**University of Plymouth**

**PEARL** **https://pearl.plymouth.ac.uk**

04 University of Plymouth Research Theses 01 Research Theses Main Collection

2021

# Estimating the background error for variational data assimilation of an ocean model using a binless analysis of innovations

Sampson, Lewis William Lloyd

http://hdl.handle.net/10026.1/17022

http://dx.doi.org/10.24382/455
University of Plymouth

*All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.*

**Copyright Statement**

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from this thesis and no information derived from it may be published without the author's prior consent.

UNIVERSITY OF
PLYMOUTH

# Estimating the background error for variational data assimilation of an ocean model using a binless analysis of innovations

by

**Lewis William Lloyd Sampson**

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

**DOCTOR OF PHILOSOPHY**

School of Biological and Marine Sciences

**March 2021**

# Acknowledgements

Each journey through a postgraduate research degree is usually focussed on one individual, but is never completed alone. I am no exception to this, the support, guidance, inspiration and confidence that I have received during my studies is the reason I am able to get as far as I have, and I'd be a fool not to acknowledge that.

The direct tutelage from my supervisory team is an obvious cause of my current knowledge and understanding, however in addition to this they have been encouraging and insightful throughout the course of my entire postgraduate education. Professor Georgy Shapiro has been supporting me from the very beginning, from offering me this amazing opportunity to begin with, to the painstaking and iterative process of helping me to write this thesis, I am grateful for your ideas, discussions and commitment. I also want to thank Dr Fred Wobus and Dr Xavier Francis for being crucial in my transformation into postgraduate education, showing me the key differences and future pitfalls, while also teaching many key nuggets of information that would become vital to my research. Dr Jose Maria Gonzalez Ondina has earned a special thanks and gratitude, for being willing to step into a position to help my studies at such a crucial time when the others could not. He has inspired me to experiment, imagine and learn, even with some of the most complicated theories that I've worked on, all while keeping a smile on my face and his. Although, Jose may be partly responsible for my current caffeine addiction (a big thanks to the baristas in the university cafes)! The remaining members of the Plymouth Ocean Forecasting Centre (Sally, Asif, Francesco, Murray, Marie, Ihor, Sanjay) had no required connection to my studies, but regardless they were all extremely friendly and polite helping me out in

multiple times of need. I very much enjoyed the multiple chats, coffees, Christmas parties and leaving dos that I was fortunate enough to be invited to.

I would then like to begin to extend my thanks to friends and family that have supported my education. My parents have always encouraged me to work hard and to push to succeed throughout all of my educations with unending support and a salient point to enjoy the journey regardless of the struggles. On the note of journey, I have to thank all the post-graduate students in the Marine Institute at the UoP, who have mutually suffered along side me in the pursuit of education, it was made all the better having you there! A special thanks to a few key members crucial in keeping the marbles from being lost; Diego, Nieves, Erin, Mark, Oli, Marcus, Maxine, and many others. I would also like to mention the Coastal Processes Research Group, the consistent meetings and discussions that I was invited to gave me a great glimpse into general research and development, inspiring much of the motivation to improve my own knowledge. I am grateful for this, as well as the Christmas parties, showing that academics do know how to have fun!

I would be mistaken to not thank the communities and companions I have found during my time in Plymouth, a brilliant city. Without knowing the extent to it, they have inspired me to continue on even in the toughest of times, to pick myself up and stride forward. My lifetime friend Annette has gone over and above supporting me, always willing to listen and pretend I am not making the conversation boring. She has been there to take me away from work whenever it was needed, and create great memories with our other friends from outside the university, Ben and Phil. One last thanks has to be given to the instructors and members at the Macmillan Martial Arts Academy, through them I learned the

importance of hard work, perseverance, dedication and efficiency, these things I will hold onto forever.

In the end, I have to say that this Phd exists due to the people mentioned before because they have always supported me, and I couldn't have done it without them. With one last big thanks to you, for taking the time to read my work.

*Our virtues and our failings are inseparable, like force and matter. When they separate, man is no more.*
*- Nikola Tesla*

*Adapt what is useful, reject what is useless, and add what is specifically your own.*
*- Bruce Lee*

*This book was written with 100% recycled words.*
*- Terry Pratchett*

# Author Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment

For my studies I was supervised by an evolving team due to circumstances of independent professional work for certain members. The director of studies has been Professor Georgy Shapiro for the duration of the PhD, with my original 1$^{st}$ and 2$^{nd}$ supervisors being Dr. Fred Wobus and Dr. Xavier Francis respectively. Mid-way through my research degree, Dr. Wobus moved to facilitate his professional position and continued to give support where possible but was physically limited. Dr. Jose Maria-Gonzalez Ondina became an additional supervisor to assist in Fred's absence. Further on through my studies Dr. Francis also moved away for work and Dr. Ondina became the primary supervisory presence with Dr. Wobus and Dr. Francis being available if requested.

This research was funded by the University of Plymouth Enterprise Limited (UOPEL), and more specifically via the support of the Plymouth Ocean Forecasting Centre (POFC).

The Phd was completed over 3 years of full-time research and a final year part-time while working for the POFC and later receiving a position at the Meteorological Office. During the 3 full-time years I attended multiple training opportunities at the University of Reading, Plymouth Marine Laboratory and the Met Office. I attend the Sharing Geosciences Online as part of the European Geosciences Union general assembly of 2020. The original intentions were to present our research at the conference, but due to the Covid-19 pandemic this was held entirely online with a new format.

The word count for this document is $\approx 37000$

Signature:

Date: 2021/03/26

# Abstract

**Author:** Lewis William Lloyd Sampson

**Title:** Estimating the background error for variational data assimilation of an ocean model using a binless analysis of innovations.

In order to create the cost function for variational data assimilation one needs to compute the background error covariance matrix, and this in turn requires information on the true state of the physical variables being modelled. This introduces a large difficulty which is often overcome by estimating the background error and then using this behaviour of the error to model the entire covariance during assimilation. A common approach for the estimation is the Hollingsworth and Lönnberg method, whereby the model forecasts subtracted from the observational data, known as the innovations, are manipulated using assumptions of correlation. This method requires the spatial binning of the innovation data and a curve fitting scheme that is applied to these bins.

During my research I have produced a new method based on the general idea of *"analysis of innovations"* similar to the Hollingsworth and Lönnberg. The new method is referred to as a binless analysis of innovations this is because I have removed the need for spatial binning. Instead the innovations are used at their exact latitude and longitude, with the error covariance model and form functions. The method then minimizes a norm of differences to find a best approximation to the background error covariance. This is done using an interpretation of the minimization for the norm which involves inner products and produces a solvable

series of equations.

The accuracy of our methods has been compared against an implementation of the Hollingsworth and Lönnberg method, and then assessed in multiple ways; first using qualitative analysis of the background standard deviation and length-scale ratio, secondly by the assessment of assimilated forecasts, and finally by reducing the observations used in the methods to represent the affect of application to sparsely observed systems. In all cases the binless analysis of innovations is similar or better then the results from currently used methods.

The positives of our method can be summarised by being able to produce an error estimate with competitive accuracy, without the use of spatial binning and requiring less free parameters leading to a small increase to the range of applications. The additional range of applications comes from the binless methods innate flexibility towards other covariance modelling scheme, as there is no spatial restrictions we are not required to use the isotropic assumption. We have been able to use the binless analysis of innovations method for a simple anisotropic application but for any operational case this would require more research and a larger set of observations. We have included our anisotropic research to demonstrate the potential use for multi-dimensional error covariance models.

# Contents

**Author Declaration**

**Abstract**

**1 Introduction** **2**

**2 Literature Review** **11**
2.1 Satellite data and in-situ observations . . . 12
2.2 Methods of data assimilation . . . 13
2.2.1 Optimal interpolation . . . 14
2.2.2 Sequential methods . . . 19
2.2.3 Variational methods . . . 24
2.3 Estimating error covariance . . . 30
2.3.1 Hollingsworth and Lönnberg method (H-L) . . . 32
2.3.2 National Meteorological Centre method (NMC) . . . 36
2.3.3 Analysis-ensemble method (A-E) . . . 39
2.4 Optimization of the data assimilation suite . . . 41
2.4.1 Minimization of the cost function . . . 41
2.4.2 Preconditioning of the analysis process . . . 47
2.5 Summary . . . 50

**3 Methods and Materials** **56**
3.1 Introduction . . . 56
3.2 Ocean modelling with assimilation . . . 57
3.2.1 Arabian Sea 1/20 Model . . . 58
3.2.2 NEMOVar . . . 59
3.3 Current error estimation methods . . . 67
3.3.1 NMC Method . . . 68
3.3.2 H-L Method . . . 71
3.4 Justification . . . 75
3.5 Binless Analysis of Innovations (BAI) . . . 77
3.5.1 Innovation covariance statistics . . . 79
3.5.2 Modelling the covariance function . . . 81
3.5.3 Minimizing the norm . . . 82
3.5.4 Anisotropic BAI . . . 85
3.6 Summary . . . 87

**4 Results and Discussions** **90**
4.1 Introduction . . . 90
4.2 Operational methodology . . . 92
4.3 Spatial comparisons - Coarse grid . . . 95
4.4 Spatial comparisons - Model grid . . . 103
4.5 Innovation subsets . . . 108
4.6 SLA analysis . . . 115
4.7 NEMOVar assimilated runs . . . 117
4.8 Anisotropic BAI method . . . 123

**5 Conclusion and future work** **132**
5.1 Conclusions . . . 132
5.2 Future works . . . 137

**A Equivalence in minimizing the norm** **145**

**B Kolmogorov's strong law of large numbers** **148**

**C Rose and Cylc** **150**

# List of Figures

1.1 Schematic of how DA adds value to observational and model information. The data shown are various representations of the ozone distribution at 10 hPa (∼30 km) on 23 September 2002, each of which has errors [Lahoz and Schneider, 2014]. . . . . 3

2.1 The difference between the Kalman filter and 4D-VAR. The Kalman filter performs an analysis at each model time step. 4D-VAR analyses all observations within a larger assimilation window simultaneously [Holm, 2008]. . . . . 22

2.2 Schematic showing the main elements of the EnKF, as implemented during the assimilation window ($t_{n-1}$,$t_n$).The blue unfilled circles to the left represent the range of the ensemble of analyses at time $t_{n-1}$; the full blue lines represent the range of ensemble forecasts using the ensemble of analyses at $tn - 1$ as the initial states; the dashed red line represents a linear combination of the forecasts (using the red star as the initial state) used to provide the final state—the analysis, at time $t_n$. The red stars filled in yellow color represent the observations used during the assimilation window. The blue unfilled circles to the right represent the range of the ensemble of analyses at time tn used for the next assimilation window. The spread of the ensemble members represents the forecast error. Based on material in [Kalnay and Yang, 2010]. Image from [Lahoz and Schneider, 2014] 24

2.3 Schematic diagram illustrating 4D-Var. The most recent observations are marked as blue stars, the previous forecast is used as the background (black dots, the background state $x_b$ is the initial condition). This updates the initial model trajectory for the subsequent forecast (red dots), using the analysis $x_a$ as the initial condition. The box to the left identifies the special case of 3D-Var [Lahoz and Schneider, 2014]. . . . . 28

2.4 Shows a graphical representation of the difference between 3D-Var and 4D-Var, the y-axis on these graphs is a 3-dimensional model space vector. [Holm, 2008]. . . . . 29

2.5 Graphical representation created to help demonstrate the process of the H-L method. The x-axis represents separation distance and the y-axis is the innovation covariances, which in turn estimates forecast and observational error. . . . . 35

2.6 Schematic illustration showing how a perturbation analysis and forecast may be generated by perturbing the inputs to the analysis system [Fisher, 2003]. . . . . 39
2.7 Schematic illustration of the analysis-ensemble method of generating fields of background difference [Fisher, 2003]. . . . . 40
2.8 The method of steepest descent for a single iteration. (a) Starting at $[-2, -2]^T$, take a step in the direction of steepest descent of f. (b) Find the point on the intersection of these two surfaces that minimizes f. (c) This parabola is the intersection of surfaces. The bottom most point is our target. (d) The gradient at the bottommost point is orthogonal to the gradient of the previous step [Shewchuk, 1994]. . . . . 44
2.9 An illustrations of the complete method for the steepest descent, starting at $[-2. - 2]^T$ and converges at $[2, -2]^T$, [Shewchuk, 1994]. 45

3.1 AS20 ocean model domain in terms of latitude and longitude, displaying the wet cells in dark blue and land in beige. The white ocean areas are not covered by the AS20 ocean model. . . . 59
3.3 NMC operational application graphic. . . . . 69
3.4 Background error SDV estimate from the NMC method using SST model forecasts during the winter season (December-January-February). . . . . 70
3.5 A graphical representation of the H-L method using real innovation data (small green circles), binned average (large black circles) and curve fitting (red line). . . . . 74
3.6 Background error SDV for temperature, produced using the H-L methods, using observations from the winter season (December-January-February.) . . . . . 74
3.7 A graphical representation of the H-L method using innovation data (small green circles), binned average (large black circles) and curve fitting (red line). . . . . 78

4.1 The forecast error SDV on the coarse grid for AS20. Produced by either the H-L or BAI methods with SST, for each season. The third plot in each row demonstrates the BAI minus H-L, and it is important to note the unique colour bar for these which is centered around zero. . . . . 98
4.2 The short forecast error LSR on the coarse grid for AS20. Produced by either the H-L or BAI methods for SST, with four independent seasons. The third plot in each row demonstrates the BAI minus H-L, and it is important to note the unique colour bar for these which is centered around zero. . . . . 99
4.3 The forecast error SDV on the model grid for AS20. Produced by either the H-L or BAI methods, with comparison to the MO error. Using SST observations for each season. . . . . 106

4.4 The short forecast error LSR on the model grid for AS20. Produced by either the H-L or BAI methods, with comparison to the MO error. Using SST observations for each season. . . . . 107
4.5 Forecast error SDV plots after applying the error estimation methods, H-L or BAI, with a percentage of the available SST observations. Producing error for the winter season December-January-February. . . . . 110
4.6 Short forecast error LSR plots after applying the error estimation methods, H-L or BAI, with a percentage of the available SST observations. Producing error for the winter season December-January-February. . . . . 111
4.9 Annual RMSD for the error SDV and short LSR produced by either analysis of innovations method for SST. This annual result is for the year 2014, with only SST observations. . . . . 115
4.10 Annual bias for the error SDV and short LSR produced by either analysis of innovations method for SST. This annual result is for the year 2014, with only SST observations. . . . . 115
4.11 The forecast error SDV on the coarse grid for AS20. Produced by either the H-L or BAI methods for SLA, with four independent seasons. . . . . 118
4.12 The short forecast error LSR on the coarse grid for AS20. Produced by either the H-L or BAI methods for SLA, with four independent seasons. . . . . 119
4.15 The background error SDV produced using the BAI method with four 2D basis functions to include anisotropy. This plot is using the SST observations for the winter season. $V$ . . . . . 124
4.16 The background error long-short length-scale weight produced using the BAI method with four 2D basis functions to include anisotropy. This plot is using the SST observations for the winter season. $w_1$ . . . . . 124
4.17 The background error anisotropic weight produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for the winter season. $v_1$ . 126
4.18 The background error longitudinal-latitudinal length-scale weight produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for the winter season. $v_2$ . . . . . 126
4.19 The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for winter 2014. . . . . 127
4.20 The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for spring 2014. . . . . 128

4.21 The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for summer 2014. . . . . 129
4.22 The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for autumn 2014. . . . . 130

C.1 Example Cylc workflow control for AS20 NEMOVar rose suite. . . 150

# List of Tables

2.1 Summary of background error estimation methods for the error covariance matrix. . . . . 54

4.1 Number of available observations for SST and SLA per season. . . 94
4.2 Mean and root-mean-square-error for the deviations of H-L minus BAI spatial plots for SDV. (Degrees Celsius) . . . . . 102
4.3 Mean and root-mean-square-error for the deviations of H-L minus BAI spatial plots for LSR. (Degrees Celsius) . . . . . 102
4.4 Table of average RMSE for NEMOVar innovations with each season. (Degrees Celsius) . . . . . 115
4.5 Table of average mean for NEMOVar innovations with each season. (Degrees Celsius) . . . . . 115
4.6 Table of average RMSE for NEMOVar innovations with each season. (Degrees Celsius) . . . . . 122
4.7 Table of average mean for NEMOVar innovations with each season. (Degrees Celsius) . . . . . 122

# List of Abbreviations

| | |
|---|---|
| 3DIncVar | Three Dimensional Incremental Var |
| A-E | Analysis-Ensemble method |
| AS20 | Arabian Sea 1/20 degree NEMO ocean model |
| BAI | Binless Analysis of Innovations |
| BLUE | Best Linear Unbiased Estimator |
| CVT | Control Variable Transform |
| DA | Data assimilation |
| ECM | Error Covariance Matrix |
| FCM | Flexible Configuration Management |
| H-L | Hollingsworth and Lönnberg |
| HPC | High Performance Computing |
| IAU | Incremental Analysis Update |
| KF | Kalman Filtering |
| LSR | Length-Scale Ratio |
| MO | Meteorological Office |
| NEMO | Nucleus of European Modelling of the Ocean |
| NEMOVar | NEMO variational data assimilation |
| netCDF | Network Common Data Form |
| NMC | National Meteorological Centre |
| NWP | Numerical Weather Prediction |
| OI | Optimal Interpolation |
| POFC | Plymouth Ocean Forecasting Centre |
| SDV | Standard Deviation |
| SLA | Sea Level Anomaly |
| SSH | Sea Surface Height |
| SST | Sea Surface Temperature |
| UAE | United Arabian Emirates |
| UoP | University of Plymouth |
| Var | Variational data assimilation |

# Chapter 1

# Introduction

The main aim of ocean modelling is to be able to use our knowledge and understanding of the physical processes and climatology associated with oceanography to produce an optimum prediction of the true ocean state. This requires optimum inputs, well defined computations and an efficient analysis; *"An analysis is the production of an accurate image of the true state of the dynamic state at a given time, represented in a model as a collection of data [Bouttier and Courtier, 1999]."* Ocean models are reliant on the most advanced methodology to give a reliable forecast, and one of the main advancements for an effective operational system, is the use of data assimilation (DA).

DA is an approach taken within forecasts to create improved outputs using observations and *a-priori* information. Currently assimilation is used in meteorology and oceanography, but there are theoretical applications for many areas of science with statistical predictions. DA is used to calculate initial conditions for dynamic models, which provides a more accurate representation of the real life system by calculating uncertainty with available data [Bin et al., 2000].

The basic concept of DA is the combination of observation and background data with the respective errors of each, these are used as weighting factors for a cost function, see figure 1.1. When this cost function is minimized we can produce

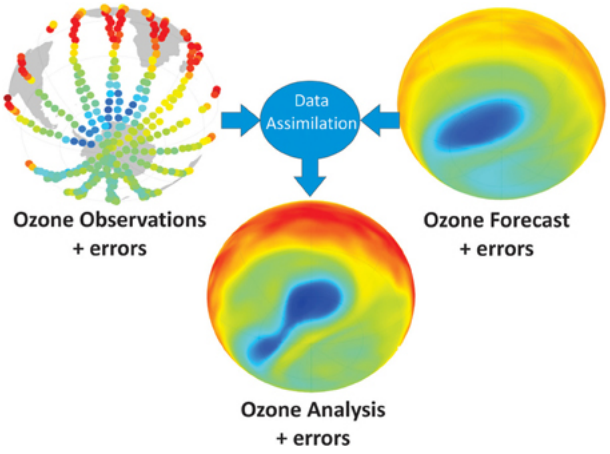



*Figure 1.1: Schematic of how DA adds value to observational and model information. The data shown are various representations of the ozone distribution at 10 hPa (∼30 km) on 23 September 2002, each of which has errors [Lahoz and Schneider, 2014].*

an improved restart file to begin a new model run and produce a more accurate forecast [Holm, 2008]. For a more complete view, DA has numerous variations to the methodology which fit different situations or different computational abilities. These variations can change core components such as the handling of observational and background data, or more subtle changes to scaling parameters within smaller equations, as a tweak to the assimilation suite definitions.

Forecast data from the model gives us background information with spatial and temporal coverage over the entire domain on its own, but is generally filled with uncertainty and larger than ideal deviations from truth. As for observational data, its not possible to get an observation for every grid point when you are looking at a potentially global domain, as will be the case for meteorological and oceanographic models [Lahoz and Schneider, 2014]. If one only used observational data, the result will be accurate information at sparse data points to then be extrapolated and cover the model area. In this case our accuracy is purely dependent on the quality of interpolation [Bouttier and Courtier, 1999], this is limited by definition and will always decrease general accuracy.

Many different methods of DA address each component of the analysis in a unique way. Two large variations for DA include sequential or variational approaches, the distinction is determined by the form of the output, which is connected to how the observations are included into the analysis [Bouttier and Courtier, 1999]. Sequential assimilation uses each observation at the time it occurred, as well as the previous observations in the assimilation window. This produces a model output that updates regularly with individual analysis increments and this series of equations is discontinuous in time. On the other hand variational data assimilation (Var) considers all observations that occur during the prescribed time window. After evaluating all of the observations, calculating observational and background error, and then minimizing the resultant cost function, the initial conditions for the analysed model are calculated [Anderson et al., 1996].

DA is used with ocean modelling to ensure that the forecast created by the model is as accurate as possible. Observations for these models are taken from the ocean via in-situ instruments for specific areas or globally for larger scale models using remote sensing. In-situ stands for *"on site"* and refers to any observations collected using an instrument in the environment. Remote sensing describes observations collected *"remotely"* from something like a satellite or an aircraft [Schowengerdt, 2006].

A popular ocean model NEMO (Nucleus of European Modelling of the Ocean), combines its model forecasts with a NEMO variational data assimilation (NEMOVar) system. Which produces forecasts for temperature, salinity, sea surface height (SSH), and wave velocity (in two latitude-longitude components). This is a package that includes many different variational DA methods (3D-Var, 4D-Var and 3D-Var FGAT) [Mogensen et al., 2012]. This then requires observations

(as a compilation of in-situ and remote sensing observations) as well as climatological information and boundary data. Using previous forecasts from NEMO, NEMOVar and the observations, an improved model output with a more accurate forecast can be produced. NEMO is the model of choice for the Plymouth Ocean Forecasting Centre (POFC) at the University of Plymouth (UoP) and is therefore the model that has been used during my research, more specifically the Arabian Sea 1/20 degree (AS20) model.

The targets for our research has been an evolving process, the original target was to review current DA and find a way to improve the state-of-the-art methods. As is typical for some PhD researchers, I would need to study the current methodology, find areas with important shortcomings, and then find solutions to these issues to improve the general understanding. What I did know about our research was that I wanted to be able to create improvements for a Var approach with a regional model. The main problem that my supervisors and I discovered inside this was the use of the isotropic assumption in the background error covariance matrix (ECM) for all operational applications of DA.

The isotropic assumption means that the background error covariance values are reliant on the 1D relative distance and assumes that there is no deterministic directional affect in the background error. This in turn led us to wonder, *(1) What is the effect of applying the isotropic assumption on the accuracy of the assimilation analysis for a regional model with active flow?*

After beginning my research to answer the previous question, I found that the addition of an anisotropic background error covariance model would require large scale changes and multiple specific research to implement operationally. This is

a topic that is of interest to many researchers, [Deckmyn and Berre, 2005],[Kucukkaraca and Fisher, 2006], [Cao et al., 2010], [Weaver and Mirouze, 2012], and there would be no easy way to give a definitive answer. In order to answer the question *(1)* I was first required to ask (and hopefully answer), *(2) What developments would be needed to implement an anisotropic background ECM into NEMOVar?* and *(3) Would the cost of developing a fully anisotropic background ECM be worth the improvement in the final analysis?*

One of the first steps that would require changes, is the error estimation process of either Hollingsworth and Lönnberg (H-L) [Hollingsworth and Lonnberg, 1986], National Meteorological Centre (NMC) [Parrish and Derber, 1992] or Analysis-Ensemble method (A-E) [Fisher, 2003]. Error covariance modelling takes place during the assimilation cycle and is initiated with some values for background error; the standard deviation (SDV) and length-scale ratio (LSR). My research into these methods has resulted in the production of a novel "binless analysis of innovations" (BAI) method, which is able to produce an anisotropic error estimation process. Once this was created I then needed to be able to give an answer to, *(4) What are the benefits and potential applications from the error estimation being replaced with an anisotropic option?*.

I have hereby posed a series of questions *(1-4)*, and have attempted to give solutions and improve the knowledge of these queries. However some of the questions posed here are too difficult and I have not been able to solve them. During this thesis I have hopefully been able to prove that I have elevated the knowledge and understanding to help provide solutions for some of the questions. With the hope that some more complete answers can be provided by bringing in previous research attempts with ours in future investigation.

In general, the aim for this research was to produce a novel method that can lead to the production of an anisotropic background ECM. This overall target required me to first research the current methodologies and then be able to reproduce them. Error estimation methods exist to produce the background error SDV and LSR, which is the ancillary component of the background ECM in operational cases. I wanted to create an analysis of innovations method for error estimation without using spatial bins, the BAI. With the principle assumptions of the general analysis of innovations and changes to statistical analysis, which uses a unique method of solving for the background error. Despite the aim being to produce an improved anisotropic representation of background error, the BAI method is also able to improve upon some of the other short-comings of the H-L analysis of innovations method.

My research lead to some public scientific documentation, the first was the *Sharing Geosciences Online* as part of the European Geosciences Union general assembly of 2020. The original intentions were to present our research at the conference, but due to the Covid-19 pandemic this was held entirely online with a new format. The online session was called: *An improved variational Data Assimilation method for ocean models with limited number of observations.* **Lewis Sampson**, Jose M. Gonzalez-Ondina, and Georgy Shapiro. The scientific concept to investigate the background error covariance of variational data assimilation methods was posed by Prof. Shapiro. Georgy also supervised the development of ideas, gave crucial feedback on early stage theory, and reviewed/edited the joint research. Dr. Ondina and I worked together on the experimental theory, refining methodology and software development of the research. In addition to what was mentioned above I was also the main contributor for the initializing

and testing of the software, creating multiple versions of the error estimation program for each methodology. Then taking the outputs of these methods into the data assimilation suite, producing the assimilation forecasts and presenting the findings. I was responsible for creating the initial conclusions with feedback from Georgy and Jose. Since I was the key author I was assigned to be the sole presenter and create the online conference presentation.

An additional research paper has been prepared for submission to the Quarterly Journal of the Royal Meteorological Society, and will be under review during the submission and examination of this thesis. The paper is also titled: *Estimating the background error for variational data assimilation using a binless analysis of innovations approach.* **Mr Lewis Sampson**, Dr Jose, M. Gonzalez-Ondina, Professor Georgy Shapiro. The work within this paper is a re-factoring of the presentation and the thesis, and the contributions for this follow the same as was EGU presentation.

The thesis is broken down into three main chapters, bookended with this introduction and a conclusion, which discusses the outcomes from our research and the possible work to follow on from here. Chapter 2 is the review of any relevant literature that has been studied during the research degree, with this record being created over the past 4 years with consistent additions and subtractions. The first section describes terminology and general knowledge for the observations that can be used in DA, section [2.1]. Then I have described the DA methods currently used with their positive and negative factors, section [2.2]. This is followed by the operational methods for producing the background ECM using error estimation approaches with the covariance modelling of NEMOVar, section [2.3]. Then finally, I have discussed some practical operations that surround the

use of the cost function, and some required optimization for operational use, section [2.4].

The third chapter of this thesis is titled Methods and Materials, and will give details for the practical uses of the DA theory and the ocean model. I have used an operational DA suite that has been compiled with the AS20 ocean NEMO model, which has a brief technical definition in section [3.2]. This is then followed with the application of both the H-L and NMC methods of error estimation for our assimilation suite, section [3.3], where I have also devoted some time to justifying our choices for the overall uses of the error estimation methods, section [3.4]. With the final part of the methods section describing the mathematical derivation and statistical justification of the Binless Analysis of Innovations (BAI) approach, section [3.5].

I have then shown the results in chapter 4, where I have used a variety of criteria to assess the ability of the BAI approach in comparison with other operational methods. The first step is a brief description of the specific set up for our experiments and the dataset used, section [4.2]. Then I begin comparing the H-L and BAI methods, first on the coarse grid, section [4.3], and then after interpolation onto the model grid, section [4.4]. The next comparison for the two methods uses a reduced observational dataset to create the error estimates, this assess their potential with smaller or less observed domains, section [4.5]. I have also included our attempt with the sea level anomaly (SLA) observations for both the H-L and BAI methods, section [4.6]. The final assessment is then on the assimilated forecasts using the error estimates as input files for NEMOVar, section [4.7]. At the end of this chapter I have shown the start of my research into anisotropic error estimation, with the initial results and findings that I have

made on this area of research, section [4.8].

This then leads into the conclusion, where the general findings of my research are presented. With references back to the original research questions that have been posed, and where my supervisors and I believe there has been answers or increased understanding. I will also discuss where future research would be required to fully investigate the possibilities of operational anisotropic DA, section [5].

# Chapter 2

# Literature Review

This literature review is a documentation of the areas I have been studying during my PhD research. The focus has primarily been on the DA methods currently in use, from my readings I have found that the field is highly populated with a lot of good mathematics and statistics and the methods are effective, but there is no such thing as a perfect. With assimilation being a computationally expensive addition to model forecasts and with many complex components, there are still areas where improvements are available, either by resulting in better representation of the errors or the faster processing of the background error covariance and minimization of the cost function. I've tried to represent both the positive and negatives on the methods where applicable, and give a general overview of my knowledge of the field.

The main topic for my literature review has been the methods of DA, in section [2.2], and the second largest topic covered is the methods of estimating error covariances, section [2.3]. With other research into the use of observations, section [2.1], and the possible optimization of operational DA, section [2.4].

## 2.1 Satellite data and in-situ observations

For any ocean model that uses DA it requires observations, these can be derived from multiple sources, the two main classifications being satellite and in-situ observations, as mentioned above. In-situ methods of measurements usually contain ground-based stations, buoys or a conductivity-temperature-depth profiler [Talley et al., 2011]. Satellite observations are taken from orbit or geostationary satellites and often referred to as remote sensing due to there being no physical contact with the observed [Schowengerdt, 2006].

"*In terms of number of observations, satellite data dominate by far the volume currently ingested by NWP DA systems*" [Thépaut, 2003]. The main providers of Earth observation satellite systems are the American, European and Japanese space agencies [Thépaut, 2003]. Due to the size of these observation datasets provided, and the ease of access, there is a preference for observations via remote sensing. More information is often considered beneficial for statistic or stochastic analysis like in DA, however there is a decrease in accuracy associated with satellite observations. Volume of observations is a good characteristics however using satellite observations is not without noticeable error and uncertainty.

Remote sensing is used in many different areas and has been much more popular in meteorology and topography than it is in oceanography [Schowengerdt, 2006]. In oceanography satellites are usually only used to measure the SSH and Sea Surface Temperature (SST), and would be difficult to accurately observe other aspects. If one wanted to model these variables on a global scale then remote sensing would be an ideal observation technique, despite remote sensing being limited to a 2D field of surface values, this is normally enough for most assimi-

lation schemes. Most assimilation schemes will create the surface layer and then extrapolate for depth using diffusion equations. However you include a possibility to create large errors, as well as generally having larger instrumental error compared to in-situ observations.

In-situ observations are usually considered to be more accurate but is a much more expensive method for gathering data of the true state. This is due to the limited amount of data each in-situ instrument can record in comparison with a satellite, increasing the cost per observation and limiting the total observations available in each dataset. The most common approach for acquiring the necessary information is to use both satellite and in-situ observations in tandem.

## 2.2 Methods of data assimilation

As mentioned before DA is the cumulation of observations (from either in-situ or satellite instruments) and background information, which will originate from previous model forecasts or knowledge of the system dynamics [Bouttier and Courtier, 1999]. There are many approaches to DA and the main characteristic that separates them is the real time assimilation or retrospective analysis more commonly referred to as variational and sequential methods of DA. Sequential assimilation only considers observations made prior to the time of analysis, where as variational (or non-sequential) methods consider all observations over the assimilation window and reduces an appropriate cost function to calculate the analysis [Bouttier and Courtier, 1999].

The main 3 methods of DA used are Optimal Interpolation (OI) [Gandin, 1965], Kalman filtering (KF) [Kalman et al., 1960] and variational methods (3D-Var and 4D-Var) [Dimet and Talagrand, 1986]. 4D-Var is the most complex of the

methods as it involves time evolution of the 3D system, this method returns the most information about the dynamic ocean with a smooth output [Bouttier and Courtier, 1999], sequential methods do have expansions that can increase the complexity and include temporal advances, but at standard it does not. As a whole, variational methods have a more favourable output due to natural smoothness of results [Lahoz and Schneider, 2014]. Since all observations are considered during the assimilation window, the new forecast created is inherently smooth but this comes at a computational cost, which will be explained properly in section [2.2.3].

### 2.2.1 Optimal interpolation

OI is considered one of the computationally cheap but relatively powerful methods of DA. Most weather centres around the world used OI throughout the 1970's and 80's [Daley, 1993]. However there has been a shift in popularity of chosen methods since then, and nowadays variational methods are more widely used. Since OI is comparatively simple its makes sense to understand OI and use this to progress ones understanding of DA.

In this section we will describe OI as it was originally founded, and how it is understood in modern terms. OI was first introduced by L.S. Gandin, during his original paper he stated *"Optimum interpolation is here understood to mean interpolation in which the mean-square interpolation error is a minimum."*. This means that the necessary condition for the approach to be optimal, is the requirement of the complete minimization of the mean-square error term [Gandin, 1965]. However this is only considered theoretical as the final result will unlikely be 100% optimal and instead is more often considered a statistical interpolation [Daley, 1993].

Gandin had a common starting point for interpolation, using a linear combination with unknown weighting coefficients to determine the desired point, an analysis $f_o$. Sometimes the interpolation used deviations by subtracting from each element of $f$ the value of its norm at the corresponding point, this resulted in a linear combination of deviations from the norm as follows.

$$f'_i = f_i - f_i^{norm} \tag{2.1}$$

$$f'_0 = \sum_{i=1}^{n} p_i f'_i \tag{2.2}$$

In these equations $f_i$ are model elements and $p_i$ are weighting factors. To solve (2.2), Gandin used the square-mean error, which also allows us to minimize this error and obtain the OI.

$$E = \overline{(f'_0 - \sum_{i=1}^{n} p_i f'_i)^2} \tag{2.3}$$

Before Gandin minimized the error, he expanded the RHS of the equation, and added the auto correlation terms.

$$E = m_{00} - 2\sum_{i=1}^{n} p_i m_{0i} + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j m_{ij} \tag{2.4}$$

$m_{ij}$ represents the autocorrelation function of element f, this describes how grid point $i$ is connected to grid point $j$ for a specific parameter. This function has useful characteristics if certain assumptions are made (these assumptions apply to most current assimilation schemes), the assumptions of isotropy and homogeneity. Isotropy means that direction is unimportant for correlation, and homogeneity means that if the entire system is rotated the correlation values will

remain the same, mathematically speaking;

$$m_{0i} = m_f(|\vec{r}_i - \vec{r}_0|) \tag{2.5}$$

$$m_{ji} = m_f(|\vec{r}_i - \vec{r}_j|) \tag{2.6}$$

$$m_{00} = m_f(0) = \overline{f^{2\prime}} \tag{2.7}$$

Gandin then normalizes the autocorrelation function in (2.4) to become a dimensionless quantity.

$$\varepsilon = \frac{E}{m_{00}} \tag{2.8}$$

$$\varepsilon = 1 - 2\sum_{i=1}^{n} p_i \mu_{0i} + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \mu_{ij} \tag{2.9}$$

At this point it is important to notice that with the change to a dimensionless quantity we now have a different notation and associated name. $\varepsilon$ is known as the interpolation error, where as E is the mean-square error. In order to minimize this we find the partial derivative with respect to $p_i$.

$$\frac{\partial \varepsilon}{\partial p_i} = -2\mu_{0i} + 2\sum_{j=1}^{n} p_j \mu_{ij} = 0 \tag{2.10}$$

and hence

$$\mu_{0i} = \sum_{j=1}^{n} p_j \mu_{ij} \tag{2.11}$$

Gandin used (2.11) to then calculate values for a system of equations from i=1..n for the weighting coefficients which are then used in equation (2.2) to determine

the deviations at the chosen point ($f_0'$). The above equations and algorithm for OI are taken from [Gandin, 1965]

In more recent literature OI is described differently due to changes in the algorithm and in notations, I have included this because it is important to be able to understand how OI compares (in processes) to modern methods.

[Ide et al., 1997] describes the general notation used within DA in the modern day, starting with a gridded field of the state vector, $\mathbf{x}$, one can interpolate the field at the location of the observations with the observation operator [Barth et al., 2008]. This operation is explained by the matrix $\mathbf{H}$, when $\mathbf{H}$ is applied to $\mathbf{x}$ it returns the original field value in model space, for both time and position, $\mathbf{Hx}$.

For OI the vector $\mathbf{x^t}$ is the true state, $\mathbf{x^b}$ is the background data and $\mathbf{y^o}$ represents the observation vector [Barth et al., 2008]. A common equation to represent the errors in background and observation data is:

$$\mathbf{x^b} = \mathbf{x^t} + \eta^{\mathbf{b}} + \mathbf{b^b} \tag{2.12}$$

$$\mathbf{y^o} = \mathbf{Hx^t} + \epsilon + \mathbf{b^o} \tag{2.13}$$

Where $\eta$ is the background/forecast error, $\varepsilon$ is the error associated with our observations (via satellites, or *in-situ* instruments) and $b$ is the bias of the background or observation errors. Originally the errors were not explicitly considered, observations were assumed to be perfectly taken. However in current assimilation schemes we use the observations error covariance and bias correction. The error

covariance is a statistic term used to quantify the background and observational errors, the specifics of which are detailed in section [2.3].

Before the analysis certain assumptions are necessary with relation to the error terms. We must have some knowledge about the expected values of the observational and background error terms, we use the ECM to characterize these assumptions [Barth et al., 2008].

$$E[\eta^{\mathbf{b}}\eta^{\mathbf{b^T}}] = \mathbf{P^b} \tag{2.14}$$

$$E[\epsilon\epsilon^{\mathbf{T}}] = \mathbf{R} \tag{2.15}$$

$$E[\eta^{\mathbf{b}}\epsilon^{\mathbf{T}}] = 0 \tag{2.16}$$

This expresses the assumption that the background error is completely uncorrelated to the observation error and that $\mathbf{P^b}$ and $\mathbf{R}$ fully describes the distribution of error for background and observations respectively.

OI analysis is also known as the best linear unbiased estimator (BLUE) of the true state $x^t$. The linear unbiased combination of the background $x^b$ and the observations $y^o$ to produce an analysis vector $\mathbf{x^a}$ , :

$$\mathbf{x^a} = \mathbf{x^b} + \mathbf{K}(\mathbf{y^o} - \mathbf{H}\mathbf{x^b}) \tag{2.17}$$

This introduces a new component called Kalman gain, which largely determines the analysis from the covariances. From equation (2.17) an estimation of analysis

error is drawn using knowledge of the analysis equation.

$$\mathbf{P^a} = \mathbf{P^b} - \mathbf{KHP^b} \tag{2.18}$$

In order to produce a reliable evaluation of $\mathbf{K}$, we must find a representation with the lowest uncertainty. By minimizing this error covariance we are able to calculate an optimal gain, which is an equation for the Kalman gain with minimal error and is the BLUE [Barth et al., 2008] :

$$\mathbf{K} = \mathbf{P^b H^T}(\mathbf{HP^b H^T} + \mathbf{R})^{-\mathbf{1}} \tag{2.19}$$

The background ECM in the previous equations is a highly dimensional matrix, with approximately $10^{14}$ components. Therefore in operational optimal interpolation analysis, usually only the computation of the diagonal variance elements is done in order to maintain applicability. This topic of error covariance estimation is recurring and is discussed in detail in [2.3] as well as some other sections.

### 2.2.2 Sequential methods

As mentioned earlier, the defining factor for a sequential method is that it uses assimilation at the time an observation occurs, resulting in a real-time analysis, which is then repeated any time a new observation enters the system. Sequential methods have many options, but the common approach for any operational assimilation is the use of KF [Holm, 2008].

For an operational system the model will run, and then during the assimilation window when the next observation occurs the DA will use the current and the previous observations to perform an analysis. KF and OI have a lot of similarities when it comes to the construction and algorithms used, however a key difference

to note is that OI does not include a dynamic evolution of the model or model error through time [Holm, 2008].

The Kalman filter is summarised as follows:

The first step for KF is a forecast step from time n-1 to n using the previous analysis:

$$\mathbf{x_n^f} = \mathbf{M_{n-1}} \mathbf{x_{n-1}^a} \tag{2.20}$$

Then we are able to calculate the forecast error covariance for the current time:

$$\mathbf{P_n^f} = \mathbf{M_{n-1}} \mathbf{P_{n-1}^a} \mathbf{M_{n-1}^T} + \mathbf{Q_{n-1}} \tag{2.21}$$

The next step is to calculate the analysis and analysis ECM for time n, using the previous analysis:

$$\mathbf{x_n^a} = \mathbf{x_n^f} + \mathbf{K_n} \left[ \mathbf{y_n} - \mathbf{H_n} \mathbf{x_n^f} \right] \tag{2.22}$$

$$\mathbf{P_n^a} = \left[ \mathbf{I} - \mathbf{K_n} \mathbf{H_n} \right] \mathbf{P_n^f} \tag{2.23}$$

K is the Kalman gain and is described by:

$$\mathbf{K_n} = \mathbf{P_n^f} \mathbf{H_n^T} \left[ \mathbf{R_n} + \mathbf{H_n} \mathbf{P_n^f} \mathbf{H_n^T} \right]^{-1} \tag{2.24}$$

[Lahoz and Schneider, 2014]

The superscripts used, $\mathbf{f}$, $\mathbf{a}$ and $\mathbf{T}$ represents the forecast, analysis and transpose respectively. The subscripts describe which time step the component is at, and $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$ are the ECM for the forecast/analysis, the model and the observa-

tional data [Lahoz and Schneider, 2014].

**M** and **H** are non-linear model and observation operators respectively, **M** carries out an evolution forward in time and **H** has the same properties as in OI, interpolating from model space to the observation space [Lahoz and Schneider, 2014].

The method is described as sequential because the process is recursive. During the next analysis the system will use the forecast quantities $x^f$ and $P^f$ to form the background state $x^b$ and the background ECM $P^b$. The analysis terms in the equation will also change to include background data.

$$\mathbf{X_n^a} = \mathbf{x_n^b} + \mathbf{K_n}\left[\mathbf{y_n} - \mathbf{H_n x_n^b}\right] \tag{2.25}$$

$$\mathbf{P_n^a} = \left[\mathbf{I} - \mathbf{K_n H_n}\right]\mathbf{P_n^b} \tag{2.26}$$

$$\mathbf{K_n} = \mathbf{P_n^b H_n^T}\left[\mathbf{R_n} + \mathbf{H_n P_n^b H_n^T}\right]^{-\mathbf{1}} \tag{2.27}$$

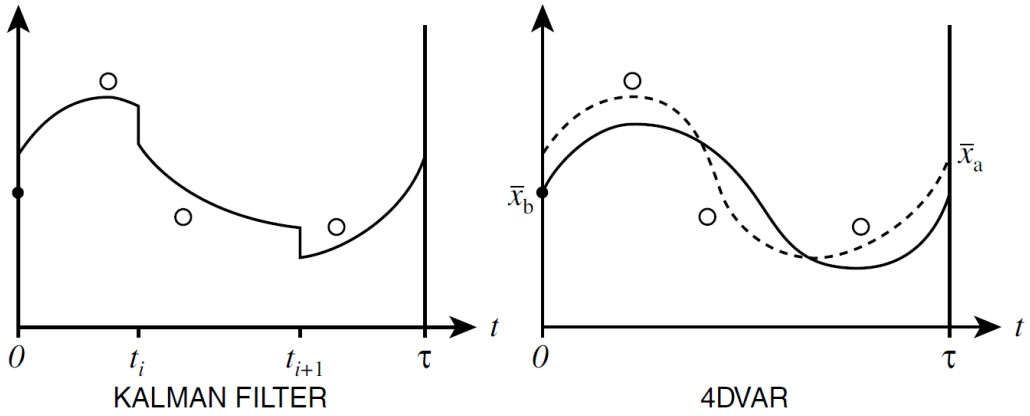



*Figure 2.1: The difference between the Kalman filter and 4D-VAR. The Kalman filter performs an analysis at each model time step. 4D-VAR analyses all observations within a larger assimilation window simultaneously [Holm, 2008].*

Nudging is a word commonly linked to DA, using the plots in figure 2.1 we can see how the analysis update can be considered a nudge for the forecast towards observations. A factor of these updates can be seen as a weakness of the KF approach by comparing the two graphs in figure 2.1. 4D-VAR is a continuous curve, whereas the KF makes a sudden jump each time an observation occurs. Since the output for KF is a sequence there are multiple discontinuities over the assimilation window. This result is generally unwanted as the known system properties don't allow for such large sudden changes, it is difficult for a model to explain a rise of 5°C in one time step without an anomaly [Holm, 2008].

There are many different approaches to KF, Extended KF (EKF), Ensemble KF (EnKF) and Smoother KF algorithms, are all improvements on the standard approach. With either more accurate error calculations or a smoothed output to better fit climatological expectations, and each with an increase to complexity by a relevant amount.

To give a brief example of one of these adaptations, the EnKF is a more statistical solution to the previously mentioned KF algorithm. EnKF creates a sample

of analyses by running the system several times for a given window, using slightly varying background data and observations [Fisher, 2001c]. Using this data an approximation for the covariance of analysis error is created reducing the need for matrix inversions, which is usually computationally expensive. We then have the covariance of background error which is used in the position of the forecast error for the next observation time and analysis [Lahoz and Schneider, 2014].

EnKF has developed into its own branch of assimilation algorithms, which aims to generate spatially and temporally varying forecast-error using Monte Carlo methods to produce ensembles [Houtekamer and Mitchell, 1998]. The variations under an umbrella of EnKF depend on their specific applications judging the different aspects of the system accordingly, and many of these families coexist in modern DA with unique use cases. Some systems will use stochastic filter for the observation errors, where as other prefer a more expensive deterministic filter to obtain the optimal gain via analysis-error covariance [Houtekamer and Zhang, 2016].

A second set of variations includes either using a sequential algorithm by assimilating the observations in sequence, or splitting the domain into a number of local areas with independent solutions. These two options are created as an approach to further reducing the numerical cost associated with the matrix inversion. The sequential option is used when the observations are considered to have entirely independent errors, whereas the local filter is more suited to a large scale application, with computational efficiency based on parallel computing architectures [Houtekamer and Zhang, 2016].

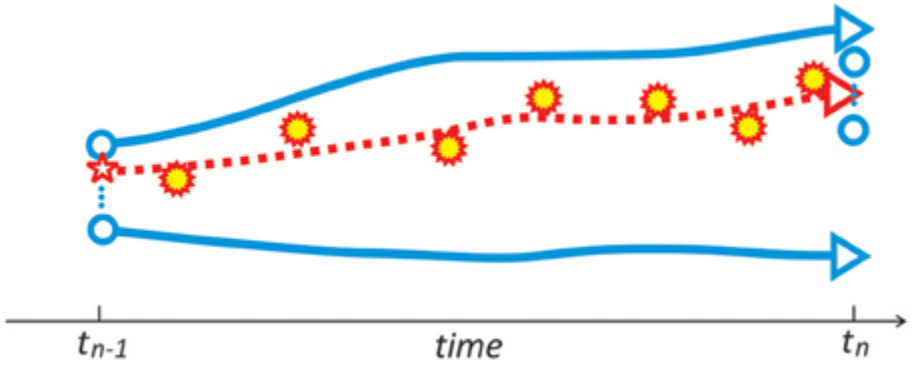



*Figure 2.2: Schematic showing the main elements of the EnKF, as implemented during the assimilation window ($t_{n-1}$,$t_n$).The blue unfilled circles to the left represent the range of the ensemble of analyses at time $t_{n-1}$; the full blue lines represent the range of ensemble forecasts using the ensemble of analyses at $tn-1$ as the initial states; the dashed red line represents a linear combination of the forecasts (using the red star as the initial state) used to provide the final state—the analysis, at time $t_n$. The red stars filled in yellow color represent the observations used during the assimilation window. The blue unfilled circles to the right represent the range of the ensemble of analyses at time tn used for the next assimilation window. The spread of the ensemble members represents the forecast error. Based on material in [Kalnay and Yang, 2010]. Image from [Lahoz and Schneider, 2014]*

### 2.2.3 Variational methods

For any assimilation that does not use real-time analysis, and instead uses all observations within a time period, it is referred to as a variational method. The frame work for the cost function used in variational assimilation stems from the use of Bayes' theorem as well as Gaussian statistics for probability. Bayes's rule is defined by Merriam-Webster as, "*The probability that an event $x$ occurs given that another event $y$ has already occurred is equal to the probability that the event $y$ occurs given that $x$ has already occurred multiplied by the probability of occurrence of event $x$ and divided by the probability of occurrence of event $y$*" [Webster, 2020].

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)} \tag{2.28}$$

In terms of DA this translates to predicting the state vector $\mathbf{x}$ given that an observation $\mathbf{y}$ has been recorded, this is what forms the framework for variational

assimilation. In a more statistical nature we call $P(x|y)$ the posterior, $P(x)$ is the prior, $P(y|x)$ is the likelihood, and $P(y)$ is the normalizing constant. With a massive difficulty introduced here since we cannot know the true state vector, instead a deviation from this vector is used instead , $\mathbf{x} - \mathbf{x}_b$, where $\mathbf{x}_b$ is the model data [Bocquet, 2014]. We now assume that the deviation is distributed according to a normalised Gaussian distribution;

$$f(\mathbf{x} - \mathbf{x}_b) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-(\mathbf{x} - \mathbf{x}_b - \overline{(\mathbf{x} - \mathbf{x}_b)})^2}{2\sigma^2}\right] \tag{2.29}$$

Since we are dealing with a large state vector, and hence deviation from the state vector are also large, the variance is then interpreted as a highly dimensional ECM, $\mathbf{B}$ or $\mathbf{R}$ for background or observational error respectively. We use the ECM to represent $\sigma^2$, and hence define the prior and likelihood probabilities using Gaussian distribution;

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{B}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x^b})^{\mathbf{T}}\mathbf{B^{-1}}(\mathbf{x} - \mathbf{x^b})\right] \tag{2.30}$$

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{R}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{Hx})^{\mathbf{T}}\mathbf{R^{-1}}(\mathbf{y} - \mathbf{Hx})\right] \tag{2.31}$$

For the above equations $\mathbf{y}$ is observational data and $\mathbf{H}$ is the observation operator which maps the state vector, $\mathbf{x}$ into a vector of observation space to be comparable with $\mathbf{y}$. We have also applied assumptions for unbiased error for observations and background data [Bocquet, 2014], where;

$$\overline{\mathbf{x} - \mathbf{x}_b} = 0 \tag{2.32}$$

$$\overline{\mathbf{y} - \mathbf{Hx}} = 0 \tag{2.33}$$

According to Bayes' rule of probability (2.28), we are then able to make a proportionality equation using (2.30) and (2.31). [Bocquet, 2014]

$$P(\mathbf{x}|\mathbf{y}) \propto exp\Big[ -\frac{1}{2}((\mathbf{x}-\mathbf{x^b})^{\mathbf{T}}\mathbf{B^{-1}}(\mathbf{x}-\mathbf{b^b}) + (\mathbf{y}-\mathbf{Hx})^{\mathbf{T}}\mathbf{R^{-1}}(\mathbf{y}-\mathbf{Hx}))\Big] \tag{2.34}$$

In order to find the optimum solution, we need to calculate the maximum of the posterior, the maximum conditional probability. This is equal to the minimum of the negative logarithm applied to the posterior probability (2.34), written as a cost function of $\mathbf{x}$ [Gandin, 1965].

$$\mathbf{J}[\mathbf{x}] = (\mathbf{x}-\mathbf{x_b})^{\mathbf{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x_b}) + (\mathbf{y}-\mathbf{Hx})^{T}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{Hx}) \tag{2.35}$$

$$\nabla\mathbf{J}(\mathbf{x}) = \mathbf{2B}^{-1}(\mathbf{x}-\mathbf{x_b}) - \mathbf{2R}^{-1}(\mathbf{y}-\mathbf{Hx}) \tag{2.36}$$

$\mathbf{J}[\mathbf{x}]$ is the cost function and $\nabla\mathbf{J}[\mathbf{x}]$ is the functions gradient. The superscripts $()^{-1}$ and $()^{T}$ represents matrix inversion and the matrix transpose respectively [Ide et al., 1997]. We also remove the common multiplier of $\frac{1}{2}$, as this factor will not affect what value of $\mathbf{x}$ that minimizes $\mathbf{J}[\mathbf{x}]$.

$\mathbf{B}$ is a weight matrix for the background error, this describes statistically the difference between the priori knowledge and the true state. In order to run the assimilation we need to calculate $\mathbf{B}$ exactly, this proves difficult as we can never know the exact value of the true state of the ocean. Many approaches are used for this calculation including; The H-L method [Hollingsworth and Lonnberg, 1986], The NMC method [Parrish and Derber, 1992] and the analysis-ensemble method [Fisher, 2001b](A-E), they are explained in detail in section [2.3].

The observation ECM $\mathbf{R}$ statistically describes the difference between the observed values and the true state, created from instrument error or error in representativity. However this is easier to calculate since we know certain properties of the observation instrument, also in most DA schemes we assume that the observations are spatially uncorrelated.

The cost function $\mathbf{J[x]}$ is comprised of 2 parts. $\mathbf{J_b}$ quantifies the misfit to the background data and $\mathbf{J_o}$ is the misfit for the observations.

$$\mathbf{J_b} = (\mathbf{x} - \mathbf{x_b})^{\mathbf{T}} \mathbf{B}^{-\mathbf{1}} (\mathbf{x} - \mathbf{x_b}) \tag{2.37}$$

$$\mathbf{J_o} = (\mathbf{y} - \mathbf{Hx})^{\mathbf{T}} \mathbf{R}^{-\mathbf{1}} (\mathbf{y} - \mathbf{Hx}) \tag{2.38}$$

In order to minimize the cost function we perform several evaluations of the gradient, initially starting at $\mathbf{J[x_b]}$, using the background data, and iteratively reducing the equation till the gradient equals zero, $\nabla \mathbf{J[x_a]} = \mathbf{0}$. The value of $\mathbf{x}$ where the minimum occurs is then our definition of the analysis $\mathbf{x_a}$ [Bouttier and Courtier, 1999]. There are many different methods that approach the problem of minimizing the cost function, and it is an area that has had a lot of interest over the evolution of assimilation schemes. The primary strategies for minimizing $\mathbf{J[x]}$ are the steepest descent method and the conjugate gradient, which are briefly described in the 2.4.1.

Figure 2.3 represents a 4 dimensional variational method (4D-Var). The separate components of the cost function are demonstrated as well as the background data, which is sometimes referred to as the forecast data, since it is formed from a previous forecast. The model will run for the time of the assimilation window,

apply the assimilation and then continue with a corrected forecast which is more representative of the true state [Lahoz and Schneider, 2014].

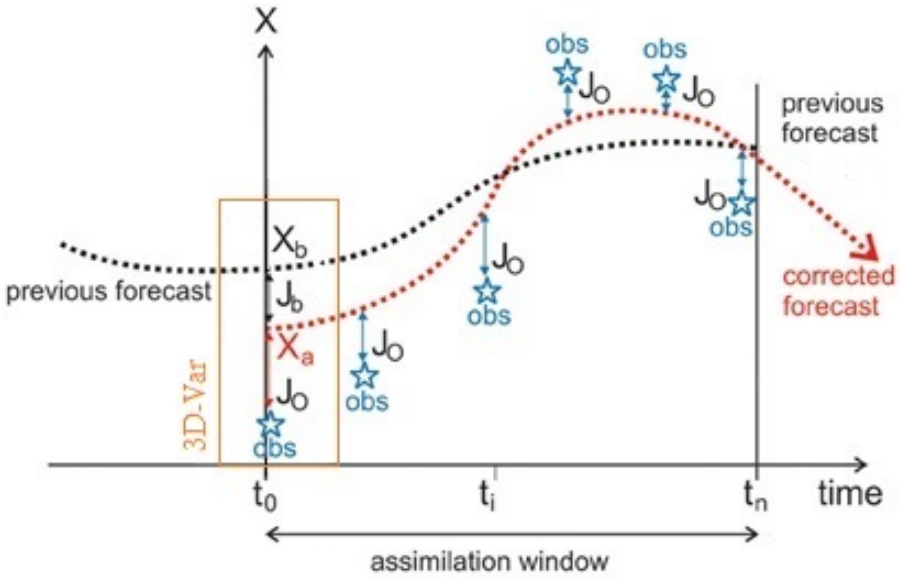



*Figure 2.3: Schematic diagram illustrating 4D-Var. The most recent observations are marked as blue stars, the previous forecast is used as the background (black dots, the background state $x_b$ is the initial condition). This updates the initial model trajectory for the subsequent forecast (red dots), using the analysis $x_a$ as the initial condition. The box to the left identifies the special case of 3D-Var [Lahoz and Schneider, 2014].*

The popularity for variational methods comes from its continuous outputs, with its relative computational ease compared with the quality of the output. 4D-Var/3D-Var also allows us to use complex observation operators, since only the operators and the adjoint's are needed [Bouttier and Courtier, 1999].

The two main forms of variational methods are 3D-Var and 4D-Var. 4D-Var includes the evolution of time within its assimilation where as 3D-Var, like OI, only applies to one time. Using figure 2.4 we can see the differences between how 3D-Var and 4D-Var interpret observations. The observations in 3D-Var are interpolated from nearby time to the average of the assimilation window.

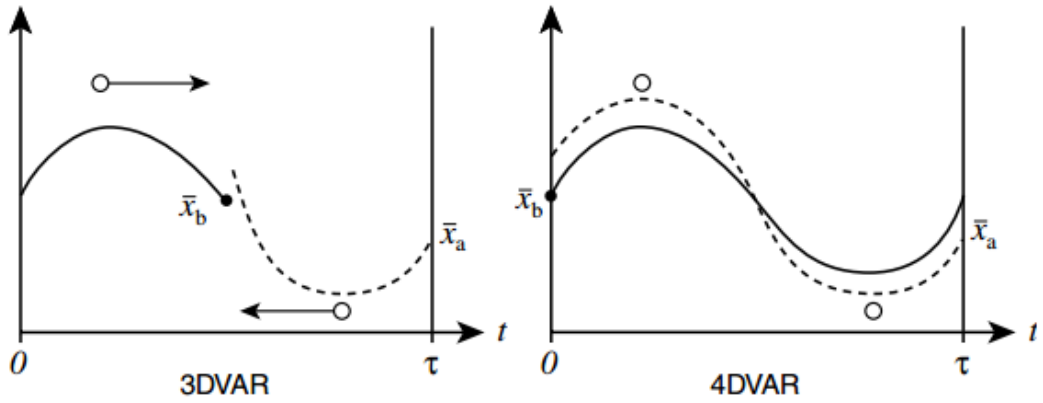



*Figure 2.4: Shows a graphical representation of the difference between 3D-Var and 4D-Var, the y-axis on these graphs is a 3-dimensional model space vector. [Holm, 2008].*

Figure 2.4 also shows a discontinuity for the 3D-Variational method, despite variational approach being a smooth assimilation method there is a single discontinuity per assimilation window. Whenever the assimilation wants to improve the forecast there will have to be an analysis update, this is an unavoidable part of the assimilation. However there is post-processing possibilities to smooth the output, or only apply the assimilation increments gradually to reduce the impact of applying to analysis [Bloom et al., 1996].

We will use the vector of $\mathbf{x_a}$ as the new initial condition for the model to then run in free mode with a singular update, similar to KF in this sense but with a single discontinuation instead of per observation.

4D-Var has an additional advantage over 3D-Var other than the propagation through time, it also considers a model error term. The cost function for 4D-Var can include an extra term for the model errors associated with the models time evolution. Typically Q is the notation used to represent the model error covariance in the cost function and the construction of this component is still

under research [Lahoz and Schneider, 2014].

## 2.3 Estimating error covariance

In DA we use the fact that the probability distribution function of an unbiased Gaussian distribution is completely described by its covariance function, in the form of a matrix [Holm, 2008]. Due to the nature of the system being predicted, using error covariances and the Gaussian assumption is necessary to determine any error statistics, and we can use the covariance matrix to represent these for the model state vector. The error covariance is described by [Weisstein, 2020] as:

$$cov(x, y) = E[(x - E[x])(y - E[y])] \tag{2.39}$$

For our use, the error covariance is a term to quantify how the error values of one grid point or parameter varies from the others. It is also worth noting that by definition the covariance for one variable with itself ($cov(x, x)$) is the variance. For DA we use an ECM where the i,j elements of the matrix are the covariance between the errors of elements $x_i$ and $y_j$ [Barth et al., 2008].

$$\mathbf{P} = E\Big[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])^T\Big] \tag{2.40}$$

There are four main covariance matrices used in DA; $\mathbf{B}$ ($\mathbf{P^b}$ or $\mathbf{P^f}$ in sequential methods) is the background or forecast ECM describing the variances of the background data to the true state. $\mathbf{R}$ is the observations background error including the effects of measurement errors, errors in design and representativeness. $\mathbf{A}$ ($\mathbf{P^a}$) represents the analysis error covariance, which is often an output of the assimilation that is used to measure the improvement of the analysis or provide an estimate of the error for the next assimilation. Then the final error

covariance that is sometimes used (whether weak or strong constraint assimilation), is $\mathbf{Q}$ the model error covariance, which describes the covariance between two errors in the model at any two locations [Holm, 2008].$\mathbf{Q}$ differs from $\mathbf{B}$, as $\mathbf{Q}$ is in charge of measuring how accurate the model propagation in time is in, rather than $\mathbf{B}$ which will predict how accurate the model is at a specific time, usually stored as a seasonal average. $\mathbf{Q}$ is more likely to be used to better understand how the model will deviate without further assimilation [Holm, 2008].

The error covariance used in DA uses an average over the time to calculate the expected value. However for most uses of the covariance in error estimation, we will not use the $E[\mathbf{x}]$ and $E[\mathbf{y}]$ terms, as we have previously assumed the Gaussian statistics to have zero mean. The individual error value for multivariate or spatial covariances, would be noted as;

$$\mathbf{B}_{12} = \langle \varepsilon_1^f \varepsilon_2^f \rangle = \frac{1}{n} \sum_{k=1}^{n} (e_{1i} - \underline{e}_1)(e_{2i} - \underline{e}_2) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \varepsilon_{1i}^f \varepsilon_{2i}^f \tag{2.41}$$

[Wikipedia, 2017]

In the above equation subscript 1/2 denotes which variable it pertains to, and an under-bar represents the sample mean.

There are many issues with using a covariance matrix for statistical weighting, the primary mathematical issue is that we cannot know the value of the true state. This means we cannot possibly calculate the error of the model/observations and hence their covariance. Current approaches to solving this problem, for the background ECM is separated into two categories; inferring error covariance from *'innovations'* $(\mathbf{y}^0 - \mathbf{H}\mathbf{x}_b)$ and creating proxies from previous forecasts and perturbations [Fisher, 2001a].

The main innovations approach was created by [Hollingsworth and Lonnberg, 1986]. The H-L method looks at the spatial covariance of differences between observations and the background, known as the innovations. Using a common assumption that the observation errors are spatially uncorrelated this method separates the observation error from the background error.

There are two mains schemes for determining background error covariance matrices by using previous forecasts and perturbations. The original is the method designed by the NMC [Parrish and Derber, 1992], this was the leading method in large operational assimilation until the analysis-ensemble method became an accessible approach. With appropriate assumptions the NMC uses the correlation of differences between 48h and 24h forecasts to predict the spatial background correlations.

The more modern and expensive approach for error covariance estimation, is the A-E method [Fisher, 2003], which involves running an ensemble of independent analysis experiments. For each experiment, the observations are perturbed by adding random noise drawn from the assumed distribution of observation error, however in comparison the computational time is much greater when you have to calculate multiple runs, 10+, for each day and is only usually implemented in operational systems that can afford the additional costs.

### 2.3.1 Hollingsworth and Lönnberg method (H-L)

The original method for determining background error covariance was provided by A. Hollingsworth and P. Lönnberg in 1986. Using radiosonde data over North America the method was applied in operational meteorological assimilation at

the ECMWF, using the homogeneous and isotropic assumptions. [Hollingsworth and Lonnberg, 1986].

The H-L method is an analysis of innovations method. This is based on calculating the difference between the observations and the background information after it has been interpolated into observation space $y_i^o - H_i x_b = d_i$ where $d_i$ is the innovation vector at observation station $i$. A core assumption for the H-L method to be possible, is that of spatially uncorrelated observation error, as well as observational and background error being fully uncorrelated [Bormann and Bauer, 2010].

$$\langle \varepsilon_{\mathbf{j}}^{\mathbf{o}} \varepsilon_{\mathbf{i}}^{\mathbf{o}} \rangle = \langle \varepsilon_{\mathbf{j}}^{\mathbf{f}} \varepsilon_{\mathbf{i}}^{\mathbf{o}} \rangle = \langle \varepsilon_{\mathbf{j}}^{\mathbf{o}} \varepsilon_{\mathbf{i}}^{\mathbf{f}} \rangle = \langle \varepsilon_{\mathbf{i}}^{\mathbf{o}} \varepsilon_{\mathbf{i}}^{\mathbf{f}} \rangle = \langle \varepsilon_{\mathbf{j}}^{\mathbf{f}} \varepsilon_{\mathbf{j}}^{\mathbf{o}} \rangle = \mathbf{0} \tag{2.42}$$

Here the angular brackets represent an averaging over time. These assumptions allow us to adapt the statistics of innovations, as a function of separation distance, in order to isolate the indicators or causes of the total error [Hollingsworth and Lonnberg, 1986]. We are able to use the innovations and manipulate the mathematics to create an estimate for the model forecast and observation errors independently, using the following:

$$\mathbf{y}^o(v, r)) = \mathbf{x}^t + \varepsilon^o(v, r) \tag{2.43}$$

$$\mathbf{x}^b(v, r) = \mathbf{x}^t + \eta(v, r) \tag{2.44}$$

$$\mathbf{y}^o(v, r) - \mathbf{x}^b(v, r) = \varepsilon^o(v, r) - \eta(v, r) \tag{2.45}$$

Equation (2.45) represents the errors for a specific variable $v$ at relative distance

$r$. The background error covariance between variable 1 at position $r$ and variable 2 at position $r + \delta r$ can be determined as a product of innovation vector. When computing this covariance the relative distance will be set to 0 ($r = 0$).

$$\langle \{\mathbf{y}^o(v_1, 0) - \mathbf{x}^b(v_1, 0)\} \times \{\mathbf{y}^o(v_2, \delta r) - \mathbf{x}^b(v_2, \delta r)\}\rangle =$$
$$\langle \epsilon^o(v_1, 0)\epsilon^o(v_2, \delta r)\rangle + \langle \eta(v_1, 0)\eta(v_2, \delta r)\}\rangle - \langle \epsilon^o(v_1, 0)\eta(v_2, \delta r)\rangle - \langle \eta(v_1, 0)\epsilon^o(v_2, \delta r)\}\rangle \tag{2.46}$$

The result is quite a long equation but by using the assumptions of correlation we can begin to simplify the equation, and result in a situation where for any case with different variables ($v_1 \neq v_2$) or for non-zero separation, the first part of the equation is assumed to be zero, and that any remaining error is cause by the background error [Bannister, 2008].

For example, when $\delta r = 0$ and $v_1 = v_2$ we simplify (2.46) to:

$$\langle \{\mathbf{y}^o(v, 0) - \mathbf{x}^b(v, 0)\} \times \{\mathbf{y}^o(v, 0) - \mathbf{x}^b(v, 0)\}\rangle = \langle \epsilon^o(v, 0), \epsilon^o(v, 0)\rangle + \langle \eta(v, 0), \eta(v, 0)\}\rangle \tag{2.47}$$

However, since the observations are spatially uncorrelated when $\delta r \neq 0$, the resulting error is purely that of background error:

$$\langle \{\mathbf{y}^o(v, \delta r) - \mathbf{x}^b(v, \delta r)\} \times \{\mathbf{y}^o(v, \delta r) - \mathbf{x}^b(v, \delta r)\}\rangle = \langle \eta(v, \delta r), \eta(v, \delta r)\}\rangle \tag{2.48}$$

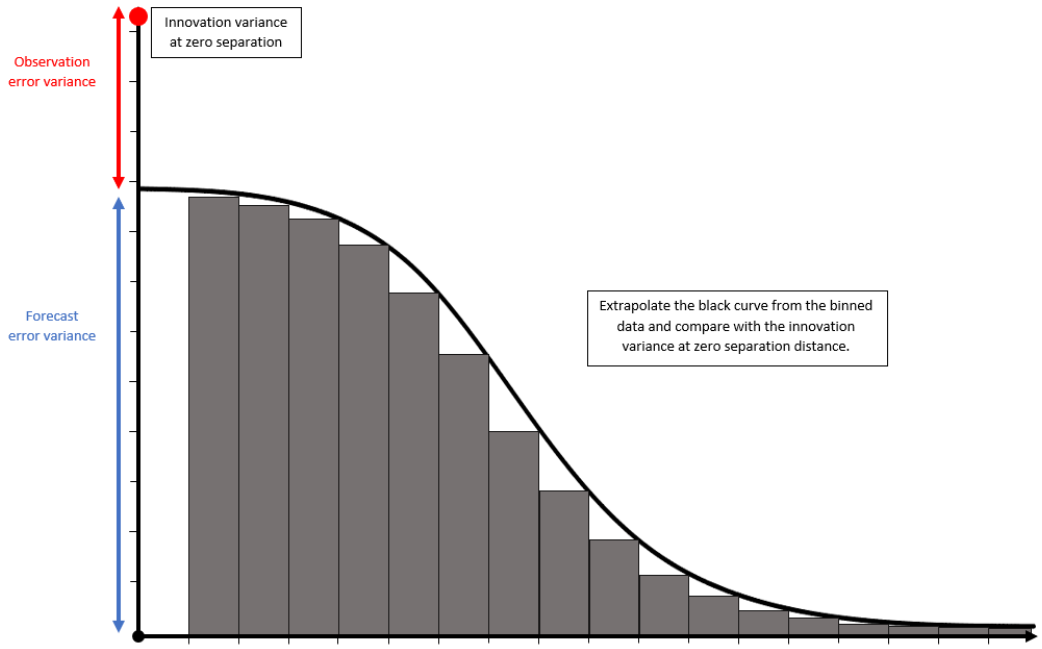



*Figure 2.5: Graphical representation created to help demonstrate the process of the H-L method. The x-axis represents separation distance and the y-axis is the innovation covariances, which in turn estimates forecast and observational error.*

Figure 2.5 represents the Hollingsworth-Lönnberg method for covariances, the addenda on the plot give some information for the components individually and how they combine together. However to further explain, the innovations are binned based on separation distance and ignoring the value at zero separation. Which innately uses the isotropic assumption, the covariance curve is then extrapolated to zero separation where the only innovation statistic has been removed. The value that the curve estimates at zero is purely based upon the forecast error, and the difference between this and the innovation statistic at zero separation is the observation error. The comments on the vertical axis help explain how the data of forecast error and observation error is represented.

For an operational system to include this approach; they begin by collecting innovations over a 2 year period, and then separate the information into seasons (winter, spring, summer, autumn), and for each season they bin the data and

apply a curve fitting scheme to the remaining covariance. The curve fitting can be a difficult part of the process due to the flexibility allowed and is often at the discretion of the scientist in charge [Bannister, 2008]. As we assume that each variable is uncorrelated in the assimilation suite it is important to remove the balanced component, in which we are referring to the climatological correlation between variables. We define the different variables within the DA suite to include correlation, typically all variables are balanced other than temperature which is treated in totality. The balanced components are calculated from knowledge of climatology and model forecasts, making assumptions about how the variables are correlated and the current state of the ocean, which in turn obtains the uncorrelated components. The effect of balanced and unbalanced error is present in all error estimation processes [Weaver et al., 2006].

This is an effective method, however it is uncertain whether it is acceptable to make the assumptions of spatially uncorrelated observations, it is specifically dubious in the case of observations like satellite radiance's where correlation or bias is possible [Lahoz and Schneider, 2014]. However, [Heilliette and Garand, 2007], and others, have used bias corrections to reduce the affect of the bias from observations, this will aim to weaken the negative affects of the assumption for uncorrelated observations. Even with the improvement H-L is still rarely used on its own in operational assimilation and is more likely to be paired with another error estimation approach.

### 2.3.2 National Meteorological Centre method (NMC)

The NMC method is a process of using model forecasts to simulate values for background error. After being published by [Parrish and Derber, 1992], the

NMC method has been implemented into some operational systems with different specifications and has the ability to produce error variances at a relatively low computational cost, the more complications added the more expensive the processes become [Waters et al., 2014].

In an operational case, the NMC method creates the error SDV using differences between 24h and 48h forecasts (which are valid at the same time) for each grid point independently. The algorithm for the NMC method is as follows [Parrish and Derber, 1992]:

$$\mathbf{x}^{48} = \mathbf{M}_{48\leftarrow 0}\mathbf{x}^{a}(t=0), \tag{2.49}$$

$$\mathbf{x}^{24} = \mathbf{M}_{48\leftarrow 24}\mathbf{x}^{a}(t=24), \tag{2.50}$$

$$\mathbf{x}^{48/24} = \mathbf{x}^{t} + \eta^{48/24} + b^{48/24}, \tag{2.51}$$

$$2\mathbf{B} = \langle(\eta\eta^{T})\rangle \tag{2.52}$$

$$\mathbf{B} \approx \frac{1}{2}\langle(\eta^{48} - \eta^{24})(\eta^{48} - \eta^{24})^{T}\rangle \tag{2.53}$$

$$\delta\mathbf{x} = \mathbf{x}^{48} - \mathbf{x}^{24} = \eta^{48} - \eta^{24} \tag{2.54}$$

$$\mathbf{B} \approx \frac{1}{2}\langle(\mathbf{x}^{48} - \mathbf{x}^{24})(\mathbf{x}^{48} - \mathbf{x}^{24})^{T}\rangle \tag{2.55}$$

The assumption is that the difference between 48h and 24h model runs, is a valid approximation to the background error variance. Which when we assume zero mean and constant bias, reflected in equation (2.51), the simplification from equation (2.53) to (2.55) becomes a stable relation. This means that we are able to extract only the background error SDV values from the model differences [Parrish and Derber, 1992].

This does not create a full background error covariance, this is only the diagonal components, the variances, of the full background error matrix, 1/nth the size of the full matrix (where n is the number of grid points). Only the diagonal components are stored to reduce the size of ancillary files, however during the assimilation the covariance can be modelled using a variety of equations. Primarily either Gaussian distributions or some form of autoregressive function is used [Martin et al., 2007]. This is the same as within the H-L method, when one needs to decide on the curve fitting scheme, they will also decide the estimated structure of the innovation covariance. However, due to the H-L method using a curve fitting scheme more weighted error correlation functions can be used as the scheme will estimate the optimum weighting as well.

Using this process creates some computational advantages, operational schemes will always contain extremely large error covariances, only storing the diagonal elements improve efficiency, and using the correlation model when it comes to assimilating without attempting to store the complete $10^{17}$ components maintains quality and makes the arithmetic's plausible.

### 2.3.3 Analysis-ensemble method (A-E)

This method was presented by [Fisher, 2003], and the basis for this method is the same as the NMC scheme, to find a surrogate quantity whose error statistics are very similar to the unknown background error. If it is safe to assume the two statistics have identical forms then we can use the surrogate to determine the correlations of background error components. In the NMC approach for this it uses a 24h and 48h forecast to create this surrogate, but it is difficult to justify that its similarity to true error is accurate.

The A-E method begins with a perturbations on every input in the analysis system, which will result in a disturbed analysis from truly random perturbations. The idea is that the analysis displacement will purely rely on the analysis error statistics. A short forecast is run from this perturbed analysis, which is the process to obtain a background field for future analysis. This component will have the statistical structure of short term forecast error, if we assume the model error itself is null [Fisher, 2003].

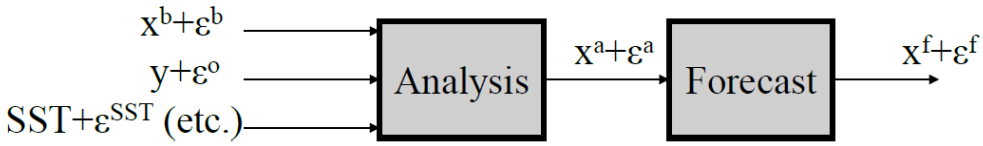



*Figure 2.6: Schematic illustration showing how a perturbation analysis and forecast may be generated by perturbing the inputs to the analysis system [Fisher, 2003].*

With this approach it is possible to use the perturbed background field for another analysis, this gives us a second perturbed analysis. A benefit of this is that after a few days (in terms of the assimilation) of repeated analysis-forecast the statistical characteristics will no longer depend on the initial perturbations that

we created at random. [Fisher, 2003] then suggests running the analysis-forecast system twice, each with statistically independent perturbations, and multiple iterations. The differences between these pairs will have the statistical structure of the differences between background error fields. They will have the correlation structure of background error, with twice the variance.

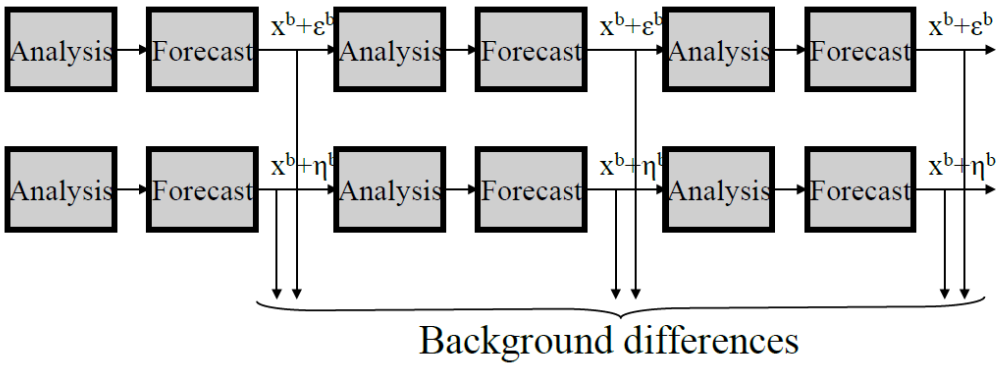



*Figure 2.7: Schematic illustration of the analysis-ensemble method of generating fields of background difference [Fisher, 2003].*

This approach gives a global field of model variables on the model grid, which is better than the output of the innovation schemes for background error covariances, specially for areas that do not have dense observations. As mentioned before it also uses a more reliable statistical assumption than the NMC methods, so it is considered the state-of-the-art method for determining background ECM, and is currently used in many operational ocean models including ECMWF (European Centre for Medium-range Weather Forecast) [Mogensen et al., 2012]. The main flaws with this scheme is the danger of feedback or the large increase in computational costs when compared to the H-L or NMC approaches. As for the feedback, there is a chance that we will introduce a bias error from the perturbations, as they are random we could possibly have an unbalanced ensemble creating an inaccurate estimate that will then compound over the numerous repetitions.

Unlike the H-L and NMC method, we have not processed the A-E method in our research, mainly due to the computational costs associated with the ensembles.

## 2.4 Optimization of the data assimilation suite

The operational application of DA has a large computational cost associated with even the simplest of assimilation suites. As an addition to an ocean model anything that does not improve the forecasts with minimal additional expense is generally unwanted. In order to make DA functionally accepted all costs must be reduced and accuracy increased. Two parts of an assimilation suite that help with this are; the optimized minimization of the cost function, and the preconditioning of the analysis process.

### 2.4.1 Minimization of the cost function

Obtaining the cost function for the 3D/4D-Var is a very expensive part of the assimilation scheme but it is not the final step for producing the necessary analysis. We need to find the minimum of the prescribed cost function. $\mathbf{J}[\mathbf{x}]$ has the possibility to be highly non-linear, with multiple minimums and maximums, this complicates finding the analysis value [Fisher, 1998].

Typically assimilation schemes will include simplifications to lower the order of $\mathbf{J}[\mathbf{x}]$, most operational assimilation uses only linear observation operators, resulting in a maximum order of two [Bouttier and Courtier, 1999]. In this section we will talk about the approach for minimizing the cost function, with linear and non-linear cost functions, as well as preconditioning methods for computational ease. It is considered logical to apply preconditioning to the minimization algorithms to increase the rate of convergence.

Mike Fisher, a prominent figure in the world of DA, published a paper in 1998, [Fisher, 1998], covering the main minimization algorithms used in operational DA. This paper focuses on solving strictly quadratic cost function, as well as a section on nearly quadratic. [Shewchuk, 1994] wrote a large article covering the minimization and preconditioning methods in great detail, as well as some operational experiments to document accuracy and applicability.

Purely quadratic cost functions are particularly important, since they appear as the functions minimized during the inner iterations of the incremental method and whenever the analysis depends linearly on the observed and background data [Fisher, 1998].

When we attempt to minimize a cost function for DA we should aim to use *conjugate gradient* or *Newton methods* for quadratic (or nearly quadratic) functions. However we need to remember we have high dimensionality as a result of the control vector, this can make storing the Hessian matrix, $\nabla \mathbf{J}[\mathbf{x}]$ close to impossible, and will require numerical methods rather than analytical methods, or potentially one would need to produce an estimated Hessian [Bouttier and Courtier, 1999].

The minimization is generally separated into two sections; direct solution methods and iterative solution methods [Fisher, 1998]. For most DA schemes direct solutions are not possible, unless the problem is split into a set of smaller problems. The separation always leads to an inaccuracy as it introduces artificial boundaries and data selection errors, however the method has been used in preconditioning DA [Fisher, 1998]. This leaves iterative solution methods as the

only successful approach for incremental variational DA.

The most popular methods for the iterative approach is the method of steepest descent, and the more accurate method of conjugate gradients [Shewchuk, 1994]. The method of steepest descent, starts at an initial condition and takes a large step, this will produce a parabola (by cutting across the 2D "dip" created by the cost function, Figure 2.8(a), we then decrease to the bottom of the paraboloid until we reach the required accuracy and take another step [Bocquet, 2014]. The direction of descent is chosen in which $f$ decreases most quickly, which is the opposite direction of $f'(x_{(i)})$, where i is the current step, starting at $f'(x_{(0)})$. [Shewchuk, 1994] describes this as $-f'(x_{(i)}) = b - Ax_{(i)}$, where $A$ and $b$ are the solutions for the minimization of $f(x) = \frac{1}{2}x^T Ax - b^T x + c$, where the ideal solution is $Ax_{(i)} = b$.

For the first step of the iterations, the direction of the steepest descent will fall on the solid line in figure 2.8(a), which will be a point following $x_{(1)} = x_{(0)} + \alpha r_{(0)}$. The issue is how large a step is needed per iteration, putting importance on the value of $\alpha$. Using calculus and minimizing directional derivatives [Shewchuk, 1994] arrives at:

$$\alpha = \frac{r_{(0)}^T r_{(0)}}{r_{(0)}^T A r_{(0)}} \tag{2.56}$$

Whereby $\alpha$ describes the size of the step with the aim that it will produce an orthogonal gradient for the next step. The best descent direction is iteratively described with the *residual*, $r_{(i)} = b - Ax_{(i)}$, and the overall algorithm of the steepest descent method becomes:

$$r_{(i)} = b - Ax_{(i)} \tag{2.57}$$

$$\alpha_{(i)} = \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}} \tag{2.58}$$

$$x_{(i+1)} = x_{(i)} + \alpha_{(i)} r_{(i)} \tag{2.59}$$

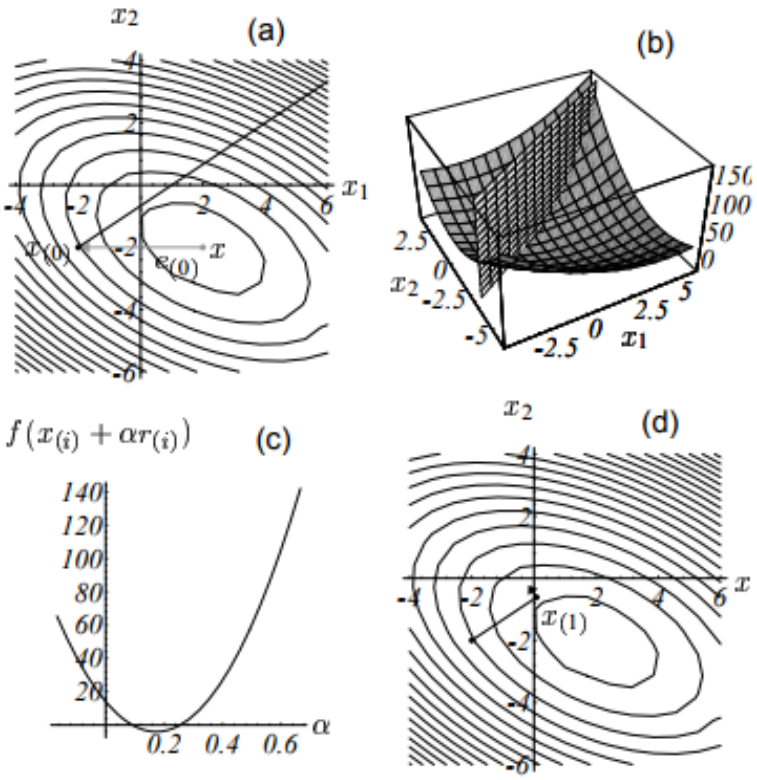



*Figure 2.8: The method of steepest descent for a single iteration. (a) Starting at $[-2, -2]^T$, take a step in the direction of steepest descent of f. (b) Find the point on the intersection of these two surfaces that minimizes f. (c) This parabola is the intersection of surfaces. The bottom most point is our target. (d) The gradient at the bottommost point is orthogonal to the gradient of the previous step [Shewchuk, 1994].*

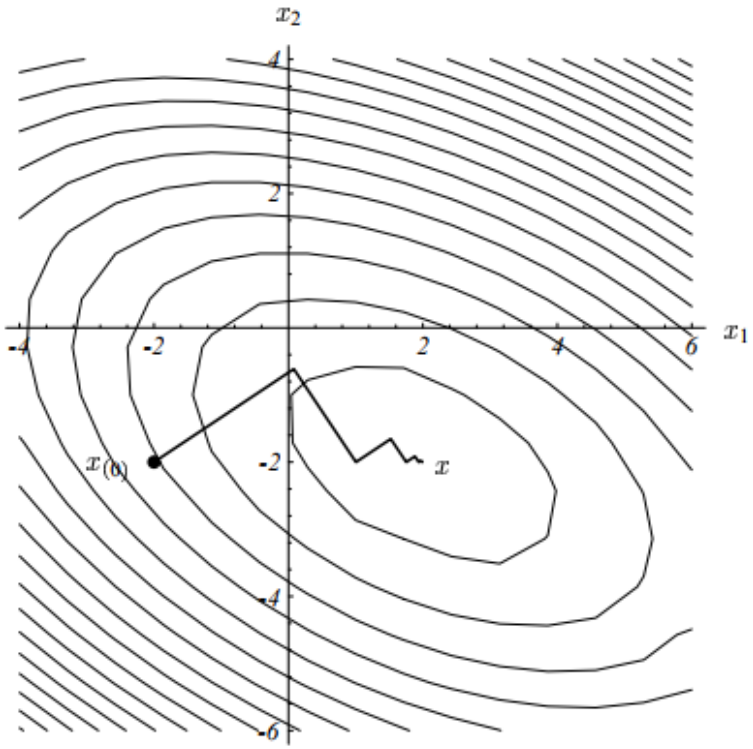



*Figure 2.9: An illustrations of the complete method for the steepest descent, starting at* $[-2. -2]^T$ *and converges at* $[2, -2]^T$*, [Shewchuk, 1994].*

The conjugate gradient algorithm is explained in-depth by [Shewchuk, 1994] and used in practice by [Fisher, 1998]. This is the method of choice when the Hessian, of J[x], is symmetric and positive definite as it usually is in the case for incremental variational DA [ECMWF, 2018].

The conjugate gradient algorithms is very similar to the steepest descent approach, but with less complexity per iteration. The direction of the first step is defined differently, instead of using the *residual* we use a search vector $d_{(i)}$. If the error was known we could define the search vector to perfectly converge to the minimum within two steps, however knowing the error would imply knowing the solution already.

In order to still apply this ideology, we say that the search directions $d_{(i)}$ are

all A-orthogonal instead of just orthogonal, if $d_{(i)}$ and $d_{(j)}$ are A-orthogonal this means $d_{(i)}Ae_{(j)} = 0$, [Shewchuk, 1994]. We then follow a very similar iterative method as before:

$$d_{(0)} = r_{(0)} = b - Ax_{(0)} \tag{2.60}$$

$$\alpha_{(i)} = \frac{r_{(i)}^T r_{(i)}}{d_{(i)}^T A d_{(i)}} \tag{2.61}$$

$$x_{(i+1)} = x_{(i)} + \alpha_{(i)} d_{(i)} \tag{2.62}$$

$$r_{(i+1)} = r_{(i)} - \alpha_{(i)} A d_{(i)} \tag{2.63}$$

$$\beta_{(i+1)} = \frac{r_{(i+1)}^T r_{(i+1)}}{r_{(i)}^T r_{(i)}} \tag{2.64}$$

$$d_{(i+1)} = r_{(i+1)} + \beta_{(i+1)} d_{(i)} \tag{2.65}$$

The derivation of the algorithm, and the Gram-Schmidt constants $\beta_{(i+1)}$ can be found in detail in [Shewchuk, 1994].

The overall algorithm is iterated for $n$ iterations, with $n$ the number of nonzero entries of $A$. This is based on the dimensions of our system or cost function, and hence for a quadratic cost function we can expect to be able to find a minimum within two iterations. This is reliant on picking the perfect polynomials and maintaining conjugacy of our search directions even in the cast of rounding point errors [Shewchuk, 1994]. This was notably an issue when the method was first proposed in 1960, but is accepted much more in modern operational sys-

tems, and is commonly the minimization of choice.

### 2.4.2 Preconditioning of the analysis process

A fast convergence is highly sought after for numerical modelling methods, like the above minimization algorithms. Using some preconditions will decrease the computation time needed for obtaining the required accuracy. As the minimization is needed for every single assimilation cycle, an increase in the convergence rate will greatly improve the overall timing for an operational assimilation. Some preconditions that are currently implemented into the assimilation cycle include; a linear transformation of the control variables in the cost function, complex conditioning of the correlation structure in the background error matrix or by measuring the Hessian matrix, of the cost function, to then assess the sensitivity of the solutions (minimization of the cost function) [Haben et al., 2011].

In operational DA minimization is very expensive with regular 3D/4D-Var, however [Courtier et al., 1994] introduced the incremental approach for assimilation. An incremental approach replaces the analysis vector within the cost function with that of an analysis increment. If we say, $\mathbf{x} = \mathbf{x}_b + \delta\mathbf{x}$, we can simplify the background part of the cost function. We then linearise the model and observation operators, to have a strictly quadratic cost function:

$$J[\delta\mathbf{x}_0] = \delta\mathbf{x}^T\mathbf{B}^{-1}\delta\mathbf{x} + \sum_{i=0}^{M}(H_i'\delta\mathbf{x}_i - \mathbf{d}_i)\mathbf{R}_i^{-1}(H_i'\delta\mathbf{x}_i - \mathbf{d}_i) \tag{2.66}$$

This linearisation allows us to more easily minimize the function, as it is the ideal condition for the minimization algorithms. We also reduce the resolution of the adjoint linear model, with these two differences the total operations of the assimilation is much less. The adjoint linear model reverses the model prop-

agation in time for the analysis increment, this is required for minimization of 4D-Var or ensemble prediction systems but has not been used in our research.

[Courtier et al., 1994] states that the incremental formulation changes time contributions of cost functions to be higher from the observations at 0 time but rapidly decreases, with the cost contributions from calculating the background components low and slowly increasing. This is because he chooses $\delta x = 0$ as the initial iteration on the minimization, so the background component is equal to zero initially. The findings show that the incremental application is not an optimal approach for reducing computational costs, in regards to a control variable transform (CVT), but the method still improves processing.

The Hessian matrix $\mathbf{A}$; $\mathbf{A} = (\mathbf{B}^{-1} + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}})$, is used as an analysis matrix for the cost function. By using the Hessian matrix we can compute a condition number and investigate its dependencies. $\mathbf{A}$ has the same structure as $\mathbf{B}$ and $\mathbf{R}$, assuming linear $\mathbf{H}$, $\mathbf{A}$ will therefore be a positive semi-definite matrix and the condition number we are seeking from $\mathbf{A}$ is:

$$k(\mathbf{A}) = ||\mathbf{A}||_2||\mathbf{A^{-1}}||_2 = \frac{\lambda_{max}(\mathbf{A})}{\lambda_{min}(\mathbf{A})} \tag{2.67}$$

This definition comes from [Haben et al., 2011], who also explains a large amount of the functionality that can be drawn from this Hessian matrix. Haben goes on to say "*We expect that the conditioning of the Hessian will be dominated by the condition number of* $\boldsymbol{B}$.". For my research this reinforces the idea that an improvement to the background error is much more likely to yield an improvement on the analysis forecast. In terms of preconditioning, the Hessian can be used to analyse the cost function and, from one cycle to the next, improve the rate of convergence by altering components. The exact specifics of this preconditioning

is mainly for 4D variational assimilation, and is not something that I explicitly used in my research.

To precondition using a change of variable, the entire cost function is expressed in terms of a CVT defined by $\delta\mathbf{x} = \mathbf{U}\mathbf{v}$. The method currently used in many operational analysis system uses the CVT to then define the background error as $\mathbf{B} = \mathbf{U}\mathbf{U}^T$. This will always be possible if we follow previous assumptions that B is defined as a semi-positive definite matrix, and hence has an adjoint available [Weaver et al., 2006]. This alters the background component of the cost function such that;

$$J_b = \frac{1}{2}\mathbf{v}^T\mathbf{U}^T\mathbf{B}^{-1}\mathbf{U}\mathbf{v} \tag{2.68}$$

$$J(\mathbf{v}) = J^b + J^o = \frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2}(\mathbf{y}^o - H(\mathbf{x}^b) - H\mathbf{U}\mathbf{v})^T\mathbf{R}^{-1}(\mathbf{y}^o - H(\mathbf{x}^b) - H\mathbf{U}\mathbf{v}) \tag{2.69}$$

The cost function is solved to determine the analysis in terms of the new variable, after which the inverse of the change of variable matrix $\mathbf{U}$ is applied, to determine the analysis in terms of model variables. Even with the preconditioning the process of minimization is time consuming relative to the other factors of the assimilation. When it comes to an operational assimilation cycle the majority of the cycle is minimization, as a large amount of the other factors required to produce the cost function can be prescribed.

## 2.5 Summary

"*DA adds value to the observations by filling in the observational gaps, and adds value to the model by constricting it with observations*" [Lahoz and Schneider, 2014]. In general DA is the process of using the observations and *a priori* data to create a forecast for a complex dynamical system. In this introduction I have shown some of the components for DA that I have been studying during the postgraduate research. These methods used with DA have influenced the path of my progression and my final research findings. For the DA methods I have found the most popular interpretations that are used or have been used in operational systems, and investigated some of the notable differences between them.

I have spent a large amount time researching the general methodologies of DA; the approaches to estimating ECM's, the minimization algorithms, and the preconditioning approaches, as well as the positives and negatives for each of these assimilation components. I have also spent some time researching the general oceanography which relates to DA, such as the ancillary climatology and the observations used within the analysis. This is only a portion of the total literature researched, and specifically only components that I believe have ended up influencing my final research outcomes. There are other aspects of DA, and specifics for operational variational suites that have not been described, such as; definitions of balance operators, covariance normalisation methods, and bias correction schemes. This is because despite being adjacent to what we have been researching we do not at any point repeat these methodologies directly and they do not have any explicit affects to the methods we have produced independently.

Sequential methods are mainly known for the popular KF method, which takes

analyses of each observation at the time they occur, with the previous observations and the relevant error covariances. KF is a recursive process and gives sequential solutions as an output, and the result can be a discontinuous curve. As KF is a real time analysis, meaning that every time a new observation becomes available the assimilation can occur. The only delay to the analysis from the real time is how long it takes to process the observations. KF is used for on-going forecasts when observations become available over time, but is less beneficial in reanalysis forecasts. This method is often used when trying to create DA suite for smaller operational systems. For large ocean modelling or numerical weather predictions (NWP) the computation expense is too much, this is due to the analysis process and the propagation of the covariance matrix [Fisher, 2001c]. There are many alterations that can be applied to the sequential methods that do improve the quality of the outputs, by smoothing the output or improving the accuracy of the error predictions. If the computational power is available KF methods can produce assimilated forecasts with accuracy matching that of a full 4D-Var suite, and has an implicit evolution of the ECM which is sometimes desired [Fisher, 2001c].

Variational methods are popular in meteorology and oceanography due to the continuous curve output from the analysis system, but also because of the way it interprets observations. When analysing a global dynamic model any process that can use the abundance of observations available is highly valuable, this is the case with any application of DA to ocean models or numerical weather prediction. Global dynamic DA can produce difficult observation operators, however fortunately for 4D-Var/3D-Var, these have the ability to compute complicated observation operators, and has multiple ways to handle any observation data. The variational assimilation methods use a minimization of the cost function,

created from background and observational data with their respective error covariances, to produce a robust and temporally smooth assimilated forecast. For operational suites variational DA often uses a incremental analysis update (IAU) to ensure that the inclusion of the assimilated forecast into the original dynamic model is relatively seamless. It has been agreed, between me and my supervisors, that due to the availability of NEMOVar (and the NEMO model) and other previously mentioned factors, that the best method for me to use during my research is the 3D-IncVar with the AS20 ocean model. This assimilation suite and model were produced by the POFC and the MO, in collaboration for a research project with the United Arabian Emirates (UAE) ocean's team.

My research has shown me many operational components and methods for ocean modelling with DA. The methods are accurate and efficient, however they are not without room for improvements. There are many assumptions made during the variational DA, which are weakly justified for dynamic systems. The focus for my work has been to create an assimilation scheme which does not assume isotropy, where our method will use the distance and direction to determine background ECM. I feel that including a horizontal component for flow-dependent background error will improve the rigidity of the scheme. This is not without difficulties, the majority of the background error covariance is created using isotropic assumptions and this would need to be replaced with functional anisotropic components.

In order to achieve this one would need to be able to create and alter the background ECM that is used in an operational assimilation. After in-depth research into DA, operational methods, and the ECM, we had decided the first step would be to replace the isotropic part within the error estimation with an

approach that has no need to assume isotropy. The second step would then be using the anisotropic error estimation with an altered background covariance flow-dependent diffusion to produce a fully anisotropic background ECM. However, there is many difficulties associated with this, mainly due to the requirements of operational assimilation systems; positive semi-definite background ECM, balance operators, and the definitions of correlation diffusion. During our applications of H-L, NMC, and the creation of the Binless Analysis of Innovations (BAI), I have been aiming to create a method that satisfies the first step, anisotropic error estimation, with the hope that this would help us progress to the second stage. In my studies I have only be able to provide an alternate analysis of innovations approach, due to the systematic restraints of NEMOVar more time and research is required to produce a fully anisotropic background ECM.

Table 2.1 summarizes the key positives and weaknesses of the error estimation methods that we have discussed throughout the literature review. Some of these points can be seen later on when I apply the H-L and NMC method to the AS20 model. Theses results from my implementations of operational methods and the notes from this literature review guided the direction of my research and will be justified in-depth in section [3.4].

Some details of the NEMOVar suite that has been use are described in the next chapter [3] for Methods and Materials. Only certain specifications have been defined, the majority of the AS20 assimilation suite has not been documented in this section since it has not directly or in-directly been altered by me during my research. Once we have defined the AS20 and NEMOVar suite, we will discuss the applications of the NMC, and H-L methods. Early experiments were run to apply these methods to our operational system, the result were used to influence

*Table 2.1: Summary of background error estimation methods for the error covariance matrix.*

| | H-L | NMC | A-E |
|---|---|---|---|
| Strengths | Accurate source data (observations).<br>Strong history of operational applications.<br>Low computational cost.<br>Able to estimate SDV and LSR.<br>Strong statistical justification for inferring background error. | More accessible data source (model forecasts).<br>Method uses only the forecast to produce estimate, no untangling of statistics.<br>Low computational cost.<br>The output is already in model grid space. | Accessible data source (model forecasts).<br>Ensemble application increases statistical accuracy.<br>Removes initial perturbations by nature of the methodology.<br>The output is already in model grid space. |
| Weakness | Reliant on assumptions of correlation.<br>Requires binning of spatial data to analyse covariance.<br>Needs large availability of data from observations. | Often recorded to underestimate the background error.<br>Unable to produce estimate of LSR without additional methods.<br>Operational applications use scaling, creating dependence on a second method. | Slower analysis due to ensembles.<br>Possibility for feedback error due to undiagnosed bias.<br>Often an inaccessible method for regional models due to increased cost versus improvement in results. |

the remainder of my research, and eventually assisting in the development of the BAI method. The methods and operational information will be included in the following section but the most of the results for these experiments will be left until chapter [4]. After this we will then summarise the results, and the general research findings in our conclusion, chapter [5].

# Chapter 3

# Methods and Materials

## 3.1 Introduction

In the previous section I have shown some of the research and theories that I have studied as part of my postgraduate degree. In this section I aim to show how I have used this knowledge, with the applications and experiments that I have conducted. This will include our initial uses of the H-L and NMC methods, the ocean model and DA that we have used, and then I will describe the novel methodology for the BAI approach.

My research supervisors are, or have been, members of the POFC, this research centre aims to provide ocean forecasts of SST and SSH with high resolution regional models, and hence they have been able to supply me with full access to the Arabian Sea 1/20 degree (AS20) NEMO ocean model. This ocean model, as well as the accompanying NEMOVar suite, is the framework for my research into DA. The assimilation was produced as a collaboration between the Meteorological Office (MO) and the POFC for their research project with the UAE ocean's team. Details for this ocean model and assimilation suite can be found in section [3.2.1] and [3.2.2].

Previously, in section [2.3] we briefly defined the theory behind both the NMC and H-L methods, we have since applied these methods to our ocean model and assimilation to produce a new error covariance estimate. This application process revealed some details that were hinted at within previous literature and the effective use of these error estimation processes, mainly for the operational uses of both the NMC and H-L. Section [3.3], will describe the differences between the theory and application of these methods, and show our initial results.

After these experiments certain decisions were made regarding the future plans for our research, we decided that focussing on using only the H-L method as a comparison for our improvements was the best idea. Our results had justified this decision, further details as well as some key results for how we came to this conclusion have been shown in section [3.4].

From here we began to work on our own novel methodology, where we have created a statistically robust alternative analysis of innovation method, which attempts to improve upon many of the known negatives of the H-L method. The derivation for the BAI method is given in section [3.5], as well as a detailed description of the fundamental differences between the H-L and the BAI. The results from these methods are then compared in the following chapter.

## 3.2 Ocean modelling with assimilation

Operational ocean modelling is often accompanied with data assimilation to take advantage of the observations and additional statistical influence that is available. In this section we will discuss the ocean model and assimilation suite that I have used for my research. These components were produced prior to the start of the research, developed to be used at the POFC for operational assimilated forecast

and not exclusively for my studies. As an operational model and assilation suite it is considered to be state of the art for context of comparing. We also discuss some of the exclusive tools and modelling procedures that are in place, this gives the reader and understanding of what was used to produce the assimilated forecasts and background error components in the later sections.

### 3.2.1 Arabian Sea 1/20 Model

The original aim for my research was to use the AS20 ocean model as a base, and then include a Var suite. Once the suite was correctly compiled, we planned to add an anisotropic component for the background error covariance, with the hope that this would improve the forecast analysis. The early steps for this were to study the DA theory, and to then get familiar with the AS20 modelling system. Since assimilation aims to use model and observational data to improve forecasting ability, its important to begin with a high quality ocean model.

The AS20 model is a high resolution regional ocean model focussed on the Arabian Sea, producing forecasts for temperature, salinity, SSH and velocity. The domain for the AS20 model covers the majority of the Arabian Sea, the Arabian Gulf, the Gulf of Oman and the Gulf of Aden. The model is fed ancillary data from rivers, tides, previous model runs and boundary data interpolated from the Indian Ocean 1/12 degree NEMO model (IND12). The model grid contains 224775 (405x555) grid points horizontally, and 56 depths levels with $\sim$ 80000 wet cells, and a resolution of $\sim$ 5 km.

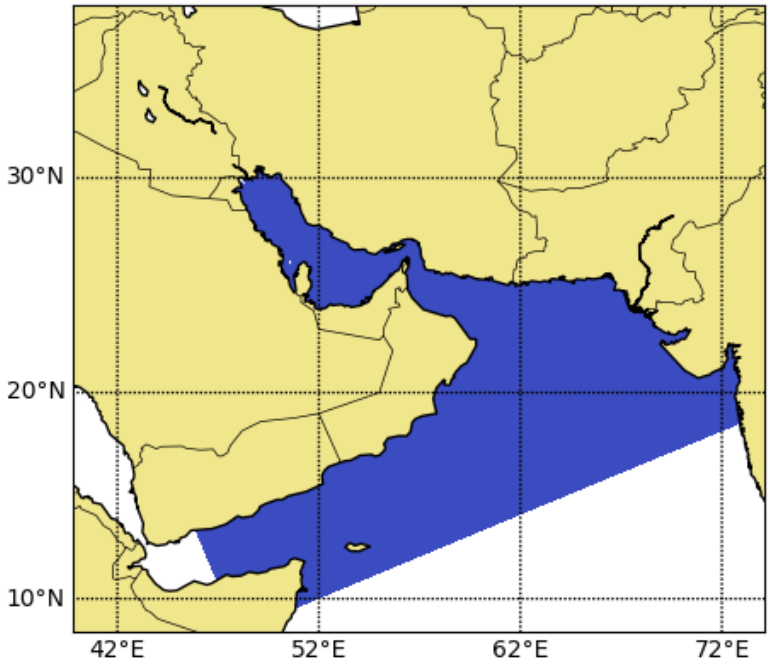



*Figure 3.1: AS20 ocean model domain in terms of latitude and longitude, displaying the wet cells in dark blue and land in beige. The white ocean areas are not covered by the AS20 ocean model.*

### 3.2.2 NEMOVar

While I was becoming familiar with the AS20 model, the POFC were beginning to work with the Meterological Office (MO) on a joint research project with the UAE ocean's team. Overtime this project lead to the production of a Var suite for the AS20 ocean model. Using interpolated ancillary files from the IND12 ocean model an initial assimilation suite was created, and eventually updated as more model and observational data became available.

The assimilation suite itself was a complied collection of Fortran 90 scripts taken from the MO's operational DA suites. This package included many options for different assimilation methods, for the AS20 a specific form was prepared to run

operationally at the UoP and also at the UAE ocean's head quarters in Abu Dhabi. This assimilation is a multivariate three-dimensional incremental Var method (3D-IncVar), using some of the aspects described in section [2.2]

3D-IncVar is commonly used for regional ocean models and operational suites which are unable to afford the computational costs of 4D DA, or those of high resolution which would not benefit enough compared to the cost associated. For high resolution models, the relative data availability is low and hence they temporal component for the observational data will have a minimal affect on the overall cost function and analysis vector.

The incremental assimilation was briefly mention in section [2.4.2], and uses a cost function as follows,

$$J[\delta\mathbf{x}] = \frac{1}{2}\delta\mathbf{x}^T\mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}(\mathbf{d} - H\delta\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{d} - H\delta\mathbf{x}). \tag{3.1}$$

Where we define the increment $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}_b$, as the difference between the state vector $\mathbf{x}$ and the model forecasts $\mathbf{x}_b$. Then the innovations, $\mathbf{d} = \mathbf{y} - H(\mathbf{x}_b)$, are the observations $\mathbf{y}$ minus the forecasts translated into observation space by the linear operator $H$.

This cost function is then minimized using an iterative preconditioned conjugate gradient method, parts of which have been mentioned in section [2.4.1]. As part of this preconditioning we apply a CVT, from section [2.4.2], this is so that the minimization requires matrix-vector calculation with $\mathbf{B}$ instead of the inverse. The CVT is $\mathbf{U}\mathbf{v} = \delta\mathbf{x}$, with $\mathbf{B} = \mathbf{U}\mathbf{U}^T$, this places a requirement on $\mathbf{B}$ to be

positive semi-definite so that $\mathbf{U}$ exists. The cost function then becomes;

$$J[\mathbf{v}] = \frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2}(\mathbf{d} - H\mathbf{U}\mathbf{v})^T\mathbf{R}^{-1}(\mathbf{d} - H\mathbf{U}\mathbf{v}). \tag{3.2}$$

Overall the incremental and control variable components are required to reduce the cost of assimilation. The operational suite also includes additional components for improving the accuracy and rigidity of the analysis vector, which includes the balance operator and the IAU. An IAU is used to introduce the analysis into the current ocean model without creating a large discontinuity between days, instead the analysis is added slowly over a period of time.

The balance operator is used within all multivariate assimilation suites, and this allows for covariances between different ocean variables to be accounted for. The balance operator $\mathbf{K}$ will transform the mutually correlated state vector, $\mathbf{x} = (T, S, \eta, u, v)$, into an unbalanced vector of uncorrelated state variables, $\mathbf{x}_U = (T, S_U, \eta_U, u_U, v_U)$ [Waters et al., 2014]. By accurately defining this operator we are able to treat each state variable independently within the assimilation and then calculate the cross-correlation component after and combine the unbalanced and balanced components. The sequence of balance operators for the incremental assimilation has been defined by [Weaver et al., 2006] as;

$$\begin{aligned}
\delta T &= \delta T, \\
\delta S &= K_{ST}\delta T + \delta S_U, \\
\delta \eta &= K_{\eta\rho}\delta\rho + \delta\eta_U, \\
\delta u &= K_{up}\delta p + \delta u_U, \\
\delta v &= K_{vp}\delta p + \delta v_U.
\end{aligned} \tag{3.3}$$

Where we have introduced two new variables, the seawater density $\rho$, and the pressure $p$. With there increments defined as;

$$\begin{aligned} \delta\rho &= K_{\rho T}\delta T + K_{\rho S}\delta S \\ \delta p &= K_{p\rho}\delta\rho + K_{p\eta}\eta. \end{aligned} \tag{3.4}$$

The balance operator itself defines the transformation from variable $j$ to $i$, $K_{i,j}$. Temperature is chosen to be the variable at the center of the balancing, and hence is treated in totality for the cross-correlations. The specific transformation from one variable to another requires climatological knowledge, [Waters et al., 2014] summarises the specific definitions for each balance operator.

In application, the balance operator is used within the production of the background ECM as so;

$$\mathbf{B} = \mathbf{K}\mathbf{B}_U\mathbf{K}^T, \tag{3.5}$$

Where $\mathbf{B}_U$ is the block diagonal matrix of unbalanced variables. Then $\mathbf{K}$ will apply the transform into full background error covariance. The unbalanced error matrix itself is produced using another series of matrix multiplication;

$$\mathbf{B}_U = \mathbf{D}^{1/2}\mathbf{C}\mathbf{D}^{1/2}. \tag{3.6}$$

Here $\mathbf{D}$ is a diagonal matrix of background error variance, and then $\mathbf{C}$ is the full correlation matrix. Instead of directly calculating the correlation matrix it is instead modelled using a normalized diffusion operator;

$$\mathbf{C}^{1/2} = \mathbf{\Gamma}^{1/2}\mathbf{L}^{1/2}\mathbf{W}^{-1/2}. \tag{3.7}$$

Where we have introduced the normalisation matrix $\mathbf{\Gamma}$, the diffusion operator $\mathbf{L}$ and the diagonal matrix of volume elements $\mathbf{W}$. The normalisation matrix ensures that the overall correlation matrix has only ones on the diagonal and decreases elsewhere. The production of this matrix is computationally expensive to calculate explicitly and instead an implicit definition from diffusion within the model is used. The estimation operationally used requires an ensemble of diffusions in a Monte Carlo statistical analysis [Weaver and Courtier, 2001]. This approach is known as the randomisation method and is relatively cheap in comparison with explicit solutions, but still takes a lot of computational power. The diffusion operator is based on a prescribed length-scale with a fixed time-step, matching that of the assimilation window. More details on how this diffusion operator is produced can be found in [Weaver and Courtier, 2001]. The volume elements represent the co-ordinate system of our ocean model, since we are using a 3D model this will be the component directions $x$, $y$, and $z$.

In order to run the operational suite, workflow and task configuration tools are used. For the NEMOVar suite we use the Rose and Cylc system, the specifics of this toolset are briefly described in the appendix C.

There are many files that the suite requires the users to supply before any assimilation can be run, and for my research the main focus of these has been the background error covariance files. Inside NEMOVar the $\mathbf{B}$ matrix is modelled using equation (3.6), which requires some value for the variance, and some length-scale for the diffusion. In practice, the diffusion uses a weighting between two fixed length-scale operators, this means that we do not need to continually

recalculate the normalisation. The diffusion operator is hence defined by;

$$\mathbf{C} = \alpha \mathbf{C}_{Mes} + (1 - \alpha)\mathbf{C}_{Syn} \tag{3.8}$$

The length-scales are chosen to represent synoptic scale and mesoscale error correlations, denoted with the subscripts $_{Syn}$ and $_{Mes}$ respectively. The user is still required to supply the LSR, $\alpha$, this is given as a 3D network common data form (netCDF) file for each season. The rose suite will then interpolate between the seasons for the current date in the assimilation. The same is done for the background error variance file, except that this is 2D, and the depth variance is parameterized during the production of the background matrix.

A varying ratio is used as opposed to changing the length-scales themselves as any change in the length-scale requires recalculation of the normalization factors. However, when changing the ratio the diffusion and hence normalisation will only be weighted by a changing scaler, the LSR. As the ratio is much easier to change, it is often used to represent seasonal changes that could originally be computed using a change of length-scale.

We first calculate the 2D surface values using SST observations from assimilated trials and then extrapolate for depths. As part of the H-L method for error estimation a curve fitting scheme is used to estimate a Gaussian with two length-scales to approximate the background error. As we are using this length-scale to model the background error with the exact same Gaussian functions, the values that the H-L uses to weight the function is stored and used as a surface layer for the ratios. In the BAI method a very similar procedure can be used to extract the LSR aswell.

Since the LSR is supplied as a 3D matrix we must also have a vertical profile, this is more difficult to approximate because we do not have the same availability of observational data to use for depth profiles. We cannot accurately estimate temperatures behaviour at depth from our observations and it is more accurate to use alternate approximations. What we want to be able to do is to understand how the correlations will change with depth and hence what changes in the LSR is an accurate representation of the error at all levels.

[Mogensen et al., 2012] talks about specifications for background error parameters in three dimensions, using model data depth profiles and affects from the mixed-layer and other oceanographic limitations, a parameterization for temperature is proposed. We use this definition to create a depth gradient for the LSR, and apply this to our surface values of LSR to create the full 3D netCDF file.

The parameterization is as follows;

$$\sigma_T^b = \begin{cases} max\{\hat{\sigma}_T^b, \hat{\sigma}_T^{ml}\} & \text{if } z \geq Z_{ml} \\ max\{\hat{\sigma}_T^b, \hat{\sigma}_T^{do}\} & \text{if } z < Z_{ml} \end{cases} \tag{3.9}$$

Where;

$$\hat{\sigma}_T^b = min\{|(\partial T^b/\partial z)\delta z|, \sigma_T^{max}\} \tag{3.10}$$

In the above, $\sigma$ is the error SDV and $Z$ denotes depths for $do$ and $ml$, the deep ocean and mixed layer respectively, with subscript $T$ for temperature. The SDV factors $\sigma_T^{max}$, $\sigma_T^{ml}$ and $\sigma_T^{do}$ are all previously defined according to research of [Mogensen et al., 2012]. $\sigma_T^{max} = 1.5°C$, $\sigma_T^{ml} = 0.5°C$, $\sigma_T^{do} = 0.07°C$.

We are then able to apply this approximation for depth using a year long assimi-

lated model forecast, this will allow us to then calculate $|(\partial T^b/\partial z)\delta z|$. The global average depth profile from this parametrisation of model data is included below. This profile was applied to the long-length, and the inverse was applied to the short length-scale. Despite the parameterization suggested above by [Mogensen et al., 2012] being for the error SDV, we believe that the gradient in depth for the error LSR will behave similarly. Since we have very few observational data points below the surface, we cannot produce innovations or estimate a depth profile from other sources, and this is one of our only options to create a depth profile. By using the same gradient file for all of our experiments we expect that any inaccuracies that this profile may introduce will be consistent for all methods, and that in general we have used an appropriate estimate for the behaviour in depth. However, if the error estimates were to be compared with externally produced versions this definition could be the cause of some discrepancies.

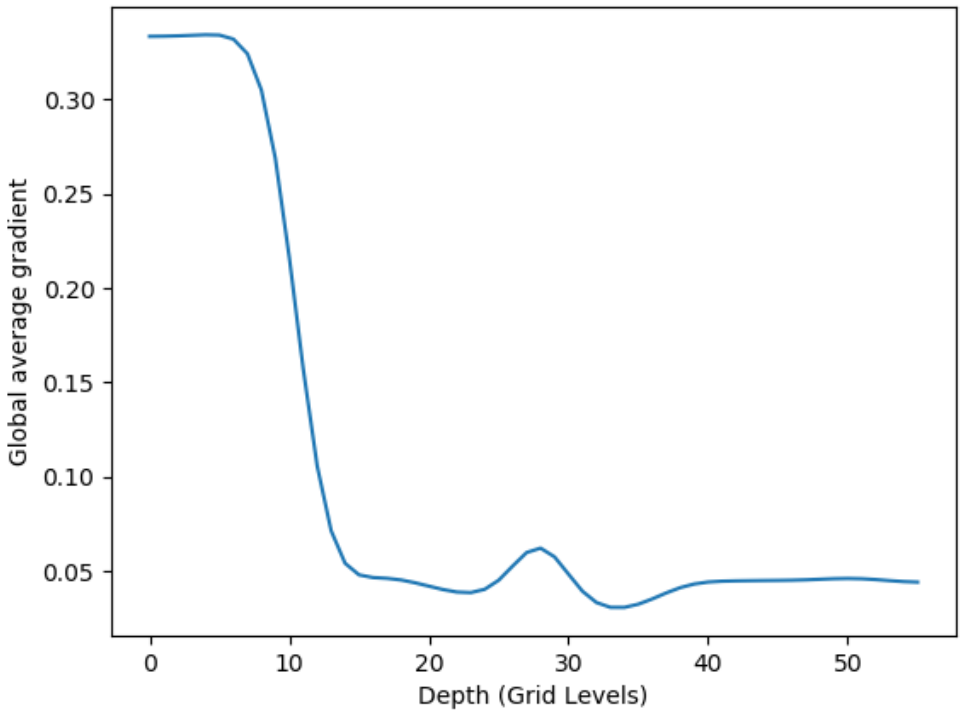



*Figure 3.2: The global average profile view of LSR parameterization in depth. This plot is for the long synoptic scale weighting parameter in winter.*

## 3.3 Current error estimation methods

In the previous section we have highlighted the processes of operational DA, and how the users are required to be able to supply NEMOVar with some netCDF files for the background error covariance.

There are many methods for estimating background error variance and LSRs, some of which have been mentioned in the previous chapter, in section [2.3]. Since my studies of these methods, I have been able to reproduce the two most popular methods, the NMC and the H-L approaches. The details of our applications have been described in the following section, with some preliminary results.

After our first practices with the two methods, we then reviewed the results and the applications. This lead to us forming decisions about our future research and the way we want to compare our new approach with previous methods. This decision and the justification has been given with our results, before we then describe how the BAI method has been developed.

### 3.3.1 NMC Method

The NMC method has been described in more detail in section [2.3.2]. I do not feel the need to recover the entire method, but the main equation that is used to calculate the background error variance is as follows;

$$\mathbf{B} \approx \frac{1}{2}\langle(\mathbf{x}^{48} - \mathbf{x}^{24})(\mathbf{x}^{48} - \mathbf{x}^{24})^T\rangle. \tag{3.11}$$

After the initial ECM had been interpolated from IND12 to our model domain, a year of assimilated AS20 forecasts has been produced by the MO. We have used these files to facilitate the NMC method. The files supplied were 2-day forecasts with hourly time steps, which could then be split into 24/48 hour model runs when needed. The basis of the NMC algorithm is that for each day in the time window, we compare the 24h and 48h forecasts and store the difference, at the end of the time window an average of the square is taken.

Figure 3.3 shows a graphical representation of the NMC algorithm, the mention of *assimilation* as the starting point is because there should be a clear distinction from assimilated model forecasts and original model forecasts, this distinction is based on whether an assimilated restart file is used. The *interpolate* flowchart box represents the process from the assimilated analysis vector into an assimilated restart file for the ocean model. The result of this process is the background

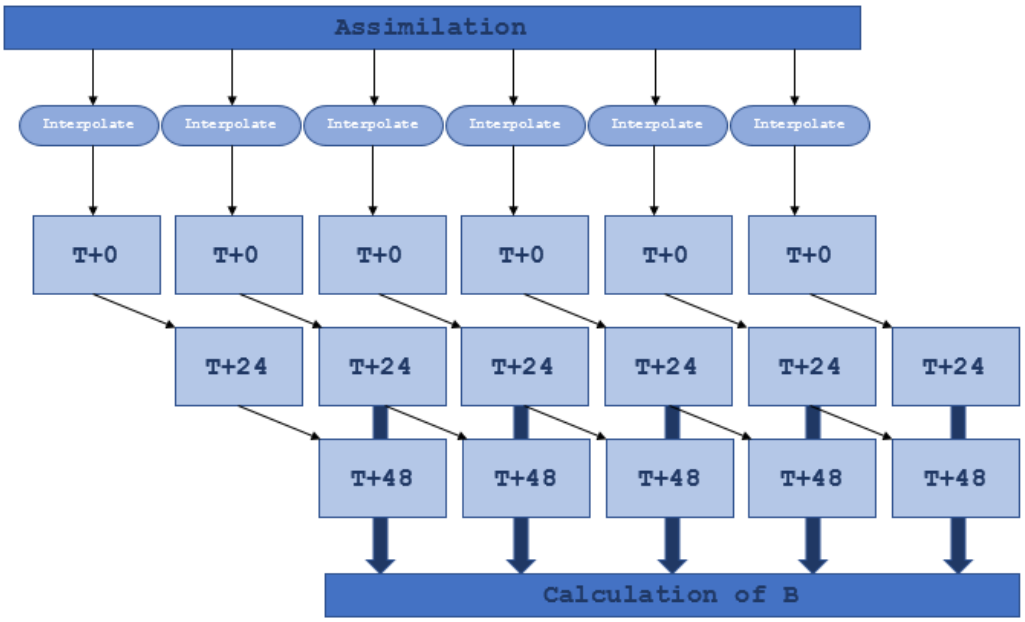



*Figure 3.3: NMC operational application graphic.*

variance in a 2D surface field over our domain.

Due to the balance operator the only variable we can do this for is sea surface temperature (SST), since the others require post-processing to remove any balanced component of the error and temperature is treated in totality. This additional step uses ocean dynamics and climatology to determine the expected correlation between variables, unfortunately I was unable to dedicate time during my research to create this. The general consensus from my supervisors was to only use temperature for any assimilation trials, we can produce background error for other variables but including them into NEMOVar is where the difficulties of balanced and unbalanced error occurs.

A common flaw of the NMC method is that it has a tendency to underestimate the background error, this is due to the increase in accuracy and resolution of ocean models the resulting difference between model runs after only one day is

reduced. Due to this trend, the NMC error variance is scaled to accommodate for the underestimation, often the H-L method can be used to determine what scaling factor the NMC needs. Since the NMC method will require scaling, and as it does not innately produce a LSR estimate it will be heavily reliant on the addition of another method for complete error estimation.

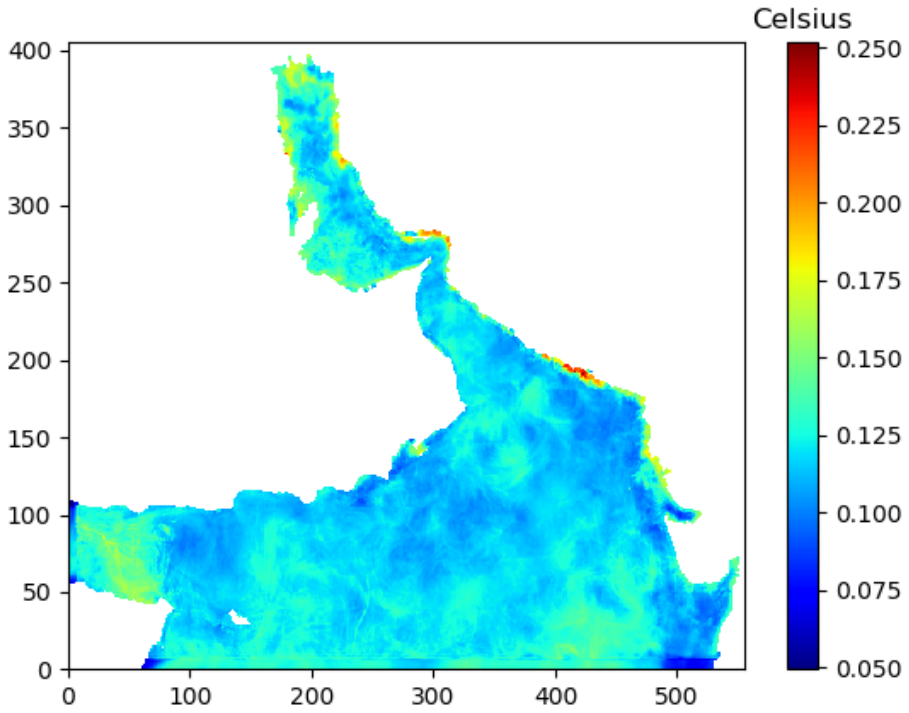



*Figure 3.4: Background error SDV estimate from the NMC method using SST model forecasts during the winter season (December-January-February).*

Figure 3.4 is an example of our applications for the NMC method, without applying any scaling factor. The range of variation for the NMC is very low, and the general average estimate is below 0.15 degrees Celsius and the maximum below 0.25. This seems to be a visualisation of the underestimate for background error that had been noted by other researchers [Bannister, 2008].

### 3.3.2 H-L Method

The H-L method has also been described in depth within the previous chapter, section [2.3.1], and this method has also been applied to the AS20 ocean model and assimilation. However, one of the main differences between the NMC and the H-L, is that the basis for this estimation is the use of innovations as opposed to model forecasts. The general assumptions associated with this method is that we are able to separate the background error and observational error since the observations are uncorrelated, and they will have zero error covariance at non-zero separation. The H-L method also requires assumptions about cross-correlation between observation and background errors, and between variables, where they are also considered to be zero. The result is that the innovations;

$$\mathbf{y}^o(v,\mathbf{r}) - \mathbf{x}^b(v,\mathbf{r}) = \mathbf{d}(v,\mathbf{r}) = \epsilon(v,\mathbf{r}) - \eta(v,\mathbf{r}), \tag{3.12}$$

Can be used to produce an equation for background error and observation error, at zero separation, and just background error at non-zero separation,

$$\overline{(\mathbf{d}(v,\mathbf{0}))(\mathbf{d}(v,\mathbf{r}))} = \begin{cases} \overline{\epsilon\epsilon} + \overline{\eta\eta} & \mathbf{r} = 0 \\ \overline{\eta(v,\mathbf{r})\eta(v,\mathbf{r})} & \mathbf{r} \neq 0 \end{cases}. \tag{3.13}$$

In the above equations, $\mathbf{r}$ is the relative position vector, which is relative to the current grid point of error estimation, and the observational error and background error are $\epsilon$ and $\eta$ respectively. We use the overbar to represent an average in time, where we are assuming ergodicity, this means the left hand side of equation (3.13) is synonymous with the *innovation covariance* since we also assume that innovations are unbiased.

$$\overline{\mathbf{d}(v,\mathbf{r})} = 0 \tag{3.14}$$

$$Cov(\mathbf{d}(v,\mathbf{0}),\mathbf{d}(v,\mathbf{r})) = \overline{\big(\mathbf{d}(v,\mathbf{0}) - \overline{\mathbf{d}(v,\mathbf{0})}\big)\big(\mathbf{d}(v,\mathbf{r}) - \overline{\mathbf{d}(v,\mathbf{r})}\big)} \tag{3.15}$$

The steps between equations (3.12) and (3.13) have been described in section [2.3.1], where we have now combined the zero-separation and non zero-separation equations in one piecewise function. These equations justify how the H-L method is able to extract the background and observational error from the innovations, however in operational applications of the method, so long as sufficient data is supplied, there is no need to do any additional steps.

For our application, the first step is to create the innovations from the assimilation feedback files. These files include observational data and model data interpolated into model space, these are one of the natural output files from a NEMOVar run. We were supplied with a year of feedback files from the AS20 assimilated model runs to produce the relevant statistics that we required.

Once we have this data we create a loop over all grid points on a coarse grid, of $\sim$30 km grid resolution and for each point calculate the innovation covariance. The innovations are then placed into 30 km bins of relative distance $r$, this is the one dimensional magnitude of their relative position. We must bin the innovations in order to be able to calculate the covariance, as the observations location changes per day and we would not have enough data consistently located to produce a robust statistic in this way.

When the innovation covariance has been made we can then apply a curve fitting scheme. In order to do this we need to supply the scheme with a function to fit the data to, for this we use an interpretation of the diffusion operator. A common

assumption is the background error covariance is accurately approximated by two Gaussian functions for mesoscale and synoptic length-scales.

$$f(r) = V\left(\alpha e^{\frac{r^2}{2L^2_{Mes}}} + (1-\alpha)e^{\frac{r^2}{2L^2_{Syn}}}\right) \tag{3.16}$$

This function is used to apply a curve fitting scheme to the innovations, but ignoring the innovation covariance from $r = 0$. Using equation (3.13) and only innovation covariances from non-zero separation, the assumption is that the function that is estimated will be done so from the background error. Then this curve is extrapolated to $r = 0$, and hence returns an estimate for background error variance and LSR, as well as the observational error variance.

This curve fitting process can be seen in figure (3.5), where the large spread of data still has an average within each bin that seems relatively consistent with the concurrent bin. This demonstrates the application of H-L and the assumptions working effectively, this can be seen as a manifestation of the central limit theorem where we may consider the innovations to be independent random variables.

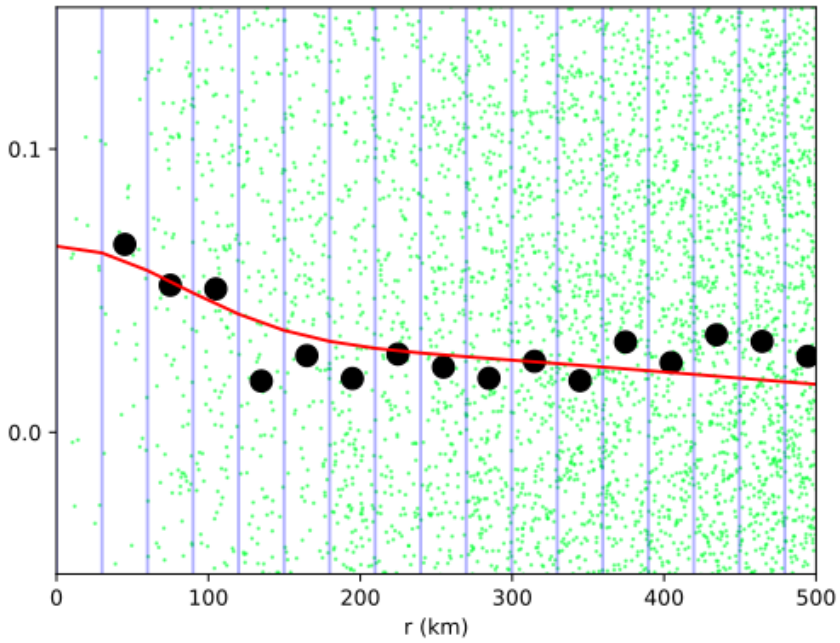



*Figure 3.5: A graphical representation of the H-L method using real innovation data (small green circles), binned average (large black circles) and curve fitting (red line).*

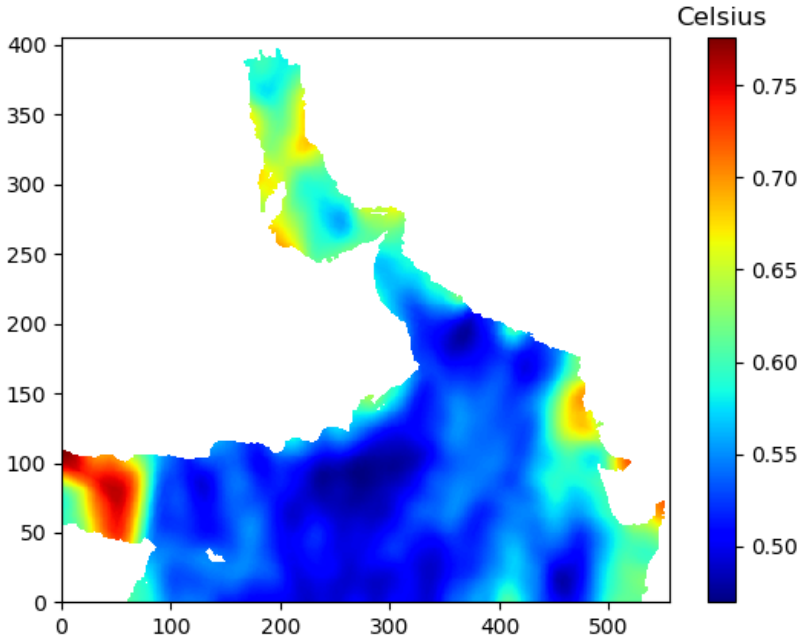



*Figure 3.6: Background error SDV for temperature, produced using the H-L methods, using observations from the winter season (December-January-February.)*

The H-L method has a noticeably higher average error SDV than the NMC method, since we do not know the true background error we do not know which is the best approximation. Following on from the literature review we do expect the H-L average to be higher than that of the NMC, as the common weakness of the NMC is to underestimate error. This may be positive evidence for the use of innovations to infer background error as a more reliable statistic.

It is important to note that this error was produced on a coarse grid, and then interpolated to this model grid using simple linear interpolation and a general smoothing filter. Despite this the SDV is able to show trends of high error and low error for different locations within the domain, and this is an expected behaviour of the background error.

## 3.4 Justification

As was mentioned in the introduction to this section, after our initial experimentation with the error estimation methods, we came to a general decision about the future of our research. Despite being able to produce the NMC error SDV, we do not think that the comparisons for this method yields enough benefit to further refine the method. Instead, for the remainder of my studies I focused on only comparing the H-L method with our novel method.

Table 2.1 in chapter 2 gives an overview of the positives and negatives of the methods in theory, but for our practical applications we had other contributing factors to our choice. The main cause of this was that we did not feel that the error estimation achieved by the NMC method was suitable for our assimilation with a high resolution regional model. The original result was very low,

suggesting that the background error from model forecasts was minuscule, and if true we would not need to apply the assimilation to improve model forecast. This would be a very unlikely situation and instead we think that the method is greatly underestimating the error. The only option for using the NMC would be to scale the result appropriately, which would require a method like H-L anyway.

Another cause for this change was due to our plan for our novel method, we were aiming to use the innovations to estimate the model error since the mathematical justification felt more grounded. By comparing one analysis of innovations method with another we are expecting to better see the benefits of our changes. If we were to compare with the NMC method some of the difference could be explained by the sources of data.

After this point in my research I continued to further develop the code for the H-L method, while working on the production of the novel method. I have optimized the code for computational cost and improved the functionality of the script by including options for the use of SLA observations and varying sizes of innovation datasets. We also made some changes to the binning procedure, as we had been binning the innovations as soon as possible for the covariance function, as this had sped up the run time, however computing $\mathbf{d}(v, \mathbf{0})\mathbf{d}(v, \mathbf{r})$ and then binning led to more accurate results. The newer version of this method is used when we mention the H-L in the results section, this is why there are differences in some of the future plots for H-L.

## 3.5 Binless Analysis of Innovations (BAI)

Now we will discuss the novel research, how we have created an analysis of innovations method for error estimation without using spatial bins, the BAI method. With the principle assumptions of the general analysis of innovations, and our changes to statistical analysis and unique method of solving for the background error. Even though the use of the H-L as an analysis of innovations method is still widely useds there are still some weaknesses; Firstly, it requires one to spatially bin the observational data, where we may be reducing the accuracy of our innovation statistics. The H-L method requires the use of bins to reduce the computational costs significantly, when using the binned data instead of the exact locations there are simplifications for the curve fitting scheme.

The H-L method also requires sufficient data in each bin to obtain a statistically robust entry for the curve fitting [Bannister, 2008]. If the there is not enough data in most of the bins, then the fitting of the model function might produce spurious values. This means we require a large amount of observational data and can make the method unstable in areas of the domain with sparse observations. The following plots show how important sufficient data is, and how outliers begin to weaken the statistical justification of the H-L when there is less innovation data.

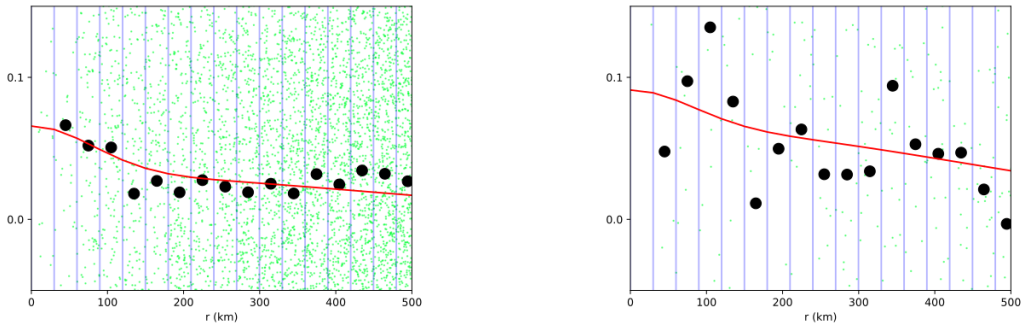



*Figure 3.7: A graphical representation of the H-L method using innovation data (small green circles), binned average (large black circles) and curve fitting (red line).*

The third weakness of the H-L method is that it is inherently one dimensional and currently there is no way of extending it to multi-dimensional anisotropic cases. This limitation is an extension of the previous two comments, even if a 2D extension for the curve fitting scheme was made, it will still require a large enough dataset to sufficiently fill the additional bins.

For the BAI method we do not use any spatial bins, and instead use the combination of inner products and basis functions to create a robust statistical analysis, which can be solved for the error variance and length-scale ratio. A specific example is used to describe the BAI method, however the method itself has applications to any general covariance modelling function, including anisotropic functions. For this example we use an isotropic covariance model with a dual length-scale Gaussian correlation, the same as was used for the H-L method. We are able to produce an error estimation, without the use of bins, and the curve fitting scheme, which generally reduces the free parameters required for analysis. However, our method still uses statistical components and there is similar requirement for the amount of available observations.

### 3.5.1 Innovation covariance statistics

As a base for all *analysis of innovations* approaches, the covariance of innovations is defined as a function of relative position between each point and all other locations [Bannister, 2008]. The innovations contain information for both the forecast and observational error such that;

$$\mathbf{d} = \mathbf{y} - H\mathbf{x}_b = (\mathbf{x}_t + \epsilon) - (\mathbf{x}_t + \eta) = \epsilon - \eta \tag{3.17}$$

The model and observational data can be seen as the true state, $\mathbf{x}_t$, plus forecast error, $\eta$ or observational error $\epsilon$. The innovations allow us to remove the true state leaving only information on the respective errors. The terms forecast error and background error can be used interchangeably for our experiments, as our only source of a-priori information is our assimilated model forecasts.

To calculate the covariance some assumptions must be made for the innovation errors. The first assumption is that both the observations and forecast error are not biased. This simplifies the equation of covariance for innovations to;

$$F(\mathbf{r}) = \overline{\mathbf{d}(t, \mathbf{0})\mathbf{d}(t, \mathbf{r})} \tag{3.18}$$

Where the overbar represents an average in time or the expected value, which are equivalent due to the application of the ergodic condition. We say $F(\mathbf{r})$ is the innovation covariance statistics, where $\mathbf{r}$ is the n-dimensional relative position vector from the target gridpoint at $\mathbf{0}$, and finally $\mathbf{d}(t, \mathbf{r})$ is the innovation at a certain location, for time $t$, in the range of all the days we have observations, $T$.

In practice, it is generally safe to assume that the forecast error and observational error are unbiased, since most advanced DA schemes include a bias correction

[Mogensen et al., 2012]. Our datasets for innovations have passed through a quality control and bias correction process before we received them, so the specifics are not known.

When we substitute equation (3.17) into (3.18) the result is;

$$\begin{aligned} F(\mathbf{r}) &= \overline{(\epsilon_{\mathbf{0}} - \eta_{\mathbf{0}})(\epsilon_{\mathbf{r}} - \eta_{\mathbf{r}})} \\ &= \overline{\epsilon_{\mathbf{0}}\epsilon_{\mathbf{r}}} - \overline{\eta_{\mathbf{0}}\epsilon_{\mathbf{r}}} - \overline{\epsilon_{\mathbf{0}}\eta_{\mathbf{r}}} + \overline{\eta_{\mathbf{0}}\eta_{\mathbf{r}}} \end{aligned} \tag{3.19}$$

By assuming the covariance of observations and forecast error are uncorrelated, and that for non-zero separation the observation errors are also uncorrelated, we get;

$$\overline{\eta_{\mathbf{0}}\epsilon_{\mathbf{r}}} = \overline{\epsilon_{\mathbf{0}}\eta_{\mathbf{r}}} = \overline{\epsilon_{\mathbf{0}}\epsilon_{\mathbf{r}}}|_{\mathbf{r}\neq\mathbf{0}} = 0 \tag{3.20}$$

$$F(\mathbf{r}) = \begin{cases} \overline{\epsilon_{\mathbf{0}}\epsilon_{\mathbf{0}}} + \overline{\eta_{\mathbf{0}}\eta_{\mathbf{0}}}, & \text{for } \mathbf{r} = \mathbf{0} \\ \overline{\eta_{\mathbf{0}}\eta_{\mathbf{r}}}, & \text{for } \mathbf{r} \neq \mathbf{0} \end{cases} \tag{3.21}$$

As the number of observations increase the innovations statistics converge in equation (3.18) to a function for forecast error covariance. Notice that from (3.21), this function is not continuous at $\mathbf{0}$, where error variances include both forecast and observation at zero separation [Hollingsworth and Lonnberg, 1986], [Bannister, 2008].

The above equations and any definitions regarding the innovation covariance are the same for the H-L method in section [3.3.2]. We have redefined them, with some additional details, to be consistent with our notation for the remainder of

the mathematical derivation. This is also to be consistent with the notation used within our scientific publication for the BAI method.

### 3.5.2 Modelling the covariance function

In order to use the innovations to find a model for the background error we must choose an estimate form function for the covariance error. An operational error covariance modelling function is already used within NEMOVar to produce the **B** matrix, and in this section we will use the same approach. However, our method can be used in more general situations, if a better model for background error is suggested (for instance, an anisotropic model). We will be using our method for a covariance model defined by a linear combination of two isotropic Gaussian functions;

$$\tilde{f}(\mathbf{r}) = m_1\phi_1(\mathbf{r}) + m_2\phi_2(\mathbf{r}) \tag{3.22}$$

With the basis functions;

$$\phi_n(\mathbf{r}) = e^{\frac{-\mathbf{r}^2}{2L_n^2}}, \qquad n \in [1, 2] \tag{3.23}$$

$L_1$ and $L_2$ are the short and long length-scales respectively. In practice, the basis functions $\phi_n(\mathbf{r})$, are implicitly computed as solutions of the diffusion equations with their respective length-scale, and were represented by matrix **L** in equation (3.7) [Weaver and Courtier, 2001].

There are some practical difficulties when computing innovation covariances, for the observations the summation over time in the covariance is impossible since the observation locations constantly change. With the H-L method, this is where the binning process occurs.

For the BAI method, we use the observations at their exact location which can change in time, as a result we cannot produce an error covariance for these locations. Instead, we delay any statistical computation until the very end when the statistics are performed for all data simultaneously, independently of their locations.

### 3.5.3 Minimizing the norm

In order to produce the background statistics without spatial binning, we need to find another process to find our function $\tilde{f}$. A solution is found by first defining a subspace $\tilde{V}$ this is a subspace of the continuous function space spanning all possible one-dimensional continuous functions. The subspace is defined as a combination of two basis functions $\phi_1$ and $\phi_2$;

$$\tilde{V} = m_1\phi_1(\mathbf{r}_i) + m_2\phi_2(\mathbf{r}_i), \qquad m_1, m_2 \in \mathbb{R} \tag{3.24}$$

The purpose of this equation is to find the best form of function $\tilde{f}(\mathbf{r})$ within the subspace $\tilde{V}$ that can approximate our background error covariance from $\tilde{F}(\mathbf{r})$. With the main difference between and the subspace being that (3.24) is for a subspace with all possible values for $m_1$ and $m_2$. The modelling function $\tilde{f}$ is the particular function and values that best approximates $\tilde{F}$. In a mathematical form this is equivalent to finding a function that will minimize the following norm;

$$\left\| \tilde{f}(\mathbf{r}) - F(\mathbf{r}) \right\| \tag{3.25}$$

This minimization is with respect to the internal weighting factors $m_1$, $m_2$ Our definition for the norm uses the inner product, $||x|| = \sqrt{\langle x, x \rangle}$, and hence we can

use a natural definition of the inner product, such that;

$$\langle g(\mathbf{x}), h(\mathbf{x})\rangle = \sum_{i=1}^{N} g(\mathbf{x}_i)h(\mathbf{x}_i) \tag{3.26}$$

With $N$ being the total number of inputs, which for our operational case is the number of innovations, and then $\mathbf{x}_i$ is the location for the innovation $i$. Now, from appendix [A] we have that the minimisation of the norm is equivalent to solving the following equation;

$$\langle \tilde{f}, \phi_n\rangle = \langle F, \phi_n\rangle, \quad \text{for } n \in \{1, 2\} \tag{3.27}$$

Since in this case we are looking at a covariance we will use the relative position $\mathbf{r}_i$ for each grid point, instead of the location $\mathbf{x}$, and we will replace $g$ and $h$ in (3.26) with our components in (3.27), and hence;

$$\sum_{i=1}^{N} \tilde{f}(\mathbf{r}_i)\phi_n(\mathbf{r}_i) = \sum_{i=1}^{N} F(\mathbf{r}_i)\phi_n(\mathbf{r}_i), \quad \text{for } n \in \{1, 2\} \tag{3.28}$$

Additional difficulties occur if we try to begin solving equation (3.28), as we know all the components of the left hand side, except the unknowns we are solving for, but we do not know the right hand side. Since $F(\mathbf{r})$ represents the innovation error covariance, this requires knowledge of $f'(\mathbf{r}_i, t_j) = \mathbf{d}(\mathbf{0}, t_j)\mathbf{d}(\mathbf{r}_i, t_j)$ for all time $t_j$, due to equation (3.18);

$$F(\mathbf{r}) = \overline{\mathbf{d}(\mathbf{0}, t)\mathbf{d}(\mathbf{r}, t)} = \sum_{j=1}^{T} \mathbf{d}(\mathbf{0}, t_j)\mathbf{d}(\mathbf{r}, t_j) \tag{3.29}$$

Since we are limited here to the available observations and their locations, we cannot compute $F(\mathbf{r})$ exactly, it must be estimated by other means. We are already assuming our innovations are independent in time when we use equation (3.18), due to ergodic conditions for the covariance. So we are able to use

Kolmogorov's strong law of Large numbers, see appendix [B];

$$\sum_{i=1}^{N} f'(\mathbf{r}_i, t_{j(i)})\phi_n(\mathbf{r}_i) \xrightarrow{N\to\infty} \sum_{i=1}^{N} F(\mathbf{r}_i)\phi_n(\mathbf{r}_i) \tag{3.30}$$

Where we have introduced a new term, $t_{j(i)}$, this is a random time in the range from 0 to $T$ for the specific index $i$, which makes $f'(\mathbf{r}_i, t_{j(i)})\phi_n(\mathbf{r}_i)$ also random. The law also guarantees that the variance of the average of $n$ random variables is $\sigma^2/n$. Where sigma is the SDV of the random variables, the variance will decrease towards zero as the sample size increases towards infinity.

From (3.28) and (3.30);

$$\sum_{i=1}^{N} \tilde{f}(\mathbf{r}_i)\phi_n(\mathbf{r}_i) = \sum_{i=1}^{N} f'(\mathbf{r}_i, t_{j(i)})\phi_n(\mathbf{r}_i) \quad \text{for } n \in \{1, 2\} \tag{3.31}$$

In order to compute the summation on the right hand side of equation (3.31) we choose $t_{j(i)}$ to be times for which we have observations. By doing this we make the statement that our choice for $t_{j(i)}$ is a random time, as this is completely unbiased. Now, since the left-hand side can be separated into its known components, we have;

$$m_1 \sum_{i=1}^{N} \phi_1(\mathbf{r}_i)\phi_n(\mathbf{r}_i) + m_2 \sum_{i=0}^{N} \phi_2(\mathbf{r}_i)\phi_n(\mathbf{r}_i) = \sum_{i=1}^{N} f'(\mathbf{r}_i, t_{j(i)})\phi_n(\mathbf{r}_i) \quad \text{for } n \in \{1, 2\} \tag{3.32}$$

Where we can create a matrix representation for (3.32);

$$M_{n,m} = \sum_{i=1}^{N} \phi_n(\mathbf{r}_i)\phi_m(\mathbf{r}_i), \quad T_n = \sum_{i=1}^{N} f'(\mathbf{r}_i, t_{j(i)})\phi_n(\mathbf{r}_i) \tag{3.33}$$

$$\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{bmatrix}^{-1} \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \tag{3.34}$$

By solving (3.34) for $m_1$ and $m_2$, we can then calculate the variance and length-scale ratios by manipulating the covariance modelling function (3.22),

$$m_1\phi_1(\mathbf{r}) + m_2\phi_2(\mathbf{r}) = V\big(\alpha\phi_1(\mathbf{r}) + (1-\alpha)\phi_2(\mathbf{r})\big), \tag{3.35}$$

and hence we can rearrange for $V$ and $\alpha$ in terms of $m_1$, $m_2$.

$$V = m_1 + m_2, \qquad \alpha = \frac{m_1}{m_1 + m_2} \tag{3.36}$$

### 3.5.4 Anisotropic BAI

As we have mentioned previously, the original goal was to be able to improve the background error to include anisotropy, so once our method had been tested and proven successful in isotropic conditions we moved to the anisotropic options.

Since the current method is able to weight multiple diffusion operators, and we are able to freely define the covariance function $\tilde{f}$, we decided to experiment with alterations to the basis functions to express anisotropy. The first logical step for this was to use the current length-scales in alternating combinations. We had tested the isotropic function with separate $x$ and $y$ for relative location, such that;

$$\tilde{f}(\mathbf{r}) = m_1 e^{\frac{\mathbf{x}^2}{2L_1^2}} e^{\frac{\mathbf{y}^2}{2L_1^2}} + m_2 e^{\frac{\mathbf{x}^2}{2L_2^2}} e^{\frac{\mathbf{y}^2}{2L_2^2}}, \tag{3.37}$$

Which was able to return the same values for $m_1$ and $m_2$ as the isotropic version

of the covariance function. This then proves that we are able to separate the relative distance into relative location, without invalidating the method. We then created our first attempt at using a simple anisotropic covariance function;

$$\tilde{f}(\mathbf{x},\mathbf{y}) = m_1\phi_{1,1}(\mathbf{x},\mathbf{y}) + m_2\phi_{2,2}(\mathbf{x},\mathbf{y}) + m_3\phi_{1,2}(\mathbf{x},\mathbf{y}) + m_4\phi_{2,1}(\mathbf{x},\mathbf{y}), \qquad (3.38)$$

$$\phi_{n,m}(\mathbf{x},\mathbf{y}) = e^{\frac{\mathbf{x}^2}{2L_n^2}} e^{\frac{\mathbf{y}^2}{2L_m^2}}, \qquad n, m \in [1, 2]. \qquad (3.39)$$

We have been able to use our BAI method with this anisotropic covariance model to produce the error coefficients. We can then use this equation to extract values for SDV and a variety of LSRs, as follows,

$$\tilde{f}(\mathbf{x},\mathbf{y}) = V\big[v\big(w_1\phi_{1,1}(\mathbf{x},\mathbf{y})+(1-w_1)\phi_{2,2}(\mathbf{x},\mathbf{y})\big)+(1-v)\big(w_2\phi_{1,2}(\mathbf{x},\mathbf{y})+(1-w_2)\phi_{2,1}(\mathbf{x},\mathbf{y})\big)\big]. \qquad (3.40)$$

Where we say that; $V$ is the error variance, $v$ is the anisotropic-isotropic weighting, $w_1$ is the short-long LSR, and $w_2$ is the longitudinal-latitudinal LSR. We show the anisotropic results in the next chapter, when we discuss the general results of both the H-L and BAI method.

This is one specific option we have for defining the covariance function $\tilde{f}$, but there is many different combinations that could be used. We have attempted a few examples and found that much further research would be needed to find the best format for quantifying anisotropy. We also found that our original aim to create an anisotropic background ECM would not be do-able within our time

frame, we feel our BAI method is a good first step to solving this, as it allows for anisotropic quantification of error, but research for an appropriate anisotropic background error covariance functions is required before we can implement this into DA.

## 3.6 Summary

In this chapter of the thesis we have described our operational use of data assimilation to create an improved ocean model forecasts. Where we are using the observational and model data in combination to create a cost function, with weightings based on the inverse of their respective error. This process was described in more detail in the introduction chapter, including how Var was originally created from Bayes theorem.

We have used the NEMOVar analysis, with our NEMO model, the AS20, to successfully improve the forecasting ability. This was created in collaboration by the POFC and the MO, with their research project for the UAE oceans team, and then I was given access to this system.

The assimilation system itself is controlled by the Rose and Cylc task configuration and work flow scheduler. Once available I have familiarised myself with this system and translated my knowledge from the theory of DA to the operational suite, noticing where the methods deviate from standard DA into the more specific 3DIncVar. Although NEMOVar has options for other types of DA methods, our plan was to try and improve the representation of the background error and not assess the abilities of different DA methods.

The original plan for our research was to create appropriate anisotropic representation within the background ECM, this task has been of interest to many researchers as the isotropic assumption is known to not be fully justified in all areas. For us to be able to do this we needed to understand the process of creating the background ECM for NEMOVar. The typical operational process for creating the background ECM will use stored error SDV and LSR to model the full matrix during the assimilation cycle. This process contains many components, including; balance operators, diffusion operators, normalisation factors and error estimation. We have spoken about this and how these components are connected to our research.

While studying these components, we began replicating the two most common operational approaches for error estimation, the NMC and the H-L methods. The NMC is a method of using model differences to create a statistical proxy and infer the background error. The H-L is known as an analysis of innovations method, where observations minus model data is used to predict the background and observational error variances, as well as the LSR for the background covariance.

Once we were able to replicate the method and produce results, we quickly found that there was a large preference for us to use the H-L method, as the analysis of innovations gave a much more appropriate estimate for the error SDV and LSR. This also led us to begin creating our own method using the basis for an analysis of innovations approach. We aimed to use the method of error estimation to expand the covariance function into two dimensions, and produce an error estimate that was able to consider anisotropic diffusion. With the overall plan for this to then lead to the production of an anisotropic diffusion operators that could be

used in the creation of the background ECM, and be weighted by the results of our 2D error estimation.

The BAI method is able to estimate the error SDV and LSR, similar to how the H-L method would, but we are using the exact locations of the observations as opposed to the binned distances. By doing this we do not produce an innovation covariance as the H-L does, but we are able to create a framework in which a minimization of the norm between the covariance function and the innovation statistics is able to lead to a solution for the error coefficients, and hence the SDV and LSR.

We have not shown any results for our BAI method as this is reserved for the results section [4], but we have been able to show the mathematical derivation. This derivation shows how robust the statistics of the BAI method are, and how we are able to use the general assumptions of the analysis of innovations to separate the background and observational error. Our derivation uses a specific case for the current NEMOVar assimilation that we are using, however the method is flexible to use any background covariance modelling function, and this is how we aimed to include anisotropy.

The general benefits of the BAI method over that of the H-L are due to; the reduction in free parameters, avoiding the need to use spatial binning, and having a straightforward option to include anisotropic components. The results will show how the methods perform when compared with each other.

# Chapter 4

# Results and Discussions

## 4.1 Introduction

During this section we want to evidence the benefits of our method in comparison with the H-L method for estimating background error variances and LSR. Previously we have mentioned in a few cases the shortfalls of the H-L method; with binning the data, applying curve fitting algorithms, and being fundamentally one dimensional.

Our focus for this thesis will be on the first two complaints of the H-L approach, while our method does allow for additional dimensions, the benefits from this will take future research to complete. We have included our experiments with anisotropy in section [4.8], but the first step is proving the standard results and validity of the BAI. We want to be able to show that by removing the data binning aspect, we are able to maintain a consistent comparison of computational time for large observational datasets, while also creating a more reliable method by removing unnecessary averaging or interpolation.

The second change that our method applies is the removal of any curve fitting processes, this is the how H-L minimizes the norm and estimates the background

error statistics, however it can be unreliable. From our examples we can sometimes create outliers from basic curve fitting schemes, and the result may require some smoothing to create a usable error estimate. Another option would be to use a more complex curve fitting algorithm, the downside to this is additional computational costs and domain specific research of the innovation statistics. By avoiding the curve fitting and instead using the inner product and basis functions in equation (3.27), we are able to produce a robust solution and avoid the issues mentioned.

Some of the issues with H-L are also issues with the BAI, we do also produce some statistical outliers, and the BAI method may suffer in accuracy if we are unable to provide a sufficient amount of observational data. In our experiments we have used a year of 24-hour forecasts and observations due to the availability, but most likely in fully operational models multiple years worth of data are used. However, we believe that the fundamental mathematics of our method is able to improve the reliability of the statistics, with the benefit of there being fewer subjective parameters to define in order to estimate the error. In the remainder of this section we hope to be able to show this by comparing the results of both methods in a few scenarios including, subsets of observations and assessing the differences in NEMOVar assimilated forecasts.

The evidence for our claims are based upon the following results; we have produced spatial plots for both the forecast SDV and LSR in section [4.3] for the coarse grid and section [4.4] for the model grid. Then we have experimented with the performance of the two methods using subsets of innovations in section [4.5]. A mention for the results of both the H-L and BAI methods using the SLA innovations is done in section [4.6]. The next result shown is the RMSE and

mean of assimilated model forecasts from NEMOVar runs, section [4.7]. Finally we have given results for our anisotropic investigations in section [4.8]

## 4.2 Operational methodology

Our research has been purely focused on the Arabian Sea (AS), this is the area of interest for the POFC due to their research project with the UAE ocean's team. This meant that models were produced, using the Nucleus for European Modelling of the Ocean (NEMO) framework, for 1/20 and 1/60 degree resolution, this is the AS20 and AG60 ocean models respectively. For this research we have only used the AS20 model and a 3D-Var suite NEMOVar. The DA component was produced as a collaboration between the POFC and the Meteorological Office (MO), under the same research project, and the innovation datasets that we have used for our estimates of the error covariance, were provided by James While and Isabella Ascione from the MO, using this same assimilation suite.

The original NEMOVar suite used interpolated (from the Indian ocean assimilation IND12) SDV and LSR to compile the assimilation, then using this as a first approximation for a year of 24-hour assimilated forecasts were produced and a second set of error estimations was created using these forecasts. For an accurate error estimation method, assimilated forecasts are required as this affects the model data that will be used, and hence the model error we wish to estimate will change. In the later sections we use the second iteration of error estimates that were supplied by the MO as a comparison for our methods on the model grid. The problem with using the MO error is that we do not know the exact methods of production, and hence it cannot be used to comment on the specific methodology.

The AS20 ocean model covers; the Gulf of Aden, the Arabian Gulf, the Gulf of Oman, and a large percentage of the Arabian Sea. The AS20 is a multi-variate model and outputs temperature, salinity, SLA and velocity, which means that the NEMOVar assimilation is also multi-variate and also produces improved forecasts for these variables. A multi-variate assimilation suite, requires an additional step to *balance* the variables and account for any cross-correlation between variables, as within the assimilation suite they are all assumed to be uncorrelated. The exception for this is the temperature variable, as this is considered the *control* model variable that the other variables are balanced around. For this reason, and due to the much larger availability of observations, the majority of our experiments are for predicting background error covariances for temperature only.

We have produced a background error estimate for the SLA as well, as our observational dataset includes SLA. However we are slightly limited, since we cannot run NEMOVar with the result due to the balance operator. Despite this limitation we are able to see some additional results for the efficiency of our method with different observations, which is described in section [4.6].

The total observations available per season have been shown in the following table, this is used to show some perspective on potential causes of changes within the final results. The reason different seasons have a different number of observations available is usually down to the available satellites, but may also be due to the quality control that is applied to the observations for that parameter. Any outliers in the observational data are removed before being used in the assimilation or the error estimation processes.

*Table 4.1: Number of available observations for SST and SLA per season.*

| | SST | SLA |
|---|---|---|
| Winter | 4736992 | 29194 |
| Spring | 5230179 | 36038 |
| Summer | 2003976 | 37790 |
| Autumn | 4250507 | 41025 |

The background error covariances have all been estimated using a series of python scripts that follow the literature in sections [2.3.1] and [2.3.2], and further details of operational applications for them has been given in the methods section [3]. For the reasons of computational costs an additional coarse grid has been produced, which is the same as the model grid of AS20 except that the grid resolution is reduced from 5km to 30km. This reduces the complexity of the algorithms (from $\sim 80000$ to $\sim 2500$ grid points) while maintaining an appropriate level of accuracy. These results were then interpolated into the model grid for use in NEMOVar, and both model and coarse grid results are shown in the following sections. For both methods the same coarse grid and spatial interpolations were used.

The error variance and LSR that are used within NEMOVar have a specific format due to the restraints that are applied by the assimilation suite itself. NEMOVar uses netCDF file format for both error estimation components, and when we interpolate from coarse grid to model grid, we also convert from .npy, saved Python array, to .nc, netCDF format. Other restrictions are applied so that we can decrease the cost for production and minimization of the cost function, the details for this were mentioned in the literature review [2]. Since the cost function uses the CVT, the background ECM must be have a real square root matrix and as a result must be positive semi-definite.

NEMOVar requires the error SDV and square-root of the length-scale ratios, instead of our variance and ratios. Before we use our error estimation outputs in any operational assimilation we must make changes to allow for this. First is that we square root the error variance, which means that we cannot allow for any negative values, and the second is to square root the LSR which is more restricted. Since both weights need to be square rooted, neither weight can be less than zero or greater than 1. These limitations are applied when converting to the model grid, but the coarse grid represents the results directly from our application of the methods before any interpolation.

## 4.3 Spatial comparisons - Coarse grid

The results of both methods are shown below, for each method they are split into the four seasons, and then also split into the error variance, and error LSR. These are the two defining parameters for the entire background ECM which cannot be stored, and instead the values of the full matrix will be modelled by the assimilation suite when required.

The outputs shown in this section are on the coarse grid for the AS20, this is the raw data that is produced from our methods. Then the general limitations briefly mentioned in section [4.2] is applied as well as some simple smoothing, this is to prepare them for use within the NEMOVar assimilation suite. It is apparent in figures 4.1 and 4.3 that both of the methods produce similar results, due to the fact that the fundamental mathematics have the same aim. To produce a statistical representation of the innovation error, and then minimize any differences between the statistics and a covariance function representing the background error. By assuming both methods are working correctly they should produce similar results, since the statistics and the aims are the same. Any differ-

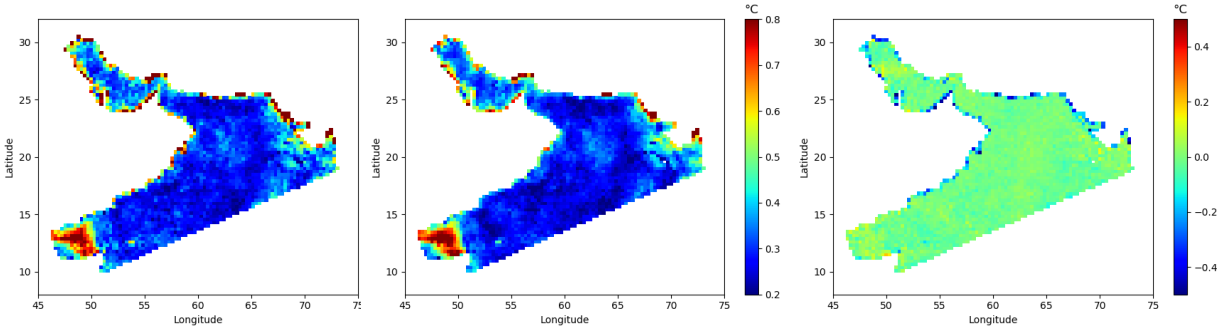
ences in the spatial comparisons will be due to the computational differences or differences between statistical reliability of the methods for the BAI and the H-L.

The following plots have been produced using all the available observational datasets for satellite data and the respective model runs, which will produce a set of innovations statistics. As we have mentioned earlier, we must be prepared to consider domains and variables which will not have this large data availability, and still produce a reliable estimate for the background error variance and LSR. We will discuss innovation subsets and their affect on productivity in section [4.5].

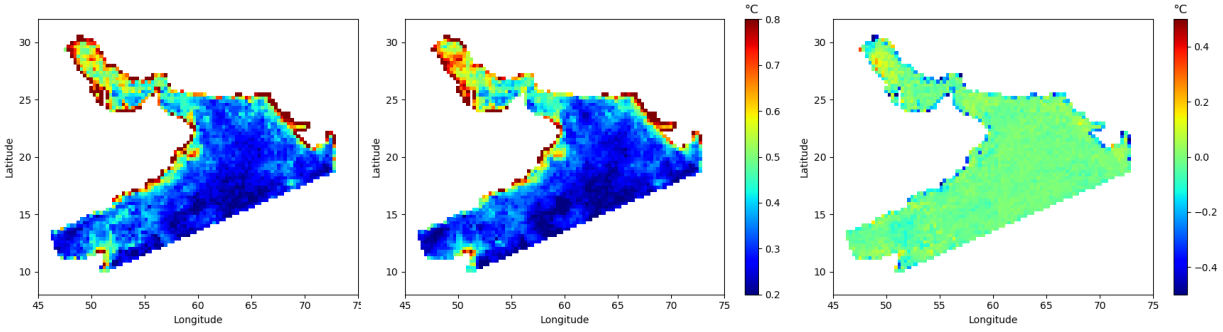
Figure 4.1 represents the SDV from both the BAI and the H-L method on the coarse grid, for there respective seasons. After creating the error covariance model we have solved for both the error variance, $V$, and LSR $w_1$. Since in NEMOVar the SDV are used instead of the variance, we have plotted the SDV $\sqrt{V}$. The contour range has been set with a minimum of 0.2 and a maximum of 0.8, this is then kept consistent throughout for all error SDV plots. Any areas which do not have observational data have been masked, as we cannot produce values for them, and any negative values have also been masked, since negative values cannot be square-rooted.

Figure 4.2 is the LSR for the short length-scale diffusion, the long length-scale weight will therefore be 1 minus the plots. I have not included this here as the important ratio between the two can be inferred from just the one image, per method, per season. In essence, for the plots in figure 4.2, where the values are closer to 1 more weight is being placed on the short scale diffusion and for the lower values the long scale diffusion has more influence. In practical covariance modelling, there is no reason we cannot support a negative weight, (or one of the

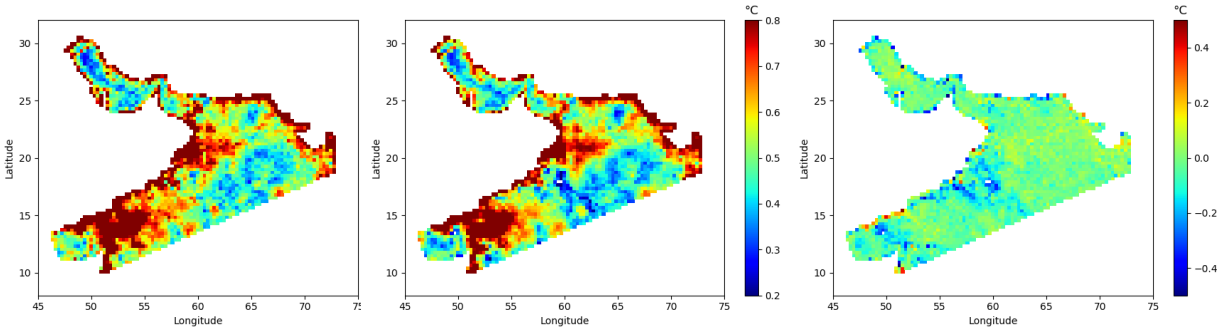
weights being greater than 1) for either short or long scale, however the current functionality for NEMOVar prohibits this. Despite this the images here have not been limited, instead we have just plotted the weightings between 0 and 1. We have done this because it makes it much easier to visualize the long scale weighting as one minus the image, $1 - w$, than the square root of one minus the values that are plotted being squared, $\sqrt{1 - w^2}$. The depth profiles for the LSR of both the BAI and the H-L were estimated using a year of model data with the parameterisation mentioned in section [3.2.2].



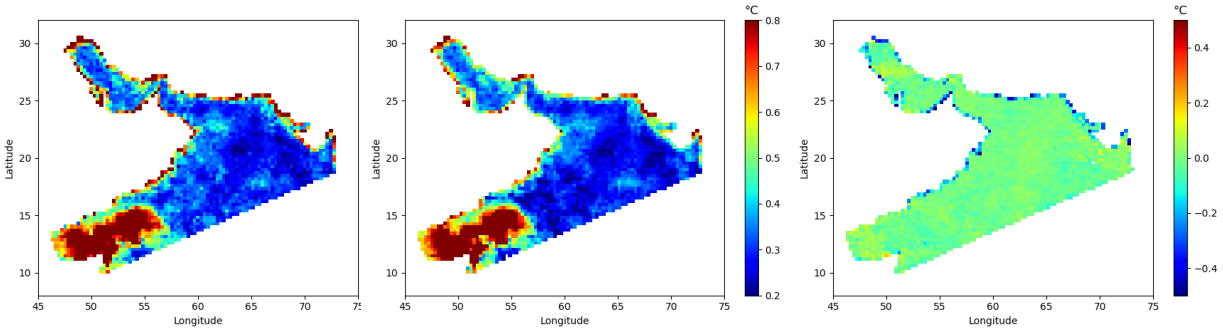

(a) Winter : H-L (b) Winter : BAI (c) Winter : BAI - H-L




(d) Spring : H-L (e) Spring : BAI (f) Spring : BAI - H-L

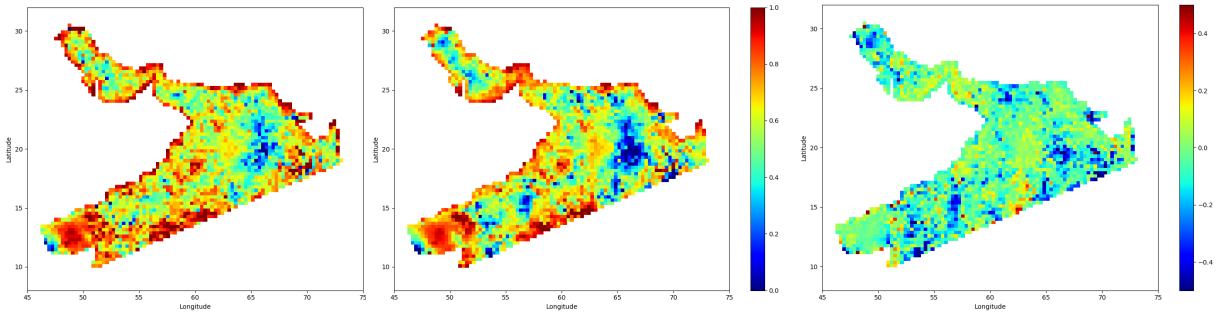



(g) Summer : H-L (h) Summer : BAI (i) Summer : BAI - H-L

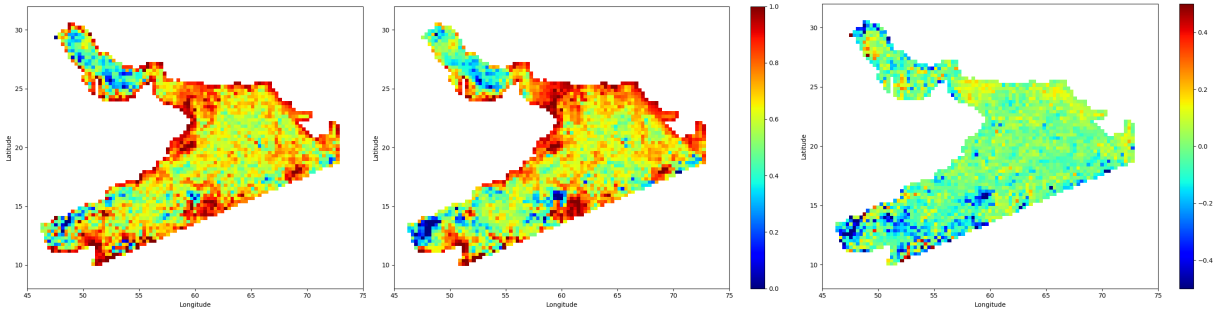



(j) Autumn : H-L (k) Autumn : BAI (l) Autumn : BAI - H-L

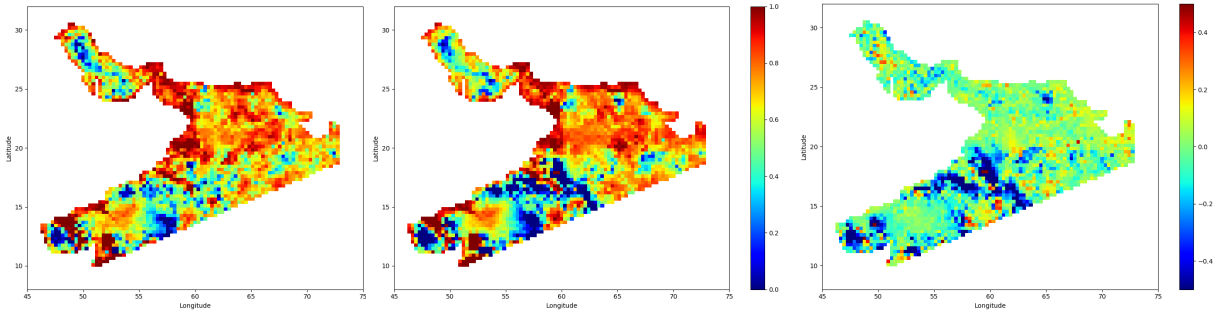
*Figure 4.1: The forecast error SDV on the coarse grid for AS20. Produced by either the H-L or BAI methods with SST, for each season. The third plot in each row demonstrates the BAI minus H-L, and it is important to note the unique colour bar for these which is centered around zero.*



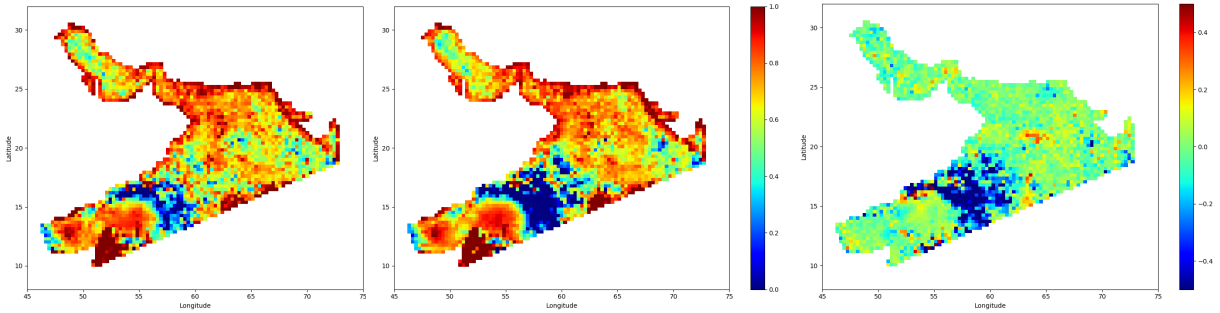
(a) Winter : H-L (b) Winter : BAI (c) Winter : BAI - H-L



(d) Spring : H-L (e) Spring : BAI (f) Spring : BAI - H-L



(g) Summer : H-L (h) Summer : BAI (i) Summer : BAI - H-L



(j) Autumn : H-L (k) Autumn : BAI (l) Autumn : BAI - H-L

*Figure 4.2: The short forecast error LSR on the coarse grid for AS20. Produced by either the H-L or BAI methods for SST, with four independent seasons. The third plot in each row demonstrates the BAI minus H-L, and it is important to note the unique colour bar for these which is centered around zero.*

For SDV the seasonal trends are very consistent through both methods, suggesting a rise of SDV for error in the warmer seasons, specifically summer. This is supported by the general increase in temperature for these months, as we would expect an increase in error to follow. This increase in error is only true in an absolute sense and would not be the case if we were to normalise the data or use a Coefficient of Variation (CoV). The CoV is a relative error and would remove any component of our estimation that may be dependent on the seasons, other than that which is due to the data availability shown in table 4.1. However, NEMOVar is configured to use the SDV and LSR to estimate error and hence we have shown these statistic instead of producing the CoV.

There is also large peaks of error SDV in the coastal regions and these areas can be more difficult to model accurately. Due to the addition of ancillary river and tide data as well as a requirement for optimum boundary conditions, there is more complexity in the model in these areas. We also believe that with less data availability in the summer months, both methods will struggle to produce an estimate for the error with the same reliability, due to the weakening of statistical justifications. This affect is expected to be minimal for the SDV, and to have greater affect on the LSR estimates as it is more sensitive to the accuracy of the statistics.

The LSR plots show a higher variability throughout the domain, again showing peaks of obscure behaviour in warmer seasons, and noticeable changes in the coastal regions. It is difficult to justify some of the changes that we can see here, but they are again generally consistent between the two methods. The model differences show that the largest deviations between the methods often occur at the high and low peaks. The initial assessment is that these represent

the sensitivity to the availability of data for the LSR, as there is less data the methods fluctuate from the ideal solution in different ways. This is support by again noting the seasonal trend, whereby the seasons with less observations have increased peaks and troughs, but also larger model differences.

The higher spatial fluctuations can be justified as still providing an optimum answer for the models error LSR, as there is no definitive reason that the 1-dimensional error covariance for a specific grid point should be approximately equal to its neighbours. Despite an assumption that the error correlation is in some way caused by an underlying error within the model, which should be related to position or model inputs, this is not a necessity and the error correlation weights can be spatially sporadic. It is also worth nothing that the dark red and dark blue colors do continue to show a relative consistent rate of change outside of the plotted range and not extreme rapid fluctuations. Although this may be a result of the manifestation of the sensitivity to data availability noted in table 4.1.

There are multiple locations where we do see some domain specific trends, like a higher weighting on the shorter length-scale for; the north-east coast, the Gulf of Oman and towards the Gulf of Aden. Which does suggest that some physical boundaries can influence the estimated optimum LSR, which we do expect in some way. However it's important to remember that due to the statistical nature of boths methods, a lack of observations will have a negative affect on the accuracy. This will explain some of the obscure behaviour that can be seen in the summer and autumn seasons or in isolated coastal regions which do not have much surrounding observations.

In the first two seasons, it is clear to see that both methods suggest a higher

weighting for the longer length-scale within the open ocean areas. However this does not seem to be the case on the southern border of the plot, where the Arabian Sea would continue to more open ocean, both approaches weight the short scale diffusion more. It is likely that this is due to the interpolation of lower resolution model data into the Arabian Sea on this border. In order to run the model, ancillary data is drawn from IND12 (Indian Ocean 1/12 degree NEMO model), and the integration of this data may be the source of the weight on the short scale error diffusion here. Assessing the AS20 model and its ability is not the scope of this research, but I wanted to justify connections where we can to comment on the accuracy of our implementations of the H-L and BAI methods for error estimation.

*Table 4.2: Mean and root-mean-square-error for the deviations of H-L minus BAI spatial plots for SDV. (Degrees Celsius)*

| SDV | Winter | Spring | Summer | Autumn |
|---|---|---|---|---|
| Mean | 0.01 | 0.02 | 0.04 | 0.02 |
| RMSE | 0.06 | 0.08 | 0.12 | 0.07 |

*Table 4.3: Mean and root-mean-square-error for the deviations of H-L minus BAI spatial plots for LSR. (Degrees Celsius)*

| LSR | Winter | Spring | Summer | Autumn |
|---|---|---|---|---|
| Mean | -0.01 | -0.01 | 0.01 | 0.01 |
| RMSE | 0.10 | 0.08 | 0.10 | 0.10 |

Using tables 4.2, 4.3, we can see that the differences between the two methods are relatively small, and it is difficult to say whether any of the changes will be an improvement or a hindrance to the estimate of the background error covariance. In general it seems that the BAI estimates lower levels of error SDV than that of the H-L, and a significant amount of these differences are visible near the northern coasts and within the Arabian Gulf. This is clearer to see by looking at the model differences in the previous figures, consistently for each season in the

SDV the coastline stands out. These changes are continuous and do not seem to be overtly reliant on the season, this is more likely to be some reflection of how the methods use the innovation error, and the inference of background error.

We can also see a general increase for the BAI method in the LSR, with the model differences gives more dark blue colours, which means that the BAI method is on average giving more weight to the shorter length-scale. Although, it can be seen in the spring and summer seasons that the BAI has an overall larger range for its weightings. This range is due to the more negative values in the south-west sea and larger positive values in the Gulf of Oman or north-eastern coasts. Some more seasonal trends can be seen in the differences plots, in which we can see the evenly spaced dark blue spots become more densely located towards the south-west. We believe this is due to a decrease in data availability in these regions over the seasons and is demonstrating a divergence between the methods based on this availability.

It does appear that an increased overall range and notable seasonal differences are a poor estimate from the BAI method however, there is no physical limitation stating that the weightings should be positive, and it is the NEMOVar assimilation that places restriction on the error LSR files. As a result we cannot be certain which trend is a better representation of the true background error SDV and LSR, and instead will need to produce some more diagnostics to judge the methods.

## 4.4 Spatial comparisons - Model grid

In this section will show the model grid results for SDV and length-scale ratios, to further support the similarities between the methods. We have also included

the set of error estimates that was given to the POFC by the MO as part of their collaborations. Figures 4.3 and 4.4 are the same as Figures 4.1 and 4.2, but for the model grid, and with the addition of the MO error estimates for comparison. For our implementations of the methods, the translation from coarse grid to model grid begins by removing any outliers, removing all values outside the *allowed range* and then interpolating the remaining values into the model grid domain. For the variance the *allowed range* is any positive value, and for the LSR the only allowed values are between 0 and 1.

After this change the same trends that were mentioned before can be seen, however the interpolation has smoothed some of the variability, by looking at the LSR its clear that we have much less fluctuations than we did previously and instead the results seem to have more location general formations. This is an expected outcome from smoothing the raw data and removing the outliers. A similar procedure is run by the MO when they convert from their coarser grid to the model domain, but the specifics are unknown.

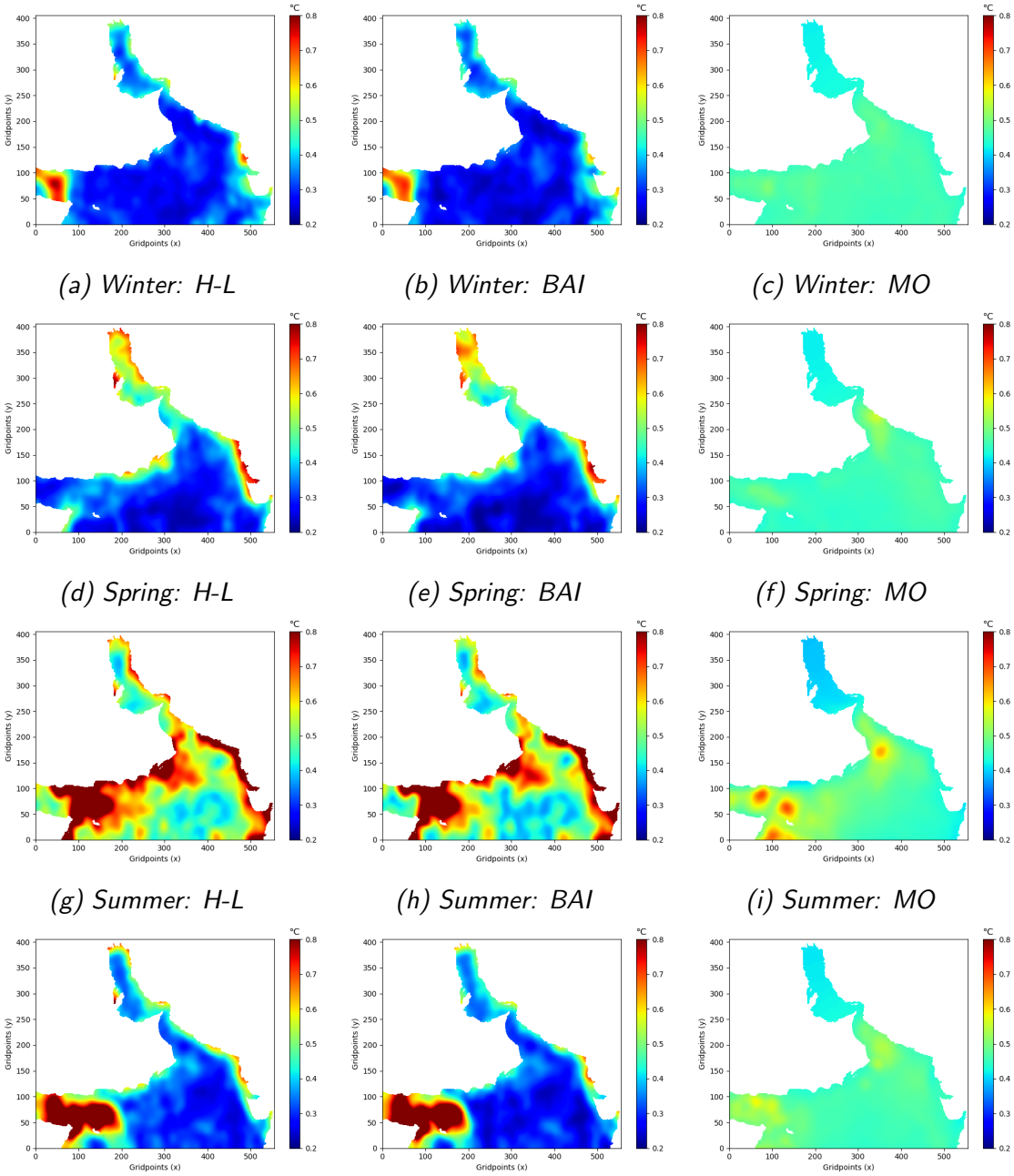
Now that a lot of the noise has been filtered out the remaining results show only the more influential trends of the error estimation results, these have been mentioned before but it is slightly easier to visualize them on the model grid. It is also easier to see the similarities of the methods, where they both seem to be highly in agreement with each other for important features of the background error.

There is no new conclusions that can be drawn from the model grids, but due to the uses of them its important that the reader is able to fully understand how the conversion from coarse grid to model grid affects the SDV and LSR. These

outputs are the ones that will be directly input into NEMOVar to produce the background error for the assimilation forecasts in section [4.7].

The error estimates we have produced are similar to each other but stark differences can be seen between our results and those provided by the MO. The most immediately noticeable of these is the difference in range, for both length-scale and SDV. It is possible that this is caused by applying a harsher smoothing filter, or by differences in application of the H-L method (we believe the method used is a combination between H-L and NMC). However, since we do not know the exact methodologies used in its production, we will mainly be comparing the results of H-L and BAI with each other. When we assess the NEMOVar assimilation, the current operational suite at POFC uses the MO error estimates, and we will use this operational procedure as a control experiment.

It is not possible to claim that any method outperforms the other as we do not know the true state of the model error, at this point we can only comment on general trends and similarities. In general the spatial comparisons here strongly suggest that we are able to successfully produce both SDV and LSR error with the original analysis of innovation approach, H-L. With the changes to this method that we have implemented to create the BAI method we are able to maintain accuracy, computational costs and can improve upon the reliability of the mathematics used. Further experiments have been completed to view the capabilities in areas of sparse observations and how the assimilation with input files from each method compares.



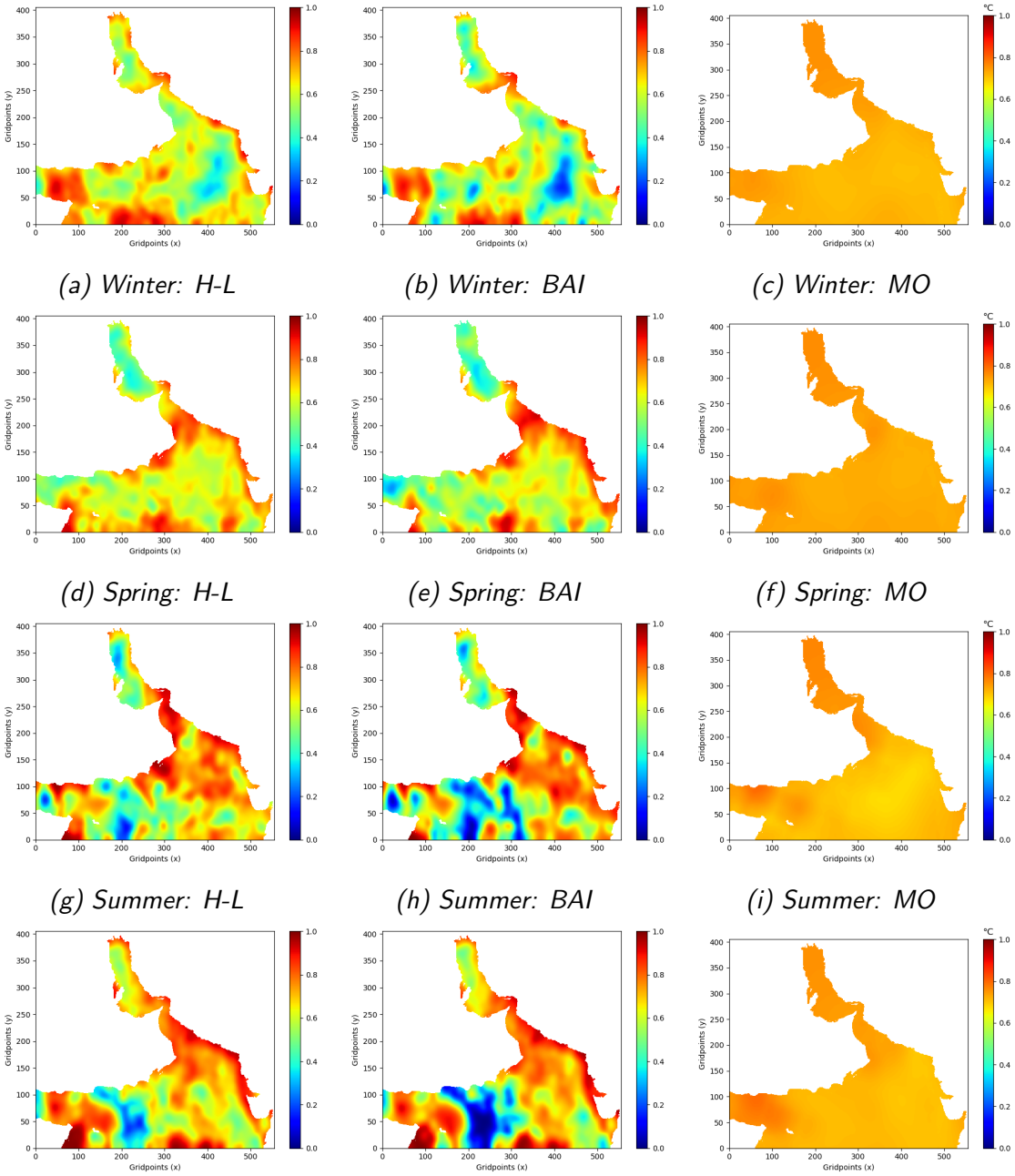

(a) Winter: H-L (b) Winter: BAI (c) Winter: MO

(d) Spring: H-L (e) Spring: BAI (f) Spring: MO

(g) Summer: H-L (h) Summer: BAI (i) Summer: MO

(j) Autumn: H-L (k) Autumn: BAI (l) Autumn: MO

*Figure 4.3: The forecast error SDV on the model grid for AS20. Produced by either the H-L or BAI methods, with comparison to the MO error. Using SST observations for each season.*




(a) Winter: H-L (b) Winter: BAI (c) Winter: MO

(d) Spring: H-L (e) Spring: BAI (f) Spring: MO

(g) Summer: H-L (h) Summer: BAI (i) Summer: MO

(j) Autumn: H-L (k) Autumn: BAI (l) Autumn: MO

*Figure 4.4: The short forecast error LSR on the model grid for AS20. Produced by either the H-L or BAI methods, with comparison to the MO error. Using SST observations for each season.*

## 4.5 Innovation subsets

We have applied our methods of error estimation to the AS20 model by using satellite observations and assimilated model forecasts. Both of the methods we have applied use statistical representations of innovation covariances to estimate the model error. Due to the use of statistics both methods have a reliance on the availability of observations within the domain to strengthen the statistical interpretations. In order to evaluate the methods and there abilities we have used a subset of innovation to re-calculate our error estimates, using only a percentage of the available observations.

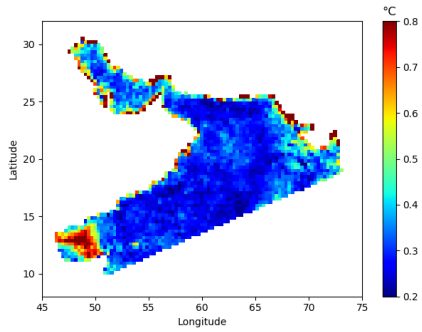
Within these samples only a percentage of random observations per day are kept and the same procedures are applied, this could be representing another domain with less observations or potentially a system with stricter observational quality controls. Both of which are possibilities and it is important to know that when applying the BAI to new domains or variables we are able to maintain accuracy.

When we are using more than 50 percent of available observations the overall results do not change much and the statistical approximations still hold well. However when we begin to use a very small percent of our available data, we see an increase in range for both methods. we have an increased number of negative variance estimates, which will become invalid values in the SDV, and there is also higher estimates of error in all the coastal regions.

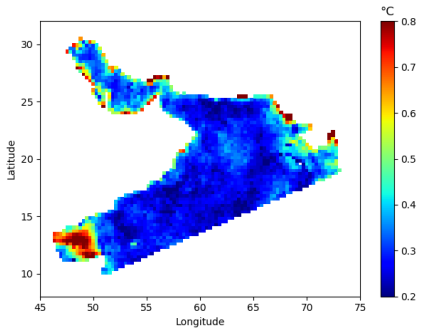
Both methods maintain their trends for reduced percentage of observations, but begin to create more noise as we use less data. The statistical analysis may deviate from the true model error as the averages are computed over a smaller

amount of inputs. This indirectly give the outliers a higher affect on the final output, and generally weakens the reliability of our statistics for innovation covariances.

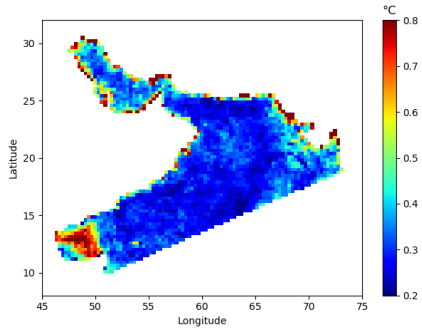
As we aim to compute averages over the seasons, and we predict that this average will tend towards the true model error during this time window, then we will wish to use all the observations that we can. It is safe to presume that for the larger amount of observations the error estimates are closer to the true model error, and we use this to create a discordance plot shown in figure 4.7 and 4.8. The discordance will show us the rate of change, based on size of innovation dataset, and give us information on the capabilities of the H-L and BAI method.



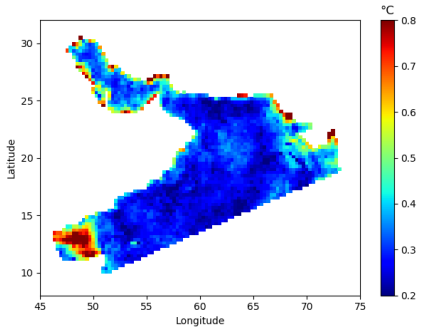

(a) 100%: H-L



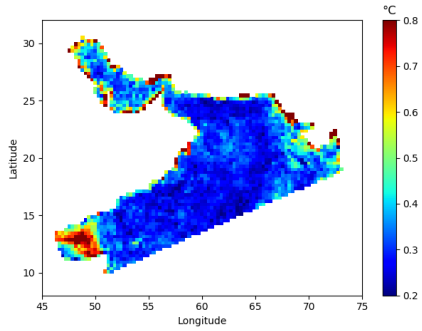

(b) 100%: BAI



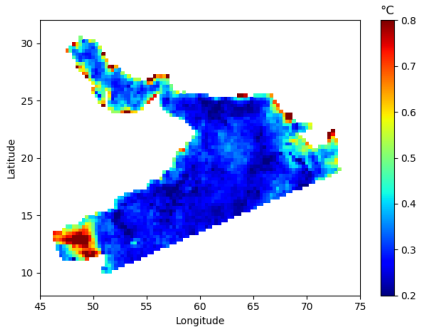

(c) 50%: H-L



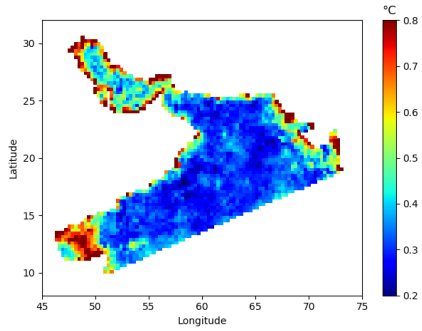

(d) 50%: BAI



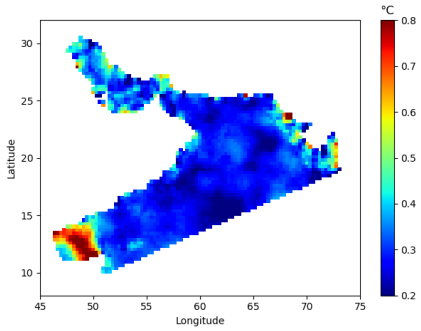

(e) 30%: H-L




(f) 30%: BAI

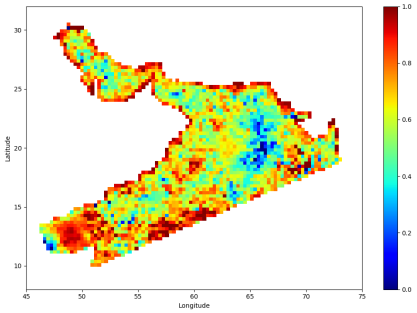



(g) 1%: H-L

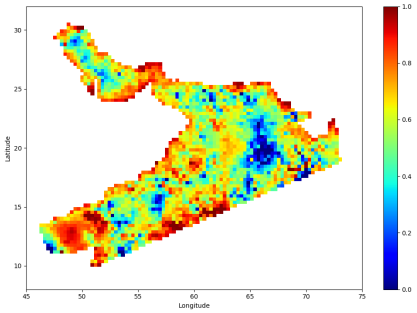



(h) 1%: BAI

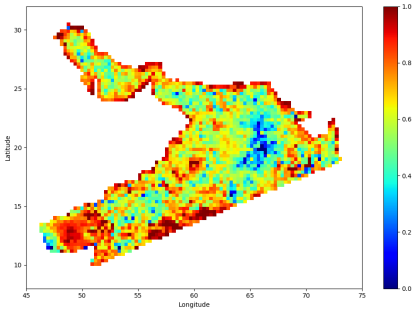
*Figure 4.5: Forecast error SDV plots after applying the error estimation methods, H-L or BAI, with a percentage of the available SST observations. Producing error for the winter season December-January-February.*



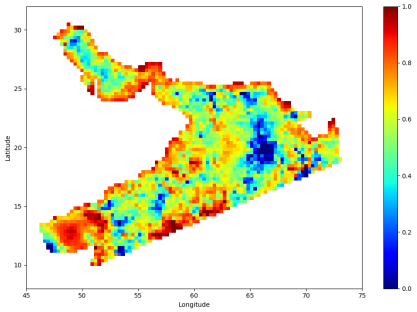

(a) 100%: H-L



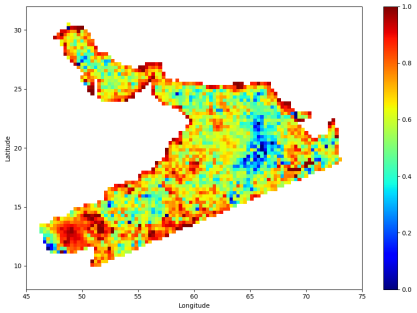

(b) 100%: BAI



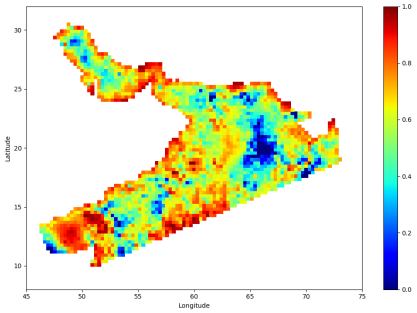

(c) 50%: H-L



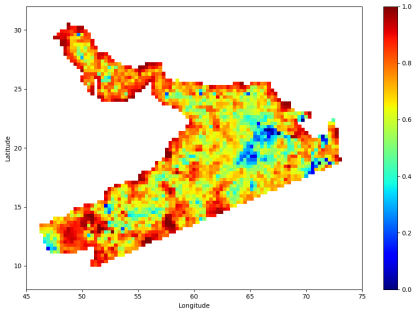

(d) 50%: BAI



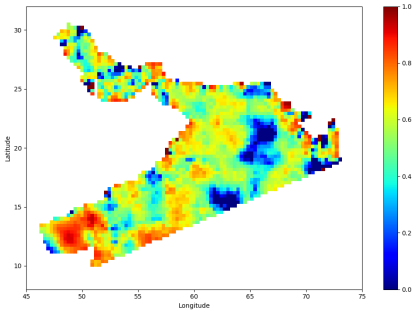

(e) 30%: H-L




(f) 30%: BAI




(g) 1%: H-L




(h) 1%: BAI

*Figure 4.6: Short forecast error LSR plots after applying the error estimation methods, H-L or BAI, with a percentage of the available SST observations. Producing error for the winter season December-January-February.*

In order to mathematically assess the quality of both methods with reduced observational datasets, we use the discordance correlation coefficient $\rho(p, q)$. Equation (4.1) uses the results from the full dataset, $p$, and each subset of innovations, $q$, with the spatial covariance $s_{p,q}$, and variance $s_p^2$. This is then shown in figures 4.7 for the SDV and 4.8 for the short LSR of our error estimation methods

$$\rho(p, q) = 1 - \frac{s_{p,q}}{s_p^2 + s_q^2 + (E[p] - E[q])}. \tag{4.1}$$

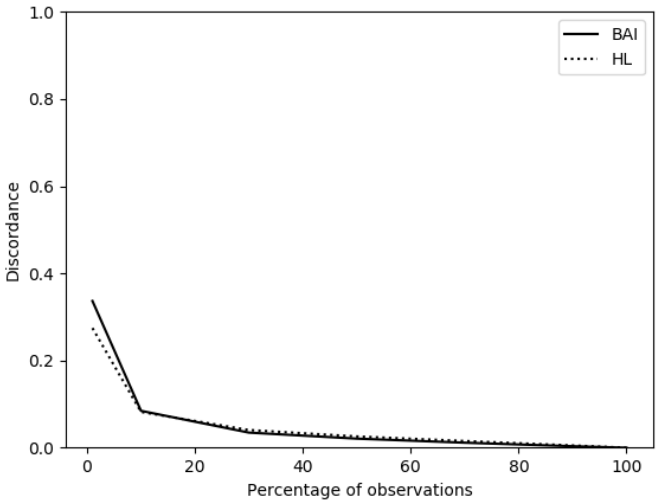



*Figure 4.7: Average seasonal discordance for the SDV produced by either analysis of innovations method. This seasonal average is for the year 2014, with only SST observations.*

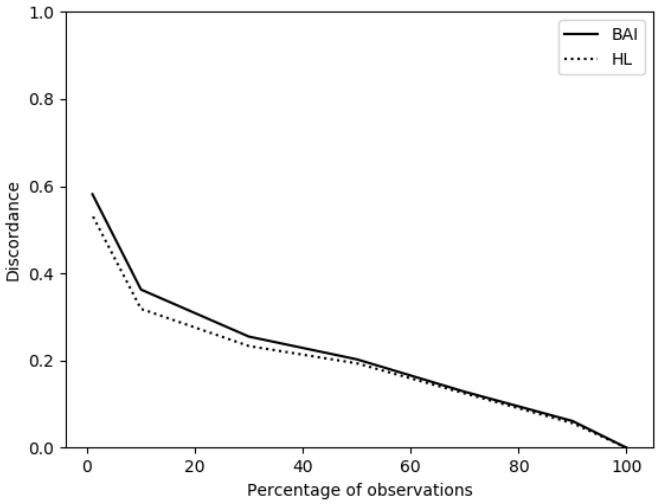



*Figure 4.8: Average seasonal discordance for the short length-scale weighting produced by either analysis of innovations method. Seasonal average is for the year 2014, with only SST observations.*

The discordance has a negative relationship with the subset percentage, as the percentage increases, the discordance decreases. This is to be expected, and reflects what can be see from the spatial plots of varying different samples. The important factor is the steepness of the regression, and how the gradient changes. A steep gradient suggests that a small increase in data set, will largely decrease the discordance, and vice versa. When the gradient is shallow, we have a plateau and we can assume that the estimates are approaching the true model error, as there is only a small increase when using more data.

Both methods have a very similar gradient curve and this means that neither can be considered to have a more rigorous statistical approximation. They both appear to be using the innovation data to produce error estimates that are converging towards the true value. It is also possible for us to state that the stronger gradient with respect to the percentage of data, for the BAI curve, that it is expected to respond faster to an increase in data availability. Both methods have a shallower gradient for SDV than the LSR as they approach the 100% dataset.

Extrapolating the curves above 100% suggests that the error estimates have begun to converge to the optimum forecast error, with the SDV converging faster than the LSR. Suggesting that by using more data we would achieve a closer estimate, but with very little change relative to the increase in data availability.

We also use the root mean square differences (RMSD) and bias of differences to compare the results from varying innovation subsets, figures (4.9, 4.10). The differences are between the results for the full innovation dataset and the ones produced using subsets of observations. For this statistical analysis the BAI method appears to perform better or equally well in comparison with the H-L, with the only exception in the length-scale ratios in the 1% subset. We believe that a possible reason behind an improvement in the results for our BAI method is due to reducing the impact of outliers by no longer using the isotropic binning of the innovations.

The bias results show some interesting characteristics of the error estimation methods, and how despite the similarities in final results, for smaller subsets of innovations the H-L begins to over estimate and produces a positive bias and the BAI method produces an underestimate. These can be seen in figure 4.10 where the points for BAI and H-L diverge.

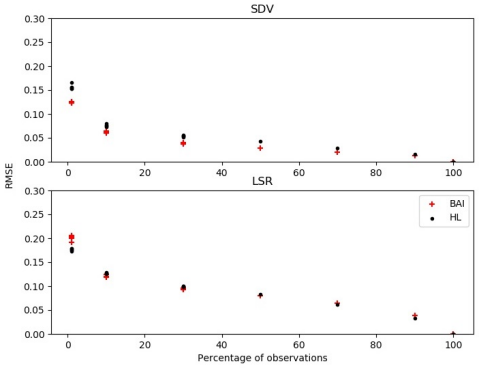



*Figure 4.9: Annual RMSD for the error SDV and short LSR produced by either analysis of innovations method for SST. This annual result is for the year 2014, with only SST observations.*

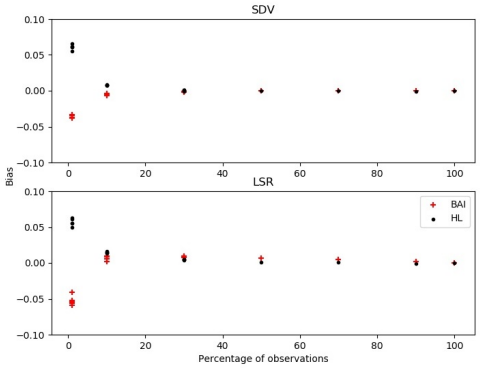



*Figure 4.10: Annual bias for the error SDV and short LSR produced by either analysis of innovations method for SST. This annual result is for the year 2014, with only SST observations.*

*Table 4.4: Table of average RMSE for NEMOVar innovations with each season. (Degrees Celsius)*

| RMSE | H-L | BAI | MO |
|---|---|---|---|
| Winter | 0.3665 | 0.3659 | 0.3633 |
| Summer | 0.8244 | 0.8228 | 0.7924 |

*Table 4.5: Table of average mean for NEMOVar innovations with each season. (Degrees Celsius)*

| Mean | H-L | BAI | MO |
|---|---|---|---|
| Winter | -0.0097 | -0.0084 | -0.0144 |
| Summer | -0.2359 | -0.2325 | -0.2057 |

## 4.6 SLA analysis

To further emphasize the impact of subsets of innovations, we have applied both the H-L and BAI method to the SLA innovations. This dataset is much less, at more than 100 times smaller dataset than the SST, which is similar to the smallest of our subsets. We use this example to demonstrate how the data size can affect the ability of the statistical analyses, when applying error estimation techniques to sparsely observed variables.

The general results from the SLA analysis are a worse estimate of the model error since a much smaller dataset is used and that the statistics will be a weak representation of the true model error. This occurs when we are approximating the background error covariance model to the innovations covariance, while we

have infinite data the statistics are an appropriate interpretation of the background error, but this is not the case for limited observation availability.

Within the SDV results, it is immediately noticeable that the SLA error estimates contain more NaN values (white cells) within the wet grid points. These NaN values are either a result of the limited data, or an issue with the analysis of innovations approach. When there is no observational data for certain grid points we cannot estimate the variance, but also, neither method aims to limit the model to only positive values of variance. Both methods will only try to find the best fit for the model to the statistics, which means that despite the majority of the error variances values being a valid, when the error is small both methods can return negative estimates. This error is more common with smaller datasets since the statistical analysis fluctuates around the true error more.

We can also see that the range for SLA error is much lower, as the error variance approaches zero there is an increased chance of NaN values. For the SLA analysis we have set the range as 0 to 0.3, this is much smaller than the temperature error variance, based on the nature of the variable. Temperature is measured in Celsius, 20-30 °C, which has a much higher range respective to the changes in SSH, which will be in metres and often less than 1 m.

The more interesting result from the SLA analysis is the short-length scale weighting, this output has many differences to our result for temperature and has a very large positive and negative weight. For us to include this into NEMOVar assimilation, the majority of the values would be removed as they are outside the range that NEMOVar accepts. The limitations for operational DA are set up so that any background covariance model that is created is guaranteed to

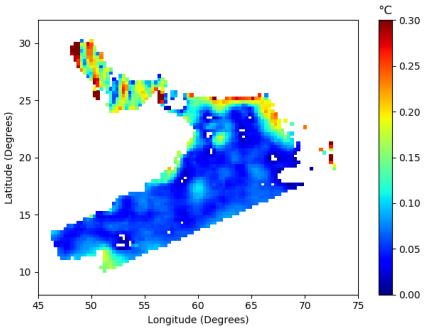
be positive semi-definite, and from our error estimates for SLA we would not be able to produce this without post-processing. In general it appears that the both of the error estimation methods have struggled to produce an acceptable error variance and LSR, and this will primarily be due the limitations in data availability.

There are also areas within the SLA results that have NaN values, before any post processing, and this is a case of grid points where we have no observations over the three month season. Without any data for certain locations its impossible to produce any error estimates here, the BAI and H-L are able to manipulate the data to produce some error here, but interpolating as part of post processing will give us a similar accuracy.

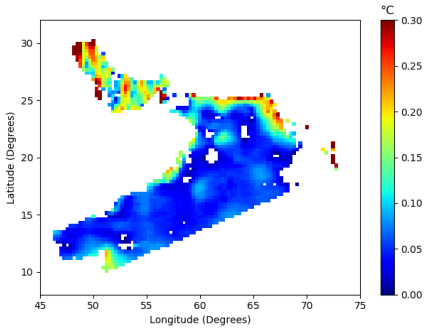
The LSR appear to be more sensitive to the availability of data, producing high fluctuations when there is insufficient statistics. The seasonal trends of the SST and SLA analysis proves this as they are consistent with the number of available observations seen in the table in section [4.2]. Further evidence can be seen by comparing the discordance plots for LSR. The gradient is generally very steep and suggests it is not converging towards the true model error LSR with small datasets. The inverse could be said for the error variance, the gradient of the discordance is still relatively steep towards the lower percentages, but appears to converge better than the LSR discordance.

## 4.7 NEMOVar assimilated runs

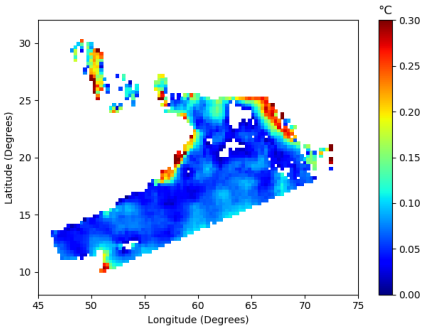
The aim for producing these error SDV and the LSR, is to then be able to run the NEMOVar assimilation. The model grid results are the graphic representation of the exact files that will be used to prepare the operational suite for AS20. We



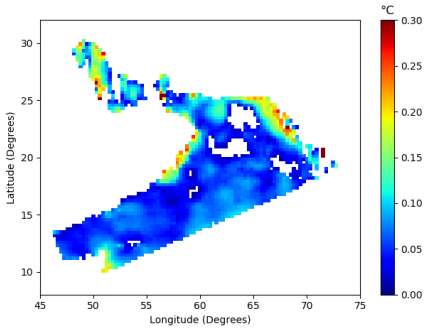

(a) Winter: H-L



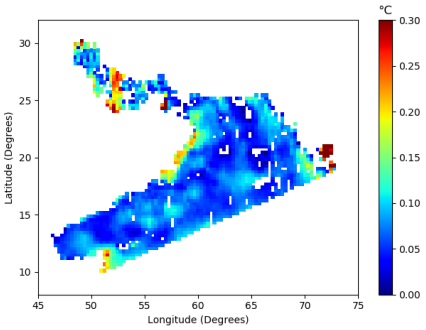

(b) Winter: BAI



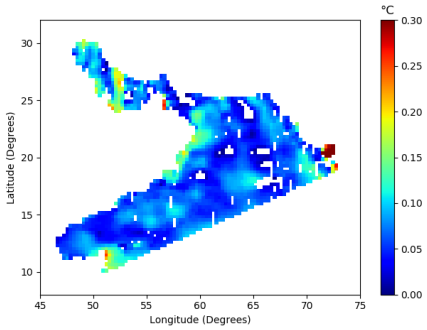

(c) Spring: H-L



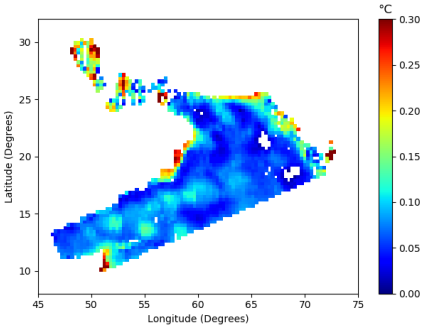

(d) Spring: BAI



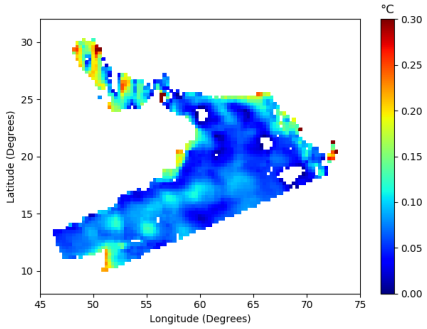

(e) Summer: H-L




(f) Summer: BAI

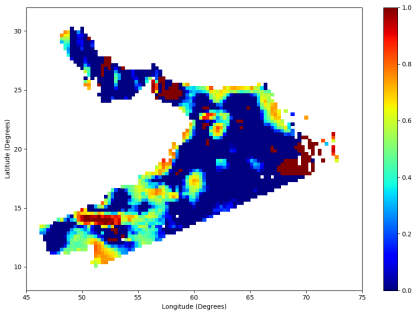



(g) Autumn: H-L

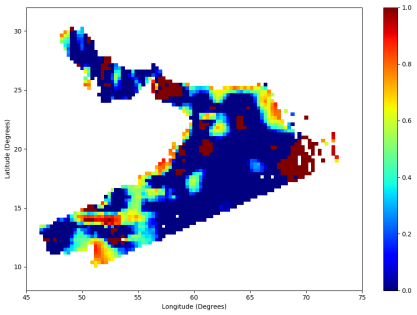



(h) Autumn: BAI

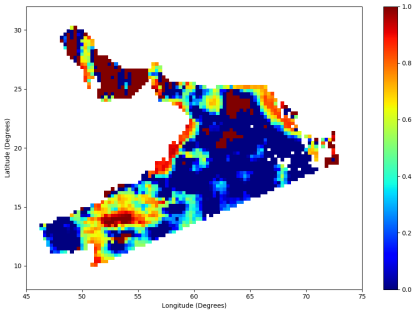
*Figure 4.11: The forecast error SDV on the coarse grid for AS20. Produced by either the H-L or BAI methods for SLA, with four independent seasons.*



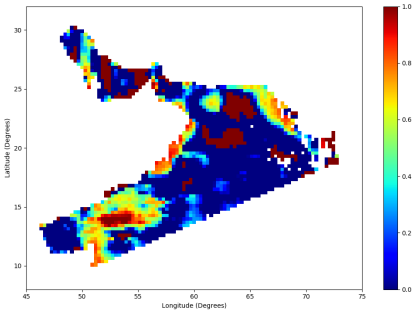

(a) Winter: H-L



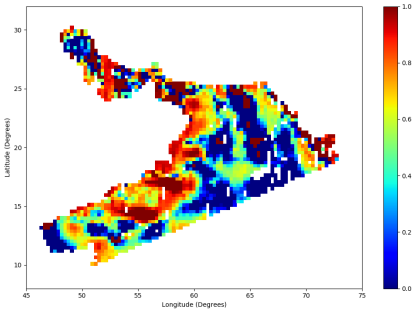

(b) Winter: BAI



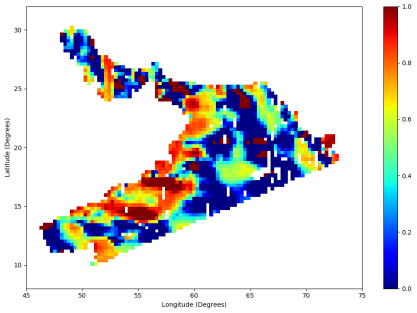

(c) Spring: H-L



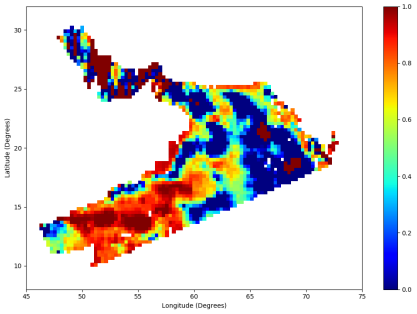

(d) Spring: BAI



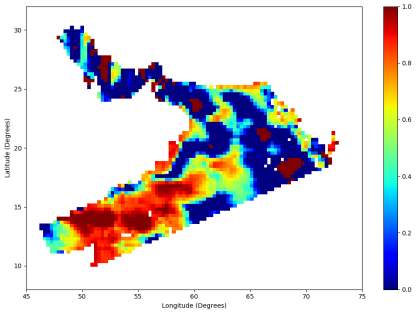

(e) Summer: H-L




(f) Summer: BAI




(g) Autumn: H-L




(h) Autumn: BAI

*Figure 4.12: The short forecast error LSR on the coarse grid for AS20. Produced by either the H-L or BAI methods for SLA, with four independent seasons.*

have run two four-week NEMOVar assimilated model runs, one for the summer and one for the winter. This produces feedback files with observational data and model data interpolated into the observation space, allowing us to compute the innovations, and we use these innovations as an assessment of forecast error. Since the observation error will be consistent for each method, any change in results between the methods will indicate a change in the forecast accuracy.

The NEMOVar runs we have done are for three types of error estimation, the two methods mentioned so far in the results section, as well as the an error SDV and LSR from the MO that is currently being used operationally at the POFC. This acts as a control result, but since we do not know the exact methods of production for this error, there is a limitation on the effective use of a comparison. The two assimilation runs are for 19/12/2017 to 16/01/2018 (winter) and 14/06/2018 to 12/07/2018 (summer), which each uses a consistent model restart file for each season with all 3 methods.

After the assimilation suite is run we use the innovations to produce a spatial mean and root mean square error (RMSE) for each day. Since the first day is produced from the same restart file all three methods will return the same forecasts. As time progresses each model run will produce its own restart for the next day, and they begin to deviate from each other.

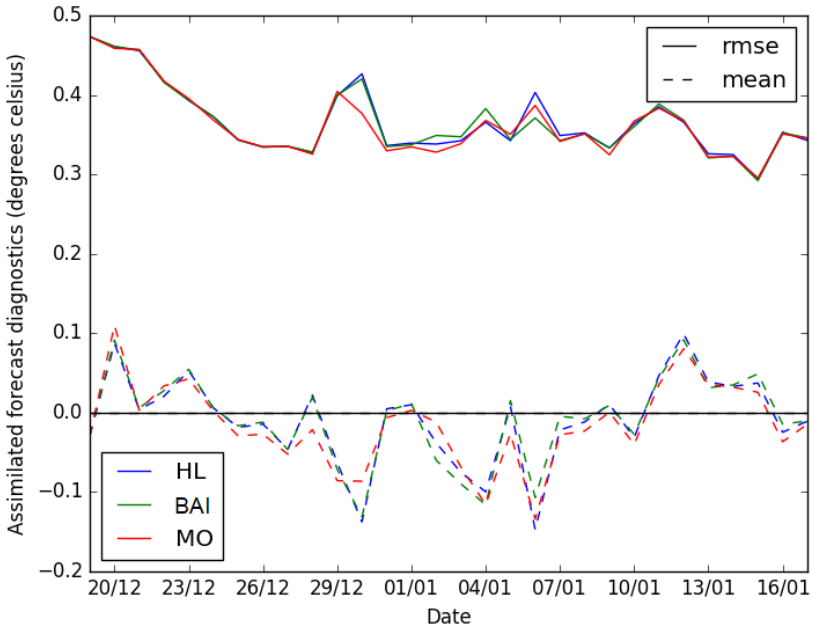



*Figure 4.13: RMSE and mean of innovations for temperature from a four week winter NEMOVar assimilated forecast.*

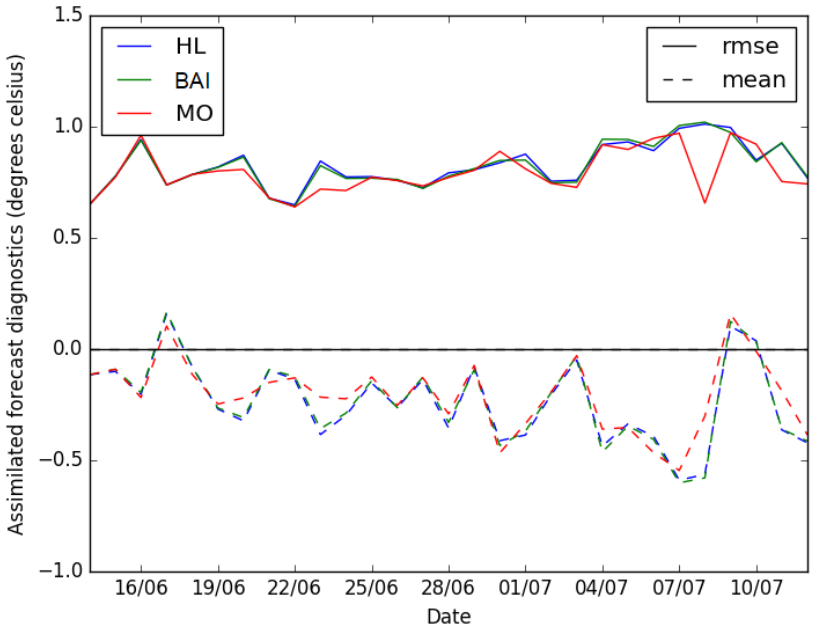



*Figure 4.14: RMSE and mean of innovations for temperature from a four week summer NEMOVar assimilated forecast.*

| RMSE | H-L | BAI | MO |
|---|---|---|---|
| Winter | 0.3665 | 0.3659 | 0.3633 |
| Summer | 0.8244 | 0.8228 | 0.7924 |

*Table 4.6: Table of average RMSE for NEMOVar innovations with each season. (Degrees Celsius)*

| mean | H-L | BAI | MO |
|---|---|---|---|
| Winter | 0.0097 | -0.0084 | -0.0144 |
| Summer | -0.2359 | -0.2325 | -0.2057 |

*Table 4.7: Table of average mean for NEMOVar innovations with each season. (Degrees Celsius)*

We have shown the average RMSE and mean in tables 4.6 and 4.7 respectively. By comparing the summer and winter seasons we can see that the average forecasting ability decreases significantly. Since the innovations have non-zero bias and the RMSE, or variability, of our error is much greater. This trend is mimicked by all three methods, and is likely a representation of our model and assimilations ability to produce forecasts. By revisiting figure 4.1, we can see that both the H-L and BAI method predicted an increase in error for the summer season, so this result supports the accuracy of our error estimation methods.

According to the average of the RMSE results over the two NEMOVar runs, the error produced by the BAI and H-L perform similarly well, with the BAI being slightly ahead. Both methods seem to be slightly worse than the MO error estimation in general, but since we do not know the exact methods or datasets used it is hard to determine the cause of this. We can also see on certain days that the differences fluctuate and sometimes this is more, since the forecast errors are based on the use of a seasonal average it is likely that the methods will more closely resemble the true error on certain days, and be less representative on other days, so a slight change in average does not discredit any method of error estimation. Instead we believe it is safer to say that the binless approach is still able to produce the error coefficients with the similar accuracy as operational error estimation methods.

## 4.8 Anisotropic BAI method

The original plans for our development of a novel error estimation method was to be able to create a system that would produce anisotropic background ECMs. This has been the aim of many researchers, as the isotropic assumption for background error is known to not be fully justified everywhere. However, there is significant difficulties in altering the Var system to allow for flow-dependent components. This is a result of the various preconditioning factors and normalisation components, as well as the outcome from the use of the diffusion scheme within the error covariance model which can become very complex when anisotropic.

We have described how we produced the anisotropic BAI method in section [3.5.4], where we create a covariance function for relative position as follows;

$$\tilde{f}(\mathbf{x},\mathbf{y}) = V\big[v\big(w_1\phi_{1,1}(\mathbf{x},\mathbf{y})+(1-w_1)\phi_{2,2}(\mathbf{x},\mathbf{y})\big)+(1-v)\big(w_2\phi_{1,2}(\mathbf{x},\mathbf{y})+(1-w_2)\phi_{2,1}(\mathbf{x},\mathbf{y})\big)\big]. \tag{4.2}$$

In this we have said, $V$ is the error variance, $v$ is the anisotropic-isotropic weighting, $w_1$ is the short-long LSR, and $w_2$ is the longitudinal-latitudinal LSR. For higher values of $v$ there should be more weight on the isotropic covariance, the $w_1$ ratio is the same as before where the closer to 1 the more weight is placed on the short-scale, and finally for $w_2$ the larger results suggest weighting a latitudinal stretch higher.

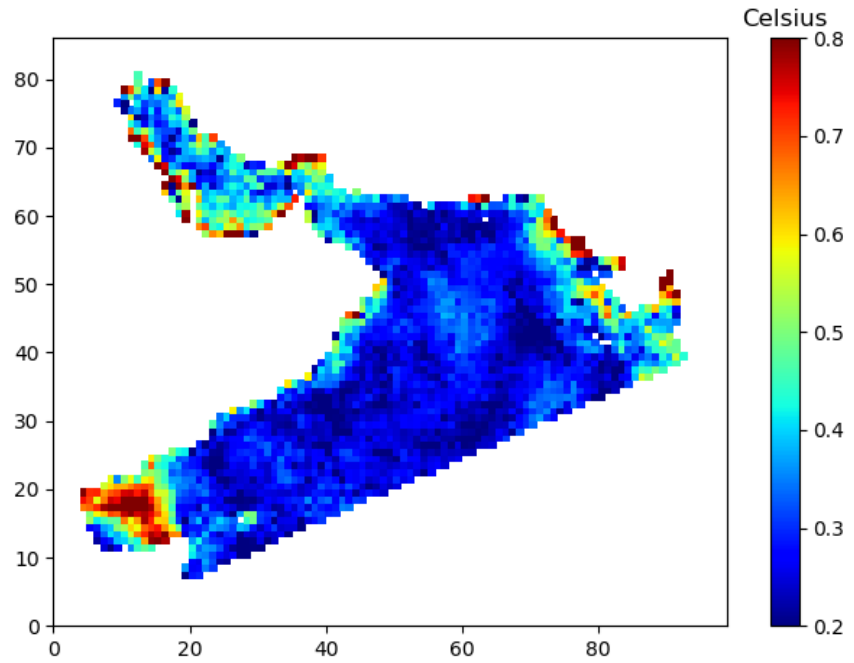



*Figure 4.15: The background error SDV produced using the BAI method with four 2D basis functions to include anisotropy. This plot is using the SST observations for the winter season.* $V$

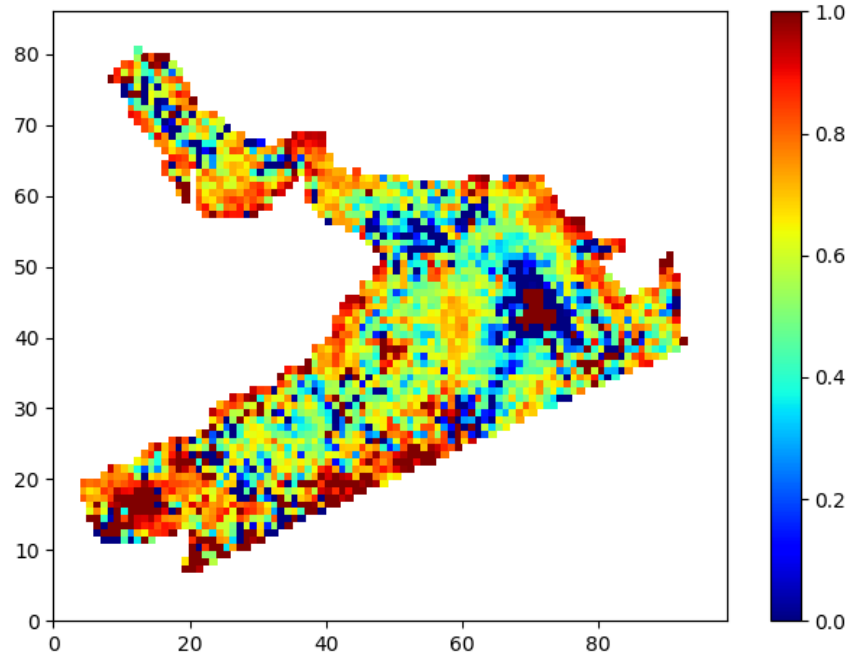



*Figure 4.16: The background error long-short length-scale weight produced using the BAI method with four 2D basis functions to include anisotropy. This plot is using the SST observations for the winter season.* $w_1$

We are able to produce the SDV and long-short LSR with reasonable similarities to the isotropic version. The biggest differences are noticeable within figure 4.16 when comparing this to figure 4.2b, despite there being some consistencies there are also stark differences. The changes to the long-short LSR suggests that the BAI method is taking consideration for the anisotropic components of the background error when we alter the covariance modelling function. This can also be seen when we look at the anisotropic weight and the longitudinal-latitudinal LSR, in figures 4.17 and 4.18 respectively.

In these plots it becomes much more difficult to make any positive conclusions for the anisotropic variation of the BAI method. The results look very sporadic, with almost no consistent shape across the domain. They also seem to show peaks that are outside the typical accepted range for weighting, and even with an extension to this range, as we have done for figure 4.18, there is still no location specific trend. This does suggest that the method is incorrectly quantifying the anisotropic weights, but we cannot say this for certain. There is a large possibility that background error anisotropy does not follow a consistent shape, and is in truth sporadic. We also have to remember that this is a simple application of anisotropy to our covariance model using equation (3.5.4), and we do not know how close to the true function this anisotropic version is.

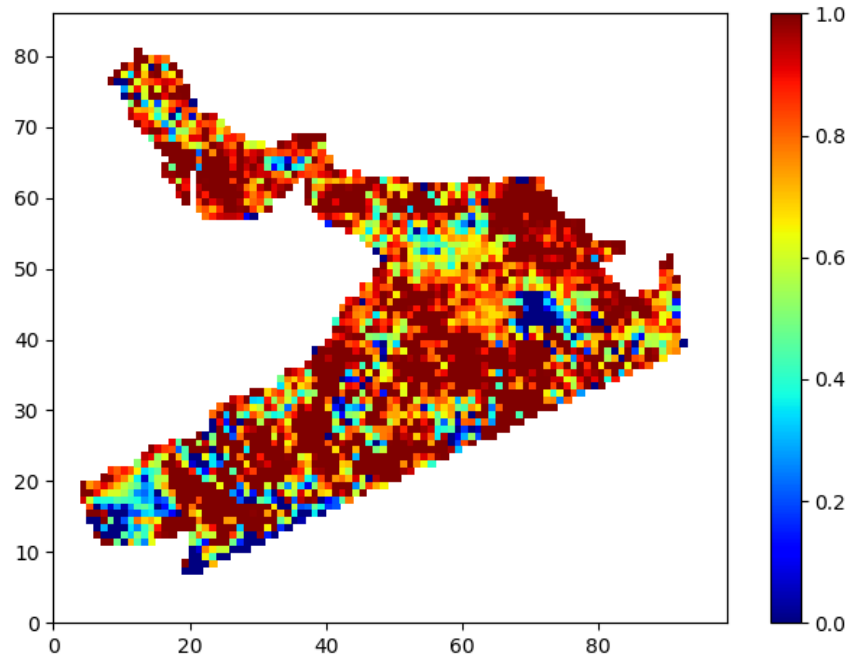



*Figure 4.17: The background error anisotropic weight produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for the winter season.* $v_1$

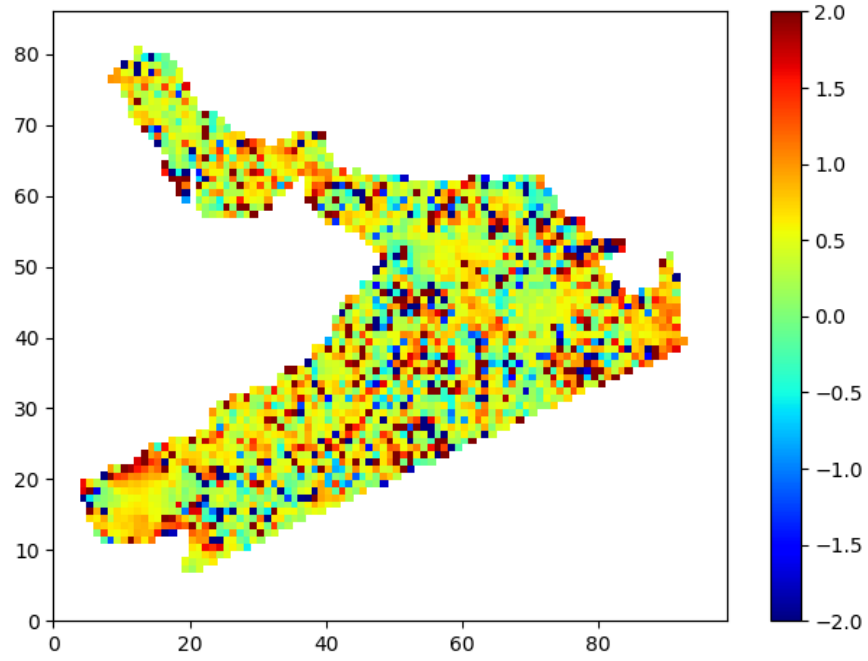



*Figure 4.18: The background error longitudinal-latitudinal length-scale weight produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for the winter season.* $v_2$

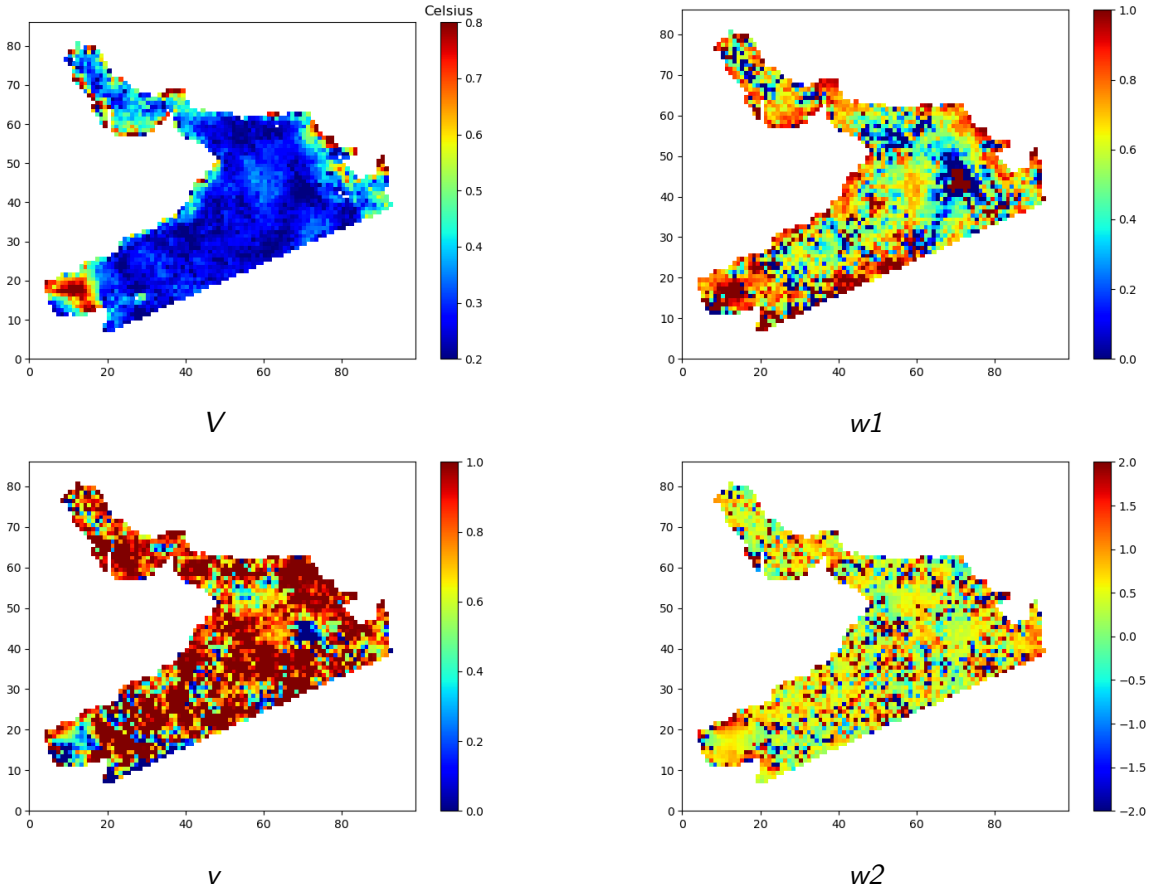



*Figure 4.19: The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for winter 2014.*

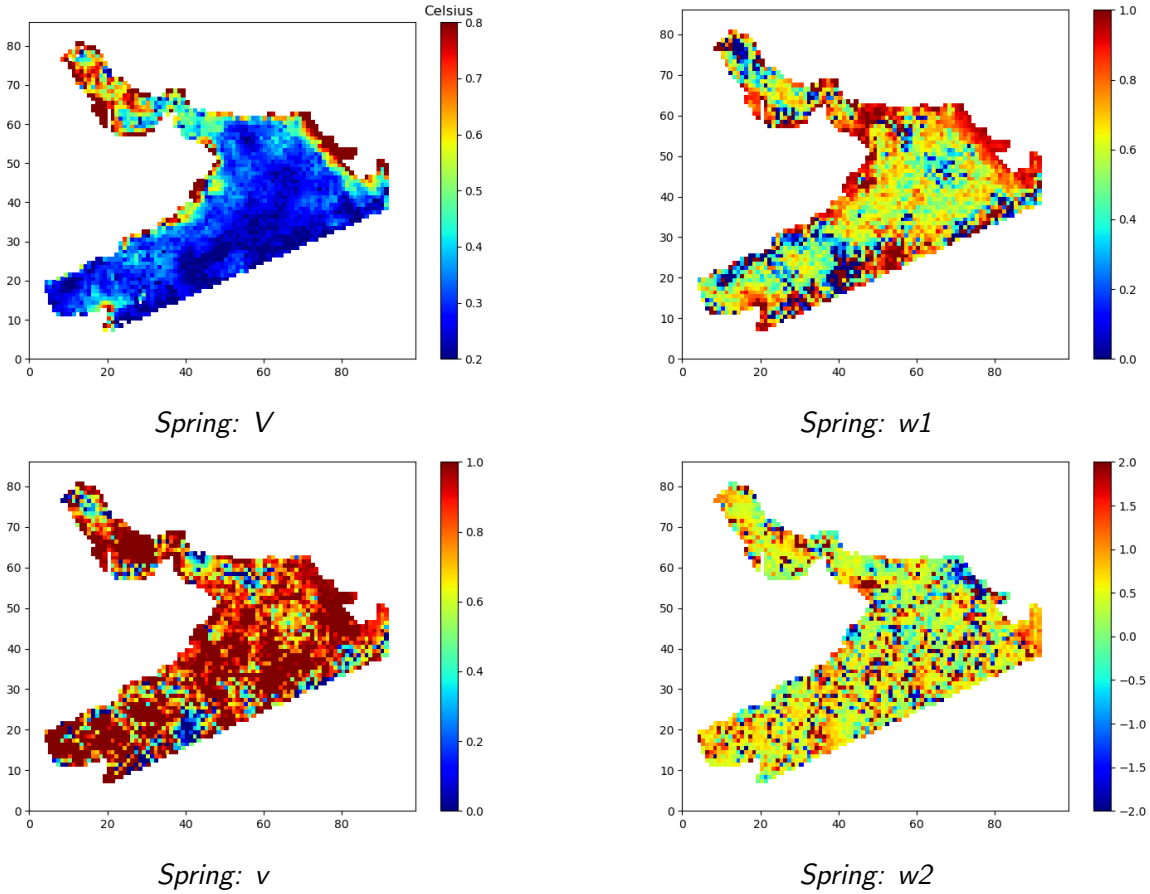
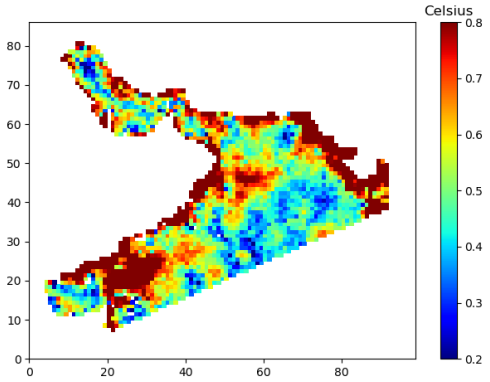
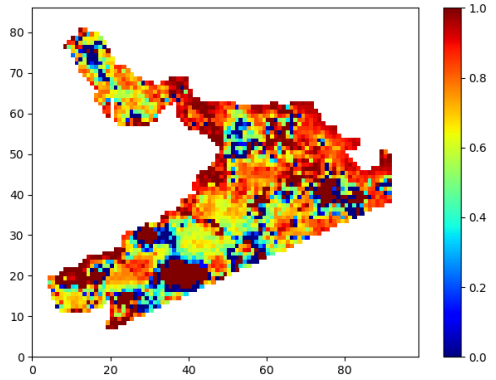


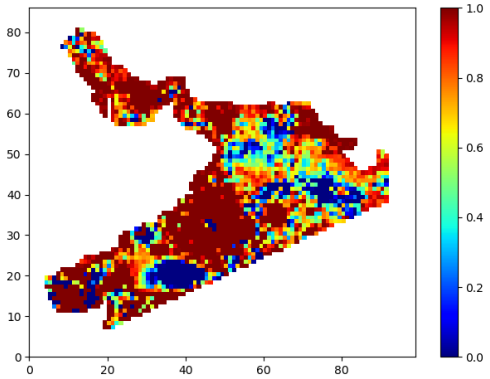

*Figure 4.20: The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for spring 2014.*



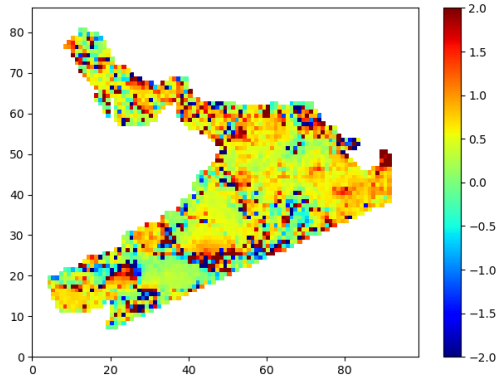

*Summer: V*




*Summer: w1*

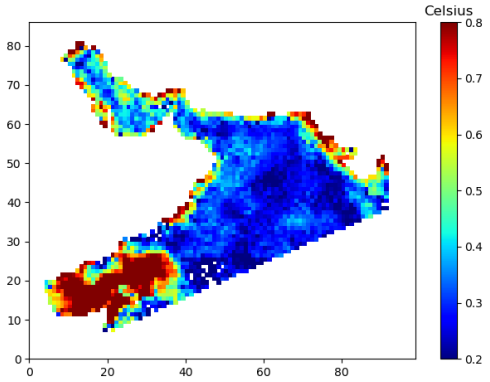



*Summer: v*

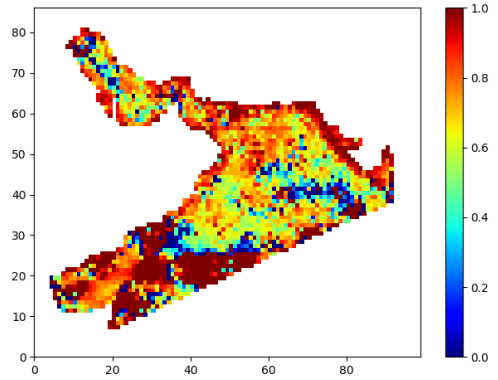



*Summer: w2*

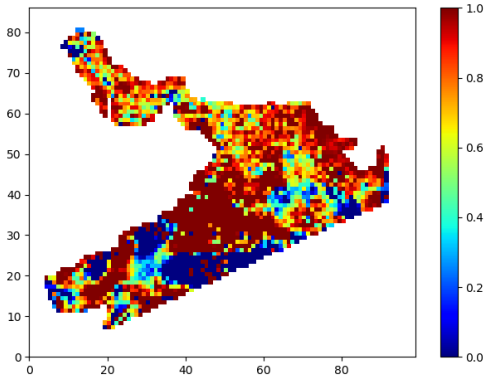
*Figure 4.21: The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for summer 2014.*



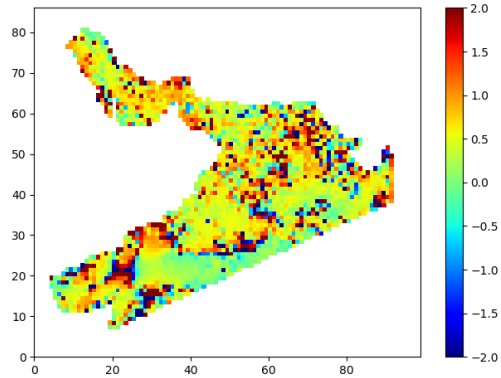

*Autumn: V*



*Autumn: w1*



*Autumn: v*



*Autumn: w2*

*Figure 4.22: The background error coefficients produced using the BAI method with four basis functions to include anisotropy. This plot is using the SST observations for autumn 2014.*

We have repeated the anisotropic application of the BAI method for all seasons and shown them in the above figure, and there are some more trends that can be noticed by viewing these results for each season. The first is the the short-long LSR seems to have a strong relationship with the anisotropic-isotropic weighting. The peaks in $w_1$ seems to coincide with low points in $v$. If our method is producing accurate estimates for the true error covariance, this relationship suggests that for a higher weight on shorter length-scales that there is a higher weight on the anisotropic error basis functions. This would then suggest that the affects of anisotropy are more applicable for shorter length-scale error covariances, due to the fact that the flow-dependent components of error will have a larger impact on localised covariance. What this means for the application of the anisotropic error covariance is still unknown and is something that will be developed in the future. What we have shown is that our method is able to produce an estimate of this form, and this is a potential approach for future anisotropic error covariances.

# Chapter 5

# Conclusion and future work

## 5.1 Conclusions

During the course of this thesis, and my research into DA techniques, we have been able to investigate some of the state-of-the-art methodologies and produce a novel error estimation approach. This required research into the core DA approaches of variational and sequential assimilation, alongside error estimation and optimization components for the NEMOVar operational suite. Gaining an understanding of the current systems used in advanced NWP for many researching institutes around the world, has allowed us to find weaknesses and limitations in the operational methods.

Once we had begun to grasp some of the core ideas and practical limitations of DA with ocean modelling, we moved onto being able to replicate some of these methods. Originally a simple DA suite was going to be independently constructed by the author, however the MO and POFC were working on a research project that was able to supply a 3DIncVar suite with the AS20 ocean model. From here we worked on understanding the suite, and becoming familiar enough with it to alter some namelist definition, and reproduce some of the input files. Specifically we wanted to be able to fully understand the background error SDV

and LSR, which are used during the assimilation suite to model the background ECM.

At this stage, while learning about the background error covariance modelling procedures of DA our general research questions began to change. Our original question we wanted to be able to answer was *(1) What is the affect of applying the isotropic assumption on the accuracy of the assimilation analysis for a regional model with active flow?* The isotropic assumption is used within Var, specifically for producing the background ECM using an isotropic diffusion operator. By improving the representation of this modelling process to include anisotropic error, we had hoped to improve the model forecast accuracy via DA. However, we discovered that there are restrictions on the background ECM, due to the construction process, that do not currently allow for anisotropic diffusion. If we wished to produce a fully anisotropic ECM we needed to ask *(2) What developments would be needed to implement an anisotropic background ECM into NEMOVar?*

We were able to subtly answer this question based on our studies of the AS20 operational NEMOVar; the largest development would be creating the anisotropic diffusion operator, this operator would then need to be able to produce positive semi-definite ECM or find an alternate method of preconditioning that didn't require this. The likelihood being that multiple diffusion operators are in a weighted combination is required for this, similar to how short and long length-scales are currently used. This diffusion scheme would need to be able to use a non-diagonal diffusion tensor, and the values of the tensor would require an appropriate anisotropic quantification. This definition for anisotropy within diffusion has been researched by [Weaver and Mirouze, 2012], where the process

of using anisotropic and homogeneous diffusion is studied. Following this, we then need to be able to normalise the diffusion. The normalisation matrix is mentioned within [3.2.2] as well, but we do not describe the details of the operational methods, as we did not replicate this. In principle, the normalisation factors do not need to change unless the diffusion operator changes. Once these were available, the next job is to estimate the error for anisotropy using a method that can calculate multi-dimensional error weightings and variance (such as the BAI). While we were investigating the required changes for a fully anisotropic ECM we began to realise the expense connected to including anisotropy, and then began to wonder, *(3) Would the cost of developing a fully anisotropic background ECM be worth the improvement in the final analysis?* This was an important aspect since there was no guarantee that the investment of time and research would result in a more accurate model forecast. To definitively answer question *(3)*, the full anisotropic ECM would be needed, but this does not mean that we cannot provide initial results which can then suggest a positive or negative outcome.

These questions, and the evolution of our research led us to studying the current error estimation methods, section [2.3], and asking *(4) What are the benefits and potential applications from the error estimation being replaced with an anisotropic option?* As we have mentioned, the background ECM is modelled using error SDV and LSR. If we wished to progress the first step would be to create a novel approach for error estimation that allowed for two dimensional analysis, as both the NMC and H-L are fundamentally 1D and isotropic. This led to the development of the BAI method in section [3.5], the isotropic application and then in turn the experimentation with simple anisotropic options for error estimation. We then ran experiments in chapter [4], where we tried to show how our alter-

nate method of error estimation is a viable option. We have removed some of the components from the H-L method, such as spatial binning and curve fitting schemes, and hence created a more flexible option for error estimation. We are able to use almost any definition for background error covariance modelling, including multi-dimensional and anisotropic.

The results for isotropic BAI method are comparable with the H-L method, (figures 4.15 and 4.16) however the anisotropic results are more interesting, (figures 4.17 and 4.17) and help us answer question *(4)*. We can see a similar value for the error SDV and LSR from the anisotropic method, with some changes in the extremes for LSR. When we view the anisotropic weights we struggle to see any consistent shape. This could suggest that either; anisotropy is sporadic and hence will be very difficult to model with a single seasonal value, our simple anisotropic representation is not appropriate for the true background error, or that the Arabian Sea and AS20 background error has no flow-dependence.

By investigating the error estimation methods and producing the BAI approach we have begun to answer *(4)*. We can see that the possibilities of anisotropic error estimation is flexible and equal on computational costs. The difficulty with the application of this is that we need to define the anisotropic covariance model, and there is freedom to define this however best suits the true background error. We have been able to state some of the requirements for the anisotropic background ECM in our answer to question *(2)*.

In our attempts to answer question *(3)*, we have discovered that it is very difficult to have only one answer. Since there is multiple ways to develop anisotropy into the background ECM, there may be other options that do not require the same

costs. For our approach, with anisotropic BAI, we believe that there is a high cost in terms of research to fully understand the optimum anisotropic covariance modelling. However, if this research is done then there should be visible improvements in the representation of the background error, and hence the accuracy of the forecasts. The positive impact from anisotropy will be purely based on the flow-dependence of the background error, if a domain and ocean model does not contain flow-dependence then the cost will not be worth the improvement in analysis.

This brings us to our attempts to answer question *(1)*, this is a large question with multiple difficulties. Once an anisotropic assimilation is created the answer would be much clearer but without a working option our answers can only be from our experiences and practices. I believe that the largest affect would be an increase in cost, as I have mentioned. Despite the error estimation being computationally comparable if additional diffusion operators are required then this will increase the processing time of NEMOVar. We must remember that it is generally known in the DA community and is stated in previous research, that the use of the isotropic assumption is not justified, and still no anisotropic approach has been developed for operational use. Since this is the case, then we would have to expect that including anisotropy in specific cases will improve analysis, but the cost of researching and developing this is known to be large. This is then devalued by the fact that there is still the possibility that the background error of any specific ocean model will only have a minor or negligible flow-dependence.

## 5.2 Future works

Since we are claiming that the BAI method is valid, then the assumption is that the weights which we receive are the best for the covariance model, but a different anisotropic covariance model could be used. We have kept in mind during this thesis that we may need to expand our research to investigate the potential for other anisotropic covariance models. which would then be implemented into the diffusion operator (the model and diffusion need to be consistent).

If the BAI method is producing the best weights for the covariance model, and we can be sure that this model is the best for the background error then we may need to investigate the application of the BAI in other regional models. By experimenting with other domains we may find evidence for the positive use of anisotropic weightings, and it may just be our domain which does not showcase flow-dependence. If the results are consistent for multiple domains and more appropriate error covariance models, then the third and final option, is that including anisotropy for the background error will not reduce the differences between our estimated ECM and the true ECM.

There is another potential research following this thesis, the first is to spend more time investigating the background error covariance of the AS20 model. If we were to spend more resources into determining an optimum anisotropic background error covariance model then this we may show more evidence of using anisotropy for error. Which will lead to us being able to improve the potential for including anisotropic diffusion into NEMOVar with the improved covariance model.

If we cannot not find any evidence for flow-dependence in the AS20 model, then anisotropy may not improve our forecast accuracy. It would be good to expand our research so that an assessment process is created to determine the flow dependence of the background error within current models. I believe that there is potential for this to be done by adapting the current stage of BAI. Since it is possible for us to quantify some aspect of the flow-dependence for background error, we could determine the anisotropy that may be represented. Using this we might find new ocean models and domains where including anisotropy would greater improve the DA forecasts.

Since question *(1)* specifies that it is for *a regional model with active flow*, One final option for would be to expand the application to a global models. The anisotropy in a global model will likely be less deterministic, however since the global resolution is lower, then the margin of background error is likely to be higher and this may lead to more visible anisotropic trends. An early option for this experimentation would be to use a global model and limit it to the domain of the AS20, and then compare the results for the BAI method and its anisotropic weightings.

# Bibliography

[Anderson et al., 1996] Anderson, D. L. T., Sheinbaum, J., and Haines, K. (1996). Data assimilation in ocean models. *Reports on Progress in Physics*, 59(10):1209.

[Bannister, 2008] Bannister, R. N. (2008). A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134(637):1951–1970.

[Barth et al., 2008] Barth, A., Azcarate, A. A., Joassin, P., Beckers, J.-M., and Troupin, C. (2008). Introduction to Optimal Interpolation and Variational Analysis. *GeoHydroDynamics and enviroment research*.

[Bin et al., 2000] Bin, W., Zou, X., and Zhu, J. (2000). Data Assimilation and Its Applications. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11143–11144.

[Bloom et al., 1996] Bloom, S. C., Takacs, L. L., da Silva, A. M., and Ledvina, D. (1996). Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124(6):1256–1271.

[Bocquet, 2014] Bocquet, M. (2014). Introduction to the principles and methods of data assimilation in the geosciences. *Master M2 OACOS & WAPE Ecole des Ponts ParisTech*.

[Bormann and Bauer, 2010] Bormann, N. and Bauer, P. (2010). Estimates of spatial and interchannel observation-error characteristics for current sounder

radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Quarterly Journal of the Royal Meteorological Society*, 136(649):1036–1050.

[Bouttier and Courtier, 1999] Bouttier, F. and Courtier, P. (1999). Data assimilation concepts and methods March 1999. *Meteorological training course lecture series. ECMWF*.

[Cao et al., 2010] Cao, X.-Q., Zhang, W.-M., Song, J.-Q., and Zhang, L.-L. (2010). Modeling Background Error Covariance in Variational Data Assimilation with Wavelet Method. *Institute of Electrical and Electronics Engineers (IEEE)*.

[Courtier et al., 1994] Courtier, P., Thépaut, J.-N., and Hollingsworth, A. (1994). A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387.

[Daley, 1993] Daley, R. (1993). *Atmospheric Data Analysis*, volume 89 of *Cambridge Atmospheric and Space Science Series*. Cambridge University Press.

[Deckmyn and Berre, 2005] Deckmyn, A. and Berre, L. (2005). A Wavelet Approach to Representing Background Error Covariances in a Limited-Area Model. *Monthly Weather Review*, 133(5):1279–1294.

[Dimet and Talagrand, 1986] Dimet, F.-X. L. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110.

[ECMWF, 2018] ECMWF (2018). *IFS Documentation CY45R1 Part II : Data Assimilation*, chapter Cy41r1, page 103. Number 2 in IFS Documentation. ECMWF.

[Fisher, 1998] Fisher, M. (1998). Minimization algorithms for variational data assimilation. *Recent Developments in Numerical Methods for Atmospheric Modelling*, pages 364–385.

[Fisher, 2001a] Fisher, M. (2001a). Assimilation techniques (3): 3dVar. *ECMWF*.

[Fisher, 2001b] Fisher, M. (2001b). Assimilation techniques (4): 4dVar. *ECMWF*.

[Fisher, 2001c] Fisher, M. (2001c). Assimilation techniques (5): Approximate Kalman filters and singular vectors. *ECMWF*.

[Fisher, 2003] Fisher, M. (2003). Background error covariance modelling. In *Seminar on Recent Development in Data Assimilation for Atmosphere and Ocean*, pages 45–63.

[Gandin, 1965] Gandin, L. (1965). *Objective Analysis of Meteorological Fields*. Quaterly journal of the royal meteorological society.

[Haben et al., 2011] Haben, S., Lawless, A., and Nichols, N. (2011). Conditioning and preconditioning of the variational data assimilation problem. *Computers & Fluids*, 46(1):252–256.

[Heilliette and Garand, 2007] Heilliette, S. and Garand, L. (2007). A practical approach for the assimilation of cloudy infrared radiances and its evaluation using airs simulated observations. *Atmosphere-Ocean*, 45(4):211–225.

[Hollingsworth and Lonnberg, 1986] Hollingsworth, A. and Lonnberg, P. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data. part I: The wind field. *Tellus A*, 38A(2):111–136.

[Holm, 2008] Holm, E. V. (2008). Lecture notes on assimilation algorithms. *European Centre for Medium-Range Weather Forcecasts*.

[Houtekamer and Mitchell, 1998] Houtekamer, P. L. and Mitchell, H. L. (01 Mar. 1998). Data assimilation using an ensemble kalman filter technique. *Monthly Weather Review*, 126(3):796 – 811.

[Houtekamer and Zhang, 2016] Houtekamer, P. L. and Zhang, F. (01 Dec. 2016). Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12):4489 – 4532.

[Ide et al., 1997] Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C. (1997). Unified Notation for Data Assimilation : Operational, Sequential and Variational (gtSpecial IssueltData Assimilation in Meteology and Oceanography: Theory and Practice). *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B):181–189.

[Kalman et al., 1960] Kalman, R. E. et al. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

[Kalnay and Yang, 2010] Kalnay, E. and Yang, S.-C. (2010). Accelerating the spin-up of ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1644 – 1651.

[Kucukkaraca and Fisher, 2006] Kucukkaraca, E. and Fisher, M. (2006). Use of analysis ensembles in estimating flow-dependent background error variances. Shinfield Park, Reading.

[Lahoz and Schneider, 2014] Lahoz, W. A. and Schneider, P. (2014). Data assimilation: making sense of Earth Observation. *Frontiers in Environmental Science*, 2.

[Martin et al., 2007] Martin, M. J., Hines, A., and Bell, M. J. (2007). Data assimilation in the FOAM operational short-range ocean forecasting system: a description of the scheme and its impact. *Quarterly Journal of the Royal Meteorological Society*, 133(625):981–995.

[Mogensen et al., 2012] Mogensen, K., Balmaseda, M. A., and Weaver, A. (2012). The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for System 4. *European centre for medium range weather forecasts techinical memoranda*.

[Parrish and Derber, 1992] Parrish, D. F. and Derber, J. C. (1992). The National Meteorological Center's Spectral Statistical-Interpolation Analysis System. *Monthly Weather Review*, 120(8):1747–1763.

[Schowengerdt, 2006] Schowengerdt, R. (2006). *Remote Sensing: Models and Methods for Image Processing*. Elsevier Science.

[Shewchuk, 1994] Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Technical report, School of computer science Carnegie Mellon University.

[Talley et al., 2011] Talley, L. D., Pickard, G. L., Emery, W. J., and Swift, J. H. (2011). Data Analysis Concepts and Observational Methods. *Descriptive Physical Oceanography*, pages 147–186.

[Thépaut, 2003] Thépaut, J.-N. (2003). Satellite data assimilation in numerical weather prediction: an overview. *Meteorological Training Course Lecture Series, ECMWF, Reading*.

[Waters et al., 2014] Waters, J., Lea, D. J., Martin, M. J., Mirouze, I., Weaver, A., and While, J. (2014). Implementing a variational data assimilation system

in an operational 1/4 degree global ocean model. *Quarterly Journal of the Royal Meteorological Society*, 141(687):333–349.

[Weaver and Courtier, 2001] Weaver, A. and Courtier, P. (2001). Correlation modelling on the sphere using a generalized diffusion equation. *Quarterly Journal of the Royal Meteorological Society*, 127(575):1815–1846.

[Weaver et al., 2006] Weaver, A., Deltel, C., Machu, E., Ricci, S., and Daget, N. (2006). A multivariate balance operator for variational ocean data assimilation. *Technical Memorandum*, (491):19.

[Weaver and Mirouze, 2012] Weaver, A. T. and Mirouze, I. (2012). On the diffusion equation and its applications to isotropic and anisotropic correlations modelling in variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 139(670):242–260.

[Webster, 2020] Webster, M. (2020). Bayes' theorem. Webpage. [online].

[Weisstein, 2020] Weisstein, E. W. (2020). Covariance - from mathworld – a wolfram web resource. Webpage. [online].

[Wikipedia, 2017] Wikipedia (2017). Estimation of covariance matrices. Accessed: 2017-04-03.

# Appendix A

# Equivalence in minimizing the norm

During section 3.5.3, we need to minimize a norm of innovation covariances, in order to find the forecast error SDV and LSR. The minimization of the norm, equation (3.25), is equivalent to solving the inner product of the innovations, (3.27). Below is the proof for this equivalence.

$$\min_{m_1,m_2} \left\| \tilde{f}(\mathbf{r}) - F(\mathbf{r}) \right\| \tag{A.1}$$

$$\langle \tilde{f}, \phi_n \rangle = \langle F, \phi_n \rangle, \quad \text{for } n \in \{1, 2\} \tag{A.2}$$

From calculus we know that the local minimum occurs when the partial derivative with respect to $m_1$ and $m_2$ are both equal to zero;

$$\min_{m_1,m_2} \left\| \tilde{f} - f \right\| \equiv \begin{cases} \frac{\partial}{\partial m_1} \left\| \tilde{f} - f \right\| = 0 \\ \frac{\partial}{\partial m_2} \left\| \tilde{f} - f \right\| = 0 \end{cases} \tag{A.3}$$

Hence by solving both derivatives we can achieve an expression for our error variance and length-scale. Before we can solve a derivative with respect to $m_n$, we use the naturally defined norm for inner product spaces, where;

$$\|x\| = \sqrt{\langle x, x \rangle} \tag{A.4}$$

Therefore;

$$\frac{\partial}{\partial m_n}\left\|\tilde{f}-f\right\|^2=\frac{\partial}{\partial m_n}\langle\tilde{f}-f,\tilde{f}-f\rangle \tag{A.5}$$

Which we use to solve equation (A.1) since;

$$\min_{m_1,m_2}\left\|\tilde{f}-f\right\|=\min_{m_1,m_2}\left\|\tilde{f}-f\right\|^2 \tag{A.6}$$

Since we are not going to directly compute the inner product in this proof, we will not define any dependencies of $f$ or $\tilde{f}$ as it is not necessary, but they must be treated as if they are consistent for the inner product to be well defined. In order to solve the derivative we can expand $\tilde{f}$, since the form function is a linear combination of $m_1$ or $m_2$, but the statistics $F$ are not;

$$\begin{aligned}\langle\tilde{f}-f,\tilde{f}-f\rangle =& m_1^2\langle\phi_1,\phi_1\rangle+m_2^2\langle\phi_2,\phi_2\rangle+2m_1m_2\langle\phi_1,\phi_2\rangle\\ &-2m_1\langle\phi_1,F\rangle-2m_2\langle\phi_2,F\rangle\end{aligned} \tag{A.7}$$

$$\frac{\partial}{\partial m_1}\langle\tilde{f}-F,\tilde{f}-F\rangle=2m_1\langle\phi_1,\phi_1\rangle+2m_2\langle\phi_1,\phi_2\rangle-2\langle\phi_1,F\rangle=0 \tag{A.8}$$

$$\langle m_1\phi_1+m_2\phi_2,\phi_1\rangle=\langle F,\phi_1\rangle \tag{A.9}$$

We can then do the same for $m_2$;

$$\frac{\partial}{\partial m_2}\langle\tilde{f}-F,\tilde{f}-F\rangle=2m_2\langle\phi_2,\phi_2\rangle+2m_1\langle\phi_1,\phi_2\rangle-2\langle\phi_2,F\rangle=0 \tag{A.10}$$

$$\langle m_1\phi_1 + m_2\phi_2, \phi_2\rangle = \langle F, \phi_2\rangle \tag{A.11}$$

Which when compared with the system of equations (A.2);

$$\min_{m_1,m_2} \left\| F - \tilde{f} \right\| \equiv \begin{cases} \langle m_1\phi_1 + m_2\phi_2, \phi_1\rangle = \langle F, \phi_1\rangle \\ \\ \langle m_1\phi_2 + m_2\phi_2, \phi_2\rangle = \langle F, \phi_2\rangle \end{cases} \tag{A.12}$$

Our method is then a problem of finding $m_1$ and $m_2$, such that both equations in (A.12) are satisfied. This creates a pair of simultaneous equations which can be solved, for simplicity we write this as the following inner product;

$$\langle \tilde{f}, \phi_n\rangle = \langle f, \phi_n\rangle, \quad \text{for } n \in \{1, 2\} \tag{A.13}$$

# Appendix B

# Kolmogorov's strong law of large numbers

In section 3.5.3 we use the strong law of large numbers with the Kolmogorov's condition. There are many mentions of this theorem and proofs online, but we have included a description of the main principles to potentially help the reader understand our use of the law.

The law of large numbers is a probability theorem that helps describe the results from performing an experiment a large number of times. The average of the results obtained from a larger number of trials should be close to the expected value and will tend to become closer as more trials are performed.

The law's general principle is formed for the case where $X_1$, $X_2$,... is a sequence of independent and identically distributed random variables, with a constant expected value. For both the strong and weak law they state that the sample average;

$$\overline{X_n} = \frac{1}{n}(X_1 + X_2 + ... + X_n) \tag{B.1}$$

will converge to the expected value. $\overline{X_n} \to \mu$ for $n \to \infty$. The strong law adds

an additional component, saying that the sample average converges strongly to the expected value;

$$\Pr\left(\lim_{n\to\infty} \overline{X_n} = \mu\right) = 1 \tag{B.2}$$

*If the summands are independent but not identically distributed, then;*

$$\overline{X_n} - E\left[\overline{X_n}\right] \xrightarrow{a.s.} 0 \tag{B.3}$$

*Provided that each $X_k$ has a finite second moment and;*

$$\sum_{k=1}^{\infty} \frac{1}{k^2}\mathrm{Var}[X_k] < \infty \tag{B.4}$$

Strong convergence is also called *Almost Sure* (a.s. for short), this version of the law of large numbers is called the strong law because random variables which converge strongly are guaranteed to converge weakly, in probability.

# Appendix C

# Rose and Cylc

Rose is a toolkit for writing, editing and running application configurations. A Rose suite is compiled to control each assimilation run, defining nameslist parameters and specifications for the high performance computing (HPC) to use parallel processing, as well as any required task definitions. Using this configuration Cylc takes control of the workflow, ensuring that no jobs begin processing if their requirements are not met. Cylc is a workflow engine, a system that automatically executes tasks according to their schedules and dependencies.

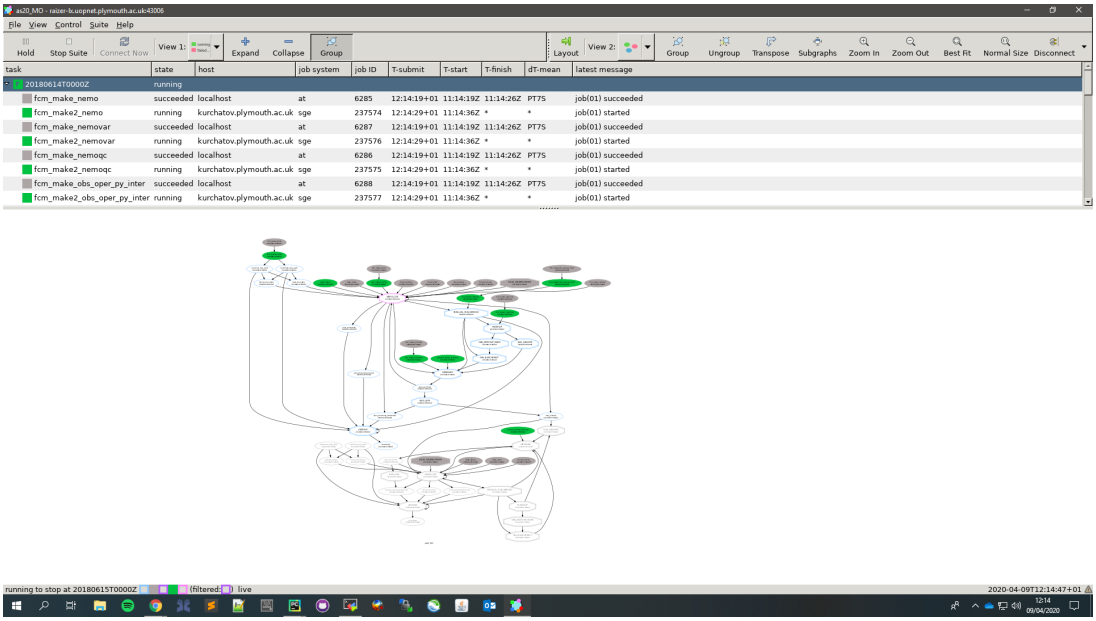



*Figure C.1: Example Cylc workflow control for AS20 NEMOVar rose suite.*

Rose also defines the jobs that will be run through the suite. Either by using

the shell scripts for each job, or by building an application using flexible configuration management (FCM). FCM is a modern Fortran building system, and will create executables for parallel processing on the HPC. The larger jobs like the ocean model and assimilation are built from Fortran scripts and then runs as single applications.

Our operational suite begins by running pre-processing scripts preparing the data and building the applications, then the first large job to be run is the NEMO observation operator. Preparing the cost function requires the observation operator and background data, so the job of this NEMO observation operator is to prepare both of these. From here the NEMOVar application is run, after some smaller processes for bias correction and pre-processing of the error covariance files. This application creates the full background ECM when needed, and then is able to produce the cost function from the rest of the data. Once the cost function is minimized the NEMOVar task outputs the analysis increments, which are systematically added to the model restart file over a specific time-period, IAU, and finally the last job is the running of the ocean model with the new restart file.