Statistical Methods for High-Dimensional Turbulent Systems with Applications

by

Jeffrey M. Covington

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Mathematics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2023

Date of final oral examination: 12/19/2023

The dissertation is approved by the following members of the Final Oral Committee:

Nan Chen, Assistant Professor, Mathematics
Sam Stechmann, Professor, Mathematics
Sebastien Roch, Professor, Mathematics
Till Wagner, Assistant Professor, Atmospheric and Oceanic Sciences

Dedicated to my family.

ACKNOWLEDGMENTS

I am grateful to my advisor Nan Chen for his dedication, support, and guidance. As a teacher and a mentor Dr. Chen has been exemplary. His deep knowledge of mathematics, research skill, and caring for his students has made him a great advisor, and I am thankful for the opportunity to work with him over the past five years.

I am also thankful to my coauthors and collaborators: Nan Chen, Monica Wilhelmus, Di Qi, Rosalinda Lopez, Bryan Xu, Evelyn Lunasin, and Steve Wiggins among others. The many discussions I've had over the years have been fruitful as well as immensely helpful in developing my own knowledge and skills. Working with so many great scholars has been amazing and an indispensable part of my experience.

My research has been partially funded by the Institute for the Foundations of Data Science (IFDS) as well as the Multidisciplinary University Research Initiatives (MURI) program of the Office of Naval Research (ONR) and I am very grateful for the opportunities and support that these programs have offered me.

I would also like to acknowledge my committee members: Nan Chen, Sam Stechmann, Sebastien Roch, and Till Wagner. Their feedback have helped shaped this dissertation into the best that it can be.

Lastly I am thankful to my family for their love and encouragement.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

High-dimensional turbulent systems appear frequently throughout science and engineering where the high-dimensionality and nonlinearity of these systems pose a persistent challenge. Due to the chaotic nature of these systems as well as the uncertainty that arises in real applications, a statistical approach is useful where the high-level statistical averages are used to analyze and model their behavior. This dissertation presents several ways in which statistical methods can be incorporated into analyzing turbulent systems, including applications to real satellite data of the Arctic. First, a method for dynamically interpolating between missing observations of sea ice floe trajectories is presented. These satellite-based observations of individual sea ice floe trajectories are often obscured by clouds, leading to gaps in the dataset. The dynamical interpolation method uses a balanced physics-based and data-driven approach to address the high-dimensionality and nonlinearity of the coupled ice-ocean-atmosphere system. Second, effective strategies for statistical control of turbulent dynamical systems are presented. Statistical control offers an efficient and robust approach to the control of turbulent dynamical systems and new strategies are developed which extend the statistical control framework to scenarios with large initial perturbations and changes in the dynamical regime. Lastly, a framework for utilizing observations of ice floe trajectories for ocean eddy identification is developed. Eddies play an important role in the Earth's ocean and climate systems yet eddy identification is typically restricted to ice-free regions of the Arctic. By utilizing observations of ice floe trajectories this framework aims to extend eddy identification capabilities to provide valuable insights into the Arctic.

1 BRIDGING GAPS IN THE CLIMATE OBSERVATION NETWORK: A PHYSICS-BASED NONLINEAR DYNAMICAL INTERPOLATION OF LAGRANGIAN ICE FLOE MEASUREMENTS VIA DATA-DRIVEN STOCHASTIC MODELS

The following chapter is adapted from an open-access paper [38] published in *Journal of Advances in Modeling Earth Systems* (*JAMES*) under a creative commons license. The authors, Jeffrey Covington, Nan Chen, and Monica M. Wilhelmus, all contributed to the research and writing of the paper.

The research of Nan Chen was partially funded by Office of Naval Research (ONR) Multidisciplinary University Initiative (MURI) award N00014-19-1-2421. Monica M. Wilhelmus was funded by the ONR awards N00014-20-1-2753 and N00014-19-1-2421. Jeffrey Covington was supported as research assistant under this grant and by the National Science Foundation award DMS-2023239 through the Institute for Foundations of Data Science (IFDS) at UW-Madison. The authors also acknowledged Dr. Georgy Manucharyan for his insightful discussions, and Dr. Rosalinda Lopez-Acosta for her work on the development of the Ice Floe Tracker algorithm.

1.1 Introduction

Sea ice plays a key role in the Arctic climate system [132, 143, 144, 84, 109, 17]. It modulates important momentum, heat, and material transfer processes between the ocean and the atmosphere [133, 129, 134]. Given the sensitivity of the sea ice cover to global warming trends, the observation and modeling of sea ice are critical for understanding global climate, including monitoring the drastic changes in the Arctic and assessing possible future climate scenarios.

Earth system models typically characterize sea ice as a continuum with viscous-plastic rheology primarily through ice concentration, volume, and thickness [66, 137, 68]. While this traditional modeling approach yields realistic results at the basin scale, at scales of $\mathcal{O}(10)$km and smaller, sea ice exhibits brittle behavior with the motion of individual fragments deviating from a continuum description. In this case, the discrete element method (DEM) [41, 40, 65], which characterizes the trajectories of individual ice floes, as opposed to clusters of ice, becomes the natural choice to describe sea ice dynamics. Compared to continuum models, the modeling of individual floes provides a richer representation of sea ice dynamics through local interactions with the oceanic and atmospheric components [86, 138]. In addition, since the DEM models are developed under Lagrangian coordinates, there is no need for an advective transport scheme to move floes between grid cells as in continuum models, significantly reducing computational costs. The DEM models can also change the spatial resolution as the geophysical situation requires, allowing greater flexibility in the study of sea ice.

The unique advantages and wide applications of models based on the DEM

highlight the need for observational sea ice data within the Lagrangian framework. Observed floe trajectories facilitate the development and calibration of DEM models and provide insight into the evolution of sea ice properties. However, despite the increase in satellite missions and the improved techniques in acquiring remote sensing observations, most existing observational products are based on Eulerian descriptions of the sea ice drift field. Exceptions include the Arctic Ocean Sea Ice Drift Reprocessed [9] and the Making Earth System data records for Use in Research Environments (MEaSUREs) programs [80]. Yet, these measurements cannot adequately resolve sea ice motion at small scales due to the spatial resolution (31.25km for Arctic Ocean Sea Ice Drift Reprocessed product) or the sampling frequency (3-day interval for MEaSUREs) of the data. On the other hand, in situ field measurements using buoys on ice floe surfaces have provided invaluable information, but trajectories are often sparse [21, 56, 82, 71, 69] as the marginal ice zone is undersampled relative to central Arctic regions.

Recently, a new Lagrangian floe tracking algorithm, called the Ice Floe Tracker [87, 148], was developed and applied to optical satellite images. It creates Lagrangian sea ice measurements in the low-sampled regions between the ice pack and the open ocean, commonly known as the Marginal Ice Zones (MIZ). This data set was the first of its kind in that it provides not only the Lagrangian trajectories but also the floe sizes and geometries together with the angular displacements of floes in the MIZ extending throughout the 21st century. Given the demonstrated link between floe rotation rates and the characteristics of the underlying small-scale ocean eddies in the Beaufort Gyre MIZ, these sea ice floe observations have

proved to be essential for recovering the state of the underlying turbulent ocean field, providing a unique insight into the multi-scale nature of the ocean [107].

Atmospheric noise is visible in optical images and leads to many one- or two-day gaps in the retrieved trajectories within the Lagrangian Ice Floe Tracker data set. See Figure 1.1 for an example. A commonly used approach to filling these gaps is to interpolate between the available observations through linear interpolation [107]. These trajectories (and their curvature) are used to characterize ocean circulation/eddy behavior. However, such an approach ignores the MIZ dynamics, leading to trajectories lacking many physical properties. Linear interpolation also fails to retrieve the curvature of the trajectories, which is essential for characterizing the turbulent ocean flow field within the meso/submeso-scale range in polar regions. Alternatively, physics-based dynamical interpolation incorporates both the available partial observations and knowledge of the coupled atmosphere-ocean-floe dynamics. While the resulting interpolated trajectories are expected to reflect reality better, traditional dynamical interpolation approaches are often extremely slow and computationally expensive due to the high dimensionality and nonlinearity of the underlying system [62, 19].

This paper presents an efficient and statistically accurate nonlinear dynamical interpolation framework for recovering the missing floe observations in Lagrangian trajectories. It exploits a balanced physics-based and data-driven construction to address the challenges posed by the high-dimensional and nonlinear nature of the coupled system. This new method involves a sequential prediction-correction procedure. The error from predicting the missing values in the coupled atmosphere-

ocean-floe system is mitigated by incorporating the available observations of floe positions and orientations via Bayesian inference. One crucial feature of the presented framework is that it exploits a data-driven reduced-order stochastic modeling strategy to advance the statistical forecast of the atmosphere and ocean fields, which are the underlying driving forces of the sea ice motion, but are not considered by the direct curve-fitting algorithms. Particularly, these simple stochastic models describe the time evolution of the leading spectral modes of the atmosphere and ocean fields, where effective stochastic forcing is adopted to characterize the fluctuations at the unresolved scales. Therefore, the resulting surrogate models significantly reduce the computational cost at the forecast step, which is the most time-consuming part in traditional dynamical interpolation approaches. It is worth highlighting that closed analytic formulae are available for expressing the statistics associated with these simple stochastic models, facilitating the systematic and efficient data-driven model calibration. The calibrated models succeed in accurately predicting the atmosphere and ocean states and the associated uncertainty. The latter is crucial in reaching the least biased state estimate using nonlinear dynamical interpolation, especially in the presence of strong turbulence, which is again completely missed by deterministic curve fitting methods. In addition, the framework allows for simultaneous estimation of several critical physical parameters that cannot be directly inferred from satellite images but are essential for the dynamical interpolation, such as the thickness of each floe, using only relatively short floe trajectories.

The rest of the paper is organized as follows. It starts with the development of the physics-based data-driven dynamical interpolation framework. Then the new

method is tested on both a synthetic data experiment and the real data set of floe trajectories in the Beaufort Sea MIZ. The focus here is on the non-interacting floes, but the framework can be easily extended to the interacting ones. The study also includes analysis of the resulting interpolated Lagrangian ice floe trajectories and angular displacements as well as the recovery of several key physical properties of the floes and their associated statistics. The recovered floe trajectory utilizing the traditional linear interpolation approach will serve as a benchmark solution.

1.2 The Reduced-Order Modeling and Nonlinear Dynamical Interpolation Framework

This section presents an overview of the new modeling and nonlinear dynamical interpolation framework consisting of four key steps. The technical details of the model, the data and the methods are included in Appendix.

This framework works with a coupled atmosphere-ice-ocean system. While this system can take the form of a coupled dynamical model, the framework also allows the atmospheric and/or oceanic components to be given as numerical data. Here, a DEM model is used to characterize the sea ice dynamics, where the individual floe shapes and sizes are drawn from a library of floe observations in the Beaufort Gyre MIZ [88]. Given that the observations contain only nearly non-interacting and shape-preserving floes, the ice floe motion can be assumed to be mainly driven by oceanic and atmospheric forcing, which are calculated from surface integrals over floe shapes. The ocean component is given by a two-layer quasi-geostrophic

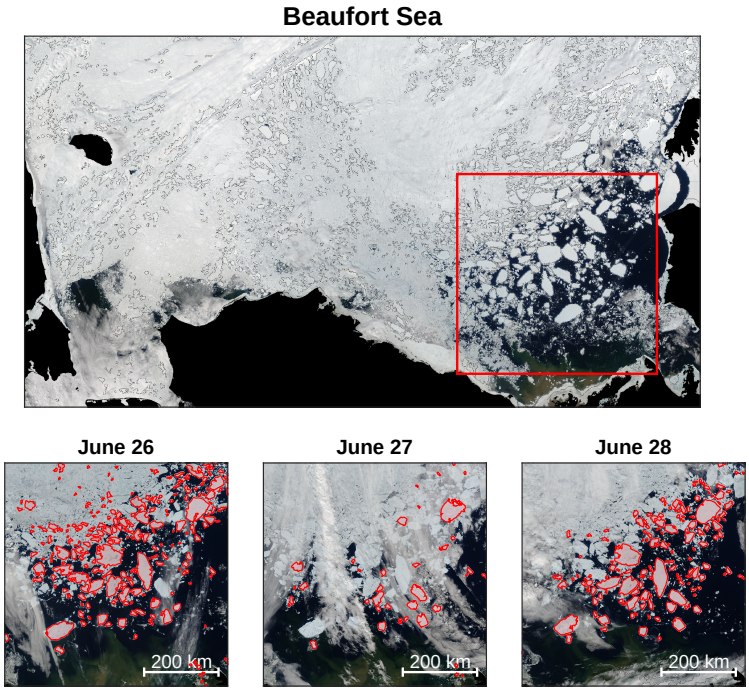

Figure 1.1: Sea ice floes in the Beaufort Sea MIZ. Top panel: Representative Moderate Resolution Imaging Spectroradiometer (MODIS) True Color image (downloaded from the NASA Worldview application) displayed in a WGS 84/NSIDC Sea Ice Polar Stereographic North 70" N projection. For only this figure, the image is oriented 90° from standard Polar Stereographic coordinates so that the top of the image is roughly north. The red box outlines the region of interest. Bottom panels: The observed MIZ of the Beaufort Sea (the box area of the top panel) is shown on three consecutive dates (26.06.2008 to 28.06.2008). Identified floes are contoured with red. On 27.06.2008, the atmospheric noise acts to blur ice floe contours impeding the effective identification of most of the floes.

(QG) model that generates eddies from baroclinic instabilities, the so-called Phillips model [139], in which the cumulative impact of many passing floes on the turbulent eddy field is represented via a quadratic surface drag. The model has been systematically calibrated to capture the key features of the interaction between the

Arctic ocean and the ice floes. The optimization criterion used here is to match the simulated and observed scale-dependency of the foe rotation variance, where the bulk vertical shear of the background horizontal velocity, the deformation radius, and the effective ratio between the top and bottom layer depths are the tuning parameters. See [107] for the detailed calibration strategy. The spatial resolution is 128×128 gridpoints. Because of the high latitude, the deformation radius is small: on the same order as the model resolution. Despite potential unresolved effects, their impact is mitigated by the relatively large floe areas and statistical averaging of the dynamical interpolation. The atmospheric component is taken from a reanalysis product (ERA5) [115, 37], which provides Eulerian wind velocity fields over the observational period. As the wind field exhibits larger-scale features, a coarser spatial resolution of 11×11 gridpoints is used. Note that the focus here is in the MIZ of the Beaufort Sea (see Figure 1.1). Hence, a double-periodic boundary condition is adopted for simplicity. The potential model error and bias introduced from the various approximations can be mitigated at the statistical forecast stage using the stochastic corrections and Bayesian inference in the framework introduced in the rest of this section. The domain size, as shown in Figure 1.1, is roughly 600km$\times$600km. Figures 1.2 and 1.3 include a schematic illustration of the main steps of the framework.

Step 1. Development of low-cost data-driven reduced-order stochastic models.

The ensemble forecast adopts a probabilistic characterization of the model state and is thus a natural way to predict complex turbulent systems [117, 136, 85]. However, the high dimensionality and nonlinearity of the coupled atmosphere-ice-

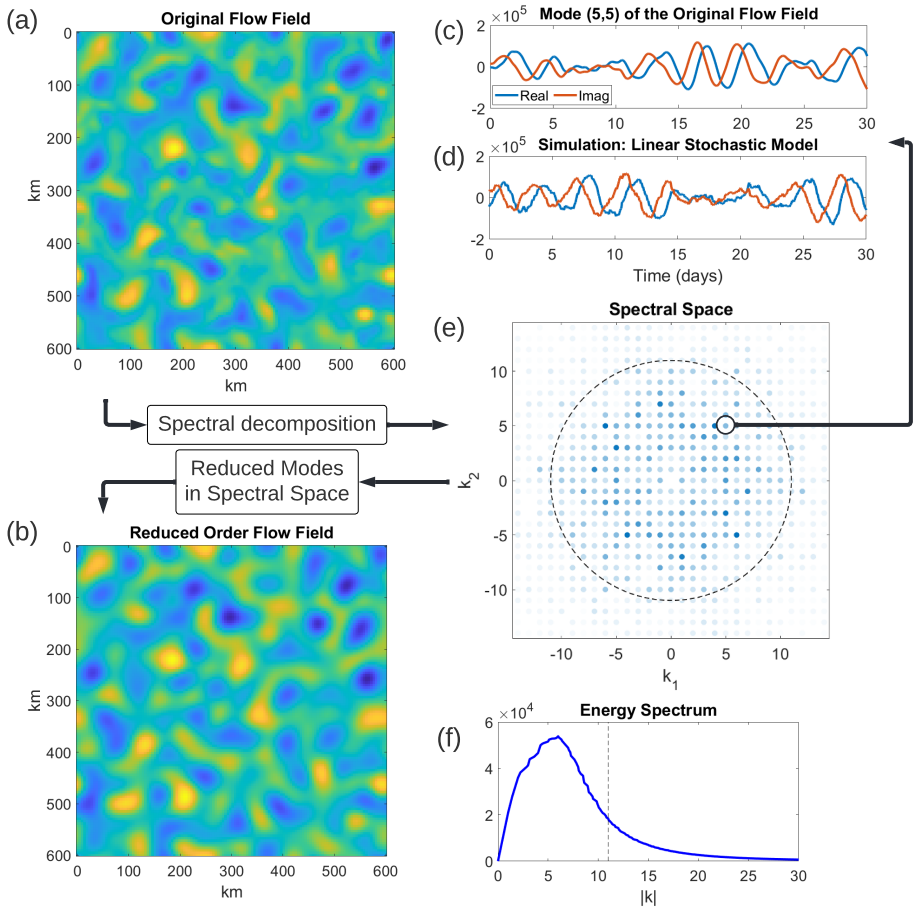

Figure 1.2: Schematic diagram of the new method. Panels (a)–(f): A spectral decomposition is applied to the output of a complicated ocean model. Only a small set of the most energetic spectral modes are retained. The governing equations of these energetic modes are modeled by the low-cost linear stochastic models, thereby significantly reducing the computational cost. The illustration also compares the true ocean flow field and its reconstructed state. The original field is generated from the two-layer quasi-geostrophic (QG) model, while the reconstructed one only uses modes for which $|\mathbf{k}| \leqslant 11$. The top right corner compares the time series of the mode $\mathbf{k} = (5, 5)$ associated with the QG model and a random realization from the calibrated linear stochastic model.

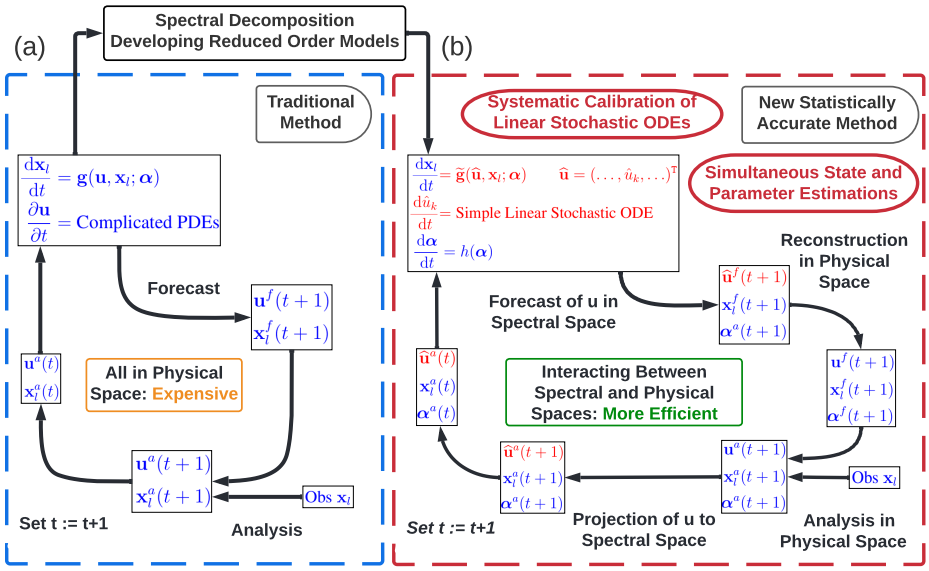

Figure 1.3: Schematic diagram of the new method. Panels (a) and (b): The dynamical interpolation is performed via nonlinear data assimilation. The traditional method requires running the original system in the physical space and is extremely expensive. Here, $\mathbf{x}$, $\mathbf{u}$ and $\boldsymbol{\alpha}$ denote the ice floes, the ocean and the atmospheric state variables, and the model parameters, respectively. In contrast, the new and efficient method for dynamical interpolation alternates between physical and spectral spaces using the reduced-order stochastic models. It has the main benefit of allowing for the simultaneous estimation of state variables and key physical parameters.

ocean system makes a single realization of the model forecast very computationally expensive, let alone the forecast of the entire ensemble. Therefore, the first step in this framework is to develop data-driven reduced-order models with the aim to significantly lower the computational cost of the forecast step.

Figure 1.2 outlines the development of such reduced-order models for the turbulent ocean field. Given a long simulation generated from the original two-layer QG ocean model, the spectral decomposition of the velocity field is used. Most of the energetic modes are concentrated within a circular area centered at $\mathbf{k} = (0, 0)$

with a relatively small radius in spectral space $|\mathbf{k}| \leqslant K$, where $\mathbf{k} = (k_1, k_2)$ is the spectral index. The reduced-order model is set up to only describe the temporal evolution of the dynamics of this small set of spectral modes. It is expected to retain most of the key features of the original ocean field but significantly lower the computational cost. Yet, given the nonlinearity of the original ocean model, the governing equation of each spectral mode is fully coupled with all other modes, including those omitted in the reduced-order model. To effectively characterize the temporal evolution of each spectral mode in the reduced-order model, a linear stochastic model is developed as a surrogate [57],

$$\frac{du}{dt} = (-a + i\omega)u + f + \sigma\dot{W}, \tag{1.1}$$

where u is a complex variable for a single spectral mode, a and ω are the damping and oscillation frequencies, respectively, f is the forcing of the system, $\dot{W}$ is a complex-valued white noise, and σ is the amplitude of the noise. In (1.1), damping and stochastic noise are adopted to parameterize the contribution of the extremely complicated, nonlinear, and deterministic part of the original governing equation, which leads to a cheaper and more effective way to reproduce the statistical forecast results [97, 52, 16, 19, 99, 121]. Independent linear stochastic models are used to characterize the temporal evolution of each mode in spectral space, which nevertheless allows a fully correlated spatial pattern in physical space. Since the QG model generates an incompressible flow field, the spectral representation of the ocean is based on the stream function. Similarly, a pair of linear stochastic models is utilized to approximate the two-dimensional velocity components associated

with each spectral mode of the atmospheric wind field.

Note that despite the independence between the linear stochastic models for different Fourier modes, the strong correlation still exists between the state variables in physical space after the spatial reconstruction via the inverse Fourier transform. The simple structure of the linear stochastic model allows for systematic model calibration and large computational savings, making an ensemble forecast feasible. It is worthwhile to highlight that the most important factor impacting the performance of the dynamical interpolation is the leading-order statistics of the ensemble forecast, rather than the dynamics of individual model trajectories. In this way, because the independent stochastic models are accurate with respect to the statistics of the QG model, the forecast produces good results.

Step 2. Systematic model calibration.

The linear stochastic model in (1.1) can be calibrated systematically by taking advantage of the analytic formulae for its four fundamental statistics: the mean, the variance, and the real and the imaginary parts of the decorrelation time. The values of these four statistics have a unique one-to-one correspondence with the four parameters α, ω, f, and σ. Therefore, once the values of these statistics are computed numerically from the time series of a single spectral mode in the original QG ocean model, these values are plugged into the closed analytic formulae, determining the four parameters in the linear stochastic model associated with that spectral mode.

Rather than simulating the 30,000 modes of the two-layer QG model, setting K to be 11 in step 1 of the framework results in a reduced-order model containing only about 400 modes. This simulation still resembles the full QG system while

being much more computationally inexpensive, roughly 30 times faster. See panels (a) and (b) of Figure 1.2. In addition, after applying the proposed calibration procedure, a random realization of the time series from the linear stochastic model is also statistically similar to the truth by capturing the mean, variance and the decorrelation time. See panels (c) and (d) of Figure 1.2 for a comparison of mode $(5, 5)$. This similarity is essential for an accurate ensemble forecast using the linear stochastic reduced-order models. Finally, the same linear stochastic models are adopted as surrogate models to describe the atmospheric wind field based on the ERA5 reanalysis data.

Step 3. Physics-based dynamical interpolation via nonlinear data assimilation.

Dynamical interpolation exploits the optimal combination of the ensemble forecast from the model and the information from the partial observations via nonlinear data assimilation. The incorporation of the underlying dynamics sets dynamical interpolation apart from pure curve fitting methods. The basic dynamical interpolation scheme used here is the ensemble Kalman smoother (EnKS), in which an ensemble of model trajectories represents the estimate of the system state. Each ensemble member contains trajectories for all state variables, including the sea ice, the ocean, and the atmosphere. The EnKS provides point estimates through the ensemble mean and quantifies the uncertainty through the ensemble spread. The resulting distribution is called the posterior distribution, which contrasts with the prior distribution solely obtained from the forecast step of the model.

The traditional EnKS contains a straightforward prediction-correction loop in physical space that requires repeatedly integrating the expensive original dynami-

cal model. In contrast, the new method here uses the linear stochastic models to approximate the ocean and atmospheric flow fields, and the prediction-correction procedure alternates between the physical and the spectral spaces. Specifically, the prediction of the ocean and atmospheric flow fields, which involves running the linear stochastic models forward, is implemented in the spectral space. On the other hand, the correction of all the state variables, which applies the Bayesian formula that optimally combines the model and observational information, is carried out in the physical space. Spectral decomposition and flow field reconstruction are adopted after each correction and prediction step, respectively. See Figure 1.3. Since only a few spectral modes are involved in the sequential prediction-correction procedure, the computational efficiency is preserved. Note that the DEM sea ice model remains highly nonlinear, which makes the entire dynamical interpolation nonlinear. To further improve the numerical stability and mitigate erroneous spurious long-term correlations, localization and fixed lag strategies are incorporated into the basic version of the EnKS [6, 49].

Each set of floe observations are processed sequentially in time. The algorithm represents the model state with an ensemble of model trajectories. During the prediction correction loop, the ensemble is forecast forward in time up to the next available observation. The sea ice variables are forecast in physical space according to the sea ice model. The ocean and atmosphere variables are forecast in spectral space according to the statistically-accurate stochastic forecast models. After the forecast of the ensemble, all variables are transformed back to physical space where they are compared to the observations in the analysis step. The ensemble is updated,

in a Bayesian sense, according to the new observations. Finally the ocean and atmosphere variables are transformed back to spectral space for the next iteration of the loop.

Step 4. Efficient parameter estimation of important physical quantities.

The main practical challenge of using general dynamical interpolation methods to analyze the sea ice cover is the lack of access to the entire parameter space from a single remote sensing instrument. For example, the thickness of the floes determines the inertia of floe motion and is crucial to the coupled system. To overcome this challenge, an efficient parameter estimation algorithm is embedded into the dynamical interpolation framework. Here, the unobserved physical quantities are treated as the augmented state variables, which are simultaneously estimated with the actual variables of the model state. The uncertainty in the estimated parameters, due to the relatively short Lagrangian trajectories, is also quantified in the algorithm.

1.3 Results of Interpolating the Floe Trajectories and Angular Displacements

Setups of the two experiments

The new dynamical interpolation framework is first applied to a synthetic data experiment and then to the real observation scenario.

The synthetic data experiment uses the two-layer QG ocean model and the

reanalysis data for the atmospheric winds to force the ice floes governed by the DEM model. The ice floe shapes, sizes, positions, and orientations are initialized from a library of floes in the Beaufort Gyre MIZ [88], which is generated from optical remote sensing imagery using the Ice Floe Tracker algorithm [87]. The thickness of each floe is randomly drawn from a background distribution [79] and is assumed to be constant during the entire observational period. See panel (c) of Figure 1.5. Note that the stochastic approximate models are not utilized to generate the synthetic data, rather they are only used to dynamically interpolate the missing floe observations. For the real data experiment, Lagrangian sea ice floe trajectories are obtained using the Ice Floe Tracker algorithm within the study area delineated in Figure 1.1 during the spring-to-summer transition of 2008. The same linear stochastic models that are calibrated for the synthetic data experiment are adopted to carry out the dynamical interpolation. See Table 1.1 for the summary of the models used to generate the true signal and those adopted to implement the dynamical interpolation in the two experiments.

The floe locations and angular displacements are the only observational information in the dynamical interpolation for the coupled atmosphere-ice-ocean system. These two quantities are obtained from the satellite images at a frequency of roughly every 24 hours. The observational uncertainty, which is used in the dynamical interpolation algorithm, is set to be 0.25 km and 5 degrees, respectively.

Both experiments contain 38 floe trajectories of various lengths in time. See the bottom panel of Figure 1.4. Excluding the first and the last point in each floe trajectory, there are in total 164 remaining candidate observational points for the

(a) Synthetic data experiment			
	Atmosphere	**Ocean**	**Sea Ice**
Truth	ERA5 reanalysis	Two-layer QG	The known DEM model
Interpolation	Calibrated LSM	Calibrated LSM	The known DEM model
(b) Real data experiment			
	Atmosphere	**Ocean**	**Sea Ice**
Truth	Not needed	Not needed	Satellite observations
Interpolation	Calibrated LSM	Calibrated LSM	The known DEM model

Table 1.1: Summary of the models used for both the synthetic and the real data experiments. In each experiment, the first row "truth" stands for the underlying systems that generate the true signal, while the second row "interpolation" indicates the model used for dynamical interpolation. The same calibrated linear stochastic model (LSM) is utilized for the real data as for the synthetic data experiments. In the real data case, the true signals of the atmosphere and ocean components are not needed. In the synthetic data case, the true atmosphere and ocean models are used to drive the DEM model to generate the observed floe trajectories and angular displacements.

38 trajectories. These 164 candidates are randomly divided into four sets, where each set contains 41 data points. Then four independent dynamical interpolation simulations are carried out. In each simulation, the 41 candidate observations in the corresponding set are artificially removed as the missing observations. Note that the missing observations referenced in the real data experiment are not the actual missing ones in the satellite images obscured by clouds, but are rather the artificially removed ones. Such a setup guarantees the true values of these missing floes are known and therefore it allows the qualitative study of the accuracy of the dynamical interpolation. Nevertheless, this $3:1$ ratio between the number of observed and missing floe observations mimics the real-world situation in the MIZ during the boreal summer. The number of ensemble members used here is 600.

Figure 1.4 displays the 38 sea ice floe trajectories in the real data experiment, which are retrieved from satellite remote sensing imagery using the Ice Floe Tracker algorithm during the spring-to-summer transition of 2008. Each floe trajectory is represented by the transition from fully transparent to opaque with each floe assigned a specific color. The index in these floes corresponds to those in Figures 1.6 and 1.8.

Results of the synthetic data experiment

Figure 1.5 illustrates the parameter estimation of the floe thicknesses from the dynamical interpolation. To quantify the uncertainty in the estimated thickness of each floe, the posterior distribution characterized by the ensemble members is also included via a violin plot. The results shown here are calculated using all the ensemble members from the four simulations, but different simulations lead to similar distributions for all the floes. The estimations are overall reasonably accurate, especially given such a small number of observations within a large domain. Particularly, the truth of all the 38 floes is consistently covered by the posterior distribution. In addition, for two-thirds of the floes, the true thickness value lies in a high likelihood region of the distribution within one standard deviation from the mean. Note that the error in the thickness estimation can offset the error from recovering the atmosphere and ocean flow fields, and therefore the overall interpolation results remain accurate, as will be seen below. It is worthwhile to highlight that both the background thickness distribution, from which the true thickness values are drawn, and the estimated posterior distributions, exhibit strongly fat-tailed

non-Gaussian behavior, as is clear in the violin plot. These non-Gaussian features are the unique outcome of the highly nonlinear dynamics of the sea ice floes. In addition, because the ensemble is transformed from the prior distribution to the posterior during each update of the nonlinear EnKS, rather than resampled, these important non-Gaussian features can be preserved through the ensemble update (see Appendix.) By contrast, simply considering the ensemble mean and standard deviation would underestimate the likelihood of large floe thicknesses while simultaneously overestimating the likelihood of small thicknesses. Notice that such a non-Gaussian feature is found in all model variables, but is especially illustrated by thickness estimation. These findings imply the necessity of incorporating both the nonlinear sea ice dynamics and the nonlinear data assimilation scheme into the dynamical interpolation framework.

Panel (a) in Figure 1.6 compares linear and dynamical interpolation for recovering the floe location and angular displacement. The ensemble mean estimate using the dynamical interpolation almost always outperforms linear interpolation in recovering the floe locations. Specifically, the absolute error using the linear interpolation is nearly three times as large as that using the dynamical interpolation. The linear interpolation also, by design, completely fails to recover the curvature and the nonlinear evolution of the floe trajectories. In contrast, the dynamical interpolation accurately captures these important physical features. In addition to the point estimate using the ensemble average, the ensemble provides the quantification of the estimated uncertainty. The uncertainty overall remains at a relatively low level, indicating the confidence of the posterior mean estimate. Among all the 164

recovered missing observations, roughly 80% of the true observations fall within two standard deviations around the ensemble mean estimate. This implies the accuracy and robustness of the dynamical interpolation. In addition, the recovery of the angular displacement using the dynamical interpolation is quite accurate.

Panel (a) of Figure 1.7 illustrates the recovered ocean field represented by the stream functions utilizing the dynamical interpolation. The result shown here is on a specific day in the middle of the entire time period. The accuracy in recovering the ocean field remains in a similar level on other days. The overall pattern correlation between the truth and the recovered ocean field is around 0.3. Nevertheless, given the fact that there are only 17 floes inside this large domain on this day, the skill of recovering the ocean field is already significant. In particular, the ocean eddies are recovered quite reasonably in the areas, where the observed floes are concentrated. The pattern correlation is above 0.3 and the amplitude of the recovered eddies is similar to the truth. This is partially due to the use of the localization technique in the dynamical interpolation, where each observation affects more towards the skill of the recovered ocean in the nearby regions. The appendix includes more sensitivity analysis, which shows the improvement of the recovered ocean field if the density of observed floes increases.

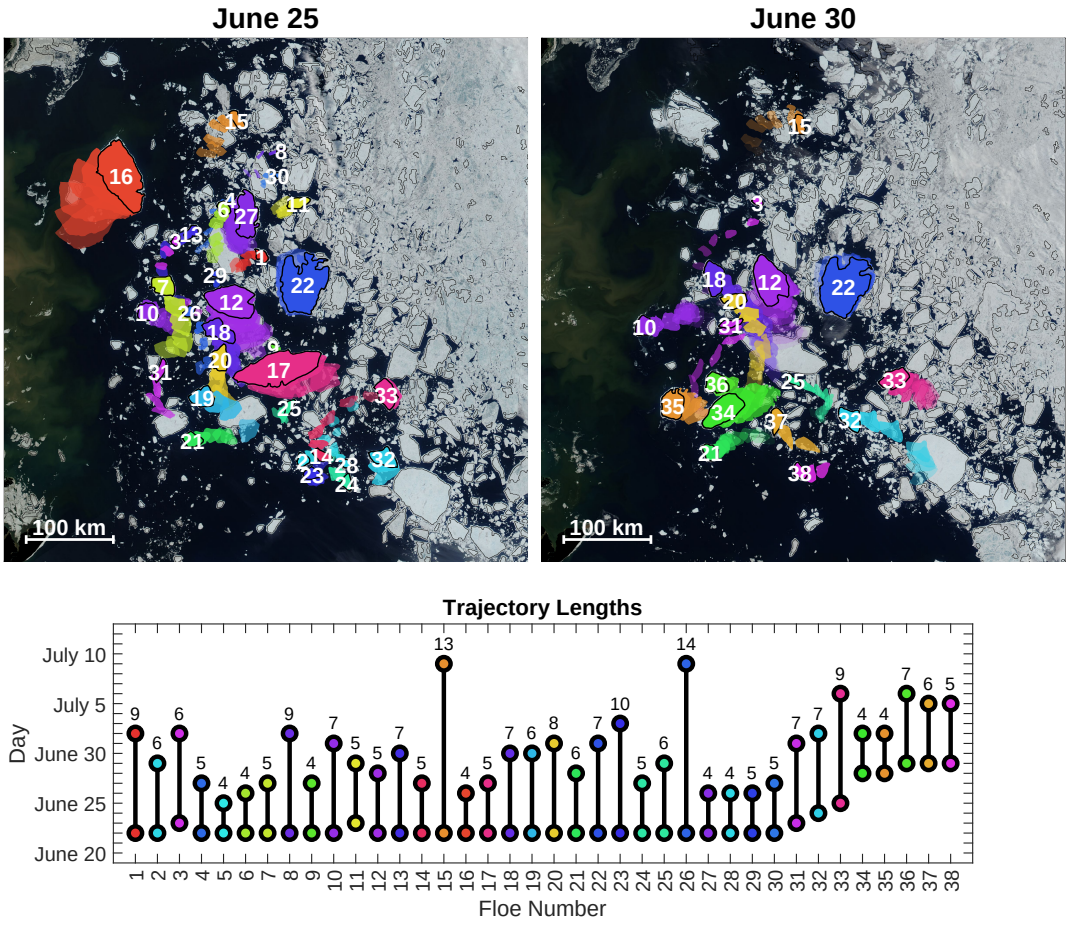

Figure 1.4: Sea ice floe trajectories retrieved from optical satellite remote sensing imagery. MODIS True Color images (downloaded from the NASA Worldview application) acquired on 25.06.2008 and 30.06.2008 are displayed in a WGS 84/NSDIC Sea Ice Polar Stereographic North 70″ N Projection, on top of which retrieved ice floe trajectories are displayed in color. In both images, the evolution of floe positions is represented as a shift in opacity from transparent to opaque objects. Final floe positions are marked using black contour lines. Note that the recovered floe trajectories have different lengths and periods. Information regarding the acquisition period of each floe trajectory is shown in the bar plot underneath. Only a sub-set of the 38 non-interacting floes used in this study are shown for clarity.

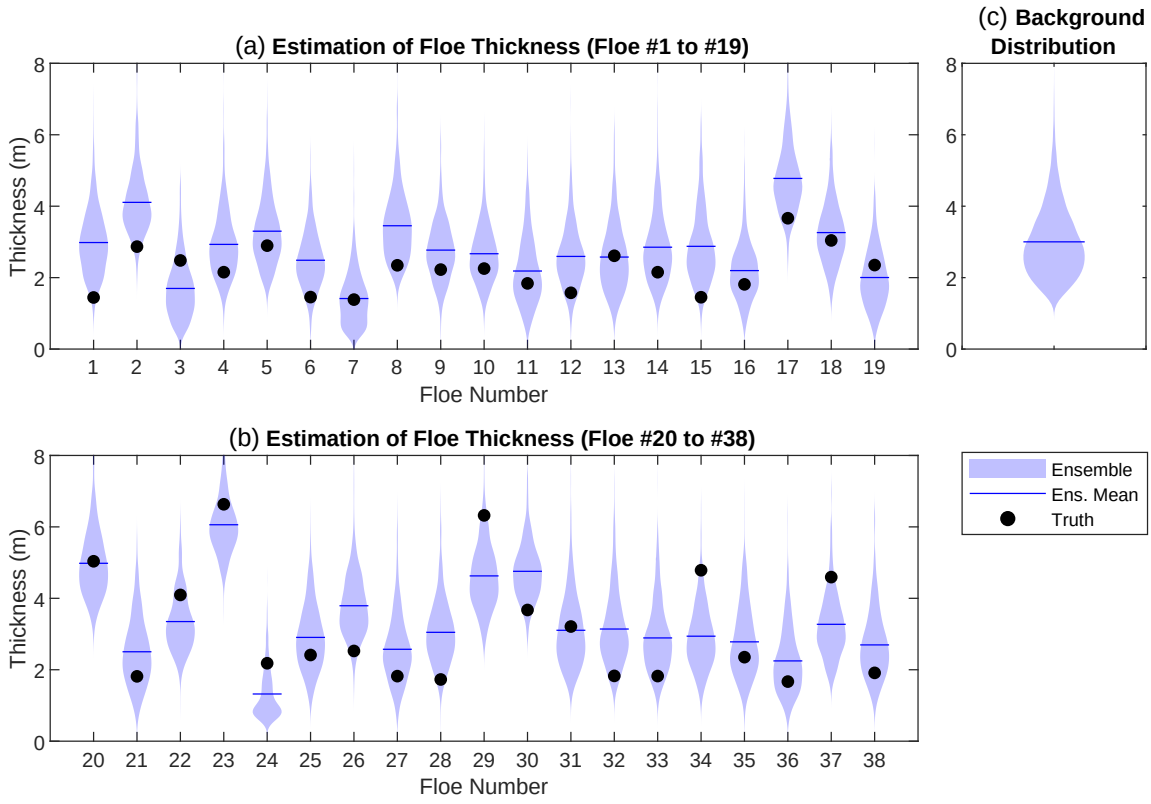

Figure 1.5: Parameter estimation of sea ice thickness in the synthetic data experiment. Panels (a) and (b): The black dots and solid lines indicate the truth and the ensemble mean estimate of each sea ice floe, respectively, while the shaded area in the violin plot indicates the estimated non-Gaussian PDF formed by ensembles. Panel (c): The background sea ice thickness distribution. The true value of the thickness for each sea ice floe is randomly drawn from such a distribution. It is also used as the initial distribution in the parameter estimation algorithm.

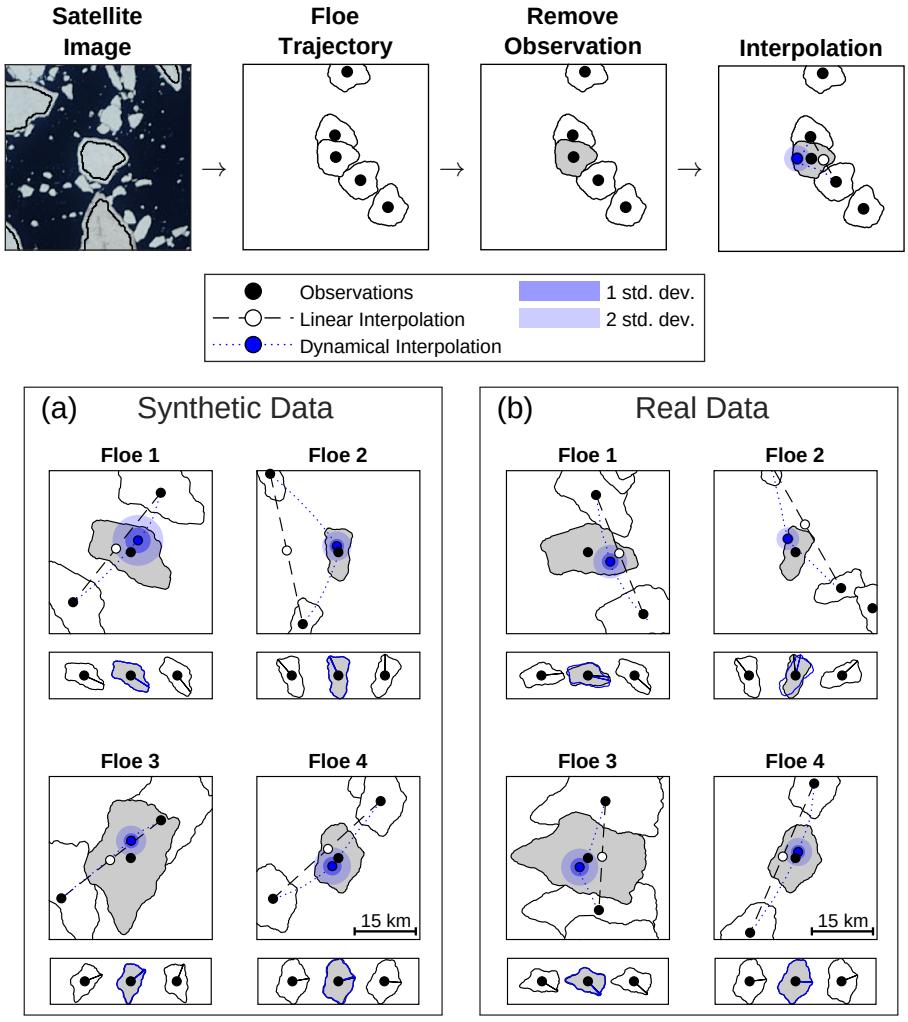

Figure 1.6: Comparison of recovering the missing observations using linear and dynamical interpolation schemes. The top panel illustrates the procedure of performing the interpolation experiments. A floe trajectory is first retrieved from the satellite imagery. Next, the observed floe on a specific day is artificially removed. The linear/dynamical interpolation framework is applied to recover this artificially removed observation. In the bottom part, panels (a) and (b) show the results from the synthetic and the real data experiments, respectively. In each panel, the top part shows the interpolated floe locations while the bottom part shows the interpolated angular displacement. Since the floes in the synthetic data experiment are taken from the library of sea ice floe observations, floes with the same index in the two experiments are identical (i.e., shapes and sizes are retained). In addition to the ensemble mean estimate presented by the blue marker, the uncertainty resulting from the dynamical interpolation is provided by the shaded areas. For the illustration purpose, only the two-dimensional Gaussian confidence interval is used to characterize the uncertainty in the dynamical interpolation.

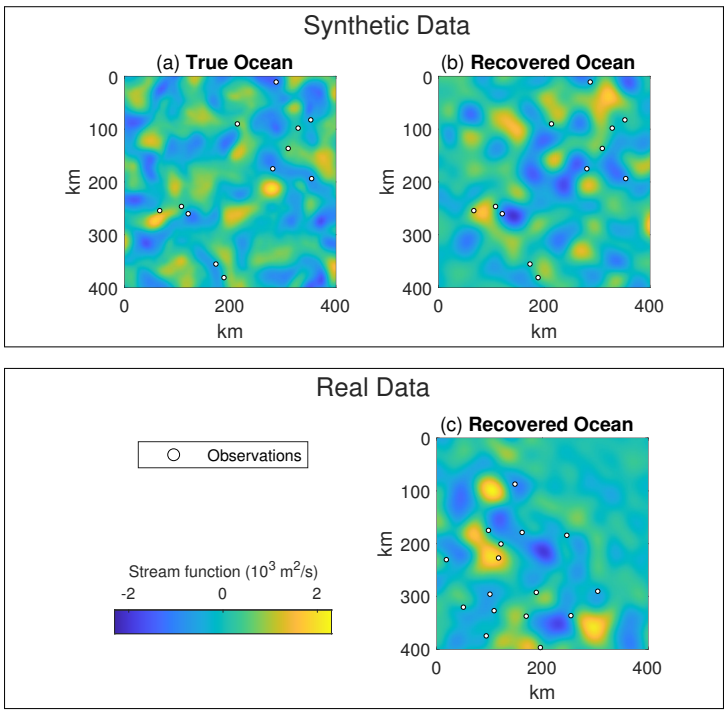

Figure 1.7: The recovered ocean flow field represented by the stream functions utilizing the dynamical interpolation. The top panel shows the truth and the recovered ocean field in the synthetic data experiment while the bottom panel shows the recovered ocean field of the real data. Since the primary focus is to resolve regions close to the ice edge, a 400km$\times$400km domain within the original 600km$\times$600km area and having the same domain center is presented. The pattern correlation between the true and recovered ocean in the 400km$\times$400km subdomain is 0.32. The results shown here are on a specific day in the middle of the study period. For the real data, it is July 1. The error in recovering the ocean field remains in a similar level on other days. The white dots mark the locations of the floes. There are in total 17 floes inside the 400km$\times$400km domain for both cases.

Results of the real data experiment

Panel (b) of Figure 1.6 includes four cases of the recovered missing floes on the real data set. Similar to the conclusion from the synthetic data experiment, the dynamical interpolation being applied to the real data set also shows significant advantages over the linear interpolation in the sense that the error in the ensemble mean is overall much smaller and the uncertainty can be systematically quantified. Comparing with the analogs from the synthetic data experiment in Panel (a), the accuracy of the results in the real data test remains comparable. Figure 1.8 includes additional case studies of the recovered missing floe trajectories from the real data experiment. Again, the dynamical interpolation provides reasonable results in most of the cases. Panel (b) of Figure 1.7 displays the recovered ocean field on July 1, 2008. Although there is no true solution for the validation of the point-wise recovery skill, the overall flow amplitudes as well as the number and the size of the eddies in the recovered ocean field all look reasonable. One interesting finding is that the recovered ocean field in the north-east corner of the domain is nearly zero due to high uncertainty, which corresponds to the area beneath the large piece of the ice cover shown in Figure 1.4.

Figure 1.9 compares the physical properties of the recovered ice floes between the real observations, the dynamically interpolated data, and the direct model simulation. Since the data set consists of discrete observations, the two metrics used are the discrete curvature and the daily angular displacement. The former is calculated using the circumscribing circle of each trio of observations while the latter is obtained by taking the difference in angle between two consecutive

observations. The results using the linear interpolation are omitted here, as the linear interpolation fails to provide any useful information of these two physical quantities. Panel (a) shows that the curvature of the floe trajectories from the direct model simulation is severely underestimated, which is a natural outcome of the model error. In contrast, the data resulting from the dynamical interpolation succeeds in reproducing the non-Gaussian distribution of the observed truth with a one-sided fat tail. Next, with respect to the angular displacement, as is shown in Panel (b), the real data set has a negative bias due to the influence of the Beaufort Gyre, something which is not reflected in the direct model simulation that is again due to the model error. Nevertheless, such a bias in the direct model simulation is almost fully corrected in the dynamically interpolated data with the help of the partial observations. These results indicate the importance of utilizing both the observations and a suitable model in the dynamical interpolation, as the model provides at least partially the access to the crucial underlying nonlinear dynamical information while the observations can largely reduce the biases from the model forecast.

1.4 Conclusions and Discussion

Model error is inevitable when studying complex systems. In the dynamical interpolation framework developed here, the sea ice DEM model remains highly nonlinear while the ocean and atmospheric components are effectively approximated by linear stochastic models. Indeed, a large error will appear if these linear

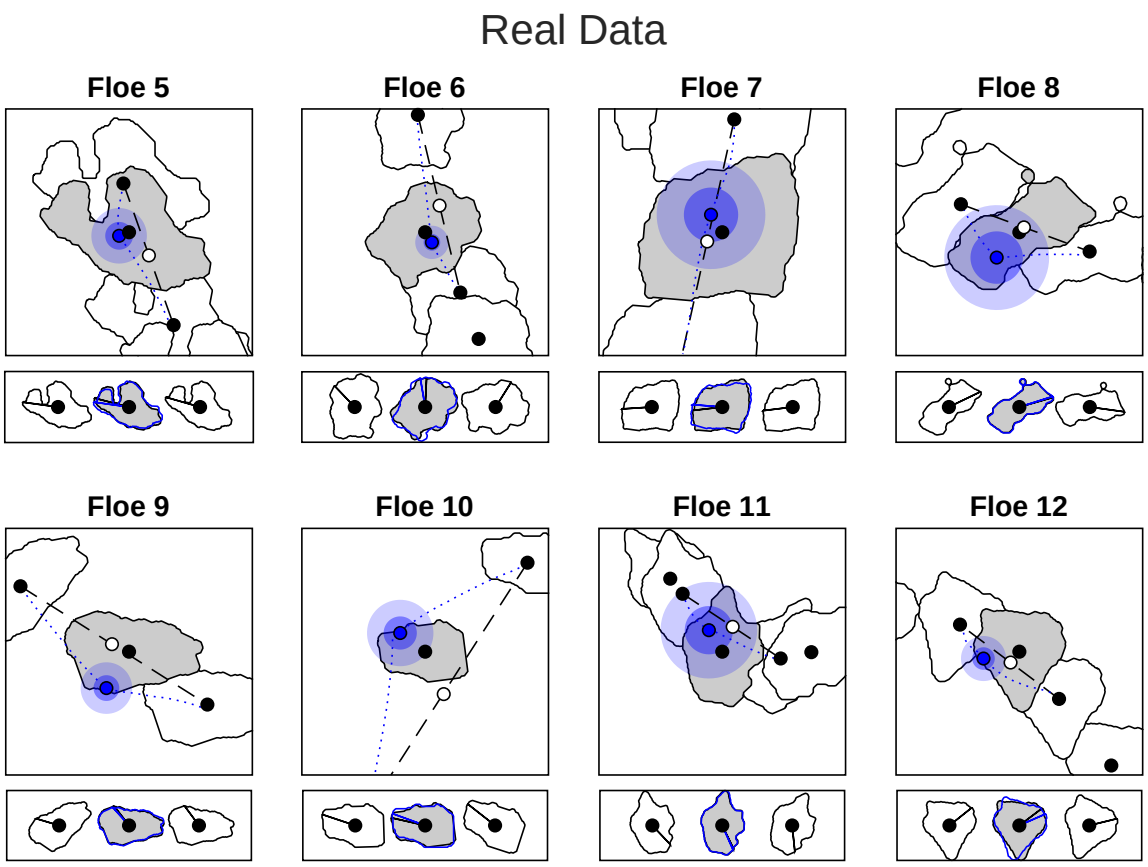

Figure 1.8: Additional results for the real data experiment, similar to those in Panel (b) of Figure 1.6.

stochastic models are used to study the dynamics associated with the ocean and atmospheric fields. Nevertheless, for the purpose of dynamical interpolation, the information needed from the model is merely some prior knowledge of the short-range statistical forecast of these fields, which are usually quite accurate due to the fact that these linear stochastic models are carefully calibrated.

The reduced-order models in the proposed framework are not limited to linear stochastic models. If the time series of the underlying flow fields exhibit strong

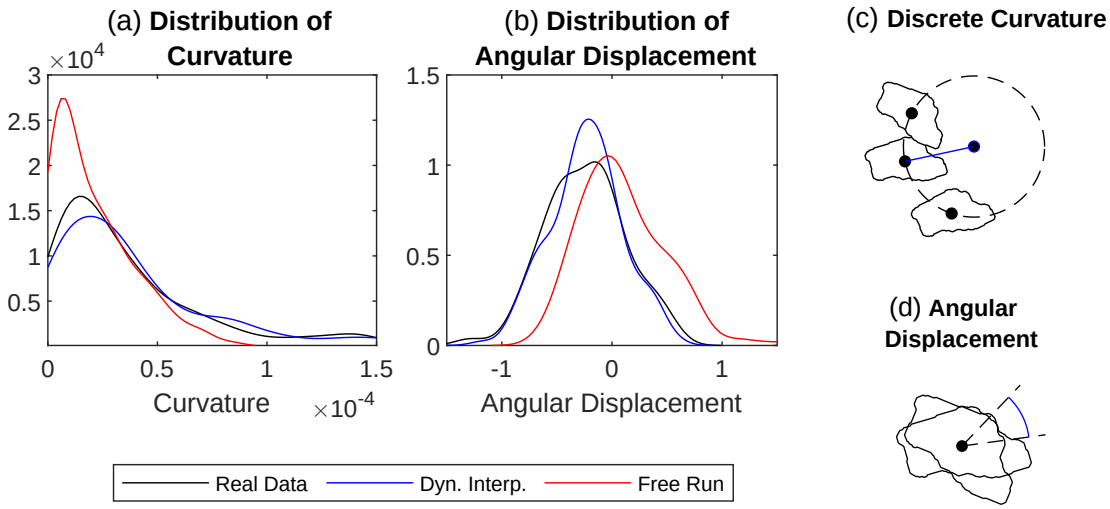

Figure 1.9: Comparison of the recovered properties using different interpolation methods. Panel (a): comparison of the distribution of the curvature of the recovered trajectories. Panel (b): comparison of the distribution of the angular displacement of the recovered trajectories. Panels (c)–(d): Schematic illustrations of the definitions of the discrete curvature and angular displacement used to compute the distributions in Panels (a)–(b).

non-Gaussian features, then suitable nonlinear or non-Gaussian surrogate models can be easily incorporated [32, 73, 58, 48, 103]. In particular, one such simple candidate is a family of linear models with multiplicative noise [11]. Another potential alternative is the multilayer stochastic models [76, 77]. On the other hand, the sea ice dynamics within the scales studied here are predominantly nonlinear. The governing equations are well understood and are crucial in the dynamical interpolation, given that the directly observed variables are sea ice floe trajectories. The strong nonlinearity in sea ice dynamics is also more deterministic and less turbulent than the atmosphere or the ocean. Therefore, linear stochastic models are

not appropriate for approximating the fully nonlinear behavior of sea ice. Since the degree of freedom in characterizing the floe trajectories is much lower than the governing equations of the atmospheric and oceanic velocity fields, the nonlinear floe dynamics are explicitly incorporated into the dynamical interpolation framework.

It is also worth highlighting the importance of the prior physical model of the ocean and the prior time series of the atmosphere, which significantly facilitate the calibration of the linear stochastic models. In the absence of a suitable prior model or data for the ocean and the atmosphere, the calibration of the reduced-order stochastic surrogate models requires a more complicated iterative expectation-maximization procedure [27]. In other words, the dynamical interpolation, the parameter estimation of the thickness, and the uncertainty quantification of the oceanic and atmospheric flow fields have to be carried out simultaneously with the floe trajectories providing the only available information. Such an iterative approach often requires an extensive observational database to ensure the accuracy of the dynamical interpolation scheme.

Another crucial point is the quantification of the uncertainty, particularly in the presence of the model error, the small number of observations, and the implications of turbulent systems. Uncertainty quantification is not available by applying direct curve fitting methods but rather a unique feature of the dynamical interpolation framework. In this study, the properties of sea ice exhibit various non-Gaussian features such as the non-symmetry in the distribution of the angular displacement, strong skewness, and fat tails with extreme events in the distributions of the curvature and the thickness. These non-Gaussian features have been shown to be crucial

in understanding the sea ice dynamics [135, 110]. Hence, to assess uncertainty, the attributes of the entire distribution are considered.

Finally, the framework developed here has several unique implications for improving our understating of Earth system science. First, new-generation climate models that accurately represent sea ice dynamics at the floe scale will require validation against Lagrangian observations of sea ice floes at high and moderate resolutions. The point-estimate recovery of missing observations and the associated estimates of the uncertainty mitigate some of the issues of Lagrangian optical remote sensing observations. The methodology presented here is also easily adaptable to analyze the output of other instruments. In this sense, it is expected that the continuous trajectories stemming from the nonlinear data assimilation can be used in more sophisticated deep-learning models for calibration and training. Second, the proposed framework allows for accurate parameter estimation of unobserved variables at unprecedented scales in a Lagrangian setting. For example, the retrieval of sea ice thickness is outlined here, which is a crucial variable in understanding the evolution of the sea ice cover in response to a changing climate. Lastly, data assimilation provides the key missing piece for understanding ocean transport and mixing processes at high latitudes. Small-scale eddies have an important role in transferring energy to larger-scale structures via an inverse cascade of energy and are thus hypothesized to be the missing energy source to close the ocean energy budget. High-resolution numerical simulations have highlighted their contribution to nutrient redistribution, oxygen transport, and biogeochemical processes. However, they are hard to observe due to the lack of resolution of space-borne

sensors and the sparsity of in situ instruments. Given the recently demonstrated connections between the rotation rate of sea ice floes and eddies with sub-surface expression in the western Arctic Ocean, it is anticipated that this method can be used to understand fundamental processes of ocean turbulence at small-to-moderate scales.

2 EFFECTIVE STATISTICAL CONTROL STRATEGIES FOR COMPLEX TURBULENT DYNAMICAL SYSTEMS

The following chapter is adapted from an open-access paper [39] published in *Proceedings of the Royal Society A* under a creative commons license. The paper is authored by Jeffrey Covington, Di Qi, and Nan Chen. The research was conducted by all the authors. Jeffrey Covington was responsible for writing the original draft of the publication with editing provided by Di Qi and Nan Chen.

The research of Nan Chen is funded by ONR (grant no. N00014-21-1-2904) and ARO (grant no. W911NF-23-1-0118). Jeffrey Covintton is partially supported by ONR (grant no. N00014-21-1-2904) as a graduate research assistant.

2.1 Introduction

Complex turbulent dynamical systems emerge throughout fields in science and technology, including in geophysics, engineering, neural science, and plasma physics, to name a few [94, 139, 97, 112, 28, 130, 146, 126, 60, 32]. These turbulent systems are characterized by high-dimensional state spaces with substantial nonlinear energy transfers between scales [120, 97, 4]. The control of such turbulent dynamical systems has grand importance and broad applications. The goal of control is to design an optimal course of action (the control) to drive the perturbed state of interest back to a desired final target state under minimum cost within a finite time window. For example, active and passive control strategies can reduce the aerodynamic

drag of vehicles such as aircraft and cargo ships [131, 81, 15, 119, 3] and increase the efficiency of liquid and gas transport through pipelines [10, 122, 74]. Various control strategies have also been designed for applications in industrial mixing and manufacturing [20, 1, 118, 78]. In addition, control of turbulent systems has significant implications for climate change mitigation involving large-scale models with high uncertainties [92, 91, 70, 46]. However, controlling turbulent systems, in general, has proven a formidable challenge. The central obstacles involve the development of efficient algorithms to deal with the genuinely high dimensionality and a large number of unstable modes. Linear control theory is a well-developed field [44, 18, 127] and linear control strategies have been applied to turbulent systems under a variety of circumstances where chaotic systems can be linearized about a fixed-point and stabilized by the control [116, 67]. For example, linear control with closed-loop feedback from the system has successfully been used to delay the transition from laminar to turbulent flow [12, 50, 13]. However, linear control techniques do not scale well computationally when the dimensionality of the system becomes large [75, 45]. Thus low-dimensional reduced-order approximations are frequently employed through model reduction and system identification [147, 125, 23, 72, 55]. Besides linear control, there have been many other innovative approaches to the control of turbulent systems. Machine learning, for example, has been used to design control laws for turbulent systems using data [45, 24, 22]. In addition, there have been open-loop approaches where a predetermined control is applied [122, 140].

Statistical control offers a fundamentally different approach compared with the

traditional trajectory control methods. Statistical control aims to control certain statistical features of the underlying system. These features can be considered as each statistical moment of the critical model state and are estimated by averaging an ensemble of model states. Note that even deterministic systems can fall under the statistical control framework due to the uncertainties in the initial conditions and observations. These initial uncertainties are propagated and amplified in time as a response to the strong instability considering the turbulent nature of the system. Statistical control has several unique advantages. First, there is no need to exploit exhausting procedures to resolve the full solution of each high-dimensional chaotic trajectory of the underlying turbulent dynamics. The control strategy is achieved effectively by considering only the contributions of the leading order moments. This significantly reduces the computational cost and avoids the randomness of individual trajectories in affecting the control results. Second, although individual trajectories are turbulent, the time evolution of the statistics is deterministic thus is usually easier to control in practice. In addition, for energy-conserving systems, the statistical energy combining the mean and total variance of the system is always stable towards its statistical equilibrium. This can be seen by noticing that the Fokker-Planck equation, which is the time evolution of the probability density function (PDF) of the state variables, is always linear despite the associated underlying dynamical system being highly nonlinear and turbulent. Using conservation of energy, the statistical energy, which bounds the mean and variance of the system, exponentially decays to the equilibrium state of the system. [96]. Therefore, the statistics are more controllable. Third, statistical control can naturally account for

uncertainties and incorporate stochastic reduced-order models [103, 104, 124]. It allows a large degree of freedom to design suitable strategies for efficient controls.

Statistical energy is a measure of the total statistical mean and variance of the system state across all scales. It is a natural scalar quantity to consider in the context of controlling turbulence [53, 96, 123]. Previous works [102, 105, 104] have demonstrated that statistical responses in the key states can be successfully controlled using the statistical energy by making use of an energy-conservation principle that appears in numerous turbulent dynamical systems [96, 97, 63]. In particular, exploiting symmetry in the total statistical energy dynamics avoids the inherent nonlinear structure containing instability. Consequently, this approach eliminates the need to track and control a large number of unstable modes. In its current formulation, this statistical control strategy is run in an open-loop manner without requiring online feedback from the system, allowing for the prescribed control forcing to be determined offline for efficient computation. In addition, using the scalar-valued total statistical energy as the control object circumvents the computational issues raised by high-dimensional systems. These factors point to promising applications of statistical turbulent control considering different dynamical features of the targeting turbulent systems.

The statistical control strategy aims to control the statistical energy from a perturbed state back to the target equilibrium state by exerting an external control forcing in the underlying turbulent system. From a high-level description, the strategy consists of two consecutive steps: calculating the optimal energy control and the inversion of a nonlocal control-forcing relation. In the first step, the explicit

dynamics of statistical energy are derived using the aforementioned statistical energy-conservation principle. The original control of the high-dimensional system gets reduced to a linear control problem for the scalar energy, which can then be solved using the Hamilton-Jacobi-Bellman (HJB) equations directly [5, 14]. In the second step, a nonlocal inversion problem is solved to find the deterministic external control forcing in the underlying system, which yields the optimal control of the original turbulent system. This nonlocal inversion problem uses the coupling of the optimal energy control found in the first step with both the deterministic external forcing and the response of the statistical mean of the system to the external forcing. Direct simulation of the response of the mean would be prohibitively expensive, so a crude first-order linear response approximation was employed in previous works instead [102, 104, 95, 100]. Notably, the second-order feedback term in the full control-forcing relation was truncated. This simplifies the analysis and remains consistent with the first-order approximation valid for small amplitude perturbations within the linear regime. This approach effectively controlled the statistical energy to the target equilibrium state from small initial perturbations.

This paper aims to develop new statistical control strategies for scenarios with more significant initial perturbations and stronger nonlinear responses, allowing the statistical control framework to be applied to a much wider range of problems. Note that the second-order term in the control-forcing relation, consisting of the product of the external forcing perturbation and the mean response to the forcing, is significant for large initial perturbations from the equilibrium state and thus cannot be neglected. While this term can be truncated for small perturbations, it

must be included to guarantee proper performance under most large perturbations. In scenarios where the initial mean perturbation is the dominant component of the initial energy perturbation, the inclusion of the second-order term is reflected by the initial external forcing perturbation prescribed by the strategy. Even when the initial mean perturbation is small, the required strong external forcing coupled with dominant nonlinear terms to efficiently control the system usually has a correspondingly strong mean response. In such a case, the external forcing perturbation is strongly influenced by the second-order term in the control-forcing relation.

Two new statistical control methods are developed in this paper to address these difficulties. First, the higher-order methods, incorporating the second-order term, is developed to fully resolve the control-forcing relation given the mean response so that the higher-order responses are considered explicitly. The corresponding changes to recovering the forcing perturbation effectively improve the performance of the statistical control strategy in most test cases. Second, the accuracy of the mean response used in the existing statistical control methods to the external forcing also dramatically impacts the performance of the statistical control strategy. With large perturbations, although linear response theory can provide reasonable results in some special cases [64, 100, 108], the assumptions justifying the use of linear response theory breaks down and the existing methods often lead to significant errors. In particular, the mean linear response is inadequate when the system is perturbed into a different dynamical regime than the equilibrium state. Due to these limitations, a *mean closure model* for the mean response as an alternative to the mean linear response is developed in this work. The mean closure model is based on the

explicit mean dynamics given by the underlying turbulent dynamical system. The dependence of the mean dynamics on higher-order moments is closed using linear response theory but for the response of the second-order moments to the forcing perturbation rather than the mean response directly. Despite still incorporating linear response theory, the introduction of explicit dynamical information from the underlying model allows the mean closure model to better reflect the properties of the perturbed regime compared to the mean linear response, which only contains information from the equilibrium statistics.

The rest of the paper is organized as follows. Section 2.2 reviews the general strategies of energy-conserving turbulent dynamical systems, including the relevant assumptions and properties needed for the statistical control strategy. The statistical energy control problem is formulated in Section 2.3 along with the strategies for recovering the optimal forcing perturbation from the optimal energy control, namely combining the low-order or high-order methods with either the mean linear response or the mean closure model. In Section 2.4, the control strategies are evaluated in detail based on two prototype models. The first model is a prototypical test model that can exhibit various behaviors and dynamical regimes. The second model is the Lorenz '96 model, which is high-dimensional and exhibits multiple dynamical regimes based on the magnitude of the external forcing. Section 2.5 discusses the results and provides guidance and suggestions for when the various strategies should be applied. Lastly, Section 2.6 concludes the paper and offers potential future research directions.

2.2 Background on Statistical Modeling

Statistical formulation of the turbulent systems

Turbulent dynamical systems with quadratic energy-conserving nonlinearity can be represented in the following general canonical form [97, 103, 99] on the state variable $\mathbf{u} \in \mathbb{R}^N$ satisfying the dynamics

$$\frac{d\mathbf{u}}{dt} = (L + D)\mathbf{u} + B(\mathbf{u}, \mathbf{u}) + \mathbf{F}(t) + \boldsymbol{\sigma}(t)\dot{\mathbf{W}}(t). \tag{2.1}$$

Above, the linear component of the operator is decomposed into two matrices: a skew-symmetric matrix representing linear dispersion effects, L, and a negative definite matrix representing dissipation effects, D. The quadratic nonlinearity is given by a bilinear operator, $B(\cdot, \cdot)$, which satisfies the following energy conservation law,

$$\mathbf{u} \cdot B(\mathbf{u}, \mathbf{u}) = 0. \tag{2.2}$$

Here "$\cdot$" denotes the Euclidean inner product. The last two terms of equation (2.1) represent the external forcing of the system, which is composed of the deterministic component of the forcing, $\mathbf{F}(t)$, and the random component, $\boldsymbol{\sigma}(t)\dot{\mathbf{W}}(t)$, where $\dot{\mathbf{W}}$ is Gaussian noise. The family of turbulent dynamical systems which can be represented by the general abstract equation (2.1) is large and diverse, including many important examples from geophysics, neural science, material science, plasma physics, and engineering [94, 139, 97, 112, 28, 130, 146, 126, 60, 32]. The statistical control strategies proposed in this paper can be applied to these practical problems,

for example, the design of effective strategies for the control of climate change systems back to its previous unperturbed equilibrium; and the control of anomalous statistics in the radial transport of plasma flows in tokmak devices that enable fusion.

It is useful to decompose the state $\mathbf{u}$ into a deterministic mean state, $\bar{\mathbf{u}}(t) = \langle \mathbf{u}(t) \rangle$, and the stochastic fluctuations about each mode

$$\mathbf{u} = \bar{\mathbf{u}} + \sum_{k=1}^{N} Z_k(t)\mathbf{e}_k, \tag{2.3}$$

where $\langle \cdot \rangle$ denotes statistical expectation, and $\mathbf{e}_k$ is the predetermined orthonormal basis. The covariance matrix of $\mathbf{u}$ is defined as $R(t) = \langle \mathbf{Z}\mathbf{Z}^* \rangle$ where $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$ and $\cdot^*$ denotes the conjugate transpose. Using the above mean-fluctuation decomposition of $\mathbf{u}$ and equation (2.1) the dynamics of the mean of $\mathbf{u}$ can be explicitly written as

$$\frac{d\bar{\mathbf{u}}}{dt} = (L + D)\bar{\mathbf{u}} + B(\bar{\mathbf{u}}, \bar{\mathbf{u}}) + \sum_{i,j=1}^{N} R_{ij} B(\mathbf{e}_i, \mathbf{e}_j) + \mathbf{F}, \tag{2.4}$$

The operator L_u incorporates the energy transfers between modes from the linear dispersion and dissipation effects

$$\{L_u\}_{ij} = [(L + D)\mathbf{e}_j + B(\bar{\mathbf{u}}, \mathbf{e}_j) + B(\mathbf{e}_j, \bar{\mathbf{u}})] \cdot \mathbf{e}_i. \tag{2.5}$$

Importantly, note that the mean dynamics given in equation (2.4) are not closed due to the dependence on the covariance through the interactions with the nonlinearity. Further, even by including the next-order covariance dynamics from the

equation. The covariance equation for R can be derived as

$$\frac{dR}{dt} = L_u(\bar{\mathbf{u}})R + RL_u^*(\bar{\mathbf{u}}) + Q_F + Q_\sigma, \tag{2.6}$$

$$\{Q_F\}_{ij} = \sum_{m,n=1}^{N} \langle Z_m Z_n Z_j \rangle B(\mathbf{e}_m, \mathbf{e}_n) \cdot \mathbf{e}_i + \langle Z_m Z_n Z_i \rangle B(\mathbf{e}_m, \mathbf{e}_n) \cdot \mathbf{e}_j. \tag{2.7}$$

Q_F is the energy flux that accounts for the energy transfer from higher-order non-Gaussian statistics, and $Q_\sigma = \sum_k (\mathbf{e}_i \cdot \sigma_k)(\sigma_k \cdot \mathbf{e}_j)$ is positive definite and gives the energy transfer from the stochastic component of the external forcing. The system with (2.6) is still not closed due to the third-order moments that appear in the energy flux term Q_F. This means that the statistical mean response to external forcing cannot be fully resolved in this hierarchical approach, so in practice, various approximations and closures are needed [103].

Statistical Energy and Response to External Forcing

One quantity of primary interest is the total statistical energy of the system, defined as a combination of the energy in the mean and total covariance

$$E = \frac{1}{2}\bar{\mathbf{u}} \cdot \bar{\mathbf{u}} + \frac{1}{2}\,\mathrm{tr}(R). \tag{2.8}$$

Here, $\bar{\mathbf{u}}$ is the mean vector of $\mathbf{u}$, and $\mathrm{tr}(R)$ is the trace of the covariance matrix. The total statistical energy incorporates both the energy contained in the mean flow of the system as well the energy contained in the fluctuations caused by turbulence, so that it encompasses the energy transfers between these features of the system [96].

By controlling this scalar statistical energy containing statistical information from all the scales, we are able to effectively control the potentially very high dimensional system with very low computational cost.

In addition to the energy conservation principle given in equation (2.2), the following assumptions, detailed in [96], are needed to precisely formulate the dynamics of the statistical energy E, namely,

$$B(\mathbf{e}_i, \mathbf{e}_i) \equiv 0, \quad 1 \leqslant i \leqslant N, \tag{2.9}$$

and

$$\mathbf{e}_i \cdot [B(\mathbf{e}_j, \mathbf{e}_i) + B(\mathbf{e}_i, \mathbf{e}_j)] = 0 \quad \text{for any } i, j. \tag{2.10}$$

The above identities characterize the general symmetry in the system that the self-interactions and the closed interactions between pairs of modes vanish under the quadratic nonlinearity. To simplify the notation used in this discussion, we also assume uniform damping $D = -dI$ with $d > 0$. Under these assumptions and using equations (2.4) and (2.6), the total statistical energy satisfies

$$\frac{dE}{dt} = -2dE + \bar{\mathbf{u}} \cdot \mathbf{F} + \frac{1}{2}\operatorname{tr}(Q_\sigma). \tag{2.11}$$

Statistical energy generally decays to the equilibrium state exponentially. Notably, the dynamics depends only directly on the external forcing and first-order mean state, not the covariance or higher-order moments. This allows for controlling the response to external forcing from determining the total statistical energy and solely

considering the external forcing and the response of the mean to the forcing.

To determine the response of the statistical energy to perturbations of the deterministic external forcing, denote the statistical energy of the system under the equilibrium distribution as E_{eq} and denote the energy perturbation as $E'(t) = E(t) - E_{eq}$. The equilibrium energy satisfies $dE_{eq}/dt = 0$, so, using equation (2.11), the equilibrium energy can be explicitly computed using only first-order mean state by

$$E_{eq} = \frac{1}{2d}\bar{\mathbf{u}}_{eq} \cdot \mathbf{F}_{eq} + \frac{1}{4d}\,\mathrm{tr}(Q_\sigma). \tag{2.12}$$

Using the above equation and further denoting the deterministic forcing perturbation as $\delta\mathbf{F} = \mathbf{F} - \mathbf{F}_{eq}$ and the corresponding mean perturbation as $\delta\bar{\mathbf{u}} = \bar{\mathbf{u}} - \bar{\mathbf{u}}_{eq}$, the energy perturbation, E', satisfies

$$\frac{dE'}{dt} = -2dE' + (\mathbf{F}_{eq} \cdot \delta\bar{\mathbf{u}} + \bar{\mathbf{u}}_{eq} \cdot \delta\mathbf{F}) + \delta\bar{\mathbf{u}} \cdot \delta\mathbf{F}. \tag{2.13}$$

Again, for simplicity, assuming that the stochastic component of the external forcing is not perturbed and does not contribute to the energy perturbation dynamics. It is useful to further decompose the response of the energy perturbation for each mode, i.e.

$$\frac{dE'}{dt} = -2dE' + \sum_{k=1}^{N} \left[\bar{u}_{eq,k} \cdot \kappa_k(t) + F_{eq,k} \cdot \delta\bar{u}_k(t;\boldsymbol{\kappa}) + \kappa_k(t) \cdot \delta\bar{u}_k(t;\boldsymbol{\kappa})\right], \tag{2.14}$$

$$E'(0) = E'_0, \tag{2.15}$$

where κ_k is the kth component of $\delta\mathbf{F}$. Likewise, $F_{eq,k}$, $\bar{u}_{eq,k}$, and $\delta\bar{u}_k$ are the kth

components of $\mathbf{F}_{eq}$, $\bar{\mathbf{u}}_{eq}$, and $\delta\bar{\mathbf{u}}$ respectively. To emphasize the central role of the forcing perturbation in each component, the mean response, $\delta\bar{u}_k$, dependence on the forcing perturbation, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)^T$, is noted explicitly in the above equation.

2.3 Methods on Statistical Control

This section describes the control strategies for high-dimensional turbulent systems, including large-amplitude perturbations. The method is generally split into two consecutive steps: i) the calculation of the optimal energy control for the total statistical energy, and ii) the attribution of the forcing contribution for each detailed spectral mode. Especially, high-order accuracy is achieved by considering different ways to approximate the mean responses and high-order feedback in the control. We illustrate the general idea in the diagram in Figure 2.1.

Optimal Control of the Perturbed Energy

As the first step of the statistical control strategy, the objective is to drive the total statistical energy from a perturbed state back to a target equilibrium state with a minimized cost through prescribing the deterministic external forcing perturbation in each mode, κ_k. The direct optimal control of the energy dynamics is achieved by controlling the much simpler scalar equation independent of the full dimensionality of the system. In the second step, the external forcing perturbation that yields this optimal control will be recovered by attributing the contribution to the total energy from each individual mode.

In this and future sections, the energy perturbation will be denoted as E to simplify the notation. Define the energy control problem as

$$\frac{dE}{dt} = -2dE + \sum_{k=1}^{N} \mathcal{C}_k \tag{2.16}$$

where the objective is to control the energy perturbation, E, back to zero over the time interval $[0, T]$ using the controls $\mathcal{C}_k$, representing the total contribution to the energy perturbation from each mode. The cost functional of this control problem is proposed as

$$\mathcal{F}_\alpha[\mathcal{C}_k(\cdot)] \equiv \int_t^T \left[E^2(s) + \sum_{k=1}^{N} \alpha_k \mathcal{C}_k^2(s) \right] ds + k_T E^2(T), \tag{2.17}$$

where α_k gives the relative weights between each control and the total energy perturbation. k_T is the cost coefficient for the final energy perturbation from the equilibrium state at time T. From the above setups, the control of total energy becomes a standard linear control problem with a quadratic cost, and so the Hamilton-Jacobi-Bellman (HJB) equation [5, 14] can be applied to find the optimal control $\mathcal{C}_k^*$. In addition, the total statistical energy is a scalar quantity, so the associated energy control problem is tractable even when the underlying system is high-dimensional.

Solving the HJB equations [105] leads to the following Riccati equation

$$\frac{dK}{dt} = \sum_{k=1}^{N} \alpha_k^{-1} K^2 + 4dK - 1, \quad 0 \leqslant t < T \tag{2.18}$$

$$K(T) = k_T \tag{2.19}$$

which is solved backward in time. The quantity K is a factor of the value function that appears in the HJB equations. The full details of this application of the HJB equations can be found in [105]. Using the solution of K, the optimal response of the energy perturbation, E^*, is given by the forward equation

$$\frac{dE^*}{dt} = -\left(2d + \sum_{k=1}^{N} \alpha_k^{-1} K\right) E^*, \quad 0 \leqslant t < T, \tag{2.20}$$

$$E(0) = E_0. \tag{2.21}$$

Finally, the optimal control in each mode, $\mathcal{C}_k^*$, can be calculated as

$$\mathcal{C}_k^*(t) = -\alpha_k^{-1} K(t) E^*(t). \tag{2.22}$$

Inversion of the Control-Forcing Relation

The goal is to find the forcing perturbation in each mode κ_k that yields the optional control discovered from the energy perturbation. For simplicity in notation, the optimal control of each mode found in the previous section will be denoted by $\mathcal{C}_k$, where the superscript "$*$" is dropped. Using equation (2.14) yields the following relation between the energy control and the forcing perturbation in each mode:

$$\mathcal{C}_k(t) = \bar{u}_{eq,k} \cdot \kappa_k(t) + F_{eq,k} \cdot \delta\bar{u}_k(t;\boldsymbol{\kappa}) + \kappa_k(t) \cdot \delta\bar{u}_k(t;\boldsymbol{\kappa}). \tag{2.23}$$

The mean perturbation response, $\delta\bar{u}_k$, depends on the forcing perturbation $\boldsymbol{\kappa}$. Inverting this relation for κ_k involves solving an ODE system which couples the

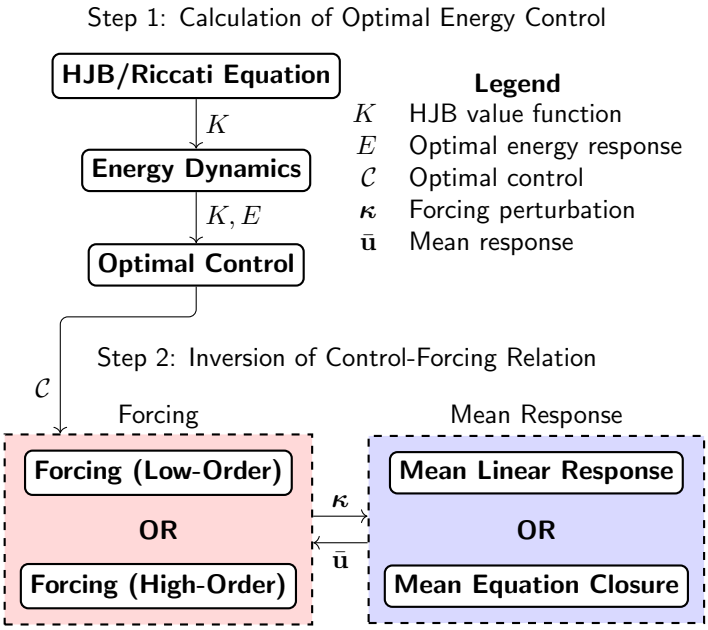

Figure 2.1: Schematic diagram of the statistical control strategy. Step 1 is the calculation the optimal control, $\mathcal{C}_k$, for each mode. First, a Riccati equation, equation (2.18), is solved. This is used to calculate the optimal energy response using equation (2.20). The optimal control is then calculated using equation (2.22). Step 2 consists of finding the forcing perturbation, κ_k, which yields the optimal control in each mode. Inverting the control-forcing relation involves solving coupled equations for the forcing and the mean response to that forcing. There are two choices for the forcing equations: the low order equations, equation (2.24), and the high order equations, equation (2.27). For the mean response there are two strategies: a linear response for the mean, equation (2.34), and the mean dynamics with a closure for the higher-order moments, equation (2.39). Choosing one strategy from each category yields four strategies total.

optimal control, $\mathcal{C}_k$, the forcing perturbation, κ_k, and the mean response, $\delta\bar{u}_k$.

The Low-Order Method

In previous works [102, 105], small forcing perturbation and mean state response are always assumed, and thus the second-order term, $\kappa_k \cdot \delta\bar{u}_k$, is omitted in equation (2.23). In this way the relation only accounts for the dominant leading-order response of the energy to the forcing perturbation. This assumption is justified by using leading order approximations for the mean response for small initial perturbations, which introduce $O(\delta^2)$ errors in the same order as the second-order perturbation term. In doing so, the inversion relation is linearized, leading to simplified analytical solutions. The ODE resulting from this linearized relation, referred to in this paper as the "low-order method", is given by

$$\frac{d\kappa_k}{dt} = \frac{1}{\bar{u}_{eq,k}} \left(\frac{d\mathcal{C}_k}{dt} - F_{eq,k} \cdot \frac{d\bar{u}_k}{dt} \right) \tag{2.24}$$

$$\kappa_k(0) = \frac{1}{\bar{u}_{eq,k}} \left(\mathcal{C}_k(0) - F_{eq,k} \cdot \delta\bar{u}_k(0) \right). \tag{2.25}$$

Here $d\mathcal{C}_k/dt$ can be explicitly calculated using equations (2.18), (2.20), and (2.22)

$$\frac{d\mathcal{C}_k}{dt} = -\alpha_k^{-1} E^*(t)(2dK(t) - 1). \tag{2.26}$$

The term $\delta\bar{u}_k(0)$ denotes the mean perturbation in the initial perturbed state. Solving this ODE system involves approximating the response of the mean, $d\bar{u}_k/dt$, to the forcing perturbation. Strategies for computing the mean responses are detailed in sections 2.3.

The High-Order Method

The second-order term in the control forcing relation was truncated in the low-order method. However, large errors might be introduced due to this omission when the perturbation amplitude grows large. Still, the same analysis can be done by including this additional term. The ODE resulting from inverting equation (2.23), including all contributing terms, is given by

$$\frac{d\kappa_k}{dt} = \frac{1}{\bar{u}_{eq,k} + \delta\bar{u}_k(t)} \left(\frac{d\mathcal{C}_k}{dt} - \left(F_{eq,k} + \kappa_k(t) \right) \cdot \frac{d\bar{u}_k}{dt} \right) \tag{2.27}$$

$$\kappa_k(0) = \frac{1}{\bar{u}_{eq,k} + \delta\bar{u}_k(0)} \left(\mathcal{C}_k(0) - F_{eq,k} \cdot \delta\bar{u}_k(0) \right), \tag{2.28}$$

where $d\mathcal{C}_k/dt$ is given by equation (2.26) and $\delta\bar{u}_k(0)$ is the initial mean perturbation in the perturbed state. Compared to equation (2.24), equation (2.27) contains one more perturbation term that accounts for the higher-order contributions to the energy response. For large perturbations from the target equilibrium state, the high-order feedback term becomes necessary to accurately recover the true forcing forms. The inversion of the full control-forcing relation will be referred to in this paper as the "high-order method" in which the response of the energy to the forcing and mean perturbations is fully resolved. On the other hand, the extra terms in the denominator may lead to additional numerical complications, but the improved performance is generally worth the additional sophistication. We will discuss the performance in the detailed numerical tests in Section 2.4.

Different Strategies to Recover Mean Responses

The inversion of the control-forcing relation given by equations (2.24) and (2.27) requires the response of the mean to the forcing perturbation. Unfortunately, a direct simulation for the mean responses would be prohibitively expensive considering the extremely high dimensional problem, so approximations for the mean response are developed instead. In this subsection, we develop two approaches to efficiently estimate the mean responses without directly solving the full equations. The first strategy uses linear response theory as a convenient way to recover the leading-order mean response based on the Fluctuation-Dissipation Theorem (FDT) [93, 83]. The second approximation provides a more accurate high-order approximation based on solving the explicit mean dynamical equation with a high-order closure.

Mean Linear Response to Forcing

Linear response theory based on the Fluctuation-Dissipation Theorem (FDT) is commonly used in statistical physics and climate science to predict the leading order statistical responses of a system to forcing perturbations without having to expensively solve the Fokker-Plank equation [93, 104, 83, 108]. Below we review the general idea and key formulae for the linear responses used in this paper with more details summarized in Appendix B.1.

According to the FDT the statistical expectation of a functional $A(\mathbf{u})$ in a perturbed state can be written as

$$\langle A(\mathbf{u})\rangle(t) = \langle A(\mathbf{u})\rangle_{eq} + \delta\langle A(u)\rangle(t) + O(\delta^2) \tag{2.29}$$

where $\langle\cdot\rangle(t)$ denotes statistical expectation under the probability density of the perturbed state $p(\mathbf{u}, t)$, and $\langle\cdot\rangle_{eq}$ denotes statistical expectation under the time-invariant equilibrium probability density $p_{eq}(\mathbf{u})$. The quantity $\delta\langle A(u)\rangle(t)$ is the leading-order response to the perturbation and is given by

$$\delta\langle A\rangle(t) = \int A(\mathbf{u})\delta p'(\mathbf{u}, t)\, d\mathbf{u} \tag{2.30}$$

where $\delta p'(\mathbf{u}, t) = p(\mathbf{u}, t) - p_{eq}(\mathbf{u})$ is the perturbation of the probability density. Under the assumption that the perturbed state is as a result of the external forcing perturbation $\delta\mathbf{F}(t) = \mathbf{w}\delta f(s)$, linear response theory states that this leading order response can be calculated as the convolution

$$\delta\langle A\rangle(t) = \mathcal{R}_A * \delta f = \int_{-\infty}^{t} \mathcal{R}_A(t-s)\delta f(s)\, ds \tag{2.31}$$

where $\mathcal{R}_A$ is

$$\mathcal{R}_A(t) = \langle A[\mathbf{u}(t)]G[\mathbf{u}(0)]\rangle_{eq} \tag{2.32}$$

and G is given by

$$G(\mathbf{u}) = -p_{eq}^{-1}\, div_{\mathbf{u}}(\mathbf{w}p_{eq}). \tag{2.33}$$

Using equation (2.31) with the functional $A(\mathbf{u}) = u_k - u_{eq,k}$, the linear mean response of the kth mode to each mode of the forcing perturbation $\kappa_\ell = \mathbf{e}_\ell \cdot \boldsymbol{\kappa}$ is computed by

$$\delta\bar{u}_k = \sum_{\ell=1}^{N} \left[\int_0^t \mathcal{R}_{\bar{u},k\ell}(t-s)\kappa_\ell(s)\, ds + \int_{-\infty}^{0} \mathcal{R}_{\bar{u},k\ell}(t-s)\delta F_{p,\ell}\, ds\right] + O(\delta^2), \tag{2.34}$$

where

$$\mathcal{R}_{\bar{u},k\ell}(t) = \langle (u_k(t) - \bar{u}_{eq,k}) G_\ell[\mathbf{u}(0)] \rangle_{eq}, \tag{2.35}$$

and

$$G_\ell(\mathbf{u}) = -\frac{\mathrm{div}_{\mathbf{u}}(\mathbf{e}_\ell \cdot p_{eq}(\mathbf{u}))}{p_{eq}(\mathbf{u})}. \tag{2.36}$$

Here $p_{eq}(\mathbf{u})$ is the equilibrium probability density of $\mathbf{u}$ and $\langle \cdot \rangle_{eq}$ denotes statistical expectation with respect to p_{eq}. The basis vector $\mathbf{e}_\ell$ corresponds with the contribution to the response from the ℓth mode. While equation (2.34) sums the contributions from each of the N modes to the mean response of the kth mode, in practice only the contributions from a few modes are needed. The forcing perturbation $\delta \mathbf{F}_{p,\ell}$ is the constant external forcing perturbation corresponding to the initial perturbed state. This is contrasted with $\boldsymbol{\kappa}$, which is the time-dependent forcing perturbation based on the control that takes over at time $t = 0$, which is why the integral in (2.34) is split into two intervals. Notably, the linear response operator $\mathcal{R}_{\bar{u},k\ell}$ only depends on the PDF of the target equilibrium state and does not depend on the perturbed state. This explicit formula provides a great computation reduction for convenient calculation of the mean responses but only in the leading order.

Then, the linear response approximation is incorporated into the low-order and high-order methods given in equations (2.24) and (2.27), which utilize the derivative of the mean response. By truncating the $O(\delta^2)$ terms and taking the derivative of equation (2.34) with respect to time, the equations of the mean linear

response approximation are given by

$$\frac{d\bar{u}_k}{dt} = \frac{d(\delta\bar{u}_k)}{dt} = \sum_{\ell=1}^{N} \left[\mathcal{R}_{\bar{u},k\ell}(0)\kappa_\ell(t) + \int_0^t \mathcal{R}'_{\bar{u},k\ell}(t-s)\kappa_\ell(s)\, ds - \mathcal{R}_{\bar{u},k\ell}(t)\delta F_{p,\ell} \right] \tag{2.37}$$

with initial condition

$$\delta\bar{u}_k(0) = \sum_{\ell=1}^{N} \left[\int_{-\infty}^{0} \mathcal{R}_{\bar{u},k\ell}(t-s)\delta F_{p,\ell}\, ds \right]. \tag{2.38}$$

While the linear response includes contributions from all modes, typically, only the contribution from a few modes is relevant, allowing the rest to be truncated to further save the computation cost.

Mean Dynamical Equation Closure for Response to Forcing

While directly using linear response theory for the mean response provides a useful approximation, as in section 2.3, the linear response is proved inadequate in many cases with large perturbations. This paper develops another strategy to incorporate the mean response based directly on the mean dynamical equation explicitly given in equation (2.4) subject to the forcing perturbation κ. For this method, the dependence of the mean dynamics on the second-order moments through the nonlinearity is closed using a linear response for the covariance. Compared to the linear response of the mean, which only incorporates statistical information from the target equilibrium distribution, the use of the mean dynamical equation incorporates crucial additional dynamical information. This can be particularly useful when the system is perturbed into a different dynamical regime.

The mean dynamical equation in the mean closure method is given by (2.4)

$$\frac{d\bar{\mathbf{u}}}{dt} = (L + D)\bar{\mathbf{u}} + B(\bar{\mathbf{u}}, \bar{\mathbf{u}}) + \sum_{i,j=1}^{N} R_{ij}(\boldsymbol{\kappa}) B(\mathbf{e}_i, \mathbf{e}_j) + \mathbf{F}_{eq} + \boldsymbol{\kappa}. \tag{2.39}$$

This equation still requires the solution of the second-order covariances R inversely dependent on the form of the forcing perturbation $\boldsymbol{\kappa}$. A closure model is constructed using linear response theory for the response of the covariance, R_{ij}, to the external forcing

$$R_{ij}(t;\boldsymbol{\kappa}) = R_{eq,ij} + \sum_{\ell=1}^{N} \left[\int_0^t \mathcal{R}_{R,ij\ell}(t-s)\kappa_\ell(s)\, ds + \int_{-\infty}^{0} \mathcal{R}_{R,ij\ell}(t-s)\delta F_{p,\ell}\, ds \right] + O(\delta^2), \tag{2.40}$$

where the linear response operator for the covariance is given by

$$\mathcal{R}_{R,ij\ell}(t) = \langle (u_i(t) - \bar{u}_{eq,i})(u_j(t) - \bar{u}_{eq,j}) G_\ell[\mathbf{u}(0)] \rangle_{eq}, \tag{2.41}$$

and

$$G_\ell(\mathbf{u}) = -\frac{\mathrm{div}_{\mathbf{u}}(\mathbf{e}_\ell \cdot p_{eq}(\mathbf{u}))}{p_{eq}(\mathbf{u})}. \tag{2.42}$$

The higher-order closure of the mean equation enables a better characterization of the mean responses respecting its explicit nonlinear dynamics. The errors from the linear response approximation then appear in the second-order covariances rather than the first-order mean. The linear response in (2.41) will require the computation of lagged third moments, adding more non-Gaussian information

into the approximation. As in Section 2.3, a quasi-Gaussian approximation is used for the linear response operator to efficiently compute the response operators.

Strategies for Calculating the Linear Response Operators

The linear response operators given in equations (2.32), (2.34), and (2.40) be computationally expensive to calculate in practice due to non-Gaussian features and high-dimensionality of the equilibrium probability distribution. In this paper, we adopt the quasi-Gaussian approximation [83, 98, 59] where the equilibrium probability density, $p_{eq}(\mathbf{u})$, is approximated by a Gaussian distribution to directly calculate $G_\ell(\mathbf{u})$, in which case the equilibrium probability density is given by

$$p_{eq}(\mathbf{u}) = (2\pi)^{N/2} \det(\mathbf{R})^{1/2} \exp\left(-(\mathbf{u} - \bar{\mathbf{u}}_{eq})^T \mathbf{R}_{eq}^{-1} (\mathbf{u} - \bar{\mathbf{u}}_{eq})\right). \tag{2.43}$$

Then equations (2.36) and (2.42) can be calculated explicitly using this approximation as

$$G_\ell(\mathbf{u}) = \mathbf{e}_\ell \cdot \mathbf{R}_{eq}^{-1} (\mathbf{u} - \bar{\mathbf{u}}_{eq}). \tag{2.44}$$

This is then utilized in equation (2.32), giving

$$\mathcal{R}_A(t) = \langle A[\mathbf{u}(t)](\mathbf{e}_\ell \cdot \mathbf{R}_{eq}^{-1}(\mathbf{u}(0) - \bar{\mathbf{u}}_{eq}))\rangle_{eq} \tag{2.45}$$

where $A[\mathbf{u}]$ is the appropriate functional. For equation (2.34) the functional is $A[\mathbf{u}] = u_k - u_{eq,k}$ and for (2.40) it is given by $A[\mathbf{u}] = (u_i - u_{eq,i})(u_j - u_{eq,j})$. Equation 2.45 can then be calculated numerically from a model trajectory. Notice

this approximation is quasi-Gaussian since the higher-order moments could still be involved due to the probability expectation in (2.35) based on the true solution rather than only the Gaussian closure.

There are many other strategies for approximating the linear response operator [104]. While the quasi-Gaussian approximation is sufficient for the purposes of this paper, other strategies can be considered for calculating the linear response operator. The previous works on the statistical control strategy [102, 105] have considered an exponential fit to approximate the linear response operator

$$\mathcal{R}_A(t) \approx \exp(-\gamma_k t) \tag{2.46}$$

where the complex parameter γ_k can be systematically chosen based on an information theory criterion. This use of an exponential fit is justified from the case with a quasi-Gaussian approximation and a diagonal covariance matrix where the linear response operator reduces to an autocorrelation function of $\mathbf{u}$, which frequently have exponential and oscillatory structures. While this exponential fit is not utilized in this paper, it is noted for its potential application as a useful approximation in practical settings and for the theoretical convenience of having an explicit form of the linear response operator.

2.4 Numerical Results

We examine the performance of the methods developed in Section 2.3 using detailed numerical tests. Combining the high-order or low-order methods with the

mean equation closure or mean linear response gives a total of four approaches for inverting the control-forcing relation to compare. These methods concern different aspects of component approximations of the statistical control strategy, and the choice to implement each one can be made accordingly based on the specific problem. It is important to note that while the energy response to a theoretically optimal forcing perturbation is known, inverting the control-forcing relation in equation (2.23) necessarily utilizes approximations, and so the actual energy responses produced by the forcing perturbations calculated by each strategy need to be compared to the optimal energy response.

These four strategies are evaluated on two complex nonlinear models exhibiting various dynamical and statistical behaviors. The first test model is a prototypical triad nonlinear model [103] focusing on a generic coupling between three modes of a turbulent system. It can exhibit a wide variety of nonlinear and non-Gaussian behaviors. Two regimes of this model are considered, including a nearly Gaussian regime with nonlinear energy transfers between modes and a non-Gaussian regime exhibiting an energy cascade representing the transition to turbulence. The second test model is the classic Lorenz '96 model [89] with 40 dimensions which shows multiple dynamical regimes depending on the magnitude of the external forcing. Large perturbations inducing regime switching will be the primary consideration. These test models will illustrate the differences between the statistical control strategies.

The experimental setup is as follows. First, the external forcing perturbation is calculated offline using different statistical control strategies. The optimal energy

control is calculated using equations (2.18), (2.20), and (2.22). The forcing perturbation which yields this control is found using either the high-order method given in equation (2.27) or the low-order method given in equation (2.24). The mean response to the forcing perturbation is approximated by the mean equation closure model described in equations (2.39) and (2.40) or by the mean linear response in equations (2.34) and (2.37). Second, the external forcing perturbation is applied to a Monte Carlo simulation of the underlying dynamical system. An initial ensemble of model trajectories of size $M = 1 \times 10^4$ is drawn from the initial distribution. The perturbed initial state is created by a deterministic forcing perturbation. The deterministic component of the external forcing $\mathbf{F}_{eq}$ is perturbed by a constant forcing amplitude $\delta\mathbf{F}_p$ to drive the statistics of the ensemble into a perturbed state away from the original statistical equilibrium.

At the time $t = 0$, the control strategy takes over. Thus, the previous forcing perturbation is replaced by the statistical control forcing, $\boldsymbol{\kappa}$. The response of the statistical energy to the forcing is calculated from the ensemble and is tracked as the system is controlled back to the equilibrium state. These energy responses from each strategy are compared to the theoretically optimal energy response in equation (2.20). The uncontrolled case where no control forcing perturbation is applied, in which the energy naturally decays back to the original equilibrium state, is also used as a point of comparison. Other quantities, such as the mean response, variance response, and empirical control, are also compared for tracking the performance.

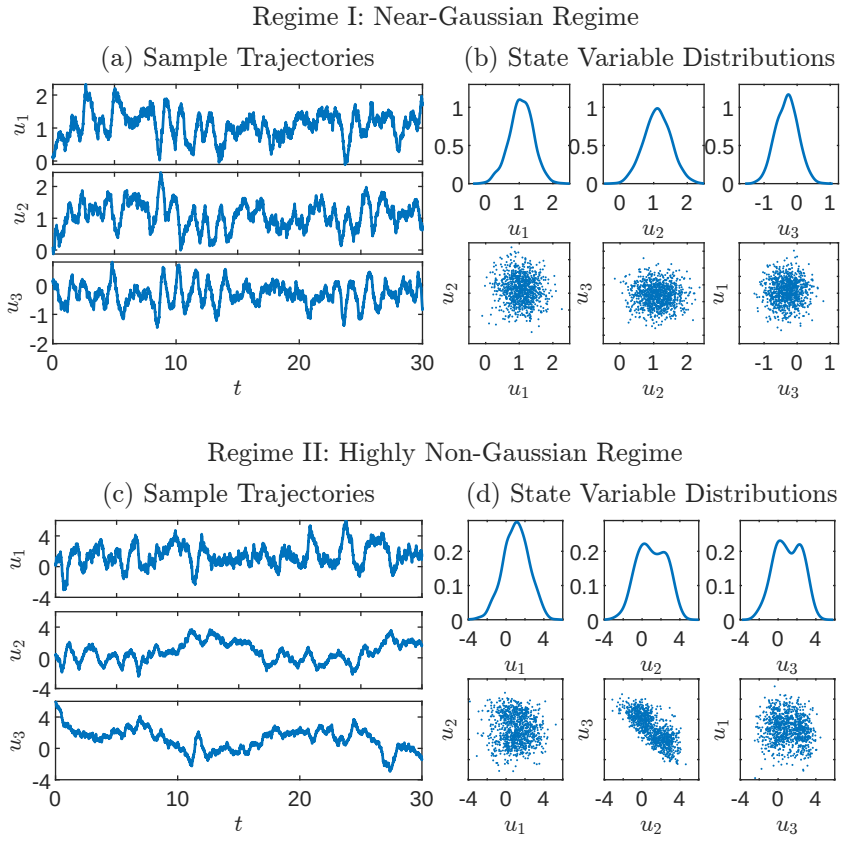

Figure 2.2: The dynamics and equilibrium distributions of the prototype triad model under two different regimes. Panels (a) and (c) show sample trajectories of each regime of the model. Panels (b) and (d) show the equilibrium marginal distributions of the state variables as well as their pairwise joint distributions in each regime. The nonlinearity in the model produces non-Gaussian distributions in each regime. In particular Regime II exhibits intermittency and highly non-Gaussian statistics.

A Prototype Nonlinear Triad Model

The first test model is a prototypical 3-dimensional model equipped with a quadratic energy-conserving nonlinearity, which is referred to in this paper as the *triad model*.

The triad model represents a generic nonlinear coupling between three variables universal in turbulent flows [103, 101, 35]. Such a triad interacting structure would emerge as the bare truncation of three identified modes from a high-dimensional turbulent model. The triad model can also generate a wide variety of nonlinear and non-Gaussian behaviors due to the dominant role of the nonlinear coupling term. This sets a desirable first test model to evaluate the skills of the different proposed approaches considering the high-order contributions in the mean state and the energy equation. Despite the nonlinearity, the triad model is analytically tractable in terms of the equilibrium statistics when the linear parts have certain special structures [101], making it an appropriate test model used to have an in-depth study of the different features of the four proposed strategies.

The state variables of the triad model are represented by $\mathbf{u} = (u_1, u_2, u_3)^T$ with governing differential equations:

$$\frac{du_1}{dt} = L_2 u_3 - L_3 u_2 - d_1 u_1 + B_1 u_2 u_3 + F_1 + \sigma_1 \dot{W}_1, \tag{2.47}$$

$$\frac{du_2}{dt} = L_3 u_1 - L_1 u_3 - d_2 u_2 + B_2 u_3 u_1 + F_2 + \sigma_2 \dot{W}_2, \tag{2.48}$$

$$\frac{du_3}{dt} = L_1 u_2 - L_2 u_1 - d_3 u_3 + B_3 u_1 u_2 + F_3 + \sigma_3 \dot{W}_3. \tag{2.49}$$

In addition, the quadratic coupling coefficients satisfy

$$B_1 + B_2 + B_3 = 0, \tag{2.50}$$

which ensures the general energy-conserving property (2.2) of the turbulent dy-

namical system framework. Figure 2.2 shows two typical regimes of the triad model used to evaluate the strategies: one has near-Gaussian statistics, while the other is highly non-Gaussian.

Control on a Near-Gaussian Regime

The first test regime for the triad model is a near-Gaussian regime which nonetheless contains strong nonlinear energy transfers between modes to reach the equipartition of energy. Sample trajectories and equilibrium distributions for this regime are pictured in Regime I in Figure 2.2. Nearly Gaussian and weakly non-Gaussian features are common in practice, such as in fully turbulent flow with strong mixing. The damping coefficients for this regime are $d_1 = d_2 = d_3 = 1$. The linear dispersion coefficients are $L_1 = 3$, $L_2 = 2$, and $L_3 = -1$. The nonlinear quadratic coupling coefficients are $B_1 = 1$, $B_2 = -0.6$, and $B_3 = -0.4$. The deterministic external forcing for the equilibrium state is given by $F_1 = F_2 = 1$ and $F_3 = -1$ while the stochastic external forcing coefficients are given by $\sigma_1 = \sigma_2 = \sigma_3 = 0.5$. To perturb the model state, F_3 is perturbed by $\delta F_{p,3} = -4$ until time $t = 0$. While only F_3 is perturbed initially, all modes are used to control the system back to the equilibrium state.

The control of the system from the perturbed state back to the equilibrium state under different strategies is shown in Figure 2.3. All the methods show faster convergence than the no-control scenario to efficiently return the unperturbed equilibrium state. Notably, the high-order method achieves the most accurate near-optimal performance, particularly the high-order method with a mean equation closure. In contrast, the low-order methods overshoot the response incurring

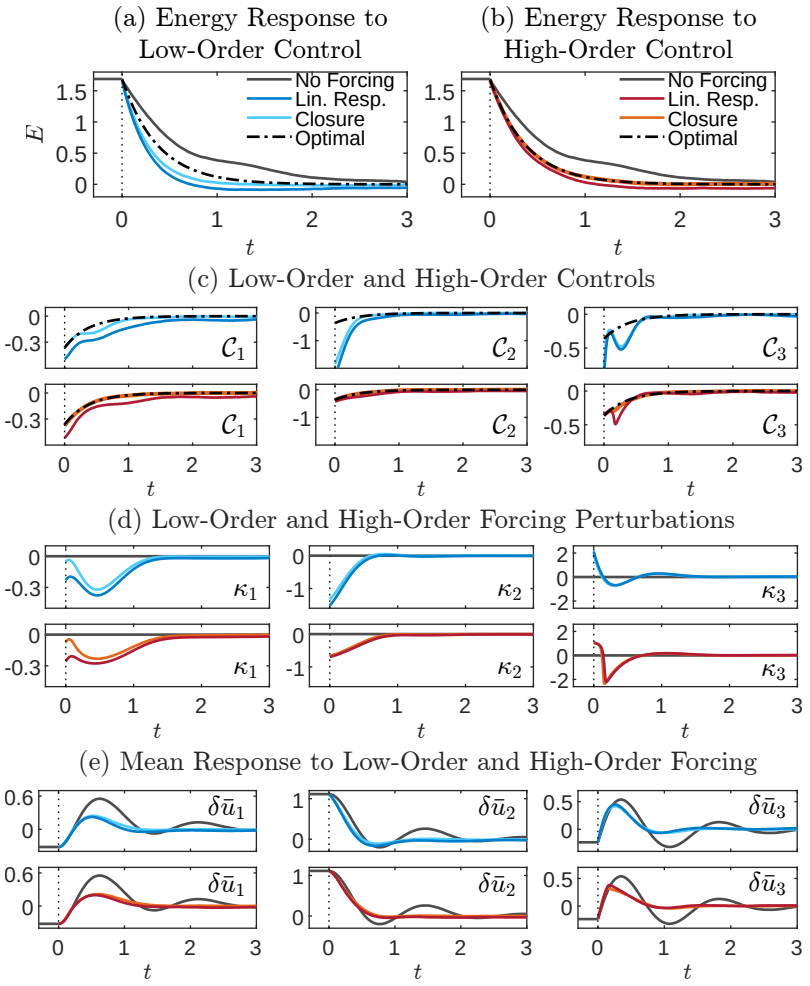

Figure 2.3: The control of the prototype triad model from the perturbed state back to the equilibrium state in the near-Gaussian regime. Panels (a) and (b) show the energy response to the forcing, including the response from no control, the linear response and equation closure strategies, as well as the theoretically optimal response. Panel (a) shows the energy response for the low-order strategies: using a mean linear response and using a mean equation closure model. Panel (b) shows the same with the high-order strategies. Panel (c) compares the controls realized by the various strategies to the optimal control. Panel (d) shows the forcing perturbations prescribed by each strategy. Note that the control-forcing relation cannot be inverted exactly, so there is not forcing perturbation that corresponds to the theoretically optimal energy response. Panel (e) shows the responses of the mean under each strategy.

relatively higher costs. This is the first confirmation of the crucial role of the higher-order correction in the energy equation when nonlinearity is dominant, as in the triad system. There is a stark difference in the control forcing perturbations between the high-order and low-order methods which can be seen by κ_2 and κ_3 in panel (d) of Figure 2.3. The initial forcing perturbation for κ_2 differs entirely from the low-order and high-order methods. This is because the corresponding initial mean perturbation, $\delta\bar{u}_2$, pictured in panel (e), is relatively large; thus, the initial contribution of the second-order $\kappa_2 \cdot \delta\bar{u}_2$ term to the control-forcing relation is quite significant. Indeed, the low-order method overcompensates for lacking this high-order term with a large initial forcing perturbation. In contrast, the initial forcing perturbation for the high-order method is significantly smaller. For the third mode, κ_3, the initial forcing perturbation is comparable between the high-order and low-order methods due to the initial mean perturbation $\delta\bar{u}_3(0)$ being relatively small. However, the high-order method produces a stronger forcing perturbation in κ_3 shortly after the initial time. This is explained by the observation that in all methods, the mean perturbation response, $\delta\bar{u}_3$, quickly takes on a large value. So the second-order $\kappa_3 \cdot \delta\bar{u}_3$ term in the control-forcing relation is non-negligible. Because of this, only the high-order method can account for the contribution of the second-order term to the energy response. One can also see a difference between the initial forcing perturbation for κ_1 between the mean linear response and mean equation closure methods. This is an example where the mean linear response does not accurately produce the initial mean perturbation, while the mean closure model directly incorporates the initial mean perturbation.

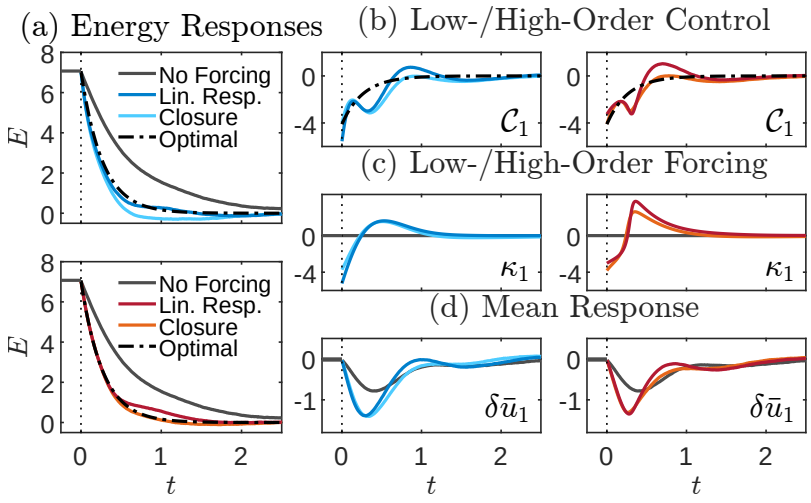

Figure 2.4: Example of controlling a highly non-Gaussian regime in the prototypical triad model. Panel (a) shows the response of the energy to the forcing perturbation for all strategies. Panel (b) shows the control for each strategy for the u_1 mode. Panel (c) shows the forcing perturbation in the u_1 mode. Panel (d) shows the mean response in the u_1 mode.

Control on a highly non-Gaussian Regime

The second test regime, shown in Regime II of Figure 2.2, has highly non-Gaussian statistics and intermittency. It features an energy cascade from u_1 to u_2 and u_3 reminiscent of the transition to turbulence. The damping coefficients are $d_1 = d_2 = d_3 = 1$, the quadratic nonlinear coupling coefficient are $B_1 = 2$ and $B_2 = B_3 = -1$, and the linear dispersion coefficients are $L_1 = 0.03$, $L_2 = 0.02$, and $L_3 = -0.01$. The unperturbed deterministic external forcing is given by $F_1 = F_2 = F_3 = 2$ and the stochastic external forcing coefficients are $\sigma_1 = 2$ and $\sigma_2 = \sigma_3 = 1$. The perturbed state is achieved through constant deterministic forcing perturbations $\delta F_{p,1} = \delta F_{p,2} = \delta F_{p,3} = 2$ until time $t = 0$. Figure 2.4 shows the results of applying

the control strategies to Regime II of the triad model. We focus on the performance of the dominant mode u_1. The other two modes $u_2 u_3$ have qualitatively similar performance and are omitted for a cleaner representation. Similar to the near-Gaussian regime, there is a strong forcing perturbation in κ_1 for the high-order method after the initial time. The initial forcing using different methods is comparable due to the relatively small initial mean perturbation $\delta\bar{u}_1(0)$ but the mean response shortly after the initial time requires the high-order method to capture the subsequent response by the energy. In addition, this example illustrates how the mean equation closure model can produce more accurate forcing perturbations under a strongly nonlinear non-Gaussian regime even when the initial perturbation is similar. As expected, stronger non-Gaussianity requires a more accurate calibration of the mean responses taking into account the higher-order statistics. The high-order equation closure gains a more accurate estimation of the mean state, thus leading to the most accurate result. For low-order methods, lacking the higher-order correction term often leads to larger errors. We suspect that the agreement in the low-order linear response approach comes as an accidental cancellation of errors.

A High-Dimensional Model with Multiple Regimes

The Lorenz '96 model is a standard test model which mimics geophysical waves and exhibits phenomena such as mid-latitude baroclinic instability [89]. The model is defined in a 40-dimension vector state by

$$\frac{du_j}{dt} = (u_{j+1} - u_{j-2})u_{j-1} - u_j + F, \quad j = 1, \ldots, 40 \tag{2.51}$$

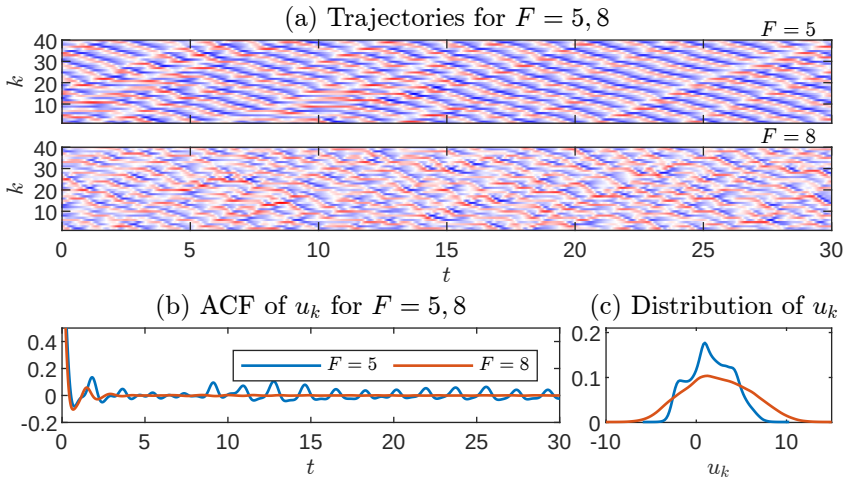

Figure 2.5: Sample trajectories and distributions of the 40-dimensional Lorenz 96 model for both the $F = 5$ (weakly chaotic; highly non-Gaussian) and $F = 8$ (strongly chaotic; nearly Gaussian) regimes. Panel (a) shows sample trajectories for one sample of each regime in the form of the Hovmoller diagram. Panel (b) shows the autocorrelation functions (ACFs) for each regime. Note that the ACF for the $F = 5$ regime exhibits long-term oscillatory behavior while the ACF of the $F = 8$ regime decays very fast. Panel (c) shows the equilibrium distribution for each regime. The $F = 5$ regime is highly non-Gaussian while the $F = 8$ regime is nearly Gaussian.

Here, the variables are indexed periodically, e.g., $u_{41} = u_1$. The external forcing F is the same for each mode, and so the equilibrium statistics of the system are spatially invariant. Note that quadratic nonlinearity satisfies the energy conservation law given in equation (2.2), which can be shown by the symmetry of the nonlinearity in equation (2.51).

A key property of the Lorenz '96 model is that it exhibits a variety of dynamical and statistical regimes by altering the value of F. Multiple dynamical regimes are typical of complex turbulent systems and represent a classic obstacle to effective

control. Figure 2.5 exhibits two dynamical regimes corresponding to $F = 5$ and $F = 8$. The $F = 5$ regime is weakly chaotic. It has highly non-Gaussian statistics and a long decorrelation time. Meanwhile, the regime corresponding to $F = 8$ features near-Gaussian statistics and strongly chaotic dynamics with a correspondingly short decorrelation time. Previous results [102, 105] have shown the statistical control strategy to be effective at controlling small perturbations in the Lorenz '96 model back to the equilibrium state.

In the current experiment, we show the efficacy of the strategies on a large perturbation which drive the system into a different dynamical regime. This leads to a much more challenging problem since the model state goes through a statistical transition between two distinctive regimes. The linear response estimation is no longer valid since the model moves far beyond the linear and near-Gaussian regime. The higher-order corrections become necessary to guarantee effective control performance. In this case, the $F_{eq} = F = 5$ regime is taken as the equilibrium state and $\delta F_p = 3$ so that the perturbed state is in the $F = 8$ regime. Figure 2.6 shows the control strategies for this large perturbation. In this case, the high-order method with the mean dynamical equation closure shows the most effective strategy. In fact, it shows that combining both the high-order and the mean closure methods is essential to achieve good performance. This is a typical example to confirm the necessity of including high-order corrections when nonlinear and non-Gaussian features become dominant.

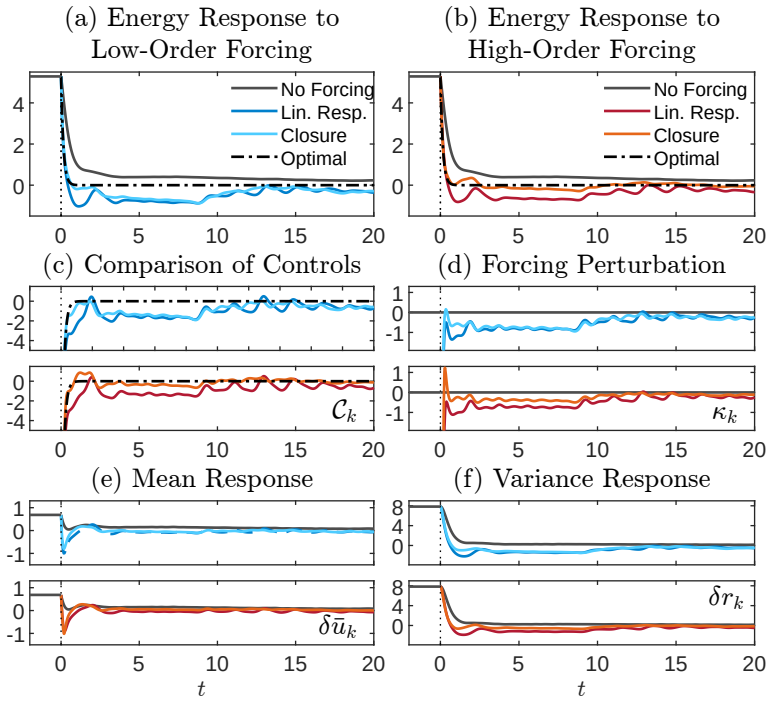

Figure 2.6: The control of the Lorenz 96 model from the perturbed state of $\mathsf{F} = 8$ back to the equilibrium state of $\mathsf{F} = 5$. Note this is a large perturbation into a regime with very different dynamics and statistics from the equilibrium. Panels (a) and (b) show the response of the energy perturbation to the control forcing for the low-order strategies and high-order strategies respectively. The energy perturbation is normalized by the dimension of the system. Panels (c)-(f) show the controls, forcing, mean response, and variance response for each mode. Note the system is translationally invariant, so the corresponding values for each mode are the same.

2.5 Further Discussions

The statistical control strategies extend the effective control methods beyond the small perturbation scenario and demonstrate promise for applications to a broader range of turbulent situations. To summarize, they enjoy several attractive features. Using the total energy as the object of control, there is no need to track and control a large dimension of instabilities due to the energy-conservation principle. Thus, the computational cost is significantly reduced and independent of the dimensionality of the system. Further, the control can be determined entirely offline and only requires statistical information about the target equilibrium state, which is usually available from history observation data in many realistic applications.

Here, we discuss several key features in the new high-order control strategies based on the observations from the numerical experiments.

The High-Order Correction

The control-forcing relation, which encodes the energy response to the external deterministic forcing perturbation, has a second-order perturbation term, $\delta\mathbf{F} \cdot \delta\bar{\mathbf{u}}$, which accounts for the higher-order contributions of the deterministic forcing perturbation to the energy response. This term consists of the product of the forcing perturbation and the mean perturbation in response to the forcing. Under the circumstances with small perturbations from the equilibrium state, both the mean perturbation and the forcing perturbation are small, so this term can be truncated without compromising the accuracy of the energy response. This method, where

only the leading-order contributions to the energy response are considered, is the low-order method. However, for most large perturbations, the second-order term becomes large and significantly affects the energy response. The high-order approach incorporates this second-order term in the inversion of the control-forcing relation, fully resolving the energy response given the forcing perturbation and mean response. We explore several circumstances where the second-order perturbation term significantly impacts the energy response; thus, adding the high-order method can yield significant improvements over the low-order method.

When the initial mean perturbation is large, the initial value of the external forcing perturbation is greatly affected by the presence of the second-order term. This can be seen in equations (2.24) and (2.27) where the initial condition for the high-order method includes an extra term for the initial mean perturbation $\delta\bar{\mathbf{u}}(0)$. The effect of the initial mean perturbation on the resulting forcing perturbation can also be seen in κ_2 of Figure 2.3 in the control of Regime I in Section 2.4. However, a large initial mean perturbation is unnecessary, and the second-order term can still have a significant effect even when the initial mean perturbation is relatively small if there is still a large initial energy perturbation due to the variance. Because the initial energy perturbation is large, the relative balance of the mean and variance in the total energy perturbation can shift over time, resulting in a potentially large mean perturbation after the initial time. In this case, the second-order term significantly impacts the evolution of the forcing perturbation even with the same initial conditions. Several examples of this phenomenon can be seen in the numerical tests in Sections 2.4 and 2.4, especially in Figure 2.6, where a drastic phase transition is

shown.

As a further comment, whether to use the low-order or high-order methods can be made independently for each mode. For example, in a multiscale system, the effect of the second-order term may be small relative to the total energy response for small-scale modes and only be significant for larger-scale modes. In this case, the high-order method could be applied to only a subset of large-scale modes, while the low-order method is used for the rest, simplifying the dynamics in those small-scale modes without compromising performance.

The Mean Closure Equation

The response of statistical energy to the external forcing depends directly on the mean response to the forcing. It is indirectly linked to the higher-order moments through the mean dynamical equation. This property is critical to formulating the statistical control strategy, allowing for attributing an external forcing perturbation to the optimal control by solving the control-forcing relation. Therefore, accurately approximating the mean response to external forcing is vital to the success of the statistical control strategy. Linear response theory effectively approximates the mean response for small perturbations, and it can perform well for larger perturbations in some cases when non-Gaussian statistics is not so important. However, its skill degenerates when the system is largely perturbed to a different dynamical regime where the linear response operator, based solely on the unperturbed dynamics, can provide very little information for the future perturbed state.

Using a mean closure equation for the mean response, which directly incorpo-

rates dynamics from the model, is expected to show improved performance when the initial perturbation spans multiple dynamical regimes. A mean dynamical closure equation is based on the explicit mean dynamics given in equation (2.4), in which the dependence on higher-order moments is closed using a suitable approximation. The closure considered in this paper utilizes a linear response for the higher-order contribution of the covariance described in equations (2.39) and (2.40). While this mean closure equation still relies on the linear response for the covariance, the mean dynamics still provide more information about the perturbed regime than the mean linear response.

The initial forcing perturbation is affected by the choice of mean response. The mean linear response cannot directly use the initial mean perturbation and instead must use the initial mean perturbation predicted by the linear response to a constant forcing perturbation. This is necessary to guarantee the convergence of the forcing perturbation to zero in the linear response case. The mean closure model, however, can utilize the initial mean perturbation directly. This is illustrated by κ_1 in Figure 2.3 in Regime I of Section 2.4 where the initial mean response differs between the linear response and mean closure methods. In this case, the mean equation closure achieves better performance. Even when the initial mean perturbation is similar to the initial mean perturbation predicted by the linear response, the mean closure model can provide more accurate dynamics in many cases. In Section 2.4, the mean equation closure method performs better among all cases, especially in Regime II. In Section 2.4, the Lorenz '96 model is perturbed from a non-Gaussian regime to a near-Gaussian turbulent regime. The mean equation closure again performs better

than the linear response in this case with multiple dynamical regimes.

Convergence to the Equilibrium State

The total statistical energy bounds the total mean and variance. Ideally, one hopes that the efficient control of the equilibrium energy will also achieve the efficient control of the mean and variance back to the equilibrium state. While this appears to be the case in most applications, this is not mathematically guaranteed by the statistical control strategy. Figure 2.7 illustrates an example where the optimal energy response is achieved through the high-order mean closure strategy, but the external forcing perturbation converges to a constant non-zero state. Essentially the system converges to a different equilibrium with a different constant external forcing but the same equilibrium energy.

The cost functional given in equation (2.17) only penalizes the strength of the direct energy control $\mathcal{C}_k$ rather than the external forcing perturbations κ_k that yield that control. In addition, the control-forcing relation given in Equation (2.23) admits multiple solutions in the limit in both the low-order and high-order formulations. One natural fix for such an issue is incorporating additional terms into the cost function, for example, explicitly excluding the mean state.

2.6 Conclusion

An efficient method of controlling the complex turbulent system with energy conserving nonlinearity is achieved through control of the total statistical energy from

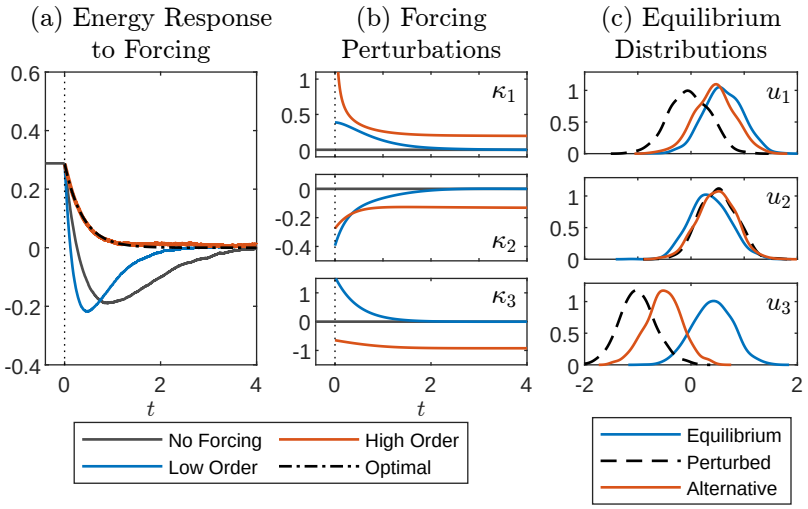

Figure 2.7: An example where the optimal energy response is achieved, but the system is forced to a different equilibrium state. The triad model has parameters $d_1 = d_2 = d_3 = 1$, $L_1 = L_2 = L_3 = 0$, $B_1 = 1$, $B_2 = -0.6$, $B_3 = -0.4$, $F_1 = F_2 = F_3 = 0.5$, and $\sigma_1 = \sigma_2 = \sigma_3 = 0.5$. The perturbed state has $F_3 = -1$. Panel (a) shows the energy response to the low-order and high-order control strategies. Panel (b) shows the forcing perturbations in each mode. Note that the forcing perturbation for the high-order method does not converge to zero. Panel (c) shows the equilibrium distribution, the perturbed distribution, and the alternative distribution achieved by the high-order method that yields the same statistical energy.

a perturbed state back to equilibrium without controlling the large number of multiscale and potentially unstable modes. This paper proposed new statistical control strategies overcoming the inherent limitations in previous works [102, 105], which had been restricted to scenarios with small perturbations from the equilibrium state. Incorporating the high-order term in the control-forcing relation accounts for the second-order contribution of the perturbations to the statistical energy, allowing the strategy to account for the response of the energy more accurately to large amplitude external forcing perturbations. Additionally, introducing a mean

dynamical closure model allows the statistical control strategy to better account for large perturbations that drive the system into different dynamical regimes whose dynamics cannot be adequately reflected directly by a mean linear response approximation. These strategies allow for the practical application of the statistical control strategy to a wider variety of perturbations and regimes than previously possible.

The field of statistical control theory remains relatively underexplored, providing many promising research directions. The results presented in this paper could be further refined by developing more sophisticated methods for incorporating the mean and covariance dynamics into the mean response. Besides, other designs for mean closure models could be considered, several of which are described in [103]. Reduced order stochastic models, such as those described in [124], could also be used to design suitable mean response operators and closures. It may also be possible to incorporate other more suitable statistical functionals into the control strategy in addition to the statistical energy. This would allow, for example, the direct control of the statistical mean, giving finer control over the response of the system. Lastly, while the current statistical control strategy is conducted in an open-loop fashion, determining the control offline, a natural extension would be incorporating feedback from the system into a closed-loop statistical control strategy. This could be accomplished, for example, by combining estimates of the actual mean response of the system obtained through data assimilation into the inversion of the control-forcing relation. The mean response operator then can be effectively estimated from the output of the data assimilation scheme based

on the observation data, such as the ensemble Kalman filter. The computational advantages of the statistical control strategy would be very advantageous in such a closed-loop formulation, which requires real-time incorporation of the model feedback.

3 PROBABILISTIC EDDY IDENTIFICATION WITH UNCERTAINTY QUANTIFICATION USING ICE FLOE TRAJECTORIES

This chapter presents research conducted by Jeffrey Covington under the guidance of Nan Chen, Steve Wiggins, and Evelyn Lunasin.

3.1 Introduction

Mesoscale eddies are a major component of the Earth's ocean and climate systems. For example, mesoscale eddies are important in momentum, heat, and mass transfer in the ocean [26, 36]. Identification and tracking of ocean eddies is therefore a vital component in the study of ocean dynamics and of the climate system as a whole. The majority of eddy identification methods utilize high-resolution satellite altimetry data which have provided global datasets of eddy tracks. However, in the arctic regions, sea ice interferes with satellite altimetry. In addition, in-situ measurements are sparse and expensive in the arctic. Because of this, eddy identification in the arctic is generally restricted to ice-free regions and depends on the seasonality of sea ice. In order to extend the capabilities of eddy identification methods in the arctic, satellite observations of free floating sea ice floe trajectories can be used. However, utilizing observations of ice floe trajectories is very different from utilizing satellite altimetry data, most notably in accounting for the inherent uncertainty from estimates made from sea ice data. The objective of the chapter is to build a framework for utilizing ice floe trajectories for eddy identification and for applying

eddy diagnostics under uncertainty in the ocean velocity field.

In order to perform eddy identification it is necessary to estimate the ocean velocity field. Data assimilation is used to accomplish this from observations of ice floe trajectories. Applying data assimilation combines the observed trajectories with a coupled ice-ocean model to estimate the unobserved ocean state. This provides not just a mean estimate of the ocean state, but also the associated uncertainty. Due to the nature of ice floe observations, a certain level of uncertainty is always present. Therefore, it is necessary to consider traditional eddy identification methods in the context of this uncertainty.

Section 3.2 presents a coupled ice-ocean model that incorporates complex ice floe dynamics including the highly nonlinear contact forces between floes. Section 3.4 presents a data assimilation framework that can utilize the observed floe trajectories along with the highly nonlinear ice-ocean model to optimally estimate the unobserved ocean state in a computationally efficient way. Section 3.5 discusses eddy identification and considers two eddy diagnostic tools. The identification of eddies under uncertainty from the data assimilation is explored.

3.2 Sea ice model

Ice floes are formed when ice sheets break up. Especially during the summer, ice floes can be free floating, driven by the local characteristics of the ocean. In order to utilize satellite observations of ice floe trajectories for data assimilation it is necessary to incorporate a model which can account for intricate ice floe dynamics. In this

section we present a model of ice floe dynamics which captures complex sea ice and ice-ocean interactions which, despite its high-dimensionality and nonlinearity, can be used in data assimilation applications.

For this ice floe model [42, 29] the ice floes are assumed to have cylindrical shape with uniform thickness. The ice floes are subject to ocean drag forces in a one-way interaction where the ocean drag force on the ice floes is calculated using a quadratic drag approximation. Further, the ocean forces and torques acting on the floes are assumed to be uniform over the shape of the floe, allowing for the forces to be explicitly calculated without the need to calculate surface integrals in the floe dynamics. Finally, the contact forces between floes can be approximated by taking advantage of the cylindrical floe shapes. Despite these simplifications, this model can capture many of the rich features of sea ice dynamics that are necessary to account for in a real observational setting.

3.3 Ice floe dynamics

We consider an idealized discrete element method (DEM) model of individual sea ice floes. In this model [42] each floe has an index $\ell = 1, \ldots, L$ and is characterized by the floe position at the centroid, $\mathbf{x}^\ell = (x^\ell, y^\ell)^\mathsf{T}$, its translational velocity $\mathbf{u}^\ell = (u^\ell, v^\ell)^\mathsf{T}$, its rotational orientation relative to the initial condition, Ω^ℓ, and its angular velocity ω^ℓ. Each floe also has two associated parameters, the floe radius R_ℓ, and the floe thickness h_ℓ. Using the ice density, ρ_{ice}, and ocean density, ρ_{ocean}, the mass and moment of intertia of each floe can be calculated. The governing equations for

the floe dynamics are then given by Newton's second law

$$\frac{d\mathbf{x}^\ell}{dt} = \mathbf{u}^\ell \tag{3.1}$$

$$m_\ell \frac{d\mathbf{u}^\ell}{dt} = \mathbf{F}^\ell_{\text{contact}} + \mathbf{F}^\ell_{\text{ocean}} \tag{3.2}$$

$$\frac{d\Omega^\ell}{dt} = \omega^\ell \tag{3.3}$$

$$I_\ell \frac{d\omega^\ell}{dt} = \tau^\ell_{\text{contact}} + \tau^\ell_{\text{ocean}} \tag{3.4}$$

where $\mathbf{F}^\ell_{\text{contact}}$, $\mathbf{F}^\ell_{\text{ocean}}$, $\tau^\ell_{\text{contact}}$, and τ^ℓ_{ocean} are the forces and torques induced by the ocean drag and contact forces between floes. The mass of each floe is given by $m_\ell = \rho_{\text{ice}} h_\ell \pi R_\ell^2$ and the moment of inertia is given by $I = \rho_{\text{ice}} h_\ell \pi R_\ell^4$. The ocean forces and torques are given in equations (3.30) and (3.31) detailed in section 3.3. The total contact forces and torques between ice floes are calculated in equations (3.8) and (3.12) detailed in section 3.3.

Ocean drag

The ice-ocean interaction of the model occurs through the ocean drag, where the drag force acts to bring the floe velocity and angular velocity closer to the ocean velocity and vorticity. Note that this can speed up a floe in addition to slowing a floe down.

The drag force and torque induced by the ocean on the ℓth floe are calculated using a quadratic drag approximation

$$\mathbf{F}^\ell_{\text{ocean}}(\mathbf{x}, t) = \alpha_\ell \left(\mathbf{u}_o(\mathbf{x}, t) - \mathbf{u}^\ell\right) \left|\mathbf{u}_o(\mathbf{x}, t) - \mathbf{u}^\ell\right| \tag{3.5}$$

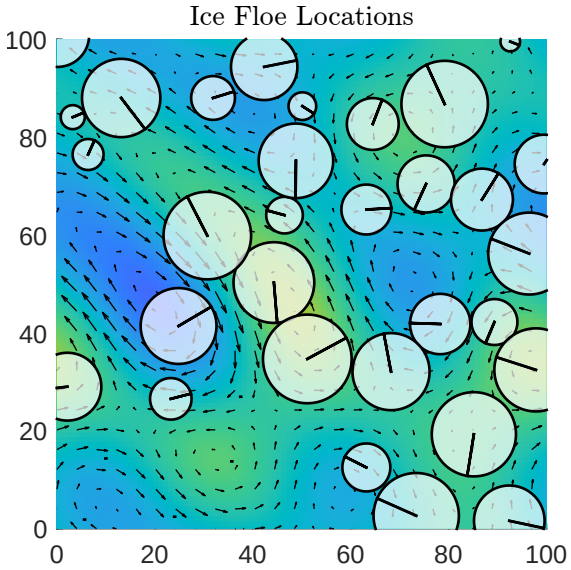

Figure 3.1: A snapshot of the ice floe model. The floes in the model are cylindrical. A line is plotted to indicate the orientation of the floe. The underlying ocean velocity field and stream function is also plotted.

and

$$\tau^{\ell}_{\text{ocean}}(\mathbf{x}, t) = \beta_{\ell} \left(\frac{\nabla \times \mathbf{u}_o(\mathbf{x}, t)}{2} - \omega \right) \left| \frac{\nabla \times \mathbf{u}_o(\mathbf{x}, t)}{2} - \omega \right|. \tag{3.6}$$

The ocean velocity, $\mathbf{u}_o$, is given by equation (3.14). Here $\nabla \times \mathbf{u}_o = \frac{\partial v_o}{\partial x} - \frac{\partial u_o}{\partial y}$ is the ocean vorticity, which is twice the angular velocity, and is given by equation (3.15). The coefficients $\alpha_{\ell} = d_{ocn} \rho_{ocn} \pi R_{\ell}^2$ and $\beta_{\ell} = d_{ocn} \rho_{ocn} \pi R_{\ell}^4$ incorporate the ocean drag coefficient, d_{ocn}, the ocean density ρ_{ocn}, and the floe cross-sectional area πR^2. The coefficient β also has an extra factor of R^2 which comes from integrating the torque over the floe area, whose integrand depends on the radius from the centroid.

Floe contact forces

The total contact force on the ℓth floe is calculated from the sum of the contact forces between the floe and all other floes

$$\mathbf{F}^{\ell}_{\text{contact}} = \sum_{k=1}^{L} \left(\mathbf{F}^{\ell k}_{n} + \mathbf{F}^{\ell k}_{t} \right) \tag{3.7}$$

where the contact force from each other floe is divided into the normal component, $\mathbf{F}^{\ell k}_{n}$, and the tangential component, $\mathbf{F}^{\ell k}_{t}$. The contact forces are only nonzero if the floes are overlapping, that is if the distance between the floes is less than the sum of the floe radii.

The normal component of the contact force follows Hooke's law

$$\mathbf{F}^{\ell k}_{n} = E c \mathbf{n} \tag{3.8}$$

where E is Young's modulus, c is the length of the chord of intersection calculated by equation (3.10), and $\mathbf{n}$ is the unit normal direction. Due to the cylindrical shape of the floes the normal direction is parallel to the vector between the floe centroids.

The force in the tangential direction given by

$$\mathbf{F}^{\ell k}_{t} = G c v_{t} \mathbf{t} \tag{3.9}$$

where G is the shear modulus, v_t is the difference in tangential floe velocity, and $\mathbf{t}$

is the tangential direction. The length of the chord, c, can be calculated by

$$c = \frac{1}{d}\sqrt{4d^2R_{max}^2 - \left(d^2 - R_{min}^2 + R_{max}^2\right)^2} \tag{3.10}$$

where d is the distance between the floes, R_{max} is the larger of the two floe radii, and R_{min} is the smaller. The difference in tangential velocity is given by

$$v_t = (\mathbf{u}^k - \mathbf{u}^\ell) \cdot \mathbf{t} + (R_k\omega^k + R_\ell\omega^\ell) \tag{3.11}$$

where $\mathbf{u}^k$, ω^k, and R_k are the velocity, angular velocity, and radius of the other floe respectively, $\mathbf{t}$ is the tangential direction perpendicular to the normal direction, and "$\cdot$" is the dot product. Note that the difference in tangential velocities depends on the translational velocities of the floes as well as the angular velocities of the floes. The tangential velocity component due to the angular velocity is proportional to the floe radii. Also note that while (3.11) takes the difference in the floe velocities, the tangential velocity of floe k is opposite in sign from the perspective of floe ℓ, hence there is a sum instead of a difference in the second term.

The total contact torque acting on the ℓth floe is calculated from the tangential contact force as

$$\tau_{contact} = \sum_{k=1}^{\ell} R_\ell f_t^{\ell k} \tag{3.12}$$

where $F_t^{\ell k} = f_t^{\ell k}\mathbf{t}$.

Ocean model

The ocean model is incompressible and periodic on the domain $[0, x_{max})^2$. The ocean state is characterized by its Fourier coefficients, $\hat{u}_{\mathbf{k}}(t)$, where $\mathbf{k} = (k_1, k_2)^T \in \mathbb{Z}^2$ is the two-dimensional wave number and the coefficients satisfy $\hat{u}_{-\mathbf{k}} = \overline{\hat{u}_{\mathbf{k}}}$, the complex conjugate. The real-valued stream function, ψ, is given by

$$\psi(\mathbf{x}, t) = \sum_{\mathbf{k} \in \mathbb{Z}^2} \hat{u}_{\mathbf{k}}(t) \exp\left(-\frac{2\pi i}{x_{max}}(\mathbf{k} \cdot \mathbf{x})\right) \tag{3.13}$$

where the level sets of the stream function are the streamlines of the velocity field, so the corresponding ocean velocity is given by $\mathbf{u}_o = (u_o, v_o)^T = \left(\frac{\partial \psi}{\partial y}, -\frac{\partial \psi}{\partial x}\right)$. From the Fourier coefficients the ocean velocity is reconstructed as

$$\mathbf{u}_o(\mathbf{x}, t) = \sum_{\mathbf{k} \in \mathbb{Z}^2} \hat{u}_{\mathbf{k}}(t) \exp\left(-\frac{2\pi i}{x_{max}}(\mathbf{k} \cdot \mathbf{x})\right) \mathbf{r}_{\mathbf{k}} \tag{3.14}$$

with eigenvectors $\mathbf{r}_{\mathbf{k}} = \left(\frac{2\pi i k_2}{x_{max}}, -\frac{2\pi i k_1}{x_{max}}\right)^T$. In addition, the ocean vorticity, $\nabla \times \mathbf{u}_o = \frac{\partial v_o}{\partial x} - \frac{\partial u_o}{\partial y}$, is given by

$$\nabla \times \mathbf{u}_o(\mathbf{x}, t) = \sum_{\mathbf{k} \in \mathbb{Z}^2} \hat{u}_{\mathbf{k}}(t) \exp\left(-\frac{2\pi i}{x_{max}}(\mathbf{k} \cdot \mathbf{x})\right) \left(\frac{2\pi |\mathbf{k}|}{x_{max}}\right)^2. \tag{3.15}$$

The Fourier coefficients, $\hat{u}_{\mathbf{k}}(t)$, can depend arbitrarily on t. Here we consider a stochastic ocean model where the Fourier coefficients are governed by independent

Ornstein-Uhlenbeck (OU) processes

$$\frac{d\hat{u}_{\mathbf{k}}}{dt} = (-a_{\mathbf{k}} + \phi_{\mathbf{k}} i)\hat{u}_{\mathbf{k}} + f_{\mathbf{k}} + \sigma_{\mathbf{k}} \dot{W}_{\mathbf{k}}(t) \tag{3.16}$$

with complex Gaussian white noise $\dot{W}_{\mathbf{k}}$ and real parameters $\sigma_{\mathbf{k}} > 0$, $a_{\mathbf{k}}$, $\phi_{\mathbf{k}}$, and $f_{\mathbf{k}}$. Note that even though the OU processes for the Fourier coefficients given in equation (3.16) are independent, the resulting velocity field in physical space still has spatial correlations. In order to ensure that the ocean velocity is real-valued it is necessary that $\hat{u}_{-\mathbf{k}} = \overline{\hat{u}_{\mathbf{k}}}$, so the parameters and noise have the additional restriction that $a_{-\mathbf{k}} = a_{\mathbf{k}}$, $\phi_{-\mathbf{k}} = -\phi_{\mathbf{k}}$, $f_{-\mathbf{k}} = \overline{f_{\mathbf{k}}}$, and $\sigma_{-\mathbf{k}} = \sigma_{\mathbf{k}}$ with $\dot{W}_{-\mathbf{k}} = \overline{\dot{W}_{\mathbf{k}}}$.

Parameter fitting based on equilibrium statistics

The equilibrium mean, covariance, and decorrelation time of each independent OU process given in (3.16) can be written explicitly in terms of the parameters $a_{\mathbf{k}}$, $\phi_{\mathbf{k}}$, $f_{\mathbf{k}}$, and $\sigma_{\mathbf{k}}$ as

$$\text{Mean}(\hat{u}_{\mathbf{k}}) = \frac{f_{\mathbf{k}}}{a_{\mathbf{k}} - i\phi_{\mathbf{k}}} \qquad \text{Var}(\hat{u}_{\mathbf{k}}) = \frac{\sigma_{\mathbf{k}}^2}{2a_{\mathbf{k}}} \qquad T_{\text{corr}} = \frac{1}{a_{\mathbf{k}} - i\phi_{\mathbf{k}}} \tag{3.17}$$

where the decorrelation time is defined as

$$T_{\text{corr}} = \int_0^\infty \frac{\mathbb{E}[(\hat{u}_{\mathbf{k}}(0) - \text{Mean}(\hat{u}_{\mathbf{k}}))(\hat{u}_{\mathbf{k}}(t) - \text{Mean}(\hat{u}_{\mathbf{k}}))]}{\text{Var}(\hat{u}_{\mathbf{k}})}\, dt. \tag{3.18}$$

Inversely, the equations given in (3.17) can be inverted to write the parameters in terms of a given a set of equilibrium statistics

$$a_{\mathbf{k}} = \mathrm{Re}\left[\frac{1}{T_{\mathrm{corr}}}\right] \qquad \phi_{\mathbf{k}} = -\,\mathrm{Im}\left[\frac{1}{T_{\mathrm{corr}}}\right] \tag{3.19}$$

$$f_{\mathbf{k}} = \frac{\mathrm{Mean}(\hat{u}_{\mathbf{k}})}{T_{\mathrm{corr}}} \qquad \sigma_{\mathbf{k}} = \sqrt{2\,\mathrm{Var}(\hat{u}_{\mathbf{k}})\,\mathrm{Re}\left[\frac{1}{T_{\mathrm{corr}}}\right]}. \tag{3.20}$$

In this way the parameters can be systematically calibrated to a specified equilibrium mean, variance, and decorrelation time, allowing the model to have specified properties or for the model to be tuned to data, either to real data or a free run of sophisticated ocean model.

3.4 Lagrangian data assimilation of ice floe observations

In order to utilize observations of ice floe trajectories for eddy identification, we apply data assimilation to incorporate the observations with a coupled ice-ocean model to estimate the underlying ocean state. Data assimilation provides an estimate of the underlying ocean through the posterior distribution, which is the probability distribution of the ocean state conditioned on the ice floe observations, in a Bayesian sense. There are a number of advantages to using data assimilation for eddy identification. First, remotely-sensed measurements of the ocean, including ice floe observations, are inherently noisy. Many eddy identification diagnostics are sensitive to noise, requiring various preprocessing methods. Data assimilation,

which takes a probabilistic viewpoint, naturally accounts for observational noise in the system. Second, the posterior distribution provides not just a point estimate of the system state, but also the uncertainty in the estimate. This is especially important in the case of ice floe measurements where the spatial sparsity of observations leads to large uncertainties in the estimated ocean state. Rather than only relying on the point-estimate of the ocean, the uncertainty quantification provided by data assimilation can be used to account for the uncertainty appropriately.

This is an example of Lagrangian data assimilation, where the observations, in this case the ice floe trajectories, are given in Lagrangian coordinates. The nonlinear translation between the Lagrangian coordinates of the observations and the Eulerian coordinates of the velocity field poses a challenge to many traditional data assimilation methods [7]. Because of this nonlinearity, this study utilizes the conditional Gaussian (CG) framework for data assimilation [32]. The CG framework applies to a wide variety of nonlinear models in geophysics and has been applied to Lagrangian data assimilation [29]. In addition, the CG framework has a number of advantageous properties. First, the posterior distribution is Gaussian, despite the nonlinearity, and so the posterior distribution is fully characterized by the posterior mean and covariance, precluding the need for approximation of the posterior distribution with an ensemble. Second, the posterior mean and covariance are given through analytic formulae, making the method computationally efficient. Third, the CG framework allows not just for the sampling of the posterior distribution at a fixed time instant, but also provides a method for sampling entire model trajectories conditioned on the observations. This is important for eddy diagnostics like the

Lagrangian descriptor, which utilizes the temporal component of the velocity field.

Conditional Gaussian data assimilation

A dynamical system falls under the conditional Gaussian (CG) framework if it can be written in the following form:

$$\frac{d\mathbf{X}}{dt} = \mathbf{A}_0(\mathbf{X}, t) + \mathbf{A}_1(\mathbf{X}, t)\mathbf{Y} + \mathbf{B}(\mathbf{X}, t)\dot{\mathbf{W}}_{\mathbf{X}}(t), \tag{3.21}$$

$$\frac{d\mathbf{Y}}{dt} = \mathbf{a}_0(\mathbf{X}, t) + \mathbf{a}_1(\mathbf{X}, t)\mathbf{Y} + \mathbf{b}(\mathbf{X}, t)\dot{\mathbf{W}}_{\mathbf{Y}}(t) \tag{3.22}$$

where $\mathbf{X}$ is the state vector of observed variables and $\mathbf{Y}$ is the state vector of unobserved variables. Here, $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{B}$, $\mathbf{a}_0$, $\mathbf{a}_1$, and $\mathbf{B}$ are matrices that can depend on t and $\mathbf{X}$ arbitrarily nonlinearly. While this dependence is always assumed, it may not be denoted in the subsequent discussion for notational efficiency. $\dot{\mathbf{W}}_{\mathbf{X}}$ and $\dot{\mathbf{W}}_{\mathbf{Y}}$ are Gaussian white noise.

Because of the potentially highly nonlinear interactions with the observed variables, the CG framework applies to a wide range of nonlinear systems and an even wider range of systems have suitable nonlinear CG approximations. However, despite the nonlinearity, the coefficient matrices only depend on the observed variables, $\mathbf{X}$, and do not depend on the unobserved variables, $\mathbf{Y}$, and so the system is linear in $\mathbf{Y}$ given $\mathbf{X}$. This property ensures that the conditional distribution of $\mathbf{Y}$ given a trajectory of $\mathbf{X}$ is Gaussian as the name implies. Because of this property the posterior distributions for filtering and smoothing are characterized by their posterior mean vectors and covariance matrices which are given by analytic

formulae.

The posterior distribution of $\mathbf{Y}(t)$ at time t given a trajectory of $\mathbf{X}(s)$ over the interval $[0, t]$ is known as the filter posterior distribution and is given by

$$p(\mathbf{Y}(t) \mid \mathbf{X}(s), s \leqslant t) \sim \mathcal{N}(\boldsymbol{\mu}_f(t), \mathbf{R}_f(t)) \tag{3.23}$$

where $\boldsymbol{\mu}_f$ and $\mathbf{R}_f$ are the filter mean and covariance respectively. The filter mean and covariance can be calculated using the forward equations

$$\frac{d\boldsymbol{\mu}_f}{dt} = (\mathbf{a}_0 + \mathbf{a}_1\boldsymbol{\mu}_f) + (\mathbf{R}_f\mathbf{A}_1^*)(\mathbf{B}\mathbf{B}^*)^{-1}\left(\frac{d\mathbf{X}}{dt} - (\mathbf{A}_0 + \mathbf{A}_1\boldsymbol{\mu}_f)\right) \tag{3.24}$$

$$\frac{d\mathbf{R}_f}{dt} = \mathbf{a}_1\mathbf{R}_f + \mathbf{R}_f\mathbf{a}_1^* + \mathbf{b}\mathbf{b}^* - (\mathbf{R}_f\mathbf{A}_1^*)(\mathbf{B}\mathbf{B}^*)^{-1}(\mathbf{A}_1\mathbf{R}_f) \tag{3.25}$$

where the initial condition is a Gaussian distribution with mean $\boldsymbol{\mu}_f(0)$ and covariance $\mathbf{R}_f(0)$. Here "$\cdot^*$" denotes the conjugate transpose.

The posterior distribution of $\mathbf{Y}(t)$ at time $t \in [0, T]$ given a trajectory of $\mathbf{X}(s)$ over the entire interval $[0, T]$ is known as the smoother posterior distribution [33] and is given by

$$p(\mathbf{Y}(t) \mid \mathbf{X}(s), s \in [0, T]) \sim \mathcal{N}(\boldsymbol{\mu}_s(t), \mathbf{R}_s(t)) \tag{3.26}$$

where $\boldsymbol{\mu}_s$ and $\mathbf{R}_s$ are the smoother mean and covariance respectively. Compared to the filter posterior distribution, the smoother posterior distribution incorporates both past and future observational observation.

To calculate the smoother mean and covariance, first the filter mean and covariance are calculated. The condition of the smoother mean and covariance equations

at the final time T is given by the filter mean and covariance at time T, that is $(\boldsymbol{\mu}_s(t), \mathbf{R}_s(t)) = (\boldsymbol{\mu}_f(t), \mathbf{R}_f(t))$. The following equations are then integrated backwards in time from time T to time 0:

$$\frac{d\boldsymbol{\mu}_s}{dt} = -\mathbf{a}_0 - \mathbf{a}_1\boldsymbol{\mu}_s + (\mathbf{b}\mathbf{b}^*)\mathbf{R}_f^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_s) \tag{3.27}$$

$$\frac{d\mathbf{R}_s}{dt} = -(\mathbf{a}_1 + (\mathbf{b}\mathbf{b}^*)\mathbf{R}_f^{-1})\mathbf{R}_s - \mathbf{R}_s(\mathbf{a}_1^* + (\mathbf{b}\mathbf{b}^*)\mathbf{R}_f^{-1}) + \mathbf{b}\mathbf{b}^*. \tag{3.28}$$

The smoother equations calculate the matrix inversion $\mathbf{R}_f^{-1}$ which can be computationally inconvenient for high dimensional systems. While $\mathbf{R}_f$ is full in general, an approximation can be used where the full matrix $\mathbf{R}_f$ is used in equations (3.24) and (3.25) but only the diagonal entries are saved and used to calculate $\mathbf{R}_f^{-1}$ in equations (3.27) and (3.28). In addition, if only the smoother mean is required, equation (3.28) does not need to be calculated as equation (3.27) does not depend on $\mathbf{R}_s$.

While the smoother mean and covariance can be used to sample $\mathbf{Y}$ at a fixed time instant, the CG framework can also be used to sample entire trajectories of $\mathbf{Y}$ conditioned on $\mathbf{X}$ on the interval $[0, T]$. Such sampled trajectories are important for calculating eddy diagnostics such as the Lagrangian descriptor, which incorporate temporal information from the velocity field. Sampling these trajectories is accomplished via the backward sampling equation where an initial $\mathbf{Y}(T)$ is drawn from $\mathbf{Y}(T) \sim \mathcal{N}(\boldsymbol{\mu}_f(T), \mathbf{R}_f(T))$ and then its trajectory is calculated using the following

stochastic equation integrated backward in time

$$\frac{d\mathbf{Y}}{dt} = -\mathbf{a}_0 - \mathbf{a}_1\mathbf{Y} + (\mathbf{b}\mathbf{b}^*)\mathbf{R}_f^{-1}(\boldsymbol{\mu}_f - \mathbf{Y}) + \mathbf{b}\dot{\mathbf{W}}_{\mathbf{Y}}(t). \tag{3.29}$$

Note that when sampling multiple trajectories of $\mathbf{Y}$ conditioned on $\mathbf{X}$, the same filter mean and covariance are used in equation (3.29) and the dynamics of the sampled trajectories differ only in the sampled initial condition at time T and the realizations of the Gaussian white noise $\dot{\mathbf{W}}_{\mathbf{Y}}(t)$. In particular $\mathbf{R}^{-1}$ can be calculated once on the interval $[0, T]$ and then reused to calculate each sampled trajectory. The same diagonal approximation for $\mathbf{R}_f^{-1}$ that can be used for the smoother equations is also suitable for the backward sampling equation.

Data assimilation of the sea ice model

The sea ice model considered in section 3.2 along with observations of the floe trajectories nearly falls within the conditional Gaussian framework, where the observed variables, $\mathbf{X}$, consist of the floe positions and angular displacements, and the unobserved variables, $\mathbf{Y}$, consists of the floe velocities and the Fourier coefficients of the ocean. Notably the highly nonlinear contact forces depend nonlinearly only on the observed variables.

The quadratic ocean drag used in section 3.3 means that the model does not strictly fall under the conditional Gaussian framework. One solution is to use linear

drag, where equations (3.30) and (3.31) are replaced by linear drag

$$\mathbf{F}^{\ell}_{\text{ocean}}(\mathbf{x}, t) = \alpha_{\ell} \left(\mathbf{u}_o(\mathbf{x}, t) - \mathbf{u}^{\ell} \right) \tag{3.30}$$

and

$$\tau^{\ell}_{\text{ocean}}(\mathbf{x}, t) = \beta_{\ell} \left(\frac{\nabla \times \mathbf{u}_o(\mathbf{x}, t)}{2} - \omega \right). \tag{3.31}$$

The coefficients $\alpha_{\ell} = d_{ocn}\rho_{ocn}\pi R_{\ell}^2$ and $\beta_{\ell} = d_{ocn}\rho_{ocn}\pi R_{\ell}^4$ remain the same as in the quadratic drag case.

However, it is not necessary to linearize the floe dynamics. Instead, a highly accurate approximation can be used where the dynamics are linearized about the posterior mean $\boldsymbol{\mu}_f$. To see this consider a variation of equations (3.32) and (3.33) that considers a nonlinear coupling between $\mathbf{X}$ and $\mathbf{Y}$

$$\frac{d\mathbf{X}}{dt} = \mathbf{A}_0(\mathbf{X}, t) + \mathbf{A}_1(\mathbf{X}, t)\mathbf{Y} + \mathbf{B}(\mathbf{X}, t)\dot{\mathbf{W}}_{\mathbf{X}}(t), \tag{3.32}$$

$$\frac{d\mathbf{Y}}{dt} = \mathbf{F}(\mathbf{X}, \mathbf{Y}, t) + \mathbf{b}(\mathbf{X}, t)\dot{\mathbf{W}}_{\mathbf{Y}}(t) \tag{3.33}$$

The nonlinear function $\mathbf{F}$ can then be lienarized about $\boldsymbol{\mu}_f$ as

$$\mathbf{F}(\mathbf{X}, \mathbf{Y}, t) \approx \mathbf{F}(\boldsymbol{\mu}_f, \mathbf{Y}, t) + \mathbf{J}_{\mathbf{Y}}(\boldsymbol{\mu}_f, \mathbf{Y}, t)(\mathbf{Y} - \boldsymbol{\mu}_f) \tag{3.34}$$

where $\mathbf{J}_{\mathbf{Y}}$ is the Jacobian with respect to $\mathbf{Y}$.

Traditional Lagrangian data assimilation utilizes massless tracers which are analogous to ocean drifters. It has been shown that with a sufficient number of tracers the ocean can be estimated to arbitrary precision [34]. However, there are

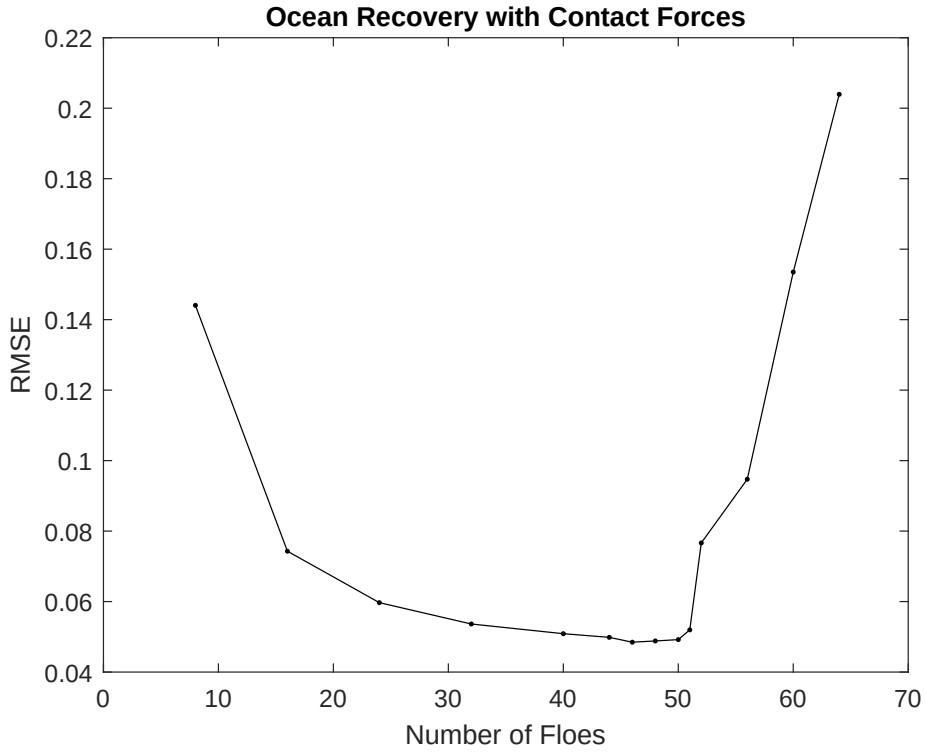

Figure 3.2: The Root Mean Squared Error (RMSE) of the ocean recovery under contact forces by the number of observed floe trajectories. The radius of all floes is 6 km so that the overall area covered by ice is proportional to number of floes. The error generally decreases as the number of floes increases. However, once a density of floes is reached the contact forces interfere with the recovery of the ocean.

important differences between floe and tracer dynamics. For one, as the number of floes increases, the influence of contact forces between floes also increases. Figure 3.2 shows that at a certain threshold of ice concentration, the contact forces dominate over the ocean drag forces, severely reducing the effectiveness of the ocean estimation. Other important differences include the influence of floe momentum, which reduces the impact of short time scales in the ocean velocity field. In a real

observational scenario there is also the influence of model error. These differences mean that utilizing floe observations always entails significant uncertainty in the posterior distribution. This is contrary to the assumptions normally made in eddy identification, where high resolution data is typically available.

3.5 Eddy diagnostics

There are many methods for characterizing and automatically identifying oceanic eddies [128, 47]. Broadly speaking, eddy diagnostic methods are based on the 2D ocean velocity field,

$$\mathbf{u}_o(\mathbf{x}, t) = (u_o(\mathbf{x}, t), v_o(\mathbf{x}, t))^T, \tag{3.35}$$

which is estimated from observations. Two methods we consider here are the widely used Okubo-Weiss parameter and a Lagrangian descriptor based on the modulus of vorticity. In addition to the methods considered here, there are many other approaches to eddy identification such as ones based on the geometry of the velocity field [111], wavelet analysis of the Sea Surface Height (SSH) [43], the geometry of the SSH [25, 51], finite time Lyapunov exponents (FTLE) [113] and Lagrangian-averaged vorticity deviation (LAVD) to name a few.

Okubo-Weiss parameter

A classical, Eulerian, approach to eddy identification is the Okubo-Weiss Parameter [114, 145]. The Okubo-Weiss parameter is defined by

$$\mathrm{OW} = s_n^2 + s_s^2 - \omega^2 \tag{3.36}$$

where the normal strain, the shear strain, and the relative vorticity are given by

$$s_n = \frac{\partial u_o}{\partial x} - \frac{\partial v_o}{\partial y}, \qquad s_s = \frac{\partial v_o}{\partial x} + \frac{\partial u_o}{\partial y}, \qquad \omega = \frac{\partial v_o}{\partial x} - \frac{\partial u_o}{\partial y} \tag{3.37}$$

respectively. When the Okubo-Weiss parameter is negative, the relative vorticity is larger than the strain components, indicating vortical flow. The Okubo-Weiss parameter is widely used in part thanks to its physical interpretability. The Okubo-Weiss parameter is an example of an Eulerian quantity, based solely on a snapshot of the ocean velocity field.

In the OW parameter framework, eddies are characterized by enclosed areas where the relative vorticity dominates over the strain components. This is normally defined as being when the OW parameter is below a certain (negative) threshold. The precise choice of this threshold partially subjective and region-dependent, but a commonly chosen threshold is $-0.2\sigma_{\mathrm{OW}}$ where σ_{OW} is the standard deviation of the OW parameter. Figure 3.3 shows an example of the OW parameter for a certain velocity field along with the eddies as defined by the framework.

In practice the OW parameter is sensitive to observational noise. For example, when using SSH to estimate the velocity field, calculating the OW parameter in-

volves taking two derivatives of SSH [25, 36]. Because of this the OW parameter is often smoothed before using for eddy identification [54].

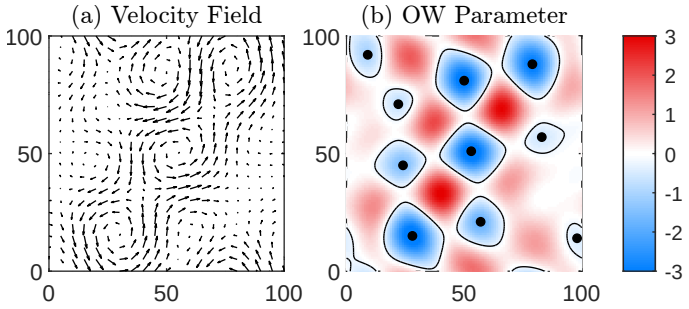

Figure 3.3: Panel (a) shows a particular velocity field and panel (b) shows the associated Okubo-Weiss (OW) parameter as well as the eddies identified by the OW criteria. Positive (red) values of OW indicate strain-dominated regions while negative (blue) values indicate vorticity dominated regions. Here the OW parameter is normalized so it has a standard deviation of 1 and the threshold value for eddy identification was chosen to be $0.2\sigma_{OW}$.

Lagrangian descriptor

Unlike the OW parameter, which considers only an instantaneous snapshot of the velocity field, the Lagrangian descriptor incorporates the time-dependent component of the velocity field [106].

The Lagrangian descriptor based on the modulus of vorticity [141, 142] is defined as

$$M_V(\mathbf{x}^*, t^*)_\tau = \int_{t^*-\tau}^{t^*+\tau} W(\mathbf{x}(t), t)\, dt = \int_{t^*-\tau}^{t^*+\tau} \left| \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right| dt \tag{3.38}$$

where the modulus of vorticity is

$$W(\mathbf{x}, t) = |\nabla \times \mathbf{u}_o(\mathbf{x}, t)| \tag{3.39}$$

and $\nabla \times \mathbf{u}_o$ denotes the ocean vorticity. The parameter τ indicates the time window over which equation (3.38) is integrated. In Equation (3.38), $\mathbf{x}(t)$ is the trajectory of a massless Lagrangian tracer such that $\mathbf{x}(t^*) = \mathbf{x}^*$ and

$$\frac{d\mathbf{x}}{dt} = \mathbf{u}_o(\mathbf{x}, t) \tag{3.40}$$

where $\mathbf{u}_o = (u_o, v_o)^T$ is the velocity of the ocean.

Note that while the modulus of vorticity given in equation (3.39) is an Eulerian quantity only depending on a fixed snapshot of the ocean, the Lagrangian descriptor given in equation (3.38) incorporates both spatial and temporal information from the ocean state, giving a comprehensive dynamical picture of the ocean.

Here an eddy is defined as a local maximum of M_V, the Eddy core, surrounded by the largest closed contour surrounding the Eddy core that only contains one local maximum. Figure 3.4 shows an example of the modulus of vorticity and the Lagrangian descriptor based on the modulus of vorticity, M_V.

Eddy identification under uncertainty

Lagrangian data assimilation gives the posterior distribution of the ocean state conditional on the observed ice floe trajectories. This posterior distribution can be then used to estimate the ocean velocity field, and hence can be used to apply various

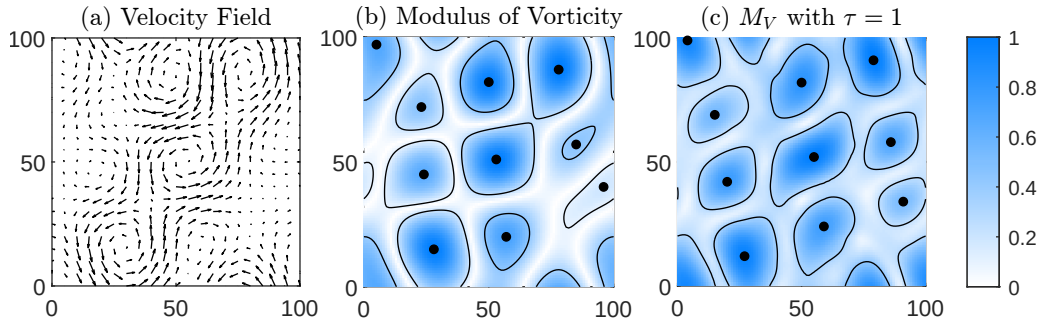

Figure 3.4: Panel (a) shows a particular velocity field. Panel (b) shows the associated modulus of vorticity as well as the eddies identified by the criteria. Panel (c) shows M_V, the Lagrangian descriptor based on the modulus of vorticity. The values are normalized to the range $[0, 1]$.

eddy diagnostics like the Okubo-Weiss parameter and the Lagrangian descriptor based on the modulus of vorticity. However, due to the inherent uncertainty associated with using ice floe observations, it is necessary to consider the entire posterior distribution of the ocean state rather than just the posterior mean. Therefore, it is important to consider the impact of uncertainty on the eddy diagnostics and eddy identification.

When calculating Eddy diagnostics, there is a distinction between calculating the diagnostic using the posterior mean of the ocean and calculating the expected value of the eddy diagnostic under the entire posterior distribution. Under the posterior distribution, the eddy diagnostic becomes a random variable from which its statistical expectation can be calculated. For example, the Okubo-Weiss parameter is a nonlinear function of the ocean state and therefore the expected value of the Okubo-Weiss parameter is not equal to the Okubo-Weiss parameter of the posterior mean ocean velocity in general. The expected value of the OW parameter can be

numerically calculated by sampling possible ocean states from the posterior distribution $\mathcal{N}(\boldsymbol{\mu}_s(t), \mathbf{R}_s(t))$ given by equations (3.27) and (3.28), calculating the OW parameter for each sampled ocean state, and then taking the mean of the sampled OW parameters. Figures 3.5 and 3.6 show the differences between these quantities for the OW parameter. The posterior mean of the ocean, especially under high uncertainty, has a lower amplitude than the true ocean. This is because for this ocean model the statistical equilibrium mean of the ocean is zero which is reflected most clearly in figure 3.6.

On the other hand the Lagrangian descriptor M_V is also a nonlinear function of the ocean state and its statistical expectation was considered in [31]. Because M_V is a Lagrangian quantity depending on dynamical information, in order to numerically calculate its expected value it is necessary to sample ocean trajectories rather than the ocean state at a fixed time. To accomplish this, the backward sampling formula given in equation (3.29) is used to sample ocean trajectories from which M_V can be calculated. The expectation of M_V is shown in figure 3.7 Under high uncertainty the expectation of the Lagrangian descriptor based on the modulus of vorticity, M_V, tends towards the equilibrium mean value of M_V which is nonzero.

It should be noted that while there is a canonical "true" ocean state used to generate the observed floe trajectories, the ocean state is not uniquely determined by the observations. The conditional Gaussian framework gives the optimal posterior distribution for the ocean state conditional on the observations, and so the true ocean state can be considered as being one sample from this posterior distribution. From this perspective, the statistical expectation of an eddy diagnostic is not an

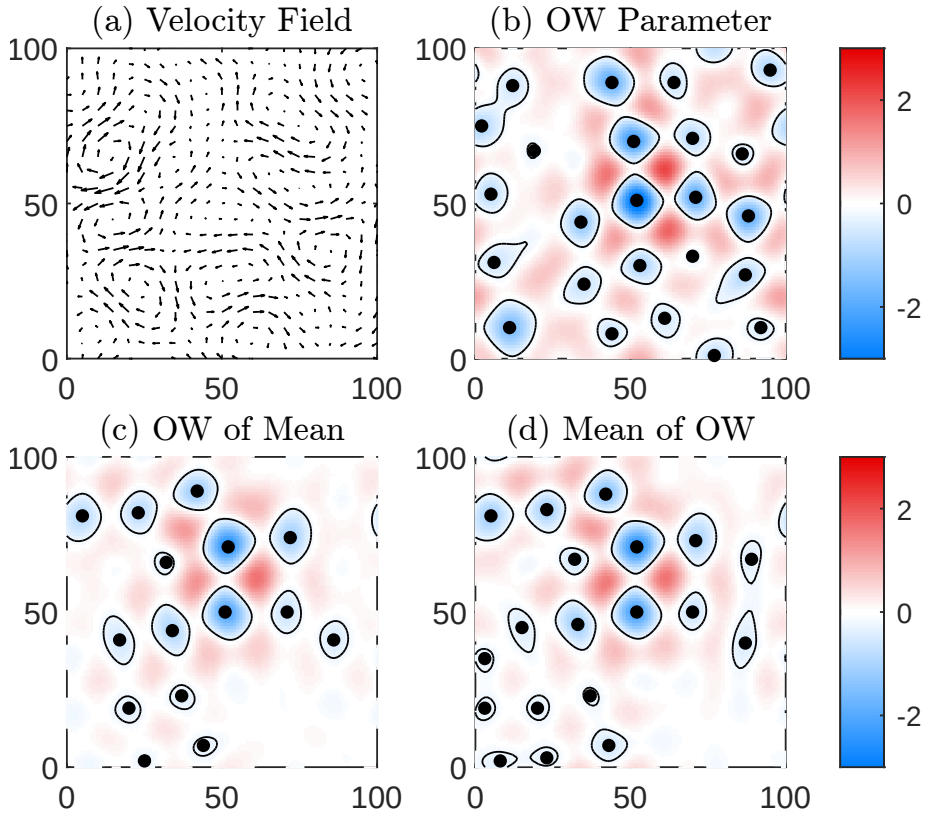

Figure 3.5: A comparison of using the Okubo-Weiss parameter under the posterior distribution of Lagrangian data assimilation using 4 observed floe trajectories. Panel (a) shows the true velocity field and panel (b) plots the associated Okubo-Weiss parameter and eddies. Panel (c) shows the Okubo-Weiss parameter calculated using the posterior mean of the ocean. Panel (d) shows the statistical expectation of the OW parameter under the entire posterior distribution. There are subtle differences between panels (c) and (d). In particular panel (d) shows several more eddies that meet the negative threshold used for eddy characterization.

estimation of the true state, but rather a description of the distribution of possible states.

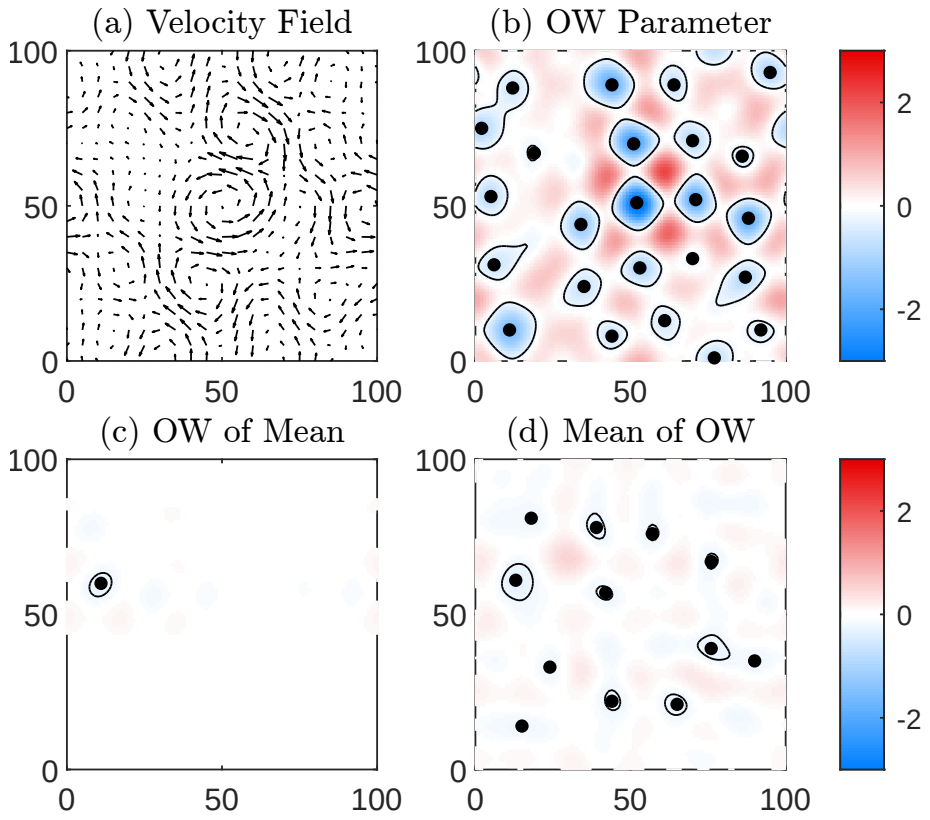

Figure 3.6: An example of the Okubo-Weiss parameter under high uncertainty. This is an extreme example where the ocean recovery is based on a single floe trajectory. Panel (a) shows the true velocity field with associated Okubo-Weiss parameter and identified eddies in panel (b).

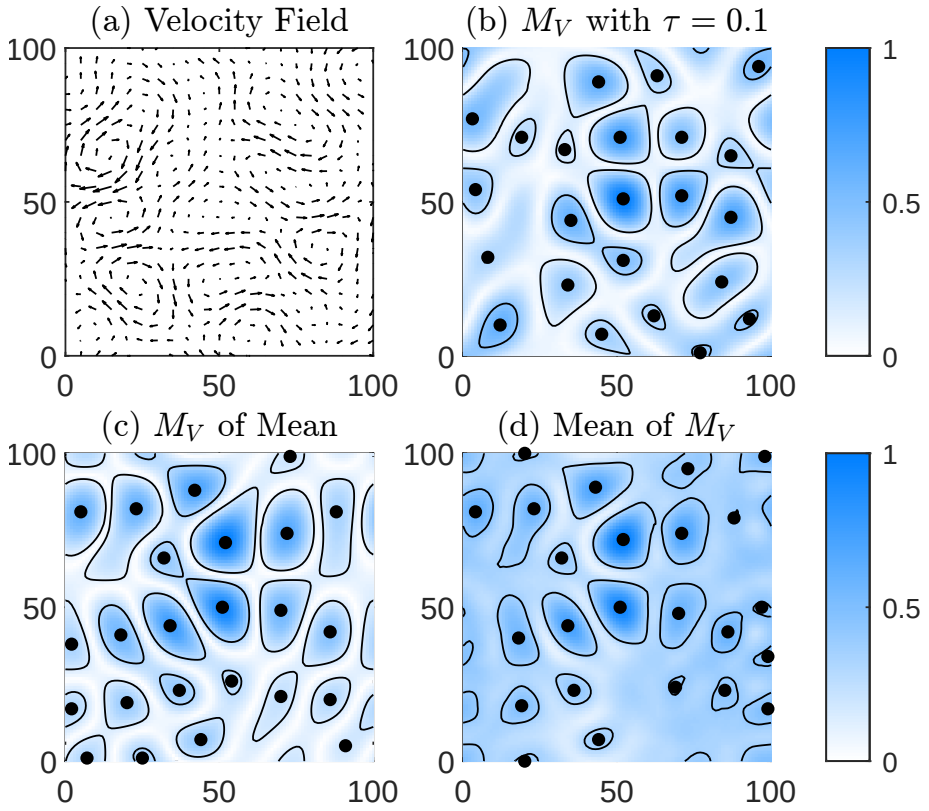

Figure 3.7: A comparison of using the Lagrangian descriptor based on the modulus of vorticity, M_B, under the posterior distribution of Lagrangian data assimilation using 4 observed floe trajectories. Panel (a) shows the true velocity field and panel (b) plots M_V and the eddies. Panel (c) shows M_V calculated using the posterior mean of the ocean. Panel (d) shows the statistical expectation of M_V under the entire posterior distribution. There are subtle differences between panels (c) and (d). Note that in regions of high uncertainty, the expectation of M_V tends towards a nonzero mean value.

3.6 Conclusion

While sea ice is typically an obstacle to eddy identification, this chapter presented a framework for utilizing ice floe trajectories for eddy identification. By utilizing Lagrangian data assimilation on a sea ice model, the underlying ocean state can be estimated from observed ice floe trajectories. However, the estimated ocean inherently contains uncertainty and so it is not enough to only utilize the mean estimate of the ocean state for eddy identification. Therefore, by sampling from the posterior distribution, traditional eddy identification were adapted to incorporate this uncertainty. Using this method the statistical expected value of two eddy diagnostics, the Okubo-Weiss parameter and the Lagrangian descriptor based on the modulus of vorticity, were calculated from the posterior distribution, giving the optimal point estimates of these diagnostics. In this way the probabilistic nature of utilizing floe observations can be naturally and rigorously incorporated into traditional eddy identification frameworks. This framework could play a key role in extending the capabilities of eddy identification further into the arctic region, allowing for a more comprehensive understanding of ocean dynamics in the arctic.

A APPENDIX TO BRIDING GAPS IN THE CLIMATE OBSERVATIONAL NETWORK

A.1 The coupled atmosphere-ice-ocean system

The DEM model

The sea ice floes are described by a DEM model. The floes can have arbitrary 2D shapes with their movements and rotations determined by the surface integrals of the ocean and wind velocities over these shapes. Both the full QG model and reanalysis data or and the approximate stochastic models can be used to drive this ice floe model. In the DEM model utilized here, the shape and thickness for each floe are assumed to be unchanging over time. Since the non-interacting floes are the primary focus of this work, the contact forces are not included in the model presented here, which greatly reduces the computational cost.

The dynamics of a single ice floe is described as follows [107, 30]. Let $\mathbf{X}_{\text{ice}} = (x_{\text{ice}}, y_{\text{ice}})$ be the centroid of the floe and Ω its the angular displacement about the centroid. Also denote $\mathbf{V}_{\text{ice}} = (u_{\text{ice}}, v_{\text{ice}})$ to be the velocity of the floe and ω is the angular velocity. Then ice floe-ocean interactions are calculated using surface

integrals over the area of the floe:

$$\dot{\mathbf{X}}_{\text{ice}} = \mathbf{V}_{\text{ice}} \tag{A.1}$$

$$\dot{\Omega} = \omega \tag{A.2}$$

$$m\dot{\mathbf{V}}_{\text{ice}} = \iint_A \mathbf{F}_{\text{total}}\, dA \tag{A.3}$$

$$I\dot{\omega} = \iint_A \tau\, dA \tag{A.4}$$

where A is the area of the floe. $\mathbf{F}_{\text{total}}$ is the total force on the ice floe induced by the ocean, atmosphere, and other sources. τ is the resulting torque, calculated from the force.

The total force on the floe by the ocean consists of ocean drag, atmosphere forcing, Coriolis force, and the pressure gradient

$$\mathbf{F}_{\text{total}} = \mathbf{F}_{\text{ocn}} + \mathbf{F}_{\text{atm}} + \mathbf{F}_{\text{Coriolis}} + \mathbf{F}_{\text{pressure}}. \tag{A.5}$$

To incorporate the ocean turning angle, define the rotation matrix $\mathbf{R}_\theta$ as

$$\mathbf{R}_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{A.6}$$

The force induced by the ocean drag at the point $\mathbf{X}_{\text{ocn}}$ within A is given by

$$\mathbf{F}_{\text{ocn}} = \rho_{\text{ocn}} C_{\text{ocn}} \|\mathbf{V}_{\text{ocn}} - \mathbf{V}_{\text{ice}}\| \mathbf{R}_\theta (\mathbf{V}_{\text{ocn}} - \mathbf{V}_{\text{ice}}) \tag{A.7}$$

where θ is the fixed ocean turning angle. Similarly the force induced by the atmosphere at the point $\mathbf{X}_{atm}$ within A is given by

$$\mathbf{F}_{atm} = \rho_{atm} C_{atm} \|\boldsymbol{V}_{atm} - \boldsymbol{V}_{ice}\| (\mathbf{V}_{atm} - \mathbf{V}_{ice}). \tag{A.8}$$

No atmosphere turning angle was used in this study. The Coriolis force induced on the floe is constant over the area of the of the floe and is given by

$$\mathbf{F}_{Coriolis} = \rho_{ice} f_c L_{ice} \mathbf{R}_{-\pi/2} \mathbf{V}_{ice} \tag{A.9}$$

where f_c is the Coriolis coefficient and L_{ice} is the thickness of the ice floe. The force induced by the pressure gradient of the ocean acting on the floe depends on the ocean velocity and so varies over the floe area, A

$$\mathbf{F}_{pressure} = \rho_{ice} f_c L_{ice} \mathbf{R}_{\pi/2} \mathbf{V}_{ocn}. \tag{A.10}$$

The pressure gradient of the ocean is in geostrophic balance with the Coriolis force induced on the ocean. The torque induced on the floe at the grid point $\mathbf{X}_{torque}$ is given by

$$\tau = (\mathbf{X}_{torque} - \mathbf{X}_{ice}) \times \mathbf{F}_{total} = (x_{torque} - x_{ice}) F_y - (y_{torque} - y_{ice}) F_x \tag{A.11}$$

where $\mathbf{F}_{total} = (F_x, F_y)$ are the components of the total force.

The two-layer QG model

The ocean model is a two-layer Quasi-Geostrophic (QG) model with periodic boundary conditions on a square domain. The ocean state is characterized by the stream functions $\psi_i(x, y)$ and potential vorticities (PV) $q_i(x, y)$ of each layer $i = 1, 2$. These quantities, ψ and q, are deviations from a background mean state. The level curves of the stream function, ψ_i, correspond to streamlines of the velocity field, which guarantees an incompressible flow. The ocean velocity field for each layer can thus be calculated as

$$(u_i, v_i) = \left(-\frac{\partial \psi_i}{\partial y}, \frac{\partial \psi_i}{\partial x}\right), \quad i = 1, 2. \tag{A.12}$$

The formulation of the QG equations follows the version in [8]. The PDEs which govern the time evolution of ψ_i and q_i are as follows:

$$\frac{\partial q_1}{\partial t} + \overline{u_1}\frac{\partial q_1}{\partial x} + \frac{\partial \overline{q}_1}{\partial y}\frac{\partial \psi_1}{\partial x} + J(\psi_1, q_1) = \text{ssd} \tag{A.13}$$

$$\frac{\partial q_2}{\partial t} + \overline{u_2}\frac{\partial q_2}{\partial x} + \frac{\partial \overline{q_2}}{\partial y}\frac{\partial \psi_2}{\partial x} + J(\psi_2, q_2) = -R_2\nabla^2\psi_2 + \text{ssd}. \tag{A.14}$$

Here "ssd" represents small-scale dissipation, which are higher-order derivative terms that are ignored. J is the Jacobian

$$J(\psi, q) = \frac{\partial \psi}{\partial x}\frac{\partial q}{\partial y} - \frac{\partial \psi}{\partial y}\frac{\partial q}{\partial x}. \tag{A.15}$$

The stream functions further satisfy

$$q_1 = \nabla^2\psi_1 + \frac{(\psi_2 - \psi_1)}{(1+\delta)L_d^2} \qquad q_2 = \nabla^2\psi_2 + \frac{\delta(\psi_1 - \psi_2)}{(1+\delta)L_d^2}. \tag{A.16}$$

where $\delta = H_1/H_2$, H_i is the depth of each layer, and L_d is the deformation radius.

$\overline{u_1}$ and $\overline{u_2}$ are the mean ocean velocities for each layer. $\partial\overline{q}_1/\partial y$ and $\partial\overline{q}_2/\partial y$ are the mean of the PV gradients for each layer and are given explicitly by

$$\frac{\partial\overline{q_1}}{\partial y} = \frac{\overline{u_1} - \overline{u_2}}{(1+\delta)L_d^2} \qquad \frac{\partial\overline{q_2}}{\partial y} = \frac{\delta(\overline{u_2} - \overline{u_1})}{(1+\delta)L_d^2}. \tag{A.17}$$

The final parameter, R_2, is the decay rate of the barotropic mode

$$R_2 = \frac{f_0 d_{Ekman}}{2H_2} \tag{A.18}$$

where f_0 is the Coriolis parameter and d_{Ekman} is the bottom boundary layer thickness. Note that in this formulation we use a constant Coriolis force throughout the domain.

Table 2 summarizes the parameters in the DEM and two-layer QG models.

The atmospheric wind velocity data

The fifth generation ECMWF reanalysis data product (ERA5) [115, 37] for the global climate and weather is implemented for describing the atmospheric wind that is used to calibrate the atmospheric component of the linear stochastic models.

Parameter	Value
Ocean density	$\rho_{ocn} = 1027 kg/m^3$
Ice density	$\rho_{ice} = 920 kg/m^3$
Air density	$\rho_{atm} = 1.2 kg/m^3$
Ocean drag coefficient	$c_{ocn} = 5.5 \times 10^{-3}$
Atmosphere drag coefficient	$c_{atm} = 1.6 \times 10^{-3}$
Coriolis coefficient	$f_c = 1.4 \times 10^{-4}$
Top layer mean ocean velocity	$\overline{u_1} = 2.58 km/day$
Bottom layer mean ocean velocity	$\overline{u_2} = 1.032 km/day$
Top layer mean potential vorticity	$\frac{\partial \overline{q_1}}{\partial y} = 0.0265 km^{-1} day^{-1}$
Bottom layer mean potential vorticity	$\frac{\partial \overline{q_2}}{\partial x} = -0.0212 km^{-1} day^{-1}$
Coriolis parameter	$f_c = 12 day^{-1}$
Coupling parameter	$R_1 = 6.9 \times 10^{-5} km^{-1}$
Decay rate of the barotropic mode	$R_2 = 1 day^{-1}$
Deformation radius	$L_d = 5.7 km$
Ratio of upper-to lower-layer depth	$\delta = 0.8$
Turning angle of the ocean	$\theta = \pi/9$
Ensemble size	600
Localization radius	200 km
Observational noise in location	250 m
Observational noise in angular displacement	5°

Table A.1: Parameters in the DEM, the two-layer QG models and the EnKS.

A.2 Sea ice floe observations and the processing of satellite images

Remote sensing measurements were retrieved from Moderate Resolution Imaging Spectroradiometer (MODIS) optical imagery (Level 1B 250M). The data is open-access through the Earth Observing Sysetm Data and Information System (EOSDIS) Worldview platform (`https://worldview.earthdata.nasa.gov`). In summary, both Corrected Reflectance True and False Color images were pre-processed

to reduce the imprint of atmospheric noise allowing the segmentation of sea ice floes ranging from 4 to 75 km in length scale as individual objects. Ice floes were then tracked in a three-stage process involving comparing geometrical parameters in successive images, finding potential matches, and selecting the best candidates based on the assessment of a similarity metric and surface area differences. The reader is referred to [87] for a detailed description of the pre-processing, segmentation, and tracking routines.

A.3 Calibration of Stochastic Forecast Models

Statistically accurate stochastic models are used for the ocean and atmosphere components of the forecast model. These models can be systematically calibrated based on a 20-year simulation of the two-layer QG model in the case of the ocean component and the reanalysis data set in the case of the atmosphere component. In both cases, the system state is represented in spectral space and the evolution of each spectral mode is governed by a linear stochastic model (1.1). Only the modes with a wave number less than a certain radius are kept: $|k| \leqslant 11$ in the case of the ocean and $|k| \leqslant 5$ in the case of the atmosphere for a total of 337 and 81 modes respectively.

Recall the linear stochastic model in (1.1), which is also known as the complex Ornstein–Uhlenbeck (OU) process [57]. In (1.1), a, ω, and σ are real-valued parameters with $a, \sigma > 0$, f is a complex-valued parameter, and $\dot{W}$ is a complex-valued white noise. The equilibrium distribution of this OU process is Gaussian

and its mean and variance are given in terms of the model parameters:

$$\bar{u} = \frac{f}{a - i\omega} \qquad \mathrm{Var}(u) = \frac{\sigma^2}{2a}. \tag{A.19}$$

The decorrelation time is defined as

$$T = \int_0^\infty \frac{\mathbb{E}\left[(u(t) - \bar{u})(u(t+\tau) - \bar{u})^*\right]}{\mathrm{Var}(u)}\, d\tau \tag{A.20}$$

where the expectation is taken over t. The decorrelation time is also given in terms of the model parameters

$$T = \frac{1}{a - i\omega}. \tag{A.21}$$

Using these equation for the equilibrium mean, variance, and decorrelation time, the four parameters of the OU process, d, ω, f, and σ, can be written explicitly in terms of these equilibrium statistics as in

$$d = \mathrm{Re}\left[\frac{1}{T}\right] \qquad \omega = -\,\mathrm{Im}\left[\frac{1}{T}\right] \qquad f = \frac{\bar{u}}{T} \qquad \sigma = \sqrt{2\,\mathrm{Var}(u)\,\mathrm{Re}\left[\frac{1}{T}\right]}. \tag{A.22}$$

Using these formulae for the model parameters, an OU process can be fit to match a given set of equilibrium statistics. In the case of the ocean model, an independent OU process is fit to each spectral mode of the stream function using the mean, variance, and decorrelation time of a 20-year QG model simulation. For the atmosphere model, a pair of independent OU processes is fit to each spectral mode of

the two-dimensional velocity field using the statistics of the velocity field from the ERA5 data set.

A.4 Ensemble Update

Let K denote the number of days of observations and denote the time of each day of observations by t_k for $k = 1, \ldots, K$. Denote the M-dimensional system state at time t by $\boldsymbol{\psi}(t)$ for $t_1 \leqslant t \leqslant t_K$. Then define the vector $\mathbf{d}_k$ of observed floe locations and orientations at time t_k by

$$\mathbf{d}_k = \mathcal{M}_k \left[\boldsymbol{\psi}(t_k)\right] + \boldsymbol{\epsilon}. \tag{A.23}$$

$\mathcal{M}_k$ returns only the subset of system variables corresponding to the observed floe positions and orientations at time t_k. The dependence on k allows for a changing number of observed floes at each observation time. $\boldsymbol{\epsilon}$ is a small Gaussian observational noise, corresponding to the resolution of the satellite images.

The EnKS, and all smoothing algorithms, estimate

$$f(\boldsymbol{\psi}(t) \mid \mathbf{d}_k, \ldots, \mathbf{d}_0), \tag{A.24}$$

the probability distribution of the state variable $\boldsymbol{\psi}(t)$ at a time $t_0 < t < t_k$ given the available observations $d_0, \ldots, d_k$. The estimation of the variable at any time t in the interval $[t_0, t_k]$ contrasts smoothing with filtering, which only estimates the state at the present time t_k. This distribution at time t_k can be calculated exactly

using Bayes' theorem

$$f(\boldsymbol{\psi}(t_k) \mid \mathbf{d}_0, \ldots, \mathbf{d}_k) \propto f(\boldsymbol{\psi}(t_k) \mid \mathbf{d}_0, \ldots, \mathbf{d}_{k-1}) f(\mathbf{d}_k \mid \boldsymbol{\psi}(t_k)), \tag{A.25}$$

a process called Bayesian inference. The left-hand side of equation (A.25) is the posterior distribution while the right-hand side consists of two factors: the prior or forecast distribution and the observational model. The EnKS represents the distribution of the state variable with an ensemble of model trajectories. This ensemble is then iteratively forecast and updated for each set of observations, which are processed sequentially in time, using equation (A.25). Note that the difference between the ensemble Kalman filter (EnKF) and EnKS is that the former only uses the information up to the current time instant while the latter uses all the available observational information including the future one. Thus, the EnKF is more appropriate for the providing an improved initialization of the forecast while the EnKS applies to dynamical interpolation.

Let N be the size of the ensemble and denote individual ensemble members by $\boldsymbol{\psi}_k^{(n)}(t)$ for $n = 1, \ldots, N$. The superscript "(n)" distinguishes individual ensemble members from the true system state denoted by $\boldsymbol{\psi}(t)$. The subscript k denotes that the ensemble has been updated using the first k observations. To compute $\boldsymbol{\psi}_k^{(i)}(t)$ for $t > t_k$, the forecast model is used. While only the value of $\boldsymbol{\psi}_k^{(i)}(t_{k+1})$ is required to perform the ensemble update, the ensemble at any prior time t for $t_1 \leqslant t \leqslant t_{k+1}$ can be stored in memory and updated for each new observation.

Once the ensemble has been updated for observation k, to assimilate observation

$k+1$, the $M \times N$ matrix of the ensemble members is formed

$$\mathbf{A}_k(t) = \left(\boldsymbol{\psi}_k^{(1)}(t) \quad \boldsymbol{\psi}_k^{(2)}(t) \quad \cdots \quad \boldsymbol{\psi}_k^{(N)}(t) \right). \tag{A.26}$$

The forecast ensemble matrix, $\mathbf{A}_{k+1}^f = \mathbf{A}_k(t_{k+1})$, is calculated using the forecast model. Then the updated ensemble is calculated using an $N \times N$ linear transformation, $\mathbf{T}$, of the ensemble

$$\mathbf{A}_{k+1}(t) = \mathbf{A}_k(t)\mathbf{T} \tag{A.27}$$

for any $t_1 \leqslant t \leqslant t_{k+1}$ where T is formed using the Kalman filter equations from the forecast ensemble, $\mathbf{A}_{k+1}^f$, and the observations, $\mathbf{d}_{k+1}$.

The above method is modified slightly to utilize localization, which leverages the spatial structure of the system to reduce the negative effects of spurious correlations. During the update at time t_k, each variable in $\boldsymbol{\psi}$ is associated with a location in physical space. For variables of an observed floe, their assigned location is the position of the observation. For an unobserved floe, the forecast mean position is used. For ocean and atmosphere variables, the spectral representations are transformed to physical space and the location for each grid point is used.

To update under localization, each state variable in $\boldsymbol{\psi}$ is updated individually. An ensemble corresponding to each state variable is formed from taking the values from the ensemble of model trajectories. This ensemble is then updated using only observations that are within a fixed radius of that variable's associated spatial location. The matrix $\mathbf{T}$ is formed using the full forecast ensemble, $\mathbf{A}_k^f$, and the observation vector $\mathbf{d}_k$ containing only the observations within the localization

radius. Then the $1 \times M$ row vector of the variable's ensemble is updated using this localized version of $\mathbf{T}$. This process is repeated to update all localized variables in $\boldsymbol{\psi}$.

A.5 Parameter Estimation

Unknown parameters, such as individual floe thicknesses, can be estimated within the proposed framework. The parameters are appended to the state vector and treated as non-dynamical variables. In other words, the evolution equations of the parameters θ are $d\theta/dt = 0$. The initial values of the parameters in each ensemble member are drawn from a background distribution (displayed in panel (c) of Figure 1.5). Therefore, the initial ensemble includes the uncertainty of the parameters. During the ensemble forecast, the parameter values are kept constant. However, during the ensemble update, the distribution of parameter values among the ensemble members is updated as well according to the same linear transformation. In this way the distribution of parameter values changes over the course of the algorithm even though the values do not change during the forecast. Since the parameters are non-dynamical, trajectories of these parameters do not need to be considered and only the current values of the parameters need to be stored.

A.6 Sensitivity Analysis of the Ocean Recovery

The ocean is represented in the system state by the stream function, which encodes an incompressible flow. The synthetic data was generated using a quasi-geostrophic (QG) incompressible ocean model. For nonlinear data assimilation, a statistically accurate stochastic forecast model is used to forecast the ensemble of model trajectories. This ensemble is then updated after each set of observations, allowing any model variables, including the ocean stream function, to be recovered at any point in time. For these experiments, the ocean is recovered for the day of July 1, which is roughly in the middle of the set of observations. Note that the ocean evolves slowly with time and so this is representative of the overall ocean recovery. In addition to the standard setup in the main text, the following two additional experiments of the ocean flow field recovery are included.

Figure A.1 shows the recovery of the ocean stream function using synthetic data without atmosphere forcing, i.e. atmosphere component of the sea ice model is removed for both the generation of synthetic data and in the forecast model. In this case the dynamics of the floes are fully accounted for by ocean forces and the total degree of freedom of the unobserved variables is reduced. Therefore, a better recovery of the ocean state is expected due to the lack of interference from the atmosphere. Indeed, the results of the ocean recovery seem to be improved. However, the lack of atmosphere forcing, which is responsible for the large-scale floe movements, leads to floe trajectories that travel a shorter distance, and hence cover a smaller portion of the region. This impacts localization during the ocean ensemble update, which removes the influence of observed floes that fall outside

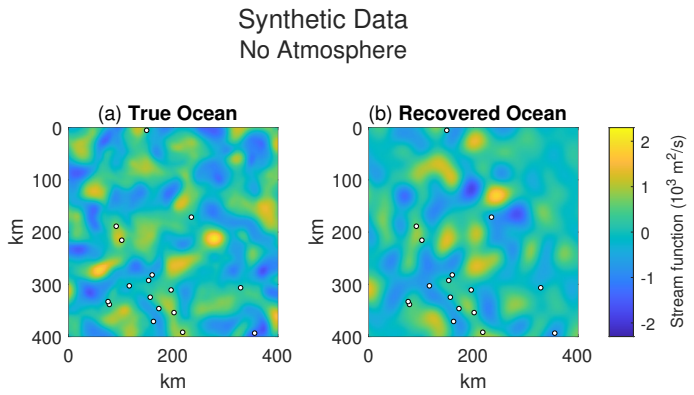

Figure A.1: A comparison of the true and recovered ocean on July 1 for a synthetic data experiment with no atmosphere forcing. Panel (a) shows the stream function of the true ocean generated from a QG ocean model. Panel (b) shows the recovery of the stream function using the ensemble mean. The pattern correlation between the true and recovered ocean in the 400km$\times$400km subdomain is 0.43. Each plot shows white dots indicating the position of the observed floes. While a 400 km by 400 km region is plotted, the ocean is extends another 100 km on all sides to reduce the impact of the periodicity in the ocean model.

the localization radius of each ocean grid point. In this case the amplitude of the recovered ocean is nearly zero in some areas, which reflects the uncertainty of the ensemble rather than an actual amplitude of individual ensemble members.

In another experiment, Figure A.2 shows the ocean recovery for synthetic data that is generated with atmospheric forcing, but using a 200 km by 200 km ocean (rather than a 600 km by 600 km ocean) that is extended periodically to the entire domain. This is analogous to a case in which a much higher density of observations is available. No localization is utilized here, as the typical localization radius of 200 km covers the entire ocean. In this case the qualitative recovery of the ocean is quite good, demonstrating that a sufficiently high density of floe observations can recover ocean features well.

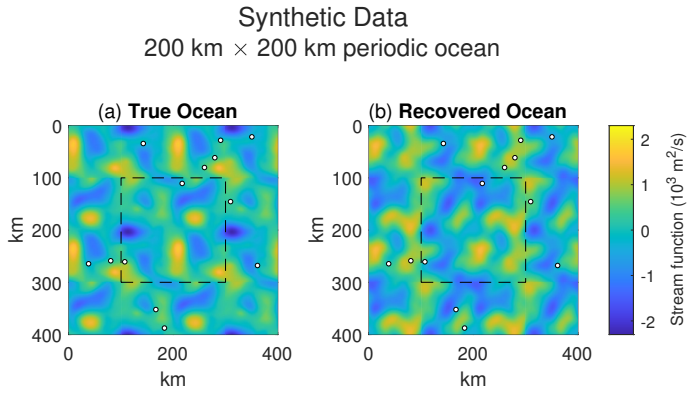

Figure A.2: Compares the true and recovered ocean stream functions for a synthetic data experiment using a 200 km by 200 km ocean extended periodically to the whole domain. Panel (a) shows the true ocean, generated from a QG model on a 200 km by 200 km domain. The dashed black line indicates the size of the ocean which is then periodically extended to the whole domain. Panel (b) shows the ensemble mean stream function. The pattern correlation between the true and recovered ocean is 0.60. The floe positions are indicated with white dots. Note that observations outside of the 200 km by 200 km ocean region influence the ensemble update of the ocean just as much as observations inside the region due to the periodicity of the ocean.

A.7 Behavior of Ensemble Members in the Dynamical Interpolation

The ensemble mean provides a complete trajectory and is the best point estimate of a floe's position available from the dynamical interpolation method. However, despite the fact that each ensemble member is based on the model forecast corrected by the partial observations, their average — the ensemble mean — is often not a physically consistent trajectory of the model. Therefore, it is important to look at each individual ensemble member and understand the additional physical properties that are not fully reflected in the ensemble mean.

Figure A.3 shows an example of a continuous floe trajectory generated from the sea ice model, along with the daily observed floe positions. The trajectories between daily observations are then interpolated using different methods. Note that this floe trajectory contains a loop midway through, which is compared qualitatively with the interpolation using different methods. While the ensemble mean provides better physics than linear interpolation and contains nonlinear evolution of the trajectory, it still fails to capture this loop. On the other hand, more than half of the ensemble members recover such a loop around the correct location, something which is otherwise lost during averaging. The finding here reveals the potential role that utilizing individual ensemble members can play in identifying properties of floe trajectories that are not otherwise captured by the ensemble mean. Such a result also highlights the importance of considering the uncertainty represented by the ensemble members in addition to the mean state estimation.

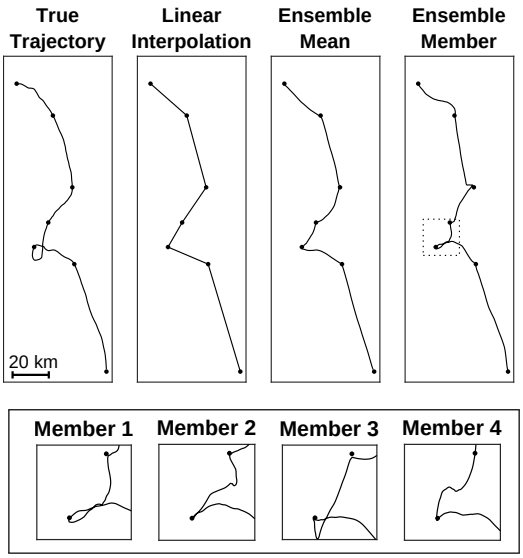

Figure A.3: Comparison of different methods for the interpolation of floe trajectories. The first panel shows a floe trajectory generated from the sea ice model and sampled every time unit at the black dots. This trajectory contains a loop. The next three panels show various forms of interpolation. The first shows linear interpolation. The second shows the ensemble mean. The third shows a sampled ensemble member. The bottom panels shows other sampled ensemble members.

B APPENDIX TO EFFECTIVE STATISTICAL CONTROL STRATEGIES FOR COMPLEX TURBULENT DYNAMICAL SYSTEMS

B.1 Statistical linear response theory for turbulent dynamical systems

Here, we give a brief description for the theories and strategies in predicting model statistical responses using statistical linear response theory [93, 98, 103]. The statistical response theory and the Fluctuation-Dissipation Theorem (FDT) offer a convenient way to compute leading-order statistical approximation about model responses to various external perturbations [93, 108]. Assume that the perfect model of the turbulent dynamical system is

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}). \tag{B.1}$$

The ideal equilibrium state associated with (B.1) is the invariant probability density $p_{eq}(\mathbf{u})$ that satisfies $\mathcal{L}_{FP} p_{eq} = 0$ with $\mathcal{L}_{FP}$ the corresponding Fokker-Planck operator. The equilibrium statistics of some functional $A(\mathbf{u})$ are determined by

$$\langle A \rangle_{eq} = \int A(\mathbf{u})\, p_{eq}(\mathbf{u})\, d\mathbf{u}. \tag{B.2}$$

Next, perturb the system (B.1) by the external forcing perturbation written in separation with temporal and spatial variables,

$$\delta \mathbf{F}(\mathbf{u}, t) = \mathbf{w}(\mathbf{u})\, \delta f(t), \tag{B.3}$$

where $\mathbf{w}(\mathbf{u})$ defines the spatial dependence and $f(t)$ defines the temporal variability of the forcing perturbations. We are interested in the statistical response in the perturbed state $\mathbf{u}^\delta$ subject to the perturbation form (B.3) in the perturbed equation

$$\frac{d\mathbf{u}^\delta}{dt} = \mathbf{F}\left(\mathbf{u}^\delta\right) + \mathbf{w}(\mathbf{u})\, \delta f(t). \tag{B.4}$$

The resulting perturbed probability density function p^δ due to the general forcing perturbation $\delta\mathbf{F}$ then can be asymptotically expanded according to the equilibrium and the correction to perturbation, that is

$$p^\delta(\mathbf{u}, t) = p_{eq}(\mathbf{u}) + \delta p'(\mathbf{u}, t), \quad \int p_{eq}(\mathbf{u})\, d\mathbf{u} = 1, \int \delta p'(\mathbf{u})\, d\mathbf{u} = 0. \tag{B.5}$$

Accordingly, the statistical expectation of any functional $A(\mathbf{u})$ to the perturbation is formulated accordingly under the equilibrium measure and leading-order correction,

$$\langle A(\mathbf{u})\rangle = \langle A\rangle_{eq} + \delta\langle A\rangle(t) + O\left(\delta^2\right), \tag{B.6}$$

using the measure asymptotic decomposition (B.5) and (B.2), and $\delta\langle A\rangle = \int A(\mathbf{u})\, \delta p'(\mathbf{u})$ according to the perturbation correction $\delta p'$.

Linear response theory gives the leading order prediction for the statistical re-

sponse functional computed from the convolution with the linear response operator when the perturbation amplitude δ is small enough

$$\delta \langle A \rangle (t) = \mathcal{R}_A * \delta f(t) = \int_0^t \mathcal{R}_A (t-s)\, \delta f(s)\, ds. \tag{B.7}$$

The derivation of the above formula is concluded from separating the leading order dynamics of the Fokker-Planck equation for the perturbed probability function [93]. The pointed-bracket above denotes the statistical average under the solution p from the Fokker-Planck equation. $\mathcal{R}_A(t)$ is called the *linear response operator* corresponding to the functional A, which is calculated through correlation functions in the unperturbed statistical equilibrium only

$$\mathcal{R}_A(t) = \langle A[\mathbf{u}(t)]\, G[\mathbf{u}(0)] \rangle_{eq}, \quad G(\mathbf{u}) = -p_{eq}^{-1} \mathrm{div}_{\mathbf{u}} \left(\mathbf{w} p_{eq} \right). \tag{B.8}$$

The above linear response formula for statistical responses (B.7) is shown to have high skill for the mean response and some skill for the variance response for a wide variety of turbulent dynamical systems even with nonlinearity [98, 2, 100, 90, 61]. In addition, rigorous theories [64, 95] have been provided for the validity of the linear response theory.

Still, the linear response operator $\mathcal{R}_A$ is difficult to calculate by directly using (B.8) for general systems considering the complicated and unaccessible equilibrium distribution p_{eq}. A variety of Gaussian approximations for p_{eq} and improved algorithms have been developed for computing the linear response operators [83, 93, 100, 95]. In many situations, it is found that a Gaussian PDF from the

Gibbs invariant measure can offer a quite accurate characterization of the unperturbed distribution of the system. In this way, the *quasi-Gaussian* (qG) closure, $p_{eq} \sim p_{eq}^{G}$, provides a desirable approximation of the equilibrium measure. Then the linear response operator (B.8) can be computed directly from the autocorrection functions. The qG FDT has shown effective skill in predicting the statistical responses especially in the mean state, thus we adopt this approximation in the main text for designing the control strategies.

BIBLIOGRAPHY

[1] Ole Morten Aamo and Miroslav Krstic. *Flow control by feedback: stabilization and mixing*. Springer Science & Business Media, 2003.

[2] Rafail V Abramov and Andrew J Majda. Blended response algorithms for linear fluctuation-dissipation for complex nonlinear dynamical systems. *Nonlinearity*, 20(12):2793–2821, 2007.

[3] Mohammad Ahmadzadehtalatapeh and Majid Mousavi. A review on the drag reduction methods of the ship hulls for improving the hydrodynamic performance. *International Journal of Maritime Technology*, 4:51–64, 2015.

[4] Alexandros Alexakis and Luca Biferale. Cascades and transitions in turbulent flows. *Physics Reports*, 767:1–101, 2018.

[5] Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[6] Jeffrey L Anderson. Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, 140(7):2359–2371, 2012.

[7] A Apte and CKRT Jones. The impact of nonlinearity in lagrangian data assimilation. *Nonlinear Processes in Geophysics*, 20(3):329–341, 2013.

[8] Brian K Arbic and Glenn R Flierl. Baroclinically unstable geostrophic turbulence in the limits of strong and weak bottom Ekman friction: Application to midocean eddies. *Journal of Physical Oceanography*, 34(10):2257–2273, 2004.

[9] Fanny Ardhuin and Cédric Prévost. Arctic ocean sea ice drift reprocessed, 2020.

[10] Franco Auteri, Arturo Baron, Marco Belan, Gabriele Campanardi, and Maurizio Quadrio. Experimental assessment of drag reduction by traveling waves in a turbulent pipe flow. *Physics of Fluids*, 22(11), 2010.

[11] TA Averina and SS Artemiev. Numerical solution of systems of stochastic differential equations. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 3(4):267–286, 1988.

[12] Shervin Bagheri and Dan S Henningson. Transition delay using control theory. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1940):1365–1381, 2011.

[13] Shervin Bagheri, Dan S Henningson, J Hoepffner, and Peter J Schmid. Input-output analysis and control design applied to a linear model of spatially developing flows. *Applied Mechanics Reviews*, 62(2):020803, 2009.

[14] Martino Bardi, Italo Capuzzo Dolcetta, et al. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*, volume 12. Springer, 1997.

[15] DW Bechert, M Bruse, W vd Hage, JG Th Van der Hoeven, and G Hoppe. Experiments on drag-reducing surfaces and their optimization with an adjustable geometry. *Journal of fluid mechanics*, 338:59–87, 1997.

[16] Judith Berner, Ulrich Achatz, Lauriane Batte, Lisa Bengtsson, Alvaro de la Cámara, Hannah M Christensen, Matteo Colangeli, Danielle RB Coleman, Daan Crommelin, Stamen I Dolaptchiev, et al. Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, 98(3):565–588, 2017.

[17] Uma S Bhatt, Donald A Walker, John E Walsh, Eddy C Carmack, Karen E Frey, Walter N Meier, Sue E Moore, Frans-Jan W Parmentier, Eric Post, Vladimir E Romanovsky, and William R Simpson. Implications of Arctic sea ice decline for the Earth system. *Annual Review of Environment and Resources*, 39:57–89, 2014.

[18] Shankar P Bhattacharyya, Aniruddha Datta, and Lee H Keel. *Linear control theory: structure, robustness, and optimization*. CRC press, 2018.

[19] Michal Branicki, Andrew J Majda, and Kody JH Law. Accuracy of Some Approximate Gaussian Filters for the Navier–Stokes Equation in the Presence of Model Error. *Multiscale Modeling & Simulation*, 16(4):1756–1794, 2018.

[20] Robert Brodkey. *Turbulence in mixing operations: theory and application to mixing and reaction*. Elsevier, 2012.

[21] Charles Brunette, L. Bruno Tremblay, and Robert Newton. A new state-dependent parameterization for the free drift of sea ice. *Cryosphere*, 16(2):533–557, 2022.

[22] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.

[23] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[24] Michele Alessandro Bucci, Onofrio Semeraro, Alexandre Allauzen, Guillaume Wisniewski, Laurent Cordier, and Lionel Mathelin. Control of chaotic systems by deep reinforcement learning. *Proceedings of the Royal Society A*, 475(2231):20190351, 2019.

[25] Dudley B Chelton, Michael G Schlax, and Roger M Samelson. Global observations of nonlinear mesoscale eddies. *Progress in oceanography*, 91(2):167–216, 2011.

[26] Dudley B Chelton, Michael G Schlax, Roger M Samelson, and Roland A de Szoeke. Global observations of large oceanic eddies. *Geophysical Research Letters*, 34(15), 2007.

[27] Nan Chen. Learning nonlinear turbulent dynamics from partial observations via analytically solvable conditional statistics. *Journal of Computational Physics*, 418:109635, 2020.

[28] Nan Chen. *Stochastic Methods for Modeling and Predicting Complex Dynamical Systems: Uncertainty Quantification, State Estimation, and Reduced-Order Models*. Springer Nature, 2023.

[29] Nan Chen, Shubin Fu, and Georgy Manucharyan. Lagrangian data assimilation and parameter estimation of an idealized sea ice discrete element model. *Journal of Advances in Modeling Earth Systems*, 13(10):e2021MS002513, 2021.

[30] Nan Chen, Shubin Fu, and Georgy E Manucharyan. An efficient and statistically accurate lagrangian data assimilation algorithm with applications to discrete element sea ice models. *Journal of Computational Physics*, 455:111000, 2022.

[31] Nan Chen, Evelyn Lunasin, and Stephen Wiggins. Lagrangian descriptors with uncertainty. *arXiv preprint arXiv:2307.04006*, 2023.

[32] Nan Chen and Andrew J Majda. Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. *Entropy*, 20(7):509, 2018.

[33] Nan Chen and Andrew J Majda. Efficient nonlinear optimal smoothing and sampling algorithms for complex turbulent nonlinear dynamical systems with partial observations. *Journal of Computational Physics*, 410:109381, 2020.

[34] Nan Chen, Andrew J Majda, and Xin T Tong. Information barriers for noisy lagrangian tracers in filtering random incompressible flows. *Nonlinearity*, 27(9):2133, 2014.

[35] Nan Chen, Andrew J Majda, and Xin T Tong. Rigorous analysis for efficient statistically accurate algorithms for solving Fokker–Planck equations in large dimensions. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1198–1223, 2018.

[36] Yu-Hsin Cheng, Chung-Ru Ho, Quanan Zheng, and Nan-Jung Kuo. Statistical characteristics of mesoscale eddies in the north pacific derived from satellite altimetry. *Remote Sensing*, 6(6):5164–5183, 2014.

[37] Copernicus Climate Change Service (C35). ERA 5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, 2017.

[38] Jeffrey Covington, Nan Chen, and Monica M Wilhelmus. Bridging gaps in the climate observation network: A physics-based nonlinear dynamical interpolation of lagrangian ice floe measurements via data-driven stochastic models. *Journal of Advances in Modeling Earth Systems*, 14(9):e2022MS003218, 2022.

[39] Jeffrey Covington, Di Qi, and Nan Chen. Effective statistical control strategies for complex turbulent dynamical systems. *Proceedings of the Royal Society A*, 479(2279):20230546, 2023.

[40] Peter A Cundall. Formulation of a three-dimensional distinct element model—Part I. A scheme to detect and represent contacts in a system composed of many polyhedral blocks. In *International journal of rock mechanics and mining sciences & geomechanics abstracts*, volume 25, pages 107–116. Elsevier, 1988.

[41] Peter A Cundall and Otto DL Strack. A discrete numerical model for granular assemblies. *geotechnique*, 29(1):47–65, 1979.

[42] Anders Damsgaard, Alistair Adcroft, and Olga Sergienko. Application of discrete element methods to approximate sea ice dynamics. *Journal of Advances in Modeling Earth Systems*, 10(9):2228–2244, 2018.

[43] AM Doglioli, Bruno Blanke, Sabrina Speich, and Guillaume Lapeyre. Tracking coherent structures in a regional ocean model with wavelet analysis: Application to cape basin eddies. *Journal of Geophysical Research: Oceans*, 112(C5), 2007.

[44] Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.

[45] Thomas Duriez, Steven L Brunton, and Bernd R Noack. *Machine learning control-taming nonlinear dynamics and turbulence*, volume 116. Springer, 2017.

[46] John A Dykema, David W Keith, James G Anderson, and Debra Weisenstein. Stratospheric controlled perturbation experiment: a small-scale experiment to improve understanding of the risks of solar geoengineering. *Philosophi-*

cal Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 372(2031):20140059, 2014.

[47] Francesco d'Ovidio, Jordi Isern-Fontanet, Cristóbal López, Emilio Hernández-García, and Emilio García-Ladona. Comparison between eulerian diagnostics and finite-size lyapunov exponents computed from altimetry in the algerian basin. *Deep Sea Research Part I: Oceanographic Research Papers*, 56(1):15–31, 2009.

[48] Wouter Edeling and Daan Crommelin. Reducing data-driven dynamical subgrid scale models by physical constraints. *Computers & Fluids*, 201:104470, 2020.

[49] Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, Berlin Heidelberg, Germany, 2009.

[50] Nicolo Fabbiane, Onofrio Semeraro, Shervin Bagheri, and Dan S Henningson. Adaptive and model-based control theory applied to convectively unstable flows. *Applied Mechanics Reviews*, 66(6):060801, 2014.

[51] James H Faghmous, Ivy Frenger, Yuanshun Yao, Robert Warmka, Aron Lindell, and Vipin Kumar. A daily global mesoscale ocean eddy dataset from satellite altimetry. *Scientific data*, 2(1):1–16, 2015.

[52] Brian F Farrell and Petros J Ioannou. Stochastic forcing of the linearized Navier–Stokes equations. *Physics of Fluids A: Fluid Dynamics*, 5(11):2600–2609, 1993.

[53] Brian F Farrell and Petros J Ioannou. Statistical state dynamics: a new perspective on turbulence in shear flow. *arXiv preprint arXiv:1412.8290*, 2014.

[54] Katharina Franz, Ribana Roscher, Andres Milioto, Susanne Wenzel, and Jürgen Kusche. Ocean eddy identification and tracking using neural networks. In *Igarss 2018-2018 IEEE international geoscience and remote sensing symposium*, pages 6887–6890. IEEE, 2018.

[55] Kai Fukami, Takaaki Murata, Kai Zhang, and Koji Fukagata. Sparse identification of nonlinear dynamics with low-dimensionalized flow representations. *Journal of Fluid Mechanics*, 926:A10, 2021.

[56] A. Gabrielski, G. Badin, and L. Kaleschke. Anomalous dispersion of sea ice in the Fram Strait region. *Journal of Geophysical Research: Oceans*, 120(3):1809–1824, March 2015.

[57] Crispin Gardiner. *Stochastic methods*, volume 4. Springer, Berlin, Heidelberg, Germany, 2009.

[58] Boris Gershgorin, John Harlim, and Andrew J Majda. Improving filtering and prediction of spatially extended turbulent systems with model errors through stochastic parameter estimation. *Journal of Computational Physics*, 229(1):32–57, 2010.

[59] Boris Gershgorin and Andrew J Majda. A test model for fluctuation–dissipation theorems with time-periodic statistics. *Physica D: Nonlinear Phenomena*, 239(17):1741–1757, 2010.

[60] Michael Ghil and Stephen Childress. *Topics in geophysical fluid dynamics: atmospheric dynamics, dynamo theory, and climate dynamics*, volume 60. Springer Science & Business Media, 2012.

[61] Andrey Gritsun, Grant Branstator, and Andrew Majda. Climate response of linear and quadratic functionals using the fluctuation–dissipation theorem. *Journal of the Atmospheric Sciences*, 65(9):2824–2841, 2008.

[62] Ian Grooms and Andrew J Majda. Efficient stochastic superparameterization for geophysical turbulence. *Proceedings of the National Academy of Sciences*, 110(12):4464–4469, 2013.

[63] Federica Gugole and Christian LE Franzke. Numerical development and evaluation of an energy conserving conceptual stochastic climate model. *Mathematics of climate and weather forecasting*, 5(1):45–64, 2019.

[64] Martin Hairer and Andrew J Majda. A simple framework to justify linear response theory. *Nonlinearity*, 23(4):909, 2010.

[65] R Hart, Peter A Cundall, and J Lemos. Formulation of a three-dimensional distinct element model—Part II. Mechanical calculations for motion and interaction of a system composed of many polyhedral blocks. In *International journal of rock mechanics and mining sciences & Geomechanics abstracts*, volume 25, pages 117–125. Elsevier, 1988.

[66] W DIII Hibler. A dynamic thermodynamic sea ice model. *Journal of physical oceanography*, 9(4):815–846, 1979.

[67] Marisa Holmes. Structure. In *Organizing Occupy Wall Street: This is Just Practice*, pages 129–143. Springer, 2023.

[68] Elizabeth C Hunke and John K Dukowicz. An elastic–viscous–plastic model for sea ice dynamics. *Journal of Physical Oceanography*, 27(9):1849–1867, 1997.

[69] J. K. Hutchings, P. Heil, A. Steer, and W. D. Hibler. Subsynoptic scale spatial variability of sea ice deformation in the western Weddell Sea during early summer. *Journal of Geophysical Research*, 117(C1):C01002, 2012.

[70] Peter J Irvine, Ben Kravitz, Mark G Lawrence, and Helene Muri. An overview of the earth system science of solar geoengineering. *Wiley Interdisciplinary Reviews: Climate Change*, 7(6):815–833, 2016.

[71] Polona Itkin, Gunnar Spreen, Bin Cheng, Martin Doble, Fanny Girard-Ardhuin, Jari Haapala, Nick Hughes, Lars Kaleschke, Marcel Nicolaus, and Jeremy Wilkinson. Thin ice and storms: Sea ice deformation from buoy arrays deployed during N-ICE2015: THIN ICE AND STORMS. *Journal of Geophysical Research: Oceans*, 122(6):4661–4674, June 2017.

[72] Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A*, 474(2219):20180335, 2018.

[73] M Amin Khodkar and Pedram Hassanzadeh. Data-driven reduced modelling of turbulent Rayleigh–Bénard convection using DMD-enhanced fluctuation–dissipation theorem. *Journal of Fluid Mechanics*, 852, 2018.

[74] John Kim. Physics and control of wall turbulence for drag reduction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1940):1396–1411, 2011.

[75] John Kim and Thomas R Bewley. A linear systems approach to flow control. *Annu. Rev. Fluid Mech.*, 39:383–417, 2007.

[76] Dmitri Kondrashov, Mickaël D Chekroun, and Michael Ghil. Data-driven non-Markovian closure models. *Physica D: Nonlinear Phenomena*, 297:33–55, 2015.

[77] Dmitri Kondrashov, Mickaël D Chekroun, Xiaojun Yuan, and Michael Ghil. Data-adaptive harmonic decomposition and stochastic modeling of Arctic sea ice. In *Advances in nonlinear geosciences*, pages 179–205. Springer, 2018.

[78] Suzanne M Kresta, Arthur W Etchells III, David S Dickey, Victor A Atiemo-Obeng, et al. *Advances in industrial mixing: a companion to the handbook of industrial mixing*. John Wiley & Sons, 2015.

[79] R Kwok. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environmental Research Letters*, 13(10):105005, oct 2018.

[80] Ronald Kwok. Radarsat-1 data (csa). dataset: Lagrangian sea-ice kinematics.

[81] Changhoon Lee, John Kim, David Babcock, and Rodney Goodman. Application of neural networks to turbulence control for drag reduction. *Physics of Fluids*, 9(6):1740–1747, 1997.

[82] Ruibo Lei, Dawei Gui, Petra Heil, Jennifer K. Hutchings, and Minghu Ding. Comparisons of sea ice motion and deformation, and their responses to ice conditions and cyclonic activity in the western Arctic Ocean between two summers. *Cold Regions Science and Technology*, 170(November 2018):102925, 2020.

[83] Cecil E Leith. Climate response and fluctuation dissipation. *Journal of Atmospheric Sciences*, 32(10):2022–2026, 1975.

[84] Matti Leppäranta. *The drift of sea ice*. Springer Science & Business Media, Berlin, Heidelberg, Germany, 2011.

[85] Martin Leutbecher and Tim N Palmer. Ensemble forecasting. *Journal of computational physics*, 227(7):3515–3539, 2008.

[86] RW Lindsay and HL Stern. A new Lagrangian model of Arctic sea ice. *Journal of physical oceanography*, 34(1):272–283, 2004.

[87] R Lopez-Acosta, MP Schodlok, and MM Wilhelmus. Ice floe tracker: An algorithm to automatically retrieve lagrangian trajectories via feature matching from moderate-resolution visual imagery. *Remote Sensing of Environment*, 234:111406, 2019.

[88] R. Lopez-Acosta and Monica M Wilhelmus. Library of sea ice floe remote sensing observations in the Beaufort Sea Marginal Ice Zone, May 2021.

[89] Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1. Reading, 1996.

[90] Nicholas J Lutsko, Isaac M Held, and Pablo Zurita-Gotor. Applying the fluctuation–dissipation theorem to a two-layer model of quasigeostrophic turbulence. *Journal of the Atmospheric Sciences*, 72(8):3161–3177, 2015.

[91] Douglas G MacMartin, Ken Caldeira, and David W Keith. Solar geoengineering to limit the rate of temperature change. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2031):20140134, 2014.

[92] Douglas G MacMartin, Ben Kravitz, David W Keith, and Andrew Jarvis. Dynamics of the coupled human–climate system resulting from closed-loop control of solar geoengineering. *Climate dynamics*, 43:243–258, 2014.

[93] Andrew Majda, Rafail V Abramov, and Marcus J Grote. *Information theory and stochastics for multiscale nonlinear systems*, volume 25. American Mathematical Soc., 2005.

[94] Andrew Majda and Xiaoming Wang. *Nonlinear dynamics and statistical theories for basic geophysical flows*. Cambridge University Press, 2006.

[95] Andrew Majda and Xiaoming Wang. Linear response theory for statistical ensembles in complex systems with time-periodic forcing. *Communications in Mathematical Sciences*, 8(1):145–172, 2010.

[96] Andrew J Majda. Statistical energy conservation principle for inhomogeneous turbulent dynamical systems. *Proceedings of the National Academy of Sciences*, 112(29):8937–8941, 2015.

[97] Andrew J Majda. *Introduction to turbulent dynamical systems in complex systems*. Springer, Switzerland, 2016.

[98] Andrew J Majda, Rafail Abramov, and Boris Gershgorin. High skill in low-frequency climate response through fluctuation dissipation theorems despite structural instability. *Proceedings of the National Academy of Sciences*, 107(2):581–586, 2010.

[99] Andrew J Majda and Nan Chen. Model error, information barriers, state estimation and prediction in complex multiscale systems. *Entropy*, 20(9):644, 2018.

[100] Andrew J Majda, Boris Gershgorin, and Yuan Yuan. Low-frequency climate response and fluctuation–dissipation theorems: Theory and practice. *Journal of the Atmospheric Sciences*, 67(4):1186–1201, 2010.

[101] Andrew J Majda and John Harlim. Physics constrained nonlinear regression models for time series. *Nonlinearity*, 26(1):201, 2012.

[102] Andrew J Majda and Di Qi. Effective control of complex turbulent dynamical systems through statistical functionals. *Proceedings of the National Academy of Sciences*, 114(22):5571–5576, 2017.

[103] Andrew J Majda and Di Qi. Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Review*, 60(3):491–549, 2018.

[104] Andrew J Majda and Di Qi. Linear and nonlinear statistical response theories with prototype applications to sensitivity analysis and statistical control of complex turbulent dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(10):103131, 2019.

[105] Andrew J Majda and Di Qi. Using statistical functionals for effective control of inhomogeneous complex turbulent dynamical systems. *Physica D: Nonlinear Phenomena*, 392:34–56, 2019.

[106] Ana M Mancho, Stephen Wiggins, Jezabel Curbelo, and Carolina Mendoza. Lagrangian descriptors: A method for revealing phase space structures of general time dependent dynamical systems. *Communications in Nonlinear Science and Numerical Simulation*, 18(12):3530–3557, 2013.

[107] Georgy E Manucharyan, Rosalinda Lopez-Acosta, and Monica M Wilhelmus. Spinning ice floes reveal intensification of mesoscale eddies in the western Arctic Ocean. *Scientific Reports*, 12(1):1–13, 2022.

[108] Umberto Marini Bettolo Marconi, Andrea Puglisi, Lamberto Rondoni, and Angelo Vulpiani. Fluctuation–dissipation: response theory in statistical physics. *Physics reports*, 461(4-6):111–195, 2008.

[109] Wieslaw Maslowski, Jaclyn Clement Kinney, Matthew Higgins, and Andrew Roberts. The future of Arctic sea ice. *Annual Review of Earth and Planetary Sciences*, 40:625–654, 2012.

[110] Woosok Moon and John S Wettlaufer. A stochastic dynamical model of Arctic sea ice. *Journal of Climate*, 30(13):5119–5140, 2017.

[111] Francesco Nencioli, Changming Dong, Tommy Dickey, Libe Washburn, and James C McWilliams. A vector geometry–based eddy detection algorithm and its application to a high-resolution numerical model product and high-frequency radar surface velocities in the southern california bight. *Journal of atmospheric and oceanic technology*, 27(3):564–579, 2010.

[112] Dwight Roy Nicholson and Dwight R Nicholson. *Introduction to plasma theory*, volume 1. Wiley New York, 1983.

[113] Peter J Nolan, Mattia Serra, and Shane D Ross. Finite-time lyapunov exponents in the instantaneous limit and material transport. *Nonlinear Dynamics*, 100(4):3825–3852, 2020.

[114] Akira Okubo. Horizontal dispersion of floatable particles in the vicinity of velocity singularities such as convergences. In *Deep sea research and oceanographic abstracts*, volume 17, pages 445–454. Elsevier, 1970.

[115] Jon Olauson. ERA5: The new champion of wind power modelling? *Renewable energy*, 126:322–331, 2018.

[116] Edward Ott, Celso Grebogi, and James A Yorke. Controlling chaos. *Physical review letters*, 64(11):1196, 1990.

[117] Tim Palmer. The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145:12–24, 2019.

[118] Edward L Paul, Victor A Atiemo-Obeng, and Suzanne M Kresta. *Handbook of industrial mixing*. Wiley Online Library, 2004.

[119] Jens Pfeiffer and Rudibert King. Multivariable closed-loop flow control of drag and yaw moment for a 3d bluff body. In *6th AIAA Flow control conference*, page 2802, 2012.

[120] Stephen B Pope. *Turbulent flows*. Cambridge university press, 2000.

[121] Di Qi and Andrew J Majda. Predicting extreme events for passive scalar turbulence in two-layer baroclinic flows through reduced-order stochastic models. *Communications in Mathematical Sciences*, 16(1):17–51, 2018.

[122] Maurizio Quadrio. Drag reduction in turbulent boundary layers by in-plane wall motion. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1940):1428–1442, 2011.

[123] Valentin Resseguier, Long Li, Gabriel Jouan, Pierre Dérian, Etienne Mémin, and Bertrand Chapron. New trends in ensemble forecast strategy: uncertainty quantification for coarse-grid computational fluid dynamics. *Archives of Computational Methods in Engineering*, 28:215–261, 2021.

[124] Valentin Resseguier, Agustin M Picard, Etienne Mémin, and Bertrand Chapron. Quantifying truncation-related uncertainties in unsteady fluid

dynamics reduced order models. *SIAM/ASA Journal on Uncertainty Quantification*, 9(3):1152–1183, 2021.

[125] Clarence W Rowley. Model reduction for fluids, using balanced proper orthogonal decomposition. *International Journal of Bifurcation and Chaos*, 15(03):997–1013, 2005.

[126] Sarah A Sheard and Ali Mostashari. Principles of complex systems for systems engineering. *Systems Engineering*, 12(4):295–311, 2009.

[127] Sigurd Skogestad and Ian Postlethwaite. *Multivariable feedback control: analysis and design*. john Wiley & sons, 2005.

[128] JMAC Souza, Clement de Boyer Montégut, and Pierre-Yves Le Traon. Comparison between three implementations of automatic identification algorithms for the quantification and characterization of mesoscale eddies in the south atlantic ocean. *Ocean Science*, 7(3):317–334, 2011.

[129] Vernon A Squire. Ocean wave interactions with sea ice: A reappraisal. *Annual Review of Fluid Mechanics*, 52:37–60, 2020.

[130] Steven H Strogatz. *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[131] Mohd Nizam Sudin, Mohd Azman Abdullah, Shamsul Anuar Shamsuddin, Faiz Redza Ramli, and Musthafah Mohd Tahir. Review of research on vehicles aerodynamic drag reduction methods. *International Journal of Mechanical and Mechatronics Engineering*, 14(02):37–47, 2014.

[132] David N Thomas. *Sea ice*. John Wiley & Sons, Hoboken, New Jersey, USA, 2017.

[133] Jim Thomson, Stephen Ackley, Fanny Girard-Ardhuin, Fabrice Ardhuin, Alex Babanin, Guillaume Boutin, John Brozena, Sukun Cheng, Clarence Collins, Martin Doble, Chris Fairall, Peter Guest, Claus Gebhardt, Johannes Gemmrich, Hans C Graber, Benjamin Holt, Susanne Lehner, Bjôrn Lund, Michael H Meylan, Ted Maksym, Fabien Montiel, Will Perrie, Ola Persson, Luc Rainville, W Erick Rogers, Hui Shen, Hayley Shen, Vernon Squire, Sharon Stammerjohn, Justin Stopa, Madison M Smith, Peter Sutherland, and Peter Wadhams. Overview of the Arctic sea state and boundary layer physics program. *Journal of Geophysical Research: Oceans*, 123(12):8674–8687, 2018.

[134] Mary-Louise Timmermans, John Toole, and Richard Krishfield. Warming of the interior arctic ocean linked to sea ice losses at the basin margins. *Science advances*, 4(8):eaat6773, 2018.

[135] Srikanth Toppaladoddi and John S Wettlaufer. Theory of the sea ice thickness distribution. *Physical Review Letters*, 115(14):148501, 2015.

[136] Zoltan Toth and Eugenia Kalnay. Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125(12):3297–3319, 1997.

[137] L Bruno Tremblay and LA Mysak. Modeling sea ice as a granular material, including the dilatancy effect. *Journal of Physical Oceanography*, 27(11):2342–2360, 1997.

[138] Jukka Tuhkuri and Arttu Polojärvi. A review of discrete element simulation of ice–structure interaction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2129):20170335, 2018.

[139] Geoffrey K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, 2017.

[140] Reinier van Buel and Holger Stark. Active open-loop control of elastic turbulence. *Scientific Reports*, 10(1):15704, 2020.

[141] Rahel Vortmeyer-Kley, Ulf Gräwe, and Ulrike Feudel. Detecting and tracking eddies in oceanic flow fields: a lagrangian descriptor based on the modulus of vorticity. *Nonlinear Processes in Geophysics*, 23(4):159–173, 2016.

[142] Rahel Vortmeyer-Kley, Peter Holtermann, Ulrike Feudel, and Ulf Gräwe. Comparing eulerian and lagrangian eddy census for a tide-less, semi-enclosed basin, the baltic sea. *Ocean Dynamics*, 69:701–717, 2019.

[143] Wilford F Weeks and Steven F Ackley. The growth, structure, and properties of sea ice. In *The geophysics of sea ice*, pages 9–164. Springer, 1986.

[144] Jérôme Weiss. *Drift, deformation, and fracture of sea ice: A perspective across scales*, volume 83. Springer, Dordrecht, Netherlands, 2013.

[145] John Weiss. The dynamics of enstrophy transfer in two-dimensional hydrodynamics. *Physica D: Nonlinear Phenomena*, 48(2-3):273–294, 1991.

[146] David C Wilcox. Multiscale model for turbulent flows. *AIAA journal*, 26(11):1311–1320, 1988.

[147] Karen Willcox and Jaime Peraire. Balanced model reduction via the proper orthogonal decomposition. *AIAA journal*, 40(11):2323–2330, 2002.

[148] Robert E Wolfe, David P Roy, and Eric Vermote. MODIS land data storage, gridding, and compositing methodology: Level 2 grid. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4):1324–1338, 1998.