UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Thinning Algorithms for Remote Sensing Observations in Support of Ocean Data Assimilation

Permalink

https://escholarship.org/uc/item/583091vq

Author

Matranga, Juliana

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THINNING ALGORITHMS FOR REMOTE SENSING OBSERVATIONS IN SUPPORT OF OCEAN DATA ASSIMILATION

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE

in

SCIENTIFIC COMPUTING AND APPLIED MATHEMATICS

by

Juliana Matranga

June 2021

The thesis of Juliana Matranga
is approved:

_____
Professor Christopher Edwards

_____
Professor Andrew Moore

_____
Professor Nicholas Brummell

_____
Professor Pascale Garaud

_____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

CONTENTS

LIST OF FIGURES

Thinning Algorithms for Remote Sensing Observations in Support of Ocean Data Assimilation

By Juliana Matranga

Abstract

Data assimilation is a method to combine the numerical solutions calculated by circulation models with observations from different platforms in order to obtain an optimal estimate of the state of the system. The computational and numerical difficulties associated with processing the increasing number of observations necessitate the use of thinning techniques to reduce the number of data assimilated. In this thesis the impact of thinning two types of particularly voluminous data sets on the overall performance of an ocean data assimilation system were evaluated. In particular, an analysis of the ocean circulation along the U.S. West Coast with a 10 km resolution grid was performed, spanning a period of one year. Two different thinning methods were tested: an intelligent data thinning (IDT) algorithm to thin gridded satellite sea surface temperature (SST) and a simple thinning algorithm to reduce the volume of radial sea surface velocity measured by a network of coastal high frequency radars. The SST data were thinned by discarding data in regions with low spatial variability while retaining data in regions of high spatial variability. Conversely, the radar observations were averaged to create "super observations" consistent with the resolution of the model grid and prior assumptions about observation errors. The full and thinned data sets were assimilated using a 4-dimensional variational (4D-Var) data assimilation algorithm in the Regional Ocean Modeling System (ROMS). A statistical analysis of the diagnosed background and observation errors showed that the thinning experiments were well-behaved. Furthermore, the innovation and residual vectors (*i.e.* the difference between each observation and its prior and posterior model

counterpart) in all cases generally satisfied the assumption of Gaussian distributions. Additionally, the topology of a the total error covariance matrix of the data assimilation system was explored via its eigen space. The thinning experiments amplified the eigen spectrum, modified the condition number, and in particular thinning SST changed the aspect ratio of the hyperellipse defined by the covariance matrix to change. Lastly the impact of each type of observation on the analyses was quantified for the different thinning methods, suggesting that the radial velocities thinning was perhaps too severe, while the thinning of SST leveled the impact different observations had on the DA analysis. Overall results showed that the thinning did not significantly degrade the analysis, hence the next step will be to test these algorithms in a near-real time forecasting system.

Acknowledgments

I'd like to thank my family and friends for their unconditional support. I'd also like to thank my professors at San Francisco State University, in particular Dr. John Monteverdi and Dr. Piero Mazzini, because their dedication to teaching and their encouragement helped me to become more confident in myself.

I'm grateful to all my professors in the SciCAM program, because their classes were unbelievably helpful and enlightening. In particular, I'd like to thank Dr. Pascale Garaud and Dr. Nicholas Brummell for their guidance throughout these years and for being part of my thesis committee. Also, I feel extremely fortunate to have had the opportunity to work on this project with Dr. Christopher Edwards and Dr. Andrew Moore, whose exceptional wisdom and kindness exceeded all my expectations. I'd also like to acknowledge with gratitude the financial support from the National Oceanographic and Atmospheric Administration (NOAA) through its Coastal Ocean Modeling Testbed grant (NA18NOS0120161).

And last but not least, this thesis is dedicated to the memory of Dr. Dave Dempsey, for kindly introducing me to the wondrous world of geophysical fluid dynamics and numerical models. Dave's mentorship gave me the confidence and the courage to embark on this adventure, and I will be forever thankful.

1. Introduction

1.1. Data Assimilation and Observation Thinning

The process of data assimilation (DA) was originally created to deal with the chaotic nature of the atmosphere. Both the atmosphere and the ocean can be described as forced, dissipative dynamical systems with simple physical variables, e.g. temperature, pressure, wind or current velocity, etc., but complicated by the fact that these variables are inter-related and change from one point to another, so very large grids are needed to characterize them (Lorenz, 1995). In general, the two main sources of information needed to study the state and evolution of geophysical fluid dynamics are provided by numerical models and observations. The goal of DA is to combine all available information to calculate an optimal approximation of the state of the atmosphere or ocean (Talagrand, 1997; hereafter T97). In a broad sense, DA consists of interpolating a prior or *background* solution from the model to the observation points, computing the difference between both values -called the *innovation*- and then interpolating back to the model grid points to correct the background values, to obtain a posterior solution, known as the *analysis* (T97). In this context, there are two primary objectives for the assimilation of data: the first is to improve the accuracy of the initial conditions for numerical weather prediction (NWP) and ocean forecasting, and the second, and more widely used in physical oceanography, is to reasonably describe the state of the system during a given period of time, via a so-called *reanalysis* of historical data (T97).

Although the DA problem has a specific formulation (refer to section 2.2), the large dimension of the state variables and the inexact nature of observations and discretized physical equations ensure that the solution will have an associated uncertainty (T97;

Moore et. al., 2019; hereafter M19). Furthermore, many simplifying assumptions need to be made to identify a numerically tractable solution. In particular, the errors in the background state and the observations are generally assumed to be unbiased and described by Gaussian distributions with known covariances (M19; Lahoz et. al., 2010). Currently there are two different approaches to DA: sequential and variational assimilation (T97; Edwards et. al., 2015; M19), which act respectively as filters and smoothers of the data. In sequential methods, such as the Kalman Filter, data are assimilated as and when observations become available, i.e. the analysis is performed at the observation times and the model is integrated between each successive analysis (T97; Edwards et. al., 2015). On the other hand, the goal of variational assimilation is to adjust a model background solution to all observations available during the assimilation window (T97). In 3-dimensional variational (3D-Var) DA, the time window is very short, and the observations are considered to correspond to one single common time much like the Kalman Filter, while in 4-dimensional variational DA (4D-Var) the observations times are respected, making it necessary to dynamically interpolate the ocean state vector in space and time (Figure 1) (Edwards et. al., 2015; M19).

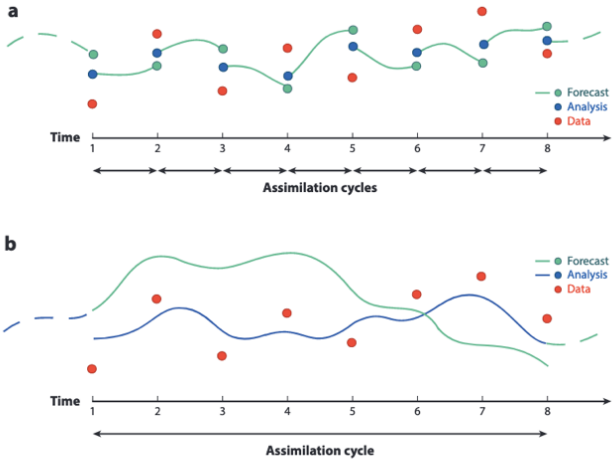


Figure 1 - Representation of sequential and variational data assimilation schemes (from Edwards et. al. 2015). a) Represents a sequential data assimilation approach, such as a Kalman Filter, in

which the model is integrated forward in time until observations are available, and an analysis is then performed. The result in this example is 7 assimilation cycles to span a period of 8 time units. b) Represents 4D-Var, in which for the same 8 time units there is only one assimilation window, and the data throughout the entire time window contributes to the analysis, i.e. the analysis at any time in the window is affected by observations prior and posterior to that time, hence 4D-Var acts as a smoother.

The DA analysis procedure is first briefly presented here to address the motivation for this thesis, although further details are given in section 2. Following the generally accepted DA notation of Ide et. al. (1997), the ocean state-vector is denoted $\boldsymbol{x}$ and it comprises all the grid-point values of the model prognostic variables. The background and analysis are respectively denoted $\boldsymbol{x}_b$ and $\boldsymbol{x}_a$. The goal of DA is to find the best, linear, unbiased estimate (BLUE), which can be expressed as:

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{B}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R})^{-1}\left(\boldsymbol{y}^o - H(\boldsymbol{x}_b)\right) \quad (1)$$

where $\boldsymbol{y}^o$ is the observation vector consisting of all the observations from different platforms, and $\boldsymbol{B}$ and $\boldsymbol{R}$ are respectively the background error and observation error covariance matrices (T97; Moore et. al., 2011; hereafter M11). The non-linear observation operator H transforms the model state vector to the observation locations in space and time, and $\boldsymbol{H}$ is its tangent linearization (M11). The BLUE can then be interpreted as a correction applied to the background, and this correction takes the form of a weighted sum of the innovations $\boldsymbol{d} = \left(\boldsymbol{y}^o - H(\boldsymbol{x}_b)\right)$. The weights are given by $\boldsymbol{K} = \boldsymbol{B}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R})^{-1}$ which is known as the Kalman gain matrix (M11).

The number of observations N in ocean DA, can very easily exceed 10^6, thus the $N \times N$ matrix inversion $(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R})^{-1}$ can be a major practical challenge. While state-of-the-art observation platforms such as earth-orbiting satellites and radar networks yield valuable information about the earth system, the computational cost, bandwidth and storage constraints imposed by large data volumes can present significant challenges (Ochotta et. al., 2005; Ramachandran et. al. 2005). In addition, the large volume of data presents

further challenges for the DA system. High correlations in the background or observation errors can produce different problems. For example, for remote sensing observations closely separated in space and time, $\boldsymbol{HBH}^T$ can have columns very similar to each other, becoming an ill-conditioned matrix. Furthermore, the inversion of $(\boldsymbol{HBH}^T + \boldsymbol{R})^{-1}$ is calculated by solving an equivalent system of linear equations using an iterative conjugate gradient method (M11). Given the heterogeneous nature of the physical observations, preconditioning is a critical step in calculating the BLUE, and it is usually achieved by using the equivalent matrix $(\boldsymbol{R}^{-1}\boldsymbol{HBH}^T + \boldsymbol{I})^{-1}$ (M11). The evaluation of $\boldsymbol{R}^{-1}$ is trivial if $\boldsymbol{R}$ is a diagonal matrix, but this implies an unrealistic assumption of uncorrelated observation errors.

All of the practical challenges outlined above can, to a certain extent, be ameliorated by data thinning which in NWP is considered to be an essential step before assimilating the data. Essentially, there are two classes of thinning algorithms: simple and intelligent. Simple algorithm techniques include random subsampling or combining several observations into one so-called "*superobservation"*. In this case, observations that are closely separated in space and/or time are combined into a single datum, which reduces considerably the dimension of $\boldsymbol{HBH}^T$ in (1) and improves the conditioning of the matrix inversion. Furthermore, the observations can be "*superobed*" or thinned in space to the extent that the resulting observation errors are approximately uncorrelated, which justifies the use of a diagonal $\boldsymbol{R}$ in (1).

In contrast to simple algorithms, intelligent data thinning (IDT) algorithms seek to reduce the number of observations while retaining as much information as possible (Lazarus et. al. 2010). In this study, two different thinning algorithms for ocean surface observations

were explored in a domain encompassing the California Current System (CCS). How the thinning affect the overall performance of DA as implemented in the Regional Ocean Modeling System (ROMS) was then analyzed. One simple "*superobing"* technique was applied to thin radial sea surface velocity measured by a network of coastal high frequency (HF) radars, while an IDT algorithm was implemented to thin SST satellite observations. The IDT technique retains data in regions of high spatial variance, and discards data in regions of low spatial variance, as described in Ramachandran et. al (2005).

1.2. The California Current System

The focus of this study is the California Current System (CCS). The CCS is one of five Eastern Boundary Currents (EBCs) in the world ocean (the others being the Humboldt, Canary, Benguela and Leeuwin currents). Upwelling-favorable winds in all but the Leeuwin Current (off the west coast of Australia) promotes high primary production that sustain large marine ecosystems which in turn support economically important fisheries and that have a significant influence on the global carbon cycle (Carr and Kearns, 2003). In particular, the CCS, the system of currents in the North East Pacific subtropical gyre is an upwelling region of great interest not only because of its economic and environmental impacts, but also because much of what has been learnt here will be applicable to the other EBC regions.

The CCS can be described as a confluence of different currents (Figure 2), some of which vary significantly between the winter and summer seasons. At the surface we find the equatorward California Current (CC), which is broad (~1000 km wide and 500 m deep) and slow (peak speed is ~10 cm/s) (Marchesiello et. al., 2003). The CC is part of the equatorward branch of the subtropical gyre (Talley, 2011) and its eastern side flows above

the California Undercurrent (CUC), which is narrow (between 10 km and 40 km) flowing poleward over the continental slope (Marchesiello et. al., 2003). The nearshore currents are strongly seasonal, in response to the winds that drive them. During Spring/Summer there is a strong equatorward coastal jet, that forms in response to the characteristic upwelling of this region; and during the winter a poleward coastal current (the Davidson Current) develops.

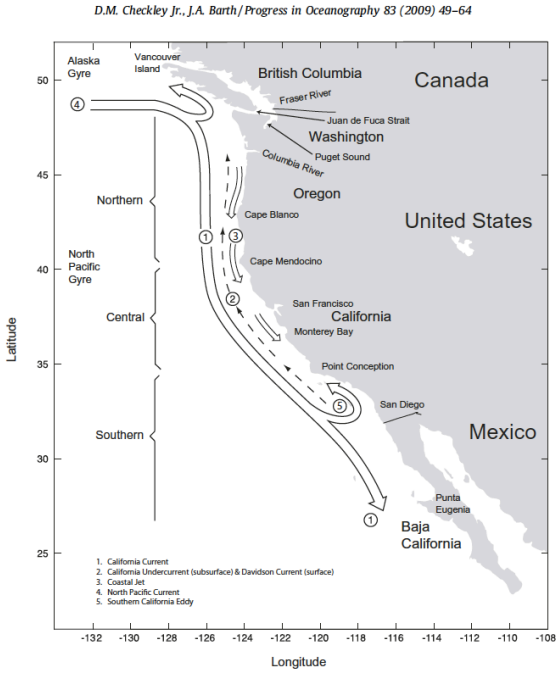


Figure 2: A schematic representation of the currents that form the CCS (from Checkley and Barth, 2009)

Three different water masses coincide in the CCS: Pacific Subarctic Water, North Pacific Central Water and Southern Water (Marchesiello et. al., 2003). The confluence of these different water masses, in addition to the coastal topographic features of the region, add to the complexity of the circulation described above. In particular, the phenomenon of coastal upwelling has been widely observed and studied, and its influence is far reaching, e.g. on climate and primary production. Upwelling of cold water plays a significant role in

climate, interacting with the atmospheric boundary layer via the marine layer and clouds (Dorman et. al., 2013). The atmospheric response to the cold water at the surface is the flow of off-shore thermally-driven winds (Dorman et. al., 2013). The seasonal upwelling of cool, nutrient-rich waters also gives rise to high levels of primary production, making this region seasonally and environmentally variable (Fiechter et. al. 2018).

1.3. Data Sets

The data sets used for this project span the CCS and were downloaded and post-processed by the Ocean Modeling Group at the UCSC Ocean Sciences Department. Our input data comes as Network Common Data Form (NetCDF) files, a widespread format in array-oriented scientific data (https://www.unidata.ucar.edu/software/netcdf/), and in particular the format used for ROMS input and output data streams. Observations of sea surface height (SSH), salinity, ocean temperature and sea surface velocity were assimilated. All of the observations of different variables to be assimilated are merged to create a single observations NetCDF file. Observations were obtained from different sources; for example SSH and salinity are from the Copernicus Marine Service (CMEMS, https://marine.copernicus.eu/), and temperature observations are available from several platforms such as satellites, ocean gliders and free-drifting ARGO profiling floats. Our thinning efforts are focused on the voluminous subsets conformed by gridded SST, obtained from satellite data from the Operational Sea Surface Temperature and Ice Analysis (OSTIA, https://ghrsst-pp.metoffice.gov.uk/ostia-website/index.html) system, and radial sea surface velocity, obtained from a network of HF radars operated by the National Oceanic and Atmospheric Administration's (NOAA) Integrated Ocean Observing System (IOOS, https://ioos.noaa.gov/). Figure 3 shows the typical extent of the HF radar coverage

along the U.S. West Coast and the HF radar antenna locations. Figure 4 shows a close up of the coverage in Central California

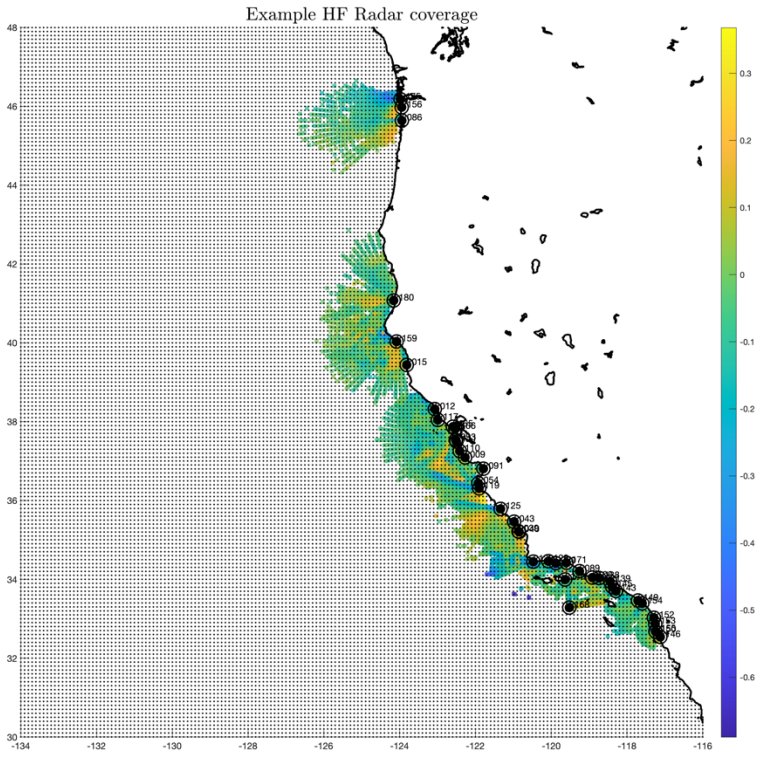


Figure 3 – An example of the typical HF radar coverage along the U.S. West Coast. The location of the individual antenna is indicated by the black circles, while the colors show the measured radial velocities in m s^{-1}.

As part of the post-processing stage, SST data was averaged resulting in one observation per grid point per day, and sea surface velocity was averaged to obtain a daily map of observations. Data in this format are the input for the thinning algorithms described in section 2.3. The assimilation was performed using observations for a period of one year, from 2017-12-31 to 2019-01-02, with an assimilation window of 4 days (i.e. analogous to the assimilation cycle described in figure 1b), resulting in 92 consecutive assimilation cycles. The analysis for the final time in each assimilation cycle is used as the initial condition for the background of the following assimilation cycle. The geographical domain for this study encompasses the U.S. West Coast, extending between latitudes 30º N to 48º

N and longitudes 134º W to 116º W (see Figure 3). This domain is designed to study the CCS. Although some of the assimilated observations correspond to different ocean depths, the data thinning is focused on the surface level.

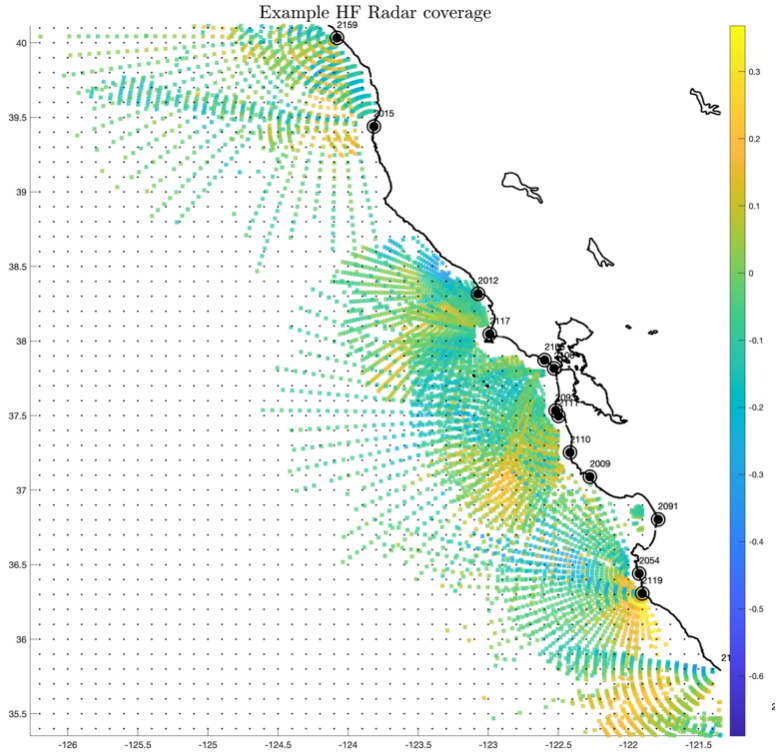


Figure 4 – A close-up example of HF radar coverage in Central California. The location of the individual antenna is indicated by the black circles, while the colors show the measured radial velocities in m s^{-1}.

The thesis is organized in the following way: Section 2 describes the model (ROMS), further details about the 4D-Var method used, and the implemented thinning algorithms. A set of preliminary single cycle experiments, and subsequent multiple cycles experiments are described in Section 3. Section 4 explores different ways to analyze and compare results of the different experiments, specifically calculating innovation statistics, performing a priori and a posteriori error consistency checks, analyzing the topology of the total error covariance matrix, and calculating the impact of each type of observation in the analysis.

Finally, section 5 summarizes the conclusions and outlines possible future avenues of work to continue researching the topics presented here.

2. Methods

2.1. The Regional Ocean Modeling System (ROMS)

2.1.1 Model Description

ROMS is a 3-dimensional, free-surface, terrain-following coordinate numerical model that solves the Navier-Stokes equations using the hydrostatic and Boussinesq approximations (Haidvogel et. al., 2008). The dynamical primitive equations are time-discretized using the third-order accurate linear multistep method leapfrog and Adams-Moulton corrector (www.myroms.com). The spatial discretization consists of a staggered grid that results from the vertical topography following coordinate z=z(x,y,σ), and the Arakawa C-grid in the horizontal (Figure 5) (Haidvogel et. al., 2008). The terrain-following σ-coordinate represents the vertical distance from the surface as a fraction of the local water column thickness, and it can be combined with nonlinear stretching (Shchepetkin and Williams, 2004).

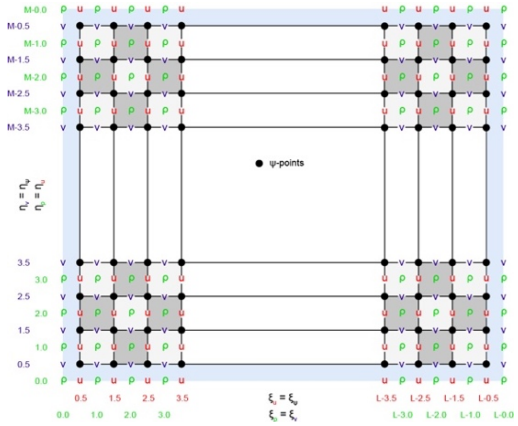


Figure 5 – The ROMS staggered grid based on Arakawa C-grid (https://www.myroms.org/wiki/File:4dvar_staggered_grid_rho_cells.png)

Different modules constitute the dynamical kernel of ROMS. The nonlinear model (NLROMS) numerical design and algorithm are thoroughly described in Shchepetkin and Williams (2004). Of special interest for the DA application are the tangent linear model (TLROMS) and the adjoint of the tangent linear model (ADROMS). TLROMS represents a first-order Taylor expansion of NLROMS and yields the Jacobian of the dynamical operators of the nonlinear model (Moore et. al., 2004). ADROMS is the adjoint of the Jacobian operator and provides information about the sensitivity of the model solution to variations in the model variables, boundary conditions and parameters (Moore et. al., 2004). TLROMS and ADROMS are critical components of the 4D-Var data assimilation system. The prognostic variables in the ROMS primitive equations are potential temperature (T), salinity (S), horizontal velocity (u,v), and sea surface displacement (ζ) (M11). Since the model is hydrostatic, vertical velocity is computed from the divergence of the horizontal velocity. As noted in section 1.1, the state vector $\boldsymbol{x}(t_i) = (T, S, z, u, v)^T$ consists of all the values of the prognostic variables at every ocean grid point at time t_i and it is propagated forward in time by NLROMS. The resulting circulation depends also on the surface forcing and lateral open boundary conditions, denoted respectively as $\boldsymbol{f}(t_i)$ and $\boldsymbol{b}(t_i)$), and, following M11, can be represented as: $\boldsymbol{x}(t_i) = M(t_i, t_{i+1})(\boldsymbol{x}(t_{i-1}), \boldsymbol{f}(t_i), \boldsymbol{b}(t_i))$.

2.1.2 Model Configuration

The ROMS model configuration used for this project consists of a 184x179x42 grid with a mean resolution dx~8.5km and dy~15km and 42 vertical levels and extends between latitudes 30º N to 48º N and longitudes 134º W to 116º W (Figure 6). Although some of the assimilated observations correspond to different ocean depths, the data thinning discussed in section 1.1 is focused on the surface level. The surface forcing boundary conditions were

computed from atmospheric fields from the Monterey Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS, https://www.nrlmry.navy.mil/coamps-web/web/view) and boundary conditions at the ROMS open boundaries from the Hybrid Coordinate Ocean Model (HYCOM, https://www.hycom.org/) global DA product.

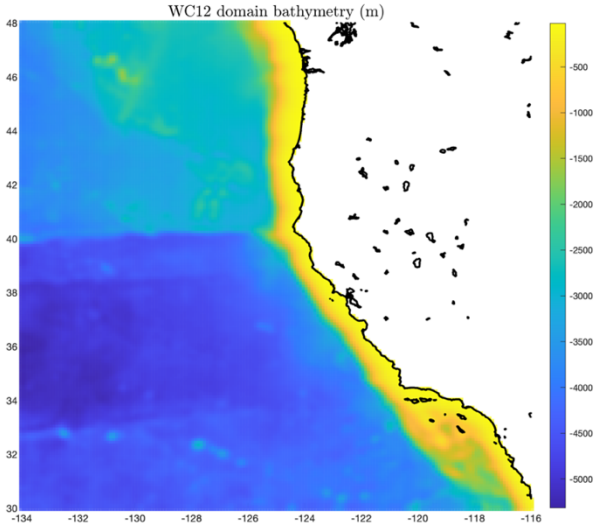


Figure 6 – The ROMS bathymetry and extent of the model domain for the CCS.

2.2. 4-dimensional Variational Data Assimilation (4D-Var)

Variational DA can be interpreted as an application of Bayes' theorem, although representations in control theory and other fields lead to the same general results (T97; Edwards et. al., 2015; M19). According to Bayes' theorem, the posterior conditional probability for the state vector x, given the observations y is given by:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{2}$$

where $p(x)$ denotes the prior distribution, $p(y|x)$ is the data distribution, and $p(y)$ is a normalizing constant (Edwards et. al., 2015). In the case of Gaussian distributions for the

background errors and observation errors, then $p(x|y) \propto \exp(-J_{NL})$ where, following the notation introduced in section 1.1:

$$J_{NL} = (\mathrm{x} - \mathrm{x_b})^{\mathrm{T}}\mathbf{B}^{-1}(\mathrm{x} - \mathrm{x_b}) + [\mathrm{y} - \mathrm{H(x)}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathrm{y} - \mathrm{H(x)}]\,. \tag{3}$$

The scalar J_{NL} is usually referred to as the cost function and the goal of DA is to identify the vector x that maximizes $p(x|y)$, and hence minimizes J_{NL} (Edwards et. al., 2015).

M11 have developed and described three variants of 4D-Var algorithm that are supported by ROMS. In this study a dual formulation of 4D-Var that evolved from the physical-space statistical analysis system (4D-PSAS) (see the flow chart in Appendix B) has been used. The goal of 4D-Var is to find the BLUE over a finite time interval. Specifically, the observation operator H in (1) includes NLROMS, while the linearized operator $\boldsymbol{H}$ and the transpose $\boldsymbol{H}^T$ rely on TLROMS ad ADROMS respectively. In this way, information is dynamically interpolated in space and time via the dynamics that control the ocean circulation (M19).

ROMS 4D-Var uses an iterative Gauss-Newton method based on an incremental approach to 4D-Var, and consists of calculating the increments $\delta\mathrm{x}_n = (\mathrm{x}_n - \mathrm{x}_{n-1})$ that minimize a sequence (known as inner-loops) of linearized approximations of J_{NL}, that can be expressed as:

$$J_n(\delta\mathrm{x}_n) = \delta\mathrm{x}_n{}^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathrm{x}_n + (\mathbf{d_n} - \mathbf{H_n}\delta\mathrm{x}_n)^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{d_n} - \mathbf{H_n}\delta\mathrm{x}_n) \tag{4}$$

where $\mathbf{d_n}$ is the innovation vector introduced in section 1.1, and the initial guess, x_0, is the background state-vector (M11) so that:

$$\mathbf{d_n} = \big(\mathbf{y} - \mathrm{H}(\mathrm{x}_{n-1})\big). \tag{5}$$

The solution x is updated in during so-called outer-loops according to $\mathrm{x}_n = \mathrm{x}_{n-1} + \delta\mathrm{x}_n$ (M11). In the 4D-Var experiments presented here 7 inner-loops and 2 outer-loops were used based on the operational real-time analysis-forecast that is run at UCSC in support of IOOS using the same ROMS configuration.

At the end of the final outer-loop, the solution (posterior/analysis state vector) is then given by:

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{\delta x}_a \tag{6}$$

The solution of the 4D-Var problem consists of estimating the increment $\boldsymbol{\delta x}_a$, which in its dual form can be expressed as:

$$\boldsymbol{\delta x}_a = \boldsymbol{BH}^T(\boldsymbol{HBH}^T + \boldsymbol{R})^{-1}\big(\mathbf{y} - \mathrm{H}(\boldsymbol{x}_b)\big) \tag{7}$$

which leads to (1).

As discussed in section 1.1, inverting the matrix $(\boldsymbol{HBH}^T + \boldsymbol{R})$ represents one of the major computational challenges of 4D-Var. $(\boldsymbol{HBH}^T + \boldsymbol{R})$ is also referred to as the stabilized representer matrix and it represents the total error covariance in the observations space (Moore et. al., 2021; hereafter M21). According to (3) (or (4)) 4D-Var can be interpreted as a least squares problem that seeks to minimize the difference between the model and the observations, given certain background and prior errors (M11). The dual form approach refers to the fact that the minimization is solved in the observation space, which in general has a dimension several orders of magnitude smaller than the state space. This means that instead of estimating the state space increment $\boldsymbol{\delta x}_a$, the counterpart in observation space $\boldsymbol{w}_a$ is identified, solving the system:

$$\boldsymbol{w}_a = (\boldsymbol{HBH}^T + \boldsymbol{R})^{-1}\mathbf{d} \tag{8}$$

As noted in section 1.1, the inverse in equation (8) is evaluated using a conjugate gradient algorithm which yields a reduced-rank factorization of the stabilized representer matrix.

Specifically, the Lanczos formulation of the conjugate gradient algorithm is applied to the rescaled stabilized representer matrix $\boldsymbol{P}$ (M11) so that:

$$\boldsymbol{P} = \boldsymbol{R}(\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^{T} + \boldsymbol{I}) \tag{9}$$

$$\boldsymbol{P}^{-1} \approx \mathbf{V}_{\boldsymbol{m}}\mathbf{T}_{\boldsymbol{m}}^{-1}\mathbf{V}_{\boldsymbol{m}}^{\boldsymbol{T}}(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^{T}\boldsymbol{R}^{-1}). \tag{10}$$

To account for the fact that the state vector represents a variety of variables, with different physical units, the matrix $\boldsymbol{P}$ is preconditioned with $\boldsymbol{R}^{-1}$, which yields a non-dimensional matrix denoted here as $\widetilde{\boldsymbol{P}}$, such that:

$$\widetilde{\boldsymbol{P}_{\boldsymbol{m}}} \approx \mathbf{V}_{\boldsymbol{m}}\boldsymbol{T}_{\boldsymbol{m}}\mathbf{V}_{\boldsymbol{m}}^{\boldsymbol{T}}(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^{T}) \tag{11}$$

where $\mathbf{V}_{\boldsymbol{m}}$ is the matrix of Lanczos vectors, with one vector calculated in each inner-loop as shown in the flow chart in Appendix B (M11, M21). In the experiments presented here, m=7. The matrix $\boldsymbol{T}_{\boldsymbol{m}}$ is symmetric, positive definite and tridiagonal, and it has the same eigen spectrum, and proportional eigenvectors (also referred to as Empirical Orthogonal Functions, EOFs, in this context) to $\widetilde{\boldsymbol{P}}$, and its computation is trivial in any numerical computing environment, since $\boldsymbol{T}_{\boldsymbol{m}}$ is in this case an mxm matrix (M21). Equation 10 shows that $\boldsymbol{P}^{-1}$ can be interpreted as a gain matrix applied to the innovation vector to calculate the BLUE of the circulation in observation space, so its properties are of considerable interest, and are explored in Section 4.3.

Currently the UCSC Ocean Modeling and Data Assimilation group runs a daily near real-time 4D-Var analysis of the CCS using the previous 4 days to produce an estimate of the physical state of the ocean (https://oceanmodeling.ucsc.edu/ccsnrt/). Figure 7 shows the typical output of the model for horizontal velocity, sea surface height, temperature and salinity for 2021-05-14.

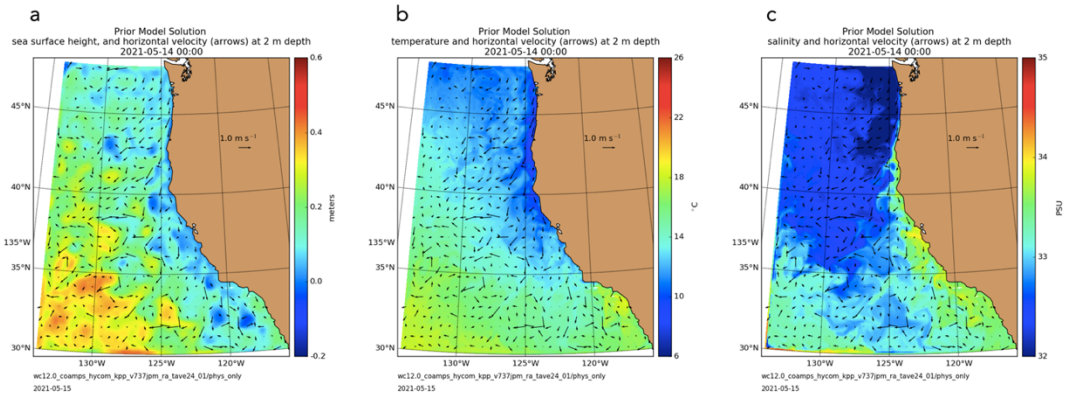


Figure 7 – Typical output of the near real-time 4D-Var assimilation performed daily by the UCSC Ocean Modeling team https://oceanmodeling.ucsc.edu/ccsnrt/

2.3. Thinning algorithms

2.3.1. Radar Superobs

Since radar is the instrument that provides the best observations to study storms, the problem of assimilating radar data has been historically studied in the context of NWP (Fabry and Meunier, 2020). In general, there are two types of operationally available radar data for NWP: reflectivity of hydrometeors and doppler radar winds (Fabry and Meunier, 2020). Although both types of observations are difficult to assimilate in NWP, reflectivity is particularly challenging because its relation to the model state variables is complicated, and also due to the convective-scale, rapid and unstable nature of storms (Fabry and Meunier, 2020). On the other hand, doppler velocity is more widely used by many operational NWP centers to improve weather forecasts (Fabry and Meunier, 2020). In addition to the NWP foundations of radar DA, the installation of HF radar systems to monitor the coastal oceans around the world has generated multiple studies on the assimilation of surface currents in ocean models (Shulman & Paduan, 2008). Sea surface velocities measured by a network of 48 HF radars along the U.S. West Coast produce a high-density distribution of observations (Figures 3 and 4). Given the radial nature of

radar observations, then close to the radar stations the density of observations is high, resulting in a large number of observations per model grid cell (Bick et. al. 2016; hereafter B16). Because on this, and in order to attenuate the effects of observation correlation, *superobing* discussed in section 1.1 is a common strategy that is used to thin radar data (Gustafsson et. al., 2018).

In order to create an even distribution of observations in the surface velocity dataset described in section 1.2, and to minimize some of the potential issues discussed in section 1.1, the superobing procedure described by B16 and Waller et. al. (2019) for NWP applications was adopted in the experiments described here. Superobing consists of averaging several radial bins to a single so-called *superobservation*, and is generally applied to each radar site individually (B16). The algorithm was implemented to allow changes in the resolution of the resulting superobservations by defining a set of parameters. Figure 9 shows a flow chart of the steps applied to each HF radar site. With reference to Figure 9, to increase the resolution the parameter *finer_grid_switch* can be turned on, in which case superobservations are also assigned to intermediate points in between model grid points. Resolution can also be decreased by increasing the parameter *grid_step*, which allows subsampling of the model grid while assigning superobservations. For each model grid point a pie-wedge area is defined (Figure 8) with an angle and range that depends on the distance from a point to the radar antenna location (B16). First, the radar bin closest to that grid point is located, which will be the location of the superobservation (B16). The value of the superobservation is calculated as the average of all the observations located inside the pie-wedge area (B16). The size of the area surrounding the grid point at a distance r_0 from the radar site is given by the range and the azimuth angle, calculated as:

- Width of range interval: $L_x\sqrt{2}$

- Azimuths at range r_0 : $\varphi = 2\arctan\{(L_x\sqrt{2}/2)/r_0\}$

where L_x is the grid is a tunable parameter that sets the spatial scale of averaging (B16). The default resolution is defined by the model configuration grid (corresponding to *Lx*=0.1 degrees). The parameter *min_number_obs* defines the minimum number of observations required to create a superobservation. Figure 10 shows a flow chart describing the algorithm applied to every grid point in the sub-grid defined for each HF radar site.

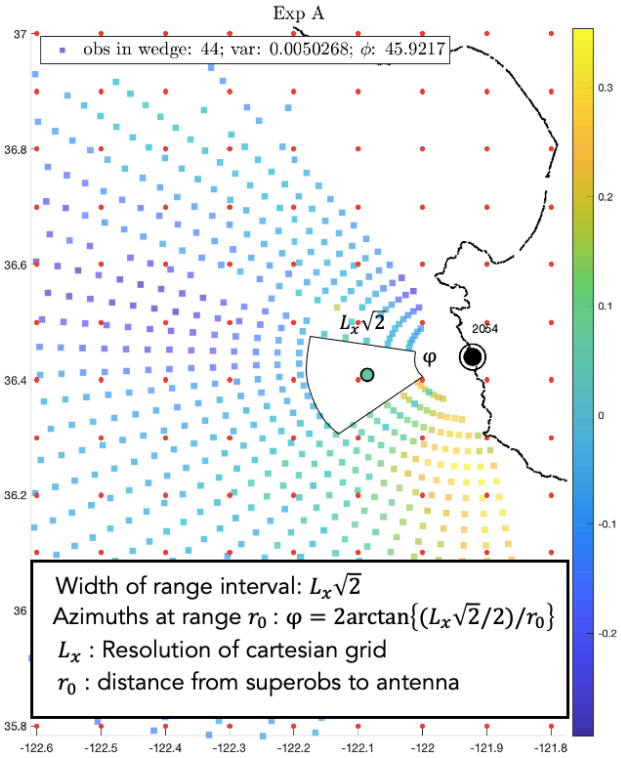


Figure 8 – Example of wedge-pie area used to calculate one superobservation in the vicinity of a radar site south of Monterey Bay.

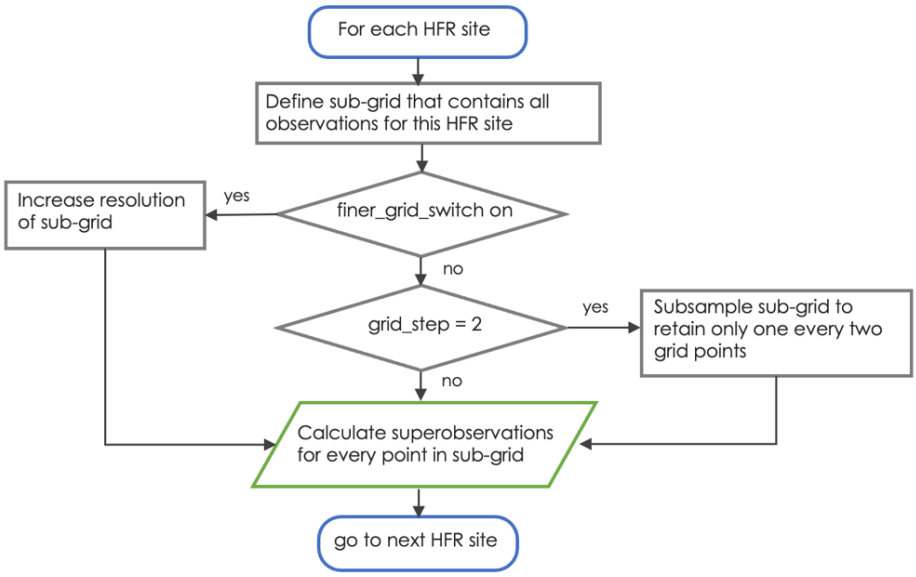


Figure 9 – Flow chart of the logic applied to every HF radar site in the superobing algorithm.

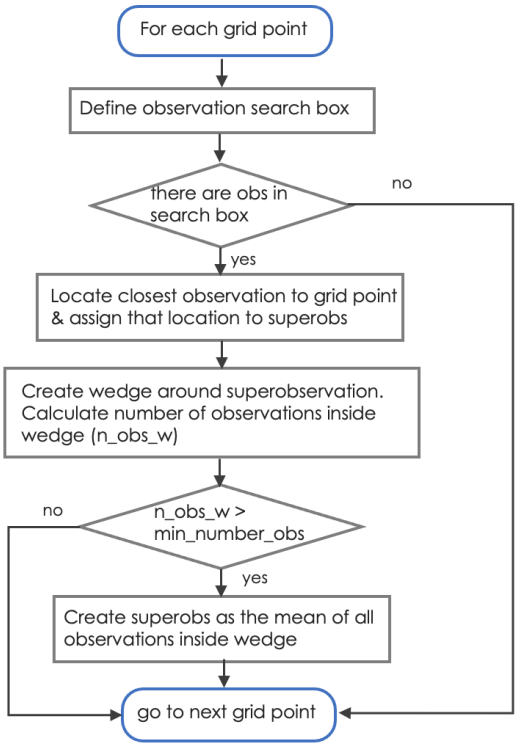


Figure 10 – Flow chart applied at each grid point in the vicinity of an HF radar site to calculate a superobservation.

2.3.2. An IDT algorithm for SST

For the SST dataset an IDT algorithm based on Ramachandran et. al. (2005; hereafter R05) was implemented to retain data in regions with a high content of information (*i.e.* regions where the spatial variance is high, such as in the vicinity of temperature fronts) while eliminating redundant observations by subsampling regions with low variance. The data thinning is based in a quadtree decomposition of the model domain, in which data are recursively divided into four quadrants (R05). In a preprocessing step, data is normalized to the range (0,1), and a global mean for the dataset is calculated (R05). This global mean multiplied by a cutoff threshold results in an acceptable standard deviation that yields the cutoff variance within each quadrant (R05). For each quadrant an F-test is performed to statistically compare the variance of the quadrant (varQ) with the acceptable cutoff variance (varT) (R05). If the quadrant has a variance greater than the cutoff, the algorithm divides the region into 4 quadrants again, otherwise the recursive decomposition is terminated and the central data point of the quadrant is used as a thinned representative value for the whole quadrant (R05). The pseudocode for this algorithm as presented in R05 is in Appendix A, and Figure 11 shows a flow chart of the logic applied to each quadrant under analysis.

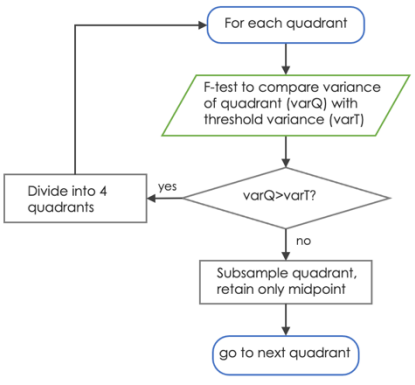


Figure 11 – Flow chart applied to each quadrant to detect regions of high spatial variance.

An example of how the quadrants are divided and the F-test calculated for a simulated velocity field is shown in Figure 12. Figures 12 a and 12 b respectively show the full and the thinned data set before and after applying the algorithm. The domain was divided into 4 quadrants and the F-test was applied to the first quadrant (in this case lower left, Figure 12 d). Since the variance of the quadrants was higher than the cutoff variance, this quadrant was divided into quadrants again (Figure 12 e). The F-test was applied again to the first of the new set of quadrants, resulting in the decision to divide into quadrants again. Finally in Figure 12 f, the variance of the quadrant is lower than the threshold variance, so the middle-point is retained and all the other values are discarded. Then the same logic was applied to the remaining quadrants.

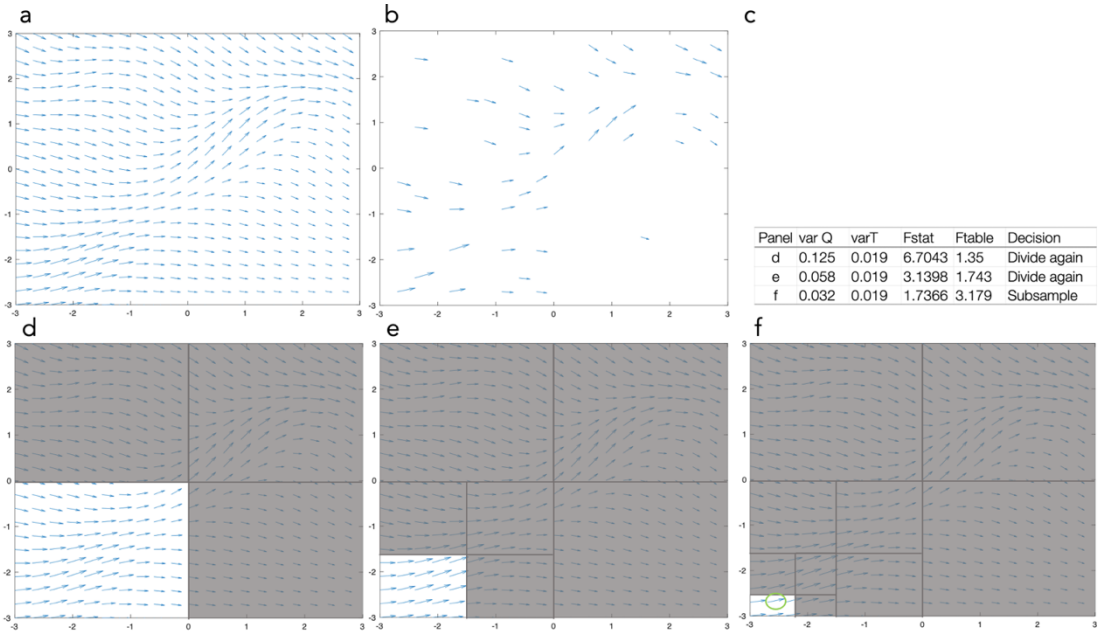


Figure 12 – An example on how the IDT algorithm recursively divides the domains into quadrants until the spatial variance is low enough to subsample the quadrant and move on to the next one. The 2-dimensional velocity field was created as a random linear combination of 2D Gaussian distributions.

The IDT algorithm was initially identified as a potential candidate for thinning gridded surface velocity estimates based on the HF radial measures. It is for this reason, the IDT

algorithm was first tested for u and v velocity fields in the form of random linear combinations of 2-dimensional Gaussian distributions illustrated in Figure 12, and then applied to gridded u and v fields from ROMS. Figure 13 shows the velocity fields before (Fig. 13 a) and after (Fig 13 b) the thinning. Figure 13 b shows that after the thinning the remaining data still captures eddies and other circulation features that are characterized by high spatial variance while discarding information in more uniform areas.[1]

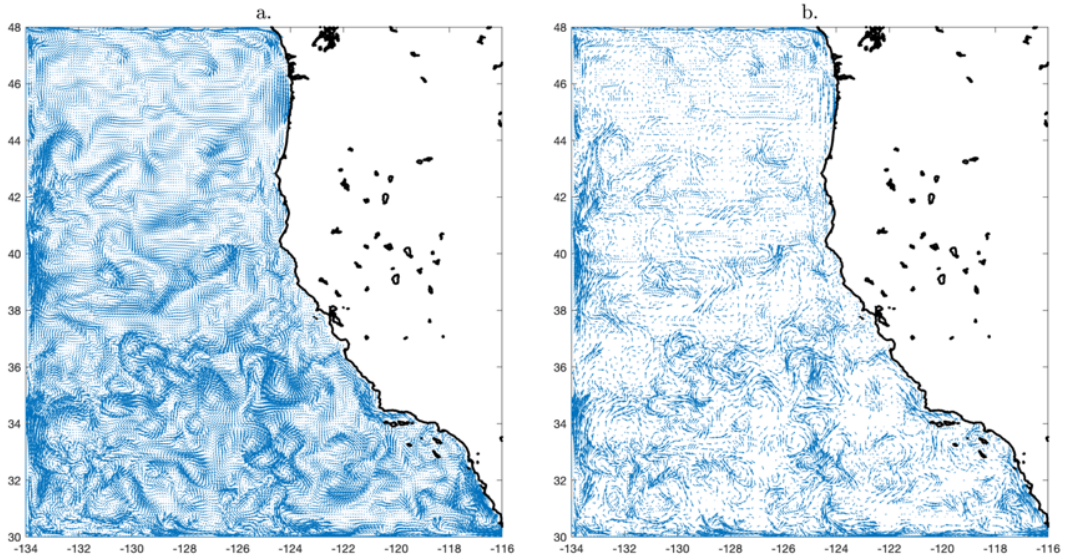


Figure 13- IDT thinning algorithm applied to a ROMS velocity field.

For the SST data different cutoff thresholds were tested to thin a dataset representative of a single day, and it was found that as the threshold increases (hence varT increases), the percentage of the total observations retained by the thinning algorithm decays

[1] Since HF radar radial observations would require a transformation to u and v coordinates in order to apply the IDT algorithm, and because the preliminary experiments based on the superobservation approach showed promising results, we decided to use the superobing approach for radar observations, and to extend the scope of this thesis to evaluate thinning of SST, a 2-dimensional field that was already part of the datasets used for this study.

exponentially (Figure 14). Increasing the threshold corresponds to the situation where data are retained based on a larger upper-limit for the variance.

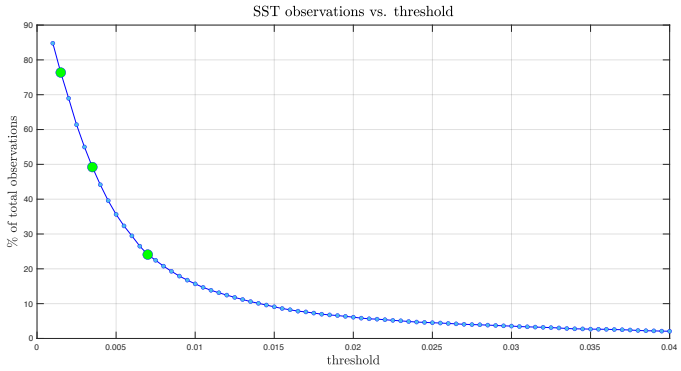


Figure 14: The percentage of the observations retained after thinning decreases exponentially as the threshold for the variance increases. The thresholds corresponding to 75%, 50% and 25% of total observations are highlighted in green, single cycle experiments were performed for these thresholds.

3. Experiments

3.1. Single cycle experiments for radar observations

As a preliminary step before running experiments for the entire dataset spanning a full year, a series of experiments for the single DA cycle for the 4-day period 2017-12-31 to 2018-01-03 were first performed. This allowed for algorithm testing and debugging and for exploring their fundamental properties before applying them to a longer time interval. The *control* experiment (sc.1) consists of assimilating all of the available data for the cycle, i.e. no thinning or superobing takes place in this case. A first set of experiments for radar observations (experiment sc.2 to sc.4) was based on a simple thinning, uniformly subsampling the radar observations. The following set of experiments (sc.A to sc.D) consisted of creating superobservations for each radar site at each observation time during the 4D-Var DA cycle, using different criteria for the grid resolution and the minimum number of observations used to compute a superobservation. Table 1 (Appendix C)

summarizes the criteria employed in each experiment, and Table 2 (Appendix C) shows the parameters described in section 2.3.1 that were defined for experiments sc.A to sc.D. Figure 15 shows examples of the different strategies used for computing superobservations using observation from a single radar site south of Monterey Bay. Experiments sc.A and sc.B (Figure 15 a) retain one superobservation per grid point, while in experiment sc.B a superobservation is retained only if three or more observations were used to calculate it. The latter approach filters out data in regions where the observation error is known to be larger (*i.e.* far away from the radar). Figure 15 b shows the result of superobing with lower (experiment sc.C) or higher resolutions (experiment sc.D).

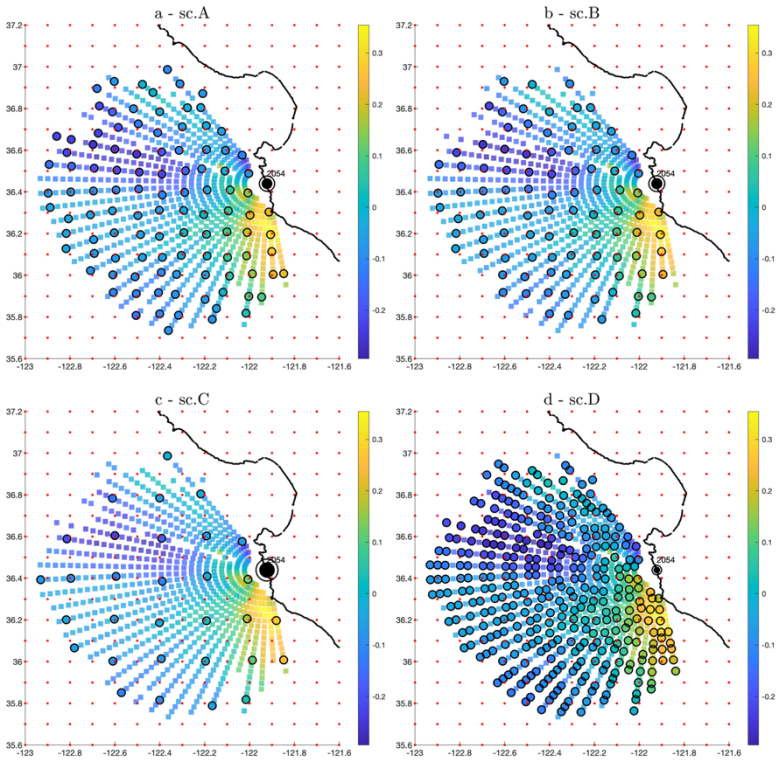


Figure 15. Visual representation of superobing to the model grid resolution for a single HF radar site located south of Monterey. Squares represent the full data set for this radar site at this survey time, and circles represent the resulting superobservations.

A challenge in operational NWP and ocean forecasting is that the true state of the system is never known, so it can be difficult to quantify the relative performance of different approaches. We will return to thus issue in section 4. For now, one way to evaluate the impact of the different superobservation strategies on the 4D-Var circulation estimate is to calculate the root-mean-square (rms) difference between the model interpolated to the observation location and the corresponding observation (Edwards et. a., 2015). A similar approach was followed here, although the standard deviation of the innovation vector was calculated instead of the rms. For a large number of observations p and a mean that tends to zero the standard deviation yields a similar measure of the error, and it can be compared with the square root of the expected error, which also represents an error in the form of a standard deviation. The differences between model and observations were calculated before and after the assimilation. These differences are respectively referred to as the innovation (Eq. 11) and residual (Eq. 12), and their standard deviations were calculated as shown according to equations (13) and (14).

$$\boldsymbol{d}_b = \boldsymbol{y} - H(\boldsymbol{x}_b) \tag{12}$$

$$\boldsymbol{d}_a = \boldsymbol{y} - H(\boldsymbol{x}_b + \boldsymbol{\delta x}_a) \tag{13}$$

$$\sigma_{innovation} = \sqrt{\frac{1}{p-1}\sum_{i=0}^{p-1}(d_b - \mu_b)_i^2} \tag{14}$$

$$\sigma_{residual} = \sqrt{\frac{1}{p-1}\sum_{i=0}^{p-1}(d_a - \mu_a)_i^2} \tag{15}$$

where μ_b and μ_a respectively represent the mean of the innovation and the residual vectors. p in this case represents the number of HF radar observations, and the innovation and residuals are calculated only for these observations. Figure 16 shows the results for

the single cycle experiments. Although sc.1 shows the lowest error after assimilation (in light blue), all the experiments show an improvement (i.e. lower error) after assimilation, even when keeping only 3.5% of the observations (experiment sc.C).

For reference, Figure 16 also shows (in green) the expected error in the innovation computed from the diagonal of the stabilized representer matrix ($\boldsymbol{HBH}^T + \boldsymbol{R}$) which, as shown in section 4.1, is the expected error covariance for the innovation vector $\boldsymbol{d}$. Therefore, the green bars in Figure 16 can be thought of as the standard deviation of the expected probability density function of the innovations. In all cases, the single realization of the innovation vector lies further than one standard deviation from the mean. For case sc.1 the standard deviation of the residual is lower than the expected error, and the same is true for sc.D, although the difference is smaller. For the subsampling experiments sc.2-4 the residual error increases as the number of retained observation decreases, and for the superobing experiments sc.A-D the residual error and the expected errors are very similar. This preliminary error analysis indicates that, overall, the thinning algorithm applied to the HF radial observations is well-behaved and performing correctly.

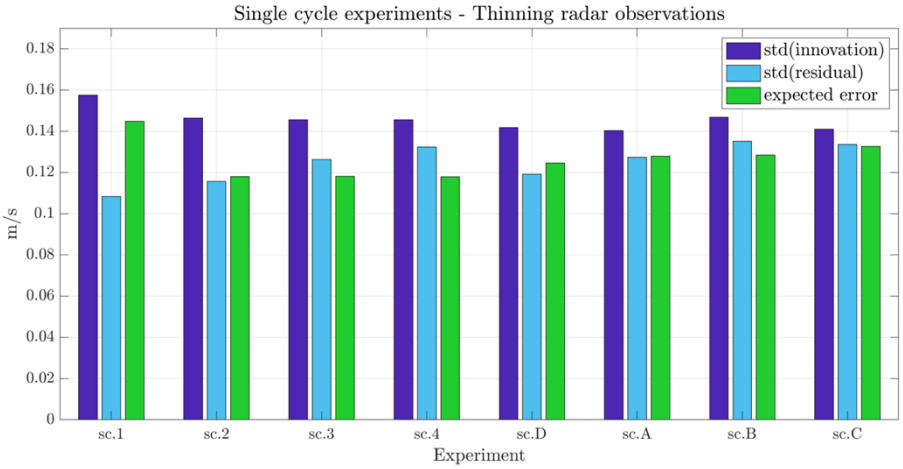


Figure 16 – Comparison between the expected error (green) and the estimated errors before (purple) and after assimilation (light blue) for the single cycle experiments thinning HFR observations.

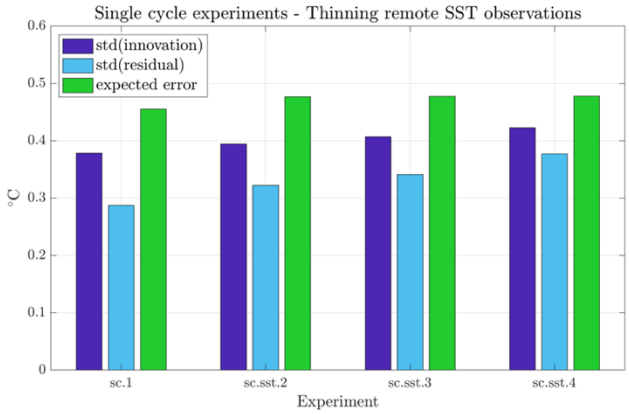


Figure 17 – Expected and estimated errors before and after assimilation for single cycle experiments thinning SST.

3.2 Single cycle experiments for SST observations.

In order to emulate the idea of thinning the dataset to different resolutions, we tested several different thresholds for the SST dataset, and after visually inspecting the results (Figure 18) three different thresholds were chosen (highlighted in green in Figure 14) as candidates DA experiments. Table 3 (Appendix C) summarizes the single cycle (2017-12-31 to 2018-01-03) experiments for SST thinning.

Expected errors for the SST thinning experiments were calculated as explained in section 3.1, using equations (11)-(14), but in this case p refers to the remote sensed SST observations. In this case, the experiment using all the observations (sc.1, Figure 17) again has the lowest error after assimilation, with the error gradually and slightly increasing as more observations are discarded (sc.sst.2-4, Figure 17). The expected error is in all cases higher than the standard deviations of $\boldsymbol{d}_b$ and $\boldsymbol{d}_a$, and again the thinning strategy is well-behaved.

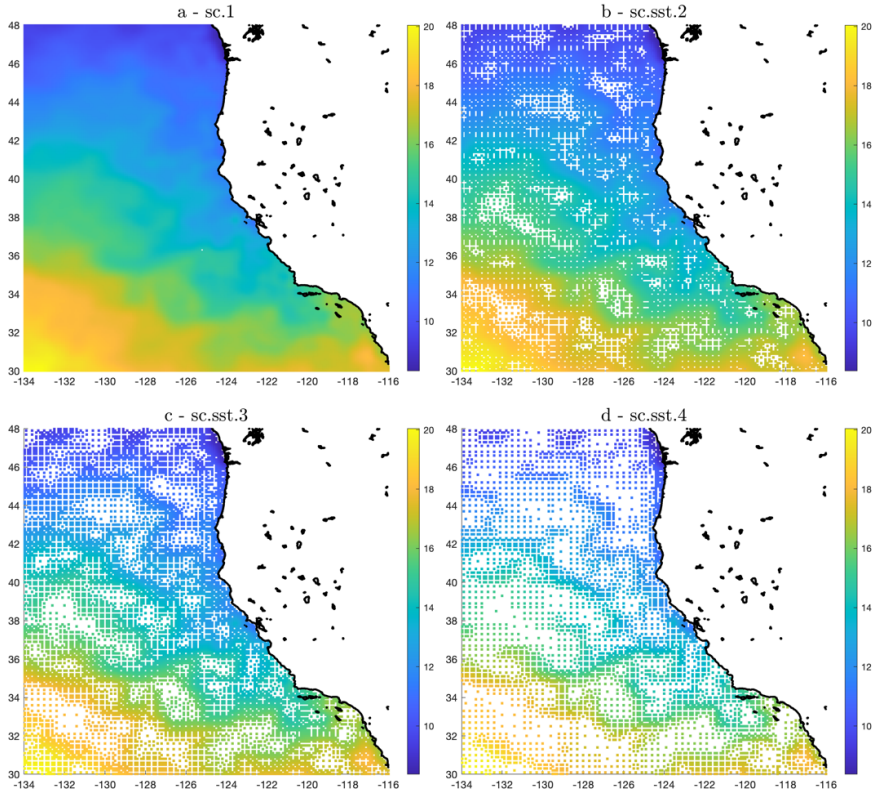


Figure 18 - Representation of the SST field for a single survey time. Panel a shows the temperature field using all the available observations. Panels b, c and d show the same field after keeping 75%, 50% and 25% of the total observations.

3.3 Multiple cycle experiments

Based on the preliminary analysis of the single cycle experiments of sections 3.1 and 3.2, a set of three DA experiments (Table 3, Appendix C) was designed using the dataset for the full year described in section 1.2. A MATLAB script to create the observation files for all the experiments was created in a way to easily create files with all the available observations, or thinned observations for either one of both variables under analysis (i.e. radial velocities and/or SST). Also, any of the experiments tested for the single cycle can be set up for the multiple cycles. Given that after the preliminary analysis, all of the single cycle experiments seemed feasible to be extended to more cycles, the criteria used to define

the multiple cycle experiments was to choose a thinning level that would allow an increase or decrease in the amount of discarded observations in future experiments. Exp01 was defined as a control test in which all the available observations were assimilated without thinning of any kind. For Exp02 and Exp03 only one of the observed variables was thinned (*i.e.* either surface radial velocities or SST) using the whole set for the other one. For Exp02 radial velocities were thinned using the same parameters described for sc.B, in principle to test the analysis while having an observation resolution similar to the model grid (i.e. keeping one observation per grid point). On the other hand, Exp03 represents an extension of experiment sc.sst.2, which keeps approximately 50% of the SST observations. The results of the multiple cycles experiments are discussed in the next section.

4. Analysis

In a strict sense, errors should be calculated as the difference between the analysis and the true state of the system. Since the true state is not known, a series of statistical analyses was performed (following the practice in NWP) to assess the reliability of the assimilation for each observing system. The covariance matrices $\boldsymbol{B}$ and $\boldsymbol{R}$ represent *a priori* estimates of the uncertainties associated with the background and the observations, and as shown in sections 1.1 and 2.2 they are essential components of the analysis state-vector (Eq. 6). The accuracy of the solution depends on the statistical distributions for the background errors and observation errors, which are assumed to be normal and unbiased, but in reality they are not well known and very difficult to estimate (Desroziers et.al., 2005; Lahoz et.al., 2010). Nonetheless, it is important to test the efficacy of the underlying assumptions of normal, unbiased errors in order to place the experimental results in their proper context.

4.1. Innovation statistics

One method to evaluate each experiment is to check if the error statistics assumed in the assimilation scheme are correctly specified based on the expected error statistics of the innovation vector, which can be rewritten as:

$$\mathbf{d} = \left(\mathbf{y} - \mathrm{H(x)}\right) \cong \epsilon_0 - \mathbf{H}\,\epsilon_b \tag{16}$$

ϵ_0 and ϵ_b are respectively the vectors of observation and background errors. Assuming ϵ_0 and ϵ_b are uncorrelated, the statistical expectation of the vector of innovations should then satisfy the following relation (Desroziers et. al., 2005; hereafter D05):

$$\mathrm{E}[\mathbf{d}^{\mathrm{T}}\mathbf{d}] \cong \mathrm{E}[{\epsilon_0}^{\mathrm{T}}\epsilon_0] + \mathbf{H}\,\mathrm{E}[{\epsilon_b}^{\mathrm{T}}\epsilon_b]\mathbf{H}^{\mathrm{T}} \quad . \tag{17}$$

Equation (17) thus provides a global check on the specification of **B** and **R** (D05) since:

$$\mathrm{E}[\mathbf{d}^{\mathrm{T}}\mathbf{d}] = \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} \tag{18}$$

Figure 19 shows a time series of the square root of the mean squared innovations (blue) and the expected error represented by $\sqrt{Tr\{\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\}/p}$ (cyan), where p is the number of observations of the variable under analysis. For radial velocities (Figure 19 a-c), Exp01 shows good agreement between the a priori expected and a posteriori calculated errors, especially between May and September. During the winter there is a sustained inconsistency where the expected total error seems to have been overestimated. Between mid-September and mid-October, and also in mid-November there are two localized periods of inconsistency. The thinning of radar observations (Exp02, Fig 19 b) exhibits inconsistencies during the summer when coastal upwelling is occurring. This suggests that perhaps different thinning strategies may be needed during different seasons to account for differences in the circulation patterns. Nonetheless, Exp03 (Fig 19 c, all radar observations) in which the expected error was adjusted based on Exp01 (as described in section 4.2) before running 4D-Var shows a better consistency throughout the year. For the remotely sensed SST, the most problematic period is the upwelling season of spring

and summer, where there are large inconsistencies for all the experiments (Figs 19 d-f), and in particular for the case with thinned SST observations (Exp03, Fig 19 f).

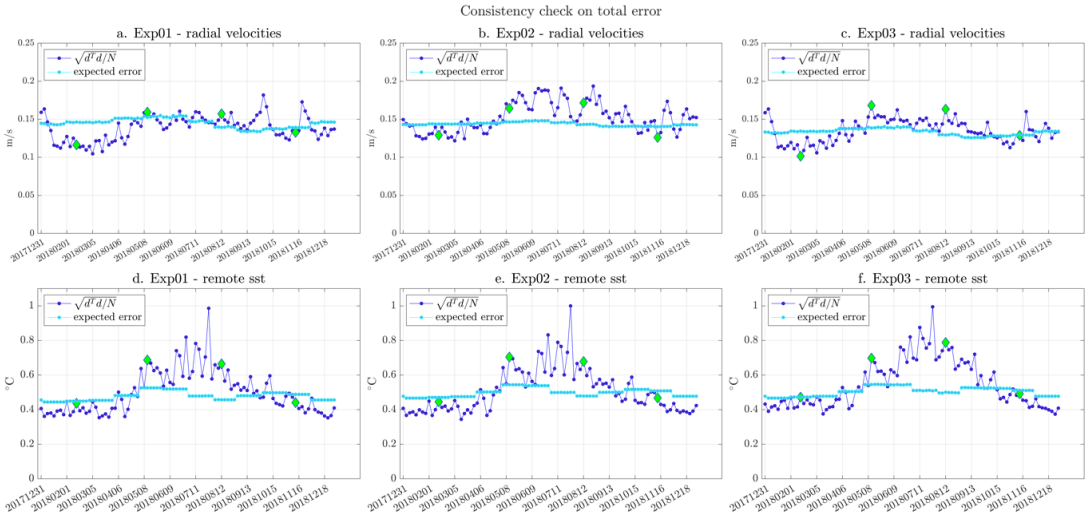


Figure 19 – Consistency check between expected errors (cyan time series) and a statistical a posteriori estimation of the actual errors assumed for each experiment (blue time series). The green diamonds highlight the cycles corresponding to the SST and surface velocity fields shown in Figures 20-22

Figure 20 and 21 show seasonal snapshots of the 4D-Var analyses of SST and surface velocity, where upwelling is noticeable in panels c) and d). Figure 20 shows that, as expected, the coldest temperatures (~10º C) are observed during the winter in the northern part of the domain (Fig. 20 a), while in the spring (Fig. 20 b) they increase ~2 ºC in that same region. During the Summer (Fig. 20 c) the majority of the domain exhibit temperatures ~18 ºC, except for the southern part, where temperatures exceed the 20º C, especially in the south-east corner of the domain with a warm tongue that reaches 24 ºC. A notable departure from the warmer summer and fall temperatures can be observed along the central coast of California, where the characteristic cold upwelling yields temperatures ~10 – 12 ºC. Figure 21 shows that the surface velocities are more energetic during the spring (Fig. 21 b) and summer (21 c)

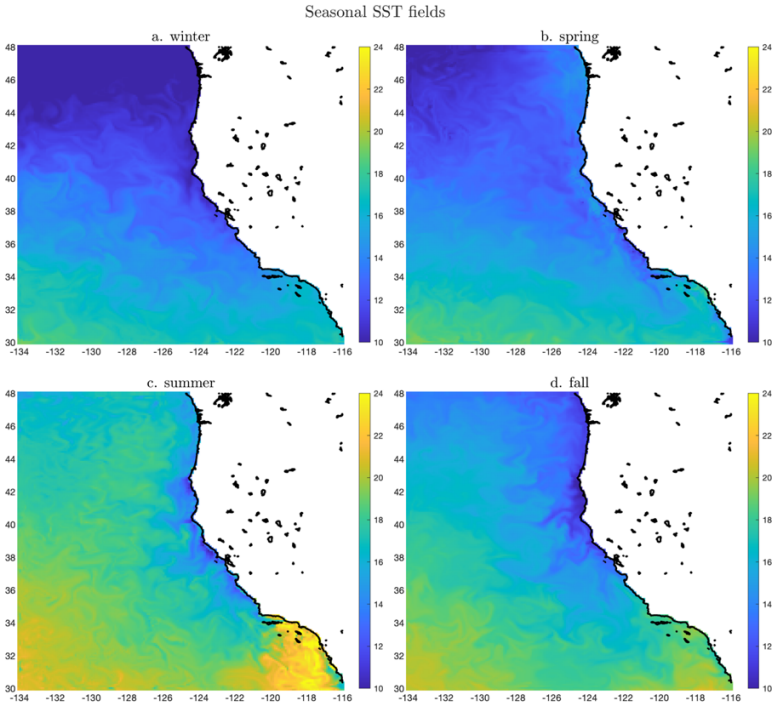


Figure 20 - Snapshots of a representative 4D-Var analysis of SST for each season. The cycles shown are a) 20180213_20180216, b) 20180512_20180515, c) 20180812_20180815 and d) 20181112_20181115 and are highlighted in as green diamonds in Figure 19.

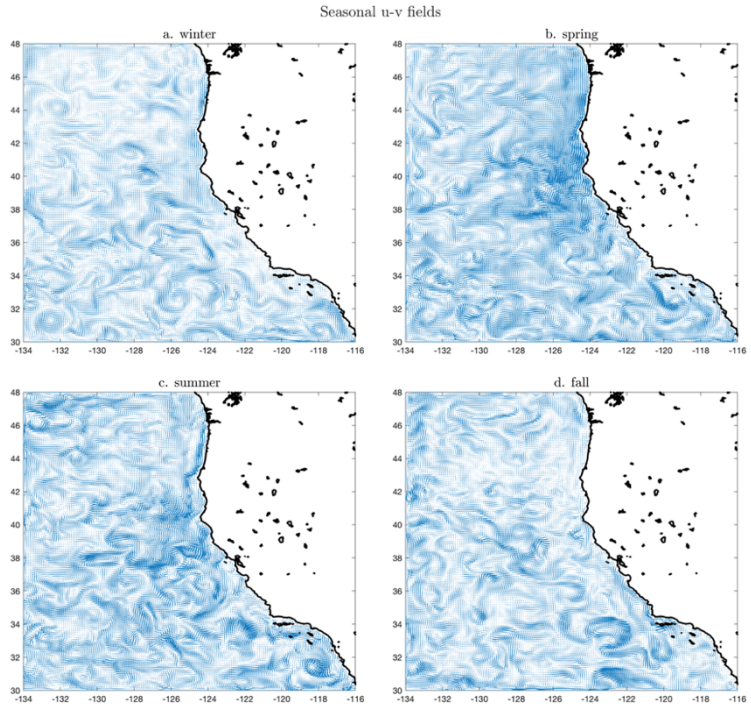


Figure 21 – Same as Fig. 20 except showing snapshots of representative 4D-Var analyses of surface velocity during each season.

As shown in Figure 3 the HF radar network provides information about the nearshore circulation, particularly the coastal upwelling regions. The influence of seasonal upwelling along the California coast is very evident in Figs. 20 and 21 in the form of cold temperatures during Summer and Fall. These features are highlighted in Figure 22 which shows the SST and surface velocities in the vicinity of the California coast. It is at these times of the year that the expected a priori and a posteriori diagnosed errors in Fig. 19 disagree the most. It is known that the model temperatures are biased at these times of the year (Veneziani et al, 2009) which is most likely a significantly contributing factor.

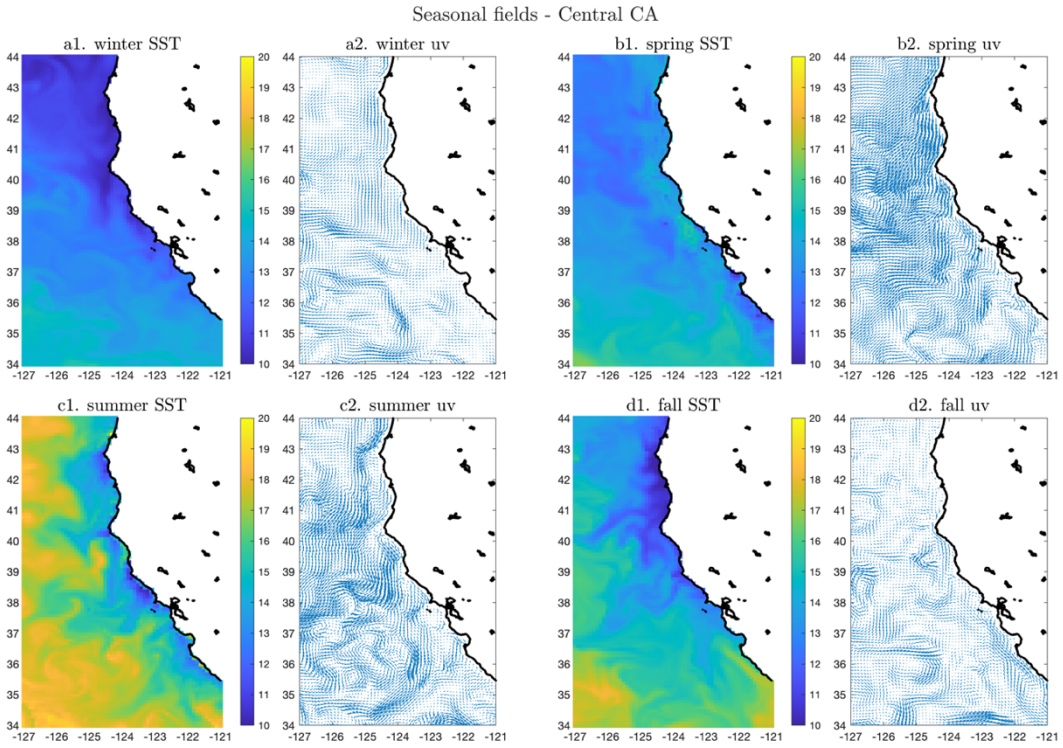


Figure 22 - Close up of seasonal the 4D-Var SST and surface current analyses along the California coast, and notably the upwelling during the summer and fall (c and d). The color bar was adjusted from Figure 20 to show variations in this part of the domain.

Since the innovations are expected to be unbiased and normally distributed, another way to compare the different experiments is to estimate the probability density functions (pdf) of the innovations, and evaluate the leading moments. In particular, given the seasonal dependencies identified in Figures 20-22, it is of interest to investigate how the structure of the pdf changes throughout the year. To do this, the pdf of the innovations was computed using a sliding window that spans five adjacent 4D-Var analysis cycles. The leading four moments of the pdfs were computed through time and are shown in Figure 23. Figure 23 suggests that for radial velocities the distribution of the innovation is fairly normal and unbiased throughout the year, while for temperatures this assumption does not hold during the summer, where the model appears to present a positive bias consistent with the findings of Veneziani et al (2009).

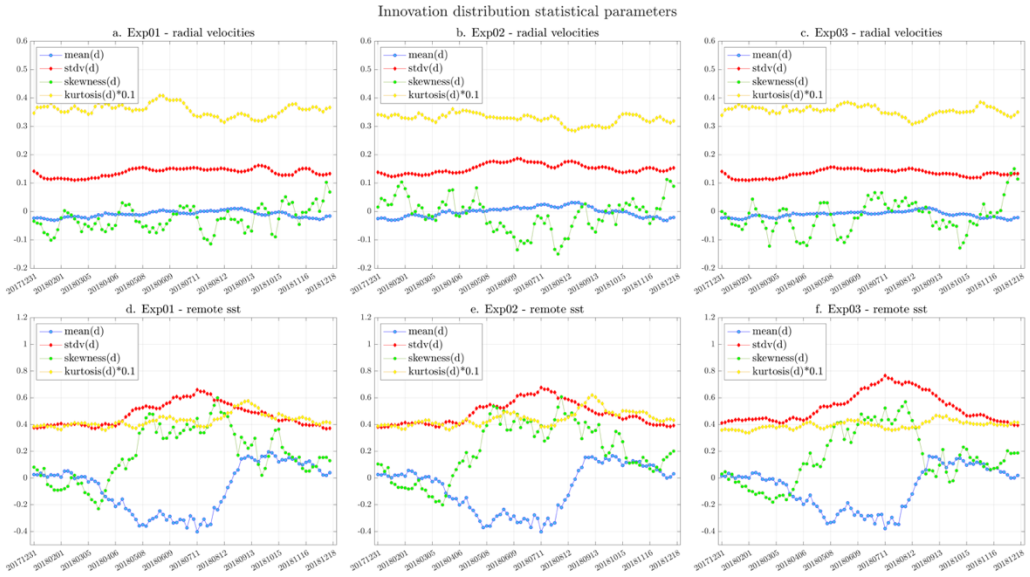


Figure 23 - Leading moments for the innovation pdfs for the multiple cycles experiments. a)-c) correspond to radial velocities innovations, and d)-e) correspond to remote SST. Note that the kurtosis (yellow line) is scaled by a factor of 0.1 for ease of presentation.

To examine seasonal differences in the innovation distributions for remote SST, the year was divided in four periods, JFM, AMJ, JAS, OND, that respectively represent winter,

spring, summer and fall, and the pdf of the innovations for each period is shown in Figure 24. Figure 24 a-e correspond to Exp01 (i.e. all observations) and Figures 24 f-j correspond to Exp03 (thinned remote SST). As expected from Figure 23, the spring and summer periods exhibit the largest departures from a normal distribution, with a higher (negative) bias and higher skewness, where in a normal distribution we would expect zero mean and zero skewness. The mean during all periods is smaller for Exp03 than Exp01, suggesting that thinning remote SST observations has attenuated the bias. The standard deviation is higher for Exp03 in all cases, and for both experiments is higher during the spring-summer period. The skewness for Exp03 is higher than in Exp01 except for the fall. The kurtosis in all cases is higher than the value of 3 expected for a Gaussian distribution, and it is higher for Exp01. Despite the described differences, overall the innovations for the experiment with all the observations (Exp01) and thinned observations (Exp03) have a similar behavior, suggesting that the thinning did not significantly improve or degrade the underlying 4D-Var hypothesis described in section 2.2.

Figure 25 shows the distribution of all the innovation and residual vectors for the whole period of one year. In all cases the innovation and residual pdfs show a better agreement with a Gaussian distribution than the SST counterparts. However, in general all the distributions are similar to a normal distribution, and the thinning experiments did not significantly change this behavior.

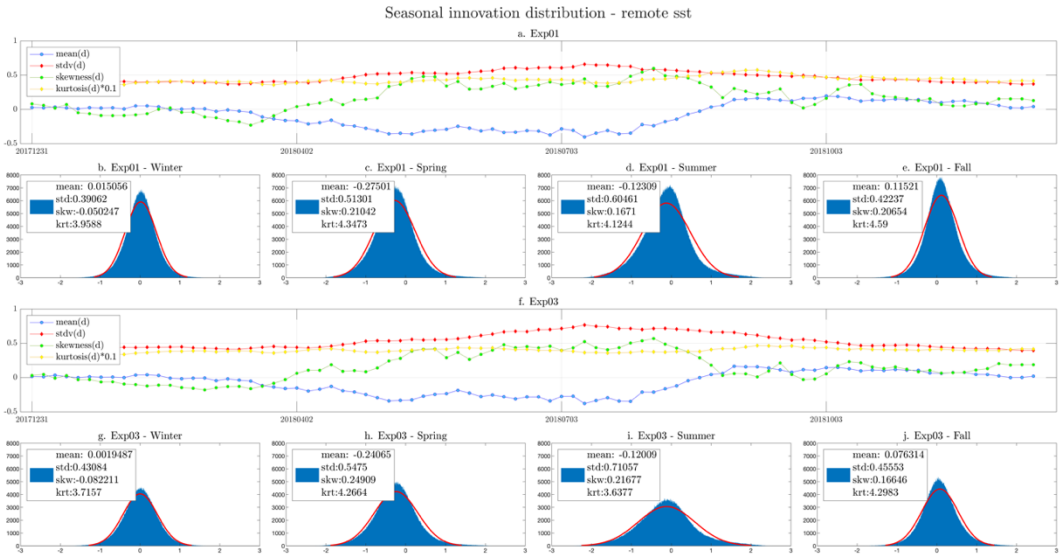


Figure 24 - In addition to the sliding window moments show in Figure 23, the actual innovation pdfs (in blue) are plotted in this case for the SST for 4 seasons for Exp01 (all the observations, a-e) and Exp03 (thinned SST observations, f-j). In red is shown the shape of a normal distribution with the same mean and standard deviation as that of each blue pdf.

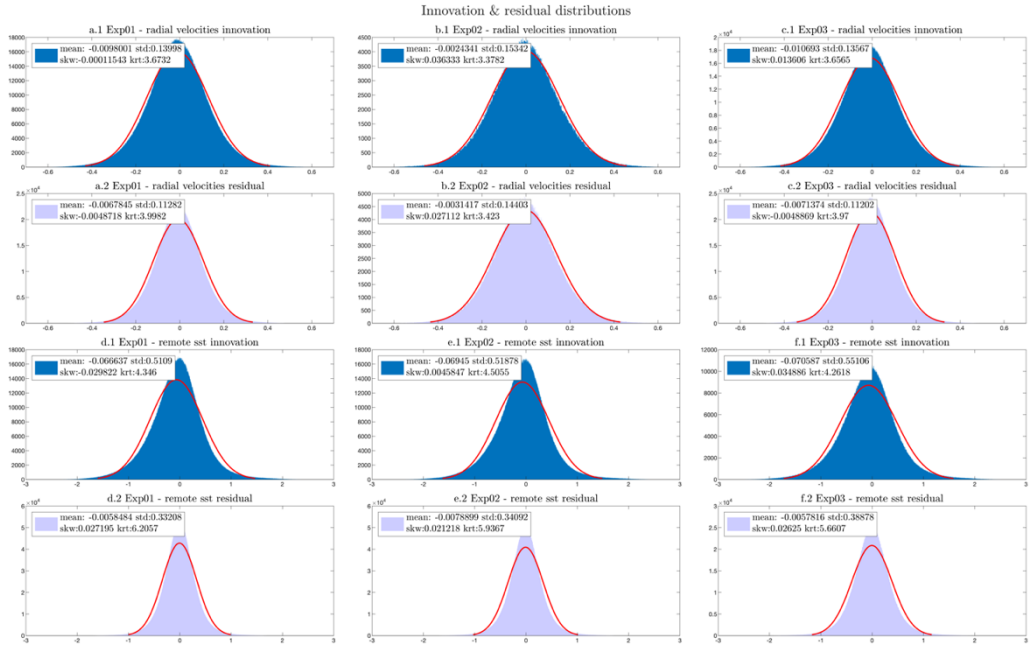


Figure 25 - Distribution for innovations and residuals as respectively defined in equations (11) and (12), for HF radar observations (a.1 - c.2) and remote SST observations (d.1 - f.2) for the whole year spanned in Exp01, Exp02 and Exp03.

4.2. Background error and observation error consistency

A widespread method used in NWP to check the consistency of the background and observation covariances individually (as opposed to the total expected error) based on the output of a data assimilation analysis has been presented by D05. Diagnosed a posteriori values for the background error and observation error variances (respectively Eq. 11 and Eq. 12) can be computed after performing the analysis, and then compared with the variances of the errors actually specified before the assimilation (D05). If inconsistencies are found (*i.e.* if diagnosed and actual variances differ significantly), the analysis can be performed again adjusting the observation and background error. Specifically, D05 showed that the average *expected* background error variance $\left(\sigma_i^b\right)^2$ and observation error variance $(\sigma_i^o)^2$ for observation type i and associated with a subset p_i of the observations are given by:

$$\widetilde{(\sigma_i^b)^2} = \sum_{j=1}^{p_i} \left(y_j^a - y_j^b\right)\left(y_j^o - y_j^b\right) \Big/ p_i \tag{19}$$

$$\widetilde{(\sigma_i^o)^2} = \sum_{j=1}^{p_i} \left(y_j^o - y_j^a\right)\left(y_j^o - y_j^b\right) \Big/ p_i \tag{20}$$

where y_j^o corresponds to the value of observation j, and the superscripts a and b refer to the analysis and background values associated with the same observation (D05).

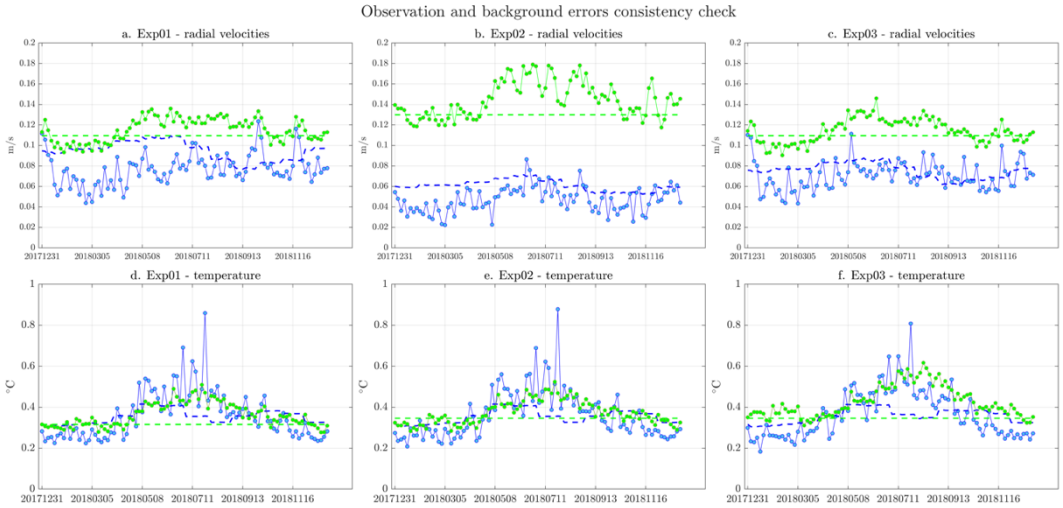


Figure 26 - Time series of the diagnosed (continuous lines) and a priori specified (dashed lines) errors for background (blue) and observations (green)

Figure 26 shows a comparison between the square root of the diagnosed values of equations (19) and (20) and the a priori specified background and observation errors. The top row corresponds to HF radar observations, and the bottom row corresponds to SST. The radial velocities errors in Exp01 (Figure 26 a) show good agreement for observations error, although this agreement is reduced in the spring and summer. The background errors are inconsistent for most of the year, where diagnosed values are in general lower than the a priori errors. These results were used to adjust the background error for radial velocities in Exp02, and Exp03, specifying them as a fraction of the initial background errors for the period, which increased consistency of the radial velocities background errors (Figures 26 b and 26 c, continuous and dashed blue lines). The observation errors for Exp02 (Figure 26 b, thinned HF radar observations) are consistent during the winter, but they show higher inconsistencies during the rest of the year, especially in the summer. This disagreement indicates that the thinning might have had an unfavorable effect during the seasons in which surface currents tend to be more energetic (*i.e.* see Figures 21 and 22), and that different thinning strategies are perhaps appropriate during different seasons. Figure 26

c shows an experiment with no thinning of HF radar observations, and after the a priori adjustment of the errors based on Exp01, the diagnostics show a better consistency. Summer is still the season with the higher differences between a priori and a posteriori estimated errors. In addition, for SST (Figure 26 d-f) all the experiments show good agreement during fall and winter, but errors are markedly inconsistent during the spring and summer.

4.3. Properties of the total error covariance matrix

As noted in sections 1.1. and 2.2, the BLUE of the ocean state-vector essentially depends on the properties of the stabilized representer matrix $\boldsymbol{P} = (\boldsymbol{HBH}^T + \boldsymbol{R})$, which can be interpreted as the total error covariance matrix in observation space. In general, the image of the unit sphere in $\mathbb{R}^n$ under any $\mathbb{R}^{mxn}$ $(m \geq n)$ matrix is an hyperellipse in $\mathbb{R}^m$ (Trefethen & Bau, 1997, pg. 25-26). In particular in statistical applications it is useful to study the properties of the hyperellipse defined by a covariance matrix (Friendly et. al., 2013). In this case, the semi-major axes of the hyperellipse are given by the eigenvectors of $\boldsymbol{P}$ and their lengths are proportional to the square roots of the eigenvalues of $\boldsymbol{P}$. It is of interest then to study how the different thinning strategies affect some of the properties of the hyperellipse defined by $\boldsymbol{P}$ since, as discussed in section 1.1, the primary motivation for thinning the observations is to alleviate potential problems associated with the behavior of $\boldsymbol{P}$.

In practice, as discussed in section 2.2, the hyperellipse defined by $\boldsymbol{P}$ can be characterized by the eigen space of its reduced rank non-dimensional counterpart $\widetilde{\boldsymbol{P}}$ (equation 10). Furthermore, the eigenvalues of $\widetilde{\boldsymbol{P}}$ and those of the tridiagonal matrix $\boldsymbol{T}_m$ are the same, and the EOFs of $\widetilde{\boldsymbol{P}}$ are given by $\boldsymbol{V}_m\boldsymbol{\phi}$ where $\boldsymbol{\phi}$ denotes an eigenvector of $\boldsymbol{T}_m$ and $\boldsymbol{V}_m$ is the

matrix of Lanczos vectors (M21). Thus, the properties of $\boldsymbol{T}_m$ obtained in each 4D-Var analysis can be calculated to obtain information about the error covariance hyperellipse. For example, the aspect ratio of the hyperellipse is given by the ratio of the largest and smallest eigenvalues which also represents the condition number of $\widetilde{\boldsymbol{P}}$. Thus, the eigenvalues of $\boldsymbol{T}_m$ provide a direct measure of how the condition number is influenced by thinning experiments presented here. Figure 27 a and 27 b respectively show how the condition number varies for the different radial velocities and SST thinning experiments. For radial velocities, the condition number for all the thinning experiments is slightly smaller than the experiment with all the observations (sc.1), which is what might be expected by reducing observation density through *superob*ing thus reducing the linear dependence of the columns of $\widetilde{\boldsymbol{P}}$. On the other hand, the SST thinning experiments present a different behavior, where the condition number for the thinning cases is larger than the case with all the observations. In this sense, the IDT algorithm seems to exhibit an undesired behavior, by detecting regions of high spatial variance, it reduces the dimension of the matrix but the retained columns are more similar to each other than before the thinning. This result puts in evidence the fact that both thinning techniques presented here use very different strategies. In the case of superobing, the initial data set has many observations in every cell defined by four adjacent grid points, and the aim of the thinning is to obtain one observation per grid point, in a restricted region of the domain (i.e. near the coast). The IDT algorithm begins instead with one SST observation for each grid point in the whole domain, and it aims to discard a lot of these observations that do not provide significant new information to the analysis.

Figure 27 c and d show how the eigen spectrum of $\widetilde{\boldsymbol{P}}$ changes after thinning the observations, suggesting that there is a change in the major axis of the hyperellipse defined

by the error covariance matrix. In this case, experiments thinning radial velocities (Fig 27 c) and SST observations (Fig 27 d) similarly amplify the leading member of the eigen spectrum. The main difference in these cases is that for the HF radar thinning experiments (Fig 23 c), the ratio of the second largest to the largest eigenvalue is conserved (~0.3 in all cases), while for the SST (Fig 23 d) it is reduced in half after thinning (~0.3 for sc.1, vs ~0.15 in all the other cases), suggesting that the aspect ratio of the hyperellipse changes in particular for the SST thinning experiments.

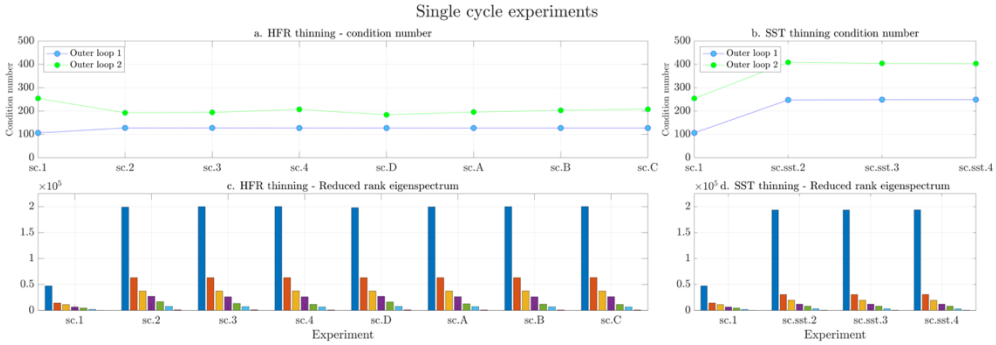


Figure 27 – The condition number of $\tilde{\boldsymbol{P}}$ for each of the single cycle experiments where thinning is applied to (a) HR radar observations, and (b) SST observations. The eigen spectra of $\tilde{\boldsymbol{P}}$ for the same series of single cycle experiments are shown in (c) and (d).

Figure 28 – rms contribution of different variables to leading EOF. The vertical axes have different scales for each row, with one order of magnitude of difference between adjacent rows.

The leading EOF is of particular interest because it defines the direction in the observation space associated with the largest fraction of total error variance. Since each dimension of $\widetilde{\boldsymbol{P}}$ is associated with an individual observation, it is possible to quantify the contribution of different observation types to the leading EOF direction. With this in mind, Figure 28 shows the rms of the elements of the leading EOF for each observation type. Although the main contributor for all cases is in situ salinity (by an order of magnitude, Figs 28 g, 28 h and 28 i), Figs 28 d, 28 e and 28 f indicate that the thinning of HF radar and SST observations causes a rotation of the leading EOF in the direction of favored more by SSH observations.

4.4 Observation Impacts

The analysis increment described in equation (7) can be also expressed as a linear combination of so-called *array modes*, leading to an alternative definition of the BLUE stated in (1):

$$\boldsymbol{x}^a = \boldsymbol{x}^b + \sum_{i=1}^{N} \alpha_i \boldsymbol{\Psi}_i \qquad (21)$$

where $\boldsymbol{\Psi}_i = \boldsymbol{B}\boldsymbol{H}^T\boldsymbol{V}_m\boldsymbol{\phi}_i$ denote the array modes that were introduced by Bennett (1985), and $\boldsymbol{V}_m\boldsymbol{\phi}_i$ are the EOFs of $\widetilde{\boldsymbol{P}}$ (M21) introduced in section 4.3. The array modes therefore represent the projection of the EOFs of $\widetilde{\boldsymbol{P}}$ into the state-vector space via the transformation $\boldsymbol{H}^T$ (M21). The weight coefficients α_i depend on the innovation vector and are given by $\alpha_i = \lambda_i^{-1}\boldsymbol{\Psi}_i^T\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{d}$. M21 have shown that the impact of each observation on the 4D-Var analysis can be quantified with the contribution of each observation. to the α_i associated with the smallest eigenvalue λ_i . Figure 29 shows the contribution of each observation type to the 4D-Var analyses from the different thinning experiments. In Exp01 (Fig 29 a) the observation impact is dominated by SST, followed by in situ salinity and HF radar

observations. In this type of analysis SST observations seem to also have a high seasonal component, where the impacts are markedly higher during the summer. The analysis for Exp02 (Fig 29 b) shows that for the chosen superobing strategy, the HF radar observations have very little impact on the analysis, suggesting that the thinning strategy used was too severe. Figure 29 c shows that when SST are thinned, their impact decreases, but it still has a noticeable component, and the impact of HF radar observations is higher than in Exp01, leading to a more even distribution of observation impacts.

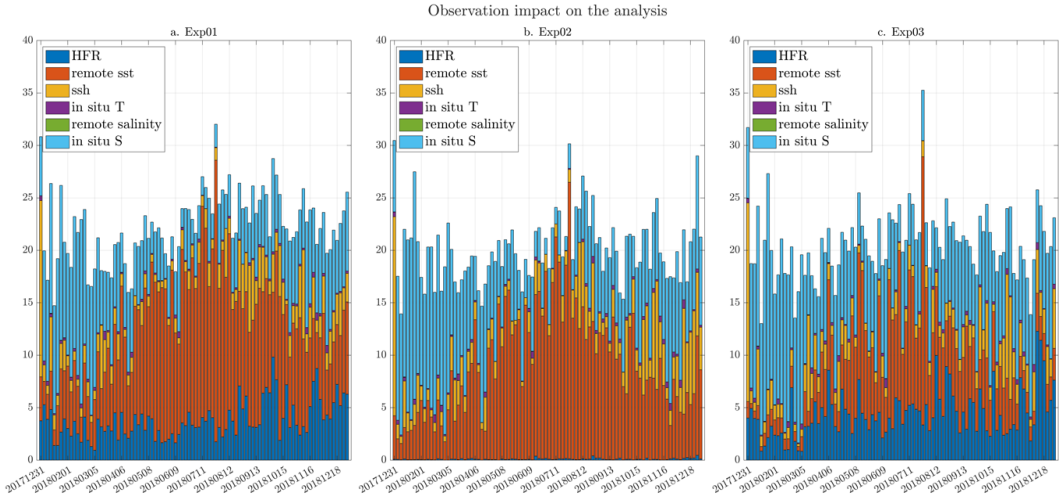


Figure 29 – Observation impact on the analysis.

5. Summary and conclusions

In this thesis we implemented data thinning algorithms that are in general applied in operational NWP but here were applied in the context of a near real-time ocean data assimilation system. We explored the effects that data thinning has on the observation space and on the 4D-Var analysis. The BLUE of the ocean state-vector calculated by the 4D-Var approach is directly related by the stabilized representer matrix, which represents the total error covariance, and therefore its dimension is determined by the dimension of the observation vector. Thinning observations changes the observation space, hence the

error covariance, and therefore it fundamentally affects the solution obtained after assimilating data.

We have focused our thinning efforts on two types of remote sensed observations: HF radar radial velocities and satellite remote sensing observations of SST which form part of a multiple platform observing system for the California Current System. The HF radar observations were thinned by combining multiple observations into one superobservation per grid point. On the other hand, SST observations were thinned by detecting regions of high spatial variance and discarding redundant observations. Different experiments were performed for a single 4D-Var cycle to test and debug the thinning algorithms. In this context, the a priori and a posteriori analysis of the errors showed that thinning in general increased the errors, but the calculated uncertainties were still within a tolerable margin as evidenced by comparing with expected error variance.

We then extended the experiments to run 4D-Var for a period of one year, which allowed a deeper exploration of the applied thinning schemes. Since evaluation of the results is neither simple or straightforward, different strategies were used to analyze and compare results among different experiments. An analysis of the innovations statistics is a useful and practical check on the behavior of the DA results, and routinely used to evaluate the performance of operational NWP systems. Since the innovations are expected to be normally distributed and unbiased, we checked that all the experiments showed a good agreement with this assumption, except during upwelling season, where the model is known to be biased. We also calculated the diagnosed statistics described by D05, where an analysis of the increments can be compared with the a priori defined background and observation errors. For the radial velocities, some inconsistencies found in Exp01 were corrected in subsequent experiments. For SST, the inconsistencies were more evident and

seasonal, suggesting that maybe a seasonal strategy should be taken to define the a priori errors. The statistical analysis showed that the thinning experiments did not seem to degrade the analysis, suggesting that these techniques could be further explored and systematically applied in this type of numerical ocean modeling system in the future.

Another approach to study the results was based on numerical linear algebra, analyzing how thinning observations affected the eigen space defined by a reduced rank approximation of the total error covariance matrix. In both sets of single cycle thinning experiments the eigen spectrum was amplified. In the case of thinning radial velocities, the aspect ratio of the hyperellipse defined by the error covariance matrix was conserved, while in the SST thinning experiments the ratio of the second largest eigenvalue to the largest eigenvalue was reduced. The results of the thinning were further explored by estimating the impact of different types of observations on the analysis with the reduced rank array modes. This analysis provided a new insight in the thinning of radial velocities, where the approach taken seems to have drastically reduced the impact of HF radar observations on the analysis. This result suggests that a different strategy using a higher resolution for the superobservations could yield better results. On the other hand, thinning SST seems to have a good effect on evening out the impact each type of observation had on the analysis.

This was a first attempt to apply techniques widely used in NWP in the context of ocean DA. The algorithms were implemented in a way to allow the use of different thinning methods for different observation platforms, and different thinning criteria could be also easily applied to different seasons. Based on the statistical analysis, a different treatment of the upwelling season could yield more satisfactory results. Furthermore, the results

obtained with the observation impact analysis suggest that different thinning strategies should be further explored, in particular for radial velocities. The thinning could be also extended to different types of observations in order to improve the characteristics of observation space that affect the 4D-Var analysis. The next step will be to evaluate the various thinning algorithms within the current near real-time analysis-forecast system that is run routinely at UCSC in support of U.S. IOOS and the Central and Northern California Ocean Observing System (CeNCOOS).

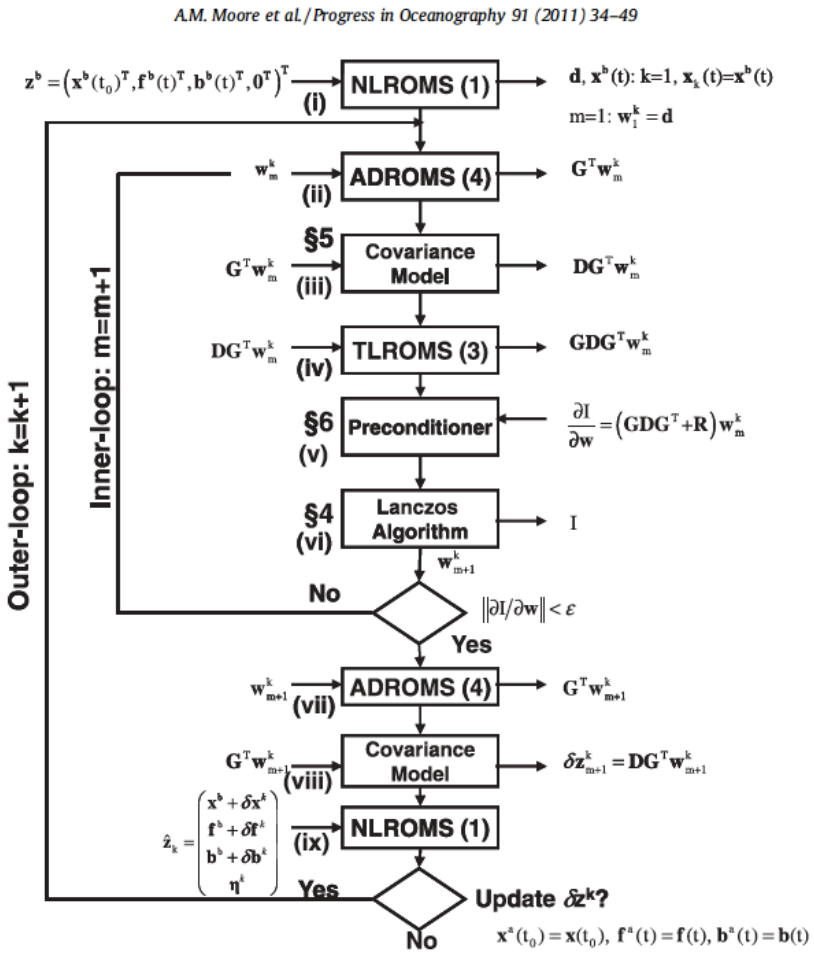
Appendices

A – IDT pseudocode

```
;Variables:
;data- original data
; ndata – thinned data
; threshold – user specified threshold
; endpoints – Max(X,Y) , min(X,Y)

function main (data , threshold)
{
        Data = Normalize(data)
        globalMean = Mean(data)
        thinnedData = IDT(data, ndata, globalMean, threshold, endPts)
        return thinnedData
}

function IDT (data,ndata, globalMean, threshold, Bounds)
{
        If ( size(data) == 4)
        {
                ndata = data
        }
        else
        {
                Q1,Q2,Q3,Q4= Divide(data) //returns quadrant end points
                Quadrant(data,ndata,globalMean,threshold,Q1)
                Quadrant(data,ndata,globalMean,threshold,Q2)
                Quadrant(data,ndata,globalMean,threshold,Q3)
                Quadrant(data,ndata,globalMean,threshold,Q4)
        }
}

function Quadrant ( data, ndata,globalMean, threshold, Q)
{
        aStdDev = globalMean*threshold
        varSample = Calc_ObjectiveMeasure(data, Q)
        if (varSample > (aStdDev*aStdDev))
                IDT(data,ndata,globalMean, threshold,Q)
        else
                ndata[midPointX,midPointY] = data[midPointX, midPointY]
}
```

Pseudocode for IDT algorithm (from R05)

B - 4D-Var PSAS flow chart



Flow chart for the PSAS variant of 4D-Var as implement in ROMS (from M11)

C - Tables

Experiment	Description	Details	Total observations (all variables)	Total radar obs	% of total radar observations
sc.1	All observations	No thinning	218,251	87,435	100.0
sc.2	Thinning	keep 1 every 2 observations	174,534	43,718	50.0
sc.3	Thinning	keep 1 every 5 observations	148,303	17,487	20.0
sc.4	Thinning	keep 1 every 10 observations	139,560	8,744	10.0
sc.A	Superobs	One superob per model grid point	141,365	10,549	12.1
sc.B	Superobs	One superob per model grid point, only if there are 3 or more observations in the vicinity of that grid point	136,957	6,141	7.0
sc.C	Superobs	Lower resolution grid	133,864	3,048	3.5
sc.D	Superobs	Higher resolution grid	156,209	25,393	29.0

Table 1 – Description of preliminary single cycle experiments for radar observations.

Experiment	Lx	grid_step	min_number_obs	finer_grid
sc.A	0.1	1	1	off
sc.B	0.1	1	3	off
sc.C	0.2	2	3	off
sc.D	0.05	1	1	on

Table 2 – Parameters defined for each experiment in radar superobing algorithm.

Experiment	Description	Details
Exp 01	All observations	No thinning
Exp 02	Radial velocities superobs - full SST	Radial velocities superobs using criteria of sc.B
Exp 03	Full radial velocities - Thinned SST	SST Thinned using criteria of sc.sst.1

Table 3 – Experiments spanning a period of one year.

References

Bennett, A. F. ,1985. Array design by inverse methods. Progress in Oceanography, 15, 129–156.

Carr, Mary-Elena, Kearns, Edward J "Production Regimes in Four Eastern Boundary Current Systems." Deep-sea research. Part II, Topical studies in oceanography 50.22-26 (2003): 3199–3221. Web.

Desroziers, G. "Diagnosis of Observation, Background and Analysis-error Statistics in Observation Space." Quarterly Journal of the Royal Meteorological Society, vol. 131, no. 613, John Wiley & Sons, Ltd., 2005, pp. 3385–96, doi:10.1256/qj.05.108.

Dorman, Clive E. "Impact of U.S. West Coastline Inhomogeneity and Synoptic Forcing on Winds, Wind Stress, and Wind Stress Curl During Upwelling Season." Journal of Geophysical Research: Oceans 118.9 (2013): 4036–4051. Web.

Edwards, Christopher A. "Regional Ocean Data Assimilation." Annual Review of Marine Science, vol. 7, no. 1, 2015, pp. 21–42, doi:10.1146/annurev-marine-010814-015821.

Fiechter, Jerome. "Wind, Circulation, and Topographic Effects on Alongshore Phytoplankton Variability in the California Current." Geophysical Research Letters 45.7 (2018): 3238–3245. Web.

Friendly, Michael, et al. "Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry." *Statistical Science*, vol. 28, no. 1, 2013, pp. 1–39. *JSTOR*, www.jstor.org/stable/43288410.

Ide, K., Courtier, P., Ghil, M. and A.C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. *Journal of the Meteorological Society of Japan*, **75**, 181-189.

Lahoz, William Albert, et al. Data Assimilation : Making Sense of Observations . Springer, 2010.

Lazarus, S. et al. (2010) Evaluation of Data Reduction Algorithms for Real-Time Analysis. Weather and Forecasting. [Online] 25 (3), 837–851. [online]. Available from: http://search.proquest.com/docview/869573065/.

Lorenz, Edward N. The Essence of Chaos . First paperback edition. Seattle: University of Washington Press, 1995. Print.

Marchesiello, Patrick, et al. “Equilibrium Structure and Dynamics of the California Current System.” Journal of Physical Oceanography, vol. 33, no. 4, 2003, pp. 753–783.

Moore, Andrew M. “A Comprehensive Ocean Prediction and Analysis System Based on the Tangent Linear and Adjoint of a Regional Ocean Model.” Ocean Modelling (Oxford), vol. 7, no. 1-2, Elsevier Ltd, 2004, pp. 227–58, doi:10.1016/j.ocemod.2003.11.001.

Moore, Andrew M., et al. “Regional Ocean Modeling System (ROMS) 4-Dimensional Variational Data Assimilation Systems Part I – System Overview and Formulation.” Progress in Oceanography, vol. 91, no. 1, Elsevier Ltd, 2011, pp. 34–49, doi:10.1016/j.pocean.2011.05.004.

Moore, Andrew M., et. al. “Synthesis of Ocean Observations Using Data Assimilation for Operational, Real-Time and Reanalysis Systems: A More Complete Picture of the State of the Ocean.(Report).” Frontiers in Marine Science, vol. 6, no. 90, Frontiers Research Foundation, 2019, pp. 1–7, doi:10.3389/fmars.2019.00090.

Moore, Andrew M., et. al. Assessing the Performance of an Ocean Observing, Analysis and Forecast System for the Mid-Atlantic Bight Using Array Modes, 2021, in press.

Ochotta, T. “Adaptive Thinning of Atmospheric Observations in Data Assimilation with Vector Quantization and Filtering Methods.” Quarterly Journal of the Royal Meteorological Society 131.613 (2005): 3427–3437. Web.

Ramachandran, R., Li, X., Movva, S., Graves, S., Greco, S., Emmitt, D., Terry, J. and Atlas, R. ‘Intelligent Data Thinning Algorithm for Earth System Numerical Model Research and Application’. In Proceedings of 21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, San Diego. American Meteorol. Soc, Boston, USA (2005)

Regional Ocean Modeling System (ROMS) - WikiROMS. (2021). Www.Myroms.Org. https://www.myroms.org/wiki/Regional_Ocean_Modeling_System_(ROMS)

Shchepetkin, Alexander F., McWilliams, James C. “The Regional Oceanic Modeling System (ROMS): a Split-Explicit, Free-Surface, Topography-Following-Coordinate Oceanic Model.” Ocean Modelling (Oxford), vol. 9, no. 4, Elsevier Ltd, 2005, pp. 347–404, doi:10.1016/j.ocemod.2004.08.002.

Talagrand, Olivier. “Assimilation of Observations, an Introduction.” Journal of the Meteorological Society of Japan, Tokyo, Japan, 1997 (1997): 191–209. Web.

Talley, Lynne D. Descriptive Physical Oceanography (6th Ed.). Academic Press, 2011

Trefethen, Lloyd N, and David Ill Bau. *Numerical Linear Algebra*. SIAM Society For Industrial And Applied Mathematics, 2000.

Veneziani, Milena & Edwards, Chris & Doyle, J. & Foley, D.. (2009). A Central California coastal ocean modeling study: 1. Forward model and the influence of realistic versus climatological forcing. Journal of Geophysical Research. 114. 10.1029/2008JC004774.