad (2.30)$$

$$= X_b^T P_b^{-1} X_b + X_b^T H^T R^{-1} H X_b \quad (2.31)$$

$$= (N - 1) I_N + Y_b^T R^{-1} Y_b \quad (2.32)$$

$$= (\tilde{P}_a)^{-1}. \quad (2.33)$$

□

Recall that we estimate the analysis error A with the ensemble through $P_a := \frac{1}{N-1} X_a X_a^T$. Thus we aim to have that P_a also satisfies (2.33) in the sense that:

$$P_a = X_b \tilde{P}_a X_b^T. \quad (2.34)$$

We use the expression as in (2.29):

$$W_a := [(N - 1) \tilde{P}_a]^{\frac{1}{2}},$$

where we define the square root W_a as the positive semi-definite symmetric (and therefore unique) matrix such that:

$$W_a W_a^T = W_a W_a = (N - 1) \tilde{P}_a.$$

This then gives us the result:

$$\begin{aligned} P^a &= (N - 1)^{-1} X_a X_a^T \\ &= (N - 1)^{-1} X_b W_a W_a^T X_b^T \\ &= X_b \tilde{P}_a X_b^T, \end{aligned}$$

as desired. The symmetric positive semi-definite square root is used because (Hunt, Kostelich, and Szunyogh 2007):

- The sum of the columns of X_a is in this case equal to 0;
- The choice ensures W_a depends continuously on $\tilde{P}_a$;
- The choice minimizes the mean square distance between the weight matrix and the identity.

The next lemma shows that also in the LETKF assimilation reduces spread:

Lemma 2.1.3 (Derived by author but could be in other textbooks). *The eigenvalues of W_a are all less than or equal to 1.*

Note: As W_a is symmetric, we can associate its biggest eigenvalue with its matrix spectral norm (defined later).

Proof. Recall the definition of W_a :

$$W_a = (N - 1)^{\frac{1}{2}} \left((N - 1)I_N + Y_b^T R^{-1} Y_b \right)^{-\frac{1}{2}}.$$

Then as R^{-1} is diagonal and has all non-negative values, it is positive semi-definite. So $Y_b^T R^{-1} Y_b$ is positive semi-definite. From this we have that $(N - 1)I_N + Y_b^T R^{-1} Y_b$ has eigenvalues all greater than or equal to $(N - 1)$, so the inverse of its square root has positive eigenvalues all less than or equal to $(N - 1)^{-\frac{1}{2}}$. In turn we see that W_a has positive eigenvalues all less than or equal to 1. $\square$

To summarise what we have derived so far, we have:

- A Kalman gain matrix based on ensembles to produce an optimal analysis estimate (theorem 2.1.2);
- A linear transformation to produce an ensemble with the correct analysis covariance (2.29).

Localisation

Localisation is a necessary feature of the LETKF that addresses the following issues:

- Small ensembles poorly estimate small covariances (Asch 2016, p. 167);
- The subspace spanned by the ensemble perturbations is far too small to cover all possible analysis vectors, or even to cover the space of all vectors that are unstable under the dynamical system defined by $\mathcal{M}$ (Asch 2016, p. 167);
- The calculations without localisation are computationally expensive (Hunt, Kostelich, and Szunyogh 2007).

Localisation is implemented in the LETKF as *local analysis*: separate weight vectors w_a and ensemble weight matrices W_a are calculated for each grid point in the model (of which there are many), and only the observations close to that grid point are incorporated into the formulae. This eliminates observations that have small covariances with the point considered (the covariances that are poorly estimated by small ensembles). Local analysis differs from covariance localisation methods in that the background covariance matrix is never directly multiplied by a localisation function - indeed the background covariance matrix is only implicitly present in the LETKF through the ensemble perturbations.

Definition 2.1.15 (Local Analysis (Houtekamer and Mitchell 1998)). *For a given data assimilation algorithm, local analysis means implementing the algorithm separately at different locations. In particular, by taking model space $\mathbb{R}^n$ to be the coordinates of the full model space that refer to the location for analysis, and at each location only including into y_o the observations that are expected to have a significant impact at that location. Resulting local analysis vectors are then concatenated to give the full analysis vector.*

In the LETKF system of the DWD local analysis is applied on what is referred to as the *coarse analysis grid*, which refers to a subset of the full model grid taken by considering only a fraction of the spatial coordinates in each of the x, y and z directions. At each point the vector x_b is of length approximately 8, with one entry each for variables such as temperature, humidity, wind speed (3 components) etc. The observation vector y_o is also much smaller, with approximately 30 conventional observations (direct observations of temperature, wind speed etc.) included for each grid point. For the gridpoints not included on the coarse grid, the analysis is then obtained by interpolating W_a and w_a between gridpoints on the coarse grid.

Concatenating local analysis vectors to make a full analysis could lead to discontinuous physical fields, for example on the boundary between a particularly influential observation being included or not. In operational data assimilation, the influence of observations is therefore tapered to gradually decrease to 0 far away from the analysis location, rather than reducing the influence immediately to 0 after a certain distance threshold. This is implemented by modifying the observation covariance matrix R , in a process we will describe shortly.

Model equivalents generally require more elements of the model state vector than are included in the local analysis, so these are calculated globally, and the same values are used in each local assimilation.

Summary of LETKF algorithm

We are now ready to describe the basis of the LETKF algorithm applied by the DWD (Hunt, Kostelich, and Szunyogh 2007).

Inputs: X_b, x_b background perturbations and mean; y_o observations; h observation operator.

1. Calculate globally the model equivalents $y_{i,b}$, and form the matrix Y_b as well as the mean y_b ;
2. Restrict the analysis perturbations to the local analysis areas (individual gridpoints);
3. For each local analysis gridpoint, calculate the distances from the point to each observation, and apply the localisation function ρ to these distances. The function ρ is generally a tapering function that is equal to 1 for observations exactly at the location considered, and tending to 0 further away. It should be equal to 0 outside of a certain horizontal range to save computational cost.
4. Again for each local analysis point, calculate $x_a = x_b + X_b \tilde{P}_a Y_b^T R^{-1} (y_o - h(x_b))$ and $X_a = X_b W_a$ for each location, but replace R^{-1} in each of the formulae with $(R^{-1}) \circ \rho$, for ρ evaluated in the previous step. Here $\circ$ refers to the elementwise/Hadamard/Schur (appendix B.1) product;
5. Concatenate the analysis vectors to get the full analysis vector.

Taking the elementwise product of R^{-1} with ρ is how the previously described tapering of observation influence is implemented. It can be interpreted as ‘artificially increasing observation variances far away from the local analysis location’, and therefore reducing the influence of the observations compared to the background. Intuitively this makes sense: by increasing the variance of the observation, we are assuming we have less certainty of the observation, so give it less weight in the assimilation algorithm.

The exact choice for ρ is typically a 5th-order piecewise polynomial function known as a Gaspari-Cohn function (Gaspari and Cohn 1999) with coefficients given in section 3.1. The Gaspari-Cohn function is first applied independently to the horizontal distance between the observation and gridpoint, and the difference between the vertical height of the observation and gridpoint (often defined in log-pressure coordinates). The localisation function is then taken to be the elementwise product of the function calculated in the horizontal and vertical distances. For this reason, localisation is generally separated into **horizontal** and **vertical** localisation, and the two are treated separately. In this thesis, we are most interested in vertical localisation.

2.2 Satellite radiances and their assimilation

The experiments in this thesis deal with satellite radiances, an observation of the intensity of radiation at the top of the atmosphere at a given wavelength. Radiances are not direct observations of atmospheric variables such as temperature and humidity, but a vertical integral of a complex structure function depending on the properties of the atmosphere at a certain time. As such, they have no specific vertical location for use in data assimilation algorithms, which is a focus of this thesis.

Particularly important is the dependence of radiances on clouds, through which satellite radiances have the potential to make significant contributions to forecast skill (McNally 2002). In this thesis we will focus on radiances at three wavelengths on board the Meteosat Second Generation satellite (Schmetz et al. 2002) and used at the German Weather Service: in the visible spectrum at $0.6\mu m$, and in the infrared spectrum at 6.2 and $7.3\mu m$.

2.2.1 Radiative transfer

Radiance is the amount of energy per unit area, time and solid angle emitted at a wavelength ν . Formally, the radiance $L(\nu)$ for an electromagnetic wave with frequency ν can be written in terms of its irradiant energy Q as:

$$L(\nu) = \frac{d^4Q}{dA dt d\Omega d\lambda} [W m^{-2} s^{-1} nm r],$$

where dA refers to unit area, dt the time interval, $d\Omega$ the direction and $d\lambda$ the wavelength. Radiance $L(\nu)$ is often converted to:

- Brightness temperature (K) by inverting the Planck equation. In our experiments this is the case for infrared radiances.
- The dimensionless quantity *reflectance* by non-dimensionalising and considering the radiance direction.

Radiances depend on atmospheric variables through the radiative transfer equation:

$$L(\nu) = \int_0^\infty B(\nu, T(z)) \frac{d\tau(\nu)}{dz} dz + \left[\begin{array}{c} \text{Contributions from} \\ \text{Surface} \end{array} \right] + \left[\begin{array}{c} \text{Reflection/Scattering} \\ \text{Terms} \end{array} \right] + \left[\begin{array}{c} \text{Cloud/Rain} \\ \text{Contributions} \end{array} \right]. \quad (2.35)$$

Where:

- Integrating over height z from 0 to ∞ represents integration from the bottom to top of the atmosphere;
- B is the radiance of a blackbody at temperature frequency ν and temperature T ;
- $\frac{d\tau(\nu)}{dz}$ represents absorption of radiation in the atmosphere;
- Temperature and absorption depend on height z .

Different frequency bands or *channels* depend in different ways on atmospheric variables through the absorption term and the surface, reflection and cloud contributions. It is considered beneficial to use several channels to give information on a range of properties of the atmosphere. When the primary absorber is a well-mixed gas (for example oxygen or CO_2), (2.35) is essentially a weighted average over the temperature profile.

The radiative transfer equation is much more complex and nonlinear when clouds are present, which presents challenges for incorporation into weather forecasts. Incorporating both cloud-affected and clear-sky radiances into weather forecasts is an active area of research.

2.2.2 Visible $0.6\mu m$ and Infrared $6.2, 7.3\mu m$ Channels

The radiance channels used operationally at the German Weather service are in the visible part of the spectrum at $0.6\mu m$ or infrared part of the spectrum at 6.2 or $7.3\mu m$. Visible (or solar) radiances measure the

amount of radiation from the sun reflected by the surface, clouds and aerosols, whereas infrared channels measure radiation emitted in the atmosphere. The $0.6\mu m$ channel is known as a ‘window’ channel - it is not sensitive to one specific gas and so gives information on a broad range of atmospheric components. The 6.2 and $7.3\mu m$ are most sensitive to water vapor and have weighting functions $\frac{d\tau(\nu)}{dz}$ that peak on average at around 450 or 550 hPa respectively.

According to Scheck, Weissmann, and Bach 2020, the channels give complementary information on the atmosphere in the following ways:

- The infrared channels are sensitive to both temperature and clouds, while visible radiances are nearly exclusively sensitive to clouds.
- The infrared channels give information on cloud top height, whereas visible channels are only weakly dependent on vertical cloud position.
- In infrared channels, low clouds are either invisible or indistinguishable from the ground or sea, whereas in visible channels they are discernible.

In general, lower values for infrared observations of brightness temperature imply the presence of clouds, as clouds absorb much outgoing infrared radiation that would otherwise be observed by the satellite. In contrast, higher values for visible reflectivity observations (in this thesis we use a threshold value of 0.4) is considered to imply the presence of a cloud, as clouds reflect a lot of incoming visible solar radiation back to the satellite.

2.2.3 Assimilating radiances

Observation Operators

Radiances can be assimilated either by making *retrievals* of atmospheric quantities or by *direct assimilation*. Making retrievals means exploiting radiation equations to infer standard quantities (for example temperature) that are used in the model, and assimilating these with a conventional observation operator (for example interpolation). In the direct method, the radiative transfer equation (2.35) is approximated and substituted into data assimilation algorithms as h , the observation operator. A big consideration here is the computational speed and accuracy of the radiative transfer operator. We consider here the direct assimilation approach, as this is what is used in the experiments later.

For infrared radiances, several methods of approximating (2.35) have been developed over the last decade. At DWD, the *Radiative Transfer for TIROS Operational Vertical sounders* package (RTTOV, Saunders et al. 2018) is used. Saunders, Matricardi, and Brunel 1999 showed that in this package (and more generally) radiative transfer is almost linearly dependent on temperature, but very nonlinearly dependent on humidity and cloud variables (e.g. cloud water content, cloud ice content).

The radiative transfer equation in the visible spectrum is much more complicated as reflection and scattering effects have a more important role, and it depends on 3D effects such as cloud inclination (Scheck et al. 2016). Until recently, there has been a lack of computationally feasible observation operators for visible radiances, so assimilation of visible radiances is relatively new. At DWD, the MFASIS (Scheck et al. 2016) program is used, which combines RTTOV with a look-up table to speed up computations. Scheck, Weissmann, and Bach 2020 showed that the operator is nonlinearly dependent on various cloud variables.

Localisation

Ensemble-based assimilation systems rely on localisation observations, but these are challenging to define for radiances. The horizontal location of radiance observations is generally clear, so distance based localisation systems are used horizontally as for conventional observations that directly observe model variables at a single point: the localisation factor is equal to one at the horizontal location of the observation and decays slowly to

0 as distance increases. As radiances depend on model variables through a vertical integral, their vertical location cannot be defined, so vertical localisation is a bigger challenge.

Localization schemes based on the sensitivity of the observation to atmospheric variables are often used. Miyoshi and Sato 2007 showed the effectiveness of using the raw sensitivity to model variables (dependent on the model state) as a localisation function instead of the Gaspari-Cohn function. It is noted however in Hutt et al. 2020 that sensitivity functions can have several peaks, and a localised function based on the sum of Jacobians (partial derivatives evaluated at different vertical heights) for different variables is proposed. The Gaspari-Cohn function is often fitted to sensitivity functions and used as a localisation function, as this is more similar to localisation functions for conventional observations. Computational methods have also been shown to have positive results (Lei, Anderson, and Whitaker 2016, Lei et al. 2020b, Anderson and Lei 2013).

Diagonal observation covariance matrix R

Although diagonal observation error matrices R are generally an acceptable assumption in data assimilation algorithms, observations from the same satellite are often physically close together and exhibit correlated errors. In many operational forecasting systems a diagonal R matrix is still used, and is a potential flaw when assimilating radiances.

2.3 Partial analysis increments and novel results on their use in observation influence analysis

The columns of the Kalman gain multiplied by their respective deviations $(y_o - h(x_b))_i$, or *partial analysis increments*, can be a useful diagnostic for the influence of a particular observation on analysis fields (Diefenbach et al. 2023). We show here a method for approximating columns of the Kalman gain using standard outputs of the DWD data assimilation system. We then give an explanation of how partial analysis increments can be used to approximate altered localisation settings in an operational data assimilation system, which will form the basis for experiments later in this thesis. In the second part of the section (subsection 2.3.2), we propose a novel characterisation for the error in the approximation of the Kalman gain with adjusted assimilation settings.

2.3.1 PAI definition and approximation with data assimilation outputs

Definition 2.3.1 (Partial Analysis Increments, PAI vectors (Diefenbach et al. 2023)). *Suppose we have a data assimilation system where $x_a = x_b + K(y_o - h(x_b))$. Partial analysis increments (PAI vectors) are the vectors $PAI = K_{:,i}(y_o - h(x_b))_i$. In words: columns of the Kalman gain multiplied by observation background departures.*

In the KENDA system (Schraff et al. 2016) of the DWD, the whole Kalman gain K is not stored as an output, so we approximate/evaluate K at each gridpoint using the following theorem. We will rely again on a linear assumption, for which we will use overbars to refer to ensemble means. In particular, refer to the ensemble mean of the background ensemble members as $\bar{x}_b = \frac{1}{N} \sum_{i=1}^N x_{b,i}$, and the ensemble mean of the background model equivalents similarly as $\bar{y}_b$.

Theorem 2.3.1 (Diefenbach et al. 2023). *Suppose in the LETKF we have that the linear approximation:*

$$h(\bar{x}_b + X_b w) = \bar{y}_b + Y_b w, \quad (2.36)$$

holds both when we take w to be the analysis weight vector w_a and when we take w to be a column of the

analysis ensemble weight matrix $(W_a)_{:i} = w_{a,i}$. Then we have at each local analysis point that:

$$K = \frac{1}{N-1} X_a Y_a^T (R^{-1} \circ \rho), \quad (2.37)$$

where ρ is the localisation function at the local analysis point.

Proof. From the definition of the LETKF, we have:

$$\begin{aligned} K &= X_b \tilde{P}_a Y_b^T (R^{-1} \circ \rho) \\ &= (N-1)^{-1} X_b (W_a W_a^T) Y_b^T (R^{-1} \circ \rho). \end{aligned}$$

From the linearity assumption (2.36) applied to the columns of W_a we have that:

$$Y_a = Y_b W_a, \quad (2.38)$$

and from the LETKF algorithm we have that $X_a = X_b W_a$. Substituting all of these into the formula for K gives the result. $\square$

The linearity assumption (2.36) on the ensemble of model equivalents (2.38) is the only assumption we have taken so far in this approximation. Although it is tenuous in the case of nonlinear h , it is consistent with the original assumptions of the LETKF. In the operational system of the DWD, Y_a is not stored but $Y_b W_a$, so as far as the linear approximation (2.38) is concerned, the expression of the Kalman gain will be exact in the experiments later in the thesis.

As a sanity check, observe that in the case of a linear observation operator represented by matrix H , we have that:

$$\begin{aligned} \frac{1}{N-1} X_a Y_a^T R^{-1} &= \frac{1}{N-1} X_a X_a^T H^T R^{-1} \\ &= P_a H^T R^{-1}, \end{aligned}$$

which is similar to the traditional Kalman gain in the form (2.3).

To produce the full partial analysis increment, we must make the approximation for K at each local analysis location, and then concatenate the result. In practise (in all the calculations in this paper and in Diefenbach et al. 2023, the paper that first proposed this method in a formal setting) this is carried out individually for each observation i and all (nearby) local analysis points, as opposed to the derivation above which was carried out at each local analysis point for all observations. The calculation of increment vectors proceeds as follows. We consider only the case of diagonal R as in the operational system of the DWD. As such, write each element of R as $R_{ii} = \sigma_i^2$.

1. Extract row i from Y_a corresponding to observation i , $(Y_a)_{i,:}$;
2. Calculate the matrix product $X_a (Y_a)_{i,:}^T$;
3. Calculate the horizontal and vertical distance from observation i to each local analysis point in model space;
4. Apply the horizontal and vertical localisation functions to the horizontal and vertical distances. Store these as vectors $\tilde{\rho}_h$ and $\tilde{\rho}_v \in \mathbb{R}^n$;
5. Take the Hadamard/Schur/elementwise product (appendix B.1) of $\tilde{\rho}_h$ and $\tilde{\rho}_v$ with the matrix product

calculated in step 2, and multiply everything by the scalar $1/\sigma_i^2$ and $\frac{1}{N-1}$:

$$\frac{1}{N-1} \frac{1}{\sigma_i^2} \tilde{\rho}_v \circ [\tilde{\rho}_h \circ ((X_a(Y_a)_i^T))].$$

6. Multiply the result by the scalar increment $(y_o - h(x_b))_i$.

Attention should be drawn to the difference between $\tilde{\rho}_h$, $\tilde{\rho}_v$, and ρ used in previous definitions for the LETKF. In previous definitions, $\rho \in \mathbb{R}^{m \times m}$ could be interpreted as a function depending on the local analysis point that multiplies each element of the inverse of the observation covariance matrix by a localisation value. The vectors $\tilde{\rho}_h \in \mathbb{R}^n$ and $\tilde{\rho}_v \in \mathbb{R}^n$ depend on the observation i in question, and multiply each element of a model space vector depending on the distance to observation i . This notation is perhaps confusing, but gives a good interpretation of how local analysis affects increments in practise.

Although in our experiments we do not make any approximation in (2.36), we make two other approximations relating to the storage of X_a and Y_a . The first is when moving between the coarse and general analysis grids. At the German weather service, the Kalman gain is calculated on a subset of gridpoints of the model, but we evaluate (2.3.1) at all gridpoints and ignore that values of W_a at many points are in fact interpolations. The second approximation is related to the storage of Y_a : for each observation, W_a at the observation location is used to calculate Y_a , not at the local analysis grid point in model space. As W_a changes smoothly between gridpoints, we expect these not to have a significant influence.

2.3.2 Using PAI to Localise Retrospectively

Partial analysis increments can be used to estimate the influence of varying localisation settings without calculating the whole analysis again. To do this we use the following approximation.

Definition 2.3.2 (Retrospective Localisation (Diefenbach et al. 2023)). *Suppose we have a diagonal R in an LETKF data assimilation system. Suppose we have used localisation function ρ_1 at a local analysis point to produce W_a , $X_a^{(1)}$ and $Y_a^{(1)}$. Approximate the Kalman gain with new localisation function ρ_2 at this point as:*

$$K_{newloc} = (N-1)^{-1} X_a^{(1)} (Y_a^{(1)})^T R^{-1} \circ \rho_2.$$

This approximation is called retrospective localisation. K_{newloc} is referred to as the retrospectively localised Kalman gain, and $(K_{newloc})_{:,i} (y_o - h(x_b))_i$ is referred to as a retrospectively localised partial analysis increment.

The approximation was verified experimentally in the case of assimilation with only one or two observation types in Diefenbach et al. 2023.

Novel characterisation of error in retrospective localisation

We propose some novel ideas to understand further where the retrospective localisation approximation might be acceptable. Intuition from these results can be applied to the experiments carried out later in this thesis, and could in future be used as a practical way of isolating where PAI-based approximations are reliable.

Refer to the Kalman gain calculated completely with the new localisation function ρ_2 as K_2 . Recall that under the linear approximation (2.36) this can be written as:

$$K_2 = (N-1)^{-1} X_a^{(2)} (Y_a^{(2)})^T (R^{-1} \circ \rho_2).$$

Define K_1 similarly. In retrospective localisation, we start with an analysis calculated with K_1 , and use K_{newloc} to approximate an analysis calculated with K_2 . In particular, we use the approximation:

$$K_2(y_o - h(x_b)) = K_1(y_o - h(x_b)) + (K_{newloc} - K_1)(y_o - h(x_b)) + [\text{Error}]. \quad (2.39)$$

We present here a formal characterisation of the error in (2.39).

Distinguish the analysis errors in ensemble space calculated with localisation functions ρ_1 and ρ_2 with the notation $\tilde{P}_a^{(1)}$ and $\tilde{P}_a^{(2)}$. Define $Y_a^{(1)}$ and $Y_a^{(2)}$ similarly. The proposed method is based on the result:

$$K_{\text{newloc}} - K_2 = \left[(N-1)^{-1} X_a^{(1)} (Y_a^{(1)})^T (R^{-1} \circ (\rho_2 - \rho_1)) \right] \tilde{E} \quad (2.40)$$

$$= (K_{\text{newloc}} - K_1) \tilde{E}, \quad (2.41)$$

where $\tilde{E}$ is a known matrix. We will give a bound on a matrix norm for $\tilde{E}$ under several conditions later. Note that the matrix $(K_{\text{newloc}} - K_1)$ in (2.41) is exactly what we remove when going from K_1 to K_{newloc} and will correspond to the PAI vectors when we multiply with the deviations. Indeed we can use (2.41) to get:

$$K_2 = K_{\text{newloc}} - (K_{\text{newloc}} - K_1) \tilde{E} \quad (2.42)$$

$$= K_1 + K_{\text{newloc}} - K_1 - (K_{\text{newloc}} - K_1) \tilde{E} \quad (2.43)$$

$$= K_1 + (K_{\text{newloc}} - K_1)(I_m - \tilde{E}), \quad (2.44)$$

where the first equality comes from (2.41). So if $\tilde{E}$ is small, we have good reason to believe that (2.39) is a good approximation.

We only give results on the Kalman gain, as results on the whole partial analysis increments with the first guess deviations included are much more difficult to obtain. For example, in the case where first guess deviations are 0 where we change the localisation and 1 everywhere else, the error is always much bigger than the PAI vector that we change.

Derivation of equation (2.41)

We have, using the same ideas as when deriving the PAI approximation previously:

Lemma 2.3.2. *If the linear assumption (2.36) holds, K_{newloc} differs from K_2 only through the analysis error in ensemble space matrix. Formally, assuming (2.36):*

$$K_{\text{newloc}} - K_2 = X_b (\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}) Y_b^T (R^{-1} \circ \rho_2). \quad (2.45)$$

Proof.

$$\begin{aligned} K_{\text{newloc}} - K_2 &= (N-1)^{-1} X_a^{(1)} (Y_a^{(1)})^T (R^{-1} \circ \rho_2) - (N-1)^{-1} X_a^{(2)} (Y_a^{(2)})^T (R^{-1} \circ \rho_2) \\ &= X_b \tilde{P}_a^{(1)} Y_b^T (R^{-1} \circ \rho_2) - X_b \tilde{P}_a^{(2)} Y_b^T (R^{-1} \circ \rho_2) \\ &= X_b (\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}) Y_b^T (R^{-1} \circ \rho_2), \end{aligned}$$

where the second equality comes from the linear assumption (2.36). □

The next proposition gives us a characterisation of the difference $\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}$. We do this by exploiting the *continuity of the matrix inverse* used to define $\tilde{P}_a$, which is expressed in a practical form in the following simple lemma.

Lemma 2.3.3. *Let $E \in \mathbb{R}^{N \times N}$ be such that $(I_N + E)$ is invertible. Then:*

$$(I_N + E)^{-1} = I_N - E(I_N + E)^{-1}. \quad (2.46)$$

Proof. Note that:

$$\left[I_N - E(I_N + E)^{-1} \right] (I_N + E) = I_N(I_N + E) - E(I_N + E)^{-1}(I_N + E) \quad (2.47)$$

$$= I_N + E - E \quad (2.48)$$

$$= I_N. \quad (2.49)$$

□

Proposition 2.3.1. *Let ρ_1 and ρ_2 be LETKF localisation functions for local analysis at a given gridpoint. Let $\tilde{P}_a^{(1)}$ and $\tilde{P}_a^{(2)}$ be the corresponding analysis error matrices in ensemble space, and $Y_a^{(1)}$, $Y_a^{(2)}$ the corresponding analysis perturbation matrices. Let E be the symmetric matrix*

$$E := (N - 1)^{-1} (Y_a^{(1)})^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_a^{(1)}.$$

If all the eigenvalues of E are all less than one in absolute value, we have that:

$$\tilde{P}_a^{(2)} = \tilde{P}_a^{(1)} - (N - 1)^{-1} (W_a^{(1)})^T E (I_N + E)^{-1} W_a^{(1)}.$$

Note: If we only change the localisation function for one observation type (in our case visible satellite observations), E only depends on observations of that type. It is therefore reasonable to expect that it will be small.

Proof. First note, starting with the definition of $\tilde{P}_a^{(2)}$:

$$\begin{aligned} \left(\tilde{P}_a^{(2)} \right)^{-1} &= (N - 1) I_N + Y_b^T (R^{-1} \circ \rho_2) Y_b \\ &= (N - 1) I_N + Y_b^T (R^{-1} \circ \rho_1) Y_b + Y_b^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_b \\ &= \left(\tilde{P}_a^{(1)} \right)^{-1} + Y_b^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_b. \end{aligned}$$

Then, noting that $W_a^{(1)}$ is symmetric:

$$(W_a^{(1)})^{-1} \tilde{P}_a^{(2)} (W_a^{(1)})^{-1} = \left[W_a^{(1)} (\tilde{P}_a^{(2)})^{-1} W_a^{(1)} \right]^{-1} \quad (2.50)$$

$$= \left[W_a^{(1)} \left((\tilde{P}_a^{(1)})^{-1} + Y_b^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_b \right) W_a^{(1)} \right]^{-1} \quad (2.51)$$

$$= \left[W_a^{(1)} (\tilde{P}_a^{(1)})^{-1} W_a^{(1)} + (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_b W_a^{(1)} \right]^{-1}. \quad (2.52)$$

The first term in the large brackets (2.52) can be resolved as:

$$W_a^{(1)} (\tilde{P}_a^{(1)})^{-1} W_a^{(1)} = W_a^{(1)} [(N - 1)^{-1} W_a^{(1)} (W_a^{(1)})^T]^{-1} W_a^{(1)} \quad (2.53)$$

$$= (N - 1) I_N. \quad (2.54)$$

For the second term in (2.52), we use again the linear assumption (2.36) to get:

$$(W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_b W_a^{(1)} = (Y_a^{(1)})^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_a^{(1)}. \quad (2.55)$$

Then substitute these two expressions (2.54) and (2.55) into the large brackets (2.52) to get:

$$(W_a^{(1)})^{-1} \tilde{P}_a^{(2)} (W_a^{(1)})^{-1} = \left((N-1)I_N + Y_a^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_a \right)^{-1} \quad (2.56)$$

$$= (N-1)^{-1} \left(I_N + (N-1)^{-1} Y_a^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_a \right)^{-1} \quad (2.57)$$

$$= (N-1)^{-1} (I_N + E)^{-1}. \quad (2.58)$$

We assumed that the eigenvalues of E are all less than 1, so $(I_N + E)$ must be invertible and we can apply lemma 2.3.3. Substitute the inverse from lemma 2.3.3 into (2.58) and multiply on the left and right by $W_a^{(1)}$ to get:

$$\begin{aligned} \tilde{P}_a^{(2)} &= W_a^{(1)} \left[(N-1)^{-1} \left(I_N - E(I_N + E)^{-1} \right) \right] W_a^{(1)} \\ &= (N-1)^{-1} W_a^{(1)} (W_a^{(1)})^T - (N-1)^{-1} (W_a^{(1)})^T E (I_N + E)^{-1} W_a^{(1)} \\ &= \tilde{P}_a^{(1)} - (N-1)^{-1} (W_a^{(1)})^T E (I_N + E)^{-1} W_a^{(1)}. \end{aligned}$$

□

From proposition 2.3.1 we can obtain the result stated above as (2.41).

Corollary 2.3.3.1. *Suppose that the conditions of theorem (2.3.1) hold (the eigenvalues of E are all less than or equal to 1 and the linear approximation (2.36) holds). Then we have that:*

$$\begin{aligned} K_{\text{newloc}} - K_2 &= \left[(N-1)^{-1} X_a^{(1)} (Y_a^{(1)})^T (R^{-1} \circ (\rho_2 - \rho_1)) \right] \tilde{E} \\ &= (K_{\text{newloc}} - K_1) \tilde{E}, \end{aligned}$$

where $\tilde{E}$ is defined as:

$$\tilde{E} = (N-1)^{-1} S_{1,2} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_2), \quad (2.59)$$

and $S_{1,2} \in \mathbb{R}^{m \times m}$ is the matrix that sends to 0 all coordinates where ρ_1 and ρ_2 are equal, and keeps all other coordinates the same (the $m \times m$ identity matrix with elements where ρ_1 and ρ_2 are the same set to 0).

Proof. Substitute the expression from proposition 2.3.1 for $\tilde{P}_a^{(2)}$ into (2.45) to get (again using the linear assumption (2.36)):

$$K_{\text{newloc}} - K_2 = X_b \left[(N-1)^{-1} (W_a^{(1)})^T E (I_N + E)^{-1} W_a^{(1)} \right] Y_b^T R^{-1} \circ \rho_2 \quad (2.60)$$

$$= (N-1)^{-1} X_a^{(1)} E (I_N + E)^{-1} (Y_a^{(1)})^T R^{-1} \circ \rho_2 \quad (2.61)$$

$$(2.62)$$

We then have, from the definition of E that:

$$(N-1)^{-1} X_a^{(1)} E = (N-1)^{-1} X_a^{(1)} \left[(N-1)^{-1} (Y_a^{(1)})^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_a^{(1)} \right] \quad (2.63)$$

$$= (N-1)^{-1} X_a^{(1)} \left[(N-1)^{-1} (Y_a^{(1)})^T (R^{-1} \circ (\rho_2 - \rho_1)) S_{1,2} Y_a^{(1)} \right] \quad (2.64)$$

$$= (K_{\text{newloc}} - K_1) \left[(N-1)^{-1} S_{1,2} Y_a^{(1)} \right]. \quad (2.65)$$

Substituting (2.65) into (2.61) gives the result.

□

For the next propositions, we look to understand the size of the errors we have characterised. For this, we will need the definition of a matrix norm. We use the spectral norm (see appendix B.3):

Definition 2.3.3 (Spectral/operator norm for matrices (Ciarlet 1989, p. 23)). *Let $A \in \mathbb{R}^{n \times m}$ be a matrix. Then define the spectral norm of A as:*

$$\|A\| = \sup \left\{ \frac{\|Av\|_2}{\|v\|_2} : v \in \mathbb{R}^m, v \neq 0 \right\}.$$

We use here $\|\cdot\|_2$ to denote the standard Euclidean norm defined by $\|v\|_2 = \sqrt{\sum_{i=1}^m v_i^2}$.

For several results we will also use the spectral radius, which is equal to the spectral norm for symmetric positive semi-definite matrices:

Definition 2.3.4 (Spectral radius (Ciarlet 1989, p. 7)). *Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then define the spectral radius of A as:*

$$\varrho(A) = \max \left\{ |\lambda_i| : i = 1, \dots, n \right\},$$

where λ_i are the eigenvalues (real or complex, simple or multiple) of A .

Proposition 2.3.2 (Ciarlet 1989, p. 27-28). *Let $A \in \mathbb{R}^{n \times n}$ be any square matrix, and $\|\cdot\|$ be the spectral norm for matrices. Then we have that:*

$$\varrho(A) \leq \|A\|.$$

For diagonalisable matrices, the result is exact:

Proposition 2.3.3. *Let $A \in \mathbb{R}^{n \times n}$ be a square diagonalisable matrix, and $\|\cdot\|$ be the spectral norm for matrices. Then we have that:*

$$\varrho(A) = \|A\|.$$

With these tools, we are now ready to prove some results on the size of our errors. As an illustration, first consider the difference $\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}$.

Proposition 2.3.4. *Suppose that $\|E\| < 1$ (as E is symmetric, this is equivalent to all of the eigenvalues of E being less than 1 in absolute value). The spectral norm of the difference $\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}$ between the analysis errors in ensemble space calculated with the two different localisation functions is bounded above by a function of the spectral norm of E . That is:*

$$\|\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}\| \leq (N-1)^{-1} \frac{\|E\|}{1 - \|E\|}. \quad (2.66)$$

Proof. Use the expression from proposition 2.3.1 to note that:

$$\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)} = (N-1)^{-1} (W_a^{(1)})^T E (I_N + E)^{-1} W_a^{(1)}. \quad (2.67)$$

Using a standard inequality for the spectral norm (appendix B, proposition B.3.1) we can split this up to get an upper bound on the norm as:

$$\|\tilde{P}_a^{(1)} - \tilde{P}_a^{(2)}\| = \|(N-1)^{-1} (W_a^{(1)})^T E (I_N + E)^{-1} W_a^{(1)}\| \quad (2.68)$$

$$\leq (N-1)^{-1} \|(W_a^{(1)})^T\| \cdot \|E\| \cdot \|(I_N + E)^{-1}\| \cdot \|W_a^{(1)}\|. \quad (2.69)$$

As $W_a^{(1)}$ is diagonalisable, proposition 2.3.3 gives that:

$$\varrho(W_a^{(1)}) = \|W_a^{(1)}\|. \quad (2.70)$$

We saw in lemma 2.1.3 the eigenvalues of $W_a^{(1)}$ are less than or equal to one, so its spectral radius and hence norm is also bounded above by 1. For $(I_N + E)^{-1}$, we use a standard bound relating to the spectral norm of an inverse (appendix B, proposition B.3.2):

$$\|(I_N + E)^{-1}\| \leq \frac{1}{1 - \|E\|}. \quad (2.71)$$

So we can bound (2.69) by $(N - 1)^{-1} \frac{\|E\|}{1 - \|E\|}$. □

We see in the next proposition that the condition of E having eigenvalues less than or equal to 1 always holds when we are just decreasing the localisation coefficients when going from ρ_1 to ρ_2 . Experimental methods are still required to show that the eigenvalues are significantly smaller than 1, or that the eigenvalues are smaller than 1 in the case of artificially adding observations into a forecast.

Proposition 2.3.5. *Suppose we have that elementwise:*

$$0 \leq \rho_2 \leq \rho_1, \quad (2.72)$$

and that the linear approximation (2.36) holds. Then the eigenvalues of E are strictly less than 1 in absolute value.

Proof. We use again the linear approximation (2.36) and write:

$$E = (N - 1)^{-1} (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_2 - \rho_1)) Y_b W_a^{(1)}.$$

Then using the fact that the matrix norm (2.3.3) is equal to the largest eigenvalue for positive semi-definite matrices in the second and last equalities (proposition 2.3.3) and that eigenvalues are preserved under cyclic permutation in a matrix product for the third equality (appendix B.5):

$$\|E\| = \|-E\| \quad (2.73)$$

$$= \|(N - 1)^{-1} (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_1 - \rho_2)) Y_b W_a^{(1)}\| \quad (2.74)$$

$$= \varrho \left[(N - 1)^{-1} (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_1 - \rho_2)) Y_b W_a^{(1)} \right] \quad (2.75)$$

$$= \varrho \left[(N - 1)^{-1} (R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} \right] \quad (2.76)$$

$$= \left\| (N - 1)^{-1} (R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} \right\|. \quad (2.77)$$

The square root $(R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}}$ is not ambiguous as $R^{-1} \circ (\rho_1 - \rho_2)$ is diagonal and non-negative. We use the square root to ensure that we are always working with diagonalisable matrices so that we can associate the spectral radius with the spectral norm.

Note that (2.77) still holds even though (2.74) and (2.77) use norms for matrices of different dimensions.

We apply now assumption (2.72) of the proposition. Let the fraction $\frac{\rho_1 - \rho_2}{\rho_1 + \mathbb{1}_{\{\rho_1=0\}}}$ be taken elementwise, where $\mathbb{1}_{\{\rho_1=0\}}$ is the indicator function for $\rho_1 = 0$ and is used to avoid considering separately the zero components introduced by localisation. From (2.72):

$$\left| \frac{\rho_1 - \rho_2}{\rho_1 + \mathbb{1}_{\{\rho_1=0\}}} \right| \leq 1. \quad (2.78)$$

To save space, define:

$$Z_{1,2} = I_m \circ \left(\frac{\rho_1 - \rho_2}{\rho_1 + \mathbb{1}\{\rho_1 = 0\}} \right). \quad (2.79)$$

Then we have that:

$$(R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} = \left[(R^{-1} \circ \rho_1) Z_{1,2} \right]^{\frac{1}{2}}. \quad (2.80)$$

Therefore:

$$(N-1)\|E\| = \left\| (R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T (R^{-1} \circ (\rho_1 - \rho_2))^{\frac{1}{2}} \right\| \quad (2.81)$$

$$= \left\| \left[Z_{1,2} (R^{-1} \circ \rho_1) \right]^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T \left[(R^{-1} \circ \rho_1) Z_{1,2} \right]^{\frac{1}{2}} \right\| \quad (2.82)$$

$$\leq \|Z_{1,2}\| \cdot \left\| (R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1)^{\frac{1}{2}} \right\| \cdot \|Z_{1,2}\| \quad (2.83)$$

$$\leq \left\| (R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1)^{\frac{1}{2}} \right\| \quad (2.84)$$

$$= \varrho \left[(R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_b W_a^{(1)} (W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1)^{\frac{1}{2}} \right] \quad (2.85)$$

$$= \varrho \left[(W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1) Y_b W_a^{(1)} \right] \quad (2.86)$$

$$= \left\| (W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1) Y_b W_a^{(1)} \right\|. \quad (2.87)$$

The first inequality is a standard result for the matrix norm (2.3.3) (appendix B, proposition B.3.1), and the second comes from (2.78).

Now note that we can diagonalise $W_a^{(1)}$ and $Y_b^T (R^{-1} \circ \rho_1) Y_b$ with the same basis as follows. Let Q_1 be an orthonormal basis of eigenvectors such that $Y_b^T (R^{-1} \circ \rho_1) Y_b = Q_1 \Lambda_1 Q_1^T$ and Λ_1 is a diagonal matrix of eigenvalues for $Y_b^T (R^{-1} \circ \rho_1) Y_b$. Then:

$$\tilde{P}_a^{(1)} = ((N-1)I_N + Q_1 \Lambda_1 Q_1^T)^{-1} \quad (2.88)$$

$$= \left(Q_1 ((N-1)I_N + \Lambda_1) Q_1^T \right)^{-1} \quad (2.89)$$

$$= Q_1 ((N-1)I_N + \Lambda_1)^{-1} Q_1^T. \quad (2.90)$$

So:

$$W_a^{(1)} = (W_a^{(1)})^T = (N-1)^{\frac{1}{2}} Q_1 ((N-1)I_N + \Lambda_1)^{-\frac{1}{2}} Q_1^T.$$

Using this basis, we can express the matrix of (2.87) as:

$$(W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1) Y_b W_a^{(1)} = (N-1) Q_1 ((N-1)I_N + \Lambda_1)^{-1} \Lambda_1 Q_1^T,$$

where we know that $((N-1)I_N + \Lambda_1)^{-1} \Lambda_1$ is a diagonal matrix with values equal to $\frac{\lambda^{(i)}}{N-1+\lambda^{(i)}}$ for $\lambda^{(i)}$ an eigenvalue of $Y_b^T (R^{-1} \circ \rho_1) Y_b$. In particular, $((N-1)I_N + \Lambda_1)^{-1} \Lambda_1$ is diagonal with entries strictly less than 1. Putting everything together, this gives:

$$\|E\| \leq (N-1)^{-1} \left\| (W_a^{(1)})^T Y_b^T (R^{-1} \circ \rho_1) Y_b W_a^{(1)} \right\| \quad (2.91)$$

$$= (N-1)^{-1} \left\| (N-1) Q_1 ((N-1)I_N + \Lambda_1)^{-1} \Lambda_1 Q_1^T \right\| \quad (2.92)$$

$$< 1. \quad (2.93)$$

In particular, the largest eigenvalue of E is strictly less than 1 in absolute value. $\square$

The proof gives quite a rough upper bound. In practise we expect eigenvalues of E to be a lot smaller than 1. Similar to this proof, we give an upper bound for the norm of $\tilde{E}$ in the case that we are just completely removing observations.

Proposition 2.3.6. *Suppose that the linear approximation (2.36) holds. Suppose that in going from assimilation with ρ_1 to assimilation with ρ_2 we are just removing observations, that is that elementwise:*

$$(\rho_2)_i = (\rho_1)_i \text{ or } 0. \quad (2.94)$$

Write the elements of the diagonal matrix R as $R_{ii} = \sigma_i^2$. Then we have that:

$$\|\tilde{E}\| \leq \frac{1}{1 - \|E\|} \cdot \max \left\{ \sqrt{\frac{\sigma_i^2}{(\rho_1)_i}} : (\rho_1 - \rho_2)_i \neq 0, (\rho_1)_i \neq 0 \right\} \cdot \max \left\{ \sqrt{\frac{(\rho_1)_i}{\sigma_i^2}} : (\rho_2)_i \neq 0 \right\}. \quad (2.95)$$

Note: The right hand side of (2.95) looks like the condition number of $R \circ \rho_1$ (the condition number of $R \circ \rho_1$ is however not defined as the localisation ρ_1 will in general send many elements to 0). In the case that we are just removing an observation type, so that either $(\rho_1)_i = (\rho_2)_i$, or $(\rho_2)_i = 0$ for all $i = 1, 2, \dots, m$, the result is saying:

- If we remove an observation type with high certainty close to the point (including the certainty implied by localisation with ρ_1), the error is smaller;
- If the remaining observations have low certainty (including the certainty implied by localisation with ρ_1), the error is smaller.

The certainty of the observation we remove will also have an effect on $\|E\|$.

Proof. Let $S_{2,1} \in \mathbb{R}^{m \times m}$ be the diagonal matrix with diagonal entries equal to 1 where $\rho_2 = \rho_1$, so that:

$$R^{-1} \circ \rho_2 = (R^{-1} \circ \rho_1) S_{2,1}. \quad (2.96)$$

Let $S_{1,2} \in \mathbb{R}^{m \times m}$ be the diagonal matrix with values equal to 1 when $(\rho_2)_i = 0$ and $(\rho_1)_i \neq 0$, so that $S_{1,2}$ is the same as in the definition of $\tilde{E}$. Let $Z_1 \in \mathbb{R}^{m \times m}$ be the diagonal matrix with entries $(Z_1)_{ii} = \frac{1}{(\rho_1)_i}$ when $(\rho_1)_i \neq 0$ and $(Z_1)_{ii} = 0$ when $(\rho_1)_i = 0$ so that:

$$S_{1,2} = S_{1,2} (RZ_1)^{\frac{1}{2}} (R^{-1} \circ \rho_1)^{\frac{1}{2}}. \quad (2.97)$$

Then incorporate these relations (2.96) and (2.97) into $\tilde{E}$:

$$\tilde{E} = (N - 1)^{-1} S_{1,2} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_2) \quad (2.98)$$

$$= (N - 1)^{-1} S_{1,2} (RZ_1)^{\frac{1}{2}} (R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_1) S_{2,1}. \quad (2.99)$$

We can then bound the spectral norm of E by the spectral norm of its factors given in (2.99):

$$\|\tilde{E}\| = \|(N - 1)^{-1} S_{1,2} (RZ_1)^{\frac{1}{2}} (R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_1) S_{2,1}\|. \quad (2.100)$$

$$\leq (N - 1)^{-1} \|S_{1,2} (RZ_1)^{\frac{1}{2}}\|. \quad (2.101)$$

$$\cdot \|(R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_1)^{\frac{1}{2}}\|. \quad (2.102)$$

$$\cdot \|(R^{-1} \circ \rho_1)^{\frac{1}{2}} S_{2,1}\|. \quad (2.103)$$

The middle term (2.102) is symmetric positive semi-definite, so we can relate the spectral norm to the spectral

radius to get that:

$$\| (R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_1)^{\frac{1}{2}} \| = \varrho \left[(R^{-1} \circ \rho_1)^{\frac{1}{2}} Y_a^{(1)} (I_N + E)^{-1} (Y_a^{(1)})^T (R^{-1} \circ \rho_1)^{\frac{1}{2}} \right] \quad (2.104)$$

$$= \varrho \left[(Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)} (I_N + E)^{-1} \right] \quad (2.105)$$

$$\leq \| (Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)} (I_N + E)^{-1} \| \quad (2.106)$$

$$\leq \| (Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)} \| \cdot \| (I_N + E)^{-1} \| \quad (2.107)$$

$$\leq \| (Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)} \| \cdot \frac{1}{1 - \|E\|} \quad (2.108)$$

$$= \varrho [(Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)}] \cdot \frac{1}{1 - \|E\|}. \quad (2.109)$$

We used in this the result from appendix B, proposition B.3.2 to bound the norm of $(I_N + E)^{-1}$. The last term can be bounded above by $(N - 1)$ in the same way as at the end of proposition 2.3.5. In particular, take a matrix Q_1 such that:

$$Y_b^T (R^{-1} \circ \rho_1) Y_b = Q_1^T \Lambda_1 Q_1, \quad (2.110)$$

where Λ_1 is diagonal with non-negative entries and

$$W_a^{(1)} = (W_a^{(1)})^T = (N - 1)^{\frac{1}{2}} Q_1 ((N - 1)I_N + \Lambda_1)^{-\frac{1}{2}} Q_1^T. \quad (2.111)$$

Then

$$(Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)} = (N - 1) Q_1 ((N - 1)I_N + \Lambda_1)^{-1} \Lambda_1 Q_1^T. \quad (2.112)$$

So in particular:

$$\varrho [(Y_a^{(1)})^T (R^{-1} \circ \rho_1) Y_a^{(1)}] \leq N - 1, \quad (2.113)$$

and we have a bound of $\frac{N-1}{1-\|E\|}$ for the middle term (2.102). As the other two terms (2.101) and (2.103) are norms applied to diagonal matrices, we can just consider the largest entry of the arguments to get:

$$\| S_{1,2} (R Z_1)^{\frac{1}{2}} \| = \max \left\{ \sqrt{\frac{\sigma_i^2}{(\rho_1)_i}} : (\rho_1 - \rho_2)_i \neq 0, (\rho_1)_i \neq 0 \right\}, \quad (2.114)$$

and:

$$\| (R^{-1} \circ \rho_1)^{\frac{1}{2}} S_{2,1} \| = \max \left\{ \sqrt{\frac{(\rho_1)_i}{\sigma_i^2}} : (\rho_2)_i \neq 0 \right\}. \quad (2.115)$$

Then substituting our upper bounds for (2.101), (2.102) and (2.103) respectively into our bound for $\|\tilde{E}\|$ in equations (2.100) to (2.103) gives the result. $\square$

In conclusion, we have derived:

- A characterisation for the change in analysis covariance in ensemble space when changing localisation, $\tilde{P}_a^{(2)} - \tilde{P}_a^{(1)}$, in proposition 2.3.1;
- A characterisation for the error when approximating K_2 with K_{newloc} in terms of the columns changed in the form $K_{\text{newloc}} - K_2 = (K_{\text{newloc}} - K_1) \tilde{E}$, in corollary 2.3.3.1;

- Bounds for relevant norms in these characterisations in propositions 2.3.5 and 2.3.6, and a potential condition for reliability of retrospective localisation through proposition 2.3.6.

Note there is no way to prove in general $\tilde{E}$ has a small spectral norm, and counter examples can be easily created where the spectral norm of $\tilde{E}$ is close to 1.

In our experiments, $Y_b W_a^{(1)}$ is stored as an output, so we can calculate E and the error through $(K_{\text{newloc}} - K_1)\tilde{E}$ at various locations for visible satellite observations. In particular, we calculate E where the difference $\rho_2 - \rho_1$ is equal to the horizontal localisation function for radiances, and equal to 0 everywhere else, as this gives an upper bound on the eigenvalues of E . Experiments showed that eigenvalues of E were much smaller than 1. The errors calculated through $\tilde{E}$ were on the scale of the PAI vectors, but generally not too much bigger.

Note that the only approximations made in this derivation are:

- The linear approximation (2.36);
- The assumption that stored analysis model equivalent perturbations are close to the actual analysis model equivalent perturbations.

2.4 Metrics and notation

Having defined PAI vectors and analysed the use of PAI vectors in retrospective localisation, we present several diagnostics that will later be used to understand the assimilation system with the PAI method. The most important of these metrics is the percentage improvement attributed to an observation with a given localisation function (2.118).

Model and observation vectors

Use superscripts to specify which model variable or observation is being referred to. In particular, x_b^T and x_b^q refer to background temperature and specific humidity values, and y_o^{VIS} and y_o^{RASO} refer to radiance observations in the visible channel and the (unassimilated) radiosonde observations. The subscripts follow the convention of the theory section (o for observations, b for background and a for analysis).

PAI vectors

PAI vectors for observations are calculated using the method explained in section 2.3 on partial analysis increments. Unless otherwise specified:

- PAI vectors are calculated for all observations of one type (for example radiances in the visible channel) and then the sum of all these vectors is taken and used. Because of localisation, only observations within a certain horizontal distance of each model gridpoint must be calculated.
- Vectors are calculated for one model variable at a time and on a subset of the whole model domain.

Write:

$$\text{PAI}^{y \rightarrow u},$$

to denote the PAI vector for the influence of observation y (for example the sum of PAI vectors for visible radiances VIS) on model variable u (for example the PAI vector just for specific humidity q). Write the concatenated PAI vector for observation y and all model variables as PAI^y .

We will often plot the difference between the final analysis and PAI vector to approximate the analysis with an observation removed. In particular, for some variable u and observation y , the vector:

$$x_a^u - \text{PAI}^{y \rightarrow u},$$

will be used to approximate the analysis calculated without the observation y (as in (2.39)).

For a PAI vector with localisation function ρ retrospectively incorporated, write $\text{PAI}_{\rho}^{y \rightarrow u}$. The vector $\text{PAI}^{y \rightarrow u}$ without subscript refers to PAI with the original localisation function. We approximate the analysis in the case that ρ was used instead of the original localisation function as:

$$x_a^u - \text{PAI}^{y \rightarrow u} + \text{PAI}_{\rho}^{y \rightarrow u}.$$

It is important to note which approximations we are making here:

- Approximating the Kalman gain columns for observation y through the PAI approximation (note from before this consists only of the approximation when moving from the coarse analysis grid to the full model grid, and the approximation relating to the storage of Y_a);
- An approximation of the PAI vectors for observation y with retrospective localisation;
- Approximating the analysis with observation y not assimilated as the analysis with y included subtract the partial analysis increments for observation y . This includes an approximation of the columns of the Kalman gain that do not correspond to y , where y has a different localisation function.

We can use our error analysis from section 2.3.2 to understand the last two of these.

Relative and specific humidity

The model and verifying observations use different measures of humidity, and we adapt some notation to account for this. In the model, specific humidity - the mass of water vapour per unit mass of air (kg/kg) - is used. The verifying radiosondes take measurements of relative humidity - the ratio of vapour pressure in the air to the vapour pressure in saturated air at the given pressure and temperature. Relative humidity can be calculated from specific humidity given the pressure and temperature at a given time. Write for relative humidity r , temperature T , pressure p and specific humidity q :

$$r = g(q, T, p),$$

and define g applied to vectors element-wise. The function g is non-linear, so we often plot the modified PAI vector for relative humidity:

$$\widetilde{\text{PAI}}^{y \rightarrow r} := g(x_a^q, x_a^T, x_a^p) - g(x_a^q - \text{PAI}^{y \rightarrow q}, x_a^T - \text{PAI}^{y \rightarrow T}, x_a^p - \text{PAI}^{y \rightarrow p}),$$

instead of calculating the PAI vector for relative humidity in the same way as other variables. Similarly we estimate the relative humidity with an observation removed as:

$$g(x_a^q - \text{PAI}^{y \rightarrow q} + \text{PAI}_{\rho}^{y \rightarrow q}, x_a^T - \text{PAI}^{y \rightarrow T} + \text{PAI}_{\rho}^{y \rightarrow T}, x_a^p - \text{PAI}^{y \rightarrow p} + \text{PAI}_{\rho}^{y \rightarrow p}).$$

Note also that this means retrospective localisation is applied before calculating relative humidity.

Verifying radiosondes

Refer to the unassimilated radiosondes as y_o^{RASO} . As there were no assimilated radiosondes in the experiment, there should be no confusion as to whether this is an assimilated observation.

To compare model outputs to unassimilated radiosondes (the observations used for verification), we need to interpolate to the location of the unassimilated radiosondes and calculate relative humidity. Define the observation operator h_{RASO} to be the operator taking model output to model equivalents for the unassimilated verifying radiosondes. This consists of interpolation for temperature T , and interpolation followed by calculation of relative humidity for relative humidity r . Use superscripts to refer to the restriction of this operator h_{RASO}^T and h_{RASO}^r to temperature and relative humidity respectively.

From this operator, for a vector x in model space, we can calculate the error to radiosondes as:

$$e_{\text{RASO}}(x) = (h_{\text{RASO}}(x) - y_o^{\text{RASO}}). \quad (2.116)$$

Benefit

As a diagnostic to see where vertically an observation improves an analysis, we can look at the difference between absolute error for the analysis with the observation removed and the absolute error with the observation assimilated. We approximate this with PAI at the verifying radiosondes by taking absolute values element-wise to give:

$$b(x_a) = |e_{\text{RASO}}(x_a - \text{PAI}^y)| - |e_{\text{RASO}}(x_a)|.$$

Refer to this as the *benefit* of observation y . We can easily adapt this to approximate the benefit of using new localisation settings as:

$$\tilde{b}(\rho) = |e_{\text{RASO}}(x_a - \text{PAI}^y)| - |e_{\text{RASO}}(x_a - \text{PAI}^y + \text{PAI}_{\rho}^{y \rightarrow x})|.$$

Cost function for new localisation settings

To measure how well an analysis x performs compared to the verifying radiosondes, we would be interested to take a norm of the vector $e_{\text{RASO}}(x)$. To weight this norm to allow for the different scales and units of the verifying variables, we can use the radiosonde error estimates contained in the observation error matrix R of the forecast system. Let R_{RASO} refer to the (diagonal) observation covariance matrix for radiosondes. We define the cost function J for a localisation function ρ :

$$J(\rho) = \|e_{\text{RASO}}(x_a - \text{PAI}^y + \text{PAI}_{\rho}^y) R_{\text{RASO}}^{-\frac{1}{2}}\|_2^2 = \sum_{\substack{\text{Radiosonde} \\ \text{obs } i}} \frac{1}{\sigma_i^2} |h_{\text{RASO}}(x_a - \text{PAI}^y + \text{PAI}_{\rho}^y) - y_{o,i}^{\text{RASO}}|^2, \quad (2.117)$$

where multiplication with $R_{\text{RASO}}^{-\frac{1}{2}}$ is standard matrix multiplication, and σ_i^2 are the diagonal entries of R_{RASO} . This cost function is similar to that of Diefenbach et al. 2023. The use of radiosonde error values is new in this thesis, and the notation is slightly different to emphasise the approximation of the columns of the Kalman gain not representing y . Note that the radiosonde observation error should also affect the values of the cost function.

Percentage improvement attributed to observation

So that the cost function (2.117) can easily be compared between different observation types and data sets, we use the cost function when the PAI vector is completely removed (corresponding to a localisation function of $\rho = 0$ constant, approximating the removal of the observation) as a benchmark. In particular, we will often discuss the percentage improvement when including a PAI vector with a specific retrospective localisation compared to removing the PAI vector completely. This can be written as:

$$\Delta J(\rho)\% = \frac{J(\rho) - J(0)}{J(0)} \times 100\%, \quad (2.118)$$

where $J(0)$ is the cost function for a retrospective localisation with function $\rho = 0$, that is:

$$J(0) = \|e_{\text{RASO}}(x_a - \text{PAI}^y) R_{\text{RASO}}^{-\frac{1}{2}}\|_2^2. \quad (2.119)$$

We aim to find localisation functions with low (negative) values of $\Delta J(\rho)\%$, where negative values suggest that with localisation function ρ , the observation y makes an improvement to the analysis at the assimilation time when verified against radiosondes.

2.5 Optimisation methods

It was chosen to minimise the cost function (2.117) (resp. percentage improvement in cost function (2.118)) against non-assimilated radiosondes as the interaction between the effect of radiosondes in the assimilation and optimisation is difficult to understand. It is expected that after assimilating radiosondes the analysis should be much closer to the radiosondes, so assimilating other observations at these locations would be correcting noise, an effect that we are not interested in studying. Note that there is still an error in the radiosonde observations present in our analysis and calculations, but as the radiosonde error should be statistically independent from the errors relating to the assimilation system and forecast model, it should not affect conclusions.

2.5.1 Optimisation of the Gaspari Cohn function

To assess the performance of each set of parameters for the Gaspari Cohn function used for vertical localisation of radiances, we used retrospective localisation as in section 2.3.2 to approximate the analysis with a wide range of localisation parameters, and evaluated the cost function (2.117) for each set of parameters. In particular this is calculated for each channel at each location and variable u as follows:

1. Calculate the sum of all PAI vectors for the observation at the location, PAI^y ;
2. Approximate the analysis without the observation as $x_a^u - \text{PAI}^{y \rightarrow u}$;
3. Approximate the error without the observation as $\|e_{\text{RASO}}(x_a^u - \text{PAI}^{y \rightarrow u})R_{\text{RASO}}^{-\frac{1}{2}}\|_2^2$. This will be used as a benchmark to compare other localisation functions to;
4. For each set of parameters: nominal height z and width σ , calculate the Gaspari-Cohn function $\rho = \rho(z, \sigma)$;
5. Approximate the analysis with each of these localisation functions as $x_a^u - \text{PAI}^{y \rightarrow u} + \text{PAI}_{\rho}^{y \rightarrow u}$;
6. Approximate the error with each of these localisation functions using the cost function (2.117):

$$\|e_{\text{RASO}}(x_a^u - \text{PAI}^{y \rightarrow u} + \text{PAI}_{\rho}^{y \rightarrow u})R_{\text{RASO}}^{-\frac{1}{2}}\|_2^2.$$

The percentage improvement compared to the approximation for observation removal (2.118) is then plotted as a more easily interpretable metric.

2.5.2 Empirical Localisation Function (ELF) Derived from PAI

We can also look at some idealised localisation functions to simplify the optimisation process and make more conclusions. In particular, the cost function (2.117) was optimised directly for piecewise linear functions of the index of the model level (as opposed to the pressure value at model levels), as this can be carried out analytically. The optimisation process is comparable to linear regression, and will be used as an indicator of the certainty of the Gaspari-Cohn optimisation at different levels.

Formally, define the idealised localisation functions as follows: for each vertical level we always multiply by the same factor, regardless of pressure value. We can consider the localisation function $\rho \in \mathbb{R}^{N_v}$ to be a vector for N_v the number of vertical levels in the model, where for all local analysis points on the i th vertical level of the model, we localise radiances vertically with the same factor of ρ_i .

For these idealised localisation functions, the cost function is (convex) quadratic and can be minimised by analytically deriving the gradient and setting it to zero. Optimising localisation functions directly is similar to the method of Anderson and Lei 2013, where *empirical localisation functions* are derived from simulations of the observing system.

Derivation of the optimal idealised localisation function

Recall our cost function (where SAT refers to PAI for a given (SATellite) radiance channel) defined in 2.4:

$$J(\rho) = \sum_{\substack{\text{Radiosonde} \\ \text{obs } i}} \frac{1}{\sigma_i^2} |h_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT}} + \text{PAI}_\rho^{\text{SAT}}) - y_{o,i}^{\text{RASO}}|^2. \quad (2.120)$$

The optimisation is based on writing this cost function (2.120) as:

$$J(\rho) = (\delta y_{\text{RASO}} - C\rho)^T R_{\text{RASO}}^{-1} (\delta y_{\text{RASO}} - C\rho), \quad (2.121)$$

for some vector $\delta y_{\text{RASO}} \in \mathbb{R}^k$ and some matrix $C \in \mathbb{R}^{N_v \times k}$ with k the number of radiosonde observations. As this is (convex) quadratic in ρ , we will then get a vector of localisation values for each vertical level as a matrix computation given as:

$$\hat{\rho} = (C^T R_{\text{RASO}}^{-1} C)^{-1} C^T R_{\text{RASO}}^{-1} \delta y_{\text{RASO}}.$$

We show formally the quadratic nature of the cost function in this case in the following proposition.

Proposition 2.5.1. *Assume that for a given observation, vertical localisation is carried out by multiplying each model level i by a different constant factor ρ_i (as described above). Let $y_i^{\text{RASO}}, i = 1, \dots, N_{\text{RASO}}$ be verifying radiosonde observations with variances $\sigma_{\text{RASO},i}^2$ combined into a diagonal covariance matrix R_{RASO} . Let h_{RASO} be the interpolation operator from model space to the radiosonde locations (ignore the case of relative humidity observations for now, as these depend non-linearly on the model space vector). Assume h_{RASO} does not depend on ρ .*

Let $J(\rho)$ be the cost function (as in Diefenbach et al. 2023 but including the verifying observation covariances and with the different localisation method), that is:

$$J(\rho) = \sum_{\substack{\text{Radiosonde} \\ \text{obs } i}} \frac{1}{\sigma_i^2} |h_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT}} + \text{PAI}_\rho^{\text{SAT}}) - y_{o,i}^{\text{RASO}}|^2.$$

Then minimising J is equivalent to a weighted least squares minimisation:

$$J(\rho) = (\delta y_{\text{RASO}} - C\rho)^T R_{\text{RASO}}^{-1} (\delta y_{\text{RASO}} - C\rho),$$

where $\delta y_{\text{RASO}} \in \mathbb{R}^{N_{\text{RASO}}}$ and $\rho \in \mathbb{R}^{N_v}$ are vectors, $C \in \mathbb{R}^{N_{\text{RASO}} \times N_v}$ and $R_{\text{RASO}}^{-1} \in \mathbb{R}^{N_{\text{RASO}} \times N_{\text{RASO}}}$ are matrices, and multiplications can be interpreted as standard matrix operations.

In particular, the optimal values for ρ are given as:

$$\hat{\rho} = (C^T R_{\text{RASO}}^{-1} C)^{-1} C^T R_{\text{RASO}}^{-1} \delta y_{\text{RASO}}.$$

Proof. First note that by linearity, we can separate out the radiosonde equivalent values in the error to get (where e_{RASO} is defined in (2.116)):

$$\begin{aligned} h_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT}} + \text{PAI}_\rho^{\text{SAT}}) - y_{o,i}^{\text{RASO}} &= h_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT}}) + h_{\text{RASO}}(\text{PAI}_\rho^{\text{SAT}}) - y_{o,i}^{\text{RASO}} \\ &= e_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT}}) + h_{\text{RASO}}(\text{PAI}_\rho^{\text{SAT}}). \end{aligned}$$

So if we define:

$$\delta y_{\text{RASO}} := e_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT}}),$$

then we just need to show that $h_{\text{RASO}}(\text{PAI}_\rho^{\text{SAT}})$ is a linear function of ρ . We present a constructive proof that

can then be used to calculate the matrix C in the final result.

First consider the case where we restrict model space to just one vertical column and one variable (perhaps re-ordering the model space vector so the model levels are in the correct order), and consider the influence of a single satellite observation indexed by i . Then the elementwise product (B.1) for vertical localisation in the PAI calculation of section 2.3 can be written as a matrix product using the following notation (where the first equality comes from the algorithm described in section 2.3, and the second from rearranging the Hadamard/Schur product):

$$\text{PAI}_\rho^i := \frac{1}{N-1} \tilde{\rho}_v \circ [\tilde{\rho}_h \circ ((X_a(Y_a)^T)_{i:})] (y_o - h(x_b))_i \quad (2.122)$$

$$= \frac{1}{N-1} (y_o - h(x_b))_i \text{diag}[\tilde{\rho}_h \circ ((X_a(Y_a)^T)_{i:})] \rho \quad (2.123)$$

$$= \text{diag}(\text{PAI}_\mathbb{1}^i) \rho. \quad (2.124)$$

Here we use $\mathbb{1}$ to refer to the vertical localisation function taking the value 1 everywhere (no vertical localisation). In particular:

$$\text{PAI}_\mathbb{1}^i = \frac{1}{N-1} [\tilde{\rho}_h \circ ((X_a(Y_a)^T)_{i:})] (y_o - h(x_b))_i$$

The relation (2.124) is true for all variables, all vertical columns, and all observations of our observation type (due to the simplified nature of such a vertical localisation), so we can generalise to all model space vectors by re-ordering the elements of the model space vector, details of which are given in the following paragraphs.

Let N_h be the number of horizontal gridpoints at which we calculate PAI. Consider for now the case of one variable. The case for model vectors with more than one variable follows by taking more columns for the other variables considered. First extend (2.124) by concatenating model space vector elements for multiple columns as follows. After re-ordering the elements of the model space vector, the retrospectively localised partial analysis increment can be written in terms of a standard matrix multiplication as:

$$\text{PAI}_\rho^i = (\widehat{\text{PAI}}_\mathbb{1}^i) \rho, \quad (2.125)$$

where $\widehat{\text{PAI}}_\mathbb{1}^i \in \mathbb{R}^{N_v N_h \times N_v}$ is a block matrix of diagonal matrices with values corresponding to PAI vectors at each horizontal point:

$$\widehat{\text{PAI}}_\mathbb{1}^i = \begin{bmatrix} \text{diag}(\text{PAI}_{\mathbb{1}, \text{gridpoint } 1}^i) \\ \vdots \\ \text{diag}(\text{PAI}_{\mathbb{1}, \text{gridpoint } N_h}^i) \end{bmatrix}.$$

Equation (2.125) is achieved by concatenating equation (2.124) along horizontal locations.

As (2.125) is true for all observations of our observation type, we can take:

$$\text{PAI}_\rho^{\text{SAT}} = \sum_{\text{Satellite obs } i} \text{PAI}_\rho^i \quad (2.126)$$

$$= \sum_{\text{Satellite obs } i} (\widehat{\text{PAI}}_\mathbb{1}^i) \rho \quad (2.127)$$

$$= \left[\sum_{\text{Satellite obs } i} \widehat{\text{PAI}}_\mathbb{1}^i \right] \rho \quad (2.128)$$

$$:= \widehat{\text{PAI}}_\mathbb{1}^{\text{SAT}} \rho. \quad (2.129)$$

We have defined here $\widehat{\text{PAI}}_\mathbb{1}^{\text{SAT}}$ as the sum of all matrices $\widehat{\text{PAI}}_\mathbb{1}^i$ for all satellite observations i .

To finish, express the operator h_{RASO} in matrix form as H to get:

$$h_{\text{RASO}}(\text{PAI}_{\rho}^{\text{SAT}}) = H \widehat{\text{PAI}}_{\mathbb{I}}^{\text{SAT}} \rho,$$

so if $C = -H \widehat{\text{PAI}}_{\mathbb{I}}^{\text{SAT}}$ we have the desired expression.

To get the optimal $\hat{\rho}$ expression in the statement, we differentiate the quadratic form of the cost function (2.5.1) and set this to 0. $\square$

Note 1: The differentiation at the end of the proof depends on h_{RASO} being independent of ρ , which is only approximately true.

Note 2: The definition of the matrix $\widehat{\text{PAI}}_{\mathbb{I}}^{\text{SAT}}$ in the previous proof may seem excessive, but it is helpful for calculating the C matrix in practise.

To understand this result better, consider the case that radiosonde observations are exactly at the level of model gridpoints (so no interpolation is used). For a given gridpoint i we then have:

$$\hat{\rho}_i = \frac{\left\langle \frac{1}{\sigma_{\text{RASO},i}^2} \cdot \delta y_{\text{RASO},i} \cdot \text{PAI}_{\text{noloc},i} \right\rangle}{\left\langle \frac{1}{\sigma_{\text{RASO},i}^2} \cdot \text{PAI}_{\text{noloc},i}^2 \right\rangle}, \quad (2.130)$$

where $\frac{1}{\sigma_{\text{RASO},i}^2}$ is the radiosonde error at a given point, and angle brackets refer to an average taken over all radiosondes at the given height level. While this is not useful in our experiments (radiosondes rarely take observations at model levels), it gives a good interpretation for the optimal values we have. In this simple case, if a partial analysis increment is positively correlated with the error excluding radiances (it is in general ‘beneficial’), then ρ will be positive, whereas if the increment is on average detrimental, ρ will be negative. The expression (2.130) is very similar to the ELF in Anderson and Lei 2013.

The expression (2.130) also suggests that if the radiance increment is strongly/negatively correlated with the error excluding radiances, $\hat{\rho}$ will could be larger than one/less than 0. Localisation values greater than one are generally considered impractical, as they interfere with the effects of inflation and choice of observation error covariance values in R , and negative localisation values would be counter-intuitive. To exclude optimal values greater than one or less than zero, we could:

- Replace all localisation values $\hat{\rho}$ with $\tilde{\rho}_i = \max(0, \min(1, \hat{\rho}_i))$. This could be sensible as optimal values greater than 1 suggest beneficial increments and negative optimal values suggest non-beneficial increments that could be removed by localisation, or:
- Optimise the cost function over the set of vectors ρ with values between 0 and 1. While this is a more complex optimisation, the results could be considered transparent as they ‘minimise the cost over all feasible ρ ’.

Both of these methods gave similar results in practise, so the first was chosen.

In the LETKF, it is important that localisation functions are suitably smooth to ensure smooth analysis fields. A quick way to enforce this would be to optimise over localisation coefficients for a subset of the vertical levels, and interpolate between these to get the coefficients for other levels. This can be easily implemented by multiplying a shorter ρ with an interpolation matrix L , so that C is replaced by $\tilde{C}$ where:

$$\tilde{C} = CL.$$

Empirical localisation functions as regression - confidence intervals

The least squares form (2.5.1) of the cost function leads us to interpret the empirical localisation function (2.5.2) as coefficients in a linear regression. Indeed if we suppose that:

$$\delta y_{\text{RASO}} = C\rho + \epsilon, \quad (2.131)$$

for $\epsilon \sim \mathcal{N}(0, \sigma R)$, $\sigma > 0$ some (unknown) error term, C constant and known from PAI, and ρ unknown, then the empirical localisation function coefficients $\hat{\rho}$ are exactly the maximum likelihood estimators for ρ in (2.131). Here we take a more general weighted least squares with unrestricted σ to account for the randomness in the PAI vector, as opposed to restricting to $\sigma = 1$ to include just the radiosonde error. The inclusion of the weights goes some way to accounting for the heteroscedasticity in the regression, but cannot account for everything.

The benefit of viewing the empirical localisation function as a least squares method is that we can then introduce confidence intervals for $\hat{\rho}$ coefficients using t-values. This gives us an (albeit rough) idea of how certain we are of the values of the coefficients.

Overview of the empirical localisation function calculations

To summarise, the empirical localisation functions were calculated as follows (where T refers to temperature and SAT refers to the channel being considered):

1. At each location, calculate the PAI vectors with the original localisation functions $\text{PAI}^{\text{SAT} \rightarrow T}$;
2. Combine $\text{PAI}^{\text{SAT} \rightarrow T}$, x_a and the radiosonde observations at all locations to calculate the vector $\delta y_{\text{RASO}} := e_{\text{RASO}}(x_a - \text{PAI}^{\text{SAT} \rightarrow T})$;
3. Calculate the PAI vector with no localisation $\text{PAI}_{\text{no loc}}^{\text{SAT}}$;
4. Expand this ‘unlocalised’ PAI vector into block matrix form $\widehat{\text{PAI}}_{\mathbb{I}}^{\text{SAT}}$
5. Apply the interpolation operator represented by matrix H for the radiosondes on the left and for the localisation coefficients with L on the right to get the matrix $C = -H\widehat{\text{PAI}}_{\mathbb{I}}^{\text{SAT}}L$;
6. Carry out a weighted linear regression of the form: $\delta y_{\text{RASO}} = C\rho + \epsilon$ for coefficients $\hat{\rho}$ with confidence intervals based on t-tests.

Chapter 3

Data and model

3.1 Data

3.1.1 Model and assimilation setup

Experiments are based on a forecast run for June 2023 with a model and data assimilation system as close as possible to that used by the DWD. Assimilation was run through the Kilometer scale ENsemble Data Assimilation system (KENDA, Schraff et al. 2016), and uses the ICOSahedral Non-hydrostatic Limited Area Model (ICON-LAM), so equations are resolved and outputs stored on a triangular grid. The model is at convective scale: the horizontal grid resolution is on average 2.1km between horizontal grid points, and 65 vertical levels were used, so processes on the kilometre scale can be resolved explicitly by the model (many approximations for smaller scale dynamics are required in lower resolution models). The system consists of a nested model over Germany - a lower resolution, less computationally expensive model is evaluated over a broader region for boundary conditions, and a higher resolution is used over Germany for the forecast. For the boundary conditions, the global ICON model with grid resolution 13km is run with a grid refinement of 6.5km over Europe.

The simulations were carried out for the month of June 2023. The weather was typically hot, and alternated between days with little clouds or rain, and days with lots of cloud cover and precipitation.

40 ensemble members were used and observations were assimilated every hour. All observations used by the German weather service in operational forecasts were assimilated except for radiosondes that were used for verification. Assimilated observations were:

- Direct observations of model variables taken at land stations, buoys and ships (SYNOP);
- Direct observations of upper regions of the atmosphere from planes;
- Satellite based direct retrievals of model variables;
- Radar observations;
- Wind profilers - measurements of wind speed at various locations using radar data.

General localisation strategy

The LETKF analysis as described at the end of section 2.1.2 was used at each point on a coarse analysis grid, where every 4th point in the horizontal and 3rd point in the vertical was used. Values of W_a and w_a were interpolated to other model grid points for forecasting.

A smooth distance-based localisation function was used for localisation, both in the horizontal and vertical

directions. In the vertical (for all observations except radiances), the width of this function was for all observation types held constant, with width values specified ‘offline’ - before starting the experiment. In the horizontal direction, adaptive covariance localisation was used for all observations except radiances: at each point in the coarse analysis grid the width of the horizontal localisation function was adapted so that the sum of observation weights was held constant where possible. As there are only 40 ensemble members, there are in turn only 40 degrees of freedom in the assimilation system, so including more than 40 (in weighted sum based on localisation factors) observations would be unlikely to benefit the forecast (Schraff et al. 2016). For radiances a constant horizontal localisation width of 25km was used.

Vertical localisation of radiances

In the visible 0.6 μm channel no vertical localisation was applied. In the infrared channel, the operational strategy of the DWD was used for vertical localisation:

For a local analysis point and one observation, the vertical localisation factor is calculated by:

1. Calculating the log of the pressure values at the local analysis point and the observation location;
2. Calculating the (absolute) difference of these;
3. Applying the Gaspari-Cohn function (Gaspari and Cohn 1999) with a given width (depending on the observation and adaptive localisation) to the difference.

The Gaspari-Cohn function is a smooth 5th order piecewise rational function with compact support that approximates a Gaussian bell curve. A Gaussian is first approximated with the function $C_0 : \mathbb{R} \rightarrow [0, 1]$ defined as:

$$C_0(x) := \begin{cases} -\frac{1}{4}x^5 + \frac{1}{2}x^4 + \frac{5}{8}x^3 - \frac{5}{3}x^2 + 1 & \text{if } 0 \leq x \leq 1, \\ +\frac{1}{12}x^5 - \frac{1}{2}x^4 + \frac{5}{8}x^3 - \frac{5}{3}x^2 - 5x + 4 - \frac{2}{3}\frac{1}{x} & \text{if } 1 < x < 2, \\ 0 & \text{if } x \geq 2. \end{cases}$$

The function C_0 is then modified to have different widths controlled by a parameter σ , where large σ refers to a wide function. In particular, for a distance d and a function with width σ , the localisation is taken to be equal to $C_0\left(\frac{\sqrt{3}}{\sigma\sqrt{10}}d\right)$.

Operationally at the DWD, the width σ is currently taken to be 0.3 for radiances. The nominal height is taken to be 800hPa for the visible channel (unlike in our experiment where no vertical localisation was applied in this channel), and an adaptive solution based on the transmission function is used for the infrared radiances, so that the nominal height varies between 200 and 500 hPa depending on the value of the infrared observation (as in our experiment).

Inflation

It is commonly observed that ensembles underestimate the variance of variables. This can be for various reasons, including deficiencies in the representation of model error and limitations in the data assimilation system (for example incorrectly specified errors, or issues relating to the normal assumption, among many others). Inflation is an ad hoc method for artificially increasing the variance of the ensemble (Asch 2016, p. 171-172).

Three different forms of inflation were used in the model runs examined here: multiplicative (adaptive) inflation, additive inflation, and relaxation to prior perturbations. We describe these here following the text of Schraff et al. 2016.

Multiplicative inflation

Possibly the simplest way to increase the spread of an ensemble is to multiply the perturbations by a constant factor ρ (we use the same letter as for localisation to be consistent with the inflation notation of the German

Weather Service), where we increase spread when $\rho > 1$. This is known as *multiplicative localisation* and was first introduced in Anderson and Anderson 1999. In the LETKF multiplicative inflation can be implemented in two ways. One option is to multiply the identity matrix in the expression for $\tilde{P}_a$ by a factor of $1/\rho$ to get:

$$\tilde{P}_a^{\text{inf}} = \left[\frac{N-1}{\rho} I_N + Y_b^T R^{-1} Y_b \right]^{-1}.$$

To see how this might increase spread, we can note (for $\rho > 1$) that the inflation would decrease the eigenvalues of the matrix within the brackets, so increase the eigenvalues of $\tilde{P}_a$ (and therefore norm as $\tilde{P}_a$ is symmetric). Intuitively, we are multiplying X_b by a factor of $\sqrt{\rho}$, so increasing spread before assimilation. As such, this is referred to as *prior* inflation.

The other way to implement multiplicative localisation in the LETKF is to apply the localisation after assimilation, known as *posterior* inflation. In this case, we apply the localisation by multiplying the W_a matrix by a factor of ρ :

$$W_a^{\text{inf}} = \rho W_a.$$

In this case, we are multiplying the analysis perturbations X_a by a factor of ρ . It is more clear why this leads to increased covariances. Note that posterior multiplicative inflation has no effect on the mean x_a .

Adaptive multiplicative inflation

In the German weather service, the coefficient ρ used in multiplicative inflation is adapted through time and space based on forecast outputs. It is based on the idea that, according to the assumptions in the data assimilation derivations, the deviations $y_o - h(x_b)$ should have covariance similar to $R + HP_bH$. The covariance is then chosen so that this is at least true on average by considering traces of these matrices as in Houtekamer et al. 2005.

Additive inflation

Additive inflation consists of adding randomly sampled perturbations to ensemble members to simulate model error (as in Zeng et al. 2019 using truncation error or Mitchell and Houtekamer 2000 using model error). Here the added perturbations are based on an estimate of the background covariance B , so draw the analysis variance towards the estimated variance. X_a (when applied after assimilation) or X_b (when applied before assimilation) is replaced with:

$$X_a + f_{\text{add}} B^{\frac{1}{2}} S,$$

where $B \in \mathbb{R}^{n \times n}$ is a constant in time estimate of the covariance matrix for X (equal to the background error covariance matrix from before), $S \in \mathbb{R}^{n \times N}$ is a randomly sampled matrix whose columns are independently taken from a $\text{Normal}(0, I_n)$ distribution. Algorithms take f_{add} as a scalar parameter chosen prior to running the model.

Relaxation to prior perturbations

Relaxation to prior perturbations consists of replacing the analysis perturbations with a convex combination of the analysis perturbations and the background perturbations (Zhang, Snyder, and Sun 2004):

$$X_a^{\text{inf}} = \lambda X_a + (1 - \lambda) X_b,$$

for some $\lambda \in [0, 1]$. As assimilation reduces spread, this then increases the spread of the analysis.

Order of inflation methods in cycled experiment

The application of inflation and calculation of model outputs was carried out in the following order:

1. Calculation of the analysis vector x_a using the uninflated W_a . As this comes first, all localisation was posterior;

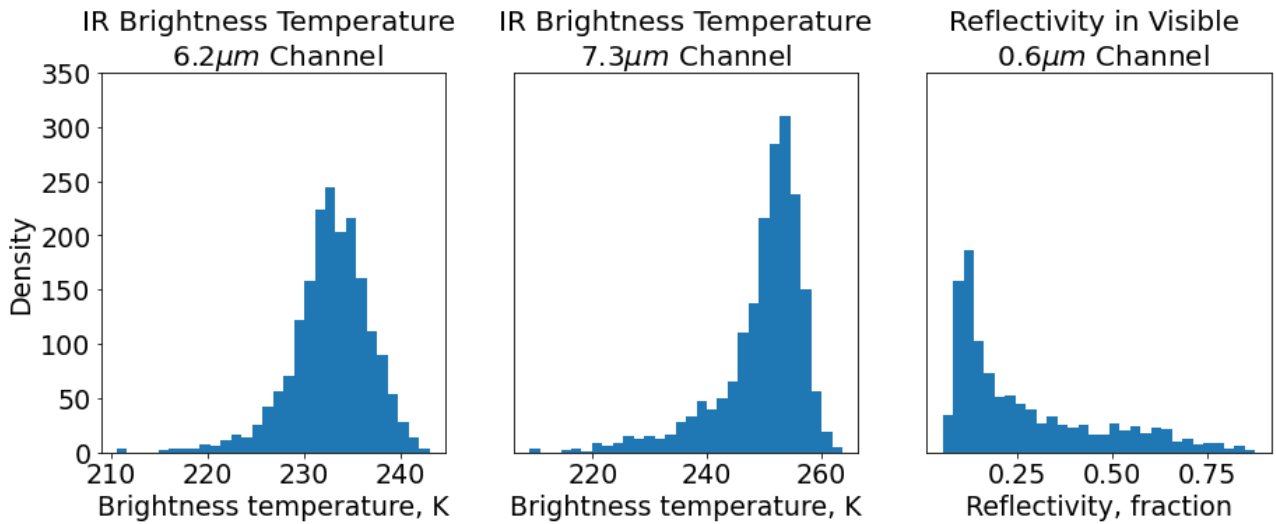


Figure 3.1: Plot of the density of infrared and visible observations at the times and horizontal locations where verifying radiosondes were observed. In general, high infrared brightness temperatures imply the presence of cloud, while lower visible observations of reflectivity suggest the presence of clouds. Note that reflectivity observations are only available in the daytime, so are considerably less.

2. Calculation of the analysis equivalent perturbation matrix $Y_a = Y_b W_a$ using the uninflated W_a ;
3. (Adaptive) multiplicative inflation applied to W_a ;
4. Calculation of the analysis perturbations as $X_a = X_b W_a$;
5. Inflation of the analysis perturbations with relaxation to prior perturbations and additive inflation.

In particular, observation equivalent analysis ensembles Y_a were not affected by inflation, whereas analysis ensembles X_a were, and contained errors relating to inflation.

3.1.2 Data analysed

Radiosondes were not assimilated and used as verification observations. Background and analysis vectors and increments were stored at 3 hour intervals, and only radiosonde observations stored at analysis times were used. On average, 28.9 ± 1.3 radiosondes were released around midnight, 6.6 ± 2 at 6am, 24.7 ± 1.5 at midday and 4.9 ± 2 at 3pm. All radiosondes at these times were compared to PAI vectors for the observations in the infrared channels, but the radiosondes at midnight were not compared to the PAI vectors for the visible channel as these were reflectivity observations which rely on sunlight.

The radiosonde locations showed a broad range of radiance values as seen in the histogram in figure 3.1. Note that the smaller counts for the visible channel represent the lack of information at midnight.

As radiances are most strongly affected by clouds compared to other model variables, and clouds are a large source of uncertainty and model error in forecasts (Geiss et al. 2021), assimilation of radiances has the most benefit when either ‘creating’ or ‘destroying’ clouds - that is when the analysis contains a cloud but the background doesn’t, or vice-versa. In our data set, there were not many verification times and locations where this occurred for the clouds observed by the visible channel, as seen in figure 3.2 - when the first guess model equivalent for visible radiances is above or below the threshold 0.4, the analysis model equivalent stays on the same side of the threshold. Although this suggests that the model is predicting clouds well - many clouds being created or destroyed would suggest a discrepancy between the model and observations - it also means that we do not have much data on the most influential assimilation events.

The infrared and visible channels used by the DWD are sensitive to clouds, so can be used to indicate whether

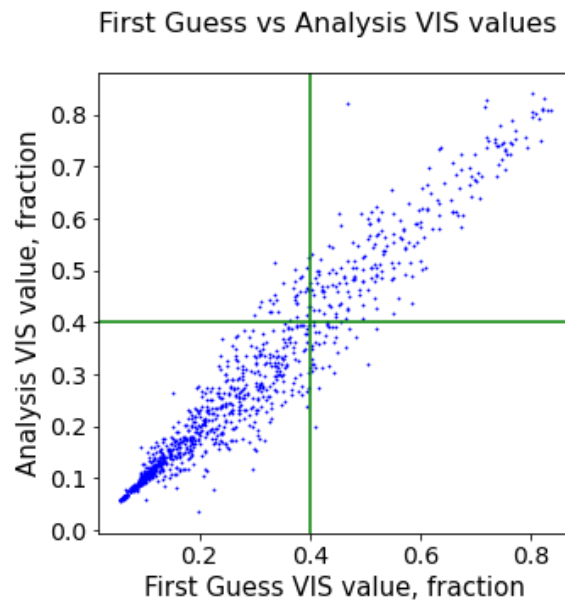
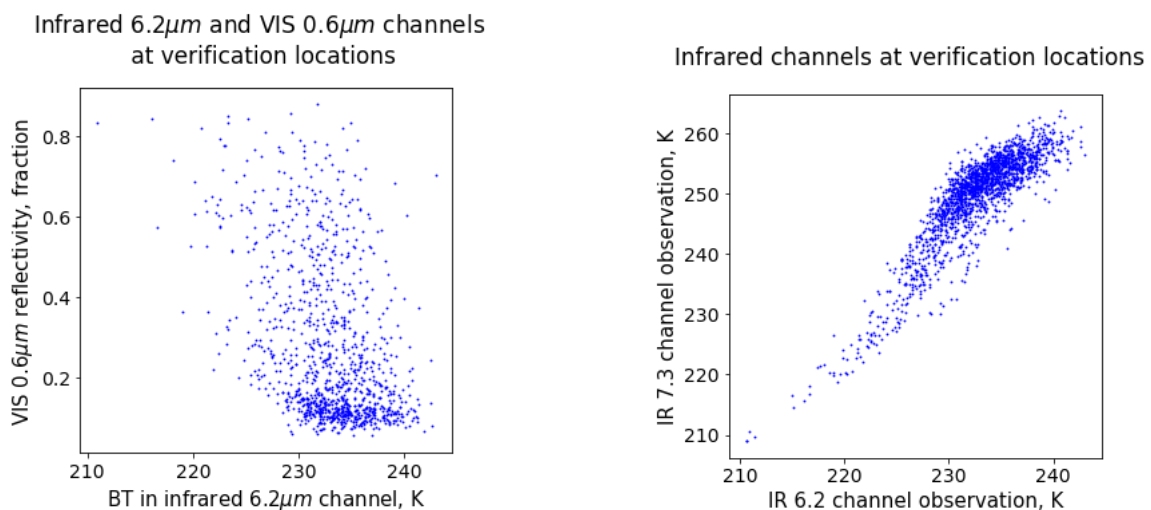


Figure 3.2: Plot of first guess versus analysis model equivalents for visible radiances. The top left and bottom right quadrants represent 'cloud created' and 'cloud destroyed' points respectively.

a certain type of cloud is likely to be present at the radiance location. As mentioned earlier, higher visible reflectance observations and lower infrared brightness temperature observations suggest the presence of clouds. Figure 3.3 shows plots of observed radiances in the different channels against each other. The infrared and visible channels show little obvious correlation: there are many radiosonde locations where the visible observation suggests a cloud but the infrared observation does not and vice versa. This does not mean that one of the observations is wrong, but that the two observations indicate the presence of a different kind of cloud: the visible channel is more sensitive to low and middle level water clouds and not very sensitive to high level ice clouds, whereas both infrared channels sense middle and high level clouds and the $6.2\ \mu\text{m}$ channel is not sensitive to lower level clouds (below 600 hPa).

Figure 3.3 should be kept in mind when partitioning the data depending on radiance cloud definitions - defining clouds with the visible channel is very different to with the infrared channels.



(a) Observed radiances in the visible channel against observed radiances in the first infrared channel

(b) Observed radiances in the second infrared channel against observed radiances in the first infrared channel.

Figure 3.3: Observed radiance values in each of the channels compared to each other. The visible channel gives a very different definition of cloud to the infrared values.

Chapter 4

Illustrative example

4.1 Using PAI to optimise localisation - example at one location

To better understand how the PAI method can be used to optimise localisation globally, we consider now a single time and horizontal location and use PAI to understand retrospectively what the best localisation conditions would have been at this individual point. We consider the vertical profile of the forecast and analysis at the location of a single Radiosonde released on June 23rd at 11am in the south-west of Thuringia, Germany.

4.1.1 Radiance assimilation at the test time and location

Before calculating any PAI vectors, we can use standard outputs of the assimilation system to understand the weather situation and the effect of radiance assimilation at the time and location. Figure 4.1 shows radiance observations in the visible and infrared channels. Areas with low infrared radiances indicate the presence of clouds: when clouds are present, much of the outgoing radiation at that frequency is absorbed and emitted at lower temperatures. Areas with high reflectivity indicate the presence of clouds, as clouds reflect a lot of incoming solar radiation in that frequency. Looking at these raw observations, we see that there were areas with clouds detected by both channels near the horizontal location of the radiosonde. Note also that figure 4.1 shows how the different satellite channels give complementary information. Over more northern parts of Germany, particularly in Saxony-Anhalt, we see how the visible channel has greater detail. The infrared channels contain details of other model variables not present in the plot.

The clouds seen in the raw observations were not correctly predicted in the initial forecast (before assimilation). Looking at the first guess departures (figure 4.2 top two panels, raw observations y_o subtract the model equivalent $h(x_b)$ for the forecast), we see that near the radiosonde (labelled 21) there were areas where the forecast both under- and over-estimated the satellite values. The values of these differences are quite high compared to other days in the cycled experiment, so the satellites have the potential to make a big improvement to the analysis here. In the data assimilation process, the first guess departures influence the analysis values through multiplication with the Kalman gain - larger first guess departures lead to a larger influence of an observation type.

In general assimilation seemed to reduce forecast errors. The differences between the analysis and background model equivalents (figure 4.4) show a similar contour to the first guess departures, suggesting that assimilation brings the analysis model equivalents closer to the observed values as expected. This happens by adjusting model variables, which in turn influence the analysis model equivalents.

The change in observation model equivalents is not the only effect of assimilation. Other variables that are related to radiances are also adjusted, either because they contribute to the observation operator (for example

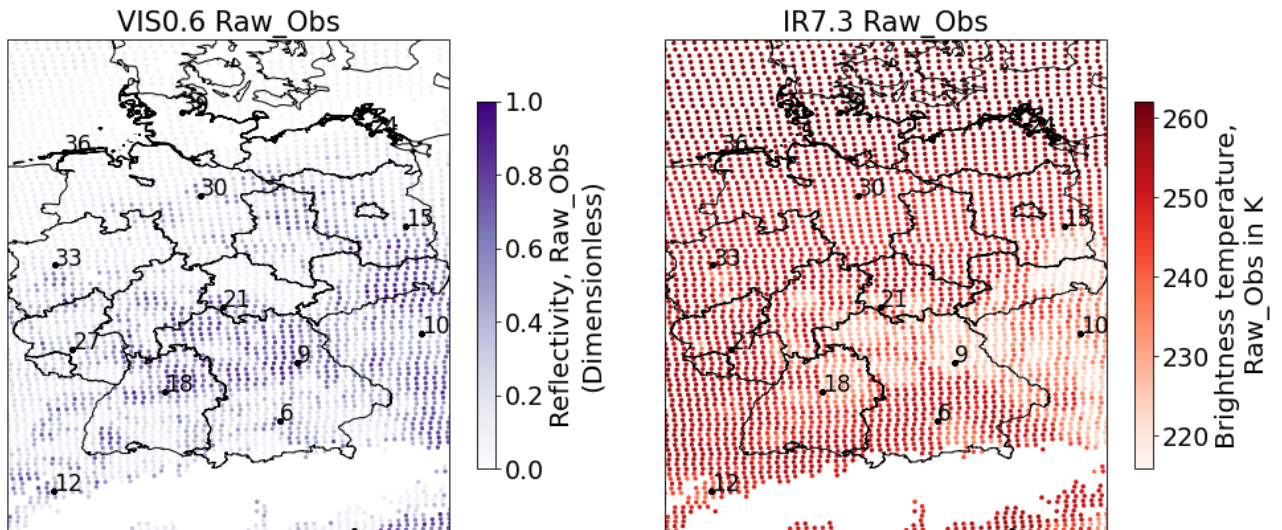


Figure 4.1: Observations

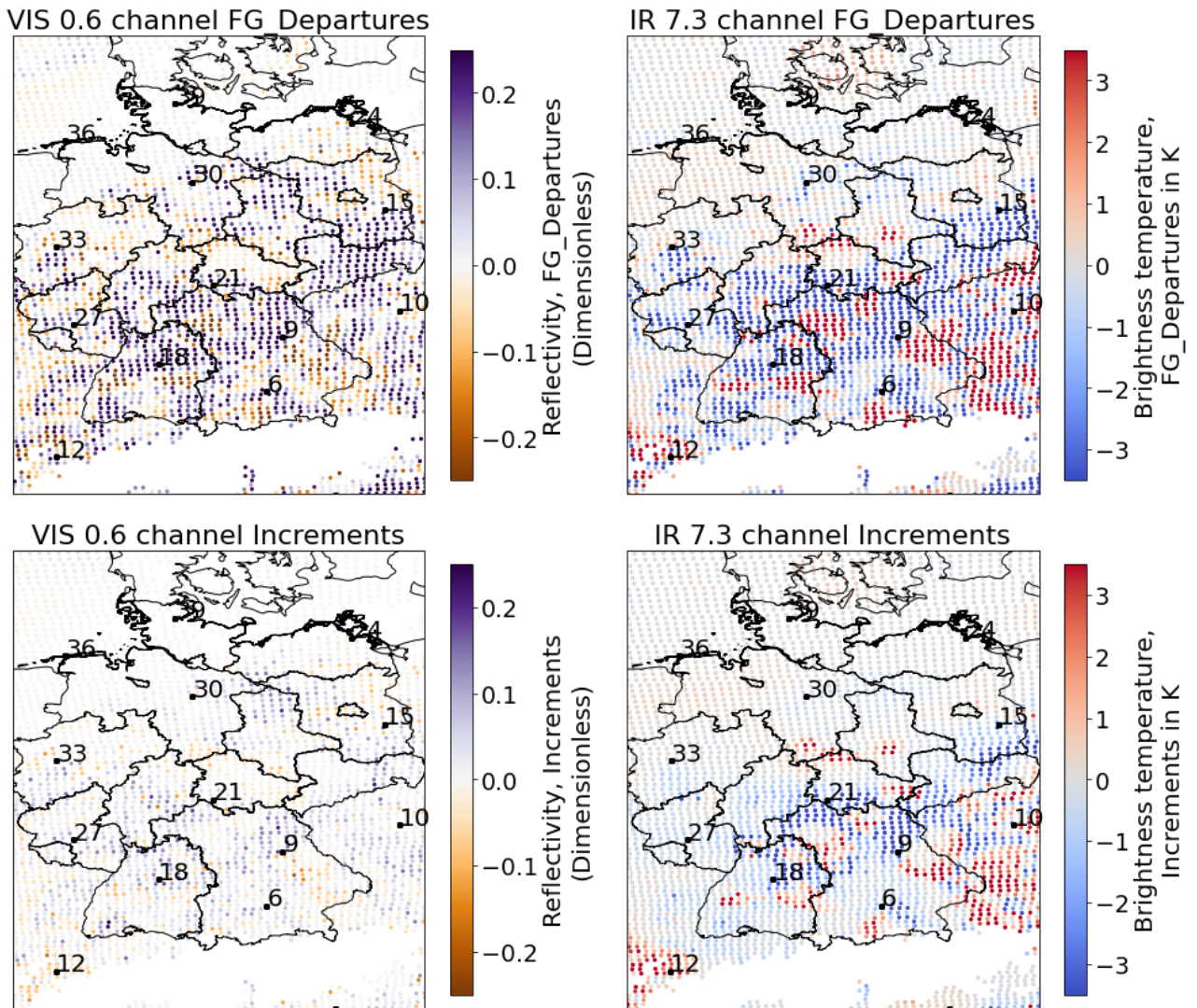


Figure 4.2: Observation – background equivalent and analysis increments. Analysis increments should correct the first guess departure (as seen by the similar contours).

water-related variables (hydrometeors) for all three channels and temperature for the infrared channels), or because they are related through covariances (for example wind speed).

4.1.2 Using PAI to understand assimilation

To understand better how assimilation of radiances affected the analysis at the test location, we can use the PAI approximation to give a diagnostic for the contribution of satellites to the analysis vector at the time and horizontal location of the verifying radiosonde. We consider now just the visible $0.6\mu m$ channel. Consider the vertical plot of the relative humidity equivalents for the background (yellow), analysis (green) and analysis with visible channel observations removed by PAI (red) compared to observed relative humidity at the radiosonde (blue) in figure 4.3. We see that there is very high relative humidity in both the model and observations at around 800hPa, suggesting that a cloud has been correctly predicted here. Further up the atmosphere, the relative humidity is less well predicted - between 300 and 400 hPa relative humidity is too high in the model and between 550 and 600 hPa too low.

In this case, assimilation (and in particular assimilation of visible radiances) does not always bring the model closer to the verifying observations. Between 300 and 450 hPa the background is closer to the radiosonde values than the full analysis. Close to 500 hPa, the analysis is closer to the radiosonde than the background, but the partial analysis increment for satellites brings the analysis further away from the radiosonde - it is not 'beneficial' here. This does not necessarily suggest a flaw in the data assimilation system, as random variation is a natural part of assimilation. We have to look at larger averages to be able to comment on trends in the assimilation system.

Note also that at this time and location, the contribution of visible radiances is relatively small (figure 4.3). This could be because other observations account for the information contained in the visible radiance observations, or because at this point in time and space visible radiances are not strongly correlated with relative humidity.

We can localise vertically to try and dampen the influence of radiance observations where they are not beneficial. Figure 4.4 shows the PAI vectors for temperature and relative humidity $PAI^{VIS \rightarrow t}$ and $PAI^{VIS \rightarrow r}$ with and without retrospective localisation for a Gaspari-Cohn function in log pressure coordinates centered at 600 hPa and with width $0.06 \ln(hPa)$. The relative humidity PAI vector corresponds to the difference between the analysis with and without satellites 'removed' in figure 4.3. The plot illustrates nicely how localisation damps the influence of an observation to 0 at locations far away from the observation.

It is worth drawing attention here to the approximation we are making when 'removing' an observation with PAI. When we subtract PAI from an analysis (subtracting vectors like those in figure 4.4 from full analysis vectors like those in figure 4.3), we are not expressing the exact analysis when calculated without the observation, and the error can be understood through the analysis in section 2.3.2.

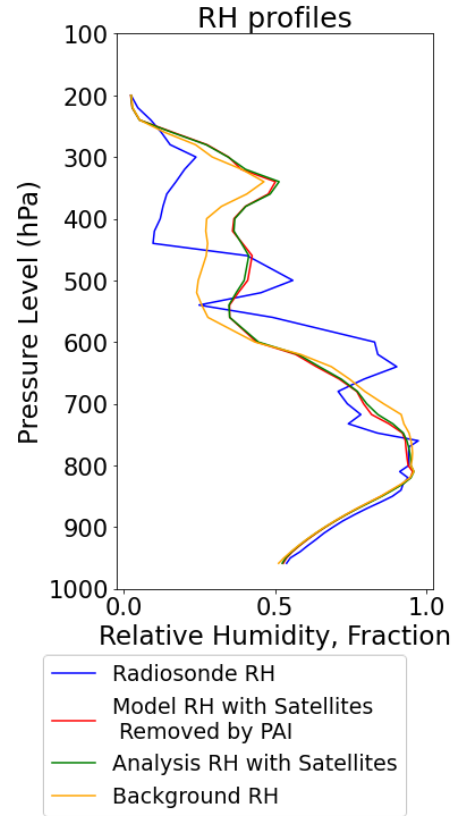


Figure 4.3: Vertical plot of relative humidity for analysis, approximate analysis without VIS, background and observations at the radiosonde location.

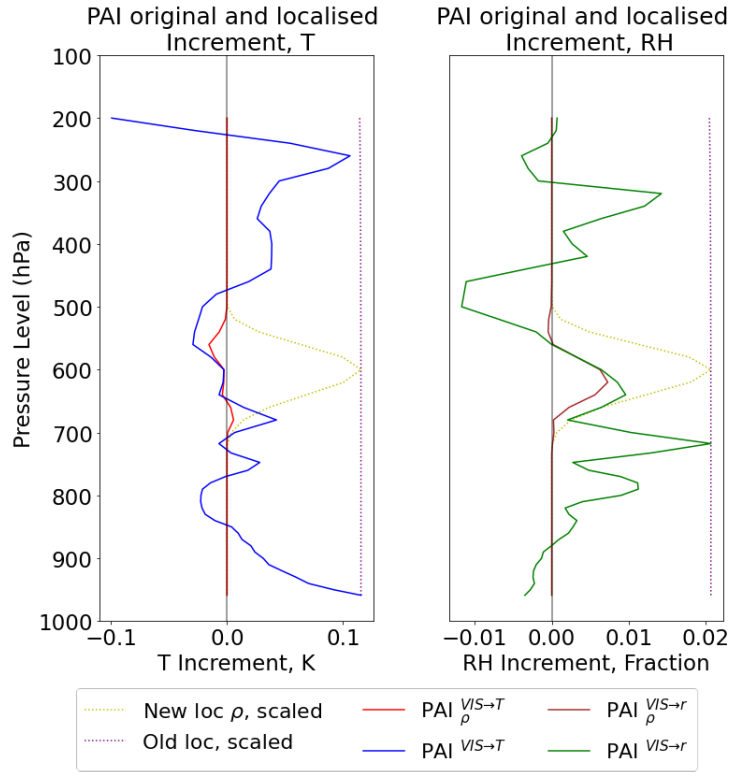


Figure 4.4: Increments with and without retrospective localisation for temperature and relative humidity.

To see more clearly where the observation was beneficial, we can plot vertically the temperature and relative humidity benefit:

$$b(x_a) = |e_{\text{RASO}}(x_a - \text{PAI}^y)| - |e_{\text{RASO}}(x_a)|,$$

(metrics and notation section 2.4) as in figure 4.5. This in general looks like the difference between the analysis and the analysis with visible radiances removed, except that the sign is adapted depending on where the observation values lie in comparison to the analysis, and is sometimes smaller when the analysis is very close to the observations. Positive values of benefit suggest that the observation reduces error at the location: the error is bigger for the approximate analysis without the observation than the error for the analysis with the observation included.

First consider the benefit for the raw PAI vector that we remove to approximate the analysis without visible radiances (labelled $|\text{Error no obs}| - |\text{Error old loc}|$). For temperature, the benefit is varied throughout the atmosphere - below 600 hPa it fluctuates a lot, and it is not clear whether radiance observations are beneficial. Between around 320 and 600 hPa, the benefit is in general positive, suggesting that the observation is useful here. So to reduce temperature error, we might aim to include the observation between 320 and 600 hPa (so that we include these beneficial parts of the increment), and remove it elsewhere. For relative humidity, the benefit varies more dramatically - at around 600 hPa the observations are beneficial, but they are detrimental for example just below 700 hPa. We might suggest removing a lot of the increment and only keeping the increment very close to 600 hPa to reduce relative humidity error.

When retrospectively localising, we set to 0 the increments away from the nominal height. This is the same approximation as used for the analysis with observations removed, so the benefit is reduced to 0. We see that localisation removes many non-beneficial increments for relative humidity, but also some beneficial increments for temperature higher up in the atmosphere (figure 4.5). It could be that this choice of localisation is too stringent, removing many potentially beneficial increments. To be able to say this quantitatively we would need to consider a cost function that combines the error in temperature and relative humidity at various height levels, which we consider later in the thesis. It makes physical sense to localise in the middle of the atmosphere: we

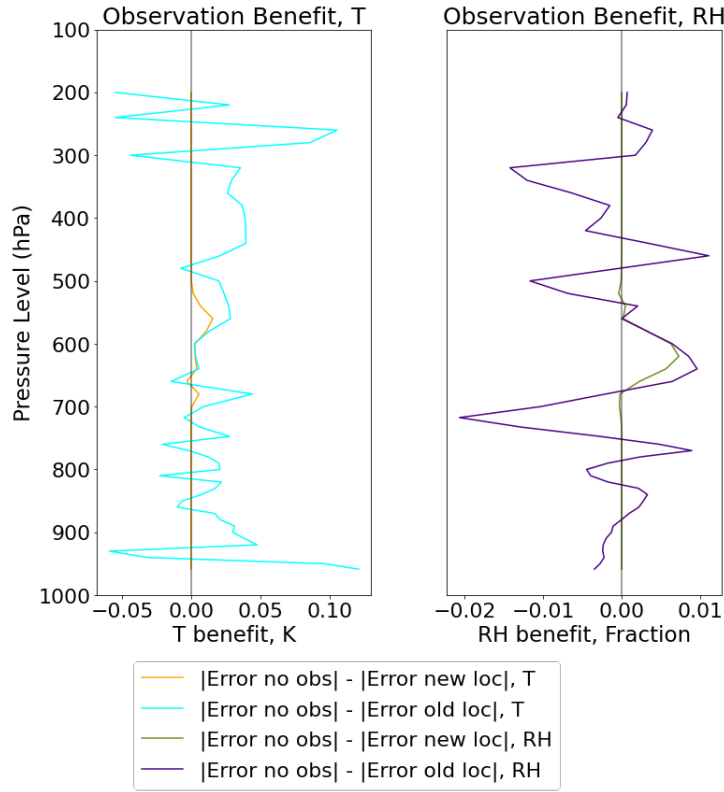


Figure 4.5: Benefit when comparing to temperature and relative humidity observations, with and without retrospective localisation. Localisation removes some non-beneficial relative humidity increments, but also some beneficial temperature increments.

are localising close to lower clouds for which radiances in the visible spectrum are considered beneficial.

Note that data assimilation is based on larger statistical effects - just because the PAI vectors are not in general beneficial at this location, it could be that they are beneficial when larger averages are taken.

4.1.3 Optimising the localisation parameters

To understand how the optimisation of localisation parameters will work later, we can calculate J for a range of nominal heights and widths for the Gaspari Cohn vertical localisation function, using just the radiosonde at this location and restricting either to temperature or relative humidity observations. We can then compare this to the approximate error when the observation is removed from the whole column (defined as $J(0)$ in the metrics section 2.4) to see if our observation was 'beneficial'. We combine this using the percentage change in cost from section 2.4. Recall that this is given as:

$$\Delta J(\rho)\% = \frac{J(\rho) - J(0)}{J(0)} \times 100\%. \quad (4.1)$$

The smaller (negative) values of $\Delta J(\rho)\%$, the better the localisation function has performed under our metric: we favour localisation functions ρ where the approximate error to radiosondes for ρ , $J(\rho)$, is smaller than the approximate error to radiosondes when the observation is removed, $J(0)$.

From the plot of benefit in figure 4.5 we saw that including observations between 320 and 600 hPa are beneficial for temperature and at other levels the benefit is ambiguous, so a localisation function with a reasonably wide width (for example $1.0 \log(\text{hPa})$) centered inbetween 320 and 600 hPa (say at 450 hPa) should have the lowest values of $\Delta J(\rho)\%$. For relative humidity, similar arguments suggest that localising with a thin width (say $0.1 \log(\text{hPa})$) at around 600 hPa should have the lowest values of $\Delta J(\rho)\%$.

Figure 4.6 shows the errors obtained plotted by colour against the width and height parameters used in the approximation. A range of 18 evenly spaced heights was used and 27 widths with higher density for lower width values. Due to the statistical nature of data assimilation, we would not expect to see any systematic effects when just considering one time and location. The optimisation should just reflect the benefit plotted in figure 4.5.

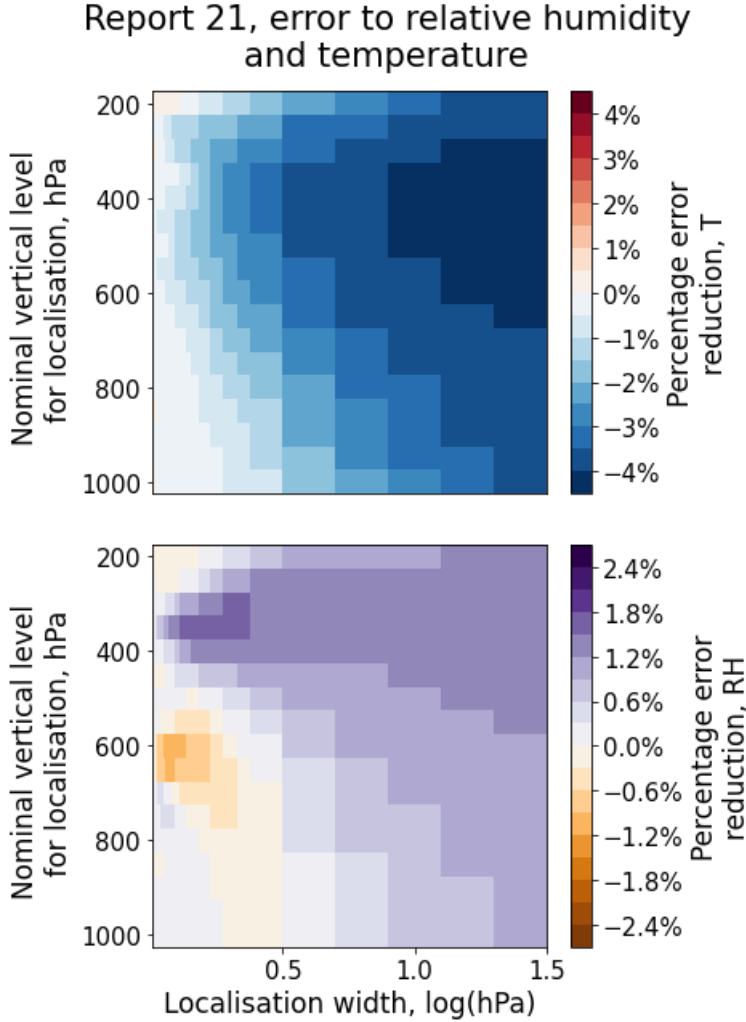


Figure 4.6: Plot of the errors to temperature and relative humidity when using different Gaspari-Cohn localisation parameters. The colour shows the cost.

The minimum error when measured against relative humidity is at a width of approximately 0.2 and a height of 600 hPa (figure 4.6). Returning to the plot of benefit (figure 4.5), we see that this makes sense - most of the beneficial increments were near 600 hPa, and the rest were non-beneficial, so restricting to increments around this area should decrease the error. For smaller localisation lengths, the error is much greater when we choose a nominal height higher in the atmosphere, say at 350 hPa, which is also in line with our intuition: we would in this case be restricting to the non-beneficial increments. For wider localisation widths it is harder to relate the plot of benefit to the plot of the cost function, as we are then averaging over more increments, so the weight at each height level plays a greater role.

Looking at the plot for percentage change in error calculated against temperature, we see that error is the least for wide localisation widths. This reflects how the temperature benefit is generally positive in the upper regions of the atmosphere, and more ambiguous in the lower regions. Using the cost function now gives us a measure of whether the observation is

overall beneficial, which was not possible in lower regions of the atmosphere using just the benefit. Again we can see how the smaller localisation widths isolate areas that are beneficial/detrimental - if we choose a very small width, say $0.02 \ln(\text{hPa})$, then we only see a benefit when we localise with a nominal height at around 400 - 450 hPa or around 550 hPa. This makes sense with the plot of benefit, as all other areas of the atmosphere have detrimental increments.

Later in the thesis we will combine the temperature and relative humidity error for all verification locations. The optimal value will be optimal in the sense that it minimises error for the whole month, at all locations, and considering both humidity and temperature.

Chapter 5

Results

5.1 Behaviour of PAI vectors in whole sample

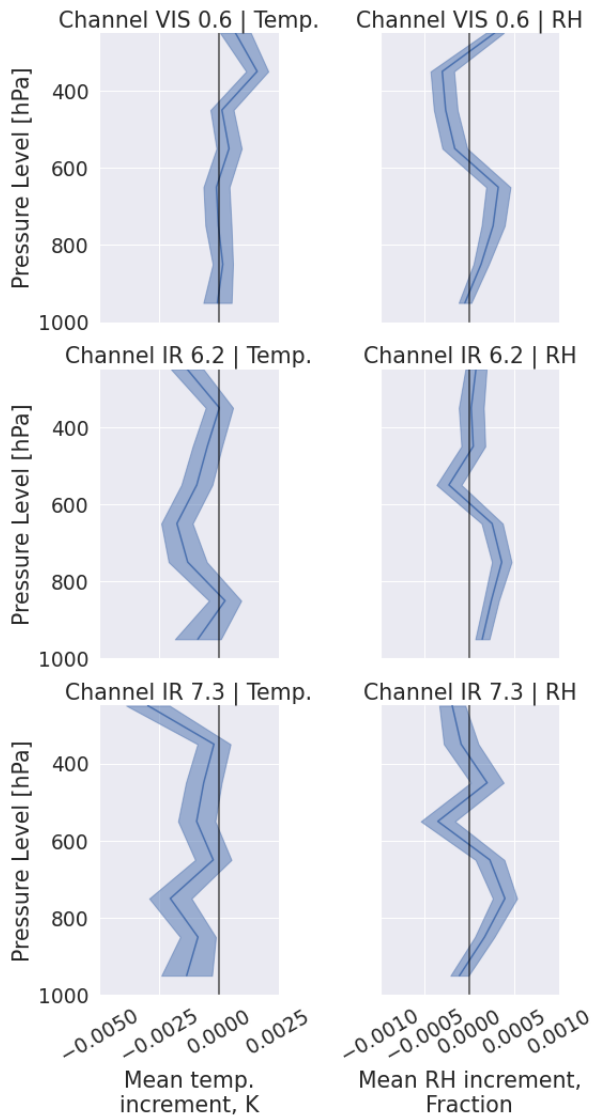


Figure 5.1: Binned averages of the increments over all the radiosonde locations. Sometimes increments were systematic, and sometimes not.

Before optimising localisation functions, we look at the behaviour of PAI vectors in the whole sample to identify systematic trends and obtain a first indication of where observations might be beneficial.

As we saw in section 4.1, vertical plots of PAI vectors give an indication of how an observation influences analyses in different ways at different vertical levels, and how beneficial these influences are. We consider now the behaviour of increments in the whole sample by taking averages over all locations where verifying radiosondes were available. The connection to optimising localisation is in this case less direct, but the vectors do give valuable information on the systematic effects of assimilation.

5.1.1 Averages of increments in the whole sample

As an indicator of how each channel affected the different variables in the whole sample, we can look at the averages of the PAI vectors over all the verification locations. Averages of PAI vectors were often non-zero, but it was difficult to attribute these effects to a specific tendency in the data assimilation system.

Figure 5.1 shows the average value of the increments over the set of verifying observations. Vertical averaging was carried out by taking the PAI vector restricted to the radiosonde location (a vertical vector of length 65 representing the temperature and humidity at the horizontal location of the radiosonde),

interpolating the increment to the vertical locations of the radiosonde, and then sorting the increment values into bins depending on height. Bins were chosen of equal size between pressure levels 100, 200, ..., 1000 hPa. Averages and standard deviations were calculated over each of the bins.

Although it may seem reasonable to expect that PAI increments have an average of zero, in most cases there were systematic tendencies. Only temperature increments below 400 hPa for the visible channel (figure 5.1, top left panel), had an average that was clearly close to 0. For the infrared 7.3 μm channel temperature increments (figure 5.1, bottom left panel), the increments below 300 hPa took values consistently a little less than 0. This could be indicative of a bias in the PAI calculation, for example relating to inflation. Below 600 hPa relative humidity increments were on average positive for all channels (figure 5.1, three panels on the right). This could be related to corrections made for clouds below the 600 hPa level. There were occasional spikes in the average increment, for example for temperature the average increment dips between 600 and 700 hPa in the infrared 6.2 μm channel (figure 5.1, middle left panel), and between 700 and 800 hPa in the infrared 7.2 μm channel (figure 5.1, bottom left panel). For relative humidity, the average increment dips between 500 and 600 hPa for both infrared channels (figure 5.1, middle and bottom right panels).

If increments are systematically non-zero at a given location, it could reflect a number of possible aspects of the data assimilation system:

- Model bias: if the forecast model regularly under-/over-estimates a given variable at a specific location, then the observations will often correct in the same way to counteract this bias;
- Biases in the observation mean - if the observations are consistently too big or small, this will affect the increments multiplying the Kalman gain columns in the PAI vectors;
- Nonlinearity of the observation operator h ;
- Errors in the observation operator h ;
- Failure of linearity assumptions such as (2.36);
- Non-Gaussian errors that violate (2.1) and (2.2).

5.1.2 Averages of benefit in the whole sample

Recall that we defined benefit in section 2.4 as:

$$b(x_a) = |e_{\text{RASO}}(x_a - \text{PAI}^y)| - |e_{\text{RASO}}(x_a)|.$$

Here we leave the benefit b as a vector. The analysis is taken to be the vector with the observation included, either approximated through the PAI method or through the initial assimilation run. The analysis with the observation removed is calculated by subtracting the PAI vector as in the definition.

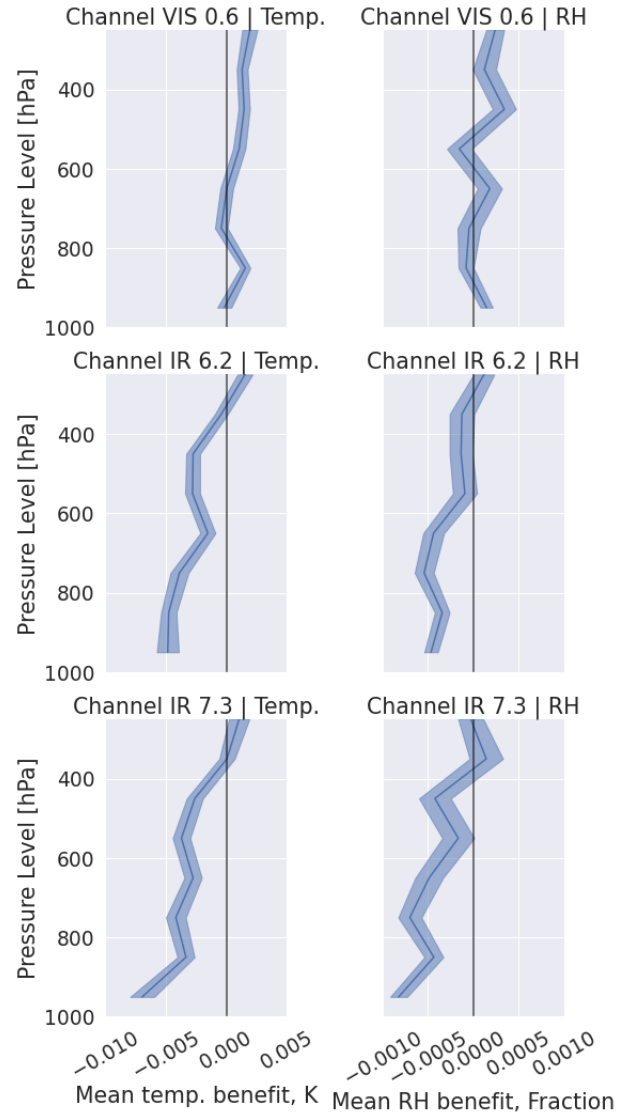


Figure 5.2: Binned averages of the increments over all the radiosonde locations. Sometimes increments were systematic, and sometimes not.

By looking at binned averages of the benefit over the whole sample (as with the increments), an indication was given of where in the atmosphere the observations were in general beneficial, and looking at standard deviations of these averages gave a first indicator of how certain we are of benefit throughout the atmosphere. Clear variation in benefit was seen between different levels of the atmosphere in all channels. As with the increments, the benefit was calculated for each observation for each radiosonde separately, and these were separated into bins of width 100 hPa where averages and standard deviations were taken.

For the visible channel, the average benefit was generally either positive or at least close to 0 (figure 5.2, top panels), whereas for the infrared channels the average benefit was generally positive above 300 hPa, and otherwise negative (figure 5.2, bottom four panels). The negative values of the infrared average benefit could be indicative of a systematic bias in PAI vectors, either through a bias in the calculation of PAI vectors induced for example by inflation, or through a bias in the forecast and assimilation system reflected by the PAI vectors, such as a flaw in the model operator $\mathcal{M}$.

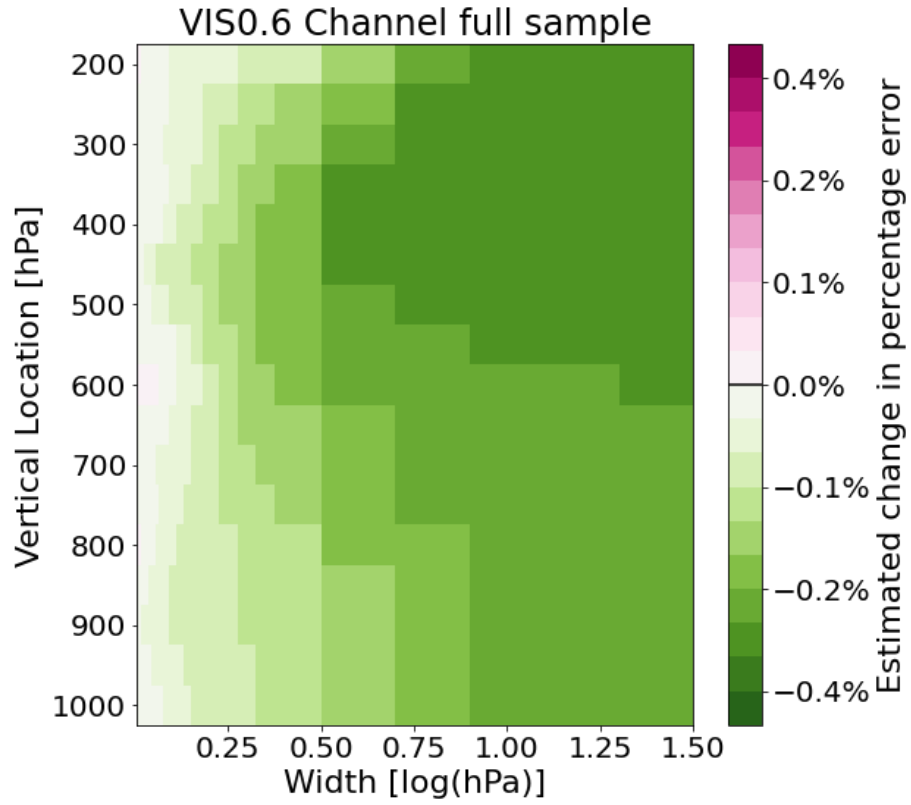
Some of the structures in the benefit functions were visible also in systematic trends in PAI vectors, but not all. For the temperature benefit in the $6.2\ \mu\text{m}$ channel (figure 5.2, middle left panel), the small increase in benefit between 600 and 700 hPa aligns with the dip in the average PAI vector (figure 5.1, middle left panel), suggesting that the systematic trend in partial analysis increment could be having a beneficial effect. Similarly, for the $6.2\ \mu\text{m}$ channel the drop in benefit for relative humidity below 600 hPa (figure 5.2, middle right panel) seems to match the increase in average relative humidity increment below 600 hPa (figure 5.1, middle right panel), suggesting that perhaps the observation operator has some issues relating to clouds below this level. For the infrared $7.3\ \mu\text{m}$ channel, there is only a slight link between the benefit and increment plots for temperature (figure 5.2, bottom two panels), but the on average negative values of temperature increment between 500 and 600 hPa seem to be beneficial (figure 5.2, bottom left panel and figure 5.1, bottom left panel).

The benefit differs from our other cost function (2.117) in that it takes the absolute value of the difference between the analysis and verifying observation, rather than the square. This means that for the cost function (2.117), increments are rewarded more for moving towards the verifying observations when the analysis before PAI is further away from the observations. We will see later that this means that some structures in the error are not noticeable when looking just at the benefit.

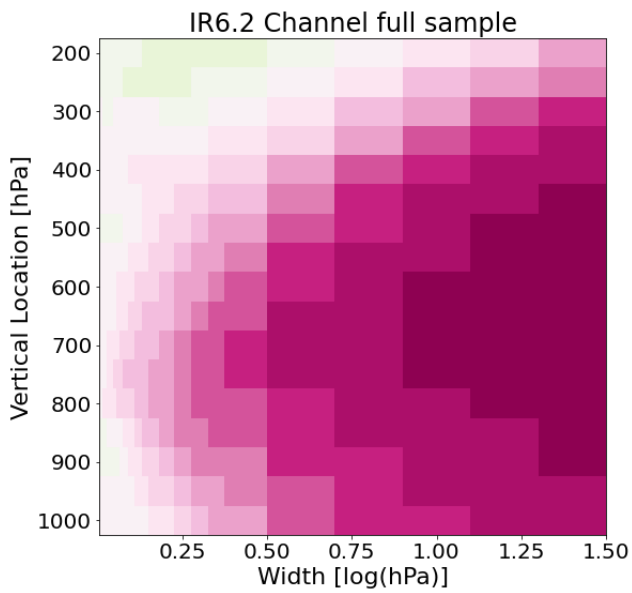
5.2 Optimisation of Gaspari-Cohn localisation function parameters

Having given an introduction to our methods and data set, we are now ready to optimise the vertical localisation of satellites. We consider first optimal parameters for a Gaspari-Cohn localisation function, first looking for optimal parameters for all situations, then considering the reliability of results and simple adaptive localisation schemes where different parameters are used depending on observation values.

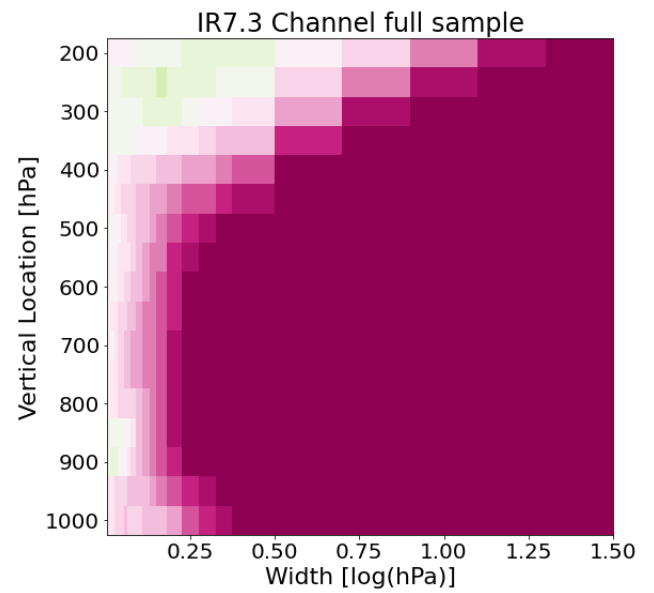
The PAI method suggested that for the visible channel it was in general best not to localise vertically (figure 5.3a), although in cloudy sky it could be beneficial to localise lower down in the atmosphere around 850 hPa (figure 5.7). For the infrared channels, a nominal height around 250-300 hPa and a width of $0.2\ \ln(\text{hPa})$ was preferred (figures 5.3b and 5.3c). Infrared observations were on average beneficial higher up in the atmosphere (figures 5.3b and 5.3c); just above the surface for the $7.3\ \mu\text{m}$ channel when looking at error to temperature (figure 5.6), and were detrimental elsewhere. As such wide localisation widths, in particular at heights lower down in the atmosphere, were not favoured for the infrared channels. Some structures from the plots of benefit (figure 5.2) were visible in the Gaspari-Cohn optimisation. There was not enough evidence to state exact optimal parameters, but broad qualitative observations were possible, particularly about the magnitude of the benefit. Random subsampling suggested that there was not always enough evidence to eliminate the possibility that the cost function was 0 (figure 5.4).



(a) Visible channel



(b) Infrared $6.2\mu m$ channel



(c) Infrared $7.3\mu m$ channel

Figure 5.3: Plot of the cost function for varying Gaspari-Cohn parameters, height (y-axis) and width (x-axis). Vertical locations with higher values suggest that observations are less beneficial/more detrimental around this point in the atmosphere, taking wider averages depending on the chosen width.

5.2.1 Global optimisation of Gaspari-Cohn parameters

For each channel, the cost function for error to unassimilated radiosondes (2.117) was evaluated for localisation functions taking each of 17 choices of nominal height between 200 and 1000 hPa (equally spaced), for 27 widths between 0.01 and 2.5 (more dense for lower values), with the aim of obtaining a broad idea of how different localisation functions may perform and making a suggestion for a global optimum. Recall that the cost function is an approximation (section 2.3.2) of error to radiosondes when the chosen localisation function is used, so also includes radiosonde error as discussed in section 2.5.

As a benchmark, the approximate cost for the analysis with the channel removed was also calculated ($J(0)$ in the notation of section 2.4), and used to normalise the values as:

$$\Delta J(\rho)\% = \frac{J(\rho) - J(0)}{J(0)} \times 100\%.$$

If $J(0)$ (equation (2.119)) is greater than $J(\rho)$ (equation (2.117)), we have that $\Delta J(\rho)\%$ is negative and we say that the localisation function is *beneficial*. We say that the localisation is *detrimental* otherwise. Considering $\Delta J(\rho)\%$ is important as it allows us to compare results for the visible channel with those for the infrared channels, which were calculated on a larger data set (including night time).

The analysis is similar to the discussion at the single location in section 4.1.3, except now we look for global optimal parameters by incorporating all radiosondes into the cost function, and we look at the combined cost function for temperature and relative humidity radiosonde observation by taking a sum. We consider separately each channel used by the German Weather Service.

The calculated values of $\Delta J(\rho)\%$ are plotted in figure 5.3 by colour against the width and height chosen.

We first make some more general statements about the choice of parameters. As in 4.1.3, looking at the cost for smaller localisation widths shows us where in the atmosphere observations were in our case beneficial: when $\Delta J(\rho)\%$ is negative for a small localisation width at a given level, we say that the observation is beneficial at this level. Using similar reasoning, the cost function for wider localisation widths shows where observations were in our case beneficial when averaging over nearby vertical levels.

The visible channel was in our case beneficial throughout the atmosphere and detrimental only very locally around 600 hPa (figure 5.3a, detriment at 600hPa very faint in figure). To understand how we can make such inferences from the plot, consider a localisation width just smaller than 0.1 ln(hPa). For this choice of width, the lowest error values are obtained when localising lower down in the atmosphere at around 850 hPa or higher up in the atmosphere around 450 hPa, and high error values are obtained when localising in the middle of the atmosphere near 600 hPa. The detrimental effect around 600 hPa is only seen for smaller choices of width, suggesting that the beneficial effect of the observation at other vertical heights dominates when wider widths are chosen. For wider localisation widths, there is lower error higher up in the atmosphere, suggesting that the benefit is of greater magnitude at higher vertical levels. This could also be related to the density of observations at different levels, however. As the visible 0.6 μm channel is a window channel, it is in line with our intuition that it is beneficial in many vertical levels in the atmosphere, and the detrimental levels probably reflect issues in the data assimilation or modelling system (for example model representation of clouds as in Geiss et al. 2021) rather than low correlations between the model levels and visible radiances.

For the infrared channels (figures 5.3b and 5.3c), looking at smaller localisation widths shows us that radiances were only beneficial higher up in the atmosphere. There is some very local benefit, for example around 850 hPa and 500 hPa for the 6.2 μm channel, but this could be an effect of random sampling. For wider localisation widths, the most detrimental choice of nominal height was around 800 hPa, as this is the height that included most of the detrimental observations lower down in the atmosphere.

As the sensitivity functions for the 6.2 and 7.3 μm channels peak at 350 and 450 hPa, we would expect

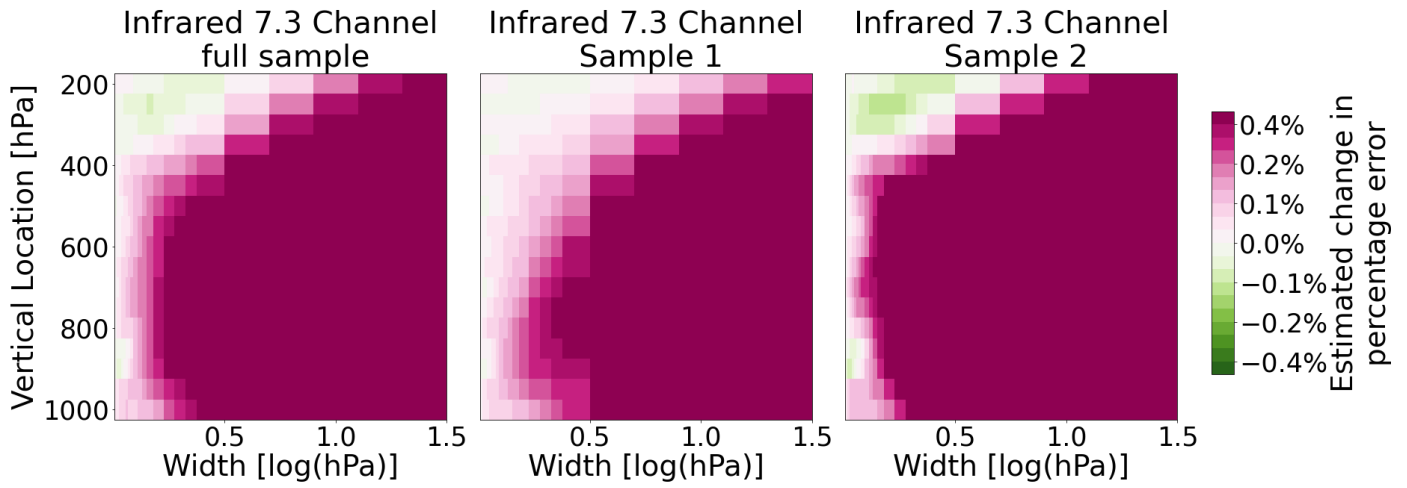


Figure 5.4: Plot of the cost function for varying Gaspari-Cohn parameters for the $7.3\mu\text{m}$ channel, height (y-axis) and width (x-axis) with random subsampling to show reliability of results.

localising around these levels to be beneficial, but this is in our plots not the case. This could be related to the fact that we have also assimilated the visible channel at this level, but not considered correlations between the channels, which could lead to detrimental increments.

5.2.2 Reliability of results - subsampling

By considering just the raw cost function, we have no measure of how sensitive the results were to random noise - it is not possible with just the raw cost function to eliminate the hypothesis that the value of the cost function has expectation 0 for a given set of parameters. By considering subsamples it was observed that this could indeed be the case for some parameter values, but for other parameter values the conclusions were more reliable. More precise methods of understanding the reliability of results are required, such as the confidence intervals for empirical localisation functions in section 5.3.

As an indicator of reliability, the cost function for each of the chosen parameters was calculated with random subsamples of the radiosonde locations. For each radiosonde, the observations were included into either of the subsamples with probability $\frac{1}{2}$. For each subsample, the sum of the costs for all locations was calculated and plotted as before as the percentage change in cost $\Delta J(\rho)\%$ (equation (2.118)). If the cost function for each of the sub-samples was very different, this is an indicator that the results were not reliable.

Figure 5.4 shows an example of this process for the infrared $7.2\mu\text{m}$ channel. In the first plot on the far left is the value of the cost function for the full sample, and the two plots on the right are the cost function for the random subsamples. In both subsamples there is a benefit when localising both higher up in the atmosphere around 250 hPa, but the magnitude of the benefit varies greatly between the subsamples. There is evidence of benefit to localising with a very small width in the first subsample at 500 and 700 hPa (not completely visible in the figure) which is not present in the second subsamples. At these locations we expect that there is insufficient evidence to suggest that the cost function really is non-zero. For this reason we are hesitant to make conclusions about small benefits such as that at 900 hPa, which does appear in both subsamples, but are more confident about the benefit higher up in the atmosphere.

Other examples of this process (not included here) showed similar results - that subsamples showed similar large-scale qualitative results, but varied for more precise conclusions about localisation functions with smaller widths. Methods inspired by standard statistical tests will later be used to understand where our results were more stable. Plotting the cost function itself is still considered useful for understanding the method and describing the magnitude of the benefit at different levels.

5.2.3 Optimising to different variables

As we saw in the plots of benefit in section 5.1.2, the vertical structure of how PAI vectors improve forecasts depends on the variable. When optimising just to one variable (either relative humidity or temperature), it was possible to isolate a few more tendencies.

For the visible channel, the detrimental effect around 600 hPa when optimising globally (figure 5.3a) was mostly attributable to the relative humidity optimisation. When we plot the cost function divided into the cost for temperature and relative humidity (figure 5.5), we see that localising with a thin width at 600 hPa is only detrimental for relative humidity (magnitude not present in the figure but stronger for relative humidity than temperature). This aligns with the plot of relative humidity benefit (figure 5.2, top right panel), where we saw a dip in benefit around this level.

For the infrared channels, restricting optimisation to single variable sometimes produced quite unstable samples, but sometimes showed new trends. We show just the choices where there were observable tendencies, with the caveat that this does imply that the results could just be a consequence of randomness in our method.

For the infrared $7.3 \mu\text{m}$ channel, there was also evidence of improved error scores to temperature when localising lower down in the atmosphere with thinner localisation widths (figure 5.6, top panel). This was not evident in the plot of benefit previously (figure 5.2, bottom left panel), showing how the cost function (2.117) can isolate different benefits - likely to be those where the analysis error was already large as mentioned.

Most noticeably in the infrared $6.2 \mu\text{m}$ channel (figure 5.6, bottom panel), but also in the $7.3 \mu\text{m}$ channel (not shown), there was benefit to localising around 500 hPa with a thin width. This is in line with the spike here in benefit in figure 5.2, middle right panel.

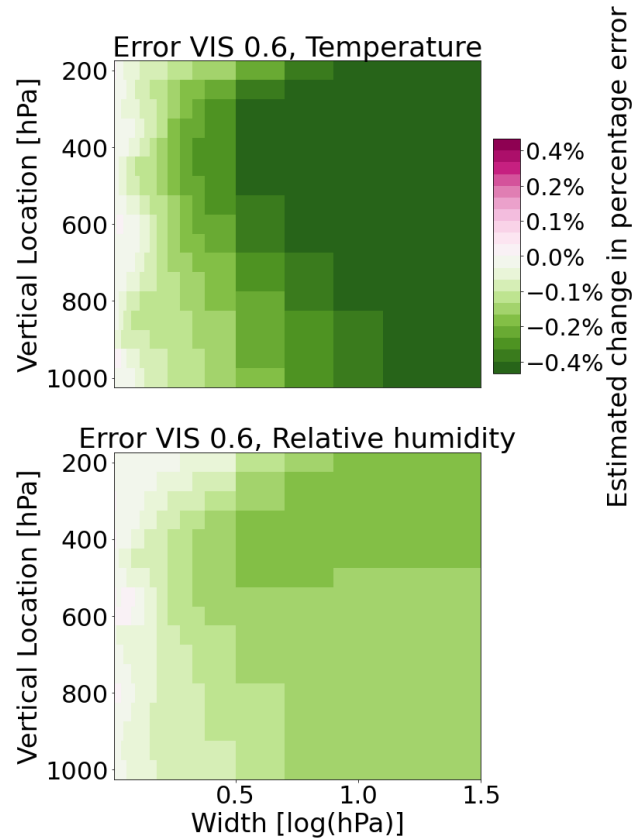


Figure 5.5: Plot of the percentage change cost function for the visible channel separated for temperature and relative humidity. Same colours used for both variables to make for easy comparison.

5.2.4 Optimal parameters for clear/cloudy sky defined by observed radiances

The observations were sorted into clear and cloudy sky based on observed radiance values, and the cost function was calculated again to determine optimal parameter values that depend on the presence of clouds. Cloud definitions were based on the channel being considered - for the visible channel cloudy sky was defined by a visible reflectance observation greater than 0.4 at the radiosonde location, and for the infrared channels by an infrared brightness temperature (in the channel being considered) less than 234 for the $6.2 \mu\text{m}$ channel and less than 251 for the $7.3 \mu\text{m}$ channel. These values were chosen so that enough data points were available for the clear and cloudy sky case, and were not based on any physical reasoning. There was insufficient evidence to suggest that there was a difference between the clear and cloudy cases for the infrared channels, so we describe only the values for the visible channel.

Figure 5.7 shows the cost function for the visible channel when divided into clear and cloudy sky. When clouds

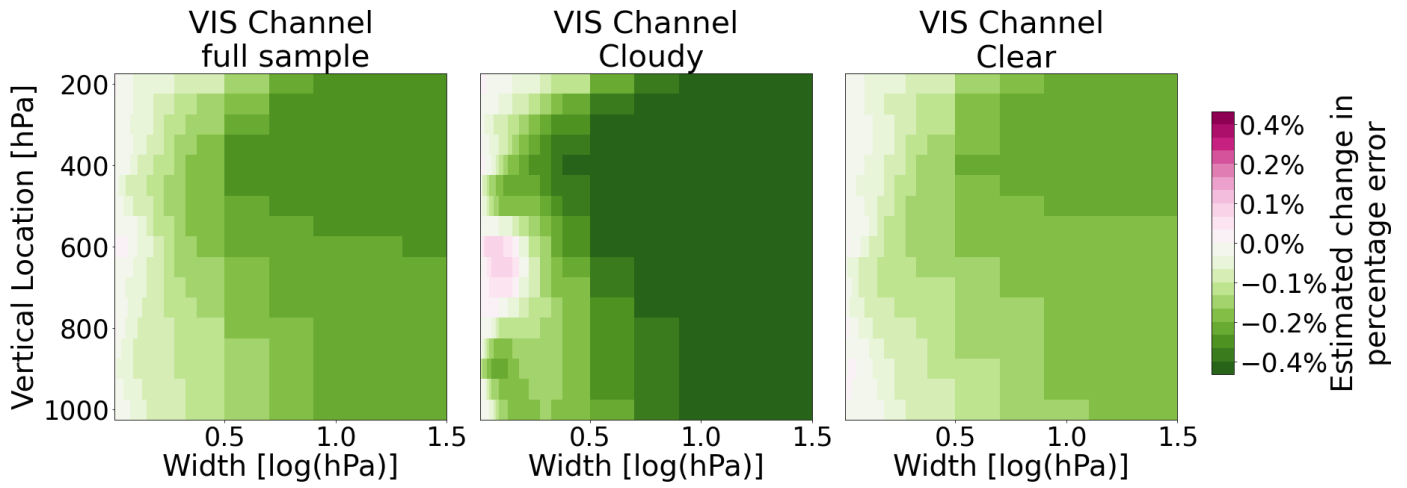


Figure 5.7: Plot of the cost function for varying Gaspari-Cohn parameters, height (y-axis) and width (x-axis), when the sample is divided into clear and cloudy sky. based on observed visible radiances. Cloudy sky shows detrimental increments in the middle of the atmosphere.

were present, increments for the visible channel were detrimental in the middle of the atmosphere between 550 and 800 hPa, although when considering wider localisation widths they were still beneficial on average throughout the atmosphere. When there was less cloud, the increments were in general beneficial in the same sense as when the full sample was considered.

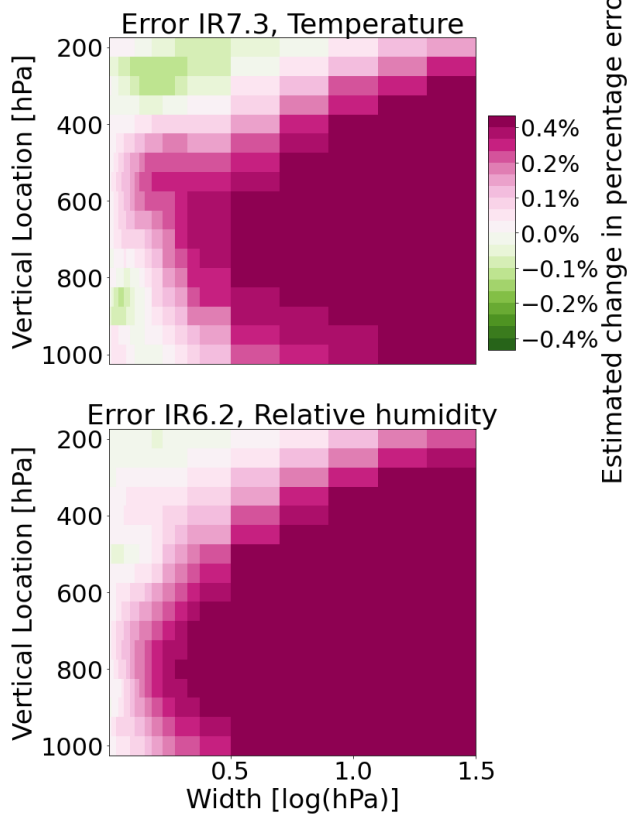


Figure 5.6: Plot of the percentage change cost function for the visible channel separated for temperature and relative humidity. Same colours used for both variables to make for easy comparison.

Figure 5.7 gives novel intuition about the assimilation of visible radiances when clouds are present - that in this model setup, visible radiances are not beneficial between 550 and 800 hPa in cloudy sky. Radiative transfer is much more difficult to model when clouds are present, so it could be that assimilation is detrimental to analysis vectors. Model representation of clouds is also difficult, however (Geiss et al. 2021), so it could be that observations such as visible radiances contain more useful information than model vector, despite issues in the assimilation system. Our results show that the observations make a beneficial contribution only in higher and lower levels of the atmosphere.

We emphasise that the detrimental increments for the visible channel in the middle of the atmosphere in cloudy sky (figure 5.7, middle panel) would be counter-intuitive for a perfect assimilation system - visible radiances should be beneficial in this region. The detrimental increments are most likely a consequence of deficiencies in the model or assimilation system.

5.3 Empirical localisation functions

5.3.1 Global coefficients

Empirical localisation functions along with weighted least squares confidence intervals as in section 2.5.2 were calculated separately for each of the satellite channels. Confidence intervals from this process give a further indicator of how certain we are of a positive cost function at each vertical level. The functions showed that the PAI vectors for the visible channel were beneficial throughout the atmosphere and that this signal was fairly strong, and that the PAI vectors for the infrared channels were only beneficial high in the atmosphere and near the surface of the earth, but that conclusions about whether increments were beneficial were only reliable near the surface.

Localisation functions were taken as piecewise linear functions, linear between every 6th vertical level. Verification was now only carried out to temperature radiosonde observations to preserve linearity in the derivation.

Figure 5.8 shows the optimal localisation functions for each channel plotted against vertical level, calculated by directly optimising the cost function as in section 2.5. A second scale is given by the average analysis pressure at this level to indicate approximately where each vertical level is in practise. The pressure value at each model vertical level is not in general constant. An optimal localisation function value of 1 at a given level suggests that increments on average move in the right direction at this level and are on average of the right magnitude (when optimising to temperature). A negative optimal localisation function suggests that increments are detrimental at the vertical level.

For the visible channel, we have evidence that observations are beneficial throughout the atmosphere (figure 5.8, left panel) as optimal values are always close to 1 (marked in the grey line). Only at the 58th vertical level, approximately at 950 hPa, are we less certain that the visible channel is beneficial, as the optimal localisation function is smaller, indicating some deficiency around this level either in the model or assimilation system. The fact that the visible observations are considered beneficial in our empirical localisation function method is in line with our optimisation of the Gaspari-Cohn function, which also suggested visible observations were beneficial throughout the atmosphere. At around the 48th and 52nd model levels (800-900hPa), we see that the optimal coefficient is a little greater than one. This is a similar level to the spike in average temperature increment between 800 and 900 hPa (figure 5.1, middle left panel), giving further evidence that this tendency in increments is beneficial.

For the infrared channels, there was not enough evidence to suggest that the increments were beneficial at most levels of the atmosphere, but some indication that they were beneficial lower down and higher up in the atmosphere (figure 5.8, middle and right panels). Both channels had positive optimal coefficients for the 10th level, reflecting previous plots (figure 5.2, middle and bottom left panels, and figures 5.3b and 5.3c) that the infrared channels were beneficial higher up in the atmosphere. Optimal localisation functions for both channels have a spike at the 52nd level, which could either suggest a benefit at this level (perhaps relating to infrared observations giving information on cloud top height) or a systematic bias. When looking at the average temperature increment for the channels, both channels have an increase in average increment between 800 and 900 hPa (figure 5.1, bottom and middle left panels). The spike in empirical localisation function around here could show that, at least for the cost function (2.117), this tendency in increments is beneficial, even if it is not shown in the average benefit. Note that the evidence that increments are beneficial lower down in the atmosphere is not present in our Gaspari-Cohn optimisation (figures 5.3b and 5.3c) as we include optimisation to relative humidity there which alters the result lower down in the atmosphere.

Note that in all cases, the wide confidence intervals for the highest vertical level is a consequence of the fact that few radiosondes made verification observations at this level.

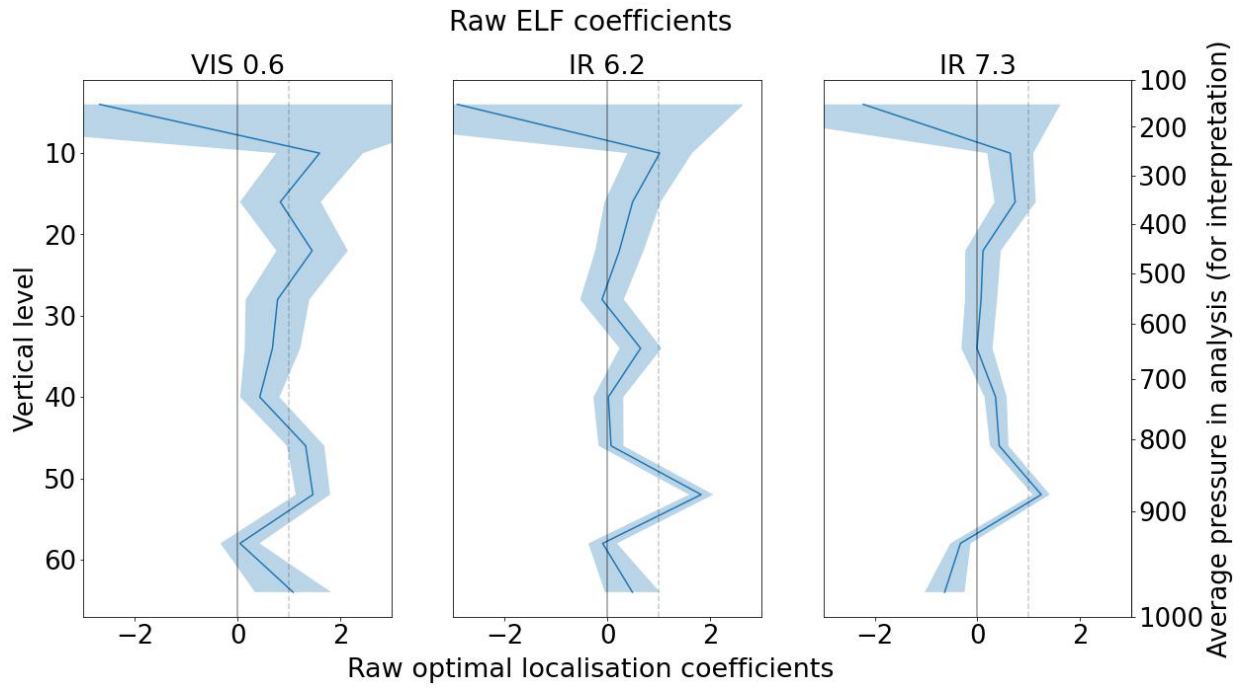


Figure 5.8: Plot of the raw empirical localisation functions calculated for each channel, along with t-test confidence intervals.

5.3.2 Clear/cloudy sky coefficients

For all channels there was a clear difference in the empirical localisation functions for clear and cloudy sky. For visible observations we defined cloudy locations to be those with an observed visible radiance reflectivity value greater than 0.4, and for infrared observations we defined cloudy sky to be locations with an observed infrared radiance brightness temperature value less than 234K and 251K respectively for the $6.2\mu m$ and $7.3\mu m$ channels. We used the observations themselves to define cloudy sky, as these should detect the clouds that most greatly affect the observations. We plot just the case of the $7.3\mu m$ infrared channel, as the result for the other infrared channel is similar.

For the visible channel, the optimal localisation function in the clear sky stayed close to 1 throughout the atmosphere, whereas in the cloudy sky case it varied drastically between vertical levels (figure 5.9). Confidence intervals for the clear sky case suggest there is little evidence that the PAI vectors for the visible channel are on average detrimental (figure 5.9, rightmost panel). This aligns with our intuition from section 5.2.4. Radiative transfer is easier to model in clear sky conditions, and model difficulties relating to representation of clouds are not present in the clear sky case, so we would expect more beneficial increments in this case. For the cloudy case (figure 5.9, middle panel), the optimal localisation function dips below 0 at around 650 hPa (34th model level) and has a spike greater than 1 at around 850 hPa (52nd model level). The low optimal function in the middle of the atmosphere suggests that the visible channel PAI vectors are here detrimental in the cloudy case, which was also observed in the optimisation of the Gaspari-Cohn function (figure 5.7, middle panel). The positive spike lower down the atmosphere suggests that the visible observations are still beneficial here in cloudy sky. Clouds are often found around this level, so this could be evidence that the visible channel gives beneficial information on clouds that then impacts temperature values through correlations. The localisation factor greater than 1 suggests that increasing the magnitude of the PAI vectors would be most beneficial at this level (similar to decreasing the assigned observation error).

For the infrared $7.3\mu m$ channel, in the cloudy sky case (figure 5.10, middle panel) the optimal localisation function is very variable - increments are in this case very detrimental near the surface at the 64th model level, beneficial at the 58th level, and otherwise there is insufficient evidence to show that the increments are either

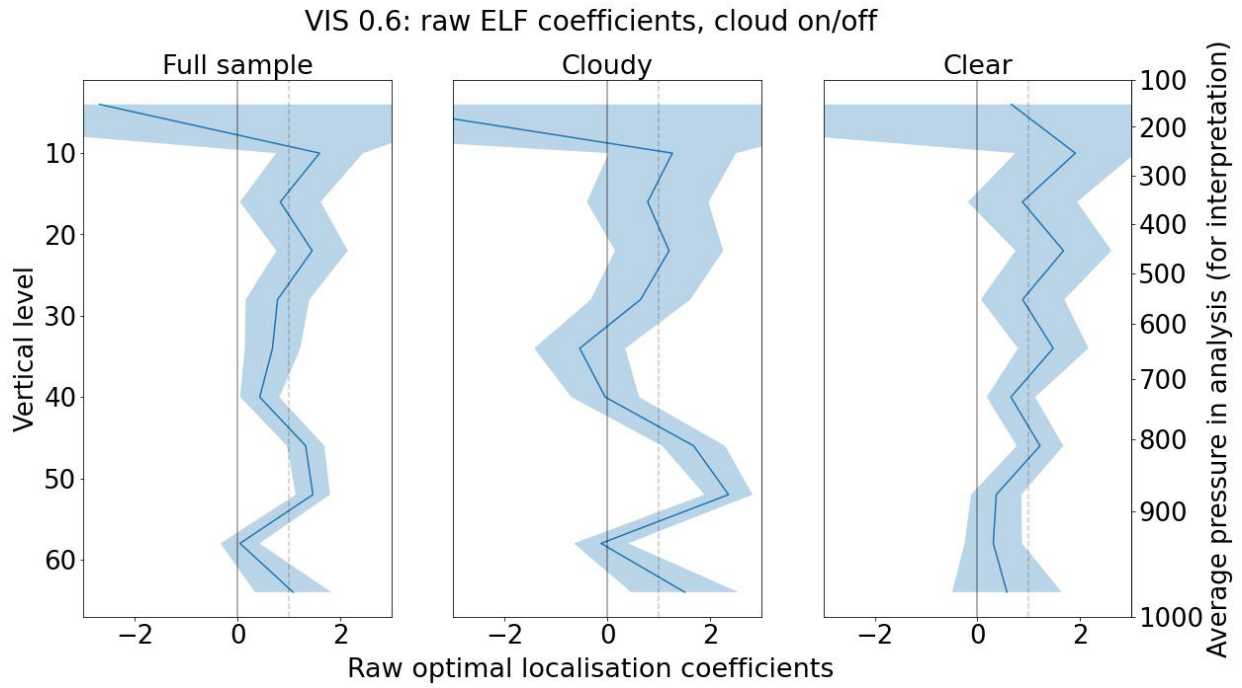


Figure 5.9: Plot of the empirical localisation functions for the visible channel when restricting the data to clear and cloudy sky.

beneficial or detrimental. For the clear sky case (figure 5.10, rightmost panel), the optimal function is similar to the optimal function over the whole sky - some benefit around 890 hPa and 270 hPa, and otherwise no evidence that the increments are either beneficial or detrimental. The strong detriment at the surface (64th level) in the cloudy sky case could suggest model deficiencies - in a perfect system, cloud information from radiances should be beneficial at the surface as clouds directly affect temperature at the surface (for example through radiative effects).

The localisation functions defined by the optimisation process for the piecewise linear functions seem too rough to propose as operational localisation functions, but do seem to motivate some novel localisation functions which we will define in the next section.

5.4 Comparison of localisation strategies

Having optimised parameters within given styles of localisation function (Gaspari-Cohn functions with and without varying coefficients in section 5.2 and piecewise linear functions of model vertical level in section 5.3), we now compare different strategies of vertical covariance localisation for satellite radiances using the percentage change in cost function (2.118). We saw that in general for the visible channel the best strategy was not to localise, although a lot of the error reduction from not localising could be recovered from a Gaspari-Cohn function whose width and nominal height varied depending on our cloud definition from visible channel observations. For the infrared channels, all the best strategies included the observations higher up in the atmosphere (above 300 hPa) and some more error reduction could be achieved by including observations lower down as well. The varying localisation width strategy of the DWD (section 3.1.1) scored best.

We first suggest a novel localisation function with two peaks based on existing results that provides a smooth version of the empirical localisation functions in section 5.3, then compare the error scores for various localisation strategies, including those currently used at DWD.

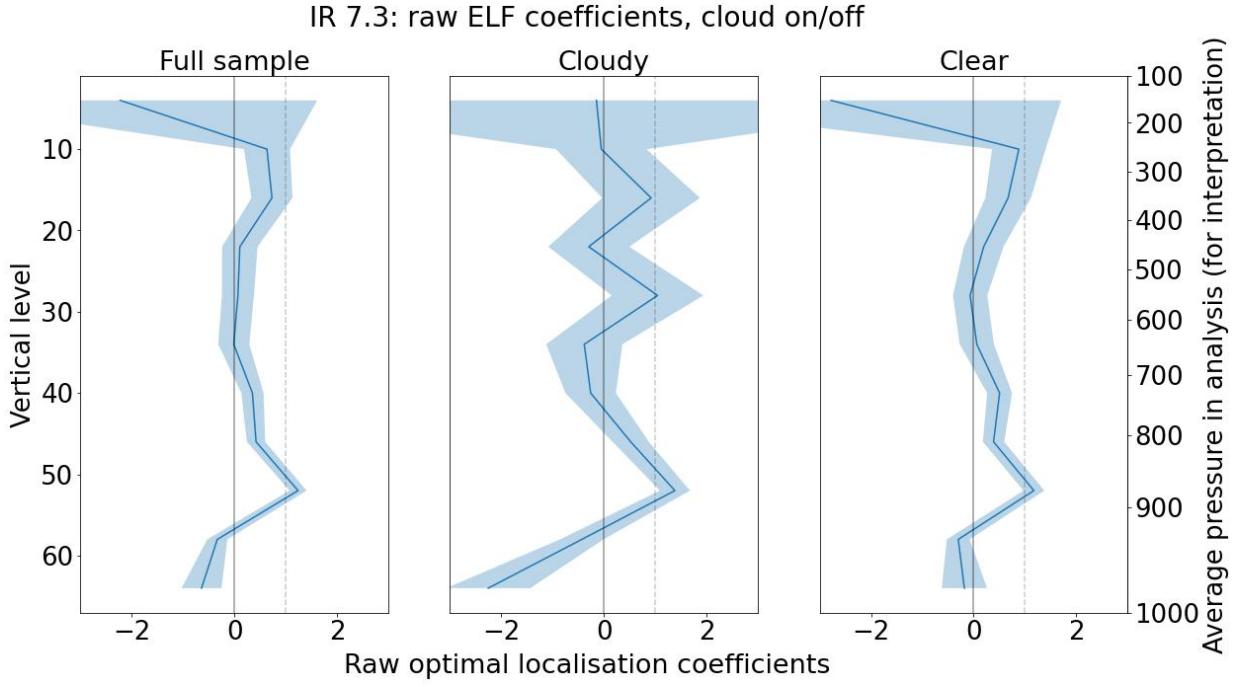


Figure 5.10: Plot of the empirical localisation functions for thesecond infrared channel when restricting the data to clear and cloudy sky.

5.4.1 Normalised sum of Gaspari-Cohn functions

We propose a simple way to normalise the sum of two Gaspari-Cohn functions with different peaks and widths so that both peaks have localisation factor 1. In this way, we can have a multi-peak localisation function that is simply related to existing localisation functions, but still does not inflate increments by taking localisation functions greater than 1.

Our results have shown that increments were often beneficial at more than one level in the atmosphere (figures 5.7, middle panel, 5.8, middle and right panels and 5.9, middle panel), so using a sum of Gaspari-Cohn functions as a localisation function would be a sensible choice. The benefit of radiances at several points in the atmosphere has also already been observed through the PAI method (Diefenbach et al. 2023), and reflects existing experimental study of covariances (Lei, Anderson, and Whitaker 2016) where optimised covariance functions sometimes had several peaks. Theoretically, it makes sense that satellites would have a beneficial influence that peaks at several points, as weighting functions for satellites can be complex and can depend on model variables non-linearly (Hutt et al. 2020), and clouds directly influence surface variables through radiative effects.

Using a multi-peak localisation function is similar to assimilating the observation twice at different levels. If we were to just use the sum of two Gaspari-Cohn functions for this multi-peak localisation function, however, then we could have localisation factors greater than 1 at some levels, which would be account for misrepresented observation errors, which is out of the scope of this thesis.

To derive our multi-peak function, consider the weighted sum of two Gaspari-Cohn functions:

$$\rho^{\text{multi-peak}}(p) = a_1 \rho_1(p) + a_2 \rho_2(p), \quad (5.1)$$

where ρ_1 and ρ_2 are Gaspari-Cohn functions in pressure coordinates with nominal height p_1 and p_2 respectively, and widths σ_1 and σ_2 respectively.

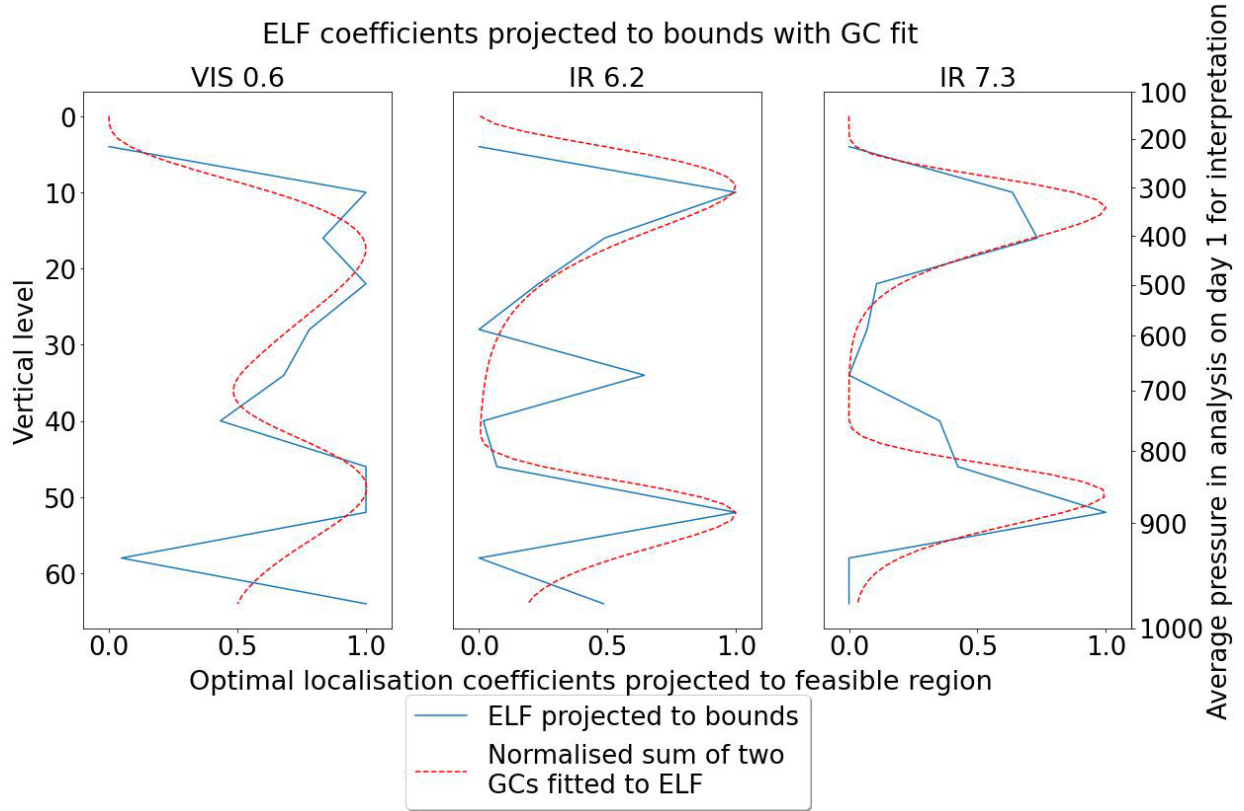


Figure 5.11: Optimal localisation coefficients projected to the set of feasible localisation values that could be used in practise. For comparison, some multi-peaked Gaspari-Cohn functions to show how some more feasible functions could be fitted to these results.

Take a_1 and a_2 to be constants such that:

$$\rho^{\text{multi-peak}}(p_1) = \rho^{\text{multi-peak}}(p_2) = 1. \quad (5.2)$$

Provided we don't choose the same nominal height for ρ_1 and ρ_2 (in which case we can just use a single Gaspari-Cohn function), the relation (5.2) defines a_1 and a_2 uniquely as:

$$a_1 = \frac{1 - \rho_2(p_1)}{1 - \rho_2(p_1)\rho_1(p_2)} \quad (5.3)$$

$$a_2 = \frac{1 - \rho_1(p_2)}{1 - \rho_2(p_1)\rho_1(p_2)}. \quad (5.4)$$

The functions $\rho^{\text{multi-peak}}$ fit well to the empirical localisation functions found in section 5.3. Figure 5.11 shows the optimal piecewise linear localisation functions projected to feasible regions for localisation coefficients (between 0 and 1) along with multi-peak functions $\rho^{\text{multi-peak}}$ with constants chosen to match the shape of the optimal localisation functions. The normalised sum of Gaspari-Cohn functions provides a seemingly good fit to the derived localisation functions. Using the function $\rho^{\text{multi-peak}}$ also seems sensible as it removes overly-complex small scale effects, such as the small peak in the middle of the atmosphere for the infrared $6.3 \mu m$ channel.

For reference, the multi-peaked Gaspari-Cohn functions fitted to the empirical localisation functions are included in table 5.2. Fits were taken by looking at plots and using intuition.

These choices also align with existing intuition - not localising in clear sky for the visible channel as it gives

	Visible 0.6 μm channel		Infrared 6.2 μm channel		Infrared 7.3 μm channel	
	Lower peak	Upper peak	Lower peak	Upper peak	Lower peak	Upper peak
Full sample	$z = 870,$ $\sigma = 0.1$	$z = 350,$ $\sigma = 0.5$	$z = 900,$ $\sigma = 0.05$	$z = 200,$ $\sigma = 0.45$	$z = 870,$ $\sigma = 0.05$	$z = 250,$ $\sigma = 0.3$
Cloudy sky	$z = 900,$ $\sigma = 0.1$	$z = 330,$ $\sigma = 0.4$	$z = 670,$ $\sigma = 0.1$	$z = 200,$ $\sigma = 0.3$	$z = 870,$ $\sigma = 0.05$	$z = 250,$ $\sigma = 0.3$
Clear sky	No localisation	No localisation	$z = 900,$ $\sigma = 0.1$	$z = 200,$ $\sigma = 0.5$	$z = 870,$ $\sigma = 0.05$	$z = 250,$ $\sigma = 0.3$

Table 5.1: Coefficients for multi-peaked Gaspari-Cohn to fit to empirical localisation functions calculated in section 5.3 and used for table 5.3.

information on all levels of the atmosphere in that case (through the information that no cloud is present), and localising lower down in the infrared 6.2 μm channel for cloudy sky, as the observation cannot see through clouds. No difference was seen between the clear and cloudy sky fits for the infrared 7.3 μm channel.

5.4.2 Cost function for different localisation strategies

Here we present and discuss cost function values for optimal parameters calculated in previous sections. Single peak Gaspari-Cohn functions were used with the following values for a global function and a varying clear/cloudy sky function:

	Visible 0.6 μm channel	Infrared 6.3 μm channel	Infrared 7.2 μm channel
Full sample	No localisation	$z = 200,$ $\sigma = 0.3$	$z = 250,$ $\sigma = 0.2$
Cloudy sky	$z = 400,$ $\sigma = 0.3$	$z = 250,$ $\sigma = 0.1$	$z = 250,$ $\sigma = 0.2$
Clear sky	$z = 300,$ $\sigma = 1.0$	$z = 200,$ $\sigma = 0.3$	$z = 900,$ $\sigma = 0.1$

Table 5.2: Coefficients for optimal Gaspari-Cohn functions derived in section 5.2 and used for table 5.3.

As with the multi-peak fits, these align somewhat with intuition - wider or lower localisation functions for clear sky, and thinner/higher localisation functions for cloudy sky. For the visible channel we do not localise.

The empirical localisation functions calculated in section 5.3 and plotted for example in figure 5.8 were first converted to feasible localisation functions by projecting the coefficients to the intervals $[0, 1]$, so look like the blue lines in figure 5.11 more than the full optimal functions. Recall that projecting to feasible regions is not the same as optimising over feasible regions, but in our case gave similar results.

Clear and cloudy sky was defined for each channel differently, for the visible optimisation by a visible observation greater than 0.4 defining cloud, and for infrared observations less than 234K and 251K for the 6.2 μm and 7.3 μm channels respectively defining cloud. The coefficients for the piecewise linear localisation functions and the Gaspari-Cohn localisation functions were varied depending on this definition in the cases defined as 'varying', where the optimal coefficients were obtained in sections 5.2 and 5.3.

For each localisation function, the percentage change in cost function (2.118) was computed and is presented in table 5.3. Differences are of the order 0.1%. Despite being small, such an improvement is still considered valuable in weather forecasting. The errors also contain radiosonde error that cannot (and should not) be removed with localisation, further reducing the percentage values. Note here we are presenting the error against radiosondes for functions that were optimised to the same radiosondes - the results should be considered as a measure of magnitude, and not as an independent verification.

Change in percentage error attributed to radiances
for varying localisation strategies

	VIS 0.6			IR 6.2			IR 7.3		
	Temp. error	RH error	Full error	Temp. error	RH error	Full error	Temp. error	RH error	Full error
Error for experiment settings	-0.59%	-0.31%	-0.43%	-0.15%	-0.04%	-0.08%	-0.19%	-0.03%	-0.10%
DWD Settings	-0.26%	-0.20%	-0.22%	-0.15%	-0.04%	-0.08%	-0.19%	-0.03%	-0.10%
No localisation	-0.59%	-0.31%	-0.43%	1.14%	0.77%	0.92%	1.44%	1.29%	1.35%
Global optimal Gaspari-Cohn	-0.59%	-0.31%	-0.43%	-0.12%	-0.05%	-0.08%	-0.18%	-0.05%	-0.10%
Optimal GC varying clear/cloudy	-0.51%	-0.30%	-0.38%	-0.12%	-0.02%	-0.06%	-0.17%	-0.05%	-0.10%
Global ELF	-0.64%	-0.32%	-0.45%	-0.12%	0.07%	-0.00%	-0.21%	0.08%	-0.04%
Multi-peak GC fitted to global ELF	-0.59%	-0.31%	-0.42%	-0.14%	-0.05%	-0.09%	-0.19%	-0.05%	-0.11%
Clear/cloudy ELF	-0.69%	-0.37%	-0.50%	-0.06%	0.16%	0.07%	-0.24%	0.00%	-0.10%
Varying multi-peak GC fitted to clear/cloudy ELF	-0.65%	-0.37%	-0.48%	-0.14%	-0.04%	-0.08%	-0.18%	-0.05%	-0.10%

Table 5.3: Estimated percentage change in error attributed to the three channels with different localisation functions.

For the visible channel, localisation functions that included observations at all levels (possibly with different weights) performed best, although a varying Gaspari-Cohn function depending on observation values could recover a lot of error reduction. Errors for relative humidity and temperature seemed to follow a similar pattern, so we consider in general just the total error.

The lowest error when changing the localisation for the visible channel was for the empirical localisation function with coefficients that were allowed to vary depending on whether a cloud was present (defined as the nearest visible channel observation being greater than 0.4). As there were many degrees of freedom when optimising this function, it could be that the coefficients are overly-fitted to our verification data. The global empirical localisation function also scored well. The errors for the empirical localisation functions were not much less than error with all observations used ('experiment analysis error'), but the empirical localisation functions have many coefficients that need to be estimated, so we do not expect that they would be better than not localising at all when applied to verification data outside of our sample. The fitted multi-peak functions performed very well for the visible channel. Although they do not save on computational cost as they still incorporate observations at all levels, they are easy to understand and adapt, so could be a good option for operational use.

Using a thin Gaspari-Cohn function to localise visible radiances (as at the DWD) we were able to recover some of the error reduction from using all observations, but the varying Gaspari-Cohn function performed much better, recovering almost all of the error reduction from using all observations. Single-peak Gaspari-Cohn functions have the advantage over wider localisation functions that they use less computational power and are easier to understand and implement than more complex localisation functions.

For the infrared channels, we see that the strongest benefit in our sample was high up in the atmosphere, above 300 hPa, as localisation functions that included these increments always performed best. There was potentially some advantage to including observations lower down as well.

Compared to other functions, the empirical localisation functions performed poorly for the infrared channels

even though there were so many degrees of freedom in the optimisation. The empirical localisation functions did, however, inspire better-performing multi-peak Gaspari-Cohn functions. For the $6.3\ \mu m$ channel, there was no improvement when using an empirical localisation function compared to an optimal Gaspari-Cohn function when optimising to any variable, and for the $7.2\ \mu m$ channel the improvement was only slight. This could be because the use of one localisation coefficient per model-level, instead of varying coefficients depending on pressure values, is detrimental. It could also be because we optimised over all possible coefficients and projected to feasible values, rather than optimising over all feasible values, which would give (slightly) different coefficients. For both infrared channels the multi-peaked Gaspari-Cohn functions performed better than the optimal single-peak Gaspari-Cohn function for temperature, and for the $7.2\ \mu m$ channel the multi-peaked Gaspari-Cohn function performed better for relative humidity as well. This suggests that it was the restrictions on the style of localisation function for the empirical localisation function, rather than the coefficients themselves, that were limiting the performance of these functions.

For the infrared channels, the DWD strategy of taking a varying nominal height depending on the observation sensitivities almost always performed better than other functions. The optimal single-peak Gaspari-Cohn functions were able to recover some of this error, but as they are by nature quite similar to the varying Gaspari-Cohn functions at DWD, this is no evidence to suggest that constant single-peak Gaspari-Cohn functions would be advantageous when localising vertically for these channels. The multi-peaked Gaspari-Cohn functions were able to recover almost all of the error reduction from the adaptive solution, despite being constant with regard to weather conditions. It could be that using an adaptive multi-peaked Gaspari-Cohn function could bring even more benefit in practise.

Chapter 6

Discussion

6.1 Optimal localisation strategies

6.1.1 Comparison to other methods for choosing vertical radiance localisation

Many possible methods have been used to try and find optimal vertical localisation strategies for satellite radiances, compared to which the approach shown here is very data based. Existing methods include localisation based on sensitivity functions (Hutt et al. 2020, Miyoshi and Sato 2007) where only locations that have a big direct impact on the model equivalents include observations, and localisation based on covariances (Fertig et al. 2007, Lei, Anderson, and Whitaker 2016, Lei et al. 2020b) that aim either to directly estimate sampling error and adjust for it, or just to include the observation only at model levels that are strongly correlated with the observation. The method here goes one step further, trying to isolate locations where the observations make a beneficial contribution to the analysis regardless of their contribution to the model equivalent or their covariance with the observation (although these are important factors).

There are many considerations to be made when choosing a vertical localisation strategy. Too broad localisation functions can lead to increased forecast error from spurious correlations (Miyoshi and Sato 2007, Zhang, Stensrud, and Clothiaux 2021). On the other hand, too narrow localisation functions can lead to errors relating to discrepancies between assimilation of different observations, and analyses that contradict various balances (commonly observed approximations for the physical laws of the atmosphere, Fertig et al. 2007, Houtekamer and Zhang 2016). When choosing the nominal height for a radiance based on sensitivity functions, there are further potential issues, for example that sensitivity functions are not always stored by radiative transfer programs, and that sensitivity functions can have multiple peaks leading to further ambiguity for the vertical location (Hutt et al. 2020). Considering just sensitivity functions also ignores important correlations between distant variables, for example radiative effects at the surface (Fertig et al. 2007). Interactions between satellite channels are also important - if localisation functions for similar channels do not overlap enough, then correlations between the channels are falsely ignored (Aravéquia et al. 2011).

Operational vertical localisation strategies generally combine many such ideas. For example, at the DWD for the infrared 6.2 and 7.3 μm channels the nominal height is chosen at the point where the transmission function (a part of the radiative transfer method) reaches half, rather than using the raw sensitivity. Further, visible radiances at the 0.6 μm level are not assimilated higher up in the atmosphere to eliminate errors associated with misrepresented cross-channel correlations.

Computational methods have also been used to calculate ‘optimal’ localisation functions. The *hierarchical filter* method (Lei, Anderson, and Whitaker 2016) uses multiple ensembles to sample the distribution of some assimilation coefficients, and uses the empirically calculated distribution of these coefficients (and in this way

the distribution of the sampling error) to find localisation coefficients that minimised sampling error. Lei et al. 2020b used similar methods during assimilation cycling to choose these optimal coefficients depending on the state of the forecast at each assimilation time. Anderson 2016 extended the hierarchical filter method to account for more errors relating to the assimilation, not just sampling errors. Empirical localisation functions (Anderson and Lei 2013) use a simulated environment to optimise localisation coefficients by modelling the forecast, the observations and the true value of the atmospheric state, and then minimising the error to the modelled truth.

Our method directly uses near-operational data and optimises so that PAI vectors correct errors as much as possible. This is a more data-based method than choosing a localisation function based on the radiative transfer model or correlations - we do not specify which potential flaws in the assimilation algorithm we aim to correct with localisation, and consider just the output of assimilation. This has the advantage, similar to Anderson and Lei 2013 or Anderson 2016, that it allows localisation to correct more errors than just sampling error. Our method uses operational data, so does not suffer from errors in modelling the true value of the atmosphere or observations that occur when calculating empirical localisation functions.

There are also disadvantages to our method. We assess only simple adaptive methods here, but adaptive methods could be the most effective, as suggested for the infrared channels in section 5.4.2. Localising based on sensitivity functions has the advantage that adaptive methods can easily be computed. Hierarchical filter methods have the advantage that localisation only corrects for spurious correlations as intended, rather than correcting for other issues in the data assimilation system, which perhaps should be left to other methods such as inflation and bias correction. Unlike other computational methods, we are limited by our verifying observations - when using a simulated truth, we can optimise error to any model variable or energy norm. When using hierarchical filter methods, there is no need for verifying observations or a model for the truth at all.

A high estimated error from the PAI method when localising at a certain location could reflect a range of issues with the data assimilation system. These include:

- Spurious correlations between the observation and the model variables at the location (the issue we aim to eliminate with localisation);
- Issues with the model operator h - if the observation operator does not represent well the contribution of model variables at a certain level to the observation, then increments may be detrimental;
- Issues with linearity assumptions on the operator h ;
- Issues with misrepresented observation error covariances, including covariances between different observations;
- Model deficiencies in the representation of clouds and their radiative effects.

Broadly, our conclusions reflect the first of these - visible radiances are on average beneficial throughout the atmosphere, reflecting that they depend on variables at many vertical levels, and infrared radiances are only beneficial higher up in the atmosphere, reflecting that they cannot see through clouds.

Our results are complementary to existing intuition about vertical localisation of radiances. If we correct only when observations are beneficial, we expect forecasts to have better scores, but in the long run there is less transparency as to what we are correcting with localisation. Our results only give information on errors at analysis time, but do suggest exactly where increments are beneficial and show the effect of localisation when all elements of the operational data assimilation system are in place. The motivation and definition for multi-peak Gaspari-Cohn functions given in sections 5.3 and 5.4.1 could be a step towards a compromise between the different considerations when choosing vertical localisation for radiances.

6.1.2 Comparison to DWD strategy

For the visible channel our results support the strategy of localising visible radiances at around 800 hPa with a width of around $0.3 \ln(\text{hPa})$, although in cloudy sky a localisation function centered slightly lower at around 850 hPa may be better and there could be benefit to including visible radiances higher in the atmosphere as well. Our results suggested that observations in the visible channel were only slightly detrimental in middle levels around 600 hPa, and this effect was mostly present in the cost function for relative humidity radiosonde observations. Localising at 800 hPa with a width of $0.3 \ln(\text{hPa})$ would dampen the detrimental increments in the middle of the atmosphere but not remove too many beneficial increments elsewhere, so is a sensible choice of localisation. In cloudy sky, our results suggested that the visible increments were detrimental for a wider range of values in the middle of the atmosphere, so localising even lower down, say with a Gaspari-Cohn centered around 850 hPa, may be even more beneficial.

For the infrared channels, our results also support the strategy of DWD - to take an adaptive nominal height based on the observation value and to localise with a width of around $0.3 \ln(\text{hPa})$ around this point. Our results showed lower errors when the infrared observations were localised with quite a thin width ($0.1\text{-}0.3 \ln(\text{hPa})$) higher up in the atmosphere. The nominal heights chosen by DWD are generally higher up in the atmosphere (although vary depending on the observation) so are in line with our results.

Complementary to the strategy of the DWD, our results suggested that radiances could be included in more levels of the model. For the visible channel, increments were on average beneficial throughout the atmosphere (figure 5.3a), so a wider localisation width than $0.3 \ln(\text{hPa})$ could be tested. For the infrared channels, some benefit was seen lower down in the atmosphere in the Gaspari-Cohn optimisation (figures 5.3b and 5.3c) and the empirical localisation functions had a multi-peaked structure (figure 5.8, middle and right panels), so a multi-peaked localisation function for these channels could lead to further error reduction. The multi-peaked Gaspari-Cohn functions performed well under our metric (section 5.4.2) and combined with an adaptive method as currently used at the DWD (section 3.1.1) could contribute to improvements in forecast skill.

6.2 Limitations and potential for further study

The results in this thesis have shown that the PAI method can be used for approximating localisation functions in a near-operational setting. Further study could be carried out to see whether the method is effective in an even broader range of cases.

Using assimilated radiosondes for verification

To quantify the error for a PAI vector we compared analyses to non-assimilated radiosonde observations. Operationally forecasts are run with assimilated radiosondes, so a new forecast had to be run in which radiosondes were not assimilated. If the method were to work similarly well when testing to assimilated radiosondes, the output of operational forecasts could be used, which is a much larger data set.

As mentioned previously, if a radiosonde is assimilated, the analysis is likely to be very accurate around that location, in which case the radiances would just be correcting random noise rather than large systematic differences such as a falsely predicted cloud. This means that by correcting to assimilated radiosondes we could be missing out some of the most important assimilation events. Further, if the radiosonde has already been incorporated into the assimilation procedure, the analysis will already have been pulled towards the radiosonde observation, and optimising localisation coefficients to these observations would be similar to including the observation 'twice', which could lead to overfitting. A new cost function accounting for overfitting could solve this issue.

Non-diagonal R

We would expect radiances from different channels to have correlated errors, so it could be that in the near future non-diagonal R -matrices will be used for assimilating radiances. The retrospective localisation approximation would have to be checked again in this case, and the error analysis would need to be generalised.

Other methods for assimilating radiances

Due to the many linearity assumptions, it could be that observation-space local analysis (as in the LETKF) is not the best way to localise radiances. For the ensemble Kalman filter, localisation can be carried out in model space by modifying the ensemble background covariance matrix P_b , so a vertical location for the satellite observations does not need to be defined. Model space localisation is generally considered a better strategy for localisation of radiances as it allows for variable-dependent localisation (Lei, Whitaker, and Bishop 2018). Although model space localisation is not currently possible in operational forms of the LETKF, some methods have been proposed, for example Z -localisation where the background ensemble is modified, similar to modifying the background ensemble covariance matrix (Kotsuki and Bishop 2022).

Chapter 7

Conclusion

This work has shown that the PAI method is a useful tool in understanding the effect of vertical localisation on analysis vectors in data assimilation with the LETKF that is complementary to other methods. First it was shown theoretically under which conditions the PAI method can approximate a new localisation function - an exact expression for the error was derived, and then a bound on this error was suggested based on some parameters that are easy to calculate. Retrospective localisation with PAI was then used with operational data to optimise localisation parameters for satellite radiances in the forecast system of the German Weather Service. In some cases our results reflected existing knowledge about localisation parameters, so gave reason to believe that the PAI method for retrospective localisation can detect qualitative concepts in data. For example, it is expected that radiances in the $0.6\mu m$ channel are beneficial throughout the atmosphere and that radiances in the infrared $6.2\mu m$ and $7.3\mu m$ channels are beneficial higher up in the atmosphere and in the boundary layer (through radiative effects). Our optimisation in section 5.2 reflected this existing knowledge. The PAI method was able to give further details that are not documented through other methods. For example, the method showed specific levels where system deficiencies led to detrimental PAI vectors, such as in the middle of the atmosphere for the visible channel in cloudy situations (figure 5.7). Our method indicated the potential benefit of a multi-peak Gaspari-Cohn function, which we tested against our verification metrics in section 5.4.2.

We started by addressing our first aim of formalising the theory of partial analysis increments and relevant concepts in the LETKF, and used this to produce novel results. We made clear the exact assumptions made when approximating the Kalman gain using LETKF outputs, in our case the approximation to the coarse analysis grid and the issue relating to the storage of Y_a , which gave a clearer idea of where inaccuracies could be in experimental work later. As our assumptions in formulating the data assimilation setup were clear we were then able to produce novel and practical results, such as the error analysis in 2.3.2.

To address our second aim of understanding the PAI approximation analytically, we used linear algebra to characterise the error when retrospectively localising. This characterisation is important for understanding the PAI method, and could be used to bound the error in retrospective localisation, for which we suggested one possible upper bound.

To address our third aim of using PAI to understand where different observations are beneficial for forecast initial conditions, we calculated PAI vectors using near-operational data. The PAI method was able to give new insight into how satellite radiances influence forecasts and allowed us to make critical remarks on operational strategies for vertical localisation of radiances. When looking at global averages of PAI vectors and benefit vectors (section 5.1), we saw that there were some tendencies in partial analysis increments that could be related to improvements associated with assimilation of radiances. For the infrared channels, partial analysis increments often caused a decrease in temperature and relative humidity at isolated levels at the atmosphere where we also saw that PAI vectors decreased error to radiosondes (figures 5.1 and 5.2, middle left panel,

middle right panel and bottom right panel in both). Having described the theory behind PAI vectors in detail, we proposed several potential reasons for this, including features of the observation operator, or model biases at these levels that were corrected by the observations.

To address our final aim of using PAI to approximate and optimise different vertical localisation settings for satellite radiances, we used several methods of optimisation, including approximations of errors with varying parameters, and direct optimisation for idealised localisation functions.

When optimising Gaspari-Cohn localisation functions to our cost function (2.117), we observed qualitative behaviour that was in line with existing studies, as well as some newly observed trends. Our optimisation method showed that PAI vectors for the visible $0.6\mu m$ channel were beneficial throughout the atmosphere except in cloudy sky (figures 5.3a and 5.7). It was expected that the $0.6\mu m$ radiances would be beneficial throughout the atmosphere as the channel gives cloud information that affects variables at all model levels. The detrimental effect of visible radiances in cloudy sky (figure 5.7) is non-trivial: there are often model deficiencies in representing clouds (Geiss et al. 2021), so visible radiances could give valuable information, but radiative transfer is also more difficult to model when clouds are present so we may also expect detrimental increments. Our results show vertically where each of these two effects dominates for this particular model and assimilation system. There were some very isolated detrimental effects of the visible channel at around 600 hPa, which our results suggest come mostly from the relative humidity error (figure 5.5).

Increments for the infrared channels were found to be detrimental right up to areas quite high in the atmosphere where we would have expected some benefit (figures 5.3b and 5.3b). This could be evidence that the observations are detrimental in these areas, or that the PAI method for retrospective localisation was false at these levels. Infrared observations were also beneficial lower down in the atmosphere for temperature but not for relative humidity, and very specifically at around 500 hPa for relative humidity and not for temperature (figure 5.6).

There was not always enough evidence to make precise statements. Sometimes the cost function for the Gaspari-Cohn parameters varied a lot when calculating for subsamples of the data set (figure 5.4). This could either be because the benefit of PAI vectors at some levels was 0, in which case the values would fluctuate, or because a bigger average was needed. Looking at empirical localisation functions based on PAI went some way to understanding this by comparing the optimisation to a least squares regression and looking at confidence intervals. For the infrared channels we saw that when optimising to temperature in many vertical levels there was not enough evidence to conclude that the optimal localisation function was non-zero, so it could be that at many levels the temperature increments were not beneficial. The benefit plots (figure 5.2) suggested more conclusively that infrared observations were detrimental lower down in the atmosphere, although this was specific to the benefit function and was not always present in the cost function (2.117).

When optimising over an idealised class of localisation functions, we were able to give an estimate for what an unrestricted localisation function optimised to the verifying radiosonde observations might look like, and proposed a new function (section 5.4.1) to give a smoother and more easily implementable fit to this. The empirical localisation functions were quite rough, reflecting how difficult it is to find a perfect solution for localising radiances vertically, but did in general show some consistent behaviour: visible observations were beneficial throughout the atmosphere so the optimal localisation function for this channel was always positive, and optimal functions for the infrared channels consisted of two peaks, one nearer the surface and one high in the atmosphere (figure 5.8). We proposed a method to normalise the sum of two Gaspari-Cohn functions to use as a fit for these ‘multi-peaked’ optimal localisation functions that seemed to match nicely (section 5.4.1). Sometimes optimised localisation functions were greater than one, suggesting the need for an adjustment of observation error values used in the assimilation system.

We were able to use the PAI method (and in particular the cost function (2.117)) to critically assess which localisation strategies are the best. By looking at raw cost functions, we saw that the strategies of the German

weather service were often very advantageous and had low error scores, particularly the adaptive localisation method for infrared radiances, but could be improved by considering functions with multiple peaks. For the visible channel, wider localisation widths were in general the best, although an adaptive strategy could recover a lot of the error reduction from not localising while reducing computational cost, and dampening increments in the middle of the atmosphere, particularly in cloudy sky, could be advantageous. For the infrared channels, only increments high in the atmosphere were clearly beneficial, although there was evidence of benefit near the surface that could be investigated further, and perhaps an adaptive function with two peaks could bring further benefit to forecasts.

Although the PAI method was able to bring new evidence to a difficult problem, it should be noted that some of our results could be related to approximations in the PAI method and not features of the forecast and assimilation system. Our calculations included the effect of inflation in our data, which could lead to systematic trends in PAI vectors as seen in figure 5.1. As the infrared and visible channels may have high cross-channel covariances with each other, the errors characterised in section 2.3.2 could be large.

Our results on the effect of localisation on an operational forecast are complementary to existing work on vertical localisation of satellite radiances. Optimal localisation functions are often rough and complex (Lei, Anderson, and Whitaker 2016), so a perfect localisation strategy is unlikely to be possible. Choices of vertical localisation must therefore be based on qualitative ideas. PAI vectors can show where strategies are effective in practise, and where existing intuition such as the behaviour of the sensitivity function is not a dominating factor. We have shown further that some simple modifications of existing localisation methods, such as adding an extra peak to the localisation function, could bring reductions of analysis error. This should be tested in the context of a forecast to see if these benefits are seen after integration. Further analysis with the PAI method, for example using different verifying observations, using fully-operational forecast output, or using the error in section 2.3.2 to calculate where retrospective localisation is reliable, could give even more detail on the effect of radiance assimilation.

Appendix A

Proofs of fundamental data assimilation results

Theorem A.0.1 (Evensen, Vossepoel, and Leeuwen 2022, p. 63-65). *Let $X \sim \mathcal{N}(x, B)$ and $Y \sim \mathcal{N}(Hx, R)$ for $H \in \mathbb{R}^{m \times n}$. Let x_b and y_o be observations of X and Y respectively. Then the maximum likelihood estimator for x is given by:*

$$x_a = x_b + K(y_o - Hx_b),$$

where

$$K = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}. \quad (\text{A.1})$$

This expression can also be written as:

$$(B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} = B H^T (H B H^T + R)^{-1}. \quad (\text{A.2})$$

The second expression (2.4) is more well known.

Proof. Let $f(x|x_b, y_o)$ refer to the posterior distribution of x . Then as the prior distribution is normal, we have:

$$f(x|x_b, y_o) \propto \exp \left[- (x_b - x)^T B^{-1} (x_b - x) - (y_o - Hx)^T R^{-1} (y_o - Hx) \right].$$

So maximising f over x is equivalent to minimising J , where:

$$J(x) = (x_b - x)^T B^{-1} (x_b - x) + (y_o - Hx)^T R^{-1} (y_o - Hx). \quad (\text{A.3})$$

As B and R are covariance matrices, they are positive semi-definite. So finding a minimum of J is equivalent to finding the argument of $\nabla J(x) = 0$. We can directly derive:

$$\nabla J(x) = 2B^{-1}(x - x_b) + 2H^T R^{-1}(Hx - y_o). \quad (\text{A.4})$$

Setting this to 0 gives that:

$$x_a = x_b + (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} (y_o - Hx_b). \quad (\text{A.5})$$

So it remains to show that this expression is equal to the one from the previous theorem. This is a matrix

relation known as the *Woodbury Identity*. We give the full derivation here. Let:

$$A = (B^{-1} + H^T R^{-1} H)^{-1}.$$

By starting with expression 2.4, we have:

$$\begin{aligned} BH^T (HBH^T + R)^{-1} &= BH^T (HBH^T + R)^{-1} \\ &= AA^{-1} BH^T (HBH^T + R)^{-1} \\ &= A(B^{-1} + H^T R^{-1} H) BH^T (HBH^T + R)^{-1} \\ &= A(H^T + H^T R^{-1} H B H^T) (HBH^T + R)^{-1} \\ &= AH^T R^{-1} (R + HBH^T) (HBH^T + R)^{-1} \\ &= (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}. \end{aligned}$$

□

Proposition A.0.1 (Kalnay 2003, p. 156-157). *Suppose we use a data assimilation algorithm with Kalman gain as in (2.4) and linear observation operator H . When all the relevant inverses are well defined, the covariance matrix for ϵ_a can be written both as:*

$$A := \mathbb{E}[\epsilon_a \epsilon_a^T] \tag{A.6}$$

$$= (B^{-1} + H^T R^{-1} H)^{-1}, \tag{A.7}$$

and:

$$A = (I_n - KH)B. \tag{A.8}$$

Proof. The proof consists of working through the different expressions for K and A .

First start with the expression (2.8) for A and derive:

$$\begin{aligned} A &= (I - KH)B(I - KH)^T + K R K^T \\ &= (I - KH)B + (-BH^T + KHBH^T + KR)K^T \\ &= (I - KH)B + (-BH^T + BH^T (HBH^T + R)^{-1} (HBH^T + R))K^T \\ &= (I - KH)B. \end{aligned}$$

Where the second line comes from expanding brackets, and the third line comes from the first expression for the kalman gain (2.4). This concludes the proof of (A.8). For (A.7), we can then substitute in the second expression for the Kalman gain (2.3) to get:

$$\begin{aligned} A &= (I - KH)B \\ &= (I - (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} H)B \\ &= (B^{-1} + H^T R^{-1} H)^{-1} (B^{-1} + H^T R^{-1} H - H^T R^{-1} H)B \\ &= (B^{-1} + H^T R^{-1} H)^{-1}. \end{aligned}$$

□

Appendix B

Required theory from linear algebra

B.1 Elementwise/Schur/Hadamard product

The elementwise matrix product is used to implement local analysis in the LETKF. In mathematical literature it is typically referred to as the ‘Hadamard’ product after mathematician Jacques Hadamard, but is also often referred to as the ‘Schur product’, perhaps after a theorem of mathematician Issai Schur. For local analysis in the LETKF with diagonal observation covariance matrix R , we only need the definition of the Hadamard product. For covariance localisation in the Ensemble Kalman filter, more properties of the Hadamard product are important.

Definition B.1.1 (Hadamard/Schur/elementwise product, Styan 1971). *Let $A, B \in \mathbb{R}^{n \times m}$ for $n, m \in \mathbb{N}$ be matrices. Define the Hadamard/Schur/elementwise product of A and B as $A \circ B \in \mathbb{R}^{n \times m}$, where $A \circ B$ has elements:*

$$(A \circ B)_{ij} = (A)_{ij}(B)_{ij}.$$

B.2 Projection maps

We use projection maps as a tool to deal with singularity of various matrices in the thesis. This section is based on Axler 2014, p. 211-216.

To define an orthogonal projection, we first need to define an orthogonal complement.

Definition B.2.1 (Orthogonal complement). *Let U be a subspace of an inner product space $(V, \langle -, - \rangle)$. Then the orthogonal complement $U^\perp$ of U is defined as the set vectors that are orthogonal to all vectors in U (except 0), that is:*

$$U^\perp = \{v \in V : \langle v, u \rangle = 0 \text{ for all } u \in U \setminus \{0\}\}.$$

Some properties of the orthogonal complement include:

Proposition B.2.1. *Suppose U is a subspace of an inner product space $(V, \langle -, - \rangle)$. Then:*

- *The set $U^\perp$ is a subspace of V ;*
- *All vectors $v \in V$ can be written uniquely as $v = u + w$ for $u \in U$ and $w \in U^\perp$;*

Proof. See Axler 2014, p. 211-216 for more details. □

The second part of B.2.1 allows us to make the following definition:

Definition B.2.2. Let U be a subspace of an inner product space $(V, \langle -, - \rangle)$. Define the projection map $P_U : V \rightarrow U$ of V onto U as $P_U(v) = u$, where $v = u + w$, $u \in U$, $w \in U^\perp$ is the unique representation of v in terms of U and $U^\perp$ elements.

It can be shown that P_U is a linear function on V .

In this thesis, we use the projection onto the span of a set of vectors - the columns of a matrix X_b . We show now some particular results relating to this space.

Proposition B.2.2. Let $X_b \in \mathbb{R}^{n \times N}$ for $n, N \in \mathbb{N}$. Let $u_1, \dots, u_l$ be an orthonormal basis for the column space of X_b , and let:

$$Q = \begin{pmatrix} | & | & | & | \\ u_1 & u_2 & \dots & u_l \\ | & | & | & | \end{pmatrix}.$$

Then QQ^T represents the orthogonal projection onto the column space of X_b .

Proof. Extend $\{u_1, \dots, u_l\}$ to an orthonormal basis of $\mathbb{R}^n$ with $\{\tilde{u}_{l+1}, \dots, \tilde{u}_n\}$. Then for any vector $v \in \mathbb{R}^n$ we can uniquely express v as:

$$v = \alpha_1 u_1 + \dots + \alpha_l u_l + \alpha_{l+1} \tilde{u}_{l+1} + \dots + \alpha_n \tilde{u}_n.$$

We can check that $\alpha_1 u_1 + \dots + \alpha_l u_l$ is then the projection of v onto the column space of X_b . But also by orthonormality of the basis:

$$Q^T v = (\alpha_1, \dots, \alpha_l)^T.$$

So then:

$$QQ^T v = \alpha_1 u_1 + \dots + \alpha_l u_l.$$

So QQ^T indeed represents the projection map. □

We now show how this projection map can be used to create an invertible matrix:

Proposition B.2.3. Let X_b , $u_1, \dots, u_l$ and Q be as in proposition B.2.2. Then the matrix

$$Q^T X_b X_b^T Q,$$

is invertible.

Proof. Suppose that for $w \in \mathbb{R}^l$, we have that

$$Q^T X_b X_b^T Q w = 0. \tag{B.1}$$

Then if $v = X_b^T Q w$, (B.1) implies that $v^T v = 0$. So $v = 0$. Then for all columns $(X_b)_{:,i}$ of X_b , we have that:

$$(X_b)_{:,i}^T Q w = 0.$$

So Qw is perpendicular to all of the column space of X_b . But Qw is by definition in the column space of X_b . So $Qw = 0$. So $w = 0$ as the columns of Q are linearly independent. So we have a contradiction, and $Q^T X_b X_b^T Q$ must be invertible. □

The matrix Q relates to the null space of X_b :

Proposition B.2.4. The vectors

$$v_i = X_b^T u_i, i = 1, \dots, l,$$

span $U^\perp$, where U is the null space of X_b .

Proof. Let $w_1, \dots, w_m, m \in \mathbb{N}$ be an orthonormal basis for the span of $\{v_1, \dots, v_l\}$, and extend to an orthonormal basis for $\mathbb{R}^N$ with $\{\tilde{w}_{m+1}, \dots, \tilde{w}_N\}$. Then we have that:

1. All of $\{\tilde{w}_{m+1}, \dots, \tilde{w}_N\}$ are in the null space of X_b ,
2. None of $w_1, \dots, w_m$ are in the null space of X_b .

To see (1), note that for all $i = 1, \dots, l$, $u_i^T X_b \tilde{w}_i = 0$. So $X_b \tilde{w}_i$ is both in the column space of X_b and orthogonal to the column space of X_b , so $X_b \tilde{w}_i = 0$.

To see (2), write each $w_i, i = 1, \dots, m$ as $w_i = X_b^T Q \hat{w}_i$ (as all w_i are in the span of the v_i). Then if $X_b w_i = 0$, we have that $\hat{w}_i^T Q^T X_b X_b^T Q \hat{w}_i = 0$. But this means $w_i^T w_i = 0$, which is a contradiction. $\square$

The following proposition is used in a derivation for the LETKF.

Proposition B.2.5. *Let $X_b, u_1, \dots, u_l$ and Q be as in proposition B.2.2. Then:*

$$w^T X_b^T Q (Q^T X_b X_b^T Q)^{-1} Q^T X_b w = (w_{||})^T w_{||}.$$

Proof. We can write w uniquely as $w = w_{||} + w_{\perp}$, where $w_{\perp}$ is in projection of w into the null space of X_b , and $w_{||}$ is the projection into the orthogonal complement of the null space of X_b . Then from proposition B.2.4 we can write $w_{||} = X_b^T Q \hat{w}$. So we have:

$$\begin{aligned} w^T X_b^T Q (Q^T X_b X_b^T Q)^{-1} Q^T X_b w &= w^T X_b^T Q (Q^T X_b X_b^T Q)^{-1} Q^T X_b (w_{||} + w_{\perp}) \\ &= w^T X_b^T Q (Q^T X_b X_b^T Q)^{-1} Q^T X_b w_{||} \\ &= w^T X_b^T Q (Q^T X_b X_b^T Q)^{-1} Q^T X_b X_b^T Q \hat{w} \\ &= w^T X_b^T Q \hat{w} \\ &= w^T w_{||} \\ &= (w_{||} + w_{\perp})^T w_{||} \\ &= w_{||}^T w_{||}. \end{aligned}$$

$\square$

B.3 The spectral norm

When analysing the error in our approximations, we use the spectral norm for matrices. Our presentation is adapted from Ciarlet 1989.

Definition B.3.1 (Spectral/operator norm for matrices (Ciarlet 1989, p. 23)). *Let $A \in \mathbb{R}^{n \times m}$ be a matrix. Then define the spectral norm of A as:*

$$\|A\| = \sup \left\{ \frac{\|Av\|_2}{\|v\|_2} : v \in \mathbb{R}^m, v \neq 0 \right\}.$$

We use here $\|\cdot\|_2$ to denote the standard Euclidean norm defined by $\|v\|_2 = \sqrt{\sum_{i=1}^m v_i^2}$.

Note that the two norms in the fraction

$$\frac{\|Av\|_2}{\|v\|_2},$$

are defined on different vector spaces.

The following results should be easy to derive from the definition of the spectral norm for matrices:

Proposition B.3.1. Let $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times m}$ be matrices and $v \in \mathbb{R}^m$ be a vector. Then:

$$\|Bv\|_2 \leq \|B\| \|v\|_2,$$

and

$$\|AB\| \leq \|A\| \|B\|.$$

The following result relates to the spectral norm of an inverse:

Proposition B.3.2 (Ciarlet 1989, p. 29-30). Let $E \in \mathbb{R}^{N \times N}$ be a matrix satisfying:

$$\|E\| < 1. \quad (\text{B.2})$$

Then the matrix $I_N + E$ is invertible and:

$$\|(I_N + E)^{-1}\| \leq \frac{1}{1 - \|E\|}. \quad (\text{B.3})$$

The next result is used several times in our analysis. It relies on the definition of the spectral radius of a matrix.

Definition B.3.2 (Spectral radius (Ciarlet 1989, p. 7)). Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then define the spectral radius of A as:

$$\varrho(A) = \max \left\{ |\lambda_i| : i = 1, \dots, n \right\},$$

where λ_i are the eigenvalues (real or complex, simple or multiple) of A .

Proposition B.3.3 (Ciarlet 1989, p. 27-28). Let $A \in \mathbb{R}^{n \times n}$ be any square matrix, and $\|\cdot\|$ be the spectral norm for matrices. Then we have that:

$$\varrho(A) \leq \|A\|.$$

For diagonalisable matrices, the result is exact:

Proposition B.3.4. Let $A \in \mathbb{R}^{n \times n}$ be a square diagonalisable matrix, and $\|\cdot\|$ be the spectral norm for matrices. Then we have that:

$$\varrho(A) = \|A\|.$$

B.4 The Neumann series expansion

The following result is used to simplify the inverse in the Kalman Gain:

Proposition B.4.1 (Werner 2018, p. 61). Let $A \in \mathbb{R}^{n \times n}$ be a symmetric square matrix such that all the Eigenvalues of A are less than 1 in absolute value. Then:

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Note we use here the convention $A^0 = I_n$.

B.5 Eigenvalues and eigenvectors

One property of eigenvalues that we use is:

Proposition B.5.1 (Proof by author but may be in other textbooks). Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$ be matrices.

Then all the non-zero eigenvalues of AB (without multiplicity) are eigenvalues of BA (without multiplicity) and vice-versa.

Proof. Let $v \in \mathbb{R}^n$ be an eigenvector of AB with eigenvalue $\lambda \neq 0$, so $ABv = \lambda v$. Then $0 \neq Bv \in \mathbb{R}^m$ (non-zero as $\lambda \neq 0$ is an eigenvector of BA with eigenvalue λ as:

$$(BA)Bv = B(AB)v = \lambda Bv.$$

We can swap m and n and A and B in the previous argument to get that the non-zero eigenvalues of AB and BA (without multiplicity) are exactly the same. $\square$

It is important to present clearly the spectral theorem for linear operators, as we use it several times in this thesis.

Theorem B.5.1 (Spectral theorem for matrices (Axler 2014, p. 245)). *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric square matrix. Then there exists some orthonormal basis Q and some diagonal matrix D such that:*

$$A = Q^T D Q.$$

Bibliography

- Anderson, Jeffrey and Lili Lei (2013). "Empirical localization of observation impact in ensemble kalman filters". eng. In: *Monthly weather review* 141.11, pp. 4140–4153. ISSN: 0027-0644.
- Anderson, Jeffrey L. (2016). "Reducing correlation sampling error in ensemble Kalman filter data assimilation". eng. In: *Monthly weather review* 144.3, pp. 913–925. ISSN: 0027-0644.
- Anderson, Jeffrey L. and Stephen L. Anderson (1999). "A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts". eng. In: *Monthly weather review* 127.12, pp. 2741–2758. ISSN: 0027-0644.
- Aravéquia, José A. et al. (2011). "Evaluation of a strategy for the assimilation of satellite radiance observations with the local ensemble transform Kalman filter". eng. In: *Monthly weather review* 139.6, pp. 1932–1951. ISSN: 0027-0644.
- Asch, Mark (2016). *Data assimilation : methods, algorithms, and applications*. eng. Fundamentals of algorithms ; 11. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics. ISBN: 9781611974539.
- Axler, Sheldon (2014). *Linear Algebra Done Right*. eng. 3rd ed. 2015. Undergraduate Texts in Mathematics. Cham: Springer International Publishing. ISBN: 3319110799.
- Ciarlet, Philippe G (1989). *Introduction to Numerical Linear Algebra and Optimisation*. eng. Cambridge University Press. ISBN: 1-139-17198-4.
- Diefenbach, Theresa et al. (2023). "Partial analysis increments as diagnostic for LETKF data assimilation systems". eng. In: *Quarterly journal of the Royal Meteorological Society* 149.752, pp. 740–756. ISSN: 0035-9009.
- Evensen, Geir (1994). "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics". eng. In: *Journal of Geophysical Research* 99.C5, pp. 10143–10162. ISSN: 0148-0227.
- Evensen, Geir, Femke C Vossepoel, and Peter Jan van Leeuwen (2022). *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*. eng. 1st Edition 2022. Springer Textbooks in Earth Sciences, Geography and Environment. Cham: Springer Nature. ISBN: 9783030967093.
- Fertig, Elana J. et al. (2007). "Assimilating non-local observations with a local ensemble Kalman filter". eng. In: *Tellus. Series A, Dynamic meteorology and oceanography* 59.5, pp. 719–730. ISSN: 0280-6495.
- Fortin, V. et al. (2014). "Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?" eng. In: 15.4, pp. 1708–1713. ISSN: 1525-755X.
- Gaspari, Gregory and Stephen E. Cohn (1999). "Construction of correlation functions in two and three dimensions". In: *Quarterly Journal of the Royal Meteorological Society* 125.554, 723–757. ISSN: 1477-870X.
- Geiss, Stefan et al. (2021). "Understanding the model representation of clouds based on visible and infrared satellite observations". eng. In: *Atmospheric chemistry and physics* 21.16, pp. 12273–12290. ISSN: 1680-7316.
- Houtekamer, P. L. and Herschel L. Mitchell (1998). "Data assimilation using an ensemble Kalman filter technique". eng. In: *Monthly weather review* 126.3, pp. 796–811. ISSN: 0027-0644.
- Houtekamer, P. L. and Fuqing Zhang (2016). "Review of the ensemble Kalman filter for atmospheric data assimilation". eng. In: *Monthly weather review* 144.12, pp. 4489–4532. ISSN: 0027-0644.

- Houtekamer, Peter L. et al. (2005). "Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations". eng. In: *Monthly weather review* 133.3, pp. 604–620. ISSN: 0027-0644.
- Hunt, Brian R., Eric J. Kostelich, and Istvan Szunyogh (2007). "Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter". eng. In: *Physica A* 230.1, pp. 112–126. ISSN: 0167-2789.
- Hutt, Axel et al. (2020). "Assimilation of SEVIRI Water Vapor Channels With an Ensemble Kalman Filter on the Convective Scale". eng. In: *Frontiers in earth science (Lausanne)* 8. ISSN: 2296-6463.
- Kalnay, Eugenia (2003). *Atmospheric modeling, data assimilation, and predictability*. eng. Cambridge : Cambridge University Press. ISBN: 9780511802270.
- Kotsuki, Shunji and Craig H. Bishop (2022). "Implementing Hybrid Background Error Covariance into the LETKF with Attenuation-Based Localization: Experiments with a Simplified AGCM". eng. In: *Monthly weather review* 150.1, pp. 283–302. ISSN: 0027-0644.
- Lahoz, W., B. Khatatov, and R. Menard (2010). *Data assimilation : making sense of observations*. eng. Berlin [u.a.]: Springer. ISBN: 9783540747024.
- Lei, Lili, Jeffrey L. Anderson, and Jeffrey S. Whitaker (2016). "Localizing the impact of satellite radiance observations using a global group ensemble filter". eng. In: *Journal of advances in modeling earth systems* 8.2, pp. 719–734. ISSN: 1942-2466.
- Lei, Lili, Jeffrey S. Whitaker, and Craig Bishop (2018). "Improving Assimilation of Radiance Observations by Implementing Model Space Localization in an Ensemble Kalman Filter". eng. In: *Journal of advances in modeling earth systems* 10.12, pp. 3221–3232. ISSN: 1942-2466.
- Lei, Lili et al. (2020a). "Adaptive Localization for Satellite Radiance Observations in an Ensemble Kalman Filter". eng. In: *Journal of advances in modeling earth systems* 12.8, n/a. ISSN: 1942-2466.
- (2020b). "Adaptive Localization for Satellite Radiance Observations in an Ensemble Kalman Filter". eng. In: *Journal of advances in modeling earth systems* 12.8, n/a. ISSN: 1942-2466.
- McNally, A. P. (2002). "A note on the occurrence of cloud in meteorologically sensitive areas and the implications for advanced infrared sounders". eng. In: 128.585, pp. 2551–2556. ISSN: 0035-9009.
- Mitchell, Herschel L. and P. L. Houtekamer (2000). "An adaptive ensemble Kalman filter". eng. In: *Monthly weather review* 128.2, pp. 416–433. ISSN: 0027-0644.
- Miyoshi, T and Y Sato (2007). "Assimilating Satellite Radiances with a Local Ensemble Transform Kalman Filter (LETKF) Applied to the JMA Global Model (GSM)". jpn. In: *Scientific online letters on the atmosphere* 3, pp. 37–40. ISSN: 1349-6476.
- Ott, Edward et al. (2004). "A local ensemble Kalman filter for atmospheric data assimilation". eng. In: *Tellus. Series A, Dynamic meteorology and oceanography* 56.5, pp. 415–428. ISSN: 1600-0870.
- Saunders, R., M. Matricardi, and P. Brunel (1999). "An improved fast radiative transfer model for assimilation of satellite radiance observations". In: *Quarterly Journal of the Royal Meteorological Society* 125.556, pp. 1407–1425.
- Saunders, Roger et al. (2018). "An update on the RTTOV fast radiative transfer model (currently at version 12)". eng. In: 11.7. ISSN: 1991-959X.
- Scheck, Leonhard, Martin Weissmann, and Liselotte Bach (2020). "Assimilating visible satellite images for convective-scale numerical weather prediction: A case-study". eng. In: *Quarterly journal of the Royal Meteorological Society* 146.732, pp. 3165–3186. ISSN: 0035-9009.
- Scheck, Leonhard et al. (2016). "A fast radiative transfer method for the simulation of visible satellite imagery". eng. In: 175, pp. 54–67. ISSN: 0022-4073.
- Schmetz, Johannes et al. (2002). "An Introduction To Meteosat Second Generation (MSG)". eng. In: *Bulletin of the American Meteorological Society* 83.7, pp. 977–992. ISSN: 0003-0007.
- Schraff, C. et al. (2016). "Kilometre-scale ensemble data assimilation for the COSMO model (KENDA)". eng. In: 142.696, pp. 1453–1472. ISSN: 0035-9009.
- Styan, G.P.H. (1971). *Hadamard Products and Multivariate Statistical Analysis*. McGill University Press. URL: <https://books.google.at/books?id=1sG4GwAACAAJ>.

- Werner, Dirk (2018). *Funktionalanalysis*. ger. 8th ed. 2018. Springer-Lehrbuch. Berlin, Heidelberg : Springer Berlin Heidelberg : Imprint: Springer Spektrum. ISBN: 3-662-55407-0.
- Zeng, Yuefei et al. (2019). "Representation of Model Error in Convective-Scale Data Assimilation: Additive Noise Based on Model Truncation Error". eng. In: *Journal of advances in modeling earth systems* 11.3, pp. 752–770. ISSN: 1942-2466.
- Zhang, Fuqing, Chris Snyder, and Juanzhen Sun (2004). "Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter". eng. In: *Monthly weather review* 132.5, pp. 1238–1253. ISSN: 0027-0644.
- Zhang, Yunji, David J. Stensrud, and Eugene E. Clothiaux (2021). "Benefits of the Advanced Baseline Imager (ABI) for ensemble-based analysis and prediction of severe thunderstorms". eng. In: *Monthly weather review* 149.2, pp. 313–332. ISSN: 0027-0644.
- Zängl, Günther et al. (2015). "The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core". eng. In: *Quarterly journal of the Royal Meteorological Society* 141.687, pp. 563–579. ISSN: 0035-9009.