

Characterizing (Sub)mesoscale Ocean Scale Interactions Through Deep Learning-Enhanced Satellite Oceanography

Scott Andrew Martin

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

Doctor of Philosophy

University of Washington
2026

Reading Committee:
Georgy E. Manucharyan, Chair
LuAnne Thompson
Patrice Klein

Program Authorized to Offer Degree:
School of Oceanography

© Copyright 2026

Scott Andrew Martin

University of Washington

Abstract

Characterizing (Sub)mesoscale Ocean Scale Interactions Through Deep Learning-Enhanced Satellite
Oceanography

Scott Andrew Martin

Chair of the Supervisory Committee:

Georgy E. Manucharyan
School of Oceanography

The ocean circulation emerges from interacting motions across a vast range of spatial scales, from basin-scale gyres and mesoscale eddies hundreds of kilometers across to kilometer-scale submesoscale fronts and ultimately three-dimensional turbulence. Although each scale is governed by distinct dynamics, understanding ocean circulation requires considering their mutual interactions. This dissertation focuses on the mesoscale–submesoscale transition, where flows shift from being largely geostrophic and two-dimensional to ageostrophic and three-dimensional, and where cross-scale energy pathways regulate how energy is generated, transferred, and ultimately dissipated in the global ocean.

Our understanding of ocean scale interactions must ultimately be grounded in observations. Yet characterizing (sub)mesoscale interactions is profoundly challenging, requiring kilometer-scale resolution alongside synoptic coverage over $O(1000 \text{ km})$ scales. Satellites provide increasingly detailed observations of sea surface height (SSH), temperature (SST), and salinity (SSS). However, inferring (sub)mesoscale surface currents from these measurements is a major challenge. Put simply, this dissertation aims to estimate surface ocean currents from satellites at significantly higher resolution than previously possible and use this to characterize (sub)mesoscale ocean scale interactions. This has been enabled by a convergence of new submesoscale-resolving satellites, high-resolution simulations, and advances in deep learning, creating new opportunities to characterize multiscale ocean dynamics from space.

In Chapter 1, we develop a deep learning framework to reconstruct gridded SSH from sparse satellite altimetry and co-located SST. Tested in the challenging Gulf Stream Extension region, the method reduces SSH mapping error by 17%, resolves spatial scales 30% smaller than objective analysis, and yields surface geostrophic currents that more closely match drifter observations, allowing us to better diagnose crucial diagnostics of mesoscale eddy dynamics.

In Chapter 2, we scale this approach to produce Neural Ocean Surface Topography (NeurOST), the first global deep learning-based high-resolution SSH and surface current dataset with state-of-the-art accuracy and resolution, and use this dataset to evaluate the role of the mesoscale kinetic energy cascade in driving seasonality at larger scales.

In Chapter 3, we develop a generative data assimilation framework (GenDA) that combines a diffusion prior trained on high-resolution simulations with sparse satellite observations to jointly reconstruct multi-variate, eddy-resolving surface states. By pursuing a generative approach where the model is trained to generate physically plausible multi-variate surface states from a simulation, we are able to target variables and scales not accessible using NeurOST, objective analysis, or physics-based data assimilation.

In Chapter 4, we extend this framework to submesoscale-resolving state estimation by retraining the prior on the submesoscale-permitting LLC4320 simulation and incorporating temporal coherence via a video diffusion model. By assimilating high-resolution SST and SSH from the recently-launched SWOT mission, we develop the first submesoscale-resolving surface state estimate in the Agulhas Retroflection, providing an unprecedented view of surface ocean dynamics and allowing us to provide observational constraints on the (sub)mesoscale kinetic energy cascade.

This dissertation delivers a step change in our ability to estimate the surface ocean state from satellite observations, suggests an increasing role for deep learning methods in ocean state estimation, and provides crucial observational constraints on (sub)mesoscale ocean scale interactions.

Acknowledgements

Behind the perhaps dry prose of this dissertation lies a formative five years of my life for which I feel immensely grateful, and for which I will now proceed to provide a series of inevitably insufficient acknowledgments.

First, to my advisor Georgy Manucharyan. Thank you for taking a chance on this physics graduate with a vaguely-articulated desire to study the ocean and a skepticism of deep learning methods and turning him into a card-carrying physical oceanographer with a penchant for neural networks. Your energy and drive pushed me to reach for my highest levels, while your personal compassion and trust in me gave me the space to develop my own instincts and process without fear of failure, and to build a fulfilling life outside of work. Each time I sit down to write I remember our long sessions together with my draft of that first paper with you always asking... ‘what are you actually trying to say here?’. The direction you set me off in for my first project has taken on a whole life of its own and your instincts for picking an important problem of lasting impact have set me up for success as I embark on my research career. I am immensely proud of the research we have carried out together. Thank you for turning me into the researcher I am today.

Second, to my co-advisor Patrice Klein. Getting to learn from you and to work with you these past five years has been one of the most joyful and unexpected aspects of my Ph.D. The curiosity-driven way you go about research is inspiring. I have left every meeting we have had feeling re-energized to do science. You have an uncanny ability to see straight to the big picture and I leave our meetings with a laundry list of exciting results that had apparently been staring me in the face in my figures. Despite having been deeply immersed in the literature for many years, you approach established ideas with an eye-opening degree of open-mindedness. You have proven to me beyond doubt that scientific research can be fruitful and of the highest quality while being thoroughly enjoyable along the way.

To my other committee members: LuAnne Thompson and Ginger Armbrust. Thank you for being there for me throughout this whole process and for constantly pushing me to articulate why I am doing what I am doing, who would care if I managed to do it, and for pushing me to be thoughtful and intentional about the career I want to build. Thanks to Aaron Donohoe for saving the day and stepping in as the GSR on my committee at a week's notice. And thanks to Arthur Nowell for spending many enjoyable sessions helping me navigate the writing minefields of my first paper and my dissertation proposal.

To the members of the Manucharyan lab group past and present: Andrei, Maddy, Brandon, Channing, Kitty, Kalyan, and an endless stream of wonderful undergraduate RAs too numerous to list here. Thank you all for building a welcoming environment to do science, for humoring my repeated basic questions about your research, and for helping me to tighten my arguments with your insightful questions and comments. Thanks to David Darr for providing round-the-clock and cheerful support on our department computing cluster, I feared for a while you still may not have retired by the time I wrote this! A particular thanks goes to my undergraduate RA mentees: Shrima, Roy, Dylan, & Maya and to my intrepid high-school mentee Niles. I immensely enjoyed working with each and every one of you and I look forward to watching from afar as you go on to achieve great things!

Ultimately, all of the research presented in this thesis was made possible thanks to the generosity of the American taxpayer. This research was funded in its entirety by NASA through Award 80NSSC21K1187 from the Ocean Surface Topography Science Team. In addition, I am grateful to have received additional funds from the Theodore & Marie Sarchin Endowed Fellowship and from the UW College of the Environment Travel Grant.

My Ph.D. career has been greatly enriched by a network of external collaborators and friends from other institutions. Thanks to Jinbo Wang for being an early advocate of our work and for helping us to translate this research into a NASA data product for the broader community. Thanks to Georgy, Jinbo, and Dhruv Balwada for many stimulating discussions during the writing of a review paper on deep learning for ocean dynamics. Thanks to Lia Siegelman, Andy Thompson, Dimitris Menemenlis, Sammy Metref, Julien Le Sommer, and David Marshall for hosting me on wonderful visits to your respective institutions. Particular thanks go to Sammy Metref and Maxime Ballarotta, your efforts developing open source community benchmarks for satellite altimetry have accelerated the research presented in this thesis by months if not years. Thanks

to Quentin Febvre, Pierre Garcia, and Tatsu Monkman for engaging research collaborations. Thanks to J. Xavier Prochaska and Peter Cornillon for enjoyable discussions and collaborations on the role of AI in oceanography these past few years. Thanks to the many friends and fellow students I have had the good fortune to get to know at conferences these past years, you know who you are. Finally, thanks to my undergraduate research mentor Caroline Terquem and all of the Physics tutors at University College, Oxford for your foundational training.

To the NVIDIA AI for Weather and Climate Simulation team (Mike Pritchard, Noah Brenowitz, Dale Durran, Jaideep Pathak, Suman Ravuri, Peter Manshausen, and Zeyuan Hu) for hosting me for a stimulating and enjoyable research internship in summer 2025. You all inspire me every day to build bigger and better things. Particular thanks to my project mentors Mike, Dale, and especially Noah, whose technical tips and tricks have fed back into the research presented in this thesis.

Beyond science, this dissertation would not have been possible without the friends and family who constantly support me in all that I do. To all of the wonderful friends I have made in UW Oceanography, thank you for building and maintaining a community that made me feel welcome from day one, especially to my cohort, I feel lucky to have shared many wonderful memories with many of you backpacking in the wilderness of WA. To my friends back in the UK for keeping me grounded, making fun of my shape-shifting accent, and for flying thousands of miles to spend time with me in my new locale. Thanks to my family near and far, though mostly far unfortunately, for your unwavering love and support. To my brother Craig for, I hate to admit it, being a role model as I followed a path you'd already laid and to both Craig and Caitlin for being there for me from the moment I landed so far from home. Finally, to my parents for instilling in me a love of science, music, and the outdoors which continue to define me and for not freaking out when I abruptly decided mid-pandemic to move to a different continent.

And finally... to Yuna. Words cannot describe what you mean to me. You were, by quite some distance, the greatest discovery of this Ph.D. Only you have seen and shared in all of the backstory to this thesis. You are the hidden hand helping shape all of what follows and I feel privileged to get to see the world through your eyes.

DEDICATION

To those who design, build, and advocate for ocean-observing satellite missions. However sophisticated our models, they are only as powerful as the observations that ground them.

Contents

Introduction	21
1 Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies	29
1.1 Introduction	30
1.1.1 Background	30
1.1.2 Previous work applying DL to SSH mapping	33
1.1.3 Problem statement and paper structure	34
1.2 Rationale for our proposed methodology	36
1.2.1 Dynamical motivation for supplementing SSH observations with SST	36
1.2.2 Rationale for deep learning	37
1.3 Datasets	38
1.3.1 Sea surface height (SSH) observations	38
1.3.2 Sea surface temperature (SST) observations	39
1.3.3 Data pre-processing: domain and coordinate system	39
1.4 Deep learning method for SSH mapping	40
1.4.1 Deep learning architecture	40
1.4.2 Formulating a loss function in the absence of gridded SSH data	42
1.4.3 Choice of training, validation, and testing datasets	43
1.5 Training performance	43

1.5.1	Determining the optimal length of the input time series and number of training ex- amples	44
1.5.2	Quality of mapped SSH field	45
1.6	Inter-comparison of existing SSH mapping methods in the Gulf Stream Extension region . .	45
1.6.1	Mapping methods and evaluation metrics	45
1.6.2	SSH validation against independent satellite observations	48
1.6.3	Surface geostrophic currents	52
1.6.4	New mesoscale eddy found by synthesizing SST and SSH observations	55
1.7	Discussion and conclusions	56
1.7.1	Summary	56
1.7.2	Physical implications of improved mesoscale SSH field	56
1.7.3	Potential for development of a global DL SSH product	57
1.8	Full performance metric comparison	59
2	Deep learning improves global satellite observations of ocean eddy dynamics	71
2.1	Introduction	71
2.2	Methods	74
2.2.1	NeurOST: Global SSH Maps from Altimetry and SST Using Deep Learning	74
2.2.2	Estimating Surface Currents from SSH	75
2.2.3	SSH Mapping Evaluation: Observing System Experiment (OSE)	75
2.2.4	Eddy Dynamics Evaluation: Observing System Simulation Experiment (OSSE) . . .	75
2.2.5	Kinetic Energy Cascade Diagnosis	76
2.3	Results	76
2.3.1	State-of-the-Art Global SSH Maps Using Deep Learning	76
2.3.2	Improved Physical Realism of Mesoscale Eddies	77
2.3.3	Seasonal Kinetic Energy Cascade	79
2.4	Conclusions	80

A	Supporting Information for Chapter 2	87
A1	Extended Methods	87
A1.1	Sea surface height and surface geostrophic currents	87
A1.2	Satellite datasets	88
A1.3	Formulating SSH interpolation as a self-supervised deep learning problem	88
A1.4	Deep learning neural network architecture	90
A1.5	Merging subdomain reconstructions to create global SSH product	91
A1.6	Regional fine-tuning experiment	92
A1.7	SSH map evaluation and inter-comparison OSE	92
A1.8	Season definitions	95
A1.9	Eddy kinetic energy	95
A1.10	Relative vorticity, strain rate, and Okubo-Weiss quantity	95
A1.11	Eddy Dynamics Evaluation OSSE	96
A1.12	KE cascade: diagnosing energy transfer between scales	97
A1.13	KE Wavenumber Spectrum Calculation	99
A2	Effective coarse-graining scale of SSH maps	100
3	Generative data assimilation for surface ocean state estimation from multi-modal satellite observations	115
3.1	Introduction	115
3.1.1	Background & Motivation	115
3.1.2	Our Contributions	121
3.2	Methods	122
3.2.1	GenDA: Generative Data Assimilation	122
3.2.2	UNet Regression: Baseline Supervised Learning Framework	127
3.3	Datasets & Experiment Set-Up	128
3.3.1	Oceanic & Atmospheric Reanalysis Products for Training and Evaluation	128
3.3.2	Experiment 1: Observing System Simulation Experiment (OSSE)	130
3.3.3	Experiment 2: Observing System Experiment (OSE)	134

3.4	Results	136
3.4.1	Experiment 1 (OSSE): Performance on Simulated Satellite Observations	137
3.4.2	Experiment 2 (OSE): Real-World Mapping Performance	142
3.5	Conclusions & Discussion	145
3.5.1	Conclusions	145
3.5.2	Discussion	145
3.6	Complete Ocean Data Challenge 2021a SSH Evaluation Metrics (OSE)	148
B	Supporting Information for Chapter 3	167
B1	Linear Ekman Regression Model for Wind-Driven Surface Currents	167
B2	Why Include Surface Winds in the State Vector?	168
B3	Neural Network Architectures and Hyperparameters	168
B3.1	GenDA Diffusion Prior Architecture	168
B3.2	GenDA Generation Hyperparameters	169
B4	GLORYS Diffusion Prior Training Results	169
B5	Coarse-Graining Scales in GenDA Observation Operator	170
B5.1	OSSE: Scales Prescribed A Priori	170
B5.2	OSE: Scales Tuned A Posteriori	171
B6	Weighted Re-Sampling Scheme for Comparing Errors For Cloudy vs Non-Cloudy SST and SWOT vs No SWOT SSH	172
B7	OSE Ageostrophic Eddy Cyclo-Geostrophy Case Study	173
B8	OSE Comparison of Ageostrophic Currents to Surface Drifters	174
4	Observational constraints on the submesoscale ocean kinetic energy cascade from generative data assimilation	181
4.1	Introduction	181
4.2	Methods	184
4.2.1	Score-based data assimilation	184
4.2.2	Learning surface ocean dynamics through video diffusion	185

4.2.3	Submesoscale-permitting ocean model simulation training data	187
4.2.4	Observation operator for submesoscale surface ocean state estimation	189
4.3	Experiments	190
4.3.1	Observing system simulation experiment	190
4.3.2	Application to real world satellite observations	192
4.4	Results	192
4.4.1	Generative data assimilation enables submesoscale-resolving surface ocean state es- timation	192
4.4.2	State estimates capture the strength and seasonality of the submesoscale cascade in LLC4320	198
4.4.3	Real world observations support the existence of a strong, seasonal upscale cascade at submesoscales	204
4.5	Discussion	208
C	Supporting Information for Chapter 4	215
C1	Extended Methods	215
C1.1	Video diffusion prior architecture and training details	215
C1.2	Score-based data assimilation hyperparameters	217
C1.3	Coarse-graining computation of kinetic energy cascade using FlowSieve	217
C1.4	Details of the satellite data products used in this study	218
C2	Validity of estimating the submesoscale kinetic energy cascade in small spatio-temporal windows	220
C3	Additional evaluations of GenLLC4320 surface state estimation	222
C3.1	Testing the effect of super-inertial SSH observation contamination on GenLLC4320 state estimates	222
5	Conclusions	225
	Conclusions	225
5.1	Summary	225

5.2	Discussion & Outlook	228
5.2.1	NeurOST: Broader Implications, Future Applications, & Refinements	228
5.2.2	Generative Data Assimilation	229
5.2.3	Ocean Scale Interactions	230
5.2.4	Final Thoughts	231

List of Figures

1.1	Study region and data availability	35
1.2	Deep learning method schematic	61
1.3	Train, validation, test partition of dates	62
1.4	Effect of increasing temporal context window	62
1.5	Example ConvLSTM SSH+SST output	63
1.6	SSH mapping performance violin plot	64
1.7	Example along-track profile	65
1.8	Along-track spectral SSH metrics	66
1.9	Temporal spectral characteristics	66
1.10	Seasonality of SSH mapping performance	67
1.11	Mesoscale eddy dynamics diagnosed from ConvLSTM SSH+SST	68
1.12	Geostrophic vorticity comparison between methods	69
1.13	Eddy kinetic energy and enstrophy comparison between methods	70
1.14	New eddy revealed thanks to SST	70
2.1	NeurOST method schematic	82
2.2	Global SSH mapping metrics	83
2.3	Snapshots of mesoscale eddy dynamics from NeurOST compared to DUACS	84
2.4	Observing system simulation experiment using GLORYS data	85
2.5	Seasonal mesoscale kinetic energy cascade observed in subtropics in NeurOST	86
A.1	Regional SSH mapping metrics method comparison across a range of regions	102

A.2	SSH mapping performance comparison to DUACS for 2-altimeter observing system	103
A.3	Comparison of NeurOST against regional SSH mapping schemes in Gulf Stream Extension .	103
A.4	Effective resolution and effective coarse-graining scales of SSH maps	104
A.1	Global maps of SSH mapping performance compared to MIOST	105
A.2	Validation of surface geostrophic currents against drifters	106
A.3	Global maps of SSH mapping performance compared to GLORYS	106
A.4	Global maps of change in seasonality of eddy dynamics in NeurOST compared to DUACS .	107
A.5	Absolute values of seasonal eddy dynamics diagnostics	107
A.6	Seasonality of mesoscale kinetic energy cascade across a range of regions	108
A.7	Comparison of NeurOST kinetic energy cascade to GLORYS	109
A.8	Partitioning of dates into training, validation, and testing	109
A.9	Improved agreement of surface currents with drifter observations when cyclogeostrophic balance is applied	110
A.10	Learning curves for NeurOST global training	111
A.11	Validating absence of interpolation artifacts using enstrophy spectrum	112
A.12	Effect of including SST in NeurOST on diagnosed mesoscale kinetic energy cascade	113
3.1	Multi-modal satellite observational sampling of N. Atlantic	150
3.2	Schematic of generative data assimilation training phase	151
3.3	Schematic of generative data assimilation inference phase	152
3.4	Multi-modal surface state estimation GLORYS observing system simulation experiment setup	153
3.5	Example GenDA predictions on synthetic observations	154
3.6	Magnitude of SST gradients of reconstructed fields from synthetic satellite observations . . .	155
3.7	Spectra of GenDA reconstructions compared to GLORYS ground truth	156
3.8	Geostrophic joint vorticity-strain PDFs	157
3.9	Geostrophic Okubo-Weiss reconstructed from simulated observations	158
3.10	Generative data assimilation channel synthesis capabilities	159
3.11	GenDA ensemble spread	160
3.12	GenDA rank histograms	161

3.13	GenDA predictions from real world satellite observations	162
3.14	Evaluation of GenDA real world state estimates against withheld satellite observations . . .	163
3.15	GenDA real world state estimate spectra	164
3.16	Geostrophic Okubo-Weiss quantity reconstructed from real world satellite observations . . .	165
B.1	GenDA UNet backbone architecture schematic	176
B.2	GLORYS diffusion learning curves	176
B.3	GenDA OSSE spectra	177
B.4	A posteriori tuning of L4 observation coarse-graining scales in real world setting	177
B.5	Tuning SST coarse-graining scale using cloud-free snapshots	178
B.6	Sensitivity of GenDA spectra to varying coarse-graining scales	179
B.7	Comparison of ageostrophic currents to cyclogeostrophic balance	180
4.1	Illustration of training domain.	188
4.2	Example GenLLC4320 state estimation in OSSE	193
4.3	GenLLC4320 ensemble spread in relation to SWOT sampling	194
4.4	GenLLC4320 OSSE state estimation metrics	195
4.5	Example GenLLC4320 state estimate using real observations	196
4.6	Real world GenLLC4320 state estimation metrics	197
4.7	GenLLC4320 eddy dynamics in the Ring Path in OSSE	199
4.8	GenLLC4320 eddy dynamics in the Return Current in OSSE	200
4.9	Enstrophy spectra in the OSSE setting	201
4.10	Comparison of GenLLC4320 estimated submesoscale kinetic energy cascade and the LLC4320 ground truth	202
4.11	Co-location of the downscale cascade at small scales with strong buoyancy fronts	203
4.12	Eddy dynamics reconstructed from real world observations in the Ring Path region	205
4.13	Eddy dynamics reconstructed from real world observations in the Return Current region . . .	206
4.14	Enstrophy spectra of real world reconstructions and their seasonality	207
4.15	Real world submesoscale kinetic energy cascade and its seasonality from GenLLC4320 . . .	209

C.1	Testing validity of estimating submesoscale cascade in small space-time windows	221
C.2	Effect of assimilating super-inertial SSH on state estimation metrics in OSSE setting	222

List of Tables

1.1	SSH mapping metrics comparison between methods	60
3.1	GenDA GLORYS simulated observations reconstruction accuracies	139
3.2	Ocean Data Challenge 2021a SSH mapping metrics for the OSE. See Section 3.3.3 for metric details.	149
B.1	Effect of assimilating ERA 5 winds on GenDA (ensemble mean) reconstruction skill for ageostrophic currents.	174
B.2	Gaussian coarse-graining scales applied to SSH, SST, and SSS to generate simulated low-resolution L4 satellite products in the OSSE.	175
B.3	Gaussian coarse-graining scales applied to SSH, SST, and SSS in GenDA observation operator in OSE after a posteriori tuning.	175
B.4	Evaluation of surface ocean currents compared against surface drifters for the OSE.	175
C.1	Training and architecture hyperparameters	216
C.2	SDA inference hyperparameters	217
C.3	Coarse-graining scales for coarse satellite products	219

Introduction

... the oceans are fully turbulent, involving a broad range of scales from at least 2 km to 5,000 km. All these scales are now known to strongly interact, leading to significant energy exchanges between scales, in particular in the upper ocean. Resulting ocean scale interactions impact the Earth climate in counterintuitive ways.

— Klein et al. [2019]

Ocean dynamics span orders of magnitude in space and time. Ocean circulation is sustained by interactions between energy sources and sinks operating at vastly different scales [Ferrari and Wunsch, 2009]. At the largest scales, wind forcing energizes the basin-scale circulation; at the smallest, boundary friction and diffusion drive irreversible energy dissipation. Between these limits is a continuum of motions from mesoscale eddies ($\mathcal{O}(100 \text{ km})$) to submesoscale eddies and fronts ($\mathcal{O}(1\text{-}10 \text{ km})$) and smaller-scale turbulence ($\mathcal{O}(<1 \text{ km})$) that cascade energy between scales [Scott and Wang, 2005; Sasaki et al., 2014; Klein et al., 2019] and regulate exchange with the atmosphere [Seo et al., 2023; Strobach et al., 2022]. Characterizing the interactions between the ocean’s different scales, and their dynamical drivers is thus essential for the development of accurate theoretical and numerical models of the global ocean and its role in climate.

Submesoscale–mesoscale interactions form a critical yet poorly understood link in the ocean energy cycle, linking forcing and dissipation scales. At mesoscales, $\mathcal{O}(100 \text{ km})$ in spatial scale, the ocean circulation is dominated by mesoscale eddy turbulence, comprising a rich field of coherent, mostly two-dimensional, eddies which together make up 80% of the kinetic energy of the global ocean circulation [Ferrari and Wunsch, 2009]. Mesoscale eddies are predominantly formed through baroclinic instability of large-scale ocean currents [Tulloch et al., 2011], and they are mostly in geostrophic balance. Theory predicts that non-linear interactions between geostrophic eddies should drive an ‘upscale’ kinetic energy cascade [Salmon, 1980], transferring energy from smaller to larger scales, which is the reverse of the downscale cascade familiar from

3D turbulence. This upscale cascade at mesoscales has been confirmed observationally [Scott and Wang, 2005], emphasizing that kinetic energy at mesoscales feeds back to the large-scale circulation, an effect that requires parameterizations in coarse climate simulations where mesoscale eddies are typically not explicitly resolved [Hewitt et al., 2020].

At submesoscales ($\mathcal{O}(1\text{--}10\text{ km})$), increasingly ageostrophic balanced eddies, fronts, and filaments coexist with unbalanced internal gravity waves [Taylor and Thompson, 2023]. While the direction of the kinetic energy cascade at scales on either side of the submesoscale is clear — downscale at smaller scales due to the increasingly 3D nature of the flow and upscale at mesoscales through interactions between geostrophic eddies — the diversity of dynamical regimes that proliferate at submesoscales means it can support both an upscale and downscale cascade, with both likely coexisting and competing in a ‘dual’ cascade [Balwada et al., 2022]. Results from high-resolution numerical simulations suggest that interactions between coherent submesoscale eddies drive an upscale cascade [e.g. Capet et al., 2008a; Schubert et al., 2020] which potentially plays a significant and previously under-appreciated role in energizing mesoscale eddies [Sasaki et al., 2014; Qiu et al., 2014]. Meanwhile, three-dimensional ageostrophic motions at fronts and filaments [D’asaro et al., 2011; Schubert et al., 2020; Srinivasan et al., 2023] and potentially interactions between eddies and internal waves [Barkan et al., 2017] appear to drive a downscale cascade. Characterizing the crucial submesoscale-mesoscale link in the energy cascade demands quantifying how the balance between upscale and downscale submesoscale processes varies geographically and seasonally and how this interacts with the mesoscale.

One striking characteristic of submesoscale flows noted in both simulations and observations is a strong seasonal cycle, with submesoscale kinetic energy peaking in the wintertime [e.g. Sasaki et al., 2014; Callies et al., 2015; Thompson et al., 2016]. This seasonality is largely driven by enhanced surface cooling and wind-driven mixing during the winter which deepen the mixed layer, setting up a reservoir of available potential energy which is converted into submesoscale kinetic energy by instabilities such as mixed layer instability, restratifying the mixed layer [Boccaletti et al., 2007; Taylor and Thompson, 2023]. This strong seasonal injection of kinetic energy at submesoscales points to a complex and seasonally varying ocean energy cycle. Numerical simulations suggest that a substantial fraction of the kinetic energy injected at submesoscales in wintertime undergoes an upscale cascade to mesoscales, leading to peaks in kinetic energy

at progressively larger scales throughout the winter-summer transition [Sasaki et al., 2014; Qiu et al., 2014; Uchida et al., 2017; Schubert et al., 2020; Ajayi et al., 2021; Schubert et al., 2023]. This seasonal upscale cascade from submesoscales to mesoscales could partly explain the observed seasonality of kinetic energy at mesoscales, where energy peaks in the summertime in many regions [Sasaki et al., 2014; Qiu et al., 2014]. Disentangling the submesoscale and mesoscale contributions is complicated in some regions by a seasonal modulation of mesoscale baroclinic instability leading to a springtime injection of kinetic energy at the lower end of the mesoscale range [Qiu et al., 2014].

While simulations point to a seasonal (sub)mesoscale upscale kinetic energy cascade being a robust feature of the ocean circulation across regions, these findings must ultimately be verified in observations. This is particularly important because even high-resolution numerical models typically do not fully resolve the full range of processes required to constrain the (sub)mesoscale energy cycle in a realistic setting. For example, models typically under-resolve the mechanisms which drive submesoscale eddy formation — such as mixed layer instability, symmetric instability, and shear instability — which feed back to larger scales through the upscale cascade. Meanwhile, the frontal instabilities or eddy-wave interactions which could drive a competing downscale cascade of energy are typically under-resolved [D’asaro et al., 2011; Barkan et al., 2017; Srinivasan et al., 2023]. Parameterizations for small-scale surface boundary layer turbulence could have subtle impacts on the strength and seasonality of the submesoscale kinetic energy cascade by misrepresenting scale interactions between submesoscale and small-scale boundary layer turbulence. Finally, running global ocean simulations with submesoscale resolution for long enough to obtain an equilibrated view of the mutual interactions between submesoscales and mesoscales under realistic forcing and coupling to the atmosphere is unlikely to be computationally feasible in the near term. It is thus critical to characterize (sub)mesoscale ocean scale interactions using observations of the real-world ocean.

Observing (sub)mesoscale ocean scale interactions is immensely challenging. To properly characterize the interactions between mesoscale and submesoscale turbulence, for instance by diagnosing the strength of the kinetic energy cascade, one needs to simultaneously observe the ocean with sufficiently fine spatiotemporal resolution to resolve submesoscales but with broad coverage over large enough domains to constrain the mesoscale state. Obtaining such a high-resolution yet synoptic view of ocean currents is typically beyond the capabilities of even the most intense and coordinated in situ campaigns. Some regional studies

have sought to quantify the kinetic energy cascade in the submesoscale range from in situ observations using drifters [Balwada et al., 2022] and moorings Naveira Garabato et al. [2022], though the limited sampling in both cases makes it challenging to directly and robustly diagnose the transfer of energy between scales and the local nature of these measurements prevents extending the analysis into the mesoscale range. In contrast, satellites provide a synoptic view of the surface ocean with increasingly fine spatial resolution thanks to continued advances in remote sensing. Satellite oceanography is thus a powerful tool for the study of (sub)mesoscale ocean scale interactions [Klein et al., 2019].

Satellites observe the dynamics of the surface ocean through a range of modalities, each with differing degrees of sparsity and sensor resolution limits. Infrared radiometers observe sea surface temperature (SST) with resolution down to 1 km over broad swaths, providing detailed snapshots of the (sub)mesoscale surface state. However, clouds frequently occlude the sensors causing large gaps in the observations, and in the summertime enhanced warming from air-sea interactions imprints large-scale SST variability that can overwhelm the signature of (sub)mesoscale ocean dynamics. Microwave radiometers alleviate some of the sparsity by penetrating through clouds, but they observe SST with lower resolution (20-30 km) and so cannot be used to observe the submesoscale. Satellite altimeters observe sea surface height (SSH), variations in sea level relative to the geoid that are driven by ocean dynamics. Assuming geostrophic balance, surface ocean currents can be derived from the gradient of SSH, though this requires access to continuous 2D maps of SSH and the approximation becomes increasingly invalid at submesoscales. Satellite altimeters have until recently only observed SSH along sparse 1D tracks with along-track resolutions of 30-70 km [Dufau et al., 2016], making estimation of even mesoscale surface currents a significant challenge. In 2022, the Surface Water and Ocean Topography (SWOT) mission was launched [Morrow et al., 2019], providing novel submesoscale-resolving wide-swath observations of SSH, allowing us to probe submesoscale SSH variability, albeit within a limited 150 km swath. Finally, satellites observe sea surface salinity (SSS) which, together with SST, sets the density of the surface ocean. SSS is a more challenging measurement to make from space than SST, and thus the spatial resolution of satellite SSS observations is considerably coarser.

While satellite observables such as SSH, SST, and SSS contain rich signatures of (sub)mesoscale ocean dynamics, characterizing the kinetic energy cascade requires gap-free, high-resolution estimates of surface currents. Since satellites do not yet directly observe surface currents, we require methods to infer the un-

derlying dynamical state, including surface currents, from incomplete observations — a process we refer to hereafter as ‘surface ocean state estimation’. There are two broad families of methods that have conventionally been used for surface ocean state estimation: *objective analysis* and *data assimilation*.

Objective analysis involves applying simple statistical interpolation algorithms to fill in the gaps in sparse observations based on some assumed covariance structure for the data [Bretherton et al., 1976]. A common framework across observational oceanography is to use the best linear unbiased estimator of the field assuming a Gaussian covariance model with decorrelation length and time scales that are prescribed a priori. Variants of this approach are used to develop most widely-used gridded satellite data products such as the AVISO DUACS gridded SSH product [Le Traon et al., 1998; Taburet et al., 2019]. Applying geostrophic balance to the DUACS SSH product to infer surface geostrophic currents has been the prevailing approach to estimating surface ocean currents, however objective analysis artificially smooths out small-scale SSH variability in the observations [Ballarotta et al., 2019] which has raised questions about the validity of estimates of the mesoscale kinetic energy cascade obtained from DUACS [Scott and Wang, 2005; Arbic et al., 2013]. More fundamentally, objective analysis provides no method for inferring unobserved variables without direct physical relationships to observed fields, preventing diagnosis of the ageostrophic component of surface currents which becomes increasingly relevant at (sub)mesoscales.

Data assimilation (DA) overcomes many of the limitations of objective analysis by utilizing a dynamical model to constrain the state estimate rather than relying upon simple covariance models. Broadly speaking, DA involves guiding a numerical model of ocean dynamics, typically a GCM, using sparse observations to produce a state estimate that respects the physics of the model while matching available observations. Beyond giving a physically self-consistent state estimate, DA allows to synergize observations of diverse variables in a unified way and allows access to unobserved variables such as ageostrophic surface currents. Despite these numerous strengths, it is challenging to constrain (sub)mesoscale ocean scale interactions using DA due to the significant computational cost of performing DA with submesoscale-resolving ocean GCMs. More fundamentally, in order to satisfy the GCM physics DA must estimate the full 3D ocean state which is a substantial challenge at submesoscales since we typically only have high-resolution observations at the surface. Finally, since GCMs are imperfect representations of complex real-world ocean dynamics DA will produce ocean state estimates that retain numerous model biases.

Recent advances in deep learning, coupled with the advent of accelerated computing have spurred broadening adoption of purely data-driven deep learning methods in the natural sciences. Briefly, deep learning is built on neural networks which are complex functions with millions of free learnable parameters whose multi-layer structure allows them to approximate functions of arbitrary complexity, mapping some input to a desired output given large volumes of training data. In recent years, deep learning has driven fundamental advances in our ability to model complex natural phenomena, with protein folding [Jumper et al., 2021] and global weather forecasting [Pathak et al., 2022; Lam et al., 2023] being prominent examples. In the context of surface ocean state estimation, neural networks offer an alternative approach to overcome the respective limitations of objective analysis and DA. By being able to model complex and high-dimensional data distributions, deep learning can leverage non-linear structure in ocean fields not utilized by objective analysis while avoiding the restrictions from imposing hard physical constraints in DA.

The research presented in this thesis has been conducted at an exciting moment of opportunity for (sub)mesoscale-resolving surface ocean state estimation, powered by a convergence of new submesoscale-resolving satellite missions and the development of powerful deep learning methods. The launch of the SWOT mission in 2022 has provided an unprecedented view of submesoscale surface ocean dynamics from space, providing for the first time submesoscale-resolving SSH observations that allow us to better constrain ocean dynamics than from SST alone [Archer et al., 2025]. Meanwhile, deep learning methods have matured to a point where they are readily applicable to a wide range of scientific domains. Further, the advent of generative deep learning models [Goodfellow et al., 2014; Ho et al., 2020; Lipman et al., 2022], which can generate new samples from complex high-dimensional data distributions, opens new modeling capabilities which remain largely unexplored in physical oceanography. Given this context, the broad aim of this thesis is to utilize novel deep learning methods and submesoscale-resolving satellite oceanography to better characterize (sub)mesoscale ocean scale interactions.

The primary scientific objective of this thesis is to provide observational constraints on the strength and seasonality of the (sub)mesoscale kinetic energy cascade. High-resolution numerical simulations suggest the cascade provides a route for submesoscale dynamics to feed back to larger scales, with the wintertime injection of kinetic energy at submesoscales eventually driving the summertime peak in energy at mesoscales [Sasaki et al., 2014; Qiu et al., 2014]. We ask the following questions:

- Do satellite observations support the existence of a seasonal upscale kinetic energy cascade in the real-world ocean at (sub)mesoscales?
- How strong is the cascade and how does this compare with results from high-resolution simulations?
- To what extent is the seasonality of mesoscale eddies driven by the kinetic energy cascade?

In order to achieve these scientific objectives, we develop two novel deep learning methods for surface ocean state estimation. Put simply, the methodological goal of this thesis has been to estimate surface ocean currents from satellite observations with higher accuracy and resolution than is possible with conventional methods. This thesis is comprised of four chapters, outlined below.

Chapter 1. In this chapter, we develop a novel deep learning method for estimating high-resolution maps of SSH from along-track nadir satellite observations. By leveraging co-located SST observations as an additional input for SSH mapping, we seek to infer small-scale SSH variability in the gaps between altimeter observations that is artificially smoothed by objective analysis. *This chapter is reproduced from: Martin, S. A., Manucharyan, G. E., & Klein, P. (2023). Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies. Journal of Advances in Modeling Earth Systems, 15, e2022MS003589. <https://doi.org/10.1029/2022MS003589>.*

Chapter 2. In this chapter, we scale up the method developed in Chapter 1 to produce the first deep learning-based global maps of SSH and surface geostrophic currents and use them to re-evaluate the strength and seasonality of the mesoscale kinetic energy cascade. We assess the extent to which a seasonal upscale cascade in the small mesoscale range is sufficient to drive the seasonality of larger-scale mesoscale eddies. *This chapter is reproduced from: Martin, S. A., Manucharyan, G. E., & Klein, P. (2024). Deep learning improves global satellite observations of ocean eddy dynamics. Geophysical Research Letters, 51, e2024GL110059. <https://doi.org/10.1029/2024GL110059>.*

Chapter 3. In this chapter we develop a novel generative deep learning method, ‘generative data assimilation’, for surface ocean state estimation from multi-modal satellite observations. Generative data assimilation involves training a generative model on high-resolution numerical simulation output to generate new multi-variate surface ocean states that resemble the simulation training data. By guiding the

generation of this model with sparse, multi-modal satellite observations we seek to resolve smaller scales than can be captured with the approach from Chapters 1 & 2 and to reconstruct unobserved fields such as ageostrophic surface currents. This chapter is reproduced from: *Martin, S. A., Manucharyan, G. E., & Klein, P. (2025). Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. Journal of Advances in Modeling Earth Systems, 17, e2025MS005063. <https://doi.org/10.1029/2025MS005063>.*

Chapter 4. By adapting the generative data assimilation method proposed in Chapter 3, we create submesoscale-resolving surface ocean state estimates and use them to provide observational constraints on the strength and variability of the (sub)mesoscale kinetic energy cascade. By synergizing high-resolution SSH observations from SWOT with satellite SST observations we obtain surface state estimates that capture submesoscale and small mesoscale variability which is smoothed out by the method used in Chapters 1 & 2. We compare the strength and seasonality of the (sub)mesoscale cascade in our observation-guided state estimates with those from prior observational estimates and from high-resolution numerical simulations to provide observational constraints on the (sub)mesoscale cascade. This chapter is in preparation for submission to a journal.

Chapter 1

Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies

This chapter is reproduced from the following publication:

Martin, S. A., Manucharyan, G. E., & Klein, P. (2023). Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies. Journal of Advances in Modeling Earth Systems, 15, e2022MS003589. <https://doi.org/10.1029/2022MS003589>

1.1 Introduction

1.1.1 Background

Sea surface height (SSH) maps - i.e. time-varying gridded maps of the height of the ocean's surface referenced against the geoid - derived from satellite altimetry observations have been a crucial tool for physical oceanographers for many years [Fu et al., 2010; Abdalla et al., 2021]. Under the assumption of geostrophy, SSH is proportional to a streamfunction for the surface currents. Hence a SSH map provides easy access to an estimated map of surface geostrophic ocean currents, which would otherwise be challenging to obtain through direct in-situ observations with global coverage. SSH maps have been used to directly observe the inverse kinetic energy cascade in the global oceans [Scott and Wang, 2005], to track and study the behavior of mesoscale eddies [Chelton et al., 2011; Fu et al., 2010], and to estimate surface eddy diffusivities [Abernathey and Marshall, 2013].

To date, the only global SSH observations have come from satellite-borne nadir radar altimeters, which measure SSH along one-dimensional tracks directly beneath the satellite's orbit resolving along-track scales of $\mathcal{O}(20\text{-}100\text{km})$ [Dufau et al., 2016]. Reconstructing the full two-dimensional (2D) SSH field from altimeter observations requires significant interpolation in space and time. Hence, the length-scales 2D products can accurately resolve are coarser than the along-track resolution of the altimeters [Ballarotta et al., 2019].

The most widely used 2D SSH product is that generated by the 'Data Unification and Altimeter Combination System' (DUACS) [Taburet et al., 2019] and distributed by the 'Copernicus Marine Environment Monitoring Service' (CMEMS) (*note that this is the same product as was formerly distributed by the 'Archiving, Validation and Interpretation of Satellite Oceanographic data' (AVISO) service*). This product is created using the optimal interpolation (OI) method [Bretherton et al., 1976; Le Traon et al., 1998], which is otherwise known as 'objective mapping' or 'objective analysis'. OI provides the best least squares linear estimator of the SSH in the gaps between observations, given an a priori model for the covariance of SSH between different locations and times, and knowledge of the instrument noise covariance. In the DUACS mapping, the SSH covariance is assumed to be a single-scale Gaussian with prescribed decorrelation length- and time-scales that are geographically varying and have been tuned empirically to account for a varying number of available satellite altimeters over the 30 year altimetry record, while the instrument noise

covariance has been chosen to be suitable for each satellite altimeter mission and technology [Taburet et al., 2019].

While the DUACS SSH product continues to be valuable to the oceanography community, the OI method has been shown to introduce significant deficiencies. Amores et al. [2018] used an Observer System Simulation Experiment (OSSE) in which a DUACS-like SSH product was generated using pseudo-observations from the output of a high-resolution ocean model to study how accurately DUACS captured the model’s mesoscale SSH field. They demonstrated that the eddy fields inferred from the DUACS SSH map are significantly distorted, with often multiple smaller eddies being aliased into larger eddies. The real-world accuracy of an SSH mapping method can also be studied by generating a map with one satellite altimeter’s observations withheld and examining the errors of the mapped SSH compared to these independent along-track observations. In this way, Ballarotta et al. [2019] showed that the DUACS product only accurately resolves SSH signals down to time-scales of $\mathcal{O}(30 \text{ days})$ and length-scales $\mathcal{O}(100\text{km})$ at high latitudes, increasing to $\mathcal{O}(800\text{km})$ in the tropics. OI tends to smooth out small-scale, fast-evolving features and strong peaks/troughs in SSH where observations are scarce and the oceanic mesoscale is energetic.

The challenges involved in reconstructing surface currents from nadir satellite altimetry observations has led to the development of a number of methods for inferring surface ocean currents from satellite observations of other surface ocean quantities, most notably sea surface temperature (SST). Satellite-borne infrared sensors allow the observation of 2D snapshots of SST with high spatial resolution of $\mathcal{O}(1\text{km})$ when no clouds obscure the sensors. Since submesoscale SST features with scales $\mathcal{O}(1\text{-}10\text{km})$ are advected by mesoscale surface ocean currents, tracking the motion of submesoscale SST features in consecutive cloud-free SST observations allows for the estimation of the mesoscale surface currents [Emery et al., 1986; Bowen et al., 2002]. Others have attempted to use a sequence of cloud-free SST observations to invert the heat conservation equation to estimate the surface currents which enter this equation through the advection term [Kelly, 1989; Ostrovskii and Piterbarg, 1995; Chen et al., 2008]. Rio et al. [2016] combined the heat conservation approach with background geostrophic surface current estimates from DUACS to reduce the uncertainty in along-isotherm velocities, which cannot be estimated from SST tracer advection. The heat conservation approach requires the approximation of the main source and sink terms in the heat conservation equation. These methods allow for the retrieval of geostrophic currents smaller than the scales resolved by DUACS

as well as ageostrophic currents that cannot be retrieved directly from SSH. However, the need for consecutive cloud-free SST observations severely limits their capability for consistently estimating surface currents across all regions and seasons. Another approach utilizes surface quasi-geostrophic (SQG) theory to estimate surface currents from a single SST snapshot and has been applied with promising results to microwave SST observations, which have lower spatial resolution than infrared but do not suffer gaps due to clouds [Isern-Fontanet et al., 2006, 2014]. However, as will be discussed further in section 1.2.1, this approach relies on particular environmental conditions that are not universally met in all regions and seasons. While the retrieval of surface ocean currents from SST observations is a dynamic and growing research field, SSH maps from satellite altimetry along with their associated surface geostrophic currents remain the primary source of surface ocean current estimates for the ocean science community.

The surface currents derived from the DUACS SSH product are used extensively in oceanographic studies, so it is essential to ensure these inferred currents are as accurate and high resolution as possible using the existing satellite observing capabilities. There is therefore increasing focus on developing better methods for reconstructing the 2D SSH field from satellite altimetry observations [Ubelmann et al., 2015, 2021; Manucharyan et al., 2021; Fablet et al., 2021a; Le Guillou et al., 2021]. Efforts to improve SSH mapping methods have been aided by the the creation in recent years of community-maintained ‘Ocean Data Challenges’[AVISO, 2021]. By establishing a common set of evaluation metrics, a common study region and time, and by sharing code and results these data challenges allow for direct comparison between different SSH mapping methods.

OI can only provide the statistical best *linear* estimator of the unobserved SSH for a given a priori covariance model. While the decorrelation length- and time- scales used in this covariance model can be tuned to best fit the data, as is done to create the DUACS product, a single-scale Gaussian covariance model and its resulting linear estimator is ultimately limited in its ability to accurately represent the non-linear dynamics of mesoscale ocean turbulence. An explicitly dynamics-based approach to SSH mapping would be to use a data assimilation framework to constrain an ocean circulation model to best match the available observations, then use this constrained model’s SSH field as the estimate of 2D SSH. However, such an approach requires additional observations to accurately constrain other essential model variables such as subsurface flow and density which are not typically available at high resolution. Data assimilation using a

full global ocean circulation model is very computationally expensive, thus it is only in recent years that data assimilation with mesoscale eddy-permitting model resolution ($1/12^\circ$) has become feasible [Lellouche et al., 2021]. Despite the computational expense of $1/12^\circ$ ocean model data assimilation, even higher model resolutions may be required before such data assimilation approaches can accurately resolve the mesoscale eddy field. To circumvent the computational challenges of using a full primitive equation ocean model, an approach that uses data assimilation with an idealized one layer quasi-geostrophic (QG) model of ocean turbulence has been effectively applied to SSH mapping [Le Guillou et al., 2021]. Such an idealized model, however, is unable to capture many of the dynamical effects observed in the real-world ocean.

The recent advances in the application of deep learning (DL) methods to the earth sciences [Sonnewald et al., 2021; Sun et al., 2022] have inspired an increasing focus on the possibility of using data-driven interpolation methods to improve SSH mapping [Beauchamp et al., 2020; Manucharyan et al., 2021]. DL models can be trained to approximate highly non-linear mappings from inputs to outputs. Thus it is plausible that a DL model could be trained to recognize dynamical signatures in partial SSH observations and use these to estimate SSH in unobserved regions given sufficient training examples. By training a DL model on a large library of examples of real ocean turbulence the DL model would be able to implicitly use ocean dynamics to create a more accurate SSH mapping than would be possible with OI.

1.1.2 Previous work applying DL to SSH mapping

A number of attempts to use DL for SSH mapping have been published recently [Fablet et al., 2021a,b; Barth et al., 2022; Buongiorno Nardelli et al., 2022]. These studies have mostly focused on the OSSE setting, where the SSH observations are generated from a circulation model in which the full 2D SSH field is known. Fablet et al. [2021a] and Fablet et al. [2021b] developed a DL approach (4DVarNet) for interpolating SSH which formulates SSH mapping as a four-dimensional variational data assimilation (4DVar) problem but replaces the dynamical model and solver with trainable neural networks. They demonstrated that 4DVarNet yields a significant improvement over OI when tested in idealized OSSEs, though the improvement is more modest when applied to real-world observations.

While these results are auspicious, OSSEs can only offer a limited approximation of the real-world SSH interpolation problem since the full 2D SSH ground truth is unknown in real-world settings. It therefore

cannot be used to calculate the reconstruction error to be minimized during training. A number of studies have attempted to address this issue by developing methods that can be trained exclusively on real-world observations in a manner similar to that described here in section 1.4.2. Barth et al. [2020] introduced a Convolutional Neural Network (CNN) method, Data INterpolating Convolutional Auto-Encoder (DIN-CAE), for reconstructing sea surface temperature (SST) from partial satellite observations. In a subsequent paper, Barth et al. [2022] refined and applied the method to reconstruct SSH in the Mediterranean from partial real-world SSH and SST observations. They found that including SST yielded a lower reconstruction error than SSH observations alone. However, an error comparison to the operational DUACS method was not made, so this method’s performance is hard to compare to other SSH mapping methods. Buongiorno Nardelli et al. [2022] developed another CNN approach for reconstructing and super-resolving SSH from partial SSH and SST satellite observations in the Mediterranean with promising results. However, to split the observations into independent training and testing datasets they randomly selected days to withhold for testing, meaning adjacent days could be used for training and testing. SSH and SST fields contain significant temporal autocorrelation over time scales of days to weeks. Hence, such a sampling strategy leads to a testing dataset that could be highly correlated to the training data, calling into question the independence of the performance metrics calculated.

To date, DL SSH mapping studies remain limited to theoretical OSSE studies or limited regional experiments and the gridded SSH product of choice for oceanographers remains the OI-generated DUACS product.

1.1.3 Problem statement and paper structure

In this study, we seek to demonstrate in practice that DL can provide a more accurate SSH mapping with a more physically realistic eddy evolution than was previously possible. Further, our DL approach allows us to include co-located observations of other ocean surface variables in the input to the mapping, something that is challenging with existing SSH mapping methods. Specifically, we will demonstrate that the inclusion of SST observations as an additional source of information for the SSH mapping leads to dramatic improvements in the accuracy and resolution of the reconstructed SSH evolution. In section 1.2, we provide a dynamical motivation for why SST observations are expected to improve SSH mapping and provide

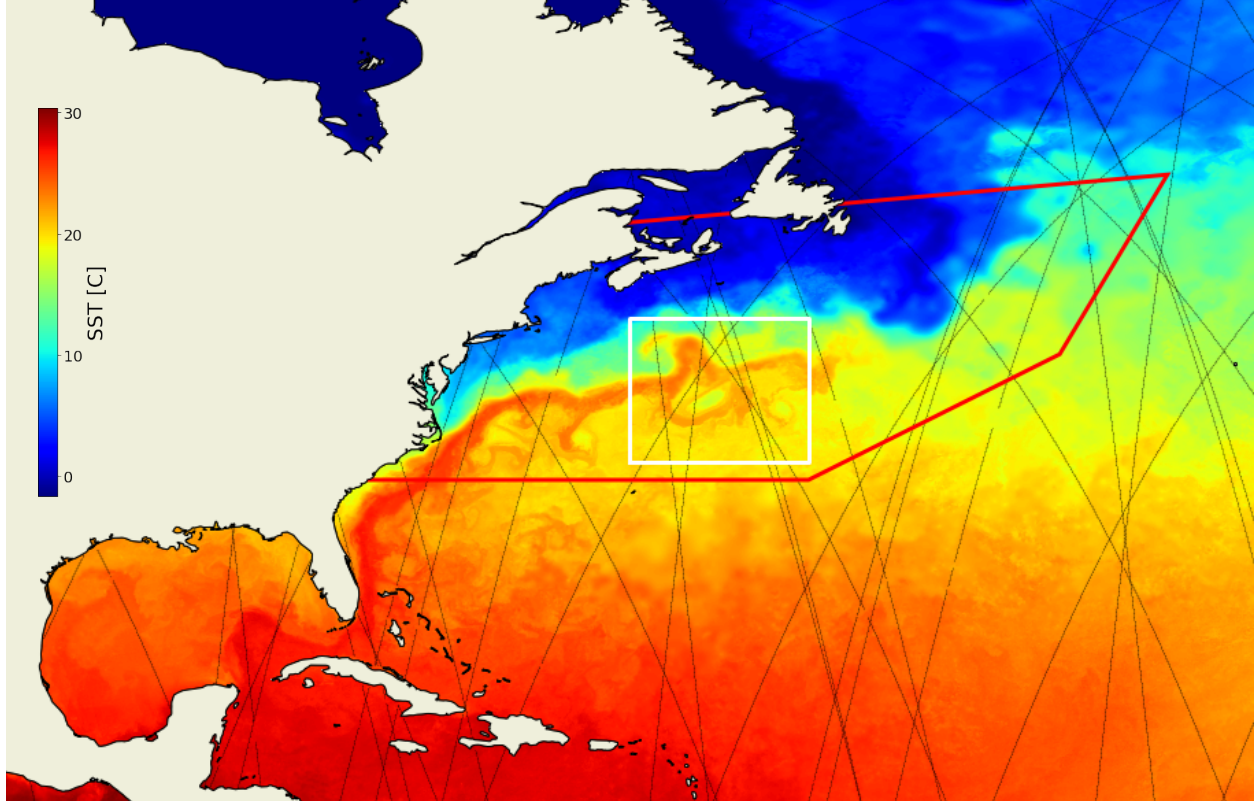


Figure 1.1: The typical data availability for one day (2017-04-11) during our testing period: the color map shows the GHRSSST MUR gridded SST, and the black tracks indicate the locations of satellite altimeter observations from all satellites available on that date (CryoSat-2, Jason-2, Jason-3, Haiyang-2A, Sentinel-3A, and SARAL/Altika). As discussed in section 1.5, we draw training examples from within the red polygon (the Gulf Stream Extension), and test our method in the white box (where other mapping methods have been applied so we can compare to their accuracy).

the rationale for our DL approach. Our method is both trained on and tested against real-world satellite observations (the datasets we use are presented in section 1.3). We present our mapping method in section 1.4 and how it performed during training in section 1.5. In section 1.6, we present the results of a regional experiment to quantitatively compare our method’s accuracy to that of the conventional and several recently proposed SSH mapping methods in the Gulf Stream Extension. In section 1.7, we discuss the implications of these results and outline a road-map to creating a more accurate global SSH product with DL.

1.2 Rationale for our proposed methodology

1.2.1 Dynamical motivation for supplementing SSH observations with SST

In addition to satellite altimetry observations, satellites routinely observe other surface ocean quantities, such as SST, sea surface salinity, and ocean color. The spatiotemporal evolution of these fields is certainly not independent from one another and from that of SSH, since they are all affected and mediated by the dynamics of the ocean (and by biological processes in the case of ocean color). Thus, if one of these other surface variables is observed in the gaps between satellite altimetry observations, and the relationship to SSH is known, then this can inform the mapping of the unobserved SSH.

In this study, we supplement satellite altimetry with SST observations to improve the SSH mapping for reasons both physical and practical. Physically, this is a natural choice since SSH and SST are known to have a close dynamical relationship. Indeed, in high eddy kinetic energy regions (e.g. western boundary currents in the northern hemisphere and the Antarctic Circumpolar Current in the southern hemisphere), temperature explains much of the density field, especially in winter [Isern-Fontanet et al., 2006], indicating that mesoscale SST anomalies can be interpreted as a dynamical part of surface-intensified baroclinic eddies [Smith and Vallis, 2001]. Hausmann and Czaja [2012] confirmed these results: SSH and SST anomalies have similar spatial scales and are characterized by a westward shift as expected from baroclinic instability, which is the main source of mesoscale eddies. In addition, a close spectral relationship between mesoscale SSH and SST can be inferred when the large-scale meridional gradient of potential vorticity at depth is linearly related to the large-scale meridional SST gradient through a function only depth-dependent [Lapeyre and Klein, 2006a], which is verified in high eddy kinetic energy regions [Lapeyre, 2009]. This relationship makes use of the surface quasi-geostrophic (SQG) approximation in which the ocean interior dynamics at mesoscale is implicitly taken into account through an “effective” Brunt-Vaisala frequency, N_{eff} [Lapeyre and Klein, 2006a; Isern-Fontanet et al., 2006; LaCasce and Mahadevan, 2006; Klein et al., 2019]. The resulting relationship in spectral space is

$$\widehat{SSH}_{SQG}(\mathbf{k}) = -\frac{f_0}{N_{eff}|\mathbf{k}|}\alpha.\widehat{SST}(\mathbf{k}), \quad (1.1)$$

where a ‘hat’ symbol indicates a horizontal Fourier transform, $\mathbf{k}$ the horizontal wavenumber vector, α the

thermal expansion coefficient, and f_0 the Coriolis frequency. Isern-Fontanet et al. [2006] proposed to use this relationship to recover SSH and therefore surface ocean currents from SST observations. Practically, satellite observations of SST have higher spatial resolution and lower missing data rates than those of SSH, which emphasizes the pertinence of using SST observations because of their close dynamical relationship with SSH.

Applying equation 1.1 to directly reconstruct SSH from SST observations is challenging in that the oceanic conditions described above do not always apply directly in all seasons and ocean regions. Other attempts have used a prescribed analytical relation between SST and SSH like that in equation 1.1 to reconstruct the unobserved SSH field [Isern-Fontanet et al., 2014; González-Haro and Isern-Fontanet, 2014]. All such methods face the challenge that the dynamical relationship between SSH and SST is non-trivial and is likely to change in space and time. Lapeyre [2009] showed that the SSH field is in some regions dominated by the SQG mode, making equation 1.1 applicable, whereas in others the first baroclinic mode dominates, meaning the SSH estimated from equation 1.1 would be inaccurate. Thus, methods that rely on the SQG framework [Isern-Fontanet et al., 2006, 2014; González-Haro and Isern-Fontanet, 2014] and methods that rely on a QG framework [Ubelmann et al., 2015, 2021; Le Guillou et al., 2021] will each be geographically and seasonally limited in their ability to reconstruct the SSH field by the pertinence of their respective dynamical assumptions. Another approach to using SST observations to inform the SSH mapping would be to use OI (as in DUACS), replacing the univariate SSH covariance function with a multivariate SSH-SST covariance function. However, the dynamical relationship between SST and SSH is sufficiently non-trivial that such an approach would likely only yield marginal mapping performance improvements.

1.2.2 Rationale for deep learning

DL models (specifically, neural networks) have been shown to be universal function approximators, given sufficient depth and training data [Hornik et al., 1989]. By employing a DL approach for SSH mapping, we implicitly make two hypotheses. *(i)*: Available satellite observations contain sufficient dynamical information that there exists a function mapping from the observations to an estimate of the dynamical state of the ocean, and by extension to the 2D SSH field. *(ii)*: There exists sufficient training data that we can train a DL model to approximate this function accurately enough to provide better SSH maps than OI. Both are

motivated by earlier work applying DL neural networks to spatiotemporal SSH interpolation in a two-layer quasi-geostrophic (QG) model given partial observations of the surface [Manucharyan et al., 2021]. George et al. [2021] also showed a DL model could learn to infer subsurface dynamics, which will, in turn, influence SSH evolution, from snapshots of the surface SSH field. The overarching conclusion from the two aforementioned studies is that the surface SSH field contains implicit signatures of the subsurface ocean dynamics and that DL models are sufficiently expressive that they can learn to recognize and exploit these signatures to more accurately reconstruct unobserved ocean variables.

The advantage of a DL approach that combines the SSH and SST fields is that a sufficiently expressive DL model, given enough training data, can, in theory, learn a dynamical relationship of arbitrary complexity. Thus, a DL model could learn to exploit the SST observations in regions and times when the dynamics makes them pertinent, while not being restricted to assuming the SQG approximation is always appropriate. The promise of a DL approach that combines SSH and SST observations to reconstruct the unobserved SSH field more accurately was also noted recently in Fablet and Chapron [2022] and Fablet et al. [2022]. However, to date, there are no operationally-used SSH reconstructions that utilize SST observations or DL, and most of the studies on the topic have been theoretical in nature.

1.3 Datasets

1.3.1 Sea surface height (SSH) observations

We use Level 3 1Hz along-track satellite altimetry SSH observations from 2010-2020 (distributed by CMEMS, [CMEMS, 2023b]). This product consists of observations from multiple satellite altimetry missions merged, calibrated, and corrected for several geophysical effects, including barotropic tides and atmospheric effects. The complete steps for producing this product are described in Taburet et al. [2019] and references therein. Note that at the Level 3 stage of the observation processing chain, the observations have not yet been mapped to a 2D grid, as shown in Figure 1.1. These along-track observations have been shown to accurately resolve the SSH signal down to length-scales $\mathcal{O}(20\text{-}100\text{km})$ with root-mean-square instrument noise of 1-4cm depending on the location, season, and satellite altimeter in question [Dufau et al., 2016].

1.3.2 Sea surface temperature (SST) observations

We use the Level 4 Multiscale Ultrahigh-Resolution (MUR) gridded SST product from the same years provided by the Group for High-Resolution SST (GHRSSST) [NASA PODAAC, 2023]. The MUR SST product is generated by combining observations from multiple satellites and in-situ sensors before interpolating to a regular grid using OI. Clouds cause gaps in the highest resolution SST satellite sensors, meaning that the OI product has high resolution in cloud-free regions but suffers artificial smoothing in cloud-occluded regions. An alternative approach would be to use the high-resolution observations directly with gaps, or to use only the lower resolution cloud-free observations provided for example by the NASA Advanced Microwave Scanning Radiometer-EOS (AMSR-E) which would both offer a uniform spatial resolution. Until a uniformly high-resolution gridded SST product is developed, there is an inevitable trade-off between using high-resolution SST with gaps, low-resolution SST without gaps, and an OI product like MUR with non-uniform spatial resolution. Despite its deficiencies, we here use the MUR SST product since it is the highest-resolution gap-free SST product currently available.

1.3.3 Data pre-processing: domain and coordinate system

We pursue a patch-based approach, where SSH is reconstructed in a local square domain using observations from only its surroundings. This approach is motivated by the fact that the dynamics governing SSH evolution is predominantly local in space and time, thus estimating the SSH at a point should only require observations from some finite space-time window surrounding that point. Our method requires the input data to be defined on a regular grid, so we bin-average the available SSH and SST observations within the domain onto regular square grids as shown in Figure 1.2. For this study, we choose the dimensions of the square domain to have a side-length of 960km with a grid resolution of 7.5km (128x128 grid points). The input SSH and SST observations are first re-scaled by subtracting the mean and dividing by the standard deviation, before being bin-averaged onto a 7.5km grid with zero padding where there are no observations. Our mapping method estimates the 2D SSH field on this same grid. This choice of grid dimensions limits the computational resources required for the reconstruction while ensuring that the domain remains large enough that a significant number of interacting mesoscale eddies fit within the model’s ‘field of view’. Defining a local orthonormal projection allows us to find the latitude and longitude for each grid point given

the coordinates of the square domain’s center. This projection ensures that the shape and size of eddies are not distorted at different latitudes, as would be the case if the square grids were defined in latitude-longitude space.

To avoid overfitting to the persistent SSH patterns in one such domain and with a view to creating a location-independent SSH mapping method, we generate the training data for our DL method by randomly selecting the coordinates on which to center the square domain for each training example. For most experiments in this study the length of the observation time series used to map SSH was chosen to be 30 days, and the effect of varying this length is explored in section 1.5.1. The input to our mapping method for a single training example therefore consists of a time series (centered on a randomly chosen date) of bin-averaged SSH and SST observations within a square domain centered on a randomly chosen point in the ocean. The SSH and SST observations are each inputted to our neural network in the form of a 3D array (2D space plus time), with the SSH observation array containing zeroes everywhere except the places where along-track SSH was observed.

1.4 Deep learning method for SSH mapping

1.4.1 Deep learning architecture

Here, we describe the DL neural network architecture we use to map SSH, a schematic diagram of our method is shown in Figure 1.2. Since the dynamical relationship between SSH and SST is non-trivial, we first use a ‘ResNet’ CNN [He et al., 2016] to encode each variable separately in a learnable latent space representation. Conceptually, each ResNet encoder learns a mapping from the SSH/SST observations to some combined latent space in which the information from the SSH and SST observations can be combined on an equal footing. Note that trying out a more straightforward approach of combining SSH and SST as two different ‘channels’ in the input to a single ResNet encoder yielded lower accuracy SSH mapping.

The ResNet architecture used for each encoder consists of alternating downsampling and residual learning blocks. A downsampling block consists of: a convolution with a stride of 2, followed by a rectified linear unit (ReLU) activation function, and finally, batch normalization. A residual learning block [He et al., 2016] consists of: a convolution with a stride of 1, ReLU activation, batch normalization, and another convolution

of stride 1. The original input to the block is then combined with the output of the convolutions through addition (the so-called ‘skip connection’ [He et al., 2016]), before a final ReLU activation function. Note that the same ResNet encoder is applied to each time step in the time series of input observations. Thus, the resulting latent space representation is one in which observations from each time step do not yet inform the representation at other time steps.

The resulting time series of latent space representations are then each passed through a bi-directional convolutional long short-term memory (ConvLSTM) layer [Shi et al., 2015]. ConvLSTM is a type of recurrent neural network widely used for problems involving regularly-spaced spatiotemporal data which has been demonstrated to capture complex dynamical relationships, for example to perform precipitation now-casting [Shi et al., 2015]. This layer must learn how partial observations from different times inform the state in the locations where observations are missing, i.e. the dynamics governing SSH/SST evolution. Separate ConvLSTM layers are applied to the SSH and SST representations, resulting in latent space representation time series for SSH and SST where each time step is informed by observations at all other times.

The SSH and SST latent space representations are combined through concatenation. This time series is then passed through another bi-directional ConvLSTM layer which can learn relationships between the SSH and SST representations to give a combined latent space representation of the dynamical state of the ocean informed by both SSH and SST.

From this latent space representation, we finally use a ResNet to decode the representation to a gridded SSH map. The decoder is like the ResNet encoder networks, but with the downsampling blocks replaced by nearest-neighbor upsampling layers, alternating with residual learning blocks. The final layer has the dimensions of the desired final target SSH grid using a linear activation. The same ResNet decoder is applied to each step in the time series. The full neural network (ResNet encoders, then separate ConvLSTMs, then joint ConvLSTM, then ResNet decoder) provides an end-to-end trainable mapping from a time series of SSH and SST observations to a time series of 2D SSH maps.

Although our method produces a time series of SSH maps for the duration of the input observations, the reconstruction error increases away from the center of the time series. This is expected, since dynamically informative observations of past (future) states are missing for days at the beginning (end) of the time series. Thus, to generate our final SSH product, we retain only the reconstruction for the central time step, creating

a time series of gridded SSH by successively shifting the input observation time series by one day.

1.4.2 Formulating a loss function in the absence of gridded SSH data

The network is trained by minimizing the mean squared error between the reconstructed SSH and the ground truth. However, the full 2D ground-truth SSH is not available when training exclusively on real-world observations. To train using real-world observations, we implement the following strategy. We first withhold some of the SSH tracks from the input and then linearly interpolate the network’s 2D reconstruction at the locations of these withheld observations to calculate the mean squared error between the reconstruction and the withheld observations. This way the error for any given training example is calculated only along a few withheld satellite tracks. Upon training, the network is forced to produce realistic 2D reconstructions throughout the domain since the withheld tracks appear at random places in the domain, so the network is unaware of where it will be evaluated.

Since there are typically several satellite altimeters operational at any time, for each training example we randomly select one of the available satellites, withhold its observations from the input, and use them as the ground truth in the loss function calculation. For each example, up to five satellite altimeters (depending on mission availability) are randomly selected to be used as the input SSH observations. The remaining satellites (or one satellite in the times when fewer than six missions are operational) are withheld for use as the ground truth when calculating the loss function.

To reduce over-smoothing and the appearance of high-frequency artifacts in the reconstruction, we include additional regularization terms in the loss function proportional to the mean squared error in the first and second along-track derivatives of SSH. Thus the cost, $\mathcal{L}$, we seek to minimize during training is given by:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{\sigma_0^2} (\tilde{\eta}_i - \eta_i)^2 + \frac{\lambda_1}{\sigma_1^2} (\partial_x \tilde{\eta}_i - \partial_x \eta_i)^2 + \frac{\lambda_2}{\sigma_2^2} (\partial_x^2 \tilde{\eta}_i - \partial_x^2 \eta_i)^2 \right), \quad (1.2)$$

where N is the number of observations, η_i is the true observed SSH for the i -th observation, $\tilde{\eta}_i$ is the corresponding mapped SSH, x is a spatial coordinate following the satellite track, σ_0^2 , σ_1^2 , and σ_2^2 are the variances of η , $\partial_x \eta$, and $\partial_x^2 \eta$ respectively, and λ_1 and λ_2 are tunable parameters controlling the relative weighting for each regularization term in the loss function. In this study, we set $\lambda_1 = \lambda_2 = 0.05$. Along-

track derivatives are estimated using first-order centered difference.

1.4.3 Choice of training, validation, and testing datasets

The data are split into training, validation, and testing datasets to prevent overfitting. The validation dataset is used to determine at what point the model is trained to its full potential. In contrast, the testing dataset is withheld and used only at the end to test the model’s performance on unseen data. Ideally, these datasets should be independent. To this end, we divide the available history of observations into time windows for training, validation, and testing as shown in Figure 1.3. The year 2017 was kept aside for testing to coincide with the testing period covered by an Ocean Data Challenge which allows us to compare our method to other proposed mapping methods (this data challenge is described in section 1.6.1). The remaining times were then split into interleaved training and validation windows. We choose interleaved windows to ensure that the training and validation datasets both capture various seasons and sample interannual variability. These windows were chosen such that 80% of the dates fall within the training dataset and 20% in the validation dataset (specifically, we broke the available dates into 60 day chunks then chose the first 4 chunks to be training, the next 2 to be validation, then the next 4 to be training, and repeated this procedure until all dates had been assigned). Because there is significant autocorrelation in the SSH field over a time scale of $\mathcal{O}(10\text{-}20\text{ days})$, it is important to choose training and validation windows that are widely enough separated in time to prevent significant autocorrelation between the training and validation data. We ensure this by leaving a gap of 30 days between training and validation windows, since we observed the autocorrelation in the SSH field of an ocean global circulation model (GCM) (specifically, the CESM 1/10° Parallel Ocean Program model with JRA55 reanalysis forcing) to drop by $\sim 50\%$ in the Gulf Stream Extension over this time.

1.5 Training performance

To provide a proof of concept for our method in this study we focus on mapping SSH in a region with energetic mesoscale dynamics, namely the Gulf Stream Extension. To this aim, we draw training and validation examples from the area shown in red in Figure 1.1. This area was chosen such that all examples feature the dynamics characteristic of the Gulf Stream Extension. In principle, we could draw training and

validation examples from anywhere in the ocean, but since there are a wide range of dynamical conditions in the ocean it is challenging for a single neural network to learn to map SSH in all regions at once (this will be further discussed in section 1.7.3). Training and validation examples are drawn from this region with the square domain centered at a randomly selected point and time (respecting the training-validation date partition outlined in section 1.4.3). This region features an abundance of mesoscale eddies and complex jet-eddy interactions between the Gulf Stream and neighboring eddies. It is thus a challenging region for SSH mapping, where the widely-used DUACS product displays relatively high mapping errors, with the error becoming comparable to the standard deviation of the along-track SSH signals.

1.5.1 Determining the optimal length of the input time series and number of training examples

The SSH reconstruction on a given day is generated using a time series of observations centered on that day. Thus, a key parameter of our approach is the length of this time series. Including observations from further into the past and future increases the aggregated spatial coverage of the SSH observations and provides information about the time variation of the surface flow. To accurately reconstruct the SSH in unobserved regions, the network needs to learn an approximation of the surface dynamics, so that observations from other days can be used to inform the reconstruction.

Physically, we expect there to be some predictability time horizon beyond which observations become decreasingly helpful in informing the reconstruction. Therefore, we expect to see diminishing improvement to the reconstruction as the length of the time series is increased beyond some time period. We tested this hypothesis by varying the length of the input time series and training our ConvLSTM model, with both SSH and SST observations for input, and finding the mean value for the cost, $\mathcal{L}$, achieved on the validation dataset (Figure 1.4a). $\mathcal{L}$ reduces with increasing time series length until about twenty days (i.e. ten days on either side of the reconstruction), beyond which further lengthening yields minimal improvement. As demonstrated in Figure 1.4b, we checked that $\mathcal{L}$ had stopped improving significantly as a function of the number of training examples for each time series length. This time-scale could be a feature of our neural network architecture rather than a physical predictability horizon.

The results in the remainder of this study use a time series length of thirty days (i.e. fifteen days on

either side of the date for reconstruction) noting that the significant computational cost of extending the time series beyond this length is unlikely to yield significant reductions in $\mathcal{L}$.

1.5.2 Quality of mapped SSH field

Figure 1.5 provides an example time series of SSH mapped using our method from the independent testing dataset (i.e. during 2017). The SSH field features separated, distinct mesoscale eddies and a strong Gulf Stream, evident in the surface geostrophic currents calculated from the SSH map. The time-evolution of the reconstructed field also looks qualitatively realistic, with a Gulf Stream meander pinching off to form a new mesoscale eddy to the north of the jet. The capability of our method to produce realistic-looking SSH fields from observations, unseen during training, gives us confidence that the model has learned a robust mapping from SSH and SST observations to the 2D SSH field, rather than overfitting to the training data.

1.6 Inter-comparison of existing SSH mapping methods in the Gulf Stream

Extension region

1.6.1 Mapping methods and evaluation metrics

We use the data distributed through the AVISO Ocean Data Challenge [AVISO, 2021] to provide a direct quantitative comparison of our method’s accuracy to that of five established and experimental SSH mapping methods. The data consist of SSH maps created for the year 2017 (the year set aside for testing) in the region covering (55° - 65° W, 33° - 43° N) using five different mapping methods, described below. Each map was created using observations from all available altimeters apart from CryoSat-2, which is withheld for use as a ground-truth for calculating the maps’ errors.

We compare our approach to five other mapping methods:

- *DUACS*: the community-standard SSH product created using OI [Taburet et al., 2019]. The full mapping method is not publicly available, but a map was generated excluding CryoSat-2 for the data challenge.
- *DYMOST*: the ‘Dynamic Interpolation’ method proposed by [Ubelmann et al., 2015] and evaluated by

Ubelmann et al. [2016] and Ballarotta et al. [2020]. This method is a variant of the OI approach. The Gaussian a priori SSH covariance model used in DUACS is replaced by a dynamically informed model based on the forward and backward in time integration of an idealized potential vorticity conservation equation. In this method the SSH evolution is assumed to be governed only by the first baroclinic mode.

- *MIOS*T: the ‘Multiscale and Multivariate Interpolation’ method proposed by Ubelmann et al. [2021]. This method is another variant of the OI approach where the single-scale Gaussian covariance model for SSH used in DUACS is replaced with a wavelet basis where the amplitude of each scale component is set to match the power spectrum observed in along-track altimetry, thus assuming isotropy.
- *BFN-QG*: the ‘Back-and-Forth Nudging QG’ method proposed by Le Guillou et al. [2021]. This method uses the back-and-forth nudging method for data assimilation to interpolate the altimetry observations while respecting the dynamics of a one-layer QG model of ocean turbulence.
- *4DVarNet*: the DL method proposed by Fablet et al. [2021a] and Fablet et al. [2021b]. This approach poses the SSH interpolation problem as a four-dimensional variational data assimilation (4DVar) problem. It replaces the dynamical model and solver with trainable neural networks. We note that this method has been used in several OSSE studies, so many different versions exist in the literature. Here, we use the map provided through the data challenge which takes only SSH observations as input. Fablet and Chapron [2022] and Fablet et al. [2022] recently showed that including SST in the 4DVarNet approach could significantly improve the SSH reconstruction. However, this has only been demonstrated in the OSSE setting where the full underlying ground-truth SSH is known during training. It has not yet been demonstrated to improve SSH mapping using real-world SST observations which suffer gaps due to cloud cover if the highest resolution observations are used. Hence, we cannot directly compare the two methods with the inclusion of SST but note that the improvements seen for both methods in their respective settings are consistent with the expectation from SQG dynamics that the SST field contains relevant information about the dynamics of the SSH field.

We do not here compare our method to a full-scale ocean GCM data assimilation reanalysis (e.g. Lellouche et al. [2021]), since none have been produced with CryoSat-2 observations withheld from the as-

simulation process, preventing a like-for-like comparison to the other methods in the above data challenge framework. The comparison of our method’s SSH mapping accuracy to a GCM data assimilation scheme will be the subject of a future study.

For our analysis, we reduce the size of the data challenge’s study region to (55° - 65°W, 34° - 42°N) to ensure the full region fits inside the ‘field of view’ of our neural network’s 960x960km output. All methods are assessed in this same region over the entire year of 2017. Figure 1.5 shows an example of the result of applying our method to map SSH in this study region.

We use the withheld CryoSat-2 observations to calculate various metrics to quantitatively compare the performance of the SSH mapping methods. These metrics are summarized below:

- *Mean RMSE (cm)*: The SSH maps are interpolated to the locations of the withheld observations and the root mean square error (RMSE) of the map is calculated for each day where observations are available, then the mean of these daily RMSEs is taken.
- *Standard deviation of RMSE (cm)*: The standard deviation of the daily RMSE values.
- *Effective spatial resolution (km)*: Calculated using the along-track SSH spectra for each satellite pass. This is the wavelength at which the power spectral density of the misfit between the map and the observations becomes comparable to that of the observations, quantifying the spatial scales accurately resolved by the map. See Ballarotta et al. [2019] for a complete discussion of this metric.

- *RMSE score (no units)*:

$$\text{RMSE score} = 1 - \frac{RMSE_{map}}{RMS_{obs}}$$

A perfect map would have a score of 1, a map that predicted zeroes everywhere would have a score of 0.

- *Standard deviation of RMSE score (no units)*: Standard deviation of the daily RMSE scores.

1.6.2 SSH validation against independent satellite observations

Root mean square error (RMSE)

All the experimental mapping methods were found to have lower mean RMSE values than DUACS and the spread in their daily RMSE is also lower, as can be seen in Figure 1.6 and Table 1.1. BFN-QG provides only a marginal improvement in RMSE, whereas DYMOST and MIOST significantly improve (a 12% reduction in mean RMSE). All the DL methods tested here (4DVarNet SSH and all configurations of our ConvLSTM method) yielded lower RMSE than all the other methods. 4DVarNet SSH showed a 14% reduction in RMSE while our ConvLSTM method respectively gave a 13% and 17% reduction in RMSE when run with just SSH (ConvLSTM SSH) and with combined SSH and SST observations (ConvLSTM SSH+SST).

To explore the eddy-jet configurations that present the most significant challenge to traditional altimetry mapping and further illustrate our method’s improvement, in Figure 1.7 we present three case studies from the testing period where we contrast the DUACS reconstruction with our ConvLSTM SSH+SST method. These cases correspond to the three CryoSat-2 satellite tracks of at least 800km in length for which the DUACS reconstruction shows its highest RMSE. In all three cases, the primary source of error is the misrepresentation of a sharp kink in the Gulf Stream in the top left corner of the domain (recall that geostrophic surface currents flow along contours in SSH). In all three cases ConvLSTM SSH+SST leads to a significantly more accurate along-track SSH profile. Physically, the eddy features in SSH are expected to evolve relatively quickly in shape and amplitude in regions where the jet (i.e. the Gulf Stream) is bent at a sharp angle, thus presenting a challenge to the simple SSH covariance model employed by DUACS. Our ConvLSTM SSH+SST method overcomes this shortcoming.

Effective along-track spatial resolution

All the experimental mapping methods were able to reconstruct smaller-scale SSH signals than DUACS, as can be seen in Figure 1.8. The spectra in Figure 1.8a show that the DUACS reconstruction underestimates the strength of the SSH signal at smaller scales compared to the observations. All other methods show spectra closer to the observations at these scales to varying degrees. The spectral coherence plot in Figure 1.8b demonstrates that the DUACS reconstruction is less accurate than all other methods at scales of 70-200km. None of the methods accurately resolve SSH below these scales due to the limited along-track

resolution of the observations and the size of the gaps between observations.

The effective spatial resolution (the scale at which each curve in Figure 1.8b crosses 0.5, also listed in Table 1.1) for each method reveals significant differences. MIOST and DYMOST only offer marginal improvements on this metric (a 7% and 12% reduction in the smallest wavelength accurately resolved), whereas BFN-QG provides a more substantial improvement (20%). Again, all DL methods yield the largest improvements on this metric, with ConvLSTM SSH, 4DVarNet SSH, and ConvLSTM SSH+SST providing 23%, 28%, and 30% improvements respectively. The SSH map produced using our ConvLSTM SSH+SST method accurately resolved SSH signals with wavelengths as small as 104km.

The significant improvement in all metrics by all the DL methods provides a compelling case for the use of DL in developing an improved 2D SSH product. The addition of SST observations also provides a clear improvement in our method’s reconstruction accuracy and resolution.

Frequency spectra of SSH fields

Another important dynamical property of a mapped SSH field is its frequency spectrum. Unlike for the wavenumber spectrum, discussed in section 1.6.2, estimating a ground truth from the independent altimeter observations is challenging since at any one location the temporal sampling is relatively sparse (CryoSat-2 observes the same $1/8^\circ$ by $1/8^\circ$ spatial bin once every 5-6 days on average in this region). Thus, we cannot calculate the frequency spectral coherence in the manner of section 1.6.2 using just the withheld altimeter observations.

Nevertheless, we can still characterize the frequency spectrum of each SSH map and draw comparisons among the maps, as shown in Figure 1.9a. Compared to all other methods, DUACS has a substantially lower power spectral density at time scales shorter than 30 days, implying that DUACS may be underestimating the short-time SSH variability, as has been noted in the literature [Ballarotta et al., 2019]. The other maps all show similar frequency spectra down to time scales of around 5 days, below which 4DVarNet and BFN QG show significantly higher power spectral density. Without knowing the ground truth frequency spectrum from observations, it is hard to discern which map’s spectrum most accurately represents the variability at these time scales. The higher variability at short temporal scales for 4DVarNet and BFN QG could be due to the appearance of non-physical artifacts along the location of the input satellite tracks (discussed more in

section 1.6.3).

To provide some estimate of the SSH variability from the observations, we also calculated the second-order structure function in time, as shown in Figure 1.9b. To do this, we used all available satellite altimetry observations from the study region and time, selected the 100 most frequently observed $1/8^\circ$ by $1/8^\circ$ bins, used the observations to construct a SSH time series for each bin, and used these to estimate the second order structure function. We thus assume that the frequency spectrum of the SSH field is approximately isotropic within this region. We used the bias-corrected and accelerated bootstrap method to estimate 95% confidence intervals.

Calculating the second-order structure function for each of the SSH maps at uniformly-spaced points throughout the region allows us to compare the maps' variability to the observational estimate. We see that DUACS is indeed underestimating the variability of the SSH field at time scales below 20-30 days as the higher variability shown by the other maps is replicated in the observations. The differences between the other maps at short (< 10 days) time scales are small and the confidence interval on the observational estimate is large since satellite altimetry provides poor sampling of the SSH field at these short time scales.

Seasonality of improvement to ConvLSTM mapping with the inclusion of SST

The utility of SST for recovering sea surface currents in theory should depend on environmental conditions which in turn vary geographically and seasonally. Enhanced surface temperature stratification in the summer months may obscure the signature of ocean eddies, while the phase shift between mesoscale SST anomalies and surface currents is known to be smaller when the mixed layer is deep, typically in the winter months [Isern-Fontanet et al., 2008]. Thus, it is anticipated that SST will be most useful in the winter months for estimating surface currents (and SSH by extension). González-Haro and Isern-Fontanet [2014] explored the geographical and seasonal variations in the accuracy of sea surface currents recovered from SST through SQG theory and for the Gulf Stream Extension region considered in this study found that the SST-derived surface currents were most accurate in the winter.

Since our DL approach does not explicitly rely on the SQG framework, it is interesting to consider the seasonality of the improvement to the ConvLSTM mapping accuracy and resolution when SST is included. We here study this seasonality for the years 2017, 2021, and 2022. These years were chosen since there were

a large number of altimeters available continuously in 2021 and 2022 (6 and 5 respectively, we excluded Jason-3 in 2022 since there was a gap in coverage) and none of the three years were used during training or cross-validation. To ensure a large enough number of independent altimeter observations for robust statistics, we generated SSH maps for each possible combination of three input satellites and evaluated each map against the remaining available altimeters; this gives 4 maps for 2017, 20 for 2021, and 10 for 2022. Rolling 2-month averages of the ConvLSTM mapping RMSE (normalized by the along-track SSH RMS variations since this varies seasonally) and effective spatial resolution were calculated for all three years with and without SST.

While there is variation between the years, the improvement in both accuracy and resolution for ConvLSTM SSH+SST compared to ConvLSTM SSH was found to be slightly higher in the summer months than during the winter (Figure 1.10). The ConvLSTM SSH+SST normalized RMSE varied from around 8% lower than ConvLSTM SSH in the winter months to around 15% lower in the summer. While the seasonal trend was less clear in the effective resolution, the maximum percentage reduction in resolved wavelength for ConvLSTM SSH+SST was found outside the winter months for all three years.

These findings run contrary to the expectations from ocean dynamics described above. One possible explanation is the impact of the seasonally varying cloud cover on the Level 4 GHRSSST MUR SST product used in this study. Since this product combines both microwave and infrared SST observations, its spatial resolution varies dramatically in space and time. The high-resolution infrared SST sensors can only be used on cloud-free snapshots whereas the lower-resolution microwave SST sensors are insensitive to cloud cover, so the MUR SST product has a high spatial resolution on cloud-free days but becomes coarser on cloudy days when it relies on the microwave observations. Cloud cover was higher in our study region during the winter months than during the summer, so the improved performance of ConvLSTM SSH+SST during summer is likely due to the higher resolution of the SST observations rather than some ocean dynamics signature. The sensitivity of our method to the choice of SST product will be explored in a future study. We note that the Level 4 GHRSSST MUR SST product provides gridded uncertainty estimates for the interpolated SST field. While in this study we did not make use of these estimates, a possible future line of inquiry would be to include them in the input to allow the neural network to put more weight on the more reliable, cloud-free SST regions.

1.6.3 Surface geostrophic currents

The primary use of SSH maps is to infer surface ocean currents through the assumption of geostrophy. The eastward and northward geostrophic surface currents (u_g and v_g respectively) are related to the SSH, η , through

$$(u_g, v_g) = \frac{g}{f} \left(-\frac{\partial \eta}{\partial y}, \frac{\partial \eta}{\partial x} \right), \quad (1.3)$$

where x and y are eastward and northward spatial coordinates respectively, f is the Coriolis frequency, and g is the acceleration due to gravity. The currents are thus proportional to the first-order spatial derivatives of η .

From these currents, other physical quantities such as relative vorticity, ω , and strain rate, γ , can be calculated by taking spatial derivatives of the velocity field (corresponding to second-order derivatives of the SSH field):

$$\omega = \frac{\partial v_g}{\partial x} - \frac{\partial u_g}{\partial y} \quad (1.4)$$

$$\gamma = \sqrt{\left(\frac{\partial u_g}{\partial x} - \frac{\partial v_g}{\partial y} \right)^2 + \left(\frac{\partial v_g}{\partial x} + \frac{\partial u_g}{\partial y} \right)^2}. \quad (1.5)$$

These higher order derivatives are an important diagnostic since they give information about the dynamics (i.e. the acceleration) of the surface motions [Hua and Klein, 1998]. Relative vorticity (and potential vorticity) are crucial quantities when studying ocean turbulence because in 2D turbulence vorticity is conserved along streamlines and potential vorticity is conserved in QG and SQG motion. As such the interaction of eddies with each other and with the mean flow, as well as their equilibration, filament formation, and frontogenesis are all processes that are viewed through the lens of vorticity and strain evolution.

One such quantity that is important to consider is the Okubo-Weiss quantity, W , defined as the difference between the square of the strain rate and the square of relative vorticity [Okubo, 1970; Weiss, 1991; McWilliams, 1984; Hua and Klein, 1998; Hua et al., 1998]

$$W = \gamma^2 - \omega^2. \quad (1.6)$$

When W is negative, relative vorticity and therefore rotation dominates. This typically occurs in the regions

inside mesoscale eddies where SSH extrema are found. However, when W is positive or close to zero, the strain rate dominates or is close to vorticity, indicating an intensification of SST fronts and associated vertical velocity. This typically occurs in regions between mesoscale eddies and at the eddy edges [Hua et al., 1998; Lapeyre et al., 1999; Klein et al., 2019].

We calculated the surface geostrophic currents, ω , and W from the DUACS and ConvLSTM SSH+SST maps. To closely match the procedure for calculating current speeds from the DUACS SSH map used to generate the surface current speeds distributed in the CMEMS Level 4 SSH product [CMEMS, 2023c], we use a 9-point stencil (as advocated in Arbic et al. [2012]) for estimating derivatives of the DUACS SSH field. Because the DUACS product is distributed on a $1/4^\circ$ ($\sim 30\text{km}$) grid whereas our SSH reconstruction is on a 7.5km grid, we take all DUACS derivatives on its native grid before linearly interpolating to our higher-resolution grid.

The ConvLSTM SSH+SST reconstruction results in narrower and stronger currents than those calculated from the DUACS reconstruction, as shown in Figure 1.11a&d. The maximum current speed for the date shown in Figure 1.11 is 40% stronger in the ConvLSTM SSH+SST reconstruction than in the DUACS reconstruction. This result is expected since the covariance model used in DUACS leads to an overly smooth SSH field, smaller-scale features being blurred out. Hence, the magnitude of the steepest gradients of the SSH field will, in turn, be weaker.

The NOAA Atlantic Oceanographic & Meteorological Laboratory (AOML) global surface drifter program provides in-situ observations of the total near-surface (15m depth) ocean currents through satellite-tracked drifters. To quantitatively test the accuracy of the surface currents inferred from the SSH maps, we here compare the mapped currents to all available AOML drifters within our study region and time. The AOML drifter data are available through the CMEMS in-situ surface and sub-surface water velocity observations product [CMEMS, 2023a]. The results shown here are after we applied the wind slippage correction distributed within the product, however, removing this correction does not change the hierarchy of the methods' errors. Both ConvLSTM SSH and ConvLSTM SSH+SST show a clear improvement in current reconstruction accuracy relative to DUACS and the other methods, with a 24% and 27% decrease in surface current RMSE respectively (surface current RMSE for each method can be found in Table 1.1). Surface drifters don't directly measure geostrophic currents, since Ekman and ageostrophic components

will be aliased onto the drifter-observed currents. However, the fact that we can more accurately estimate the surface currents through geostrophy from our improved SSH map than from the DUACS map is compelling since surface drifters are an observational dataset that is entirely independent of the satellite altimetry methodology.

The apparent difference between the DUACS and ConvLSTM SSH+SST reconstructions becomes starker for the relative vorticity and Okubo-Weiss quantity, W , (shown in Figure 1.11b,c,e,f) since deficiencies in the reconstructed SSH field are amplified as derivatives of higher order are taken. The path of the Gulf Stream is hard to identify from the DUACS relative vorticity field visually and is practically impossible to pick out in the Okubo-Weiss field. By contrast, the ConvLSTM SSH+SST map results in a relative vorticity and Okubo-Weiss fields that show a clear, physically realistic Gulf Stream trajectory. Two evident mesoscale eddies can also be seen in the Okubo-Weiss field as distinct blue regions where vorticity dominates the flow surrounded by red regions where the eddy is inducing high strain rates. One deficiency of our reconstruction is the presence of faint high frequency grid-like artifacts. Such artifacts are common when using CNNs, and their intensity decreases with increasing number of training examples. We discuss these more in section 1.7.3 but note that they only become visually apparent when taking higher-order derivatives of the SSH field.

Although several of the methods in this study exhibit similar summary statistics, there are large qualitative differences in the resulting current reconstructions. In Figure 1.12, we show the relative vorticity field calculated from each SSH map on the same date. The DYMOST and MIOST vorticity fields are qualitatively like that of DUACS so the Gulf Stream’s large-scale structure is challenging to discern. By design, BFN-QG results in a physically realistic-looking vorticity field since the SSH map is the result of solving an idealized dynamical model of ocean turbulence, albeit an overly-idealized one. 4DVarNet produces a vorticity map with similar large-scale structures to other methods but with the addition of some smaller-scale features and clear artifacts. The prominent mesoscale eddies (the two distinct red regions in the lower half of the domain) exhibit high-frequency radial features in the 4DVarNet reconstruction that are likely to be CNN-induced artifacts. In this reconstruction, there are straight-line features along which the map exhibits smaller-scale features. These lines correspond to the ground tracks of the satellites providing the input SSH observations (see movie S2 for the vorticity evolution).

These qualitative differences have impact on two key physical quantities of interest: eddy kinetic energy

(EKE) and enstrophy. Figure 1.13a shows each map’s time- and domain-averaged EKE. DUACS has a significantly lower EKE than the other maps while DYMOST and ConvLSTM SSH+SST have the highest EKE. Larger-scale flows dominate EKE, so the fact that DYMOST and ConvLSTM SSH+SST have a substantially higher EKE than other methods implies these reconstructions result in faster large-scale currents (e.g. the Gulf Stream) than the others. Figure 1.13b shows the time-and domain-averaged enstrophy for each map. Enstrophy is a quantity that is typically dominated by smaller-scale flows. DUACS has a significantly lower enstrophy than the other maps, consistent with the expectation that OI smooths out small-scale features. 4DVarNet has a very high enstrophy compared to the other methods due to the unphysical high-frequency artifacts in the relative vorticity field.

1.6.4 New mesoscale eddy found by synthesizing SST and SSH observations

To further emphasize the power of using SST observations to help map SSH, we here present one case study from our testing dataset where ConvLSTM SSH+SST was able to reconstruct a mesoscale eddy that was missed by all the other methods, which only used SSH observations as input. Figure 1.14 shows the relative vorticity fields for five consecutive days in July 2017 reconstructed using both ConvLSTM SSH and ConvLSTM SSH+SST. On July 7th in the ConvLSTM SSH+SST reconstruction there is a cyclonic mesoscale eddy in the center of the domain (boxed in blue in the figure) which is not present in the ConvLSTM SSH reconstruction or any of the other maps (not shown in the figure). None of the satellite altimeters passed over this eddy (the altimeter tracks are overlaid on the figure), but there is a clear cold signature in the SST coinciding with the reconstructed eddy. The formation process of the eddy can be seen in the SST field for the days leading up to July 7th. On July 3rd, a filament develops on the large eddy in the northeast of the domain (boxed in black in the figure). Over the following days this filament becomes unstable and sheds to form the new eddy. This formation process all happens within just 5 days and no altimeter track passes over the region during this period, making it very challenging for an SSH mapping method that only uses altimetry observations to predict this process from the available observations. Using SST observations to aid SSH mapping is thus essential if we are to capture the formation of short-lived, small-scale mesoscale eddies.

1.7 Discussion and conclusions

1.7.1 Summary

In this study, we presented a DL framework for mapping SSH from satellite altimetry and SST observations, motivated by the close dynamical relationship between these surface ocean variables. By training and testing our method on real-world satellite observations, we demonstrated that we could reconstruct the mesoscale SSH field with a high degree of physical realism. Including SST as additional observations caused a significant improvement in the accuracy of our reconstruction, in line with expectations from SQG theory. When compared against several other SSH mapping methods proposed in the literature, our method was able to map SSH in the Gulf Stream Extension with the highest accuracy and effective spatial resolution. Our study provides a roadmap towards replacing the decades-old OI method with state-of-the-art deep learning methods to produce higher-accuracy and higher-resolution global SSH maps.

1.7.2 Physical implications of improved mesoscale SSH field

The higher-accuracy and higher-resolution mesoscale SSH field that our DL mapping method provides is an exciting new dataset for physical oceanographers. The geostrophic surface currents calculated from our SSH field resulted in a substantially (25%) higher EKE than from the DUACS SSH map. This result, if replicated in other ocean regions, suggests that current satellite altimetry-derived estimates of oceanic mesoscale eddies' contribution to the global energy budget could be a significant underestimate [Thoppil et al., 2011; Xu et al., 2014; Martínez-Moreno et al., 2019]. An updated global estimate of EKE using our DL SSH mapping method is the subject of an ongoing study.

Our SSH reconstruction also resulted in relative vorticity and Okubo-Weiss quantity, W , fields that appear qualitatively to be more physically realistic, with a clearer separation between the Gulf Stream and nearby coherent mesoscale eddies. As discussed in 1.6.3, W is an important diagnostic for distinguishing between rotation-dominated regions (inside mesoscale eddies) and regions of high strain with SST front intensification and relatively strong vertical velocities. Thus, W is a prerequisite to get access to the 3D eddy field [Qiu et al., 2020; Siegelman et al., 2020b]. Another interpretation of W is that, whereas first-order SSH derivatives only give access to currents, second-order derivatives (through W) give access to accelerations

and therefore to the time evolution of the eddy field [Hua and Klein, 1998; Hua et al., 1998; Lapeyre et al., 1999]. One consequence is that improving the diagnosis of second-order SSH derivatives would lead us to better estimate the kinetic energy cascade, i.e. the balance between the merging between eddies leading to larger eddies (inverse KE cascade) and the splitting of eddies by other eddies leading to smaller eddies (direct KE cascade) [Scott and Wang, 2005; Klein et al., 2019; Storer et al., 2022]. These comments emphasize the importance of accurately diagnosing second-order SSH derivatives. Our DL SSH mapping method could allow this diagnosis to now be made from current satellite observing capabilities.

Much of what we know about coherent mesoscale eddies in the ocean comes from applying eddy tracking algorithms to SSH maps [Chelton et al., 2011; Martínez-Moreno et al., 2019, 2021]. By including SST observations in the input to the SSH mapping, we demonstrated in section 1.6.4 that we could reconstruct short-lived, small-scale coherent eddies that would otherwise not be captured in SSH maps. Thus, applying eddy tracking algorithms to our new SSH map could lead to changes in the conclusions drawn in these past studies, especially about the number of detected coherent eddies, their shapes and strengths.

1.7.3 Potential for development of a global DL SSH product

While the results of this study are promising, several challenges must be overcome before a global DL SSH map can be made available to the oceanography community. The main challenge we foresee is that different regions of the ocean exhibit a diverse range of dynamical regimes. It may therefore be challenging for a single DL model to learn to map SSH accurately in all ocean regions. An approach involving an ensemble of regional DL models could be employed to overcome this challenge where bespoke models are trained for each ocean region before the resulting regional SSH maps are merged to form a global product.

The power of DL models comes from their ability to continue to improve dramatically as they are trained on larger volumes of data. While we limited ourselves here to using real-world observations for the training dataset, future work could employ ‘transfer learning’ where the DL model is trained on a large available dataset drawn from a system with similar characteristics to that under consideration before being fine-tuned on the desired dataset [Subel et al., 2022]. In the context of SSH mapping, the abundant output from high-resolution ocean GCMs might provide a valuable dataset for pre-training our DL model, though all GCMs

contain imperfect representations of ocean physics and so the resulting SSH product would need to be used with caution if being used for observational validation of the GCM itself.

As highlighted in section 1.6.3, CNN methods such as ours and 4DVarNet can suffer non-physical high-frequency artifacts. Without enforcing hard physical constraints on the SSH reconstruction, there is to our knowledge no way to guarantee that no such artifacts would appear in the SSH map. However, in this study we found adding terms regularizing the first and second along-track SSH derivatives significantly reduced the artifacts.

For a global DL SSH product to become widely adopted in the oceanographic community, the product must have a consistent spatial resolution that does not depend on the observational sampling. We have demonstrated here that this can be achieved by constructing a training dataset where the loss function is only calculated at points not included in the input to ensure that the DL model does not learn to skillfully reproduce a high-resolution field near the input observations but with a smoother low-resolution map in the gaps between observations. The development and validation of a global SSH product using our method is the subject of an ongoing study.

The region considered in this study is particularly favorable for the use of SST to infer SSH. In regions where salinity plays a larger role in setting the surface density field and stratification, for example at high latitudes and in the tropics, the improvement due to including SST may be less pronounced. Additionally, as discussed in section 1.6.2, consideration will need to be given to the choice of SST product used to generate a global product since the effect of clouds on the MUR SST product used in this study could cause the accuracy and effective resolution of the SSH product to vary geographically and seasonally, posing challenges for end users seeking to interpret the dynamics of the gridded SSH field.

Other ocean observation datasets could be incorporated into a similar framework to that presented here to improve the mapping of SSH or other quantities of interest. Surface salinity and ocean color are routinely observed by satellites and may contain different dynamical signatures to the SST observations. ARGO profiles provide regular vertical profiles of temperature and salinity (among other variables) albeit with sparser horizontal sampling density than satellite observations of the surface. Since the vertical structure of mesoscale eddies plays a significant role in governing their evolution, it is possible that incorporating ARGO observations into a similar DL framework to that described here would lead to more accurate SSH

mapping. Although described here in the context of satellite altimetry, our DL framework could similarly be applied to other remote sensing interpolation problems.

Finally, the recent launch of the Surface Water Ocean Topography (SWOT) mission will soon offer, for the first time, high-resolution 2D snapshots of the SSH field through its wide swath. The development of a method to use SWOT observations to map high-resolution SSH is an area of active research [Beauchamp et al., 2020; Fablet et al., 2021a; Le Guillou et al., 2021] and our framework could also be adapted to suit the nature of SWOT observations. The launch of the SWOT mission certainly does not remove the need for a more accurate method for mapping SSH from traditional nadir altimetry. The nadir observations are available over a longer period than the likely duration of the SWOT mission, and so are an important dataset for studies of inter-annual variability and climate change. SSH maps and the tracked eddies inferred from them are also widely used to aid the interpretation of in-situ observations across all oceanographic disciplines, so improving the SSH maps for the pre-SWOT years would add value to many years of already-collected in-situ observations. To this end, SWOT observations in future could be used as a ground truth to further validate our nadir altimeter-derived DL SSH map or calculate the loss function during training.

The effective use of a DL framework such as that outlined in this study to map the global ocean’s SSH field from existing and future satellite observations would represent an exciting new paradigm for studying surface ocean dynamics and, by extension, global climate.

1.8 Full performance metric comparison

Mapping methods’ summary statistics						
Mapping method	RMSE [cm]	std. of RMSE [cm]	RMSE score [no units]	std. of RMSE score [no units]	Eff. resolution [km]	Drifter RMSE [m/s]
DUACS	7.7	2.6	0.88	0.055	149	0.349
DYMOST	6.8	2.0	0.89	0.047	131	0.342
MIOST	6.8	2.3	0.89	0.057	139	0.297

BFN-QG	7.5	2.6	0.88	0.053	119	0.290
4DVarNet SSH	6.6	1.8	0.90	0.046	107	0.271
ConvLSTM SSH	6.7	1.8	0.89	0.050	115	0.265
ConvLSTM SSH+SST	6.4	1.6	0.90	0.048	104	0.256

Table 1.1: Summary of the performance metrics calculated for the SSH mapping methods compared in this study. Methods above the double horizontal line are from the literature (see the text for references) and those below are those proposed in this study. See section 1.6.1 for a description of each metric.

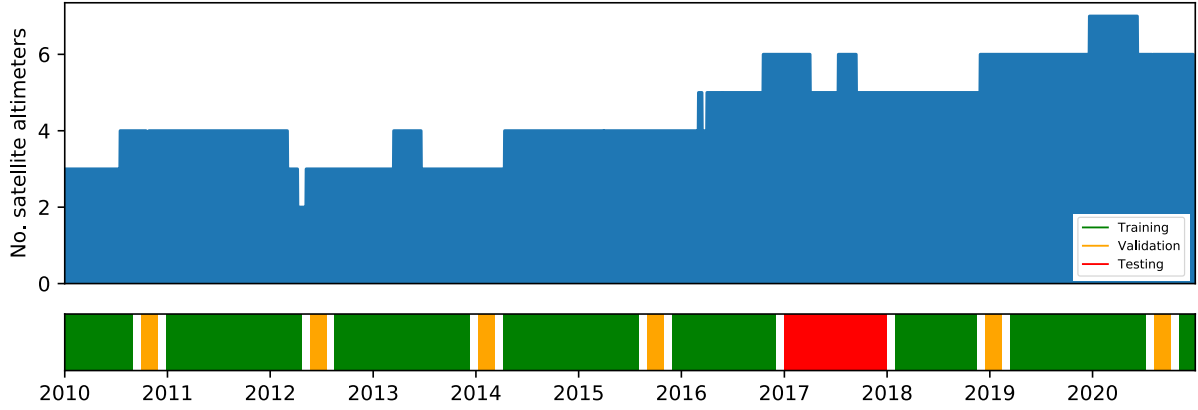


Figure 1.3: Dates are partitioned into non-overlapping and spaced out training (green), validation (orange), and testing (red) windows (the white bars represent the gaps between partitions when no examples are drawn). Examples are then drawn centered on days randomly drawn from the respective windows. The blue histogram shows the changing availability of satellite altimeters from which we can draw data.

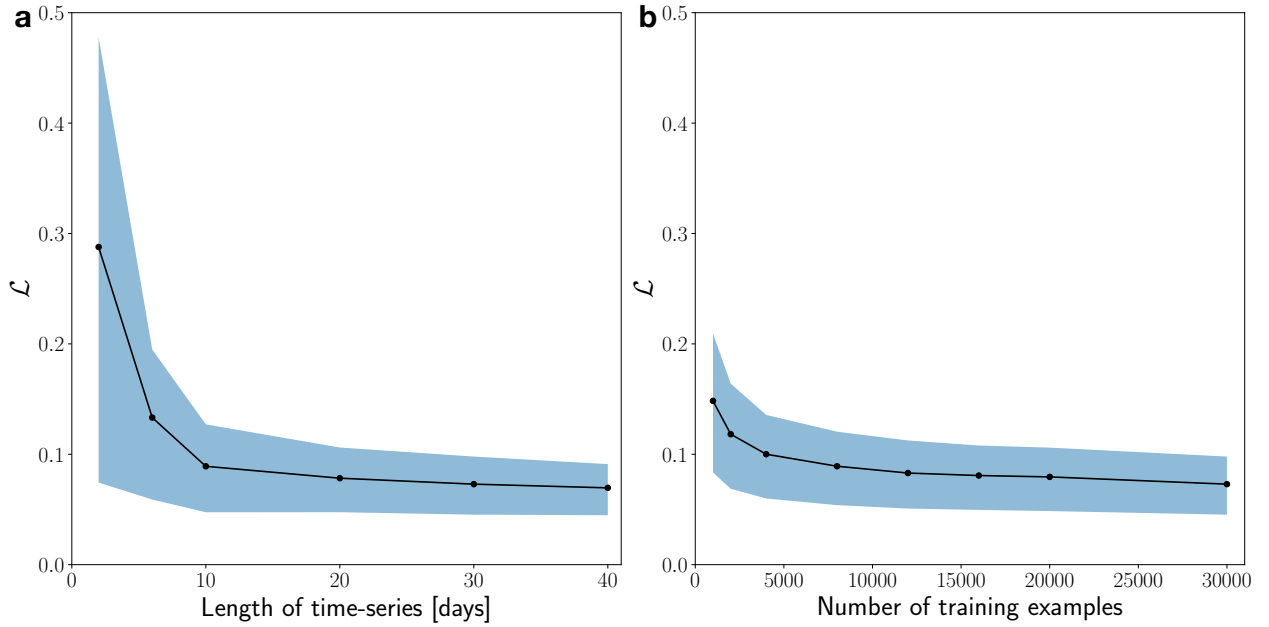


Figure 1.4: (a): the mean cost, $\mathcal{L}$, achieved on the validation dataset (containing 2000 samples) for our ConvLSTM model with SSH and SST input observations for increasing time series length. The blue shading shows the 16th and 84th percentiles, thus containing the same fraction of points as $\pm 1\sigma$ for a normal distribution. (b) the reduction in $\mathcal{L}$ with increasing numbers of training examples for our ConvLSTM model with the length of the input observation time series set to 30 days (15 days either side of the reconstruction). Shading as for (a).

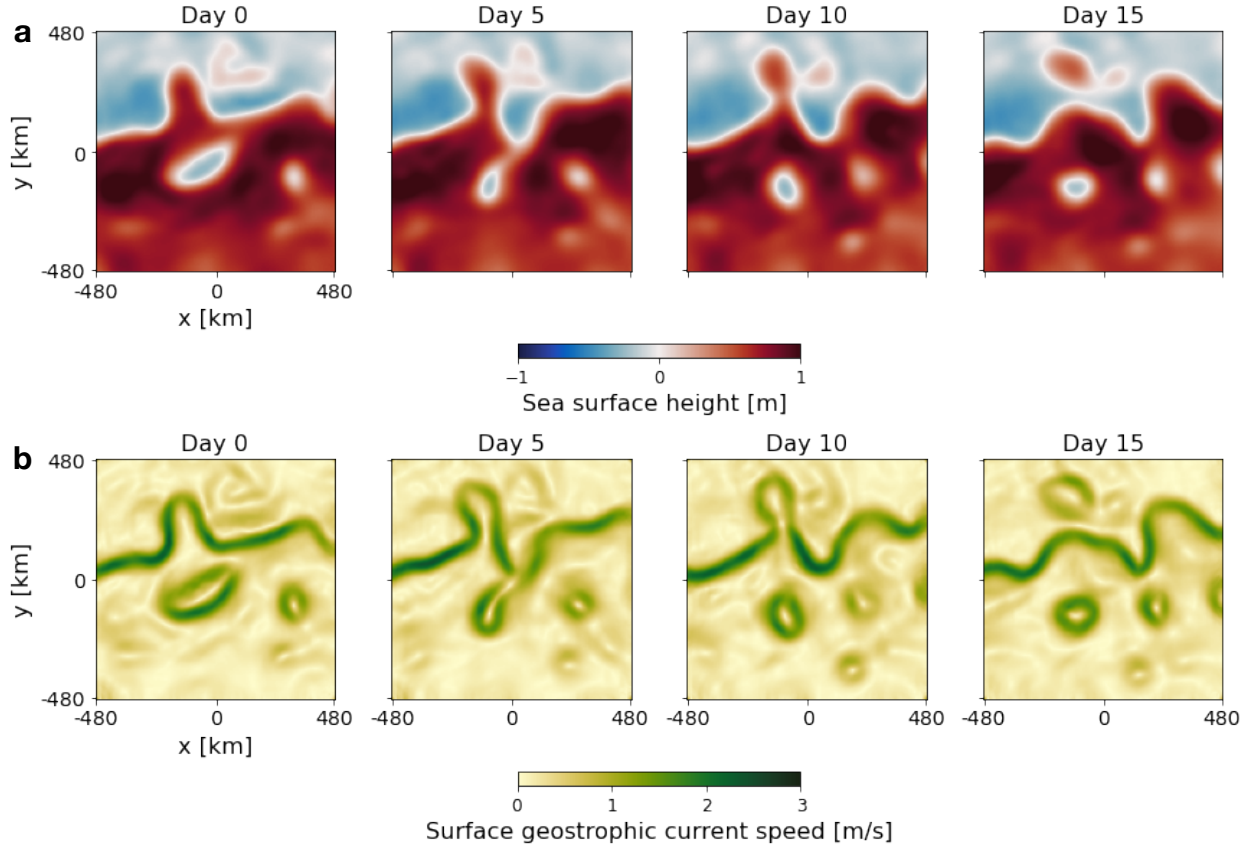


Figure 1.5: An example of the output from our ConvLSTM SSH+SST mapping method. Time series of (a) SSH (absolute dynamic topography) mapped using ConvLSTM SSH+SST and (b) surface geostrophic current speed calculated from the SSH field.

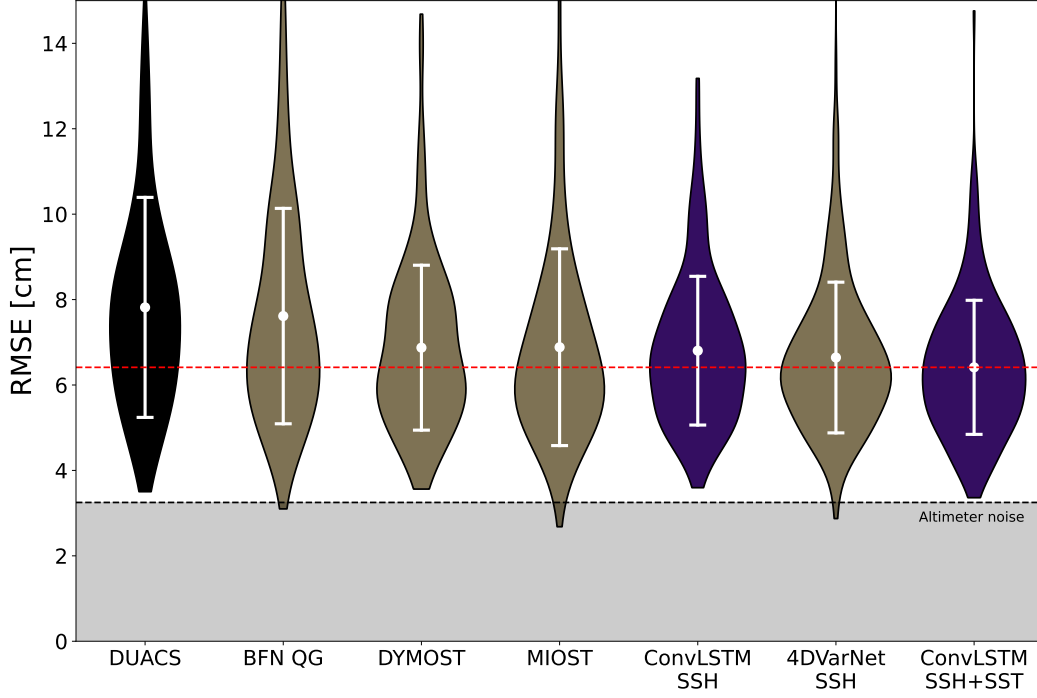


Figure 1.6: Violin plot showing the distribution of daily reconstruction RMSE for each method when compared to independent satellite altimeter observations for the test period. The black violin shows the community-standard OI product, the gold violins are experimental methods proposed by others in the literature, and the purple violins are using our new method. The width of each violin is proportional to the probability density function, the white circle indicates the mean, and the whiskers show one standard deviation. The shaded region indicates the estimated instrument noise in the observations (as reported by CMEMS).

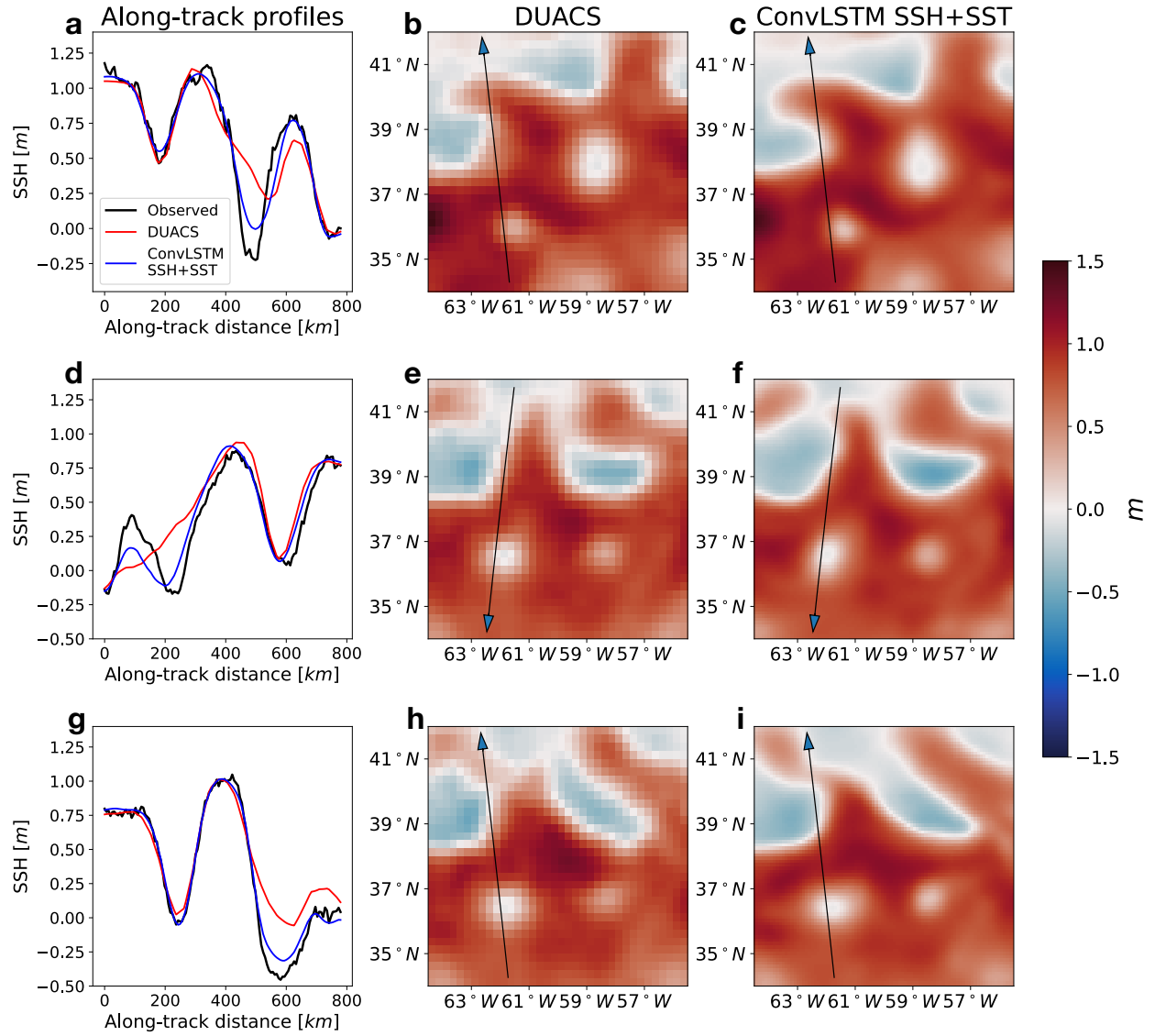


Figure 1.7: Example along-track SSH profiles. Shown are the three 800km track segments (a, d, g) where DUACS suffers the highest RMSE during 2017 along with the track's location superimposed over the DUACS reconstruction (b, e, h) and the ConvLSTM SSH-SST reconstruction (c, f, i). In each case, DUACS over-smoothes one or more large peaks/troughs in SSH. In all of these cases ConvLSTM provides a significant improvement.

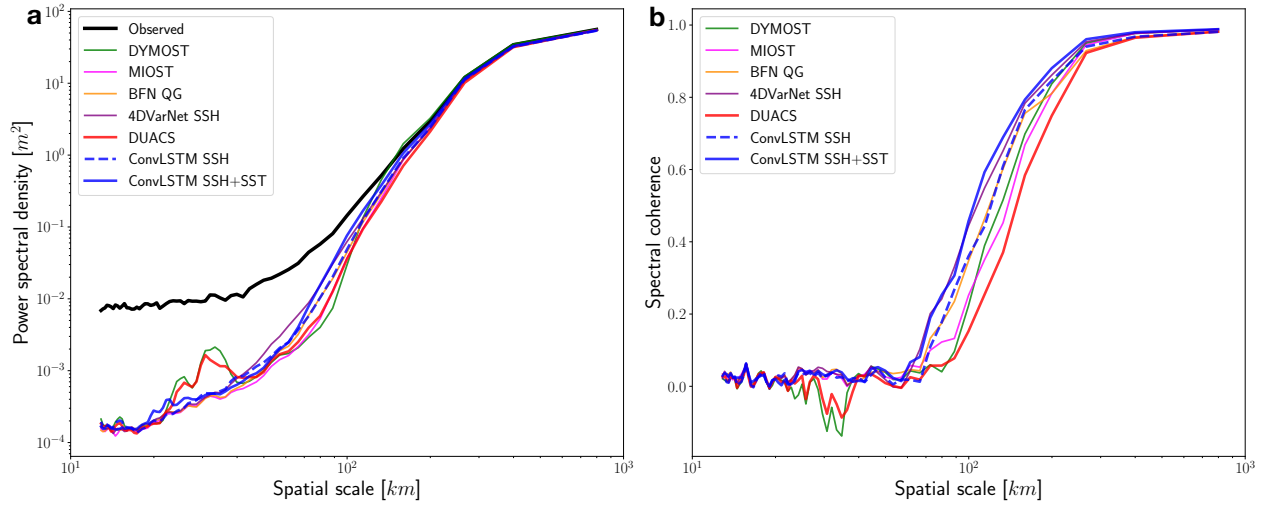


Figure 1.8: (a): wavenumber spectrum of mapped SSH along the tracks of the withheld CryoSat-2 observations for each method, compared to that of the observations. The observations show a high power spectral density for small scales due to instrument noise. (b): spectral coherence for each mapping method (approaches unity for scales at which the SSH signal is well resolved by the map).

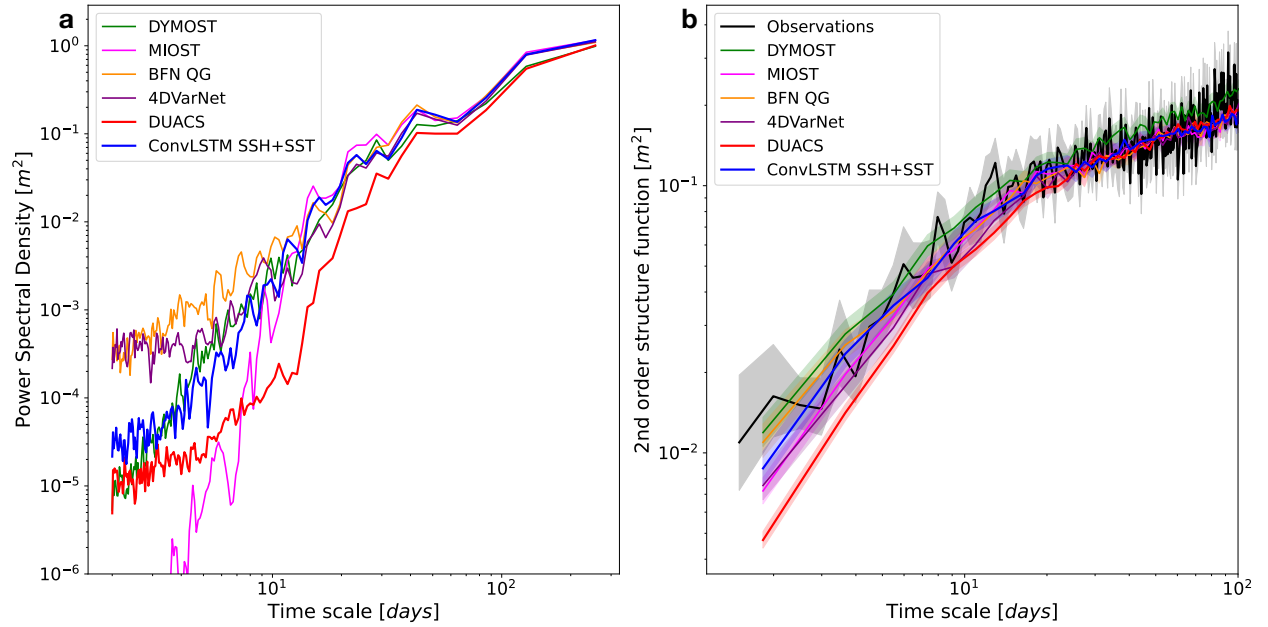


Figure 1.9: (a): frequency spectrum for each SSH map averaged over points throughout our study region. (b): second order structure function calculated from all available satellite observations (including those used to make the maps) at satellite track crossover points, compared against that of each map average over uniformly-spaced points in the domain. Error bars show the 95% confidence interval estimated using the bias-corrected and accelerated bootstrap method.

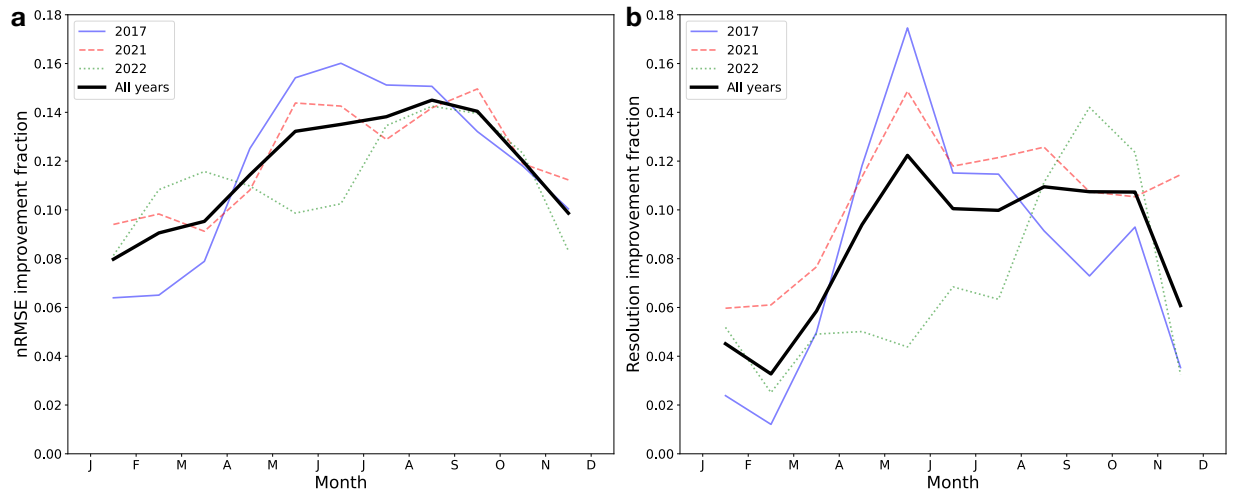


Figure 1.10: (a): fraction by which the normalized RMSE (nRMSE) was reduced with ConvLSTM SSH+SST compared to ConvLSTM SSH. Maps were created from all possible three-satellite combinations and the mean is shown for each year along with the combined mean over the three years (weighted to account for the different number of satellite configurations in each year). (b): fraction by which the effective spatial resolution was reduced for ConvLSTM SSH+SST compared to ConvLSTM SSH. Again, the mean across all three-satellite configurations is shown for each year along with the combined three-year mean.

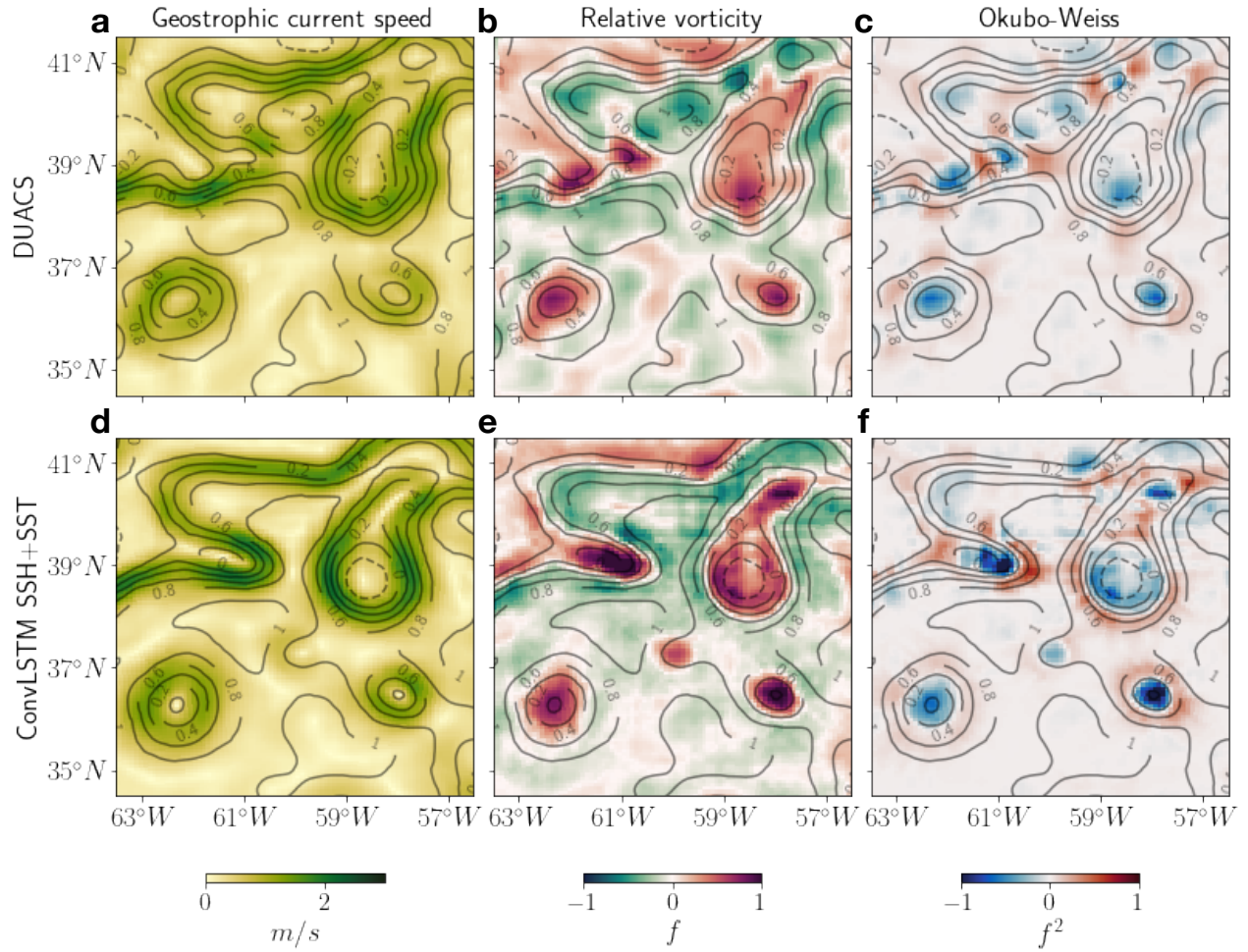


Figure 1.11: (a-c): surface geostrophic current speed, relative vorticity, and the Okubo-Weiss quantity calculated from the DUACS SSH map. (d-f): the same quantities calculated from the SSH map produced by our ConvLSTM SSH+SST method. Contours of the corresponding SSH maps (absolute dynamic topography) are overlaid in black. Relative vorticity and Okubo-Weiss are normalized by the Coriolis frequency, f , and its square respectively. A video of this figure is available for the full year in the supporting information (Movie S1).

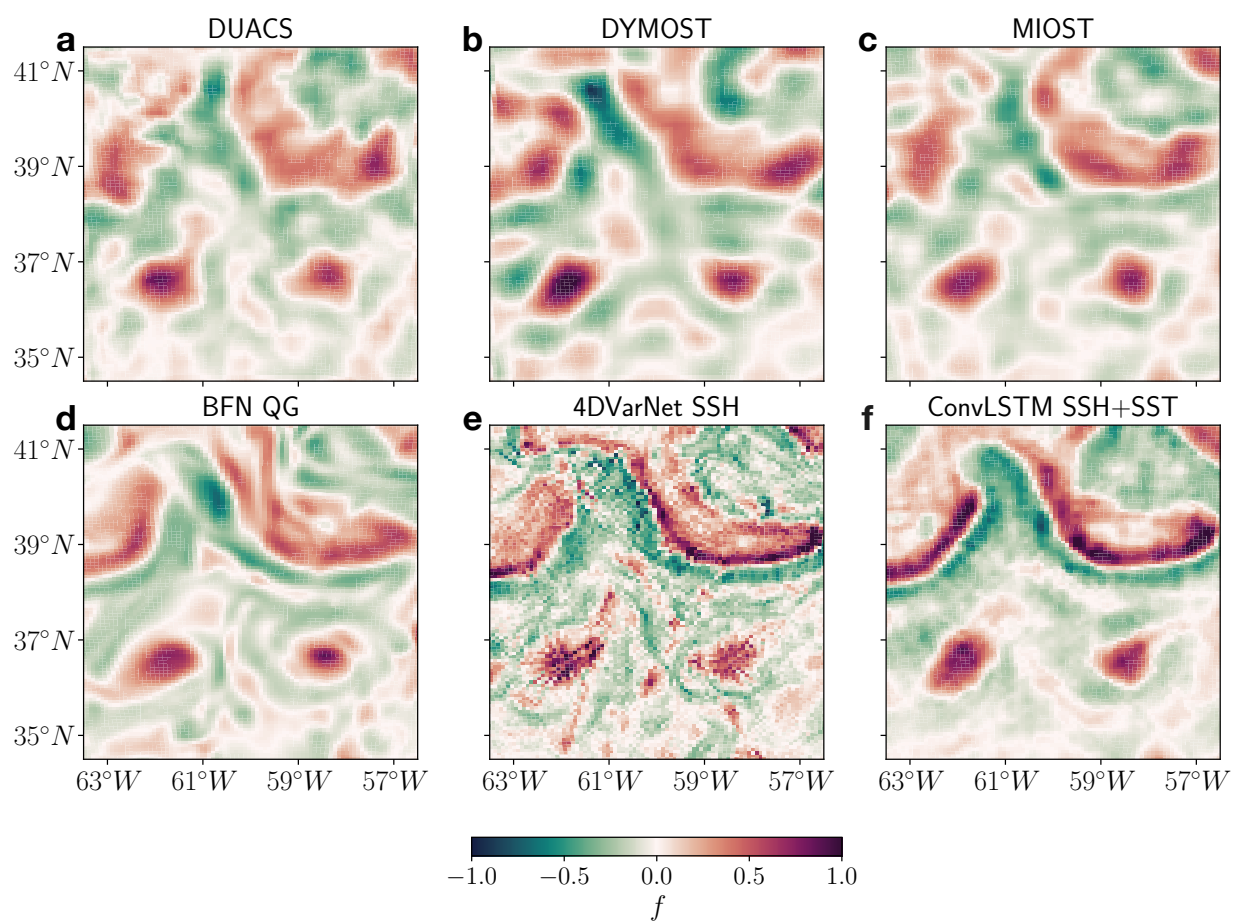


Figure 1.12: Relative vorticity calculated from the surface geostrophic currents from each method's SSH map, normalized by the Coriolis frequency, f . A video of this figure is available for the full year in the supporting information (Movie S2).

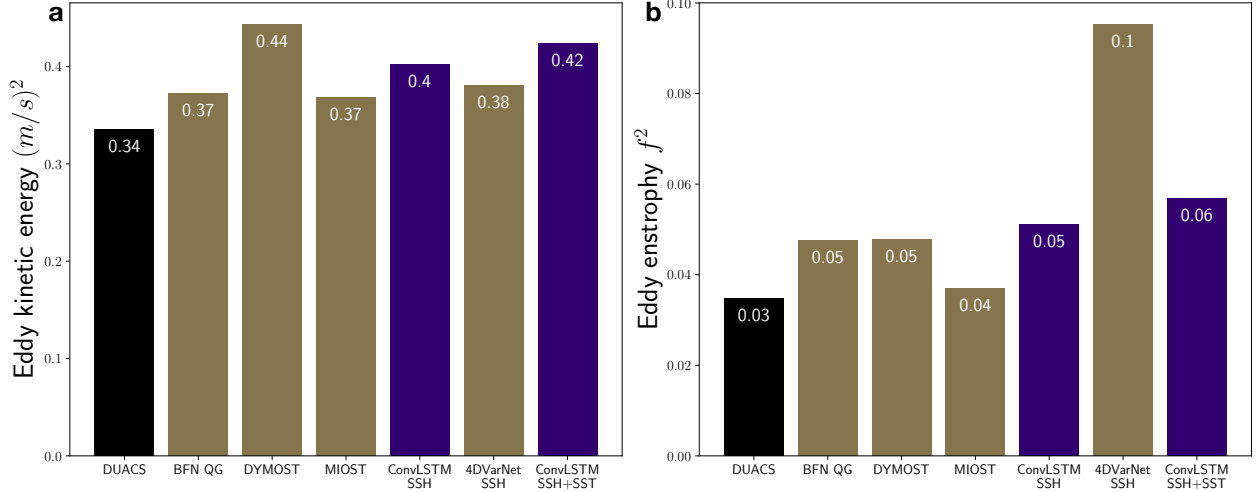


Figure 1.13: Domain- and time-averaged eddy kinetic energy (a) and enstrophy (b) for the geostrophic currents calculated from each SSH map. Enstrophy is normalized by the square of the Coriolis frequency, f . The black bars shows the community-standard OI product, the gold bars are experimental methods proposed by others in the literature, and the purple bars are using our new method.

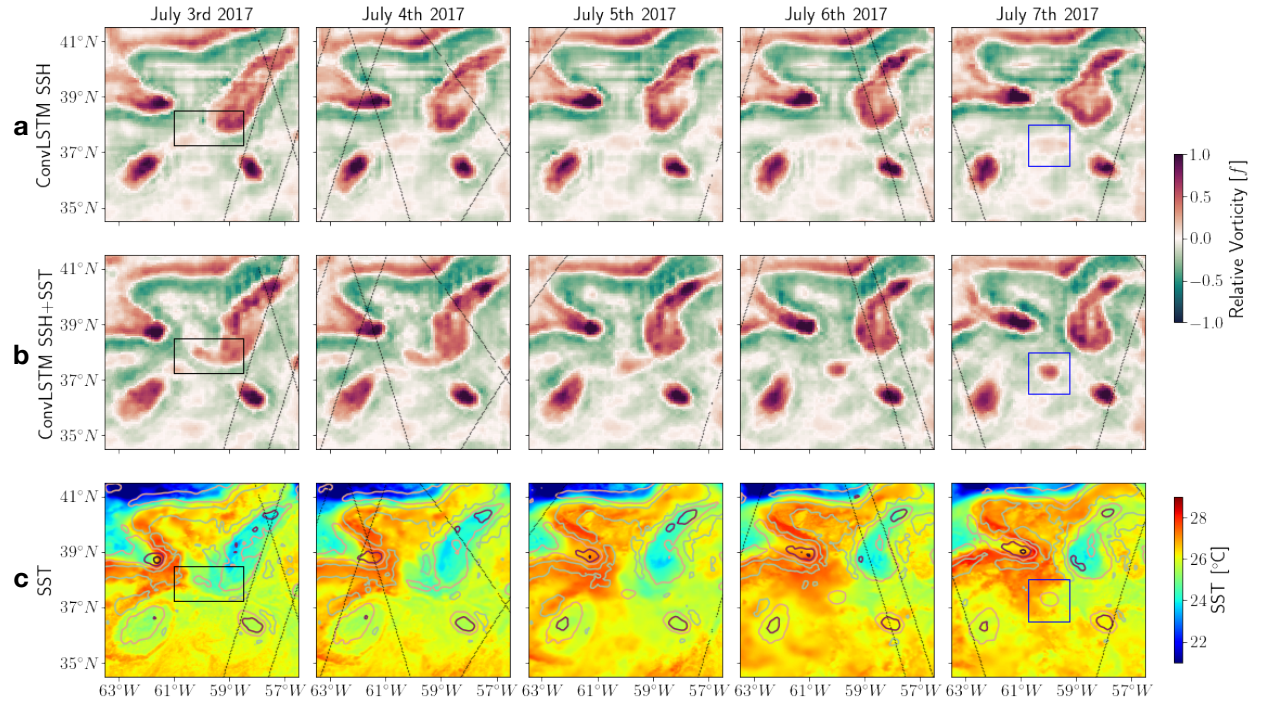


Figure 1.14: Relative vorticity from (a) the ConvLSTM SSH and (b) the ConvLSTM SSH+SST reconstructions for 5 consecutive days. GHRSSST MUR SST for the same days is shown in row (c) and the locations of the satellite altimetry observations on each day are overlaid in black. ConvLSTM SSH+SST relative vorticity contours in increments of $0.5f$ from $-1.25f$ to $1.25f$ are overlaid over the SST observations. The black box on July 3rd highlights the filament from which this eddy formed.

Chapter 2

Deep learning improves global satellite observations of ocean eddy dynamics

This chapter is reproduced from the following publication:

Martin, S. A., Manucharyan, G. E., & Klein, P. (2024). Deep learning improves global satellite observations of ocean eddy dynamics. Geophysical Research Letters, 51, e2024GL110059. <https://doi.org/10.1029/2024GL110059>

2.1 Introduction

Mesoscale eddies (50-300 km) are a critical component of the global ocean circulation, transporting dynamical and biogeochemical tracers [Wunsch, 1999; Jayne and Marotzke, 2002; Zhang et al., 2014]. Despite being the ocean's dominant reservoir of kinetic energy (KE), the sources and sinks of mesoscale eddy KE remain poorly constrained [Ferrari and Wunsch, 2009]. One major process affecting mesoscale KE is the transfer of KE between scales by non-linear eddy interactions, known as the KE cascade [Scott and Wang, 2005; Aluie et al., 2018; Klein et al., 2019]. There is growing evidence that non-linear eddy interactions induce a strongly seasonal upscale KE cascade, KE transfer from small to large scales, that is intensified in winter and spring [Sasaki et al., 2014; Qiu et al., 2014; Uchida et al., 2017; Schubert et al., 2020; Ajayi et al., 2021; Balwada et al., 2022; Naveira Garabato et al., 2022; Steinberg et al., 2022; Lawrence and Cal-

lies, 2022; Schubert et al., 2023; Storer et al., 2023]. The strength of non-linear eddy interactions and the KE cascade are set by the vorticity and strain [Aluie et al., 2018; Klein et al., 2019], which are sensitive to the geometry of eddies. There is thus a need for global observations of eddies with sufficient resolution to accurately diagnose vorticity and strain - this motivates our study.

Satellite observations of eddy surface expressions are a powerful observing system for eddy dynamics since satellites resolve a wide range of scales compared to in situ observations [Klein et al., 2019], however, inferring surface currents from satellite observables is challenging. Satellite altimetry allows the estimation of eddies by mapping their expression in sea surface height (SSH), which is used to estimate surface geostrophic currents [Chelton et al., 2001]. Conventional altimeters measure SSH and resolve mesoscale eddies along each satellite’s track [Dufau et al., 2016] but leave large gaps between tracks that must be interpolated to diagnose eddy dynamics. Meanwhile, satellites observe high-resolution 2D snapshots of sea surface temperature (SST) but there are gaps due to clouds and the relationship between SST and surface currents is complex [Isern-Fontanet et al., 2006, 2014; Rio et al., 2016].

Conventionally, surface currents are estimated either via data assimilation (DA) [Lellouche et al., 2021; Le Guillou et al., 2021, 2023] or objective analysis (OA) of SSH [Taburet et al., 2019; Ubelmann et al., 2015, 2021, 2022]. DA provides 3D state estimates approximately consistent with the physics of a numerical model, typically a general circulation model (GCM). However, state-of-the-art DA systems [Lellouche et al., 2021] use GCM resolutions that only partially resolve mesoscale dynamics and so suffer high errors in SSH and surface currents at mesoscales due to unresolved eddy dynamics and the lack of high-resolution 3D in-situ observations. In contrast, OA allows to estimate only the 2D SSH field using a statistical approach [Taburet et al., 2019; Ubelmann et al., 2015, 2021, 2022], from which currents can be estimated through geostrophy and further corrections for winds [Rio et al., 2014], equatorial dynamics [Lagerloef et al., 1999a], cyclo-geostrophy [Penven et al., 2014; Cao et al., 2023], or potentially using machine learning [Sinha and Abernathey, 2021; Xiao et al., 2023]. Surface currents derived from OA SSH fields give eddy amplitudes and configurations more consistent with observations than DA since they are not biased by unresolved dynamics. However, OA is a fundamentally statistical approach that does not guarantee a physically consistent reconstruction and biases in the covariance models used in OA could bias the reconstructions. OA artificially suppresses variance at smaller scales, smoothing and distorting eddies [Ballarotta

et al., 2019]. This leads to a significant underestimation of crucial dynamical quantities, like vorticity and strain.

Deep learning has emerged as an alternative approach for estimating surface currents. Many proof-of-concept studies demonstrate that neural networks can be trained to map SSH or surface currents from altimeter observations, through either ‘simulation learning’ using synthetic data from high-resolution GCMs [Fablet et al., 2021a; Manucharyan et al., 2021; Buongiorno Nardelli et al., 2022; Beauchamp et al., 2022; Fablet et al., 2023; Thiria et al., 2023; Febvre et al., 2024b; Archambault et al., 2024a; Kugusheva et al., 2024] or ‘observation-only learning’ from real-world satellites [Martin et al., 2023; Archambault et al., 2023]. Deep learning allows the optimal mapping to emerge objectively from the data, unlike OA which prescribes linear covariance models [Taburet et al., 2019], and allows to use SST observations as an additional input to improve the mapping between altimeter observations [Buongiorno Nardelli et al., 2022; Fablet et al., 2023; Martin et al., 2023; Archambault et al., 2023, 2024a]. Simulation learning showed promising results on synthetic observations in proof-of-concept studies. However, transferring these methods to real-world observations remains a challenge since GCMs are not exact analogs of the real world and neural networks behave unpredictably when applied to data different from that used during training. This domain gap can be partly addressed through fine-tuning on real-world observations [Febvre et al., 2024b; Archambault et al., 2024a]. Fundamentally, simulation learning blurs the boundary between observations and GCMs, much like DA. In contrast to simulation learning, observation-only learning, analogous to OA, is directly applicable to real-world observations and is uncontaminated with GCM biases. This comes at the expense of smoothing some smaller-scale features due to the limited resolution of real-world observations. Nonetheless, regional proof-of-concept studies have shown observation-only learning can give SSH maps with higher resolution than OA, leading to significant improvements in the estimation of vorticity and strain [Martin et al., 2023; Archambault et al., 2023].

Extending regional proof-of-concept studies to global SSH mapping poses a significant challenge for deep learning because the global ocean exhibits spatiotemporally diverse dynamics. Given the sparsity of the altimetry record for observation-only learning, it remains to be demonstrated that a neural network can generalize across all dynamical regimes. We hypothesize that observation-only learning can be used to create global SSH maps with enhanced resolution, and that this will radically improve global observations of

vorticity and strain - and hence of non-linear eddy dynamics. Here, we develop the first global deep learning estimates of SSH and surface currents. We evaluate the accuracy of the SSH maps, their ability to resolve vorticity and strain, and explore the resulting KE cascade. By disseminating our new global SSH product, we hope to enable more accurate studies of eddy dynamics and their impact on general ocean circulation, marine ecosystems, and climate.

2.2 Methods

2.2.1 NeurOST: Global SSH Maps from Altimetry and SST Using Deep Learning

We train a neural network to map SSH from sparse altimeter observations [E.U. Copernicus Marine Service Information (CMEMS), 2024c,b] and gridded SST [JPL MUR MEaSUREs Project, 2015]. Our approach (Figure 2.1), builds upon that described in our recent proof-of-concept study [Martin et al., 2023] and is described in full in S.I. S1.1-5.

We use ‘self-supervised’ learning, taking a time series of altimeter observations within a local subdomain (30 days by 960 km by 960 km) from all but one of the available altimeters alongside the corresponding SST as input to a neural network tasked with reconstructing 2D SSH. The objective minimized during training is the mean square error of the mapped SSH calculated against the withheld altimeter. We restrict the mapping to local subdomains since eddy dynamics are local, so a global ‘field of view’ is likely unnecessary to reconstruct eddies in any local subdomain.

We use kernel-weighted averaging to combine thousands of overlapping subdomain SSH maps together into a single global SSH map (S.I. S1.5 and Callaham et al. [2019]). Using a large set of subdomain examples drawn from across the globe, we train a single network to map SSH in all regions, achieving generalization across diverse regional dynamics. By training a single global network rather than an ensemble of bespoke regional networks we avoid arbitrarily dividing the globe into regions and learn a general and robust SSH mapping. The network was trained on observations from 2010 to 2023, with 2019 withheld for validation (Figure S8).

We refer to our method as ‘NeurOST (SSH-SST)’ (Neural Ocean Surface Topography). To assess the value of SST we also trained a network to map SSH from altimetry alone; ‘NeurOST (SSH)’.

2.2.2 Estimating Surface Currents from SSH

Large-scale currents satisfy geostrophic balance, allowing surface currents to be estimated from SSH through geostrophy (S.I. Equation 1). The limitations of geostrophy and potential empirical ageostrophic corrections [Lagerloef et al., 1999a; Rio et al., 2014; Penven et al., 2014; Cao et al., 2023] are discussed in S.I. S1.1 where we also show diagnostics of eddy dynamics (e.g. KE cascade) are only weakly sensitive to the cyclo-geostrophy correction [Penven et al., 2014]. Thus, throughout this manuscript, surface currents were calculated using geostrophy.

2.2.3 SSH Mapping Evaluation: Observing System Experiment (OSE)

We employ an observing system experiment (OSE) to evaluate SSH maps. Comparisons to existing methods are achieved using an Ocean Data Challenge [Metref et al., 2023; Metref and Ballarotta, 2023] in which developers of different methods implement them on a common experiment. In the OSE used here, we create global SSH maps for 2019 using all altimeters apart from Saral/Altika that is used for evaluation. Accuracy is evaluated using root mean square error (RMSE), and we quantify the maps’ effective spatial resolution (the smallest resolved wavelength) following Ballarotta et al. [2019]. We compare NeurOST to the community-standard ‘DUACS’ product [Le Traon et al., 1998; Taburet et al., 2019] as well as to the ‘MIOST’ method [Ubelmann et al., 2021; Ballarotta et al., 2023]. Surface geostrophic currents are evaluated using drifters [Copernicus Marine in situ TAC, 2024]. Additionally, we compare NeurOST to proof-of-concept methods in the Gulf Stream Extension using a similar Ocean Data Challenge that was regional in scope [Ballarotta et al., 2021]. (S.I. S1.6-7)

2.2.4 Eddy Dynamics Evaluation: Observing System Simulation Experiment (OSSE)

While the OSE evaluates SSH maps, we cannot use it to evaluate eddy dynamics (vorticity and strain) inferred from SSH as this requires the full 2D eddy field. We therefore conduct an observing system simulation experiment (OSSE) where we generate synthetic altimeter observations from the 1/12° GLORYS reanalysis [Lellouche et al., 2021; E.U. Copernicus Marine Service Information (CMEMS), 2024f] and use them in combination with GLORYS SST as input to NeurOST, with no re-training on GLORYS. We compare the resulting NeurOST maps to the 2D ground-truth from GLORYS to evaluate eddy dynamics diagnos-

tics, specifically surface geostrophic currents and vorticity for which we define skill scores representing the fraction of variance explained (S.I. S1.11). This point-wise comparison to GLORYS cannot be made for DUACS since this method is not open source, preventing its implementation on simulated observations.

2.2.5 Kinetic Energy Cascade Diagnosis

We use NeurOST surface geostrophic currents to diagnose the KE cascade in a range of regions through coarse-graining [Aluie et al., 2018; Storer et al., 2022; Storer and Aluie, 2023; Storer et al., 2023]. The KE cascade is diagnosed through the spectral KE flux, which quantifies KE transfer from larger to smaller scales at each wavelength. A positive flux indicates a downscale (forward) cascade, whereas a negative value indicates an upscale (inverse) cascade. (S.I. S1.12)

2.3 Results

2.3.1 State-of-the-Art Global SSH Maps Using Deep Learning

Our global SSH maps (NeurOST SSH-SST) show rich dynamical structures associated with western boundary currents, abundant mesoscale eddies in the extratropics, and large-scale equatorial waves in the tropics (Figure 2.1).

The effective resolution of our maps is improved compared to DUACS throughout the global ocean, with a pronounced improvement in western boundary currents and the subtropics where we resolve wavelengths 30% smaller (Figure 2.2b,d and Table S1). The global RMSE of the mapped SSH is 6% lower than DUACS, while reductions in the RMSE of small mesoscales (70-250km wavelengths) reach 20% in eddy-rich regions (Figure 2.2a,c,e and Table S1). Similarly, NeurOST outperforms MIOST in almost all regions, especially for small mesoscales, making NeurOST state-of-the-art in global SSH mapping (Figure S1 and Table S1).

Using SST improves the mapping of SSH throughout the global ocean (Figure 2.2f and Table S1). To assess the utility of SST, we compare the performance of NeurOST with and without SST. The mapping of small mesoscales is improved using SST, especially in the extratropics where mesoscale SSH and SST are correlated [Cornillon et al., 2019] (Figure 2.2f and Table S1). SST is especially impactful when few altimeters are available (Table S3). While observations from six altimeters were used to create the maps

compared above, for much of the altimetry era only two altimeters were operational, causing eddies in DUACS to be severely smoothed. We evaluated our network using only two altimeters and found that in most regions NeurOST SSH-SST with just two altimeters yields higher-resolution SSH than DUACS achieves with six (Table S3). This highlights the potential of deep learning and SST to extract maximum value from the thirty-year altimetry record.

NeurOST maps SSH across all regions, unlike prior studies that trained bespoke region-specific networks. While regional networks [Martin et al., 2023; Febvre et al., 2024b] in the Gulf Stream offer marginally improved SSH mapping compared to NeurOST [Ballarotta et al., 2021] (Table S2), fine-tuning on a smaller set of observations from the Gulf Stream (S.I. S1.6) brings NeurOST in line with state-of-the-art regional networks (Table S2). This shows the potential for further refinement of NeurOST by users interested only in a single region.

Surface geostrophic currents from our maps are more accurate when evaluated with drifter observations. NeurOST reduces the RMSE of surface currents significantly across the global ocean, especially in the subtropics, where RMSE is reduced by 20% compared to DUACS (Figure S2). Discrepancies between the mapped currents and drifter observations are due to both the accuracy of the mapped geostrophic current and the degree to which real-world currents are in geostrophic balance. Nonetheless, this large reduction in RMSE demonstrates the significant improvement in the mapped currents.

2.3.2 Improved Physical Realism of Mesoscale Eddies

Calculating vorticity and strain from the surface current maps appears to show a significant qualitative improvement in the realism of eddy dynamics in NeurOST (Figure 2.3b,c,d and Movie S1). NeurOST vorticity shows abundant small mesoscale eddies with clear boundaries, many of which are absent in DUACS (Figure 2.3b,c). A contrasting view of eddy dynamics emerges when comparing the eddy evolution: eddies appear to uniformly propagate westward in DUACS, while NeurOST eddies exhibit strong non-linear interactions that deform each other's vorticity cores, causing filamentation (Supplementary Movie S1). These better-resolved non-linear eddy interactions also manifest in the higher strain rate between eddies seen in NeurOST, evidenced by regions of highly positive Okubo-Weiss quantity (Figure 2.3d, S.I. S1.10). Increased strain has important implications for eddy dynamics since it is associated with enhanced frontogenesis [Hoskins, 1982;

Siegelman et al., 2020a] and a stronger KE cascade [Aluie et al., 2018]. While vorticity and strain appear qualitatively more realistic in NeurOST, their accuracy cannot be quantified using along-track SSH observations. To demonstrate that NeurOST does not introduce artificial eddies, we test its ability to reconstruct vorticity using synthetic observations from a GCM (Section 2.2.4).

NeurOST, trained on real-world observations and applied now to synthetic observations from GLORYS, skillfully reconstructs surface currents, especially in the subtropics and western boundary currents, explaining over 70% of the variance (Figure 2.4a). Its skill deteriorates somewhat in regions of low variability, at high latitudes, and near coasts, where the observational training data is likely to significantly differ from the GLORYS simulation. Since vorticity is sensitive to small-scale SSH features, its overall reconstruction skill is slightly lower than that for surface currents. Nonetheless, NeurOST reconstructs a remarkable 50-80% of vorticity variance throughout the subtropics and western boundary currents. Comparing spatial patterns of vorticity and strain, it is clear that NeurOST misses smaller-scale filaments but skillfully reconstructs larger eddies and some larger filaments and is not prone to creating artificial eddies (Fig 2.4c-f and Supplementary Movie S2). NeurOST can reconstruct features as small as 50 km (see filaments in Figure 2.4c-f). Thus, NeurOST can reasonably well reconstruct the 2D vorticity and strain fields despite being trained only on real-world along-track SSH observations and never on real or simulated vorticity/strain.

Since GLORYS contains finer-scale vorticity features than NeurOST, it can be used to estimate the KE cascade (see Storer et al. [2023] and Figure S7). However, there is a qualitative difference in eddy dynamics between GLORYS and NeurOST that clearly manifests in small-scale vorticity features. In GLORYS, vorticity features are dominated by persistent filaments, whereas in NeurOST there is an abundance of smaller-scale coherent eddies with less prominent filaments (Supplementary Movie S1). This difference could be due to the relatively coarse grid of GLORYS ($1/12^\circ$) that does not resolve the generation of small-scale eddies by submesoscale instabilities in the mixed layer, which are known to be prominent in winter. In such coarse-resolution models, large eddies stir vorticity to form small-scale filaments, but they can artificially persist and grow, being constrained only by numerical or specified model diffusion. In reality, the presence of small-scale eddies can disrupt this filamentation by large-scale eddies, and filaments often become unstable and form submesoscale eddies (e.g., Taylor and Thompson [2023]). Since there are no ground-truth observations of vorticity, one cannot definitively establish whether there is excessive filamen-

tation in GLORYS or whether NeurOST introduces artificial eddies at small scales. However, looking at the NeurOST reconstruction of GLORYS (Section 2.4), NeurOST appears to provide a coarse-grained view of GLORYS, showing no evidence of artificial small-scale eddies being introduced (compare Supplementary Movies S1 & S2). Furthermore, when reconstructing real-world SSH and surface currents, NeurOST also has significantly lower errors than GLORYS (Figure S3 and Table S1). With increased confidence that NeurOST provides a better estimation of real-world ocean eddy dynamics, we now proceed to explore its impact on our understanding of the KE cascade and seasonality of mesoscale eddies.

2.3.3 Seasonal Kinetic Energy Cascade

A distinct seasonality in eddy dynamics emerges in NeurOST that was largely absent in DUACS, with smaller scale eddies peaking in intensity in winter and spring. The mesoscale KE and strain rate throughout the subtropics are 50-100% higher in NeurOST than in DUACS in winter/spring, whereas they are comparable in summer/autumn (Figure S4 & S5). We explore this newly-resolved seasonality by focusing on the Subtropical North Pacific, which was the subject of prior studies of eddy seasonality [Qiu et al., 2014].

Eddy dynamics from NeurOST are strongly seasonal in the Subtropical North Pacific, with enstrophy (the variance of vorticity) peaking in winter/spring implying intensified small-scale eddies (Figure 2.5a). This strong wintertime peak in enstrophy in NeurOST is qualitatively consistent with submesoscale-resolving simulations [Qiu et al., 2014]. The seasonality of small-scale eddies is corroborated by the KE wavenumber spectrum (S.I. S1.13) that has a shallower slope ($\approx k^{-2}$) in winter/spring than in summer/autumn ($\approx k^{-3}$) (Figure 2.5b), meaning energy is more concentrated at small scales in winter/spring. Notably, the peak in KE for small-scales ($<125\text{km}$) leads that for larger scales by two months (Figure 2.5c). It has been hypothesized that the delayed large-scale KE peak may partly be driven by an upscale KE cascade from submesoscales, which are most energetic during winter [Sasaki et al., 2014; Qiu et al., 2014]. However, this hypothesis has not been confirmed using observations since low-resolution products like DUACS fail to resolve the small-scale eddies that proliferate in winter/spring (Figure 2.5a,c), and hence underestimate the KE cascade [Arbic et al., 2013].

The KE cascade from NeurOST is upscale throughout the mesoscale range in the Subtropical North Pacific, has a strong springtime peak, and is dramatically stronger than in DUACS (Figure 2.5d). Although

some KE sources and sinks cannot be derived from surface currents (S.I. S1.12), the magnitude of the cascade appears more than sufficient to drive the increase in large-scale KE observed over winter/spring (Figure 2.5e). In contrast, the overly smooth DUACS product significantly underestimates the springtime KE cascade, obscuring the seasonality of the cascade and its role in driving the large-scale peak (Figure 2.5e). The cascade also appears to play a crucial role in driving large-scale seasonality in other subtropical regions (Figure S6). At scales resolved in our study, there are no ground-truth data for the KE cascade to compare against [Yoo et al., 2018; Naveira Garabato et al., 2022; Balwada et al., 2022], but recent studies of along-track altimetry provide indirect evidence for the seasonal upscale KE cascade observed here [Steinberg et al., 2022; Lawrence and Callies, 2022; Schubert et al., 2023]. Note that in western boundary currents, lateral KE advection appears to dominate the cascade (Figure S6), and there is likely a substantial KE injection by instabilities of large-scale currents that cannot be diagnosed from SSH. Note that NeurOST (SSH-SST) shows a stronger seasonality in the cascade than NeurOST (SSH), likely since small wintertime eddies are hard to capture without tracking their expression in SST (Figure S12).

2.4 Conclusions

Our high-resolution SSH maps generated using deep learning represent a large stride forward for the global observation of ocean eddy dynamics, providing state-of-the-art global surface currents. Trained on real-world observations alone, NeurOST allows to diagnose eddy dynamics with greater physical realism than from existing altimetry products. NeurOST revealed the crucial role of non-linear eddy dynamics and their associated KE cascade in driving the seasonality of mesoscale eddies in many parts of the global ocean, highlighting the importance of resolving small-scale eddy dynamics in ocean models. Alongside this manuscript, we publish a NeurOST SSH product [OSTST, 2024] to facilitate future studies of eddy dynamics and the impacts of eddies on climate and marine ecosystems.

Despite the improved resolution of NeurOST, it does not yet resolve submesoscale eddies, smoothing scales below $O(100\text{km})$ (Table S4). Hence, the strength of the upscale cascade is likely still underestimated (Figure S7), and the potential presence of a downscale cascade at submesoscales cannot be quantified. The recently launched Surface Water and Ocean Topography (SWOT) satellite, the first wide-swath altimeter [Morrow et al., 2019; Fu et al., 2024], provides unprecedented 2D submesoscale-resolving SSH snapshots

that could help characterize the KE cascade at submesoscales [Klein et al., 2019; Carli et al., 2023]. However, SWOT observations present new challenges for inferring currents from SSH due to the presence of unbalanced submesoscale SSH variability and the mismatch between fast-evolving submesoscale dynamics and SWOT's long return times [Gaultier et al., 2016; Torres et al., 2018]. Deep learning methods to address these issues are under development and show promise [Febvre et al., 2022; Wang et al., 2022a; Gao et al., 2024]. As satellite oceanography enters a new submesoscale-resolving era [Morrow et al., 2019; Fu et al., 2024], further development of deep learning methods will be crucial to best monitor surface currents and other essential climate variables, like SST [Goh et al., 2023; Agabin et al., 2024].

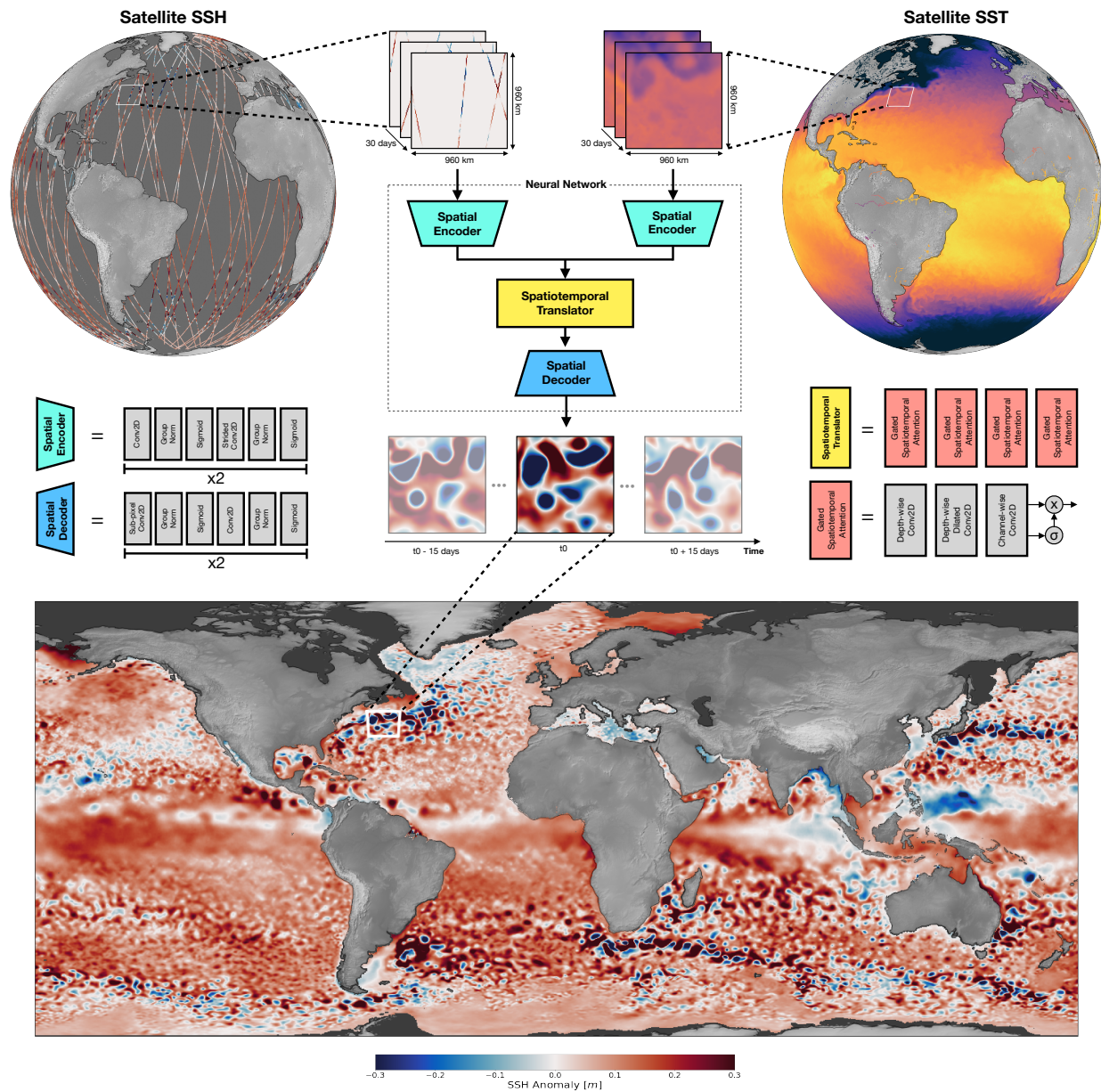


Figure 2.1: Schematic of NeurOST SSH-SST method for mapping SSH from satellite altimetry and SST.

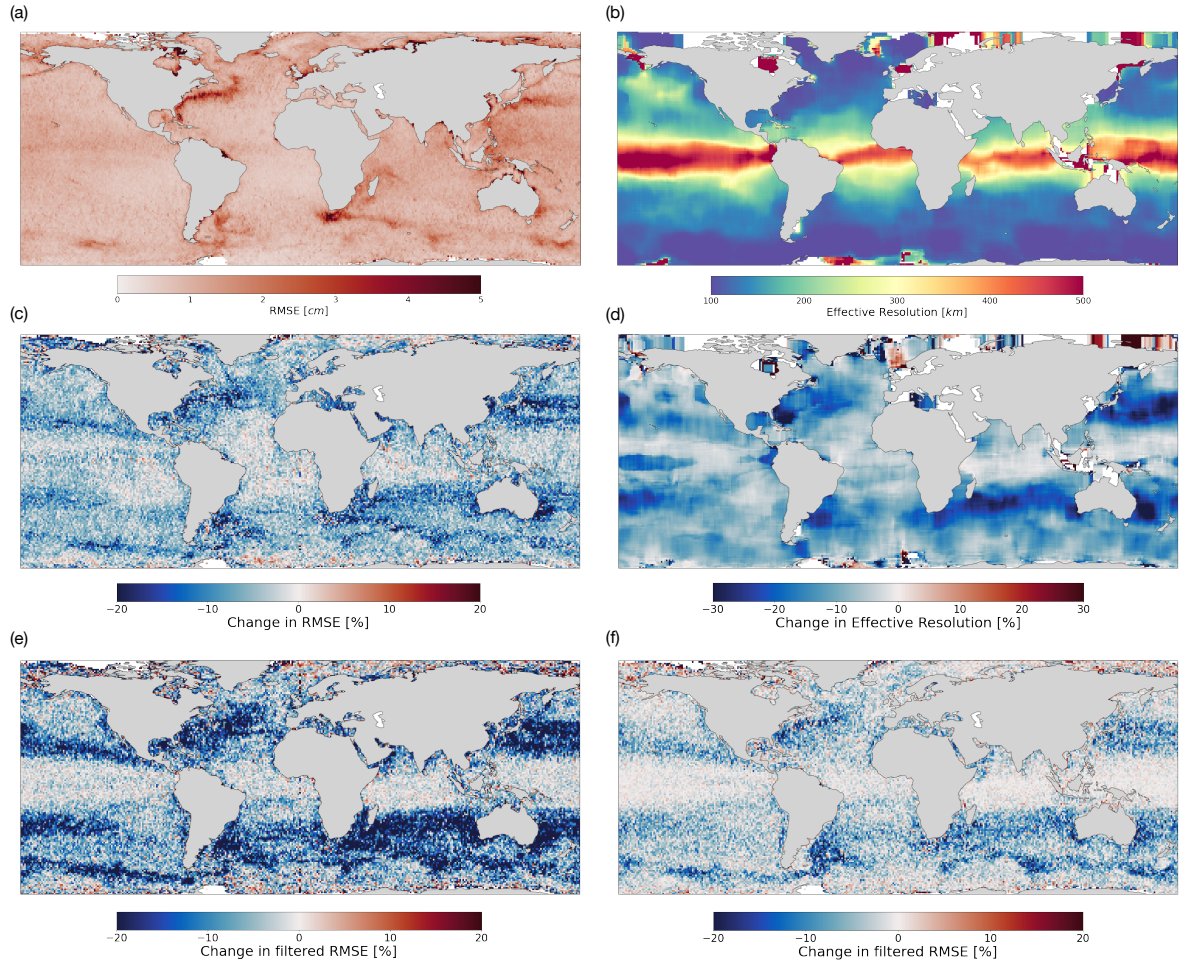


Figure 2.2: (a) RMSE of NeurOST SSH-SST compared to withheld altimeter. (b) Smallest resolved wavelengths (effective resolution) of NeurOST SSH-SST. (c) Change in RMSE of NeurOST SSH-SST compared to DUACS. (d) Change in effective resolution of NeurOST SSH-SST compared to DUACS. (e) Change in RMSE of small-scale (70-250km) signals of NeurOST SSH-SST compared to DUACS. (f) Change in RMSE of small-scale signals of NeurOST SSH-SST compared to NeurOST SSH. Blue colors indicate a relative decrease in error in panels c-f.

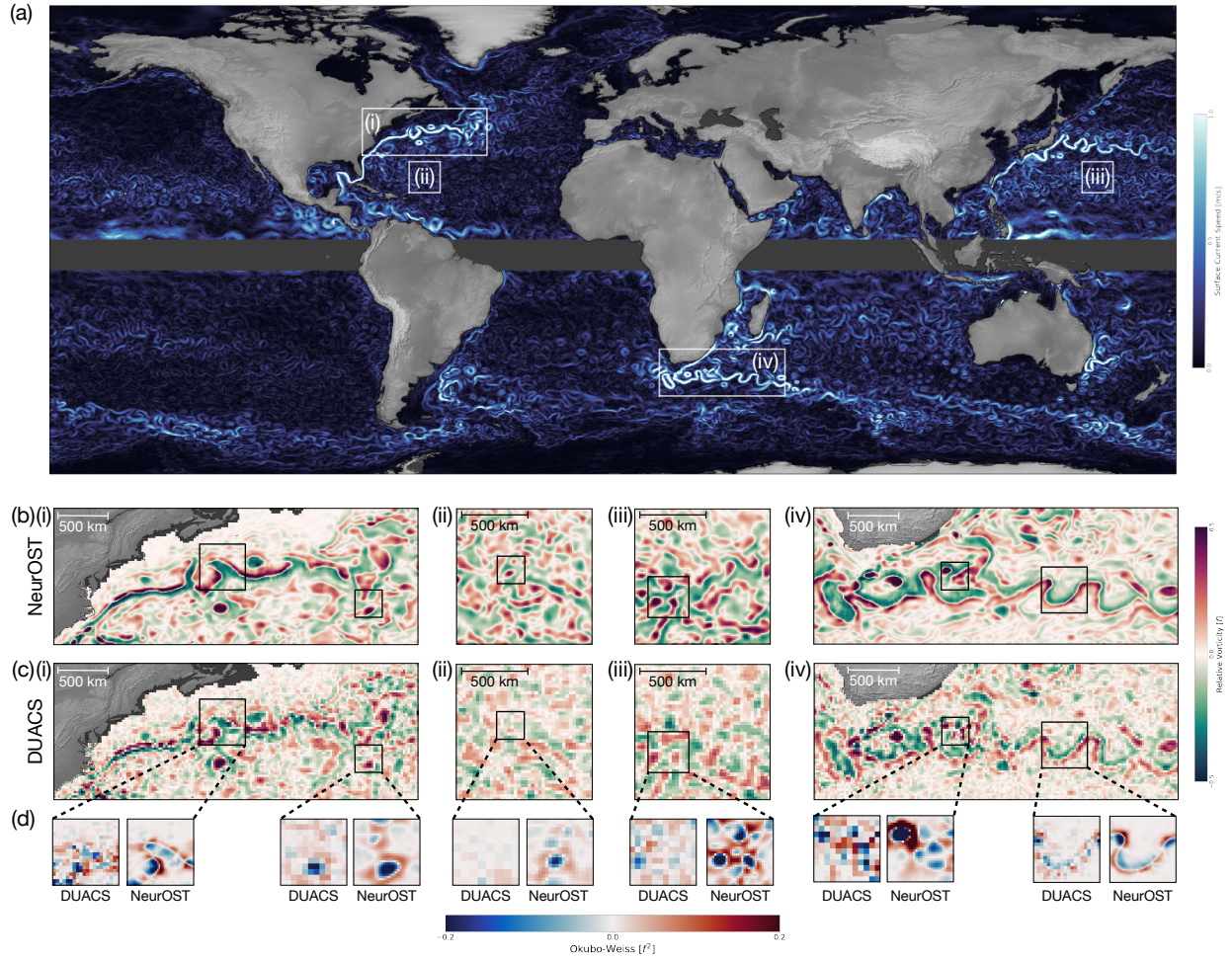


Figure 2.3: (a) Surface geostrophic current speed on March 1st 2019 derived from SSH maps made using NeurOST SSH-SST. (b) Relative vorticity from NeurOST SSH-SST. (c) Relative vorticity from DUACS. (d) Zoomed insets of Okubo-Weiss quantity for DUACS and NeurOST SSH-SST.

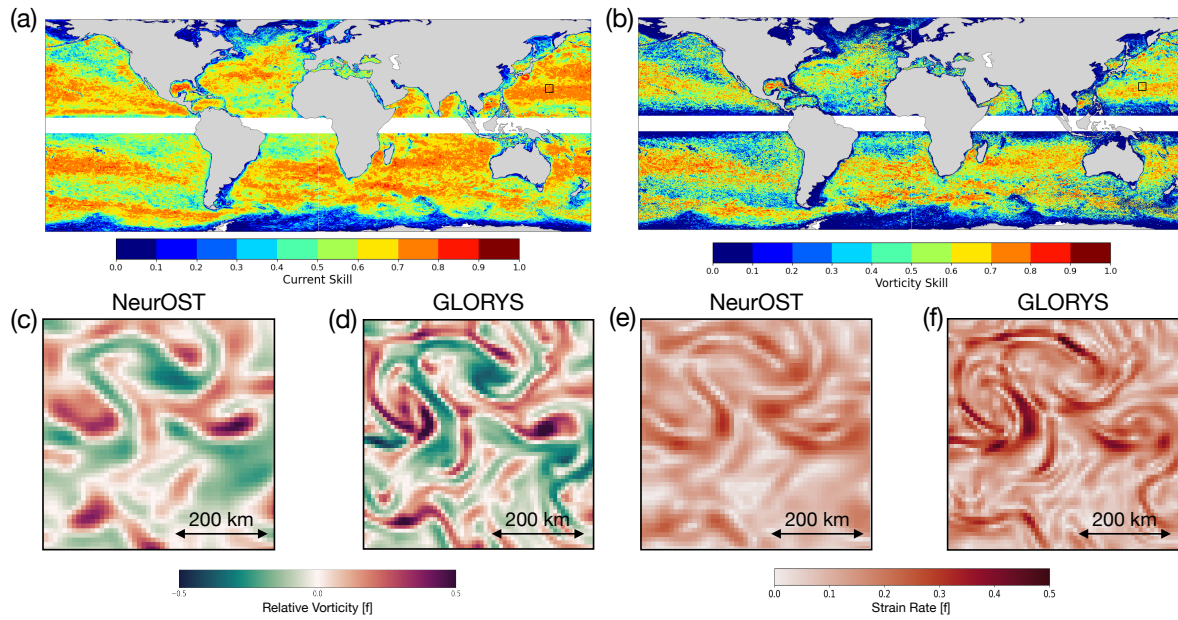


Figure 2.4: NeurOST SSH-SST reconstruction skill (explained variance) in GLORYS OSSE for (a) surface geostrophic currents and (b) vorticity. Snapshots of (c, d) vorticity and (e, f) strain from Subtropical North Pacific (boxed region in a and b) for NeurOST and GLORYS on Feb 26th 2019.

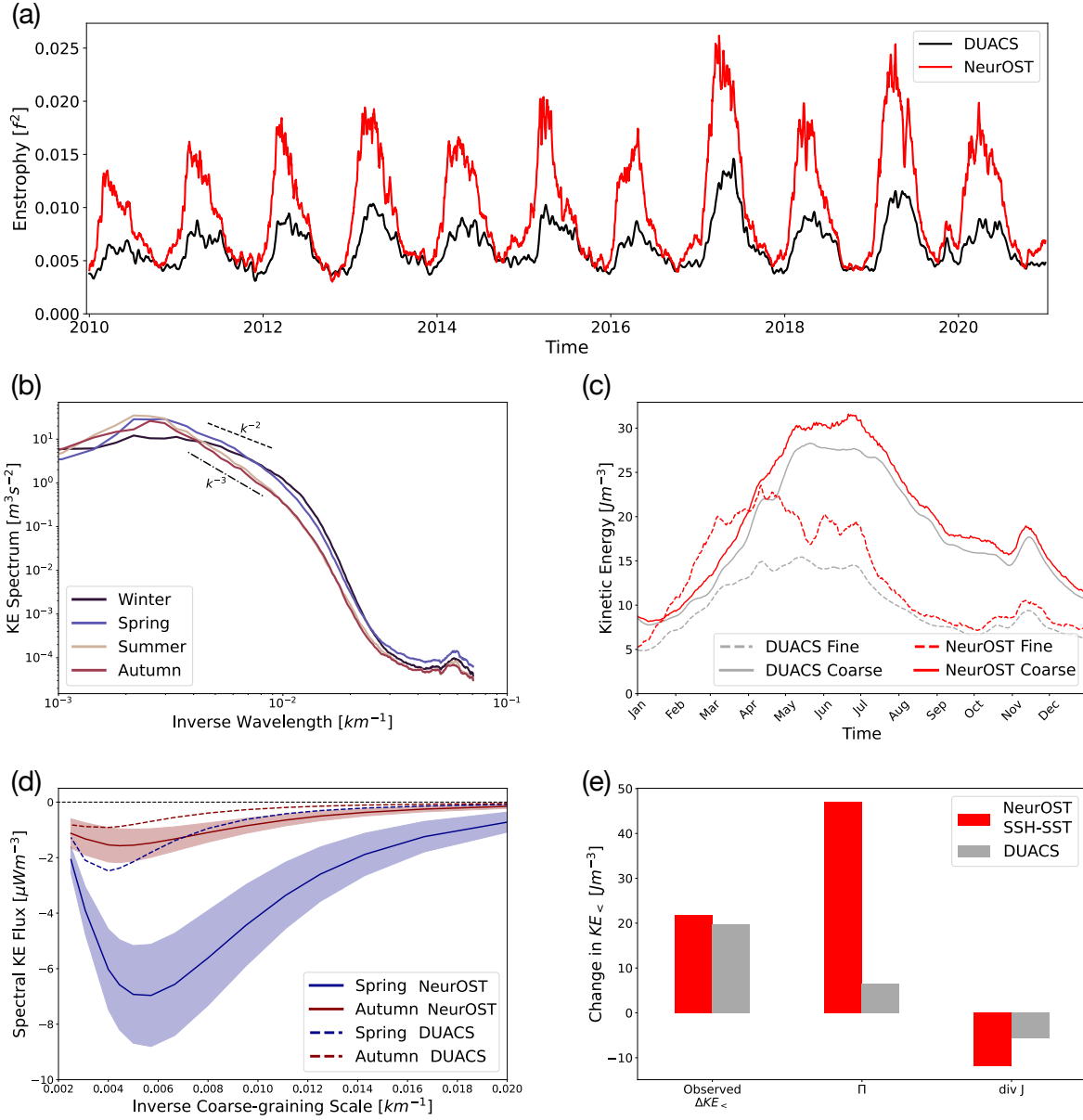


Figure 2.5: (a) Enstrophy time-series in the Subtropical North Pacific for NeurOST SSH-SST and DUACS. (b) KE spectra from NeurOST SSH-SST split by season. (c) Time-series of coarse- and fine-scale KE (above and below 125km coarse-graining scale respectively) from NeurOST SSH-SST (red) and DUACS (grey). (d) KE cascade from NeurOST SSH-SST maps for the seasons of maximum (Spring) and minimum (Autumn) upscale cascade (solid line: mean, shading: standard deviation). Dashed lines are the mean cascades from DUACS. (e) Change in coarse-scale KE ($KE_{<}$) from the winter-time minimum to the summer-time maximum compared to the diagnosed contributions of the KE cascade ($-\int \Pi dt$), and the spatial transport of coarse-scale KE ($-\int \nabla \cdot J dt$) from NeurOST SSH-SST (red) and DUACS (grey) (S.I. S1.12).

Appendix A

Supporting Information for Chapter 2

A1 Extended Methods

A1.1 Sea surface height and surface geostrophic currents

At large temporal and spatial scales, ocean currents are approximately in geostrophic balance [Vallis, 2017], meaning that currents arrange themselves such that the horizontal pressure gradient force is balanced by the Coriolis force. Surface pressure in the ocean can be directly related to sea surface height anomaly (SSH), allowing surface current velocity to be estimated from satellite altimeter observations of SSH. The surface currents are proportional to the spatial gradients of the SSH field

$$(u_g, v_g) = \frac{g}{f} \left(-\frac{\partial \eta}{\partial y}, \frac{\partial \eta}{\partial x} \right), \quad (\text{A.1})$$

where u_g and v_g are the Eastward and Northward geostrophic surface currents respectively, g is the acceleration due to gravity, η is the SSH, f is the local Coriolis frequency, and x and y are zonal and meridional coordinates respectively. This relation breaks down near the Equator where f approaches zero, so we do not calculate surface geostrophic currents within the equatorial band (5°S to 5°N).

Geostrophy breaks down at smaller scales where the impact of non-linear advection becomes significant, and does not account for wind-induced Ekman currents. Empirical corrections can be made to geostrophy at the Equator [Lagerloef et al., 1999a], to estimate Ekman currents [Rio et al., 2014], and to include the effect of non-linear advection through cyclo-geostrophic balance [Penven et al., 2014; Cao et al., 2023]. Ekman

currents and equatorial dynamics are not the focus of this study, but since cyclo-geostrophy is pertinent for mesoscale eddies we do assess the sensitivity of our results to correcting for this component using the iterative method proposed in Penven et al. [2014]. The cyclo-geostrophic correction leads to a marginally improved surface current RMSE when evaluated using drifters compared to conventional geostrophy but the change is typically well below 10%, highlighting the relatively small impact of this correction at the scales resolved here (S.I. Figure A.9). Further, we find that the KE cascade diagnosed from both NeurOST and GLORYS changes very little when the cyclo-geostrophic correction is applied (S.I. Figure A.7). The near-perfect agreement between the GLORYS cyclo-geostrophic KE cascade and that from the GLORYS 15m currents highlights the pertinence of this empirical correction, but the small correction to the currents highlights that they are to leading order geostrophic.

A1.2 Satellite datasets

The along-track SSH observations used in this study are those processed by the Data Unification and Altimeter Combination System (DUACS) and distributed by the Copernicus Marine Environmental Service (CMEMS) [E.U. Copernicus Marine Service Information (CMEMS), 2024c,b]. Specifically, we use the unfiltered, Level 3 sea level anomaly observations. At Level 3, the observations have been corrected for atmospheric effects, the barotropic tide has been removed, and the data has been adjusted to ensure consistency between the different altimeter missions.

We use the Multi-scale Ultra-high Resolution (MUR) SST analysis product to provide a gridded estimate of SST as an additional predictor variable in the mapping of SSH that combines observations from a wide range of satellite infrared and microwave radiometer observations [JPL MUR MEaSUREs Project, 2015; Chin et al., 2017]. While this product is distributed on a 1/100th degree grid, the spatial scales resolved vary in space and time due to satellite sampling and cloud cover.

A1.3 Formulating SSH interpolation as a self-supervised deep learning problem

After appropriate data pre-processing, SSH interpolation can be viewed as a video in-painting problem with an extremely high missing pixel rate ($\sim 90\%$) [Manucharyan et al., 2021; Martin et al., 2023; Fablet et al., 2021a]. We first extract satellite altimeter SSH observations in some restricted spatiotemporal subdomain

within which we seek to estimate the full SSH field. This subdomain is discretized into a regular grid in space and time onto which the observations are bin-averaged; empty voxels are padded with zeroes. This data can now be considered as a heavily-masked video, and our objective is to predict the corresponding full, unmasked video using a deep learning neural network. The objective minimized during training is the mean squared error (MSE) between the prediction and the ground-truth. When training on real-world altimetry observations, there is no full unmasked ground-truth dataset to use during training. We overcome this by randomly withholding some of the altimetry observations from the input and calculating the MSE only at the locations of these withheld observations [Martin et al., 2023]. Alternative loss functions and regularization terms (e.g. the along-track derivative regularizations used in Martin et al. [2023] and explored further in Archambault et al. [2024b] or along-track spectral regularizations) may yield further mapping improvements in future studies but we restrict ourselves to MSE here due to the computational expense of iterating on loss function choice for global mapping. Co-located estimates of gridded SST are used as an extra predictor variable by averaging them onto the same local grid and presenting this (unmasked) video as an additional input to the neural network. The dimensions of the local grid on which we map SSH were chosen to be 128x128 in the spatial domain with 7.5km grid resolution (latitude, longitude coordinates are first projected onto a local orthonormal grid to avoid distortion), and 30 frames in the temporal domain with 1 day grid resolution. Rationalization and validation of these choices is given in our previous study [Martin et al., 2023]. To train our network, we generated a training dataset of 1 million local subdomains centred on random points in space and time throughout the Global Ocean. The data are split in the temporal domain to ensure well-separated training, validation, and testing datasets, with 2019 being withheld for testing, and the remaining years from 2010-2022 split into non-overlapping training and validation periods (S.I. Figure A.8).

The mapping errors grow away from the centre of the local spatiotemporal grid due the omission of observations outside the local subdomain, therefore to produce the optimal reconstruction we use only the middle day of the predicted time-series during inference and points close to the edge of the subdomain are given low weight in our algorithm for merging subdomain reconstructions to produce a global SSH estimate (S.I. Section A1.5).

A1.4 Deep learning neural network architecture

After formulating SSH interpolation as discussed above, we are free to use any sequence-to-sequence video prediction model from the extensive computer vision literature. To ensure we employed a state-of-the-art architecture, we chose the top-performing architecture on the Moving MNIST video prediction benchmark [Srivastava et al., 2015] at the time of our study, SimVP [Gao et al., 2022; Tan et al., 2022]. SimVP is built up of three modules: a spatial encoder that learns to encode each frame of the input independently in some lower-dimensional latent space, a temporal translator that learns both spatial and temporal dependencies from the latent space, and a spatial decoder that decodes the latent space into the predicted video frames. Unlike widely-used recurrent architectures, such as ConvLSTM [Shi et al., 2015], SimVP uses convolutional neural networks (CNNs) for all three modules. Our architecture is as described in Tan et al. [Tan et al., 2022], where the temporal translator module is a gated spatio-temporal attention translator, which uses large convolutional kernels to imitate the attention mechanism allowing the translator to adaptively select informative features from the latent space. Compared with Tan et al., we removed the skip connection from the first layer of the spatial encoder to the final layer of the spatial decoder since the extreme sparsity of the input SSH frames led to the appearance of artifacts in the output coinciding with the input altimeter tracks. To synthesize SSH and SST, we use a separate spatial encoder for each variable before concatenating the encoded SSH and SST in the channel dimension and passing this to the temporal translator. Except for the temporal translator, the architecture used here is similar to the ConvLSTM-based architecture used in our earlier regional SSH mapping study [Martin et al., 2023]. During early testing we found SimVP to outperform ConvLSTM in global SSH reconstruction, which requires a more expressive architecture due to the diverse range of dynamical regimes, and its performance (when trained on global data) is comparable to our previously published values for ConvLSTM in the Gulf Stream despite the latter being trained exclusively on this region (S.I. Table A.2). Each network was trained for 50 epochs using the OneCycle learning rate scheduling policy, the Adam optimizer, and with drop-out and drop-path probabilities of 0.2 and 0.15 respectively which were selected after performing hyper-parameter optimization on 10% of the training data. Each training on the global training dataset took 7 days on a single node with four Nvidia V100 GPUs.

A1.5 Merging subdomain reconstructions to create global SSH product

Our neural network predicts gridded SSH on subdomains of size 960x960km. To produce a global gridded SSH estimate we use the trained network to predict SSH on 5615 subdomains with centres chosen to be approximately equally spaced by a distance of 250km throughout the Global Ocean. There is therefore substantial overlap between neighbouring subdomains. To merge the subdomain reconstructions into a single global SSH estimate we use the kernel-weighted averaging method described in Appendix A of Callaham et al. [Callaham et al., 2019] and outlined below.

The global 2D SSH estimate, $\hat{\mathbf{x}}$, defined on a regular 1/10th degree grid, is computed from the k subdomain estimates through

$$\hat{\mathbf{x}} = \sum_{i=1}^k \Phi_i \odot \hat{\mathbf{x}}_i, \quad (\text{A.2})$$

where Φ_i is a normalized weighting kernel for each subdomain, $\hat{\mathbf{x}}_i$ are the subdomain SSH estimates, and $\odot$ denotes the Hadamard (i.e. element-wise) product between two matrices. Note that each $\hat{\mathbf{x}}_i$ and Φ_i are filled with zeroes at all points covered by land or sea ice and at points lying outside the subdomain. All matrices were first regridded to the regular 1/10th degree grid from the original, irregular subdomain grid using bi-linear interpolation. Each weighting kernel is taken to be a Gaussian centered on the the corresponding subdomain

$$\Phi_i(\mathbf{r}) = \frac{1}{N(\mathbf{r})} \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_i|^2}{L^2}\right), \quad (\text{A.3})$$

where $\mathbf{r}$ is the position of the point being estimated, $\mathbf{r}_i$ is the position of the subdomain center, L is the characteristic width of the Gaussian kernel, and $N(\mathbf{r})$ is a normalization factor chosen such that

$$N(\mathbf{r}) = \sum_{i=1}^k \Phi_i(\mathbf{r}). \quad (\text{A.4})$$

The mapping errors in the subdomain reconstructions are expected to increase away from the center of the subdomain due to the omission of observations outside the subdomain in the mapping. Thus, for minimizing the error of the global estimate maximizing the number of subdomains is desirable. The choice to space the subdomains by 250km was made as the minimum spacing our computing resources would reasonably permit (merging the subdomain reconstructions for a single day takes ~ 3 minutes per CPU worker at

this spacing). Given this subdomain spacing, the value of L was tuned so as to minimize the mapping error for the global estimate. We found the errors to be only weakly dependent on kernel width for widths within reasonable bounds, the results presented in the manuscript were obtained using $L = 250\text{km}$. All first- and second-order spatial derivatives of the SSH field were computed first on the orthonormal subdomain grid using smooth noise-robust differentiator kernels (discussed in Arbic et al. [2012]) before being merged using the above algorithm to avoid the appearance of high-frequency numerical artifacts [Martin et al., 2023] (a similar result can be obtained by a simple low-pass filtering of the mapped SSH).

A1.6 Regional fine-tuning experiment

In S.I. Table A.2 we show that NeurOST SSH-SST trained on global observations can be fine-tuned for regional applications to bring its performance closer to that of state-of-the-art regional schemes [Febvre et al., 2024b; Martin et al., 2023]. The global model was trained on 1 million training examples drawn randomly from the Global Ocean. During fine-tuning we started training from the converged global model using the Adam optimizer with a fixed learning rate parameter of 10^{-4} on a smaller training set of 100,000 examples drawn randomly from the Gulf Stream (as in our previous study [Martin et al., 2023]) and continued training until the validation loss stopped improving. This fine-tuning took 12 hours on a single node with four Nvidia V100 GPUs. The results presented in the rest of the study use only the global trained model to limit the computational resources of the method, but in future an ensemble of bespoke regional models could be fine-tuned to further optimize the SSH mapping in each region if these maps were produced operationally by a data centre. Since previous studies used a different test year, 2017, in the Gulf Stream [Ballarotta et al., 2021], we swapped 2017 and 2019 in our training-validation-testing split (S.I. Figure A.8) to ensure 2017 was withheld during training and cross-validation (both the global and regional training were done with this updated split).

A1.7 SSH map evaluation and inter-comparison OSE

To evaluate the accuracy and resolution of the SSH signals resolved by different mapping methods, we employ an ‘observing system experiment’ (OSE) in which each method is used to generate global gridded SSH estimates using all but one of the available satellite altimeters which is then used as an independent

validation of the mapped signal. Since the existing operational SSH products are only distributed using all available altimeters, performing an OSE would typically involve re-implementation of all existing methods, which would be challenging in the case of SSH mapping since the covariance parameters used to create the community-standard DUACS product are not publicly-available. To address this, in recent years a series of ‘Ocean Data Challenges’ have been developed [Metref et al., 2023]. In each challenge a common mapping OSE problem is defined, developers of different methods implement their method and post their results, allowing a transparent performance benchmark.

To evaluate our global product, we use the recently-created global OSE challenge: ‘2023a_SSH_mapping_OSE’ [Metref and Ballarotta, 2023]. In this challenge, each method is used to create 1 year (2019) of global gridded SSH estimates using SSH observations from the satellites Jason 3, Sentinel 3A, Sentinel 3B, Haiyang-2A, Haiyang-2B, and Cryosat-2, while observations from the satellite Saral/Altika are withheld for validation. While the validation observations only sample the maps along 1D tracks, aggregating over a full year allows robust, geographically-varying error statistics to be found.

We present three SSH error metrics using the withheld altimeter, each averaged over the full year and binned into 1° bins: the root-mean-square error (RMSE) between the mapped and observed signals, the RMSE after applying a 70-250km along-track band-pass spatial filter to both the observed and mapped signals along the satellite tracks to highlight the maps’ ability to map small mesoscale eddies, and the effective spatial resolution of the mapped signal. The effective resolution is found by taking along-track segments of the withheld altimeter observations along with the mapped values at these locations and calculating the signal-to-noise ratio as a function of wavelength by dividing the power spectral density of the mapping errors by that of the observations. Concretely, the ‘effective spatial resolution’ is taken to be the wavelength at which the signal-to-noise ratio between the observed and mapped signals drops below 0.5 [Ballarotta et al., 2019].

To evaluate the surface currents inferred from each SSH map, surface drifter observations from the CMEMS global in-situ water velocity product [Copernicus Marine in situ TAC, 2024] are used. Surface drifter observations are not used in the generation of any of the surface current maps presented here, so they are an independent validation dataset. We present geographically averaged velocity RMSE values for each method.

At the time of writing, three other SSH mapping methods have been implemented world-wide and all are available in the data challenge for evaluation: DUACS [Le Traon et al., 1998; Taburet et al., 2019], MIOST (geostrophic) [Ubelmann et al., 2021], and MIOST (geostrophic + equatorial waves) [Ballarotta et al., 2023]. DUACS is the community-standard gridded SSH product that is distributed operationally by CMEMS. The DUACS system uses a linear optimal interpolation (OI) formulation [Bretherton et al., 1976], in which an *a priori* model is prescribed for how SSH covaries in space and in time, then the missing values are estimated using the best linear least-squares estimator. The assumed covariance, C , is

$$C(r, t) = \left(1 + ar + \frac{1}{6}(ar)^2 - \frac{1}{6}(ar)^3\right) \exp(-ar) \exp\left(\frac{-t^2}{T^2}\right), \quad (\text{A.5})$$

where t is the temporal separation of the observation and mapped point under consideration, T is a prescribed de-correlation time-scale, $a = 3.337$, and

$$r = \sqrt{\left(\frac{dx - C_{px}t}{L_x}\right)^2 + \left(\frac{dy - C_{py}t}{L_y}\right)^2}, \quad (\text{A.6})$$

where L_x and L_y are prescribed de-correlation length-scales in the zonal and meridional directions, dx and dy are respectively the zonal and meridional separation of the observation and mapped point under consideration, and C_{px} and C_{py} are prescribed propagation velocities. The de-correlation scales and propagation velocities are allowed to vary with geographical location and the values used are not publicly available but have been tuned over many years to best map mesoscale ocean features globally. The MIOST mapping method extends the linear mapping framework, using a wavelet decomposition to allow the construction of multiple independent components of the assumed covariance model [Ubelmann et al., 2022; Ballarotta et al., 2023]. MIOST (geostrophic) uses a single component in the covariance model intended to represent the geostrophically balanced component of SSH evolution, while MIOST (geostrophy + equatorial waves) adds an additional component to model the propagation of tropical instability waves and Poincare waves near the equator.

A1.8 Season definitions

Wherever results are split by season in this study we define those seasons in the Northern (Southern) Hemisphere: winter is January-March (July-September), spring is April-June (October-December), summer is July-September (January-March), and autumn is October-December (April-June).

A1.9 Eddy kinetic energy

The kinetic energy, KE, per unit volume of the surface currents is calculated from the surface geostrophic current maps

$$\text{KE} = \frac{\rho_0}{2} (u_g^2 + v_g^2), \quad (\text{A.7})$$

where ρ_0 is a reference density taken to be 1025 kg m^{-3} . The eddy kinetic energy, EKE, is defined as the time-varying component of the KE

$$\text{EKE} = \text{KE} - \overline{\text{KE}}, \quad (\text{A.8})$$

where $\overline{\text{KE}}$ is the time mean of the KE.

To highlight the difference in small-scale EKE between the maps we also calculate the EKE of the surface currents after the application of a 250km high-pass filter.

A1.10 Relative vorticity, strain rate, and Okubo-Weiss quantity

While first order spatial derivatives of SSH give the velocity of the geostrophic currents, second order spatial derivatives quantify the deformation and rotation induced by the flow. The relative vorticity, ω , describes the local rotation of the fluid (that is, how a patch of tracers would tend to rotate if placed at a point within the flow)

$$\omega = \frac{\partial v_g}{\partial x} - \frac{\partial u_g}{\partial y} = \frac{g}{f} \nabla_h^2 \eta, \quad (\text{A.9})$$

where ∇_h^2 is the horizontal Laplacian.

Meanwhile, the strain rate, s , defines the deformation of fluid elements by the flow (that is, how a patch of tracers would change shape due to the flow)

$$s = \sqrt{s_n^2 + s_s^2} \quad (\text{A.10})$$

where s_n is the normal component of the strain

$$s_n = \frac{\partial u_g}{\partial x} - \frac{\partial v_g}{\partial y} = -2 \frac{g}{f} \frac{\partial^2 \eta}{\partial x \partial y}, \quad (\text{A.11})$$

and s_s is the shear component

$$s_s = \frac{\partial v_g}{\partial x} + \frac{\partial u_g}{\partial y} = \frac{g}{f} \left(\frac{\partial^2 \eta}{\partial x^2} - \frac{\partial^2 \eta}{\partial y^2} \right). \quad (\text{A.12})$$

A high strain rate is associated with the stretching of patches of fluid, is common in the areas between eddies, and is associated with the generation of submesoscale filaments through frontogenesis [Hoskins, 1982] and strong transfer of kinetic energy between scales [Aluie et al., 2018]. Whereas strong relative vorticity (either positive or negative) is associated with coherent, persistent eddies and is typical in the cores of eddies.

The relative importance of relative vorticity and strain rate at each point in the fluid can be described using the Okubo-Weiss quantity [Okubo, 1970; Weiss, 1991],

$$W = s^2 - \omega^2, \quad (\text{A.13})$$

which is positive when strain dominates and negative when relative vorticity dominates (Figure 3 in main text).

Note, DUACS vorticity/strain fields appear ‘grainy’ in comparison to NeurOST (e.g. in Fig. 3) due to DUACS having a coarser grid resolution. One could first linearly interpolate the DUACS data to a higher resolution grid before estimating vorticity and strain to reduce this, but we have verified that this has negligible impact on the accuracy and spectral characteristics of the vorticity (Fig. S11).

A1.11 Eddy Dynamics Evaluation OSSE

As described in the main text, we use the GLORYS reanalysis product to implement an OSSE to evaluate the realism of eddy dynamics by generating synthetic altimetry observations from the reanalysis and inputting them to NeurOST. To ensure consistency with our OSE, we use 2019 as the test year and sam-

ple the GLORYS SSH along the locations of the 2019 altimeter tracks, applying regionally-varying white noise consistent with that described in the user guide for the along-track SSH observations [E.U. Copernicus Marine Service Information (CMEMS), 2024c,b].

The metrics we use to evaluate the realism of the mapped surface currents and vorticity normalized skill scores that give the fraction of variance explained.

For surface currents this skill is defined as

$$\text{skill} = 1 - \frac{\text{MSE}(u_g) + \text{MSE}(v_g)}{\text{Var}(u_g) + \text{Var}(v_g)}, \quad (\text{A.14})$$

where MSE is mean squared error and Var is the variance. Similarly for relative vorticity, ζ_g , this is defined as

$$\text{skill} = 1 - \frac{\text{MSE}(\zeta_g)}{\text{Var}(\zeta_g)}. \quad (\text{A.15})$$

A1.12 KE cascade: diagnosing energy transfer between scales

Energy transfers between flows of different length-scales, a characteristic property of turbulent flows, can be diagnosed using a coarse-graining analysis [Aluie et al., 2018]. By applying convolutions to the Navier-Stokes equation and neglecting small contributions from molecular viscosity, a kinetic energy (KE) budget for the coarse-grained flow (i.e. the velocity after convolution with a smoothing filter) is obtained [Aluie et al., 2018]

$$\frac{\partial}{\partial t} \rho_0 \frac{|\bar{\mathbf{u}}_l|^2}{2} = -\nabla \cdot \mathbf{J}_l^{\text{transport}} - \Pi_l + \bar{\rho}_l \mathbf{g} \cdot \bar{\mathbf{u}}_l + \rho_0 \bar{\mathbf{F}}_l^{\text{forcing}} \cdot \bar{\mathbf{u}}_l, \quad (\text{A.16})$$

where $\bar{\cdot}_l$ represents convolution with a filter with scale diameter l , $\mathbf{J}_l^{\text{transport}}$ is the spatial transport of large-scale KE (as defined in Aluie et al. [2018]), ρ_0 is a reference density (here taken to be 1025kgm^{-3}), Π_l is the transfer of energy between scales by non-linear eddy interactions defined below, $\mathbf{g}$ is the acceleration due to gravity, and $\bar{\mathbf{F}}_l^{\text{forcing}}$ is any external forcing at scales above l (e.g. by winds). The third term on the right represents the conversion of potential energy into kinetic energy.

The existence of an upscale (or ‘inverse’) cascade of KE is a characteristic property of geophysical turbulence [Vallis, 2017] that is hypothesized to play a role in setting the seasonality of mesoscale ocean eddies [Sasaki et al., 2014; Qiu et al., 2014]. We therefore here diagnose the transfer of KE between scales, Π_l ,

from surface geostrophic current maps to assess its magnitude and sign at different spatial scales, implicitly neglecting energy associated with vertical velocities (which are small at the scales considered here). We don't seek to close the energy budget in Equation A.16 as this would require precise determination of the potential energy conversion and external forcing at the same resolution as the surface geostrophic current maps.

The KE cascade, Π_l , is caused by non-linear interactions between eddies and is characterized by the interplay between the large-scale strain tensor, $\bar{S}_l$, and the subfilter-scale stress, $\bar{\tau}_l$, through [Aluie et al., 2018]

$$\Pi_l = -\rho_0 \bar{S}_{ij} \bar{\tau}_{ji}, \quad (\text{A.17})$$

where

$$\bar{S}_{ij} = \frac{1}{2} (\partial_i \bar{u}_j + \partial_j \bar{u}_i), \quad (\text{A.18})$$

$$\bar{\tau}_{ij} = \overline{u_i u_j} - \bar{u}_i \bar{u}_j, \quad (\text{A.19})$$

repeated indices are summed over, and the subscript l in the coarse-graining operation has been dropped when using index notation to avoid confusion between the coarse-graining length scale and a spatial index. The cascade term, Π_l , represents the energy transfer from scales larger than l to smaller scales due to non-linear eddy interactions, so Π_l is positive (negative), energy is transferred from scales larger (smaller) than l to smaller (larger) scales representing a downscale (upscale) cascade. Scrutiny of Equation A.17 highlights the sensitivity of the KE cascade to the strain rate, and hence to eddy geometry.

We use an open-source code, FlowSieve [Storer and Aluie, 2023], to coarse-grain the surface geostrophic current maps at a range of scales, l , and diagnose Π_l and $\mathbf{J}_l^{\text{transport}}$. While this coarse-graining can be done on global surface current fields accounting for the spherical geometry of the Earth's surface [Storer and Aluie, 2023], we here restrict our attention to a selection of open ocean regions (defined in S.I. Table A.1). This prevents the need to prescribe boundary conditions at coastlines and significantly reduces the computational requirements of the analysis. All velocities are first projected onto a local ortho-normal grid with a grid spacing 10km and side length of 2560km. We perform coarse-graining on this grid and diagnose Π_l and $\mathbf{J}_l^{\text{transport}}$ as a function of l and time at each grid point, before taking a spatial average of both quantities

over a smaller box in the centre of the domain with side length 1280km. The smoothing filter used in the coarse-graining is a smoothed top-hat, as used in previous studies [Storer et al., 2022]

$$G_l(\mathbf{r}) = \frac{A}{2} \left(1 - \tanh \left[10 \left(\frac{|\mathbf{r}|}{l/2} - 1 \right) \right] \right), \quad (\text{A.20})$$

where A is a normalization calculated numerically to ensure G_l integrates to unity and $\mathbf{r}$ is the separation between the evaluation point and the center of the convolutional kernel. The coarse-grained fields, $\bar{f}_l(\mathbf{x})$, are then defined as

$$\bar{f}_l(\mathbf{x}) = G_l * f, \quad (\text{A.21})$$

where $*$ is a two-dimensional convolution.

There is no direct correspondence between the filter scales, l , used in coarse-graining and wavelengths in a Fourier analysis (e.g. the KE spectra in Figure 5 in the Main Text). To aid interpretation of the coarse-graining results in comparison to the spectral analysis we used to evaluate the effective resolution of each SSH map, we empirically calculate associated effective coarse-graining scales for each SSH map (method described below).

When assessing the ability of the KE cascade to drive the summer-time peak in large-scale KE, we compare the change in large-scale KE from its winter minimum to its summer maximum to the time integrals of $(-\Pi_l)$ and $(-\nabla \cdot \mathbf{J}_l^{\text{transport}})$ over the same time period. This analysis neglects sources/sinks of energy at larger scales, energy lost from surface currents due to eddy barotropization, and conversion of potential energy to KE but in this study we don't seek to close the large-scale KE budget, merely to demonstrate that the diagnosed Π_l is greatly changed between maps and that its strength becomes large enough to be a significant contributor to the change in large-scale KE.

During testing we also tried estimating KE cascades using the Fourier method used in Scott and Wang [2005] but found negligible differences to the results obtained using coarse-graining.

A1.13 KE Wavenumber Spectrum Calculation

To calculate the KE wavenumber spectra in Figure 5 of the main text we calculated 2D KE wave-number spectra on a local, ortho-normal grid and azimuthally averaged to collapse the meridional and zonal dimen-

sions into a single wavenumber.

A2 Effective coarse-graining scale of SSH maps

To aid interpretation of the effective resolution metric [Ballarotta et al., 2019], we also provide a corresponding ‘effective coarse-graining scale’. This metric can be interpreted as the coarse-graining scale that best represents the smoothing induced by the SSH mapping algorithm and is useful when considering our KE cascade results which were obtained using coarse-graining.

The effective coarse-graining scale is obtained by positing that the mapped signal can reasonably be approximated as a coarsened version of the true signal where a smoothing kernel has been convolved with the observations.

For a mapped along-track signal x , and an observed signal y , the effective resolution is defined [Ballarotta et al., 2019] as the wavelength where the function, f , crosses 0.5, where

$$f(k) = 1 - \frac{\widehat{(x - y)}^* \widehat{(x - y)}}{\hat{y}^* \hat{y}}, \quad (\text{A.22})$$

where $*$ represents complex conjugation, a "hat" is the Fourier transform, and k is the along-track wavenumber.

We suppose that the mapped signal, x , can be approximated as the convolution of y with a smoothing kernel, G_l , with corresponding spatial scale, l ,

$$x = G_l * y. \quad (\text{A.23})$$

Using the convolution theorem and plugging this definition of x into Equation A.22 yields an expression for f in terms of the smoothing kernel

$$f(k) = \hat{G}_l + \hat{G}_l^* - \hat{G}_l^* \hat{G}_l. \quad (\text{A.24})$$

For any given kernel, G_l , its Fourier transform, and hence $f(k)$, depends only on the coarse-graining scale, l . Thus for each map and region we fit the function in Equation A.24 to the data for $f(k)$ for different

coarse-graining kernels, G_l , to find corresponding coarse-graining scales, l . We refer to the resulting scale as the ‘effective coarse-graining scale’ of the SSH map for each kernel.

In S.I. Table A.4, we compare the effective resolution to the effective coarse-graining scales for a Gaussian kernel

$$G_l^{\text{Gaussian}}(x - x') = \frac{1}{l\sqrt{2\pi}} \exp\left(-\frac{|x - x'|^2}{2l^2}\right), \quad (\text{A.25})$$

and the smooth top-hat kernel

$$G_l^{\text{FlowSieve}}(x - x') = \frac{A}{2} \left(1 - \tanh \left[10 \left(\frac{|x - x'|}{l/2} - 1 \right) \right] \right), \quad (\text{A.26})$$

used in FlowSieve [Storer and Aluie, 2023] that we used to diagnose the energy transfer between scales, where $|x - x'|$ is the distance between the analysis point and the kernel center and A is a normalization factor computed numerically. Note that for $G_l^{\text{FlowSieve}}$ the Fourier transform becomes oscillatory at high wavenumbers, we therefore set all values of the fitted $f(k)$ to zero at wavenumbers past the first zero crossing to ensure we only fit the physically meaningful part of the curve to the data.

Movie S1. Movie of relative vorticity in the Subtropical North Pacific (150-160°W, 20-30°N) from NeurOST SSH-SST, DUACS, and GLORYS.

Movie S2. Movie of relative vorticity in the Subtropical North Pacific (150-160°W, 20-30°N) from the NeurOST reconstruction of GLORYS and from GLORYS.

Region	Mapping Method	RMSE [cm]	RMSE (70-250km) [cm]	Eff. Res. [km]
Gulf Stream (295-305°E, 33-43°N)	DUACS	5.66	2.44	126
	MIOST (geos)	5.61	2.41	120
	NeurOST SSH	5.33	2.29	114
	NeurOST SSH-SST	5.01	2.19	107
	GLORYS	12.47	5.25	236.5
N Atlantic (322-338°E, 39-51°N)	DUACS	4.15	1.33	133
	MIOST (geos + waves)	3.86	1.25	129
	NeurOST SSH	3.85	1.19	121
	NeurOST SSH-SST	3.78	1.12	116
	GLORYS	6.45	2.31	201
Mediterranean (2-19°E, 31-44°N)	DUACS	4.24	1.02	150
	MIOST (geos + waves)	3.97	0.97	128
	NeurOST SSH	3.95	0.98	112
	NeurOST SSH-SST	3.90	0.94	112
	GLORYS	6.25	1.94	220
Kuroshio (153-167°E, 29-41°N)	DUACS	5.03	2.00	163
	MIOST (geos + waves)	4.68	1.80	150
	NeurOST SSH	4.52	1.69	138
	NeurOST SSH-SST	4.34	1.57	128
	GLORYS	8.50	3.14	219
Subtropical N Pacific (149-161°E, 19-31°N)	DUACS	3.79	1.47	197
	MIOST (geos + waves)	3.70	1.39	185
	NeurOST SSH	3.51	1.29	171
	NeurOST SSH-SST	3.42	1.22	156
	GLORYS	5.62	1.89	246
Equatorial Pacific (175-250°E, 10°S-10°N)	DUACS	3.30	1.11	490
	MIOST (geos + waves)	3.21	1.10	445
	NeurOST SSH	3.21	1.10	456
	NeurOST SSH-SST	3.19	1.09	448
	GLORYS	4.17	1.12	499
Brazil-Malvinas (305-325°E, 40-55°S)	DUACS	4.55	1.74	121
	MIOST (geos)	4.45	1.74	118
	NeurOST SSH	4.38	1.67	113
	NeurOST SSH-SST	4.09	1.47	103
	GLORYS	11.28	4.16	248
Agulhas (10-30°E, 35-45°S)	DUACS	6.09	2.55	144
	MIOST (geos + waves)	6.05	2.52	141
	NeurOST SSH	5.78	2.35	127
	NeurOST SSH-SST	5.54	2.19	118
	GLORYS	12.01	4.97	254
Subtropical S Pacific (194-206°E, 19-31°S)	DUACS	3.25	1.11	186
	MIOST (geos + waves)	3.10	1.06	174
	NeurOST SSH	3.05	0.99	162
	NeurOST SSH-SST	2.97	0.93	148
	GLORYS	4.87	1.40	224
Drake Passage (290-300°E, 55-65°S)	DUACS	4.49	1.60	115
	MIOST (geos + waves)	4.44	1.62	110
	NeurOST SSH	4.27	1.53	108
	NeurOST SSH-SST	4.14	1.44	97
	GLORYS	8.68	3.46	230

Table A.1: RMSE, filtered RMSE (signals between 70-250km), and effective resolution of the existing global SSH products compared to our deep learning method with and without SST in a selection of regions [Metref and Ballarotta, 2023]. We show only the variant of the MIOST method that gives the best RMSE in each region. Bold values indicate the best-performing method on each metric.

Region	Mapping Method	RMSE [cm]	RMSE (70-250km) [cm]	Eff. Res. [km]
Gulf Stream	DUACS	7.50 (5.66)	3.48 (2.44)	159 (126)
	NeurOST SSH-SST	5.58 (5.01)	2.39 (2.19)	115 (107)
N Atlantic	DUACS	4.64 (4.15)	1.64 (1.33)	153 (133)
	NeurOST SSH-SST	3.92 (3.78)	1.23 (1.12)	122 (116)
Mediterranean	DUACS	4.49 (4.24)	1.27 (1.02)	250 (150)
	NeurOST SSH-SST	4.01 (3.90)	1.06 (0.94)	144 (112)
Kuroshio	DUACS	6.16 (5.03)	2.65 (2.00)	197 (163)
	NeurOST SSH-SST	4.65 (4.34)	1.77 (1.57)	139 (128)
Subtropical N Pacific	DUACS	4.21 (3.79)	1.74 (1.47)	227 (197)
	NeurOST SSH-SST	3.54 (3.42)	1.31 (1.22)	169 (156)
Equatorial Pacific	DUACS	3.40 (3.30)	1.14 (1.11)	566 (490)
	NeurOST SSH-SST	3.25 (3.19)	1.10 (1.09)	472 (448)
Brazil-Malvinas	DUACS	5.98 (4.55)	2.55 (1.74)	152 (121)
	NeurOST SSH-SST	4.50 (4.09)	1.74 (1.47)	110 (103)
Agulhas	DUACS	7.85 (6.09)	3.53 (2.55)	176 (144)
	NeurOST SSH-SST	6.05 (5.54)	2.50 (2.19)	128 (118)
Subtropical S Pacific	DUACS	3.51 (3.25)	1.27 (1.11)	206 (186)
	NeurOST SSH-SST	3.03 (2.97)	0.97 (0.93)	155 (148)
Drake Passage	DUACS	4.98 (4.49)	1.96 (1.60)	131 (115)
	NeurOST SSH-SST	4.31 (4.14)	1.56 (1.44)	108 (97)

Table A.2: Evaluation of DUACS and NeurOST SSH-SST maps generated using the stable constellation of 2 satellite altimeters operational since 1993 and used in climate studies. The metrics are as in Table SA.1. Bold values indicate the best performing method on each metric. Numbers in brackets give metrics for the 6 altimeter constellation (as in Table SA.1).

Mapping Method	RMSE [cm]	Effective Resolution [km]
DUACS (Taburet et al., 2019)	7.82	152
MIOST (geos) (Ubelmann et al., 2021)	6.93	140
DYMOST (Ubelmann et al., 2015)	6.75	129
BFN-QG (Le Guillou et al., 2021)	7.69	122
4DVarNet SSH (2022) (Beauchamp et al., 2022)	6.63	110
4DVarNet SSH (2023) (Febvre et al., 2024)	6.00	100
MUSTI (Archambault et al., 2023)	6.40	115
ConvLSTM SSH (Martin et al., 2023)	6.46	114
ConvLSTM SSH-SST (Martin et al., 2023)	6.00	100
NeurOST SSH-SST (global)	6.18	114
NeurOST SSH-SST (fine-tuned)	6.04	108

Table A.3: RMSE and effective resolution of published SSH mapping methods in the Gulf Stream [Balarotta et al., 2021]. Bold values indicate the best-performing method on each metric. Note that the results for ConvLSTM shown here differ from those published in our previous work [Martin et al., 2023] as the maps used here were created by merging multiple patch reconstructions together (as described in Text S1), however, the network and weights used are the same as in our previous study.

Region	Mapping Method	Eff. Res. [km]	Eff. Scale (G_l^{Gaussian}) [km]	Eff. Scale ($G_l^{\text{FlowSieve}}$) [km]
Gulf Stream	DUACS	126	31.3	90.5
	MIOST (geos)	120	29.5	84.6
	NeurOST SSH	114	28.1	79.5
	NeurOST SSH-SST	107	26.5	76.7
N Atlantic	DUACS	133	33.4	95.7
	MIOST (geos + waves)	129	31.8	92.2
	NeurOST SSH	121	30.1	86.1
	NeurOST SSH-SST	116	28.5	83.0
Mediterranean	DUACS	150	32.7	-
	MIOST (geos + waves)	128	28.3	-
	NeurOST SSH	112	30.9	-
	NeurOST SSH-SST	112	30.9	-
Kuroshio	DUACS	163	40.1	115.7
	MIOST (geos + waves)	150	36.6	104.9
	NeurOST SSH	138	34.3	99.8
	NeurOST SSH-SST	128	31.8	92.4
Subtropical N Pacific	DUACS	197	45.9	132.3
	MIOST (geos + waves)	185	42.8	120.5
	NeurOST SSH	171	39.4	109.4
	NeurOST SSH-SST	156	36.1	103.8
Equatorial Pacific	DUACS	490	108.1	-
	MIOST (geos + waves)	445	104.2	-
	NeurOST SSH	456	103.3	-
	NeurOST SSH-SST	448	101.6	-
Brazil-Malvinas	DUACS	121	29.8	85.1
	MIOST (geos)	118	29.3	84.3
	NeurOST SSH	113	28.0	79.6
	NeurOST SSH-SST	103	25.7	73.5
Agulhas	DUACS	144	36.4	104.3
	MIOST (geos + waves)	141	35.0	102.3
	NeurOST SSH	127	31.4	91.7
	NeurOST SSH-SST	118	29.2	84.9
Subtropical S Pacific	DUACS	186	45.8	132.5
	MIOST (geos + waves)	174	42.7	120.7
	NeurOST SSH	162	39.2	109.7
	NeurOST SSH-SST	148	35.7	103.4
Drake Passage	DUACS	115	28.8	80.8
	MIOST (geos + waves)	110	27.5	78.0
	NeurOST SSH	108	27.2	77.7
	NeurOST SSH-SST	97	25.4	72.7

Table A.4: Effective resolution and effective coarse-graining scales for both a standard Gaussian kernel, G_l^{Gaussian} , and the smooth top-hat kernel used in FlowSieve, $G_l^{\text{FlowSieve}}$, as defined in Methods. Bold values indicate best performing method on each metric. Effective coarse-graining scales for $G_l^{\text{FlowSieve}}$ are not given for the Mediterranean and the Equatorial Pacific since this kernel did not provide a good fit for the observed signal-to-noise ratio (Text S2) in these regions.

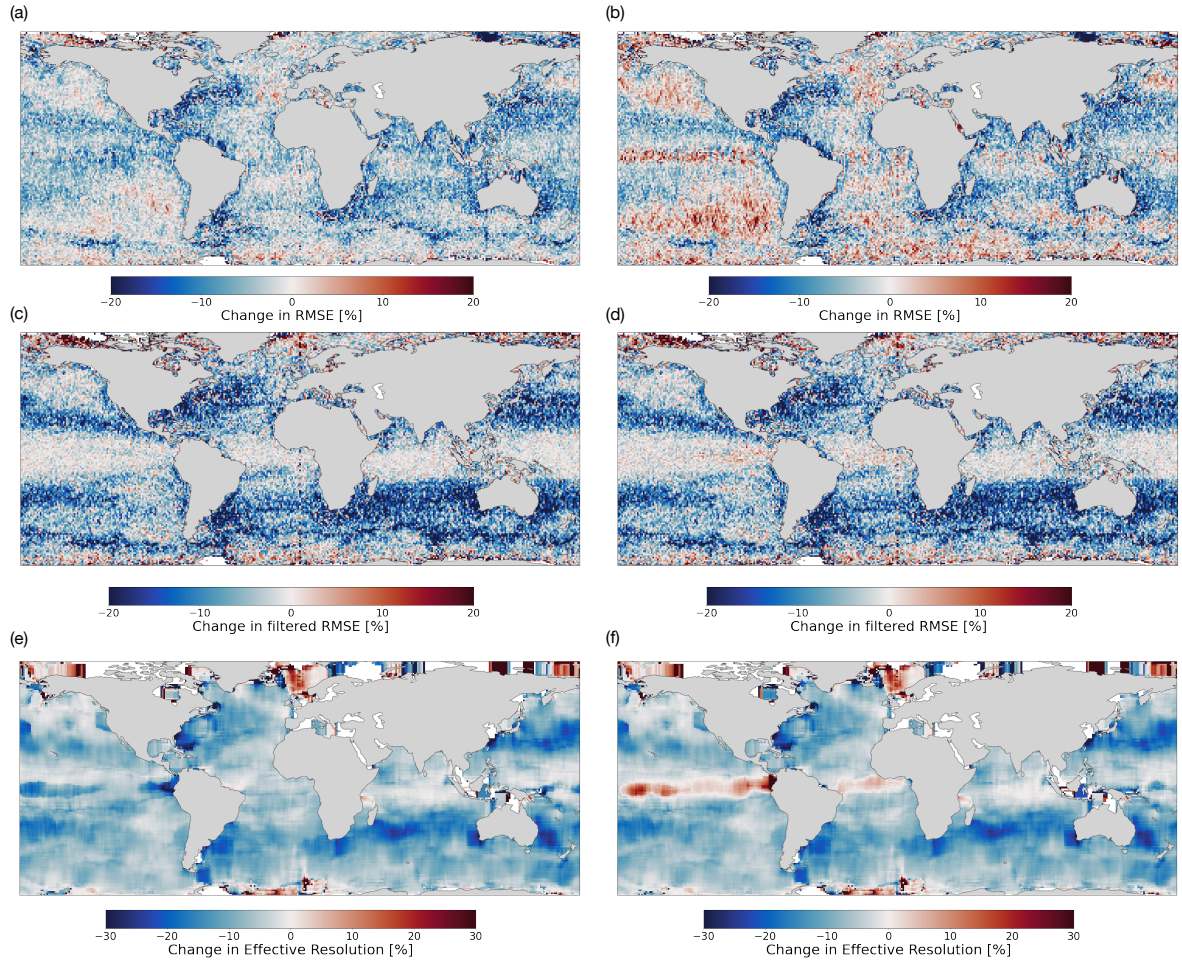


Figure A.1: (a) Change in RMSE for NeurOST SSH-SST compared to MIOST (geos.) (b) Same as (a) but for NeurOST SSH-SST compared to MIOST (geos. + waves). (c) Change in RMSE for wavelengths between 70 and 250km for NeurOST SSH-SST compared to MIOST (geos.). (d) Same as (c) but for NeurOST SSH-SST compared to MIOST (geos. + waves). (e) Change in smallest resolved wavelength for NeurOST SSH-SST compared to MIOST (geos.). (f) Same as (e) but for NeurOST SSH-SST compared to MIOST (geos. + waves).

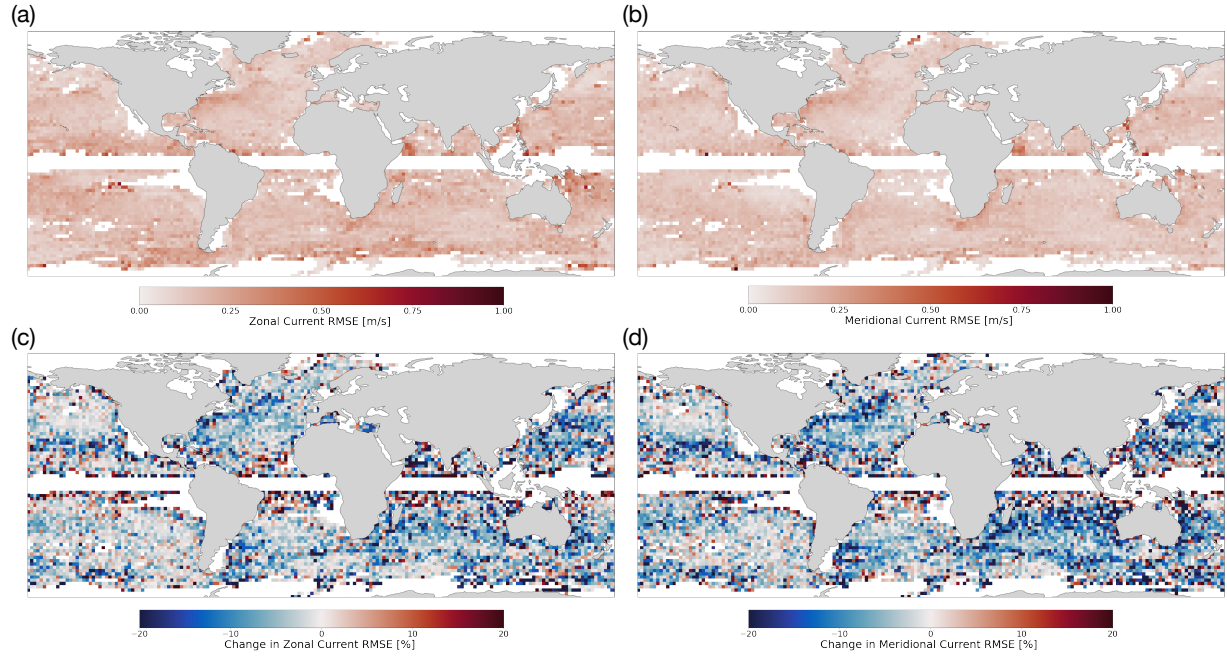


Figure A.2: (a) RMSE of NeurOST SSH-SST zonal surface geostrophic currents compared to surface drifters. (b) Same as (a) but for the meridional currents. c, Change in zonal current RMSE for NeurOST SSH-SST compared to DUACS. (d) Same as (c) but for the meridional currents.

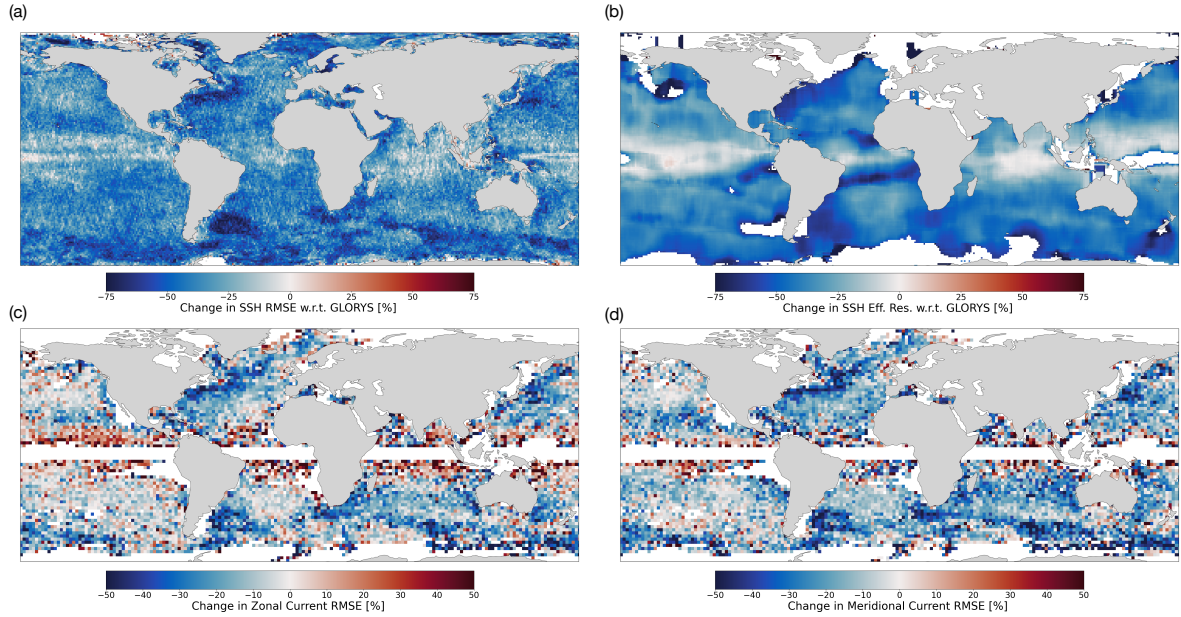


Figure A.3: (a) Change in SSH RMSE for NeurOST SSH-SST compared to GLORYS. (b) Same as (a) but for effective spatial resolution. (c) Change in RMSE of zonal surface current for NeurOST SSH-SST compared to GLORYS. (d) Same as (c) but for meridional current.

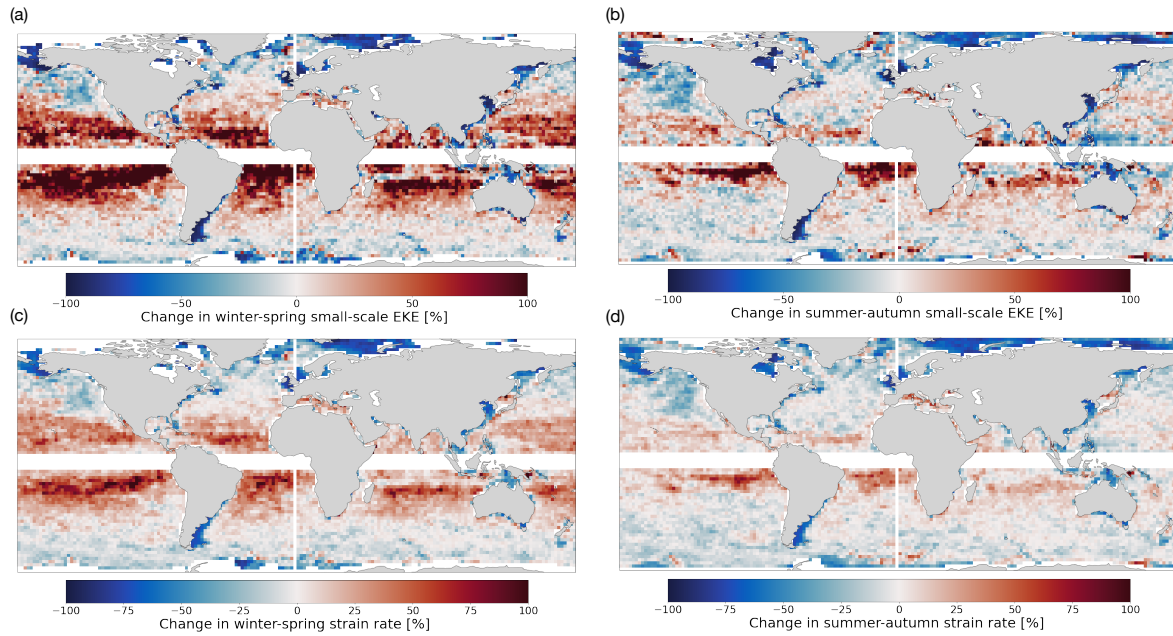


Figure A.4: Changes in diagnostics of eddy energy and dynamics from NeurOST SSH-SST currents compared to DUACS. (a) Winter-spring mean EKE for flows with wavelength below 250km. (b) Same as (a) but averaged over summer and autumn. (c) Winter-spring mean strain rate. (d) same as (c) but averaged over summer and autumn. Absolute values of these quantities are shown in S.I. Figure A.5.

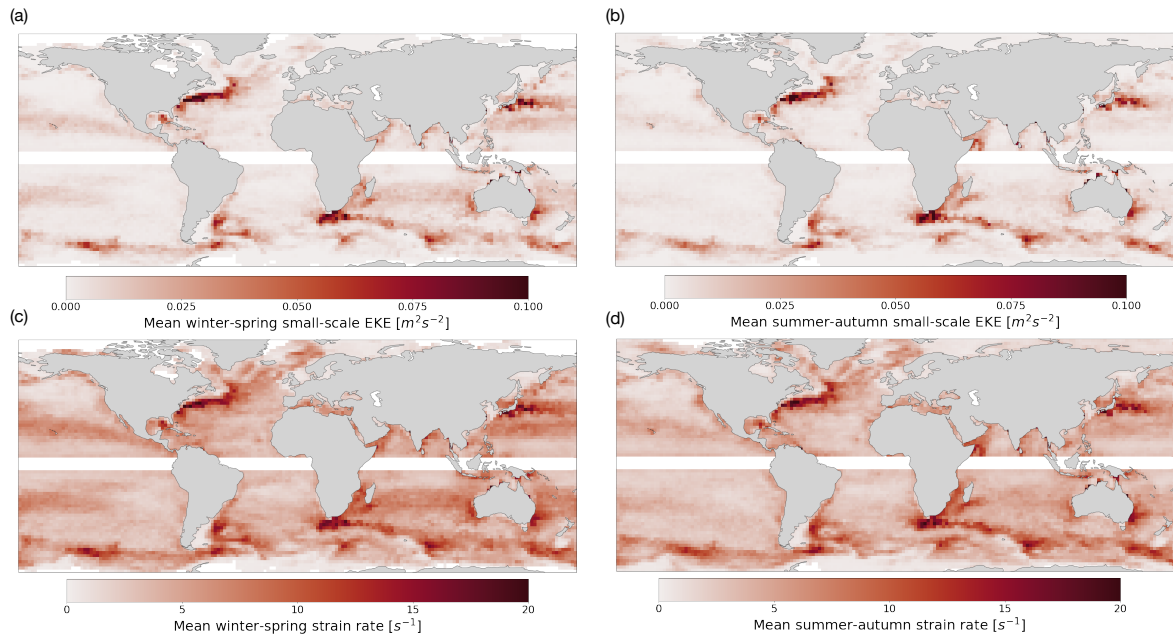


Figure A.5: Absolute values of the fields in S.I. Figure A.4 calculated from NeurOST SSH-SST maps.

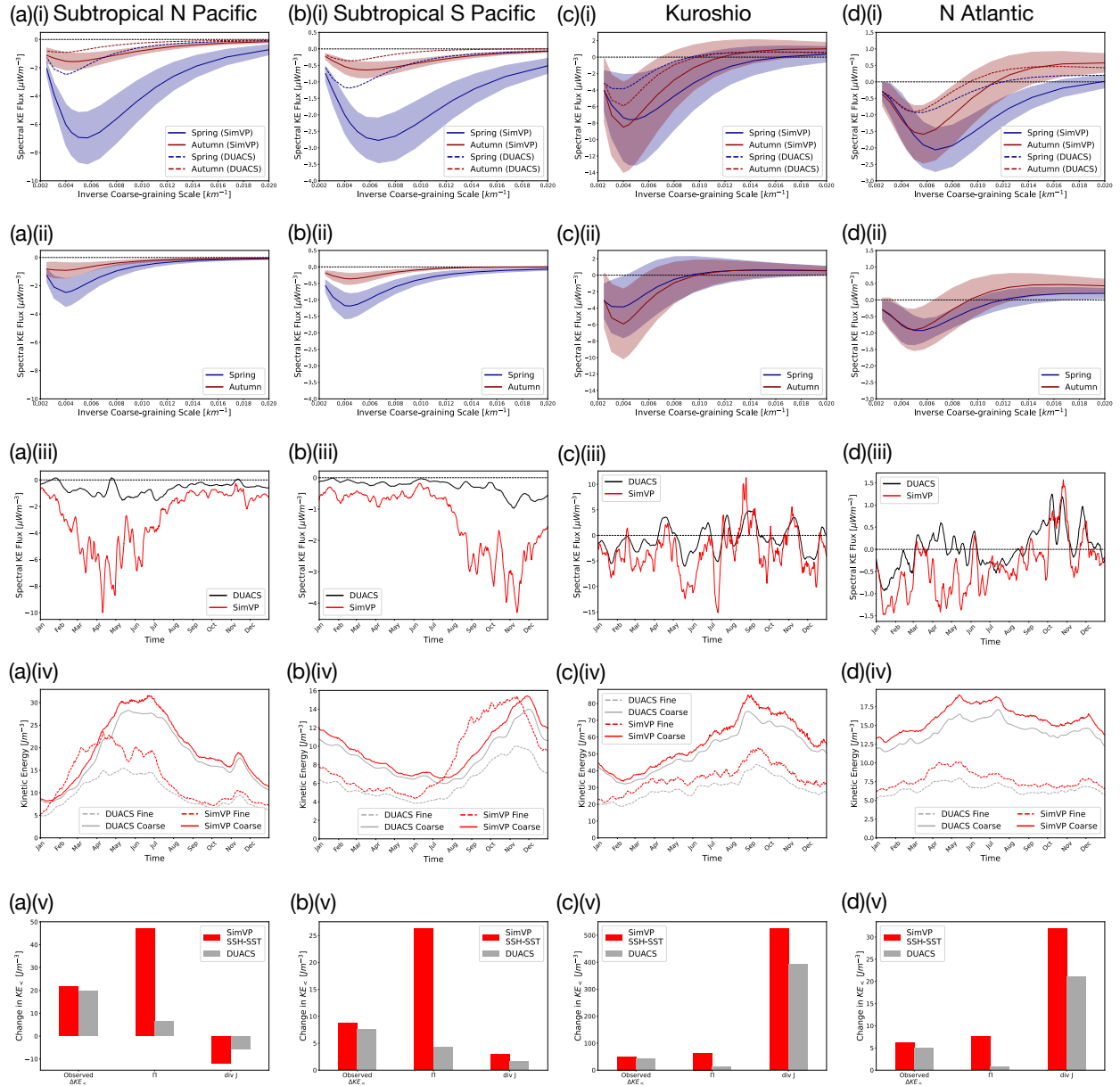


Figure A.6: (a) Subtropical North Pacific (149–161E, 19–31N). (i) Mean (solid lines) and standard deviation (shading) of KE cascade from NeurOST SSH-SST surface currents for the seasons of maximum (Spring) and minimum (Autumn) upscale cascade with mean DUACS cascades for reference (dashed lines). (ii) Same as (i) but from DUACS. (iii) Time-series of KE cascade across 250km for both NeurOST SSH-SST and DUACS. (iv) Time-series of coarse- and fine-scale KE (above and below coarse-graining scale of 125km respectively) from both NeurOST SSH-SST and DUACS. (v) Change in $KE_{<}$ (coarse-scale KE) from its wintertime minimum to its summertime maximum compared to the diagnosed contribution from the KE cascade ($-\int \Pi dt$), and the spatial transport of coarse-scale KE ($-\int \nabla \cdot J dt$) for both NeurOST SSH-SST and DUACS. (b)(i–v) Same as (a)(i–v) but for the Subtropical South Pacific (194–206E, 19–31S). (c)(i–v) Same as (a)(i–v) but for the Kuroshio (153–167E, 29–41N). (d)(i–v) Same as (a)(i–v) but for the North Atlantic (322–338E, 39–51N) and with the dividing coarse-graining scale between fine- and coarse-scale KE reduced to 80km.

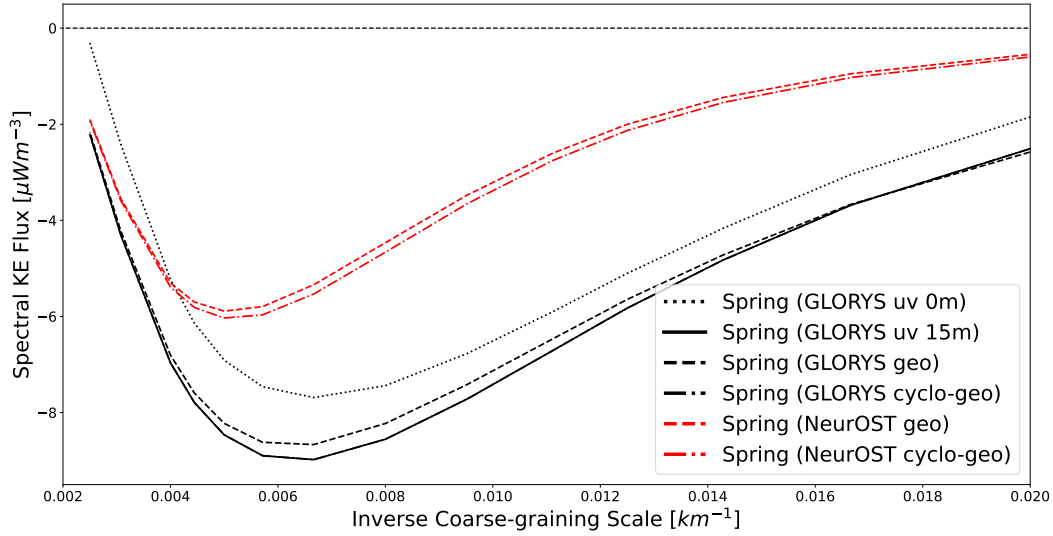


Figure A.7: Springtime spectral KE flux estimated from GLORYS and the NeurOST reconstruction of GLORYS from synthetic observations (OSSE) in the Subtropical North Pacific. Results for GLORYS are presented using the total surface currents at 0m and 15m depths (‘uv’), using surface geostrophic currents (‘geo’) calculated from SSH, and using the cyclo-geostrophic correction of Penven et al. [2014] (‘cyclo-geo’). Results for NeurOST are presented for both geostrophic and cyclo-geostrophic currents.

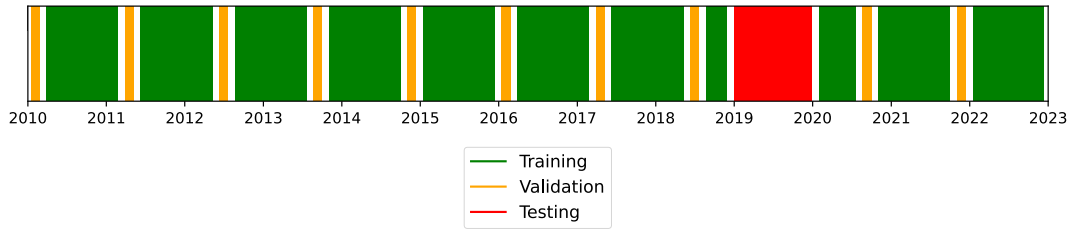
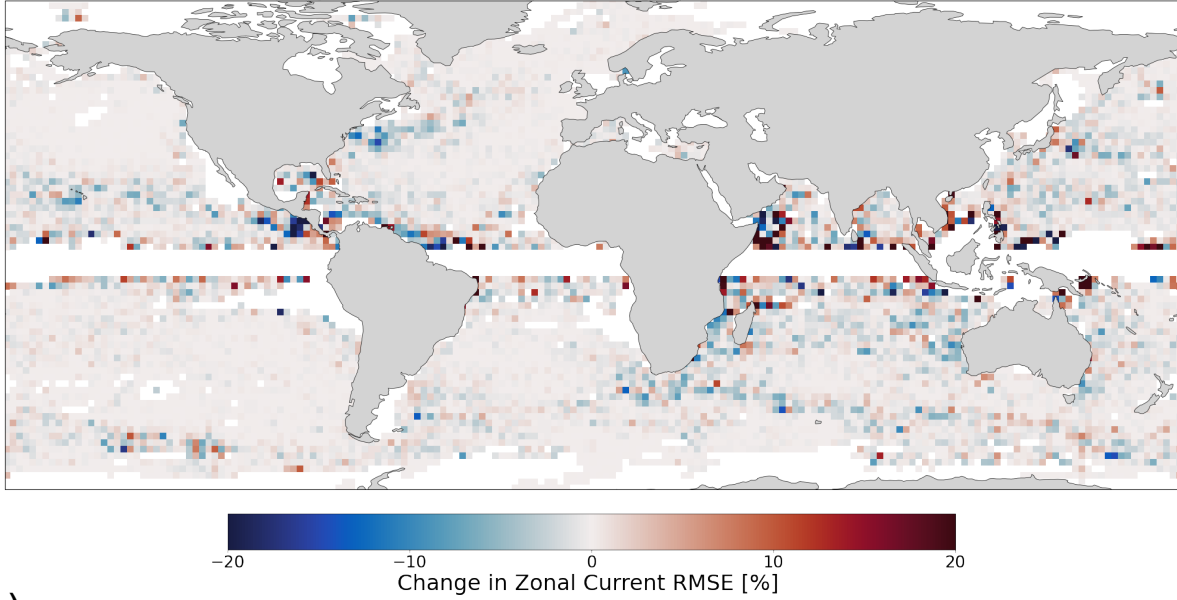


Figure A.8: Partitioning of dates between training, cross-validation, and testing when training neural network for SSH mapping.

(a)



(b)

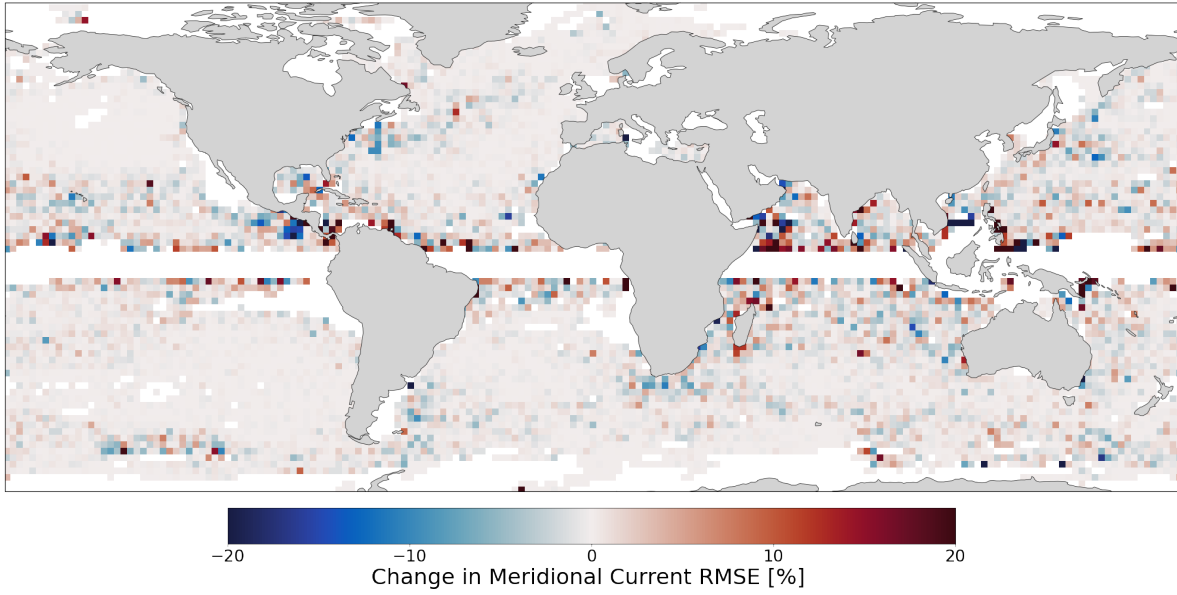


Figure A.9: (a) Change in RMSE of zonal current evaluated using drifters when the cyclo-geostrophic correction of Penven et al. [2014] is applied to geostrophic currents from NeurOST SSH-SST. (b) Same as (a) but for meridional currents.

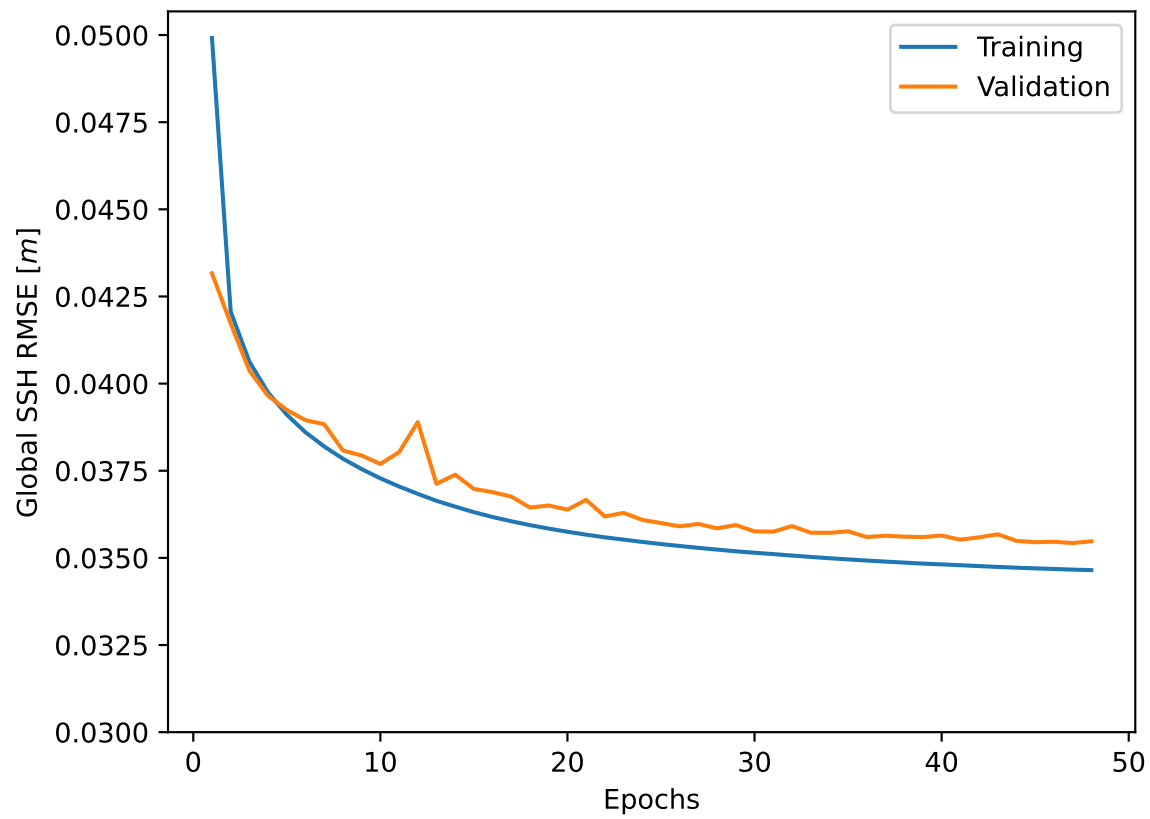


Figure A.10: Learning curves for training of global NeurOST (SSH-SST) network showing globally averaged SSH RMSE on both the training and cross-validation datasets.

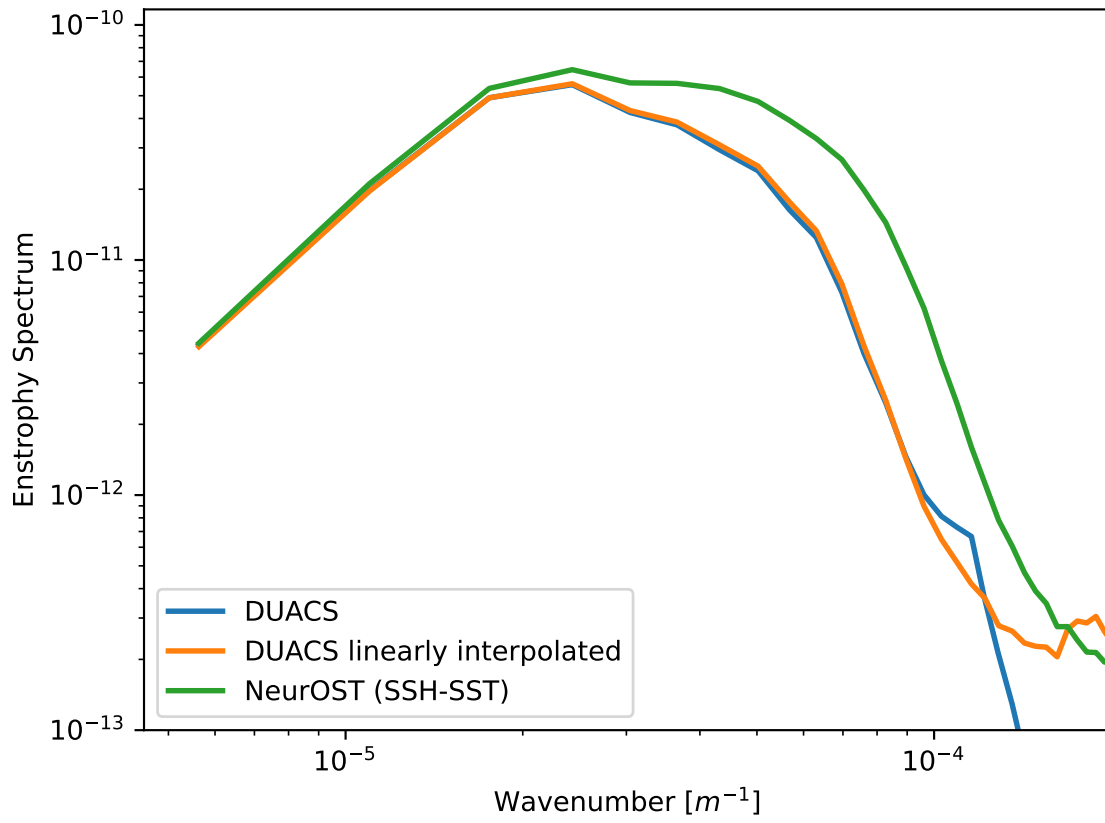


Figure A.11: Enstrophy spectra in Subtropical North Pacific for DUACS, NeurOST, and DUACS but with velocities first linearly interpolated to the NeurOST grid spacing before estimating derivatives.

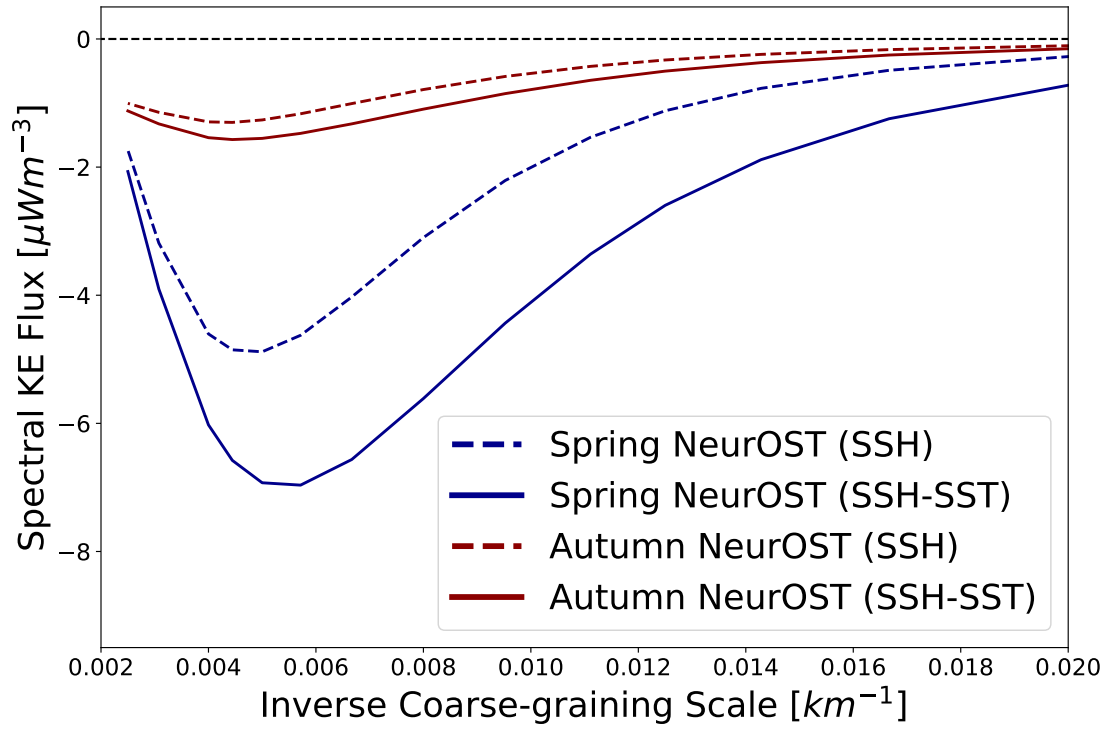


Figure A.12: Kinetic energy cascades in Subtropical North Pacific from NeurOST (SSH) and NeurOST (SSH-SST).

Chapter 3

Generative data assimilation for surface ocean state estimation from multi-modal satellite observations

This chapter is reproduced from the following publication:

Martin, S. A., Manucharyan, G. E., & Klein, P. (2025). Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. Journal of Advances in Modeling Earth Systems, 17, e2025MS005063. <https://doi.org/10.1029/2025MS005063>

3.1 Introduction

3.1.1 Background & Motivation

Conventional Methods for Surface Ocean State Estimation

Estimating the dynamical state of the surface ocean at sufficiently high resolution to resolve mesoscale eddies and their associated fronts and filaments (‘eddy-resolving’ hereafter) is fundamental for research into air-sea fluxes [Rai et al., 2021; Seo et al., 2023], vertical ocean heat transfer [Siegelman et al., 2020a], and the influence of eddies on marine ecosystems [Lévy et al., 2018; Zhang et al., 2019]. A large volume of surface ocean observations come from satellites that observe sea surface height (SSH), sea surface temper-

ature (SST), and sea surface salinity (SSS). Each variable is observed at varying resolutions and with often large spatial and/or temporal gaps between satellite passes or due to cloud occlusion (Figure 3.1). In the case of SSH, the sampling is particularly sparse, with only point-wise measurements along satellite tracks typically separated by tens to hundreds of kilometers in space and days to weeks in time. The recent launch of the first wide-swath satellite altimeter, SWOT [Fu et al., 2024], now provides groundbreaking 2D snapshots of SSH, but its 21-day orbital return time still leaves large spatiotemporal gaps [Archer et al., 2025]. Other dynamical variables, such as surface currents remain entirely unobserved by satellites and must thus be inferred either from sparse in situ observations or indirectly from SSH. The proposed ODYSEA mission could in future help to address this crucial gap in the satellite observing system [Torres et al., 2023]. Estimating the dynamical state of the surface ocean, defined here as the 2D SSH, SST, SSS, and surface current velocity fields from sparse multi-modal satellite observations (SSH, SST, & SSS) is hereafter referred to as ‘surface ocean state estimation’. While recent studies have developed approaches to reconstruct single variables at eddy-resolving scales (e.g. Martin et al. [2024]), eddy-resolving surface ocean state estimation requires jointly reconstructing SSH, SST, SSS, and surface currents while maintaining dynamical consistency between variables across scales. This is beyond the current capabilities of both dynamical and novel data-driven approaches and in this study we seek to address this.

One approach to state estimation is to assimilate observations, both satellite and in situ, into an ocean general circulation model (GCM) using data assimilation. By leveraging a GCM in the state estimation process, data assimilation ensures the reconstructed state satisfies the equations of motion for ocean dynamics while minimizing the misfit to observations. In order to satisfy the equations of motion, GCM data assimilation must reconstruct the full 3D ocean state rather than just surface fields. This requires in situ observations to constrain the ocean interior, but in situ observations are significantly more sparse than satellite observations of the surface, leading to suboptimal surface state estimation. The highest-resolution global data assimilation products (e.g. GLORYS 12 [Lellouche et al., 2021]) do not yet accurately resolve mesoscale eddies, showing high errors when compared to satellite observations of SSH or SST compared to statistical approaches like objective analysis [Lellouche et al., 2021]. Additionally, performing data assimilation with ocean GCMs of sufficiently high resolution to fully resolve ocean eddy dynamics is computationally prohibitive. For instance, the $1/12^\circ$ GCM used in GLORYS 12 is too coarse to resolve submesoscale dy-

namical processes that can affect the formation and evolution of mesoscale eddies [Taylor and Thompson, 2023]. These challenges are circumvented by data assimilation schemes for idealized, single-layer models of surface ocean dynamics [Le Guillou et al., 2021, 2023, 2024]; however, such idealized models by design make stringent assumptions about ocean dynamics that do not hold in general in the real world.

Due to the limitations of dynamical data assimilation, most state-of-the-art gridded satellite products are generated using statistical objective analysis methods, such as optimal interpolation (OI) [Bretherton et al., 1976]. In OI, linear least squares estimation predicts the missing SSH, SST, or SSS values in the gaps between observations based on a covariance model specified a priori. OI is a statistical method with no mechanism to ensure the equations of motion are respected in the resulting state estimate, in contrast to DA. By relaxing the requirement to satisfy dynamical equations, and by focusing only on surface fields, OI achieves significantly smaller errors compared to independent observations than DA. However, OI typically results in overly smooth and physically unrealistic reconstructed fields, with inaccurate representations of mesoscale eddies [Ballarotta et al., 2019]. Additionally, OI produces incoherent estimates across different observables, with different effective resolutions for each observable, which are typically each mapped separately. OI also cannot provide a mechanism for estimating variables that were never observed, like surface currents. Consequently, surface currents are usually derived from the OI-mapped SSH field under the assumption of geostrophy which holds only at mesoscale and larger [Le Traon et al., 1998; Taburet et al., 2019].

Deep Learning Approaches: Observation Vs. Simulation Learning

In recent years, data-driven deep learning methods have been increasingly adopted in satellite oceanography to better resolve ocean eddies. Deep learning approaches seek to solve satellite oceanography inverse problems using deep neural networks trained on large quantities of either real or simulated ocean data. Applications considered by past studies include: high-resolution SSH mapping from satellite altimetry and SST [Fablet et al., 2021b; Manucharyan et al., 2021; Buongiorno Nardelli et al., 2022; Beauchamp et al., 2022; Martin et al., 2023; Archambault et al., 2023; Fablet et al., 2023, 2024; Ciani et al., 2024; Febvre et al., 2024a; Archambault et al., 2024b,a; Martin et al., 2024], filling in the gaps in high-resolution infrared SST observations caused by clouds [Agabin et al., 2024; Goh et al., 2023; Fanelli et al., 2024], and inferring

ageostrophic surface currents from SSH/SST [Sinha and Abernathey, 2021; Fablet et al., 2023; Xiao et al., 2023; Fablet et al., 2024]. A number of studies have demonstrated that deep learning provides an effective way to synergize multiple satellite observables to improve performance [Sinha and Abernathey, 2021; Buongiorno Nardelli et al., 2022; Martin et al., 2023; Fablet et al., 2023; Archambault et al., 2023; Fablet et al., 2024; Ciani et al., 2024; Martin et al., 2024; Archambault et al., 2024b,a; Kugusheva et al., 2024]. While there are differences between these past studies in terms of objective, training data, and neural network architecture, they typically share the same overarching framework: supervised learning is used to train a neural network to solve a certain class of satellite oceanography inverse problem for a given observing system.

For certain satellite oceanography problems, neural networks can be trained directly on real-world observations without using synthetic data from GCM simulations in any way. For SSH mapping from nadir altimeters, it has been shown that the large volume of past real-world satellite SSH and SST observations is sufficient to train a neural network to map SSH at eddy-resolving scales [Martin et al., 2023; Archambault et al., 2023; Martin et al., 2024; Archambault et al., 2024b], recently achieving state-of-the-art global SSH mapping [Martin et al., 2024]. However, these observation-only learning methods are only applicable to tasks for which there are sufficiently dense real-world observations of the target variable to construct a training dataset. This is unlikely to be the case for many surface ocean variables, for example ageostrophic surface currents which are only sparsely observed by drifters, or SSS which is only observed at relatively low resolution from satellites. Additionally, the sparsity and sensor noise in real-world observations mean observation-only learning typically results in reconstructions that are overly smooth compared to numerical simulations (e.g. Archambault et al. [2024b]).

To address the limitations of observation-only learning, a number of studies have proposed ‘simulation learning’ approaches where synthetic data from ocean GCMs is used during training. In these approaches, a neural network is typically trained in a supervised regression setting, using pseudo-observations sampled from a GCM as inputs and the corresponding complete GCM target fields as ground truth labels [Beauchamp et al., 2022; Fablet et al., 2023; Febvre et al., 2024a; Fablet et al., 2024; Agabin et al., 2024; Goh et al., 2023; Archambault et al., 2024b,a]. During inference, the trained network receives real-world observations and generates a state estimate that mirrors the characteristics of the GCM it was trained on [Febvre et al.,

2024a]. Analogous to data assimilation, simulation learning seeks to generate a state estimate with dynamics resembling that of a GCM. However, there is no explicit mechanism to ensure the equations of motion are respected, so simulation learning does not ensure dynamical consistency in as strict a sense as dynamical data assimilation.

While simulation learning approaches are a promising avenue for surface ocean state estimation, the supervised regression framework described above has a number of limitations:

- Even subtle discrepancies between the pseudo-observations and real-world observations may propagate unpredictably through the network at inference. This often requires fine-tuning on real-world data [Archambault et al., 2024a], which is challenging for sparsely-observed variables like ageostrophic surface currents.
- Supervised regression seeks to predict a single state estimate that has the smallest average error (e.g. mean square error (MSE)). This induces a spectral bias where the large-scale signals dominate the MSE, leading to overly-smooth predictions with artificially steep spectral slopes.
- Supervised regression typically provides no natural metric of uncertainty for the resulting state estimate.
- Supervised regression requires bespoke training for each observing system, necessitating computationally expensive re-training each time the input observing system or the GCM pseudo-observation sampling strategy is altered.

Ideally, a deep learning method for surface ocean state estimation would address the above points. In addition, it would be desirable to obtain multi-variate surface state estimates where multiple variables are reconstructed jointly, preserving the dynamical consistency between variables across scales. This coherence between variables across scales is crucial for evaluating dynamical diagnostics such as eddy fluxes and frontogenesis rates.

A Generative Deep Learning Approach to Simulation Learning

Traditional data assimilation uses methods like Kalman filters or variational methods that work well for linear systems but struggle with highly nonlinear systems, high-dimensional data, and uncertainty in mea-

surements. Diffusion models [Song and Ermon, 2019; Karras et al., 2022; Croitoru et al., 2023] are generative deep learning models that can naturally learn complex probability distributions of data, making them ideal for highly nonlinear systems, such as oceanic and atmospheric turbulence. Manshausen et al. [2025] recently highlighted the strong potential of generative data assimilation based on these diffusion models for reconstructing weather from sparse observations.

To address the shortcomings of simulation learning approaches, in this study we explore the efficacy of generative diffusion models for surface ocean state estimation. Unlike in the regression formulation, generative models seek to model the full distribution of the training data, mapping from random latent vectors to ‘realistic’ examples from the desired distribution — see Buzzicotti [2023] for a review of generative models and their application to data reconstruction in complex flows. Diffusion models generate realistic examples by learning to reverse a prescribed forward process that degrades the data, typically through the addition of Gaussian noise. Applying the trained diffusion model to random noise fields then allows to generate realistic, but random, samples from the data distribution [Song and Ermon, 2019; Karras et al., 2022; Croitoru et al., 2023] (Section 3.2.1). For surface ocean state estimation, generative models could allow to retain fine-scale features and dynamical consistency between variables in the output by producing state estimates that ‘look like’ examples from the training distribution — in our case, multi-variate snapshots from high-resolution GCM simulations. The primary challenge in using generative models for state estimation is controlling the output such that the ‘realistic’ samples generated correspond well to available observations. One approach is to use conditional diffusion models, where a model is trained to generate high-resolution examples conditioned on both random noise and a low-resolution degradation of the data [Wang et al., 2024; Han et al., 2024; Ghosh et al., 2024]. However, this approach, like the supervised regression approach, requires careful design of training pairs such that the low-resolution data degradations are representative of the inputs available in the real world at inference. Instead, we here explore an approach that requires no generation of pseudo-observations from a GCM.

Score-based data assimilation (also referred to here as ‘generative data assimilation’) overcomes the challenge of controlled generation in a way that decouples the neural network training from the observing system using a two-stage strategy [Rozet and Louppe, 2023a]. First, an unconditional, score-based diffusion model is trained to generate realistic samples from a high-resolution training dataset (e.g. GCM output).

Second, the generation procedure of the diffusion model is guided in a Bayesian manner using observations (with no re-training of the diffusion model) such that the output fits sparse or degraded observations while preserving the learned characteristics of the training data [Rozet and Louppe, 2023a,b] (Section 3.2.1). Since the diffusion model is trained only on full model fields, the learned network weights are not specific to the observing system used at inference. Properties of the observing system are all encoded in the observation operator used to guide the generation at inference (Section 3.2.1). This allows training a single diffusion model then using it for inference with a wide range of observing systems with no additional re-training. Even when certain variables, like surface currents, are unobserved, the diffusion model still predicts them in a way that should remain dynamically consistent through the learned relationships to the observed variables.

Generative data assimilation was recently shown to be effective for producing 3km-resolution atmospheric state estimates from sparse weather station observations [Manshausen et al., 2025]. Notably, Manshausen et al. [2025] demonstrated this method had promising ‘channel synthesis’ capabilities that allowed estimating a completely unobserved variable (meridional winds) from observations of other atmospheric state variables with reasonable accuracy and qualitative physical realism. Channel synthesis is crucial in surface ocean state estimation since many quantities of interest (e.g. ageostrophic surface currents) are only sparsely observed by in situ platforms but have strong signatures on satellite observables like SSH, SST, and SSS [Sinha and Abernathey, 2021; Fablet et al., 2023, 2024].

3.1.2 Our Contributions

Here, we adapt the generative data assimilation method (referred to as ‘GenDA’ hereafter) developed for atmospheric reanalysis in Manshausen et al. [2025] to the closely-related problem of eddy-resolving surface ocean state estimation by designing an observation operator for assimilating multi-modal, multi-resolution satellite observations. We demonstrate that GenDA, trained on a GCM-based data assimilation product (GLORYS 12), is capable of solving realistic satellite oceanography inverse problems without bespoke training for each observing system.

Using an observing system simulation experiment, we compare GenDA to a baseline supervised learning approach trained for one specific observing system. We build upon the observation operator used in Manshausen et al. [2025] by adding coarse-graining terms to allow incorporating information from existing

low-resolution satellite SSH, SST, and SSS products (e.g. those created using OI), ensuring the accuracy at large scales of the GenDA state estimates. While previous studies have attempted to predict surface currents from SSH and SST [Sinha and Abernathey, 2021; Fablet et al., 2023, 2024], the method proposed here is the first deep learning approach, to our knowledge, that allows to jointly reconstruct the full surface ocean state vector (SSH, SST, SSS, and surface currents).

Finally, we explore the ability of GenDA to generalize to real-world observations through an observing system experiment. We demonstrate that GenDA generalizes from simulation training to real-world inference better than a baseline supervised learning approach, with lower errors and improved physical realism. The resulting GenDA real-world surface ocean state estimates preserve the dynamical characteristics of the simulation data used during training. GenDA’s generative formulation reduces the spectral bias in regression-based approaches, exhibiting realistic dynamics across scales. GenDA has significantly smaller errors against independent satellite observations than a state-of-the-art dynamical data assimilation system.

3.2 Methods

3.2.1 GenDA: Generative Data Assimilation

We seek to estimate the 2D dynamical state of the surface ocean which we represent using a state vector, x . Concretely, x here will be (SSH, SST, SSS, u_{ageo} , v_{ageo}) at each point on a regular 2D grid, where u_{ageo} and v_{ageo} are the ageostrophic components of the zonal and meridional surface current velocities respectively. Note, we reconstruct only the ageostrophic surface currents to focus GenDA on reconstructing currents not directly retrievable from SSH. We define the ageostrophic surface current to be the residual between the total surface current and that predicted from SSH assuming geostrophic balance — note this ageostrophic current will contain both balanced and unbalanced components. To estimate x , we use potentially sparse or degraded (e.g coarse-grained or noisy) observations, y , which are the result of applying an observation operator, $\mathcal{A}$, to the state vector,

$$y = \mathcal{A}(x). \quad (3.1)$$

Score-Based Diffusion Models

Diffusion models are a powerful class of generative deep learning models, trained to generate realistic examples drawn from a distribution of training data like natural images or, in our case, snapshots from ocean GCM simulations. During training, a neural network, D , is trained to ‘de-noise’ examples from the training dataset by predicting the noise-free example from its noisy counterpart [Karras et al., 2022]. In our case, D gives a prediction, $\hat{x}(t)$, of the noise-free ocean state vector, $x(t = 0)$, given its noisy counterpart, $x + \sigma(t)z$, where $\sigma(t)$ is a variable noise amplitude, z is a unit variance Gaussian noise vector with the dimensions of x , and t is a ‘time’ axis along which the amplitude of the noise added varies from 0 at $t = 0$ to σ_{max} at $t = T$ (Figure 3.2a). D is trained by minimizing MSE between $\hat{x}(t)$ and $x(t = 0)$ (Figure 3.2b).

Given a trained de-noising neural network, D , new samples from the noise-free distribution of $x(t = 0)$ can be generated from random noise, $x(t = T)$. This is achieved by simulating the time reversal of an ordinary differential equation defined to gradually transform examples from the data distribution, $x(t = 0)$, to random Gaussian noise, $x(t = T)$, when evolved forward in time

$$\frac{dx}{dt} = -\dot{\sigma}(t)\sigma(t)\nabla_x \log(p(x(t))), \quad (3.2)$$

where the ‘time’ axis, t , varies from $t = 0$ (the data distribution) to $t = T$ (Gaussian noise distribution with scale σ_{max}), and $\dot{\sigma}$ is the derivative of the noise amplitude schedule with respect to the diffusion time axis [Song and Ermon, 2019; Karras et al., 2022]. The term $\nabla_x \log(p(x(t)))$ is referred to as the ‘score function’ or ‘score network’ and represents the gradient of the log-likelihood of x . The key to score-based diffusion models lies in recognizing that a well-trained de-noiser, D , can be used to approximate the score function in Equation 3.2 through [Karras et al., 2022]

$$\nabla_x \log(p(x(t))) = \frac{D(x(t)) - x(t)}{\sigma(t)^2}. \quad (3.3)$$

Realistic (but random) ocean states, $x(t = 0)$, can thus be generated by starting with random noise, $x(t = T)$, and solving Equation 3.2 backward in time from $t = T$ to $t = 0$ using a finite difference discretization of the time axis and using the trained denoiser, D , to calculate the score function at each time step through Equation 3.3. This reverse time-stepping procedure updates the state, $x(t)$, to maximize

the likelihood $p(x(t = 0))$ which is encoded in the trained score network (also referred to hereafter as the ‘diffusion prior’). See S.I. Text S3 for details on how we implement D as a neural network.

Score-Based Data Assimilation

While reverse time-stepping Equation 3.2 using the trained score function (Section 3.2.1) provides a way to generate realistic ocean states, these states are random and are not related to any observations. Score-based data assimilation provides a way to control the generation process to push the generated state, $x(t = 0)$ to match the observations, y [Rozet and Louppe, 2023a]. Essentially, this framework seeks to replace the score function, $\nabla_x \log(p(x(t)))$, with the corresponding gradient of the *posterior* log-likelihood, $\nabla_x \log(p(x(t)|y))$.

The optimal state estimate would be the one that maximizes the posterior likelihood of the reconstructed state given the available observations, $p(x|y)$, or equivalently minimizes $-\log(p(x|y))$. The gradient of the posterior log-likelihood can be expressed through Bayes’ theorem as

$$\nabla_x \log(p(x|y)) = \nabla_x \log(p(x)) + \nabla_x \log(p(y|x)), \quad (3.4)$$

from which it follows also that

$$\nabla_x \log(p(x(t)|y)) = \nabla_x \log(p(x(t))) + \nabla_x \log(p(y|x(t))). \quad (3.5)$$

Rozet and Louppe [2023a] thus propose to replace the score function in Equation 3.2 with the posterior log-likelihood gradient given by Equation 3.5. This way, when generating ocean states from random noise by simulating the time reversal of Equation 3.2, the final state will be one that maximizes the posterior likelihood given the observations, y , and the trained diffusion prior (i.e. the score network) which encourages the reconstructed state to ‘look like’ the GCM simulation training data.

At inference, the first term in Equation 3.5 (the score function) is known (Equation 3.3) and it remains to approximate the second term. This second term is formally only known at $t = 0$, i.e. the data distribution, but Rozet and Louppe [2023a] propose that, assuming a Gaussian observing process, it can be approximated

by

$$p(y|x(t)) = \mathcal{N}(y|\mathcal{A}(\hat{x}(t)), \Sigma_y(t)), \quad (3.6)$$

where $\Sigma_y(t)$ is a heuristic variance that increases with noise level (i.e. with t), $\hat{x}(t)$ is the de-noised state predicted by the diffusion model, and $\mathcal{A}$ is the observation operator. See Rozet and Louppe [2023a] and S.I. Text S3.2 for details of the heuristic used for $\Sigma_y(t)$. Formally, this is only an approximation of the true $p(y|x(t))$, and this could introduce biases by replacing a complex and potentially multi-modal distribution with a Gaussian of heuristic variance. Nonetheless, GenDA has been shown to have impressive reconstruction abilities both for idealized quasi-geostrophic turbulence [Rozet and Louppe, 2023b] and for kilometer-scale atmospheric reanalysis [Manshausen et al., 2025].

In summary, to generate an estimate for the ocean state from observations we start with a random state vector, $x(t = T)$, and perform reverse time steps using Equations 3.2, 3.3, 3.5, and 3.6 to sample from $p(x(t = 0)|y)$ (Figure 3.3). Crucially, the updates made during reverse time-stepping are not adjustments to the neural network parameters but instead updates to the state $x(t)$. This way, generated state estimates should still qualitatively ‘look like’ examples from the simulation since we impose our simulation-trained diffusion prior on the state estimation. In this study we propose that forcing the state estimates to be qualitatively similar to the numerical simulation will allow us to preserve fine-scale features and inter-relations between observed and unobserved variables at inference. Another advantage of the GenDA method is that it is Bayesian, so rather than producing a single state estimate, GenDA produces a distribution of plausible states sampled from the posterior distribution, achieved practically by inputting different initial random noise vectors, $x(t = T)$. The dispersion of this ensemble thus appears a natural metric of uncertainty for the resulting state estimate. In this study, we generate ensemble state estimates with 24 ensemble members and explore the suitability of ensemble dispersion for uncertainty quantification.

Observation Operator for Surface Ocean State Estimation

There is significant flexibility in the choice of $\mathcal{A}$ used at inference. Since a single diffusion model is trained with no specific observing system in mind, GenDA provides a natural low-cost way to use a wide variety of observing systems to produce state estimates at inference without retraining the diffusion model. Relevant types of observing systems include highly localized point-wise measurements, like from weather stations

or satellite tracks, and coarse observations that represent relatively large-scale satellite footprints in space and/or time. Manshausen et al. [2025] explored sparse, point-wise measurements of the atmosphere from weather stations for $\mathcal{A}$, demonstrating the efficacy of diffusion models for inference from sparse observations. Here, we will develop a method allowing the inclusion of low-resolution state estimates to constrain the large-scale state in GenDA. The motivation for using coarse observing systems in conjunction with high-resolution sparse observations is that existing gridded satellite products for SSH, SST, and SSS (e.g. from OI) give a reasonable estimate of the ocean state at large mesoscales and above, whereas high-resolution, un-gridded satellite observations (say of SSH or SST) typically have large gaps on any given day. Incorporating these low-resolution gridded products implicitly incorporates observations from a longer time horizon to constrain the large-scale state, while sparse and instantaneous high-resolution observations will be used to inform smaller scales.

In practice, we compare low-resolution OI satellite products to the diffusion model output by coarse-graining the latter using an appropriate spatial scale. The coarse-graining scale is considered to be known for any OI product — see S.I. Text S5 for a discussion of how these coarse-graining scales can be selected. Concretely, the observation operator $\mathcal{A}$ generates several fields, including an instantaneous observation term, $\mathcal{A}_{inst}$, and low-resolution satellite product terms for each of the OI product, $\mathcal{A}_{smooth}$:

$$\mathcal{A}(x) = \text{concat}(\mathcal{A}_{inst}(x), \mathcal{A}_{smooth}(x; l)). \quad (3.7)$$

$\mathcal{A}_{inst}(x)$ selects the indices for the nearest pixels and variables (channels) where instantaneous observations are available and will therefore yield sparse, high-resolution satellite observations of SSH or SST. Meanwhile, $\mathcal{A}_{smooth}(x; l)$ is a coarse-graining operation, where we apply a low-pass Gaussian filter with kernel width, l , to the state estimate, x . This allows us to compare a coarse-grained view of the state estimate to a low-resolution satellite product to encourage agreement at large scales without penalizing small-scale features below the resolution limit of the low-resolution satellite products. We apply $\mathcal{A}_{smooth}(x; l)$ to all variables for which we have access to low-resolution real-world satellite products (i.e. SSH, SST, SSS) and we choose coarse-graining scales, l , for each variable to be representative of the effective resolution of the respective satellite products. Note that the coarse-graining scale, l , applied to each satellite observable in the observation operator is something we prescribe and should be chosen to reflect the effective resolution

of each satellite product being assimilated (S.I. Text S5). We treat ageostrophic current velocities as being essentially unobserved in the real world, and so neither $\mathcal{A}_{inst}(x)$ nor $\mathcal{A}_{smooth}(x; l)$ returns any values for u_{ageo} or v_{ageo} .

In practice, we find that a significant portion of the ageostrophic surface currents is driven by wind stress. This poses a problem for reconstructing ageostrophic currents from only SSH, SST, and SSS observations without information about the wind-forcing. We employ two strategies to mitigate this. Firstly, we use a linear Ekman model (S.I. Text S1) to predict the wind-driven surface currents, u_{Ek} and v_{Ek} , from the wind stress and subtract this from u_{ageo} and v_{ageo} , re-framing our state estimation problem so that we seek to estimate

$$u_{ageo} = u_{total} - u_{geo} - u_{Ek}, \quad (3.8)$$

and

$$v_{ageo} = v_{total} - v_{geo} - v_{Ek}. \quad (3.9)$$

Secondly, since the linear Ekman model doesn't capture all wind-driven variability in surface currents, we also expand our state vector, x , to include surface winds, u_{atmos} and v_{atmos} . We thus train the diffusion model to jointly generate surface ocean states alongside corresponding surface wind fields. During the assimilation process, we then provide surface wind 'observations' through $\mathcal{A}_{inst}(x)$, providing information about the wind-forcing to the state estimation and thus improving the reconstruction of ageostrophic surface currents (S.I. Text S2). Note, this method is still directly applicable to real-world observations since reanalysis winds (e.g. ERA5) are available in the real-world setting.

3.2.2 UNet Regression: Baseline Supervised Learning Framework

To provide a baseline method against which to compare and contrast the GenDA state estimates we also implement a supervised learning approach. Unlike GenDA, this supervised approach is trained specifically for the observing system under consideration by creating pseudo-observations from the high-resolution GCM simulation and training a neural network to predict the corresponding state vector, x , in one single pass through the network from these inputs. That is, we estimate the state vector as

$$\hat{x} = f_{\theta}(y), \quad (3.10)$$

where $\hat{x}$ is our estimate of x , y are the potentially sparse or degraded observations, and f_{θ} is a neural network whose parameters, θ , we seek to optimize by minimizing the MSE between $\hat{x}$ and x . Note, unlike GenDA the supervised approach provides a single prediction for each state rather than a distribution and requires bespoke training for each observing system.

In practice, we implement f_{θ} as a UNet [Ronneberger et al., 2015] (S.I. Text S3) with 7 input channels in y :

- Simulated sparse, high-resolution (i.e. ‘Level 3’ or L3) SSH satellite observations
- Simulated sparse, high-resolution (i.e. ‘Level 3’ or L3) SST satellite observations
- Zonal surface wind from reanalysis
- Meridional surface wind from reanalysis
- Simulated low-resolution, gap-free (i.e. ‘Level 4’ or L4) SSH from OI
- Simulated low-resolution, gap-free (i.e. ‘Level 4’ or L4) SST from OI
- Simulated low-resolution, gap-free (i.e. ‘Level 4’ or L4) SSS from OI

Hereafter we refer to the baseline supervised learning method as ‘UNet Regression’.

3.3 Datasets & Experiment Set-Up

3.3.1 Oceanic & Atmospheric Reanalysis Products for Training and Evaluation

In this study, we provide a regional proof-of-concept of eddy-resolving surface ocean state estimation by training GenDA on data from a high-resolution reanalysis product, GLORYS 12 [Lellouche et al., 2021; E.U. Copernicus Marine Service Information (CMEMS), 2024f]. This product is developed by assimilating satellite and in situ observations into the NEMO ocean GCM with $1/12^{\circ}$ grid resolution using a reduced-order Kalman filter. The ocean GCM is driven at the surface by atmospheric forcing from the ECMWF

Reanalysis product, commonly known as ERA 5 [Copernicus Climate Change Service (C3S, 2024; Hersbach et al., 2020)]. GLORYS 12 does not resolve submesoscale dynamics which is a key driver of ocean eddy dynamics, however, it has a sufficiently fine grid to resolve mesoscale baroclinic instability (a key formation mechanism for mesoscale eddies) and exhibits an abundance of submesoscale fronts and filaments generated by a forward cascade of enstrophy by mesoscale eddy stirring. We hence consider it a sufficiently realistic and high-resolution test bed for the GenDA method, though we emphasize that in the future, GenDA can be trained on free-running submesoscale-resolving simulations (e.g. LLC4320; Su et al. [2018]). One of the motivations of this study is that data assimilation products like GLORYS currently show large errors when compared to satellite observations at mesoscales, suggesting that optimizing the full 3D state of the ocean and ensuring its conformity to dynamical equations of motion comes at the expense of accurately placing mesoscale eddies and fronts at the surface. We hypothesize that GenDA, by optimizing only the surface fields, will reduce these biases while preserving a substantial amount of physical characteristics of the GLORYS 12 fields seen during training through the diffusion prior.

The experiments presented in this study focus on the Gulf Stream Extension region ($55\text{--}65^\circ$ W, $33\text{--}43^\circ$ N) which has one of the most energetic eddy fields in the global ocean. Focusing on this region also allows easy comparison to a number of other machine learning applications for SSH mapping through an Ocean Data Challenge [Ballarotta et al., 2021; Metref et al., 2023]. A wider region of ($70\text{--}40^\circ$ W, $25\text{--}45^\circ$ N) is used for all neural network training. The years 2010-2016 are used for network training, 2018-2020 for cross-validation, and 2017 is withheld as an independent test year on which the evaluation metrics presented throughout the rest of the study are calculated.

All variables in the 7-variable state vector, $x = (\text{SSH}, \text{SST}, \text{SSS}, u_{ageo}, v_{ageo}, u_{atmos}, v_{atmos})$, are standardized by subtracting the mean and dividing by the standard deviation of each variable calculated over the full training domain and time-series. In the case of SST and SSS, the strong seasonal cycle is removed by subtracting a monthly climatology rather than the mean. The ageostrophic surface currents are taken at 15 m depth to be more reflective of the eddy-driven ocean currents than those at the surface layer. For both GenDA and the UNet Regression baseline, the state vector, x , is estimated on a regular latitude-longitude grid at the same $1/12^\circ$ resolution as GLORYS 12 with dimensions of 128 by 128, corresponding roughly to a domain of size 1000 km.

3.3.2 Experiment 1: Observing System Simulation Experiment (OSSE)

First, we assess GenDA’s reconstruction abilities in a controlled environment where we have access to the full ground-truth fields for validation through an observing system simulation experiment (OSSE). After training GenDA on GLORYS data from 2010-2016, we generate synthetic satellite observations of GLORYS using the withheld testing year 2017 and evaluate GenDA’s ability to reconstruct the full fields, including variables unobserved by satellites (Figure 3.4).

Generation of Simulated L3 Satellite Observations

In this OSSE we consider two different sources of Level 3 (L3) input observations: SSH from satellite altimeters and high-resolution SST from satellite infrared radiometers. L3 refers to observations before they have been interpolated to a full gridded field, so L3 observations are sparse but high-resolution.

We generate synthetic L3 SSH observations by sub-sampling the GLORYS SSH field along the observation tracks from all conventional nadir altimeters available in the year 2017: SARAL/Altika, Jason 2, Jason 3, Sentinel 3A, Haiyang-2A and Cryosat-2. Nadir altimeters provide point-wise SSH measurements beneath the satellite track with an along-track spacing of approximately 7 km. In addition, we assess the impact of including wide-swath observations from the recently-launched Surface Water and Ocean Topography (SWOT) mission which provides two 60 km wide swaths of SSH measurements at 2 km resolution with a narrow gap between the swaths [Fu et al., 2024]. While SWOT has high spatial sampling it has a relatively long orbit repeat of 21 days. We sub-sample GLORYS SSH at the locations of SWOT observations by repeating its 21-day science orbit throughout the year 2017.

Satellite infrared radiometers observe SST at kilometer resolution but there are large gaps due to cloud cover. Here, we sub-sample GLORYS SST using cloud masks from the ODYSSEA SST Multi-Sensor L3 product [E.U. Copernicus Marine Service Information (CMEMS), 2024a] for the same region and dates. This ensures the applied cloud cover is realistic, with cloud cover varying on any given day between near-complete occlusion and cloud-free conditions. For both SSH and SST L3 observations we add Gaussian noise with amplitude chosen to reflect the noise for each sensor. We take the SSH noise amplitude from the relevant CMEMS product [E.U. Copernicus Marine Service Information (CMEMS), 2024e] and for SST from the average of the error estimates provided in the ODYSSEA SST Multi-Sensor L3 product [E.U.

Generation of Simulated L4 Satellite Products

In addition to high-resolution L3 observations, we provide simulated low-resolution gridded ('Level 4', L4) products for SSH, SST, and SSS. L4 refers to observations that have been interpolated to a full gridded field (typically using OI), so L4 observations are gap-free but low-resolution.

The interpolation algorithms used to interpolate between observations smooth out small-scale features, providing spatio-temporally coarse-grained estimates of the real fields. We simulate this by coarse-graining the GLORYS SSH, SST, and SSS both in space and time using a Gaussian kernel with spatial and temporal widths selected to be roughly representative of the resolutions of available L4 products (S.I. Text S5 & Table S2). These coarse-grained fields are then compared to coarse-grained output from GenDA in the observation operator when minimizing the reconstruction loss at inference. Since GenDA is applied only on single temporal snapshots, we can't replicate the temporal coarse-graining in the L4 products in the GenDA observation operator, and so resort to only coarse-graining in space. Finally, we also provide gridded values for u_{atmos} and v_{atmos} from ERA 5 with no additional coarse-graining applied since this product would also be available to use in the real-world setting.

Taken together, the simulated observations provided to GenDA and UNet Regression are:

- Along-track L3 SSH from a constellation of nadir altimeters
- Wide-swath L3 SSH from SWOT
- Cloud-obscured high-resolution L3 SST
- Coarse-grained L4 SSH
- Coarse-grained L4 SST
- Coarse-grained L4 SSS
- Surface wind velocities from ERA 5

Note, no observations are provided for u_{ageo} or v_{ageo} since these are not typically observed in the real world so we reconstruct these in an unguided manner.

Evaluation Metrics

We evaluate the point-wise accuracy for each reconstructed variable using the coefficient of determination (R^2), which represents the fraction of variance explained. For SSH and SST, we remove any pixels provided as part of the L3 observations from the evaluation metrics. Since the simulated low-resolution L4 OI products for SSH, SST, and SSS already capture a significant part of the signal, we also present R^2 for SSH, SST, and SSS calculated on the residual from the low-resolution L4 products. Note that R^2 is by definition zero for the L4 OI predictions in this setting and any positive value represents an improvement over OI. We assess the impact of cloud-free L3 SST observations on the reconstruction accuracy by computing R^2 separately for pixels that are clouded and un-clouded respectively. Since the average cloud cover has a significant spatial structure, we randomly over-sample and under-sample the clouded and un-clouded pixels using weights that ensure both datasets are uniformly drawn in latitude-longitude (See S.I. Text S6). We use the same procedure to assess the impact of L3 SWOT SSH observations on reconstruction accuracy.

We also evaluate the wavenumber spectra of each reconstructed variable and compare them to the ground truth from GLORYS. All spectra in this manuscript are calculated by estimating the 2D power spectral densities for each time step, averaging in time, then averaging azimuthally. Note, u_{ageo} and v_{ageo} are evaluated jointly through the ageostrophic kinetic energy spectrum and we also derive the geostrophic kinetic energy spectrum from the SSH reconstructions. Since we focus here on multi-modal state estimation, we also evaluate the spectra of eddy fluxes of temperature, F_T , and salt, F_S , which are sensitive to the resolution of both velocity and SST/SSS and are crucial diagnostics of the impacts of ocean eddies on climate. We focus here on the meridional eddy fluxes since these induce significant transport across the Gulf Stream in our test region,

$$F_T = v'SST', \quad (3.11)$$

$$F_S = v'SSS', \quad (3.12)$$

where primed quantities denote deviations from monthly climatology [Guo et al., 2022]. The instantaneous eddy fluxes F_T and F_S are averaged in spectral space following the same procedure as defined above for all other variables.

We also consider higher-order diagnostics of eddy dynamics that can be derived from SSH, namely the geostrophic Okubo-Weiss quantity, W , [Okubo, 1970; Weiss, 1991] which delineates strain-dominated flows from vorticity-dominated flows

$$W = (s^2 - \zeta^2) = (s_n^2 + s_s^2 - \zeta^2), \quad (3.13)$$

where

$$s = \sqrt{s_n^2 + s_s^2} = \sqrt{\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2} \quad (3.14)$$

is the strain rate and

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \quad (3.15)$$

is the relative vorticity. We further explore the accuracy of diagnosed eddy dynamics by evaluating the joint probability density function (JPDF) of s and ζ for each reconstruction. The accuracy of these diagnostics of eddy dynamics is important for studies of ocean scale interactions and are highly sensitive to the accuracy and resolution of SSH [Martin et al., 2023, 2024].

Finally, we evaluate the suitability of GenDA ensemble dispersion for uncertainty quantification through a rank histogram as is widely used in probabilistic weather forecasting [Talagrand, 1999]. This shows the probability that the ground truth falls in each rank of the ordered ensemble of predictions. Concretely, at each grid point we sort the ensemble predictions for a given variable from lowest to highest and assign the ground truth a ‘rank’ based on where in the ensemble it falls. If the truth falls below all ensemble members we assign rank of 0, if the truth is between the first and second ensemble member we assign rank of 1, and so on. For a well-calibrated ensemble of predictions, the ground truth should be indistinguishable from an individual ensemble member — the ground truth should have equal probability of falling in any rank of the ensemble. The rank histogram is constructed by aggregating the ranks over all grid points. Hence, a well-calibrated ensemble would have a flat rank histogram, an under-dispersive ensemble would have a ‘u-shaped’ rank histogram, an over-dispersive ensemble would have a ‘n-shaped’ rank histogram, and a biased ensemble would have an asymmetric rank histogram.

3.3.3 Experiment 2: Observing System Experiment (OSE)

To assess the ability of GenDA to generalize to the real-world ocean, we also perform an observing system experiment (OSE) in which we construct state estimates from real-world satellite observations. This poses a more stringent test of the method since now the observations come from the real-world ocean which potentially exhibits different characteristics to the GLORYS 12 simulated fields. Applying GenDA in the real world is akin to data assimilation, where we seek a field that qualitatively preserves the characteristics of GLORYS but that best matches available observations. Since GLORYS itself is already assimilated to observations, we can use errors against real-world satellite observations to benchmark GenDA against this state-of-the-art data assimilation product.

We conduct our OSE in the same Gulf Stream Extension region ($55\text{--}65^\circ$ W, $33\text{--}43^\circ$ N) as in the OSSE, again for the withheld test year 2017.

L3 Satellite Observations

We provide L3 observations of both SSH and SST. For SSH, the observations come from a constellation of nadir altimeters (SARAL/Altika, Jason 2, Jason 3, Sentinel 3A, Haiyang-2A) [E.U. Copernicus Marine Service Information (CMEMS), 2024d]. SWOT was yet to be launched during this test year so we provide only along-track SSH observations. To independently evaluate the accuracy of the mapped SSH fields, we withhold one altimeter, CryoSat-2, at inference and use this withheld altimeter as a ground truth for the mapped SSH. This follows the configuration of the 2021a Ocean Data Challenge [Ballarotta et al., 2021], allowing us to benchmark the performance of GenDA on real-world observations against state-of-the-art SSH mapping methods. For SST, we provide cloud-occluded infrared radiometer observations from the ODYSSEA multi-satellite product [E.U. Copernicus Marine Service Information (CMEMS), 2024a]. This product collates SST observations from a range of different satellites onto a regular 0.1° grid and includes only nighttime observations to remove diurnal variability.

L4 Satellite Products

We also provide low-resolution L4 estimates of SSH, SST, and SSS. For SSH, we use output from the deep learning model, ConvLSTM, presented in Martin et al. [2023]. This method uses a supervised, observation-

only learning regression approach to estimate gridded SSH at higher resolution than can be achieved with OI by synthesizing along-track SSH and gridded SST. In order to preserve the independence of the withheld evaluation altimeter, CryoSat-2, we ensure that the L4 SSH product was generated using all apart from the withheld altimeter and the network was never trained on observations from 2017. The ConvLSTM SSH fields have an effective resolution of 100 km (S.I. Table S3 in Martin et al. [2024]).

For SST, we use the REMSS MW-OI Global Foundation Sea Surface Temperature analysis product [Remote Sensing Systems, 2017]. This product uses OI to interpolate observations from microwave radiometer sensors onboard multiple satellites into a gridded L4 SST estimate. Note microwave radiometers have lower spatial resolution than the infrared observations we use for L3, but these sensors can penetrate clouds. The product is provided on a $1/4^\circ$ grid and we linearly interpolate it to the GenDA grid.

For SSS, we use the CMEMS Multi Observation Global Ocean Sea Surface Salinity and Sea Surface Density product [E.U. Copernicus Marine Service Information (CMEMS), 2024g]. This product is obtained through a multivariate OI algorithm that combines SSS images from multiple satellites with in situ salinity measurements and satellite SST and is provided on a $1/8^\circ$ grid which we linearly interpolate to the GenDA grid.

The effective coarse-graining scales of the real-world L4 SSH, SST, and SSS products are not known a priori, raising the question of what coarse-graining scales, l , to prescribe in the GenDA observation operator. We choose coarse-graining scales (S.I. Table S3) based on an a posteriori tuning strategy described in S.I. Text S5.2, though we acknowledge this choice is ultimately somewhat subjective.

As in the OSSE (Section 3.3.2), we also provide atmospheric wind speeds from ERA 5. No observations are provided of ageostrophic surface currents and these are reconstructed in an unguided manner.

In sum, GenDA and the UNet Regression are provided:

- L3 along-track SSH from 5 nadir altimeters
- L3 infrared SST observations occluded by clouds
- Coarse L4 SSH from ConvLSTM
- Coarse L4 SST from REMSS MW-OI product
- Coarse L4 SSS from CMEMS Multi-Obs product

- Surface wind velocities from ERA 5

Evaluation Metrics

We evaluate the mapped SSH fields along-track using the independent withheld satellite altimeter, CryoSat-2, through the metrics provided in Ocean Data Challenge 2021a [Ballarotta et al., 2021]. Specifically, we evaluate the ‘RMSE score’, μ , and the annual standard deviation of its daily mean, where

$$\mu = 1 - \frac{\text{RMSE}(\eta_{pred}, \eta_{observed})}{\text{RMS}(\eta_{observed})} \quad (3.16)$$

and η is the SSH. We also estimate the ‘effective resolution’ of the SSH maps by finding the wavelength at which the signal-to-error ratio of the maps crosses a threshold in spectral space when compared to along-track segments of SSH from CryoSat-2 following the methodology of Ballarotta et al. [2019]. We stress that the effective resolution evaluates both the magnitude and phase of the SSH spectrum. This is different to identifying a wavelength below which the magnitude of the spectrum rolls off due to smoothing. Thus, a state estimate can show significant amplitude at small scales below the effective resolution (e.g. with a power law inertial range spectrum) but these scales are not counted as ‘resolved’ since their phase is not reliably estimated.

To evaluate the SST, we artificially apply additional clouds to the L3 SST observations by randomly shuffling the cloud masks between days and applying the shuffled cloud mask on top of the already-cloudy L3 SST. We then calculate R^2 against the L3 SST using only the pixels that were masked out by the additional cloud masking.

Finally, we compute the dynamical diagnostics outlined in Section 3.3.2 to test if GenDA preserves the dynamical characteristics of GLORYS when applied in the real world.

3.4 Results

Since the focus of this study is state estimation from observations, we here focus only on the results of applying our trained diffusion prior, D , to state estimation through GenDA. The reader is referred to S.I. Text S4 for the results of training D to de-noise GLORYS 12 ocean states. After being trained on 10 million

ocean state vectors drawn randomly from our training domain, requiring 72 NVIDIA V100 GPU-hours, D can skillfully generate realistic ocean states resembling the GLORYS training data (S.I. Text S4, Figures S1 & S2). We now proceed to evaluate GenDA’s state estimation capabilities.

3.4.1 Experiment 1 (OSSE): Performance on Simulated Satellite Observations

Point-wise Accuracy of Predictions

After being assimilated to the simulated L3 and L4 observations (Figure 3.5a-e), GenDA predictions (Figure 3.5f-o) appear to visually match the GLORYS ground truth closely (Figure 3.5p-t). At large scales, the SSH, SST, and SSS predictions match those from the low-resolution L4 observations (panels c-e) both in the ensemble mean (panels f-h) and for a randomly selected individual ensemble member (panels k-m). At smaller scales, the predictions appear to resolve additional fine-scale features missing in the low-resolution L4 inputs. Where high-resolution L3 SSH and SST observations (panels a & b) were assimilated, both the ensemble mean (panels f & g) and ensemble member (panels k & l) predictions match the observations even at small scales. In contrast, where L3 observations are not available, the ensemble mean predictions smooth out small-scale frontal features compared to the ground truth (panels p & q). The ensemble member predictions (panels k & l) do contain frontal features throughout the domain, albeit with less point-wise agreement to the ground truth in the absence of L3 observations. Despite only assimilating low-resolution L4 SSS (panel e), the ensemble mean and ensemble members (panels h & m) exhibit fine-scale frontal SSS features. The ensemble mean SSS frontal features (panel h) appear to be smoothed out where L3 SST observations weren’t available (panel b), emphasizing the close relationship between SST and SSS and suggesting small-scale SSS is largely predictable from L3 SST using GenDA.

Remarkably, despite no observations of ageostrophic surface currents being assimilated, GenDA captures many of the features in the ageostrophic surface current field (compare panels i, j, n, & o to s & t). For surface currents, a clear difference emerges between the ensemble mean (panels i & j) and the ensemble member (panels n & o). The ensemble member exhibits fine-scale frontal jets with variance comparable to the ground truth, though the point-wise agreement to the ground truth for these jets appears mixed. Meanwhile, the ensemble mean exhibits lower variance and only the larger, eddy-driven currents remain. This suggests a higher degree of confidence in the predictions for the large-scale eddy-driven ageostrophic cur-

rents, which are largely due to cyclogeostrophic balance [Penven et al., 2014; Cao et al., 2023], while each ensemble member produces fine-scale frontal jets that vary in location and hence average out in the ensemble mean.

The distinction between the ensemble mean and ensemble member is further emphasized in Figure 3.6 by evaluating the magnitude of SST gradients. Despite the Eastern side of the domain being entirely obscured by clouds (panel d), the GenDA ensemble member (panel b) predicts a series of fine-scale SST fronts whereas the ensemble mean (panel c) predicts broader, weaker fronts reflecting the uncertainty in the location of each front. Note the UNet Regression (panel e) fronts in the cloud-obscured region are qualitatively similar to those in the GenDA ensemble mean, reflecting that the regression formulation leads it to predict the mean of the distribution, smoothing out smaller scales. Both the UNet Regression and GenDA ensemble mean have sharper fronts than those from OI (panel f).

When evaluated over the full withheld testing year, both GenDA and UNet Regression are able to reconstruct SSH, SST, and SSS with higher R^2 than the low-resolution L4 OI (Table 3.1). The GenDA ensemble mean outperforms the ensemble member on these point-wise accuracy metrics, emphasizing that the ensemble provides a more accurate prediction than any individual ensemble member at the expense of smoothing out fine-scale features. UNet Regression outperforms GenDA in terms of point-wise accuracy across all variables, though only marginally for SSH, SST, and SSS compared to the GenDA ensemble mean. This suggests that bespoke training for the considered observing system, and optimization of regression metrics (i.e. MSE) in the UNet Regression loss function, allows UNet Regression to learn a more accurate mapping than the more general GenDA framework, especially for the unobserved ageostrophic surface currents. However, we will show below that the advantages of GenDA become apparent when it is applied to real-world observations and when evaluating the physical consistency across scales of the reconstructions.

R^2 for each variable							
Method	SSH	SST	SSS	u_{ageo}	v_{ageo}	ζ_{geo}	ζ_{ageo}
Low Resolution L4 (OI)	0.969 (0.00)	0.923 (0.00)	0.903 (0.00)	n/a	n/a	0.530	n/a

Supervised UNet Re- gression	0.992 (0.748)	0.964 (0.529)	0.963 (0.623)	0.617	0.591	0.732*	0.436
GenDA (ensemble mean)	0.991 (0.695)	0.957 (0.444)	0.962 (0.608)	0.383	0.369	0.721	0.177
GenDA (ensemble member)	0.987 (0.598)	0.941 (0.241)	0.949 (0.473)	0.017	0.011	0.564	-0.287

Table 3.1: Coefficient of determination, R^2 , for each variable and reconstruction methods, higher values closer to 1 indicate more accurate predictions. Values in brackets for SSH, SST, and SSS are R^2 calculated on the residual from the low resolution L4 OI inputs, so any positive value indicates an improvement over OI. Ageostrophic currents and vorticity are not retrievable from the low-resolution L4 OI method since this only estimates satellite observables. The asterisk for UNet Regression geostrophic vorticity indicates that the field was smoothed using a Gaussian filter with a sigma of 1 pixel before calculating SSH gradients to remove high-frequency checkerboard artifacts, improving R^2 for ζ_{geo} .

Spectral Evaluation & Eddy Dynamics

Evaluating the spectra for the satellite observables (SSH, SST, and SSS) highlights the fact that both GenDA and UNet Regression preserve more fine-scale features than the low resolution L4 OI, with spectra closer to the ground truth GLORYS fields (Figure 3.7a-c). In all cases, UNet Regression appears to better estimate the variance at large scales than GenDA, which likely drives the improvement in R^2 seen in Table 3.1. However, for SST and SSS UNet Regression shows a spectral roll-off, underestimating variance at smaller scales with an overly-steep spectral slope compared to the ground truth. By contrast, the GenDA reconstructions show shallow spectral slopes, more like those in the ground truth data, albeit with the GenDA ensemble mean spectral slope beginning to steepen below 100 km. The fact that GenDA ensemble member reconstructions show similar spectral slopes to the ground-truth fields across all scales highlights the ability of the diffusion prior to ensure reconstructions retain the physical characteristics of the GLORYS training data across scales. At smaller scales, the UNet Regression spectra flatten due to the appearance of high-

frequency ‘checkerboard’ artifacts in the predictions which in the case of SSH overwhelm the signal below 40 km. Such artifacts don’t appear in the GenDA reconstructions likely due to its different training objective (i.e. de-noising).

Considering the kinetic energy spectra of the unobserved ageostrophic currents (Figure 3.7e), UNet Regression appears to better capture the variance at large scales, possibly indicating that GenDA is not optimal for reconstructing large-scale wind-driven surface currents that remain after subtracting the linear Ekman currents. However, UNet Regression again exhibits an overly-steep spectral slope at scales below 100 km where the regression formulation begins to smooth out fine-scale frontal jets. The GenDA ensemble member matches the ground truth spectrum more closely all the way down to the grid resolution. By contrast, the GenDA ensemble mean strongly under-estimates variance below 100 km, consistent with the smooth ensemble mean in Figure 3.5i,j. A similar picture emerges when considering the geostrophic kinetic energy spectra (Figure 3.7d), with the GenDA ensemble member coming closest to the ground truth spectra albeit still underestimating the variance at all scales. This emphasizes the advantage of GenDA’s generative formulation for reconstructing higher-order dynamical diagnostics which are sensitive to the smoothing of SSH (geostrophic kinetic energy goes as the square of the gradient of SSH). Finally, the spectra of the eddy temperature and salt fluxes (Figure 3.7f & g respectively) show how the smoothing of SSH and SST/SSS compound in the low-resolution OI reconstructions, leading to a severe under-estimation of small-scale eddy fluxes. In contrast, GenDA ensemble member fluxes have a similar spectral slope to the ground truth all the way down to the grid resolution.

The ability of GenDA ensemble members to reconstruct realistic-looking eddy dynamics is further illustrated by comparing the geostrophic vorticity-strain JPDF to that calculated from the GLORYS ground truth SSH (Figure 3.8). Compared to the ensemble mean or UNet Regression, GenDA ensemble members better capture the long tails in the vorticity and strain distributions. The improved reconstruction of vorticity and strain can also be observed visually for a snapshot through the Okubo-Weiss quantity (Figure 3.9). Accurately reconstructing geostrophic vorticity and strain, reflected in both the JPDF and the Okubo-Weiss quantity, is vital for studies of ocean eddy dynamics since they are necessary to diagnose ageostrophic dynamics, vertical velocities, frontogenesis rates, and kinetic energy cascades between scales. That GenDA has a vorticity-strain JPDF close to that from a numerical simulation is a strong indicator of ‘physical re-

alism’ since the JPDF is sensitive to whether a state estimate accurately captures the geometry of eddies and fronts, rather than merely distributing variance across scales. We emphasize though that there is no hard physical constraint in the method and future studies testing the extent to which GenDA state estimates respect known conservation principles in a more idealized setting would be valuable.

Advantages of Channel Synthesis

To emphasize the advantage of pursuing a multi-modal approach to surface ocean state estimation, we here test the extent to which GenDA can exploit observations of one variable to improve the predictions of others, i.e. its ‘channel synthesis’ capabilities. Firstly, the fact that GenDA captures significant variance and spatial structure in u_{ageo} and v_{ageo} despite assimilating no surface current observations demonstrates GenDA can exploit the physical relationships between satellite observables and ageostrophic currents learned from GLORYS during training (Figure 3.5 & Table 3.1).

GenDA ensemble mean predictions show increased R^2 across all variables in pixels where cloud-free L3 SST observations were assimilated compared to those occluded by clouds (Figure 3.10a). The improvement is (unsurprisingly) most clear for SST, as GenDA reconstructs almost all the variance in cloud-free pixels, demonstrating how closely the reconstruction converges to observations where provided. The next biggest improvement with cloud-free SST observations is for SSS, with R^2 (calculated on the residual from the low-resolution L4 SSS) almost doubling for cloud-free pixels. This makes it clear that frontal-scale SSS anomalies appear largely predictable given L3 SST and low-resolution L4 SSS. Frontal-scale SST and SSS anomalies are closely related likely because frontogenesis acts to amplify all surface tracer gradients simultaneously. There is also an improvement in R^2 for ageostrophic surface currents, likely due to the improved reconstruction of the strong frontal jets typically co-incident with strong density (and hence SST) gradients. The improvement in SSH is more limited in cloud-free pixels, possibly indicating that larger mesoscale SST signals, captured even in cloud-occluded regions by the low-resolution L4 SST input, drive most of the improvement seen in SSH mapping studies exploiting SSH-SST synergies [Archambault et al., 2023; Fablet et al., 2023, 2024; Martin et al., 2023, 2024]. Similarly, an increase in R^2 is observed across all variables in pixels where L3 SWOT SSH observations were assimilated (Figure 3.10b). SWOT SSH primarily appears to improve the reconstruction of ageostrophic currents and leads to relatively little improvement in

SST or SSS. The improvement in ageostrophic currents is likely due to their relationship to SSH through cyclo-geostrophy [Penven et al., 2014; Cao et al., 2023]. Taken together, Figure 3.10 shows the value in synergizing SSH and SST observations as they appear to provide ‘orthogonal’ information about the ocean state, with SSH containing information about the surface pressure and SST about the surface density.

Suitability of GenDA Ensemble for Uncertainty Quantification

Here we assess whether GenDA ensemble dispersion is a good metric of uncertainty. The spatial patterns of the ensemble dispersion, indicated by the ensemble standard deviation, for SSH and SST show a strong signature of the locations of L3 observations, with very low ensemble dispersion in SSH pixels observed by satellite altimeters and cloud-free SST pixels (Figure 3.11). In addition to the signature of the observing system, the SSH ensemble dispersion also appears elevated near strong SST fronts in the L3 observations. Strong SST fronts are typically associated with strong gradients in SSH, but there could be a phase shift between SSH and SST. So the elevated SSH uncertainty at well-observed SST fronts could indicate that the model has uncertainty about the SSH-SST phase shift. However, ‘u-shaped’ rank histograms for all variables show that the GenDA ensemble is under-dispersive (Figure 3.12). Concretely, the uncertainty estimate provided by the ensemble dispersion is an underestimation of the true predictive uncertainty: with the 24-member ensemble considered here, there is an up to 20% chance that the true value falls outside the range of ensemble members (adding the left and right bars in the rank histograms in Figure 3.12). Nonetheless, GenDA ensembles show no significant systematic biases even for the unobserved surface ageostrophic currents, as indicated by the symmetry of the rank histograms (Figure 3.12). Thus, despite slightly underestimating the spread, GenDA does provide a realistic-looking unbiased ensemble of eddy-scale predictions, representing an important advance compared to regression-type approaches.

3.4.2 Experiment 2 (OSE): Real-World Mapping Performance

SSH & SST Mapping Metrics

Assimilating real-world observations (Figure 3.13a-e), GenDA state estimations qualitatively look remarkably similar to those in the OSSE setting, encouraging confidence in its generalizability to real-world observations (Figure 3.13k-t). By contrast, the UNet Regression predictions (Figure 3.13f-j) appear significantly

degraded, with unrealistic SST and SSS fields and ageostrophic surface current fields with significantly stronger variance than seen in GLORYS (Figure 3.5s,t). UNet Regression behaving poorly on real-world observations suggests that the network is overly sensitive to differences between real observations and the GCM pseudo-observations used in training, likely due to different noise properties and resolutions. GenDA ensemble member predictions (panels p-t) show enhanced variance at small scales, with realistic-looking frontal features in SST, SSS, and ageostrophic currents that are mostly absent in the low-resolution L4 gridded products assimilated (panels c-e).

The GenDA ensemble mean and ensemble members exhibit performance in SSH mapping metrics comparable to state-of-the-art methods (Figure 3.14a and 3.6). This is perhaps not surprising since the ConvLSTM SSH maps were assimilated, but implies that adding four more variables to the state vector and requiring GLORYS-like dynamics does not lead to a significant drop in SSH mapping performance. By contrast, despite using the ConvLSTM SSH maps as input, UNet Regression shows severely degraded SSH mapping and fails to accurately resolve wavelengths as small as the DUACS OI method [Le Traon et al., 1998; Taburet et al., 2019]. GLORYS 12 shows significantly worse SSH mapping errors, with its smallest resolved wavelength being more than double that of GenDA. This highlights the limitations of full 3D GCM data assimilation for resolving mesoscale surface eddies and the need for surface-only methods like GenDA. A similar picture emerges when assessing the SST mapping, with GenDA outperforming OI, UNet Regression, and GLORYS 12 (Figure 3.14b,c). The reduced R^2 in SST compared to in the OSSE setting are unsurprising since here our ground truth are L3 observations with significant sensor noise whereas in the OSSE we compared directly to the noise-free ground truth.

Spectral Evaluation & Eddy Dynamics

The spectra of SSH, SST, SSS, and geostrophic kinetic energy, after assimilation to real-world observations, show significant shifts from those of the GLORYS training data and the OSSE state estimates (Figure 3.15a-d). While in the OSE setting we don't have access to the ground truth spectra, the significant shift from GenDA (OSSE) to GenDA (OSE), especially at larger scales, suggests the assimilation process allows the state estimates to adapt to real-world ocean conditions that are substantially different from the GLORYS training data. The GenDA reconstructions continue to exhibit power-law spectral slopes qualitatively similar

to those in the OSSE setting and shallower than the L4 OI inputs, implying the diffusion prior learned from GLORYS ensures GenDA preserves realistic fine-scale features. By contrast, the UNet Regression spectra show substantially different shapes to those in the OSSE, further suggesting a lack of generalizability to real-world observations.

The GenDA spectra for ageostrophic kinetic energy remain qualitatively similar to those seen in the OSSE setting (Figure 3.15e), further increasing confidence that the GenDA reconstructions remain ‘GLORYS-like’ even after assimilating real-world observations. Intriguingly, despite no observations of ageostrophic surface currents being assimilated, the ageostrophic kinetic energy spectrum shows enhanced energy at large scales when assimilated to real-world satellite observations (Figure 3.15e). This suggests even the unobserved variables respond to changes in the spectral properties of the assimilated variables, and the change in large-scale ageostrophic kinetic energy could be a response to changes in the SSH spectrum (Figure 3.15a & d) through cyclogeostrophic balance (S.I. Text S7 & Figure S6) [Penven et al., 2014; Cao et al., 2023]. The GenDA eddy flux spectra (Figure 3.15f,g) maintain their ‘GLORYS-like’ spectral slopes even when assimilated to real-world observations. Finally, the reconstructed geostrophic Okubo-Weiss fields from GenDA retain their qualitative realism from the OSSE setting (Figure 3.16), unlike the UNet Regression Okubo-Weiss which appears to deteriorate substantially. The significant differences in eddy placements in GLORYS 12 (Figure 3.16a) further highlight that full 3D GCM data assimilation and surface-only data assimilation yield substantially different results, with 3D state estimation coming at the expense of accurate placement of eddies at the surface.

Evaluating the accuracy of the ageostrophic surface currents in the real-world setting is challenging due to the sparsity of in situ observations. Nonetheless, we include a comparison of the predicted surface currents to surface drifter observations from the Global Drifter Program [Copernicus Marine in situ TAC, 2024] (S.I. Text S8 and Table S4). All methods (GenDA, UNet Regression, and GLORYS) show a slight increase in error when the predicted ageostrophic current is included compared to when just the geostrophic component is evaluated (Table S4). This highlights the challenge of accurately predicting real-world ageostrophic surface currents and suggests further work is needed to skillfully predict real-world ageostrophic surface currents from satellite observables. Nonetheless, the surface drifters further highlight that GenDA generalizes to real-world observations more reliably than UNet Regression, showing significantly lower surface

current errors than both UNet Regression and GLORYS and similar accuracy to the geostrophic currents from ConvLSTM (SSH-SST) (Table S4). It should be noted that these error metrics are calculated against just thirty available drifter trajectories which only sparsely cover the prediction domain and more comprehensive evaluation using Lagrangian rather than Eulerian metrics would be beneficial in future.

3.5 Conclusions & Discussion

3.5.1 Conclusions

Score-based data assimilation [Rozet and Louppe, 2023a,b], referred to here as GenDA, is a significant methodological departure from the regression approaches that have to date been explored in deep learning satellite oceanography studies. By decoupling the neural network training from any specific observing system, GenDA provides a flexible and robust framework for guiding inference using a simulation-informed neural network with observations. GenDA improves the generalization from simulation training to real-world observations, retaining many dynamical characteristics of the simulation used during training. Enabling the use of multi-modal and multi-resolution observations resulted in significant benefits, with GenDA being now able to reconstruct ageostrophic currents and frontal-scale SSS from real-world satellite observations in an unguided manner (i.e. without any observations of currents or high-resolution salinity). By focusing solely on surface state estimation, GenDA reconstructions outperform state-of-the-art dynamical data assimilation systems, resolving SSH wavelengths twice as small as GLORYS 12 (Figure 3.14a). GenDA thus appears to provide a surface-only, neural alternative to 3D dynamical GCM data assimilation.

3.5.2 Discussion

While the results presented in this study represent a significant advance in eddy-resolving surface ocean state estimation, a number of limitations and areas for future method development should be noted. Our implementation of GenDA here works only on single temporal snapshots, meaning the GenDA reconstructions are not temporally coherent. GenDA predictions (both ensemble members and mean) for consecutive days show large jumps between time steps, especially for the high-resolution features that are only intermittently observed with L3 SSH and SST observations (S.I. Movies S1 & S2). Future work should therefore

focus on extending the method to work on multiple time steps, potentially following the method proposed by Rozet and Louppe [2023a]. Extending GenDA to assimilate multiple time steps may remove the need to assimilate low-resolution L4 satellite products in the observation operator by improving the geographic coverage of the L3 observations assimilated. Another limitation of the results presented here is the under-dispersiveness of the ensemble predictions. Future work could explore strategies for improving the realism of the reconstructed distribution, for example, through alternative reverse diffusion time-stepping schemes or data pre-processing steps. Since real-world OI products are not necessarily well-represented as a Gaussian coarse-graining operation, a promising direction for improving the spectral properties of GenDA state estimates would be to investigate the use of alternative filtering operations in the observation operator. Such efforts would be aided by studies of the spectral properties of low-resolution L4 OI products and how these can be approximated as filtering operations.

A fundamental limitation of the GenDA methodology is that the state estimate inherits its dynamical characteristics from the GCM simulation used during training. GCMs are imperfect representations of real-world ocean physics, and it is unclear to what extent GenDA would be able to ‘un-learn’ a simulation’s biased physics during assimilation to observations. Since GenDA reconstructions are only as reliable as the GCM used during assimilation, one could consider it as a neural network analog to conventional data assimilation. Note that the high accuracy of the 2D surface ocean reconstruction is partially achieved by not formally satisfying GCM’s 3D equations of motion, which themselves are known to be imperfect. For example, submesoscale-resolving SSH observations from SWOT are beginning to reveal that even frontier submesoscale-permitting ocean GCMs like LLC4320 [Su et al., 2018] can have significant statistical biases, including under-estimation of the real-world ocean’s submesoscale kinetic energy [Archer et al., 2025]. The statistical discrepancies between simulation data and real-world observations, which are significant even in high-resolution models, emphasize the need to explore observation-only learning strategies for training high-resolution diffusion priors [Rozet et al., 2024].

There are many promising potential future research directions the GenDA method could enable. We here trained our diffusion prior on GLORYS 12 reanalysis data, but the grid resolution of the GCM used is too coarse to resolve submesoscale dynamics. In future, the GenDA diffusion prior could be trained on submesoscale-permitting free-running GCM’s such as the $1/48^\circ$ LLC4320 global simulation [Su et al., 2018]

or the $1/60^\circ$ NATL60 North Atlantic simulation [Ajayi et al., 2020]. This would offer the tantalizing prospect of ‘assimilating’ frontier submesoscale GCMs to satellite observations, something that is beyond the current capabilities of conventional dynamical data assimilation. Training on submesoscale-permitting models could improve the realism of frontal dynamics by incorporating mixed layer instabilities into the training data, leading to potential improvements in our monitoring and understanding of ocean energy cascades [Sasaki et al., 2014; Klein et al., 2019; Taylor and Thompson, 2023]. Training on submesoscale-permitting models could also help to reduce the error of the predicted ageostrophic currents compared to surface drifter observations (Table S4) since the ageostrophic currents in the GLORYS training data used in this study likely differ substantially from those in the real-world ocean where submesoscale instabilities are active. Submesoscale-permitting model training data, combined with a multi-timestep approach, could also enable GenDA in future to disentangle the effects of internal gravity waves and balanced submesoscale turbulence on SSH [Wang et al., 2022a; Gao et al., 2024], a critical science requirement for the SWOT mission [Klein et al., 2019; Fu et al., 2024]. Even when trained on the relatively coarse GLORYS model, SWOT SSH observations appear to drive a significant improvement in the estimation of ageostrophic surface currents (Figure 3.10b). We anticipate that the added value of SWOT could be even greater if the GenDA diffusion prior were re-trained on a more realistic submesoscale-permitting GCM that better captures the relation between SSH and ageostrophic currents. The ability of GenDA to skillfully predict frontal scale SSS given cloud-free L3 SST and low-resolution L4 SSS from OI could in future be used to determine the extent to which SST fronts are salinity compensated [Rudnick and Ferrari, 1999], since SSS is not observed at frontal scales from satellites and theories of frontogenesis relate to density, not just SST [Hoskins, 1982]. The flexibility to prescribe different observation operators for different observing systems make GenDA well-placed to exploit future proposed satellite missions directly measuring ageostrophic surface currents from space [Torres et al., 2023], to better exploit wide-swath altimetry [Fu et al., 2024], and to aid in future observing system design. Future studies could also consider how to assimilate other observing streams not typically assimilated in dynamical data assimilation, such as synthetic aperture radar (SAR), which contains signatures of both ocean turbulence and internal gravity waves [Ivanov and Ginzburg, 2002], or satellite ocean color observations, which show an abundance of submesoscale fronts and filaments [Lévy et al., 2018]. These observations are not immediately amenable to assimilation through GenDA since these variables are

not typically modeled in the ocean GCMs available for neural network training, and so new methods would be required to either model these variables within the GCM itself or to predict them as a post-processing step using GCM output. While GenDA provides a natural method for uncertainty quantification, further work is needed to improve the calibration of the predicted ensemble. This could be potentially achieved either through generating larger ensembles or by systematically tuning the observational noise for each observing modality in the observation likelihood term. Finally, the ability of GenDA to reconstruct un-observed variables in an unguided manner suggests future studies could consider more ambitious dynamical quantities of interest in the state vector such as vertical velocities [Zhu et al., 2023; He and Mahadevan, 2024], air-sea fluxes, or interior ocean dynamics [George et al., 2021; Manucharyan et al., 2021].

3.6 Complete Ocean Data Challenge 2021a SSH Evaluation Metrics (OSE)

Ocean Data Challenge 2021a SSH Mapping Metrics				
Method	RMSE Score, μ ($\uparrow$)	std of RMSE Score ($\downarrow$)	Effective Reso- lution [km], λ_{eff} ($\downarrow$)	Notes
GenDA (ensemble mean)	0.90	0.06	109	Trained on GLORYS, reconstructs: SSH, SSS, SST, u, v
GenDA (ensemble member)	0.90	0.07	113	Trained on GLORYS, reconstructs: SSH, SSS, SST, u, v
UNet Re- gression	0.85	0.08	188	Trained on GLORYS, reconstructs: SSH, SSS, SST, u, v
GLORYS 12	0.73	0.13	240	1/12° GCM data assimilation (evaluation altimeter assimilated)

DUACS	0.88	0.07	152	OI as used in CMEMS product (SSH only)
MIOST	0.89	0.08	139	Dynamics-informed OI (SSH only)
DYMOST	0.89	0.06	129	Dynamics-informed OI (SSH only)
BFN QG	0.88	0.06	122	Data assimilation with QG model (SSH only)
4DVarNet (v2021)	0.89	0.06	122	Data assimilation-inspired neural network regression (SSH only)
4DVarNet (v2022)	0.89	0.09	109	Data assimilation-inspired neural network regression (SSH only)
4DVarQG	0.90	0.06	106	Data assimilation with QG model (SSH only)
ConvLSTM (SSH)	0.90	0.06	113	Neural network regression (SSH only)
ConvLSTM (SSH-SST)	0.90	0.06	100	Neural network regression (SSH & SST inputs, SSH output)
NeurOST (SSH-SST)	0.90	0.06	114	Neural network regression as used in PO.DAAC product (SSH & SST inputs, SSH output)

Table 3.2: Ocean Data Challenge 2021a SSH mapping metrics for the OSE. See Section 3.3.3 for metric details.

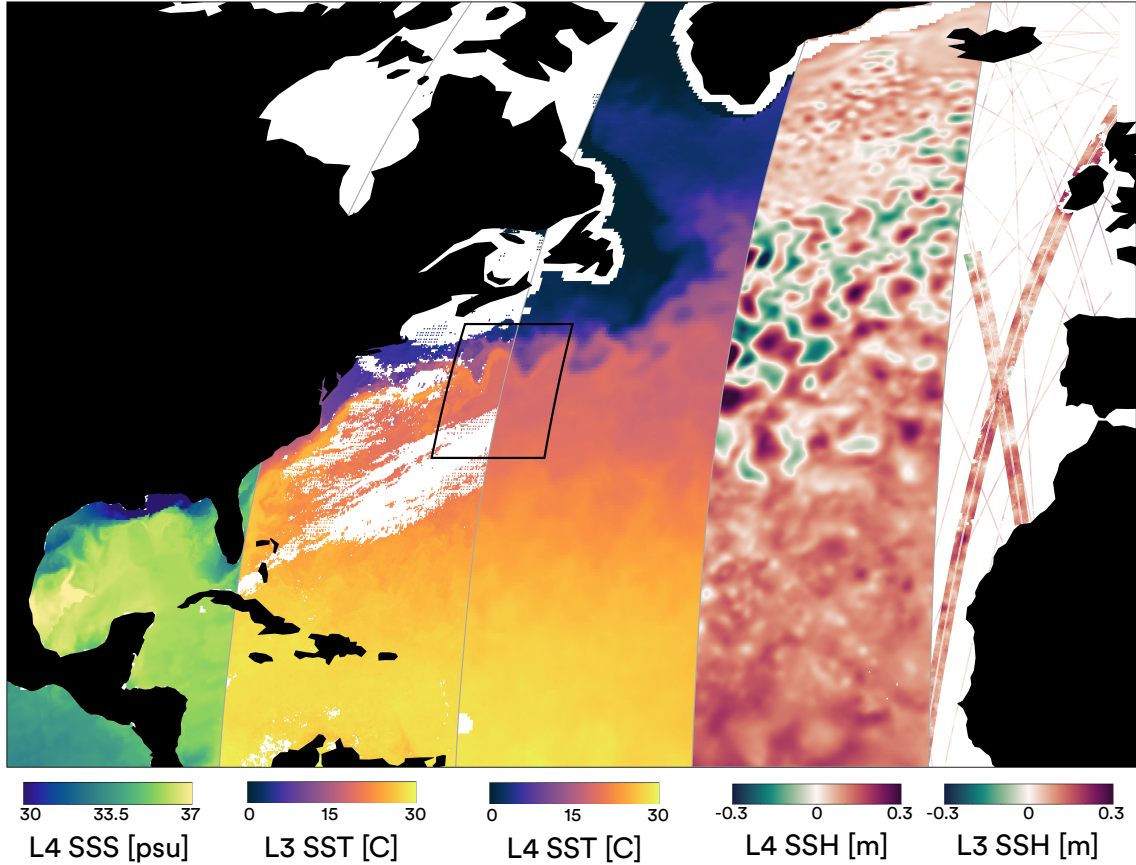


Figure 3.1: Illustration of the available multi-modal satellite observations of the surface ocean state. Satellite observations are available at both ‘Level 3’ (L3) and ‘Level 4’ (L4) processing levels. L3 observations are before any interpolation, so are sparse but high-resolution, whereas L4 has been interpolated to a gap-free grid, smoothing smaller scales (e.g. using objective analysis). From left to right: L4 SSS from E.U. Copernicus Marine Service Information (CMEMS) [2024g], L3 SST from E.U. Copernicus Marine Service Information (CMEMS) [2024a], L4 SST from Remote Sensing Systems [2017], L4 SSH from OSTST [2024], and L3 SSH from both nadir [E.U. Copernicus Marine Service Information (CMEMS), 2024d] and SWOT [AVISO/DUACS, 2024] altimeters. Observations shown here are for 2024-04-05. The boxed region is the study region considered in our experiments.

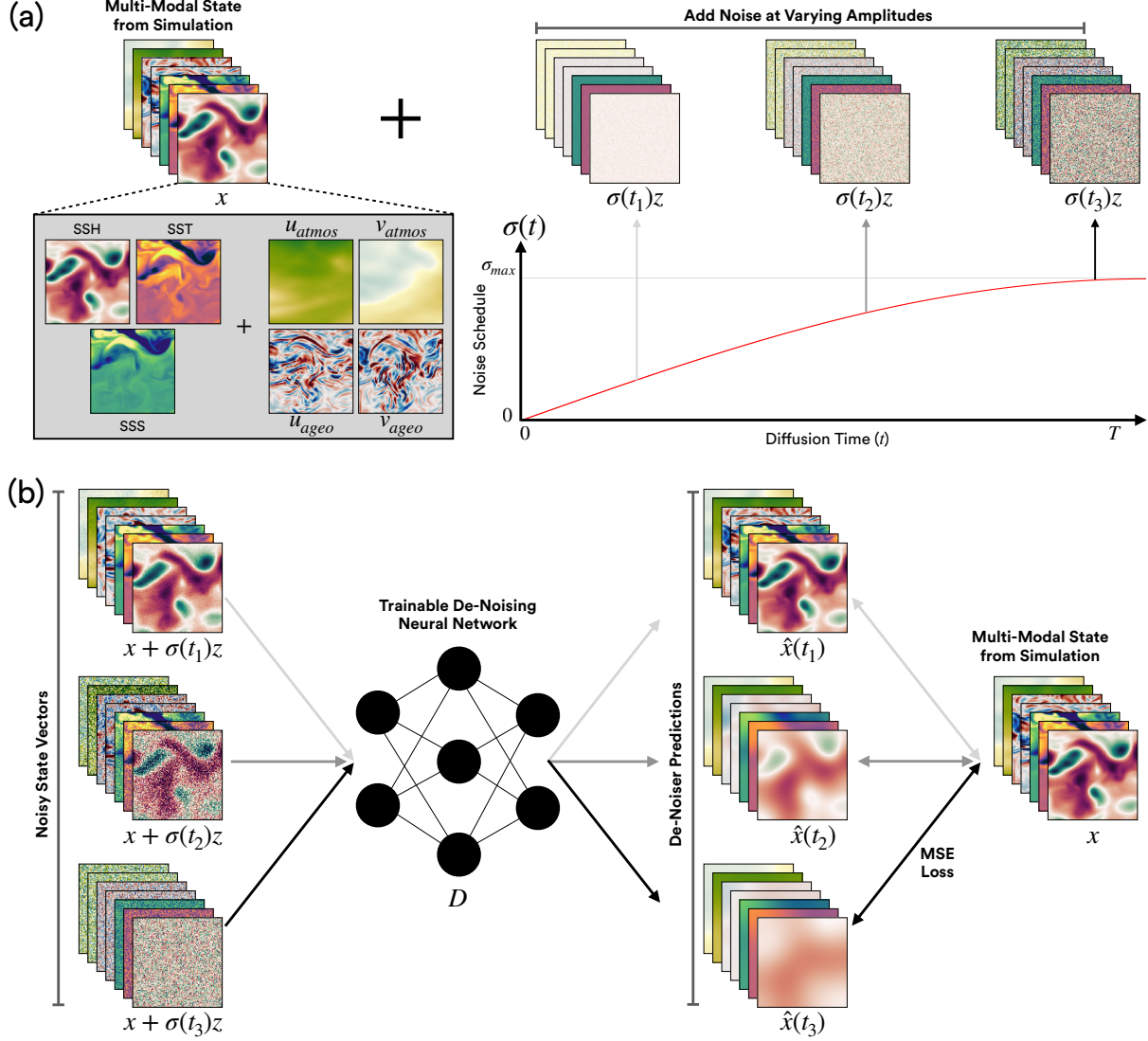


Figure 3.2: Schematic of the GenDA training phase. (a) Gap-free, multi-modal state vectors, x , are taken from the simulation training data and Gaussian white noise, $\sigma(t)z$, at varying amplitudes is added. (b) A de-noising neural network, D , is trained to map from noisy states, $x + \sigma(t)z$ to de-noised states, $\hat{x}(t)$, by minimizing MSE between $\hat{x}(t)$ and x .

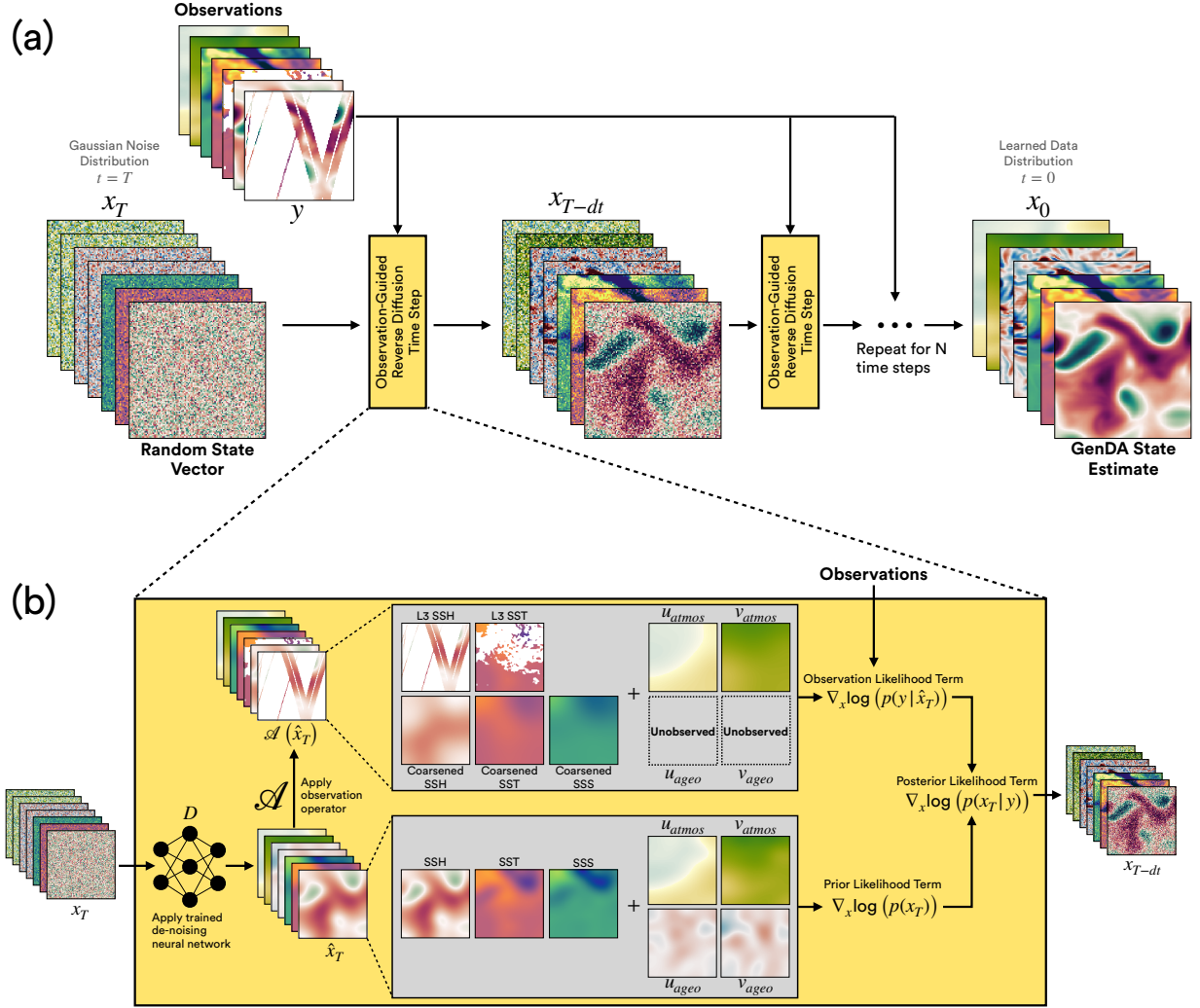


Figure 3.3: Schematic of the GenDA inference phase. (a) A random state, x_T , is mapped to a state estimate, x_0 , through repeated observation-guided reverse time steps, reversing a forward noise process (Section 3.2.1), along the diffusion ‘time’ axis from $t = T$ (noise distribution) to $t = 0$ (learned data distribution). An ensemble of state estimates is generated by inputting different random states, x_T . (b) Within each reverse time step, the trained de-noising neural network, D , is applied, predicting a de-noised state, $\hat{x}_t$, which is compared to observations, y , through the observation operator, $\mathcal{A}$. Combining the prior likelihood gradient from D , $\nabla_x(p(x_t))$, with the observation likelihood term, $\nabla_x(p(y|\hat{x}_t))$, a reverse time step is made using $\nabla_x(p(x_t|y))$ to maximize the posterior likelihood. The weights of D are kept fixed throughout this process.

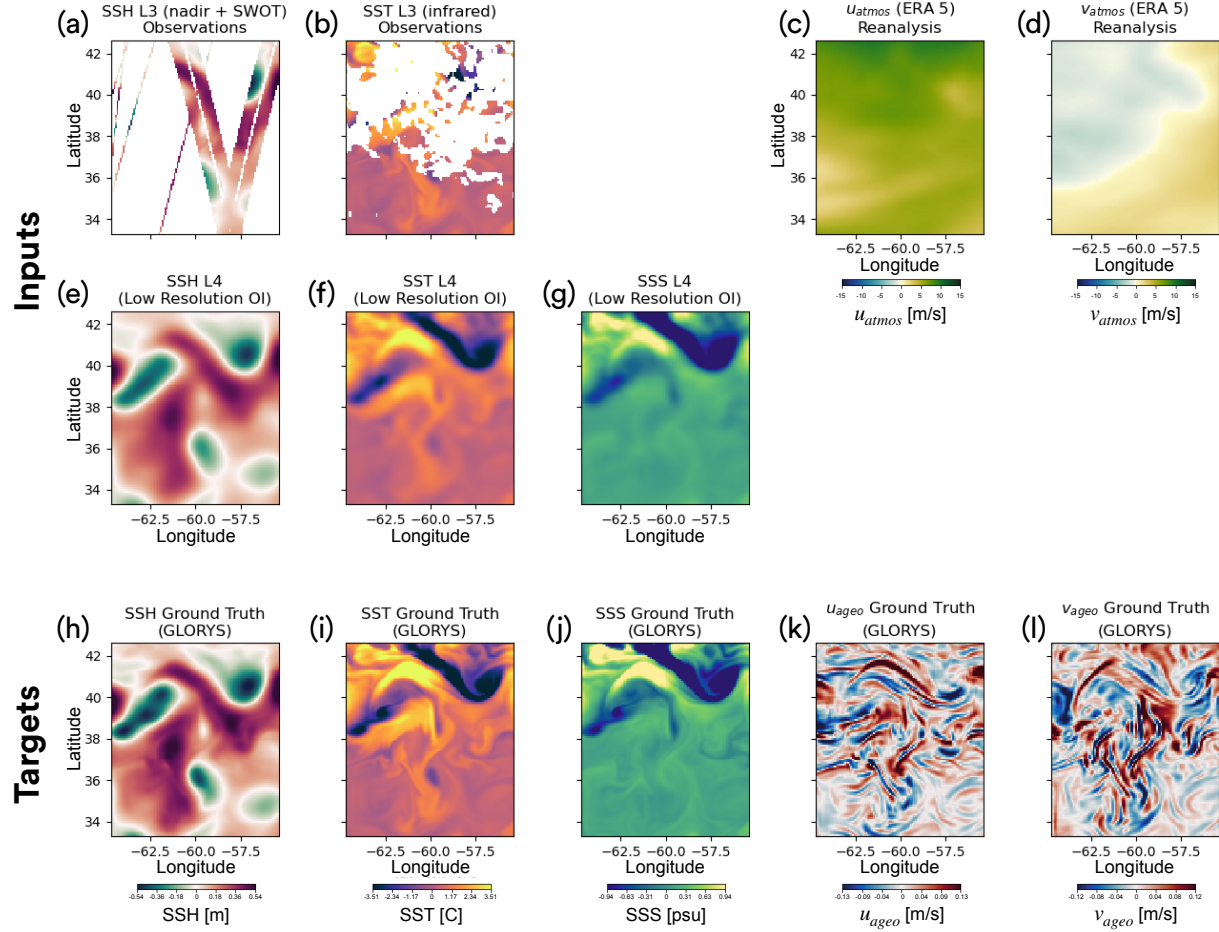


Figure 3.4: Example inputs and targets for multi-modal surface ocean state estimation simulated from GLORYS re-analysis product in our observing system simulation experiment (Section 3.3.2). The inputs are (a) simulated L3 SSH observations from a constellation of nadir altimeters and SWOT, (b) cloud-obscured L3 SST observations from infrared radiometers, (c) & (d) surface winds from ERA 5 (Section 3.2.1), (e)-(g) coarse-grained SSH, SST, and SSS to simulate existing low-resolution satellite products from OI. The target variables are the gap-free, high-resolution (h) SSH, (i) SST, (j) SSS, (k) & (l) ageostrophic surface currents. Ageostrophic currents are not observed by satellites. All values are plotted in anomaly space (Section 3.3.1).

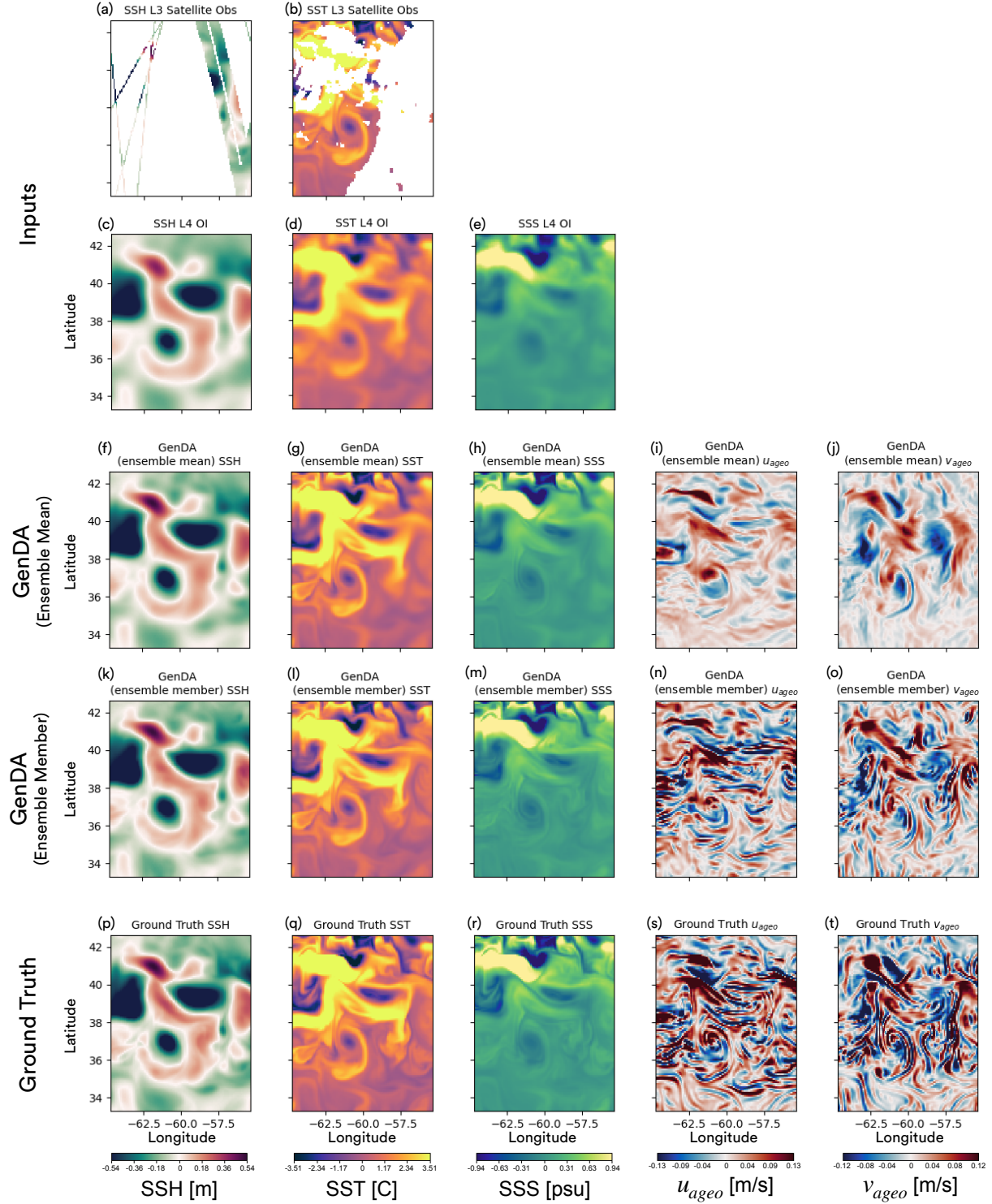


Figure 3.5: Example GenDA OSSE prediction for 2017-04-27 (See S.I. Movie S1 for all dates). First row, input L3 observations assimilated: (a) SSH & (b) SST. Second row, input low-resolution L4 satellite products assimilated: (c) SSH, (d) SST, & (e) SSS. Third row, GenDA ensemble mean predictions for (f) SSH, (g) SST, (h) SSS, (i) u_{ageo} , and (j) v_{ageo} . Fourth row, (k)-(o) GenDA ensemble member predictions. Fifth row, (p)-(t) ground truth fields (GLORYS). Note ERA 5 surface winds are also provided as input and predicted by GenDA but are not plotted since they match almost exactly.

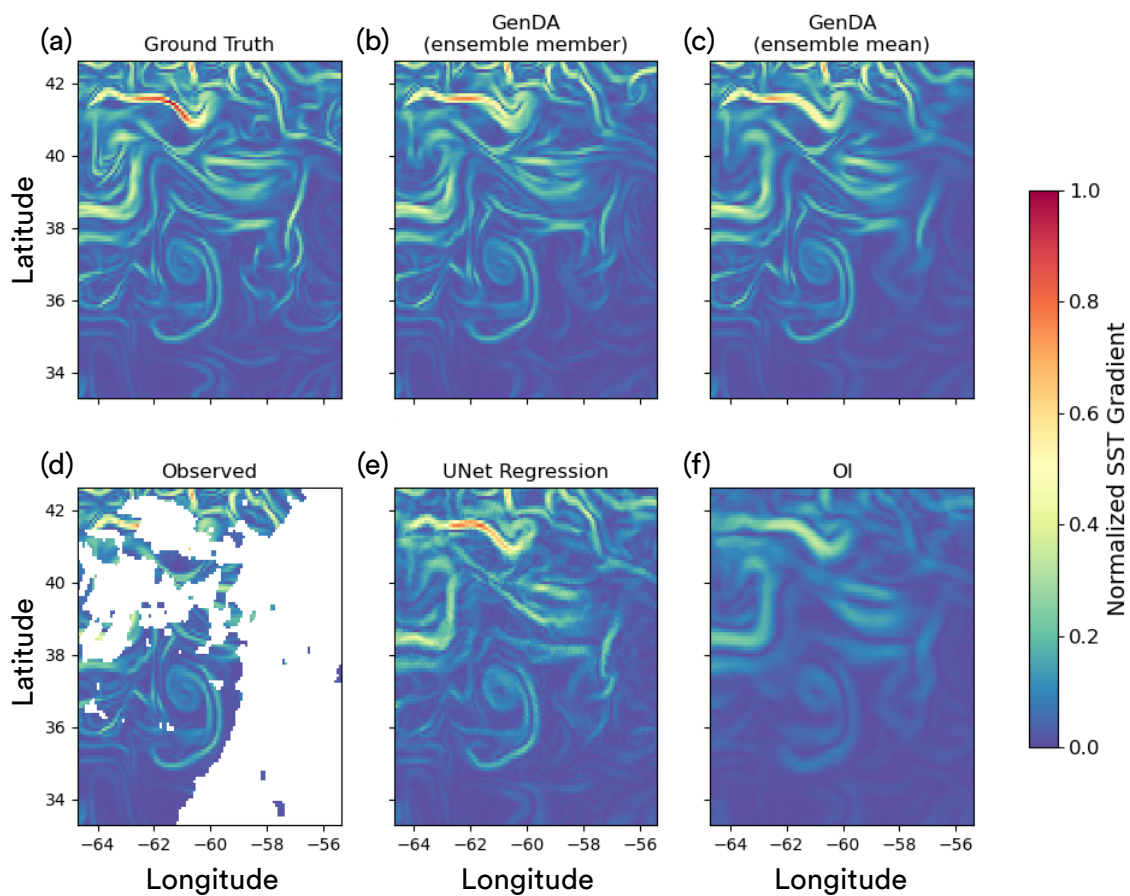


Figure 3.6: Magnitude of SST gradient, normalized by dividing by the maximum ground truth SST gradient, from (a) ground truth target field from GLORYS, (b) GenDA ensemble member prediction, (c) GenDA ensemble mean prediction, (d) cloud-obscured L3 SST observations assimilated, (e) UNet Regression, and (f) the coarse-grained L4 'OI' input assimilated.

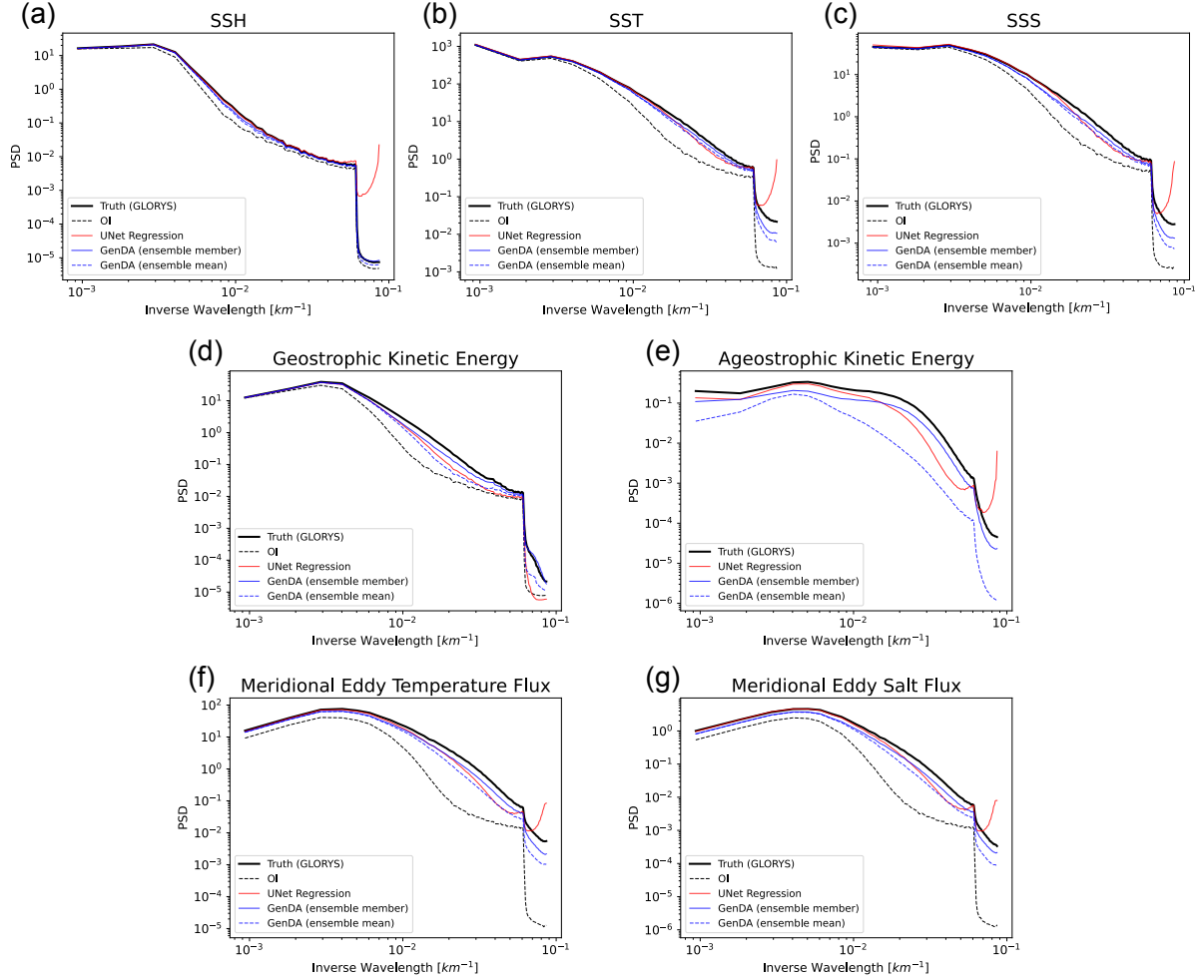


Figure 3.7: Spectra for (a) SSH, (b) SST, (c) SSS, (d) geostrophic kinetic energy, (e) ageostrophic kinetic energy, (f) meridional eddy temperature flux, and (g) meridional eddy salt flux. (solid black) ground truth target field from GLORYS, (dashed black) low resolution L4 OI input assimilated, (solid red) UNet Regression, (solid blue) GenDA ensemble member, and (dashed blue) GenDA ensemble mean.

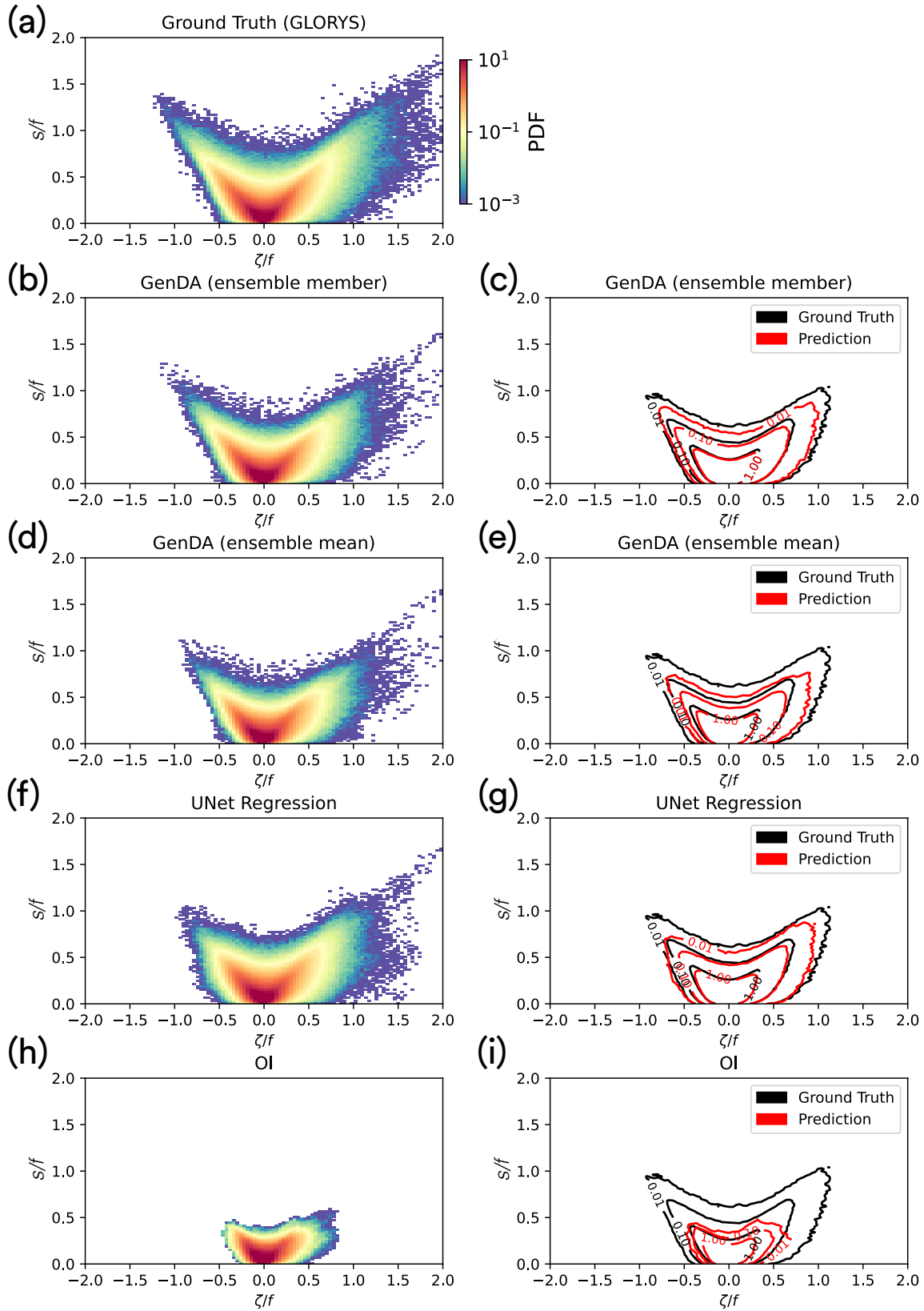


Figure 3.8: Geostrophic joint vorticity-strain probability density functions for (a) ground truth (GLORYS), (b), GenDA (ensemble member), (d) GenDA (ensemble mean), (f) UNet Regression, and (h) OI. Right panels show a comparison between the ground truth (black) and predicted (red) JPDPs at three logarithmically spaced contour levels for (c) GenDA (ensemble member), (e) GenDA (ensemble mean), (g) UNet Regression, and (i) OI.

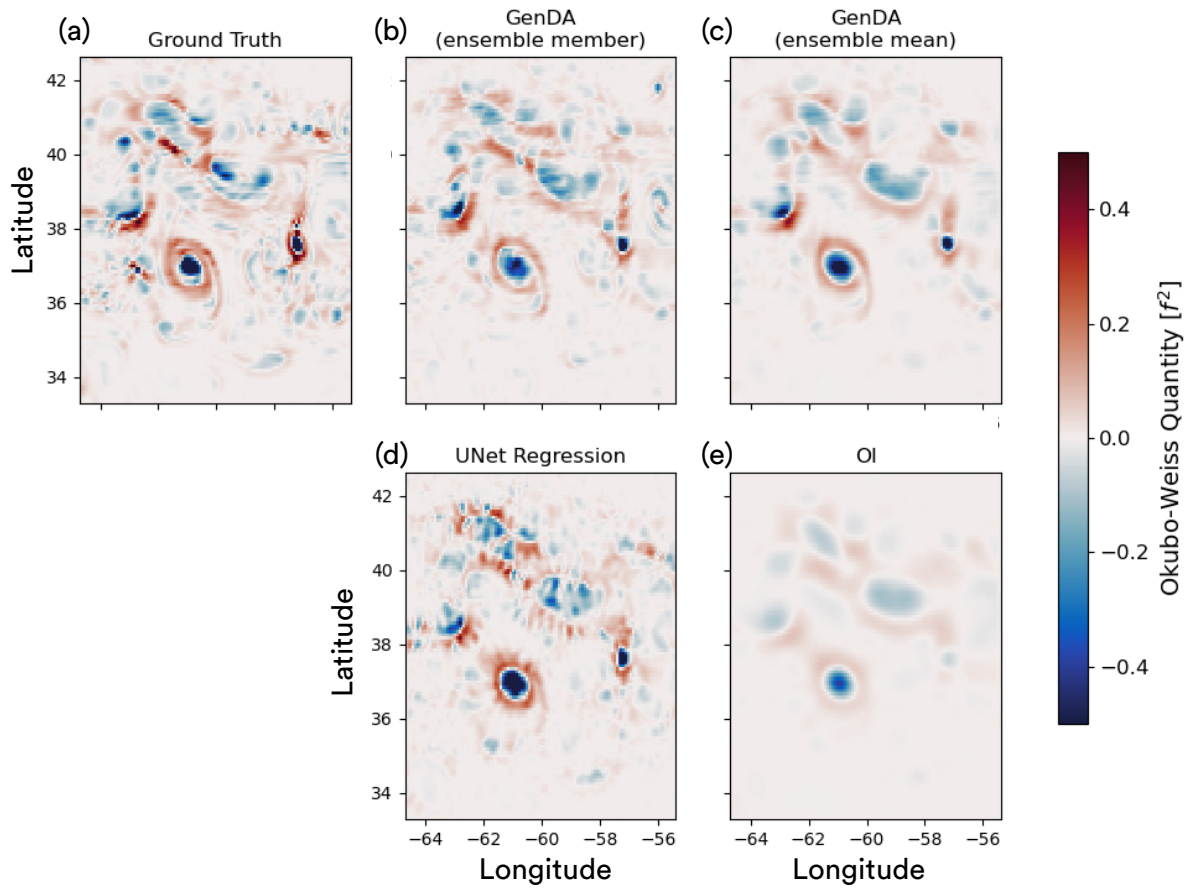


Figure 3.9: Geostrophic Okubo-Weiss quantity for (a) ground truth target field from GLORYS, (b) GenDA ensemble member, (c) GenDA ensemble mean, (d) UNet Regression, and (e) OI. Negative values indicate vorticity dominates (as in eddy cores) whereas positive values show where strain dominates (as around eddy peripheries).

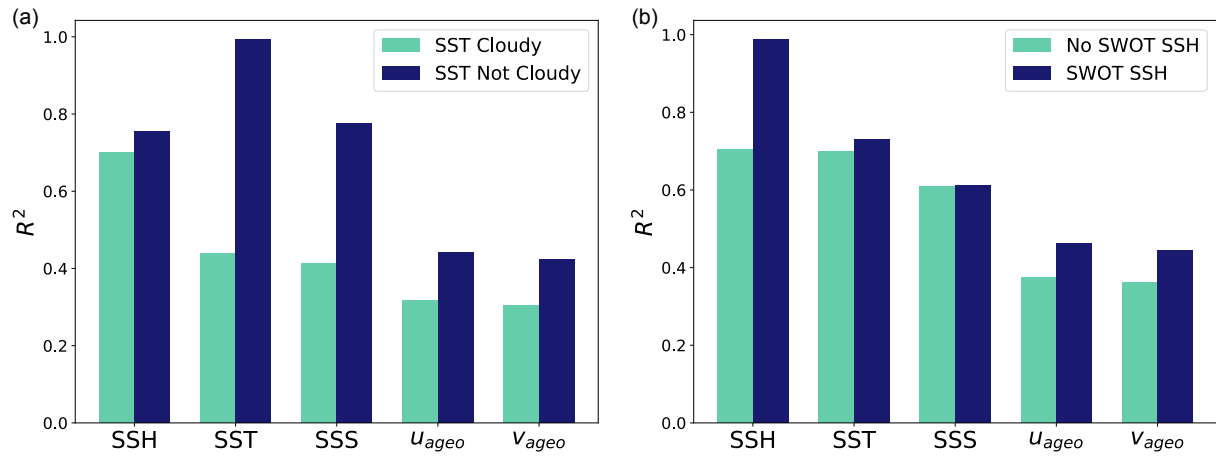


Figure 3.10: R^2 for GenDA ensemble mean OSSE predictions split by (a) whether cloud-free L3 SST observations were (dark blue) available or (teal) unavailable and (b) whether SWOT L3 SSH observations were (dark blue) available or (teal) unavailable. R^2 values for SSH, SST, and SSS were calculated on the residual from the OI predictions, so any positive value indicates an improvement over OI.

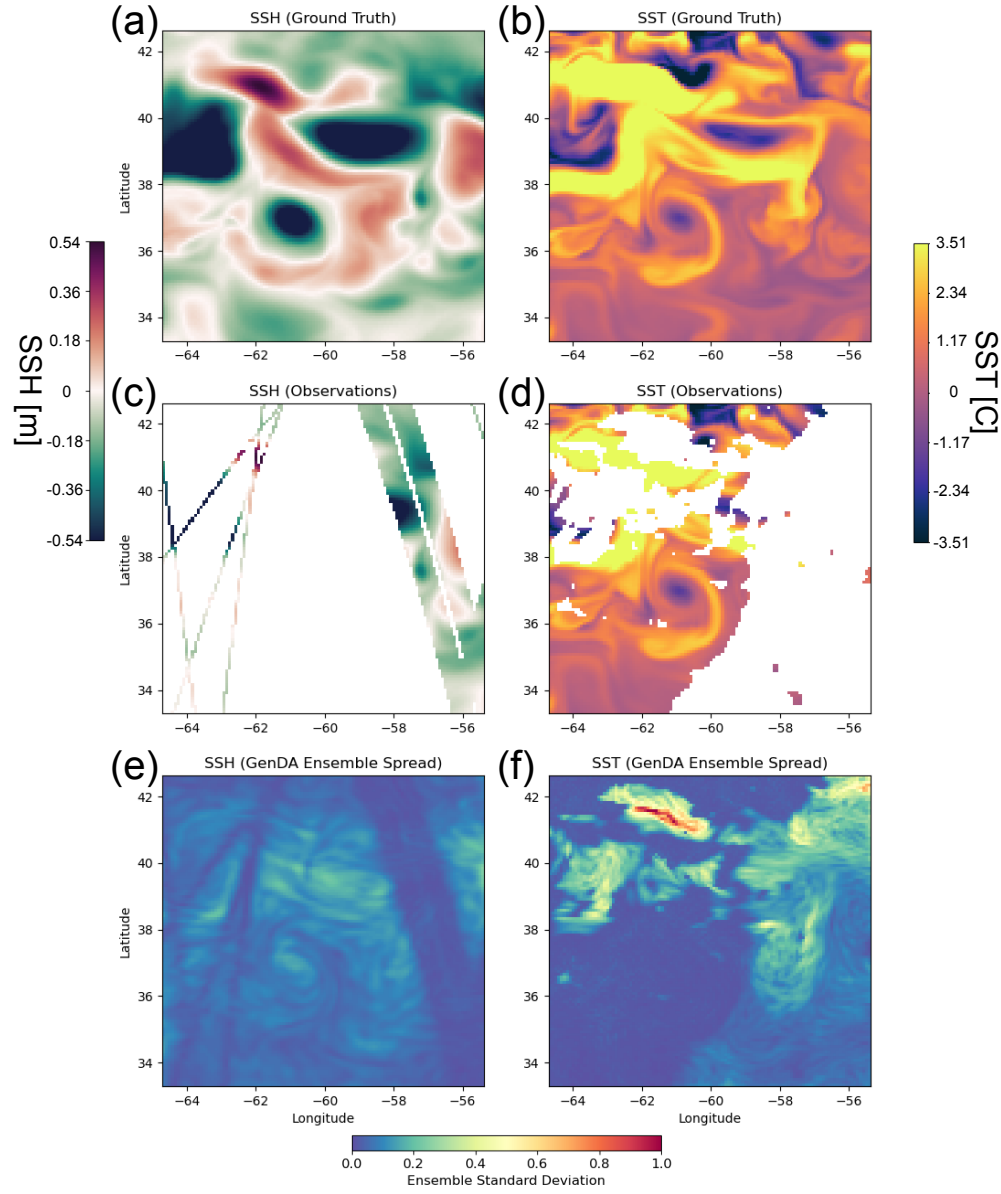


Figure 3.11: Ground truth target fields from GLORYS for (a) SSH and (b) SST. Simulated L3 observations assimilated for (c) SSH and (d) SST. GenDA ensemble standard deviation for (e) SSH and (f) SST.

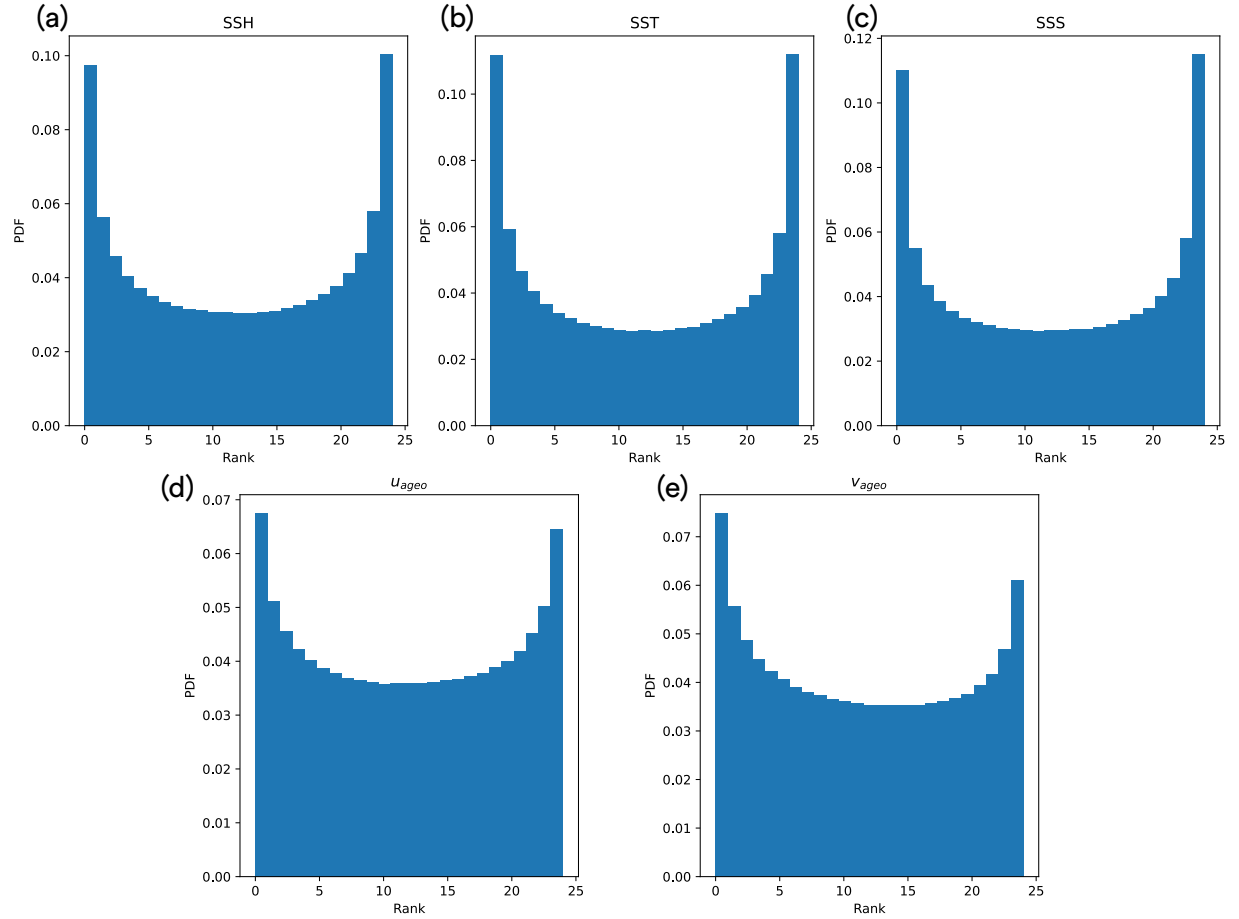


Figure 3.12: Rank histograms for GenDA OSSE predictions of (a) SSH, (b) SST, (c) SSS, (d) u_{ageo} , & (e) v_{ageo} . The x-axis represents the ordered ensemble members, while the y-axis is the probability density for the true value falling in each ensemble member bin. The first (last) bins represent truth values that lie below (above) the ensemble minimum (maximum). A well-calibrated ensemble has a flat-rank histogram.

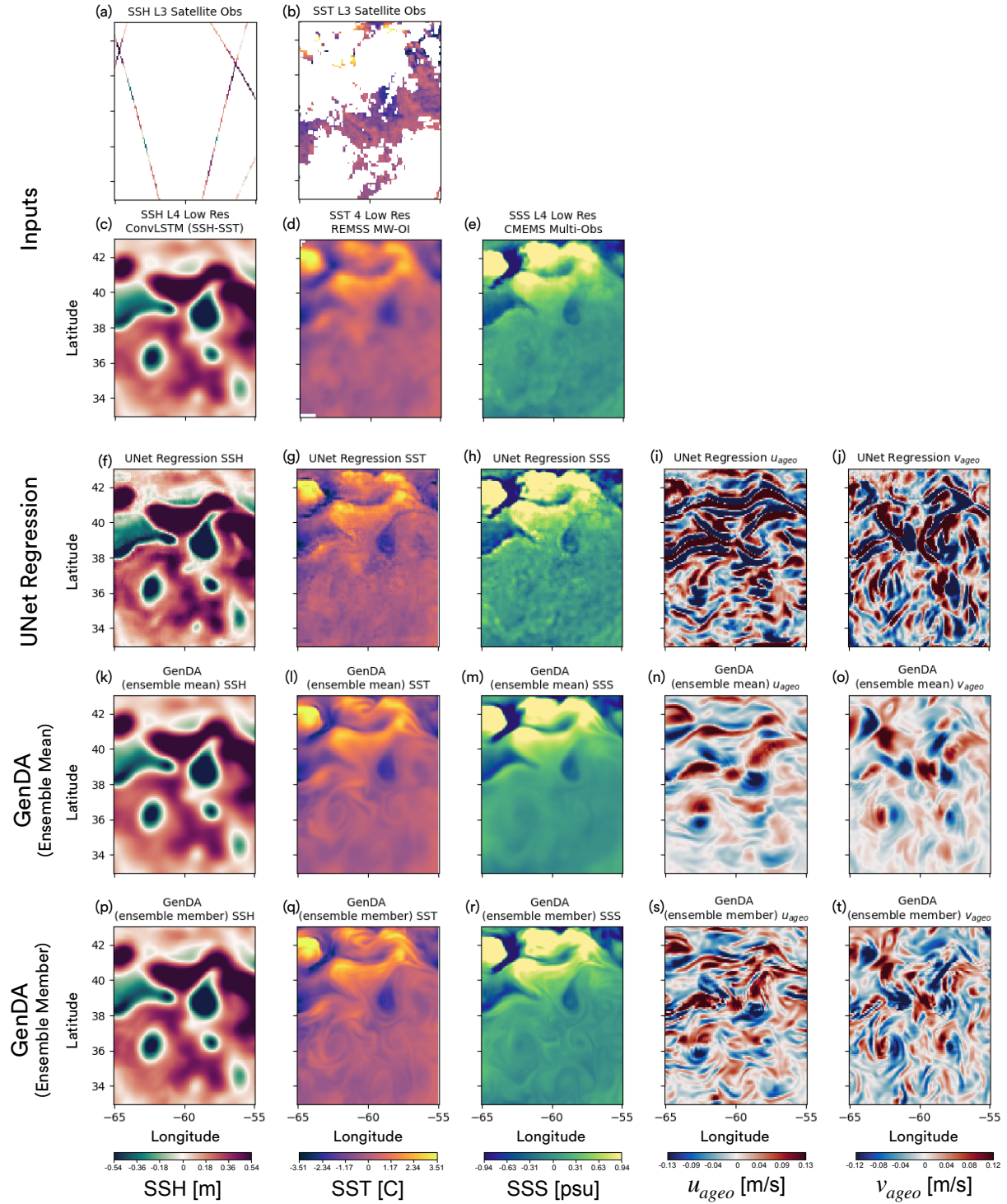


Figure 3.13: Example predictions for 2017-07-07 when assimilating real-world satellite observations in Experiment 2 (See S.I. Movie S2 for all dates). First row, input L3 observations assimilated: (a) SSH & (b) SST. Second row, input low-resolution L4 satellite products assimilated: (c), (d), & (e) SSH, SST, & SSS respectively. Third row, UNet Regression predictions for (f) SSH, (g) SST, (h) SSS, (i) u_{ageo} , and (j) v_{ageo} . Fourth row, (k)-(o) GenDA ensemble mean predictions. Fifth row, (p)-(t) GenDA ensemble member predictions. Note ERA 5 surface winds are also provided as input and predicted by GenDA but are not plotted since they match almost exactly.

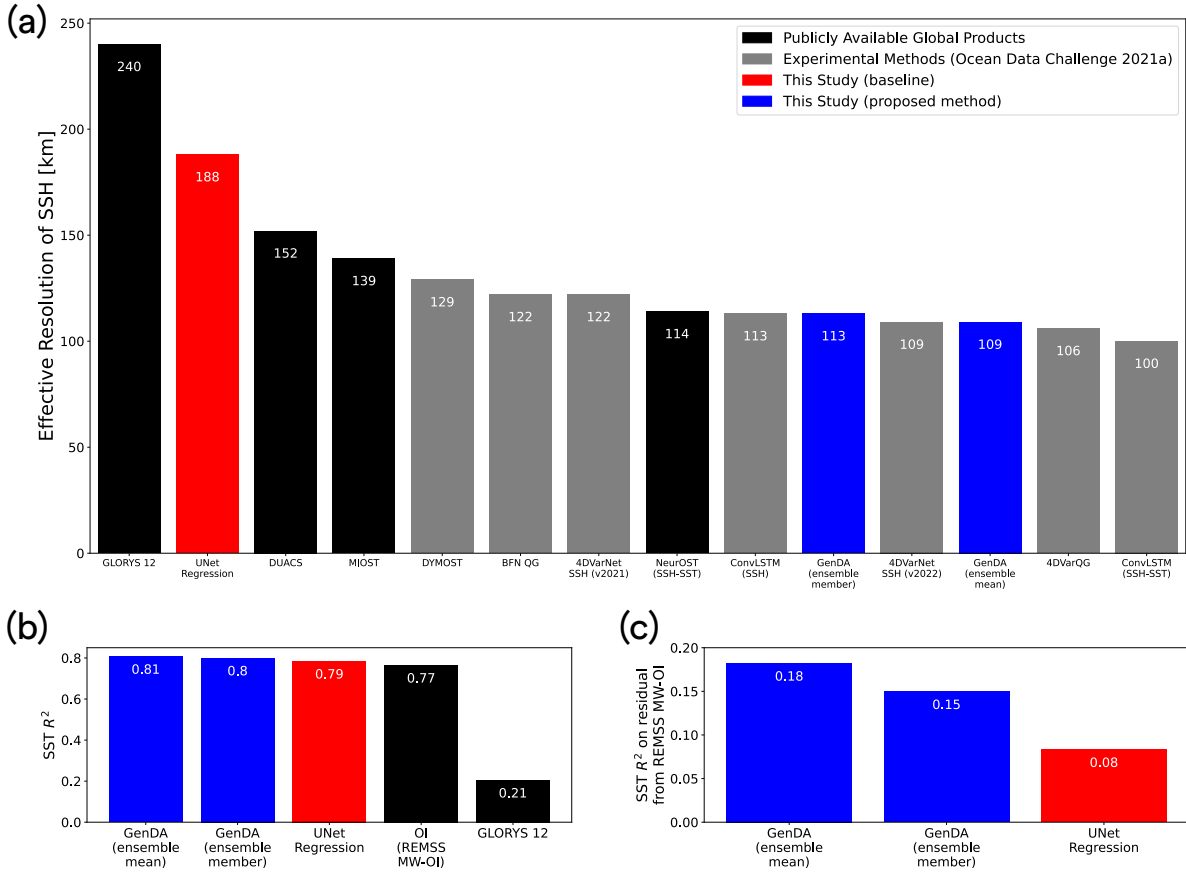


Figure 3.14: Accuracy of OSE state estimates against independent satellite observations. (a) Effective resolution, i.e. smallest resolved wavelength, of SSH for all methods in Ocean Data Challenge 2021a and those considered in this study. (Black) methods for which there are publicly available global products, (gray) experimental methods from other studies, (red) UNet Regression, and (blue) GenDA. (b) & (c) R^2 for SST compared to real-world L3 SST observations which were masked out by artificial clouds and not assimilated. (b) R^2 calculated directly on SST and (c) R^2 calculated on the residual from OI, meaning any positive value represents an improvement compared to OI.

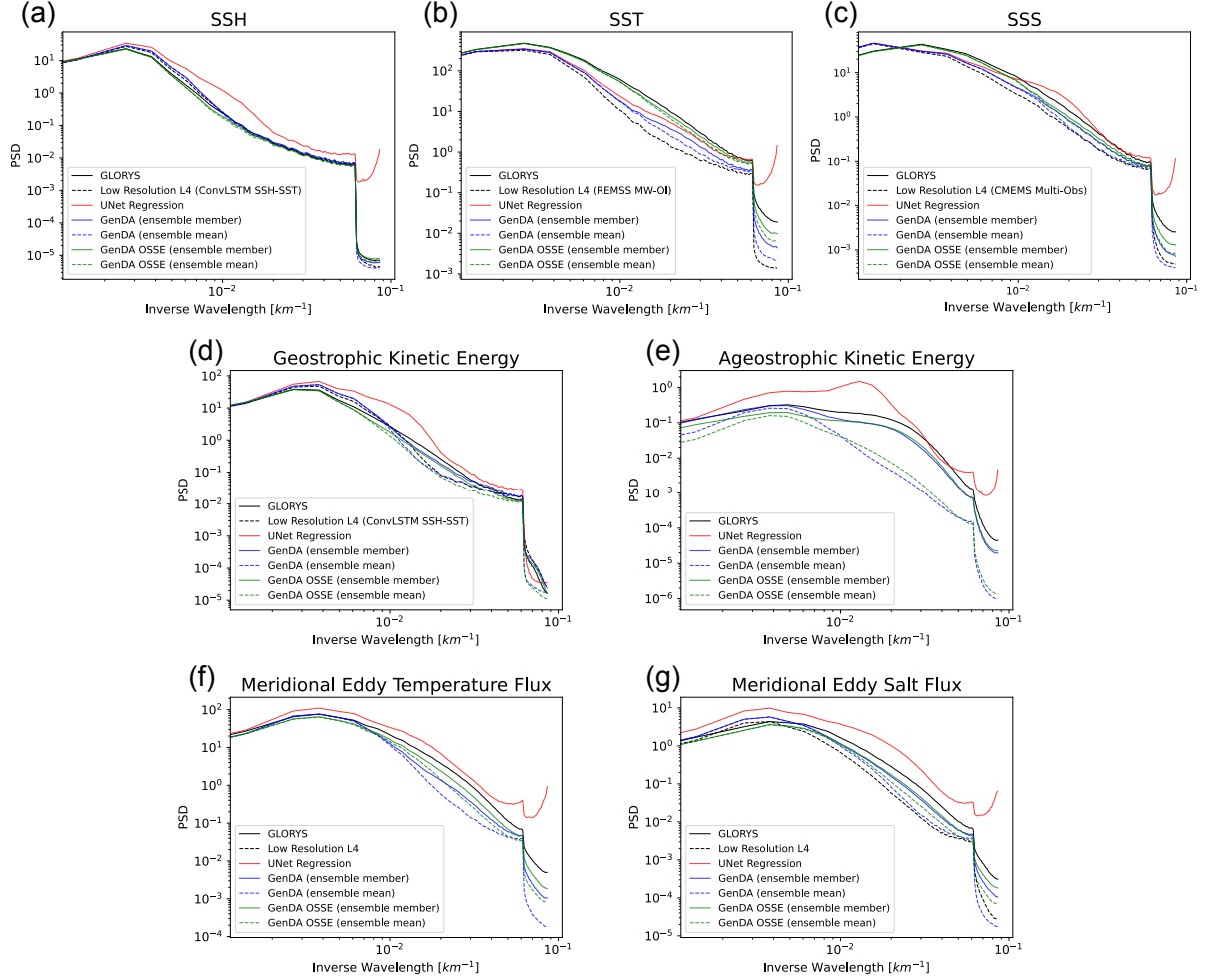


Figure 3.15: (a) SSH, (b) SST, (c) SSS, (d) geostrophic kinetic energy, (e) ageostrophic kinetic energy, (f) meridional eddy temperature flux, and (g) meridional eddy salt flux. (solid black) GLORYS reference from training (note this is no longer the ground truth when assimilating to real-world observations), (dashed black) low resolution L4 OI input assimilated, (solid red) UNet Regression, (solid blue) GenDA ensemble member, (dashed blue) GenDA ensemble mean, (solid green) GenDA ensemble member from the OSSE experiment on simulated observations, and (dashed green) GenDA ensemble mean from the OSSE experiment on simulated observations.

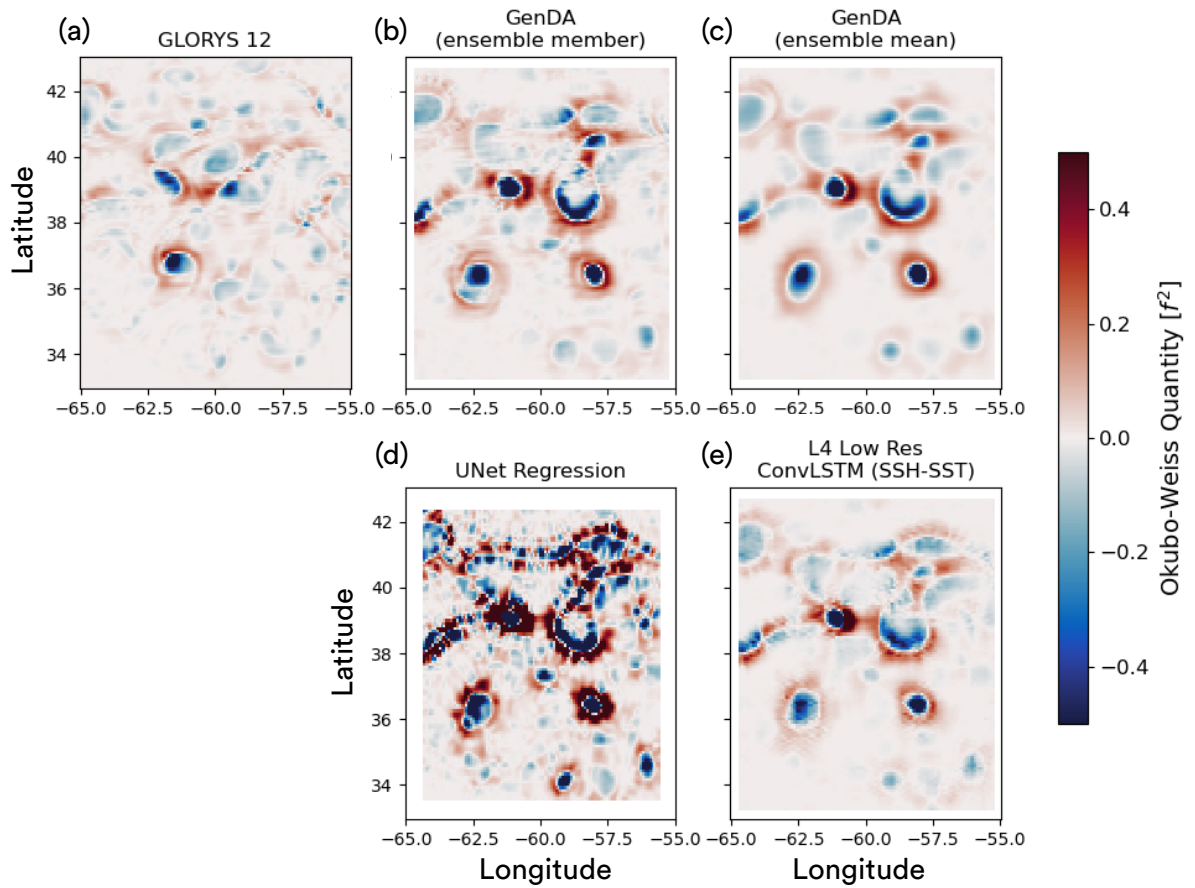


Figure 3.16: Geostrophic Okubo-Weiss quantity when assimilated to real-world observations for (a) GLORYS (no longer the ground truth in the OSE setting), (b) GenDA ensemble member, (c) GenDA ensemble mean, (d) UNet Regression, and (e) OI. Negative values indicate vorticity dominates (as in eddy cores) whereas positive values show where strain dominates (as around eddy peripheries).

Appendix B

Supporting Information for Chapter 3

B1 Linear Ekman Regression Model for Wind-Driven Surface Currents

We remove both geostrophic currents and linear Ekman wind-driven currents from the total 15 m depth currents in GLORYS 12 as a pre-processing step before training GenDA (Equations 10 & 11).

The linear Ekman wind-driven current is given by

$$u_{Ek} = \frac{\sqrt{2}}{\rho_0 f d} e^{z/d} \left[\tau^x \cos\left(\frac{z}{d} - \frac{\pi}{4}\right) - \tau^y \sin\left(\frac{z}{d} - \frac{\pi}{4}\right) \right], \quad (\text{B.1})$$

and

$$v_{Ek} = \frac{\sqrt{2}}{\rho_0 f d} e^{z/d} \left[\tau^x \sin\left(\frac{z}{d} - \frac{\pi}{4}\right) + \tau^y \cos\left(\frac{z}{d} - \frac{\pi}{4}\right) \right], \quad (\text{B.2})$$

where d is the Ekman layer depth, z is depth, ρ_0 is a reference density, and τ^x and τ^y are the zonal and meridional wind stresses respectively (Equation 8.33 from Cushman-Roisin and Beckers [2011]). The Ekman layer depth, d , is a property of local oceanographic conditions which could vary geographically. The wind stresses can be estimated from ERA 5 surface winds using bulk aerodynamical formulae

$$\tau^x = \rho_a C_d u_{atmos} \sqrt{u_{atmos}^2 + v_{atmos}^2}, \quad (\text{B.3})$$

and

$$\tau^y = \rho_a C_d v_{atmos} \sqrt{u_{atmos}^2 + v_{atmos}^2}, \quad (\text{B.4})$$

where ρ_a is an atmospheric reference density which we take to be 1.2 kgm^{-3} , and C_d is a drag coefficient which we take to be 1.2×10^{-3} [Large and Pond, 1981].

In practice, we calculate wind stress from ERA 5 winds using Equations B.3 & B.4 and re-cast Equations B.1 & B.2 as a linear regression model

$$u_{Ek} = A(x, y)\tau^x + B(x, y)\tau^y, \quad (\text{B.5})$$

and

$$v_{Ek} = C(x, y)\tau^x + D(x, y)\tau^y, \quad (\text{B.6})$$

where A , B , C , and D are regression coefficients which we allow to vary geographically but assume to be fixed in time [Lagerloef et al., 1999b]. We find these coefficients by regressing Equations B.5 & B.6 at each coordinate onto GLORYS 12 ageostrophic surface currents over our full dataset duration (2010-2021).

B2 Why Include Surface Winds in the State Vector?

To improve the reconstruction of wind-driven currents not captured by our linear Ekman model (Text SB1), we also reconstruct surface winds in the GenDA state estimate and assimilate ERA 5 winds in the observation operator. Here we show that assimilating ERA 5 winds improves the GenDA reconstruct skill for the unobserved ageostrophic surface currents (Table B.1).

B3 Neural Network Architectures and Hyperparameters

B3.1 GenDA Diffusion Prior Architecture

The neural network architecture we use is a version of the UNet architecture [Ronneberger et al., 2015] with added self-attention used widely in diffusion models [Song and Ermon, 2019; Mardani et al., 2024]. The full details of our architecture are illustrated in Figure B.1. We use this architecture also for the UNet Regression baseline with minor modifications to the numbers of channels to accommodate observations in the input and to reduce overfitting observed during training.

B3.2 GenDA Generation Hyperparameters

Throughout this study, when generating state estimates using the diffusion prior we use the following hyperparameters. We discretize the diffusion time axis into 256 steps from $t = 0$ to $t = T$, and follow the same time-stepping procedure as Rozet and Louppe [2023a] and Manshausen et al. [2025]. Although these prior studies used Langevin Monte Carlo correction steps, in this study we perform no correction steps as it was not found to improve performance in our experiments. We use the same noise schedule as Rozet and Louppe [2023a],

$$\sigma(t) = \sqrt{1 - \mu(t)^2} \quad (\text{B.7})$$

$$\mu(t) = \cos(\omega t)^2 \quad (\text{B.8})$$

$$\omega = \arccos\left(\sqrt{10^{-3}}\right) \quad (\text{B.9})$$

where we set $T = 1$.

In Rozet and Louppe [2023a] the heuristic variance for the posterior likelihood term in the main text (Equation 6) is

$$\mathcal{N}(y|\mathcal{A}(\hat{x}(t)), \Sigma_y(t)) = \mathcal{N}\left(y|\mathcal{A}(\hat{x}(t)), \Sigma_y + \frac{\sigma(t)^2}{\mu(t)^2}\Gamma\right), \quad (\text{B.10})$$

where Σ_y is the standard error of the observations, y , and Γ is a hyperparameter that controls the strength of the response to the observation likelihood term at higher noise levels. In this study we set $\Gamma = 0.1$, noting that this is slightly higher than the values used in Rozet and Louppe [2023a] and Manshausen et al. [2025]. We first tried smaller values for Γ but ran into numerical instabilities during generation so increased Γ until the generation stabilized.

B4 GLORYS Diffusion Prior Training Results

Here we present the training results for GenDA’s score-based diffusion prior, D . During training, the diffusion model is trained to de-noise surface ocean state vectors from GLORYS 12 with Gaussian noise at varying amplitudes, $\sigma(t)$, added (Figure 2). The loss function, $\mathcal{L}_{EDM}$, minimized during training is MSE between the de-noised and noise-free state with a noise amplitude-varying weighting that has been found to

improve training performance [Karras et al., 2022]

$$\mathcal{L}_{EDM} = \frac{\sigma(t)^2 + \sigma_{data}^2}{(\sigma(t) \cdot \sigma_{data})^2} (D(x(t)) - x(0))^2, \quad (\text{B.11})$$

where σ_{data} is a hyper-parameter which we set to 0.5, following Karras et al. [2022]. We observe stable training with both training loss and validation loss decreasing throughout training before leveling off after $\sim 10^7$ training examples (Figure B.2). During training we train the network at noise levels sampled using the distribution proposed in Karras et al. [2022], though we do not put a maximum limit on the noise levels seen during training (i.e. we set $\sigma_{max} = \infty$ rather than $\sigma_{max} = 80$).

Sampling the diffusion prior unconditionally, i.e. with no assimilation of observations as described in Section 2.1.1, shows that our diffusion prior is able to largely capture the distribution of the GLORYS training data (Figure B.3). For all variables, the diffusion prior appears able to produce realistic variability across all scales resolved in the GLORYS and ERA 5 training data. Intriguingly, unconditional generation appears to somewhat under-estimate the variance consistently across all scales for all variables except SSH, which should be further investigated. We note though that assimilation observations mitigates this (see the results presented in the Main Text).

B5 Coarse-Graining Scales in GenDA Observation Operator

The GenDA observation operator requires the prescription of effective coarse-graining scales that reflect the resolutions of the low-resolution L4 SSH, SST, and SSS products assimilated (Section 2.1.3). Here we describe how we choose these scales for both experiments.

B5.1 OSSE: Scales Prescribed A Priori

In the OSSE, we simulate low-resolution L4 satellite products by coarse-graining the respective ground truth GLORYS field with a Gaussian kernel of width, l , chosen to reflect what we expect to be the effective resolution of real-world satellite products. Since L4 satellite products also smooth in time, we smooth the ground truth in time with width, T , when generating the simulated L4 products. Table B.2 lists the coarse-graining scales chosen for each variable along with our rationale.

Since in the OSSE setting the effective coarse-graining scales of the L4 products are known, we set the coarse-graining scales in the GenDA observation operator to the same values used when generating the data (Table B.2). Since GenDA operates only on snapshots, we neglect the temporal smoothing in the observation operator.

B5.2 OSE: Scales Tuned A Posteriori

In the real-world (OSE) setting, we assimilate real L4 products for which the effective coarse-graining scales are not known a priori. To select appropriate coarse-graining scales we therefore employ an a posteriori tuning strategy. We select the first day of each month from one of the cross-validation years (2019) to use as a dataset for tuning the coarse-graining scales. We then generate 120 random combinations of l_{SSH} , l_{SST} , and l_{SSS} by uniformly sampling in the range [5, 40] km. We apply GenDA to the 12-day dataset for each coarse-graining scale combination and analyze the spectra of the resulting state estimates as well as compute errors against withheld SSH and SST observations.

SSH and SST reconstruction errors were found to co-vary only with l_{SSH} and l_{SST} respectively with no clear impact from l_{SSS} on either and no co-variance between SSH and SST (Figure B.4). We thus consider each variable to respond only to its respective coarse-graining scale in what follows. SSH and SST R^2 increase with decreasing l_{SSH} and l_{SST} respectively. This is likely due to the fact that decreasing the coarse-graining scale applied in the observation operator leads the GenDA state estimates to converge towards the low-resolution L4 products. The low-resolution L4 products are close to observations in terms of regression metrics, which are dominated by large-scale signals, but are overly-smooth compared to numerical simulations or L3 observations. When choosing l_{SSH} and l_{SST} , it is thus important not just to maximize R^2 , but also to consider the impact on the physical realism of GenDA at small scales.

The spectral properties of SST on a near cloud-free day, 2019-09-01, illustrate the trade-offs between setting l_{SST} too small or too large (Figure B.5). When l_{SST} is set too large, GenDA over-predicts variance at large scales relative to both OI and L3 observations to counteract the excessive smoothing applied in the observation operator (Figure B.5a & b). Whereas setting l_{SST} too small suppresses variance at small scales, pushing the GenDA state estimate towards OI and away from the L3 ground truth observations (Figure B.5a & c). While choosing l_{SST} is subjective, a rationale emerges looking at how the variance at large and small

scales varies with l_{SST} (Figure B.5b & c respectively). We pick a value that is small enough for the large-scale variance to converge towards that predicted by OI, while being large enough to avoid the small-scale variance being suppressed too much. The small-scale variance in Figure B.5c appears to roll off abruptly below $l_{SST} \sim 15$ km which we take to be a symptom of l_{SST} being too small below this threshold. To choose our OSE coarse-graining scales we thus look at how the large- and small-scale variance of GenDA SSH, SST, and SSS change with l compared to the respective OI products (averaged over all 12 days in the tuning dataset) and pick values that appear to achieve the trade-off described above (Figure B.6). This a posteriori tuning strategy leads us to values of 15 km for all variables (Table B.3), which we note are in reasonable alignment with the values used in the OSSE (Table B.2). Note, for SSH the small-scale variance didn't exhibit the clear roll-off at small l that SST and SSS do (comparing Figure B.6b to d & f), making our choice of l_{SSH} more subjective. We instead inform our choice of l_{SSH} by selecting the scale below which improvements in SSH R^2 plateau in Figure B.4a.

B6 Weighted Re-Sampling Scheme for Comparing Errors For Cloudy vs Non-Cloudy SST and SWOT vs No SWOT SSH

In Figure 10, we compare state estimation errors between cases where L3 SST is cloud-occluded and cloud-free. Since the average cloud concentration has significant geographical structure, we need to ensure both the cloud-free and cloud-occluded datasets are drawn uniformly in space to ensure a fair comparison. We achieve this by drawing N samples from each dataset with replacement using a weighting function that compensates for the average cloud concentration to ensure the drawn samples are uniformly distributed in space.

We estimate the average cloud concentration, $C(x, y)$, through

$$C(x, y) = \overline{1 - f(x, y)}, \quad (\text{B.12})$$

where the overbar indicates averaging over all days in the test year 2017, and $f(x, y)$ is one for cloud-free pixels and zero for cloud-occluded pixels.

From the cloud concentration we then design weighting functions for both the cloud-occluded dataset,

$$W_{cloudy}(x, y) = \frac{1}{C(x, y)}, \quad (\text{B.13})$$

and for the cloud-free dataset,

$$W_{non-cloudy}(x, y) = \frac{1}{1 - C(x, y)}, \quad (\text{B.14})$$

and draw N samples with replacement using these weightings to scale the probability for each pixel to be drawn.

The same algorithm is used to compare errors when SWOT SSH is and isn't available, replacing cloud concentration with the fraction of days on which SWOT SSH was not available.

B7 OSE Ageostrophic Eddy Cyclo-Geostrophy Case Study

Here we highlight evidence of learned physics in the GenDA ageostrophic surface currents in the real-world setting, where we don't have a ground-truth reference field. To do this, we zoom in on the cyclonic eddy at 62°W 36°N in Figure 13 and compare the GenDA ageostrophic currents to those from an idealized theoretical calculation.

For an axially symmetric eddy in equilibrium, the next order correction to geostrophic balance, 'cyclo-geostrophy', has a closed form solution [Penven et al., 2014]

$$V = \frac{2V_g}{1 + \sqrt{1 + \frac{4V_g}{fR}}}, \quad (\text{B.15})$$

where V is the characteristic eddy velocity (taken here to be the maximum of the azimuthally averaged eddy velocity profile), V_g is the characteristic eddy velocity calculated from geostrophy, and R is the radius of the eddy. Note V and V_g are defined to be positive for a cyclone (anti-cyclone) in the Northern (Southern) Hemisphere and negative for the opposite rotation. Thus, in the Northern Hemisphere, cyclones are expected to be weaker than geostrophy predicts, and anti-cyclones are expected to be stronger.

For the cyclonic eddy in Figure B.7a, the GenDA total surface current (we still don't add back in the Ekman current here to focus on eddy dynamics) is 10-15% weaker than the GenDA geostrophic current

(Figure B.7b-e). Applying cyclo-geostrophy to the GenDA geostrophic currents would lead to a 20-25% weakening of the current speeds (Figure B.7f). The ageostrophic surface currents predicted by GenDA therefore appear to correct the eddy velocity in the direction expected from cyclo-geostrophy (i.e. the cyclone gets weaker), albeit with weaker magnitude than predicted from cyclo-geostrophy. The discrepancy between GenDA and cyclo-geostrophy could stem both from the idealizing assumptions of axial symmetry and equilibrium made in applying Equation B.15, and from the fact that GenDA was observed to under-predict variance in ageostrophic currents in the OSSE setting (Table 1). We stress, GenDA is provided no observations of ageostrophic currents at inference and is here being assimilated to real-world satellite data, so it is predicting ageostrophic currents in an unguided manner. The simulation-informed diffusion prior therefore appears to have learned the physical relationships between unobserved ageostrophic currents and satellite observables (SSH/SST/SSS).

B8 OSE Comparison of Ageostrophic Currents to Surface Drifters

To evaluate the accuracy of the estimated surface currents in the real-world setting (OSE), we here compare them against independent *in situ* observations from surface drifters. We subset all available surface drifters from the Global Drifter Program [Copernicus Marine in situ TAC, 2024] in our study region and year and retain only drifter trajectories that remain in the domain for at least ten days — the same drifter can have multiple trajectories if it re-enters the domain at a later time. To filter out the effect of inertial oscillations we apply a 24 hour running mean to the hourly drifter data before interpolating the mapped geostrophic and ageostrophic currents to the locations of the drifters. We evaluate the RMSE of the zonal and meridional current separately. In addition, we evaluate the direction and magnitude of the predicted current jointly through the mean error vector amplitude (MEVA) defined as the mean magnitude of the velocity error vector following Garcia et al. [2025].

Table B.1: Effect of assimilating ERA 5 winds on GenDA (ensemble mean) reconstruction skill for ageostrophic currents.

ERA 5 Assimilated?	$R^2 u_{ageo}$	$R^2 v_{ageo}$
No	0.372	0.358
Yes	0.383	0.369

Table B.2: Gaussian coarse-graining scales applied to SSH, SST, and SSS to generate simulated low-resolution L4 satellite products in the OSSE.

Variable	l [km]	T [days]	Rationale
SSH	25	1.75	Error analysis in Martin et al. [2024].
SST	16	1.23	Same as chosen in Archambault et al. [2024b].
SSS	16	1.23	Assumed to be same as SST.

Table B.3: Gaussian coarse-graining scales applied to SSH, SST, and SSS in GenDA observation operator in OSE after a posteriori tuning.

Variable	l [km]
SSH	15
SST	15
SSS	15

Table B.4: Evaluation of surface ocean currents compared against surface drifters for the OSE.

Method	MEVA [cm/s]	u RMSE [cm/s]	v RMSE [cm/s]
ConvLSTM SSH-SST (Geostrophic)	23.2	19.8	18.9
GenDA Ensemble Mean (Geostrophic)	23.3	20.1	19.0
GenDA Ensemble Mean (Total)	23.7	20.3	19.5
GenDA Ensemble Member (Geostrophic)	24.2	21.0	19.7
GenDA Ensemble Member (Total)	25.2	22.1	20.2
UNet Regression (Geostrophic)	37.4	37.4	32.6
UNet Regression (Total)	42.3	44.9	35.5
GLORYS (Geostrophic)	34.8	30.2	30.3
GLORYS (Total)	36.4	31.4	31.5

Movie S1. Movie of OSSE predictions for full testing year. Figure panels correspond to those in Figure 5.

Movie S2. Movie of OSE predictions for full testing year. Figure panels correspond to those in Figure 13.

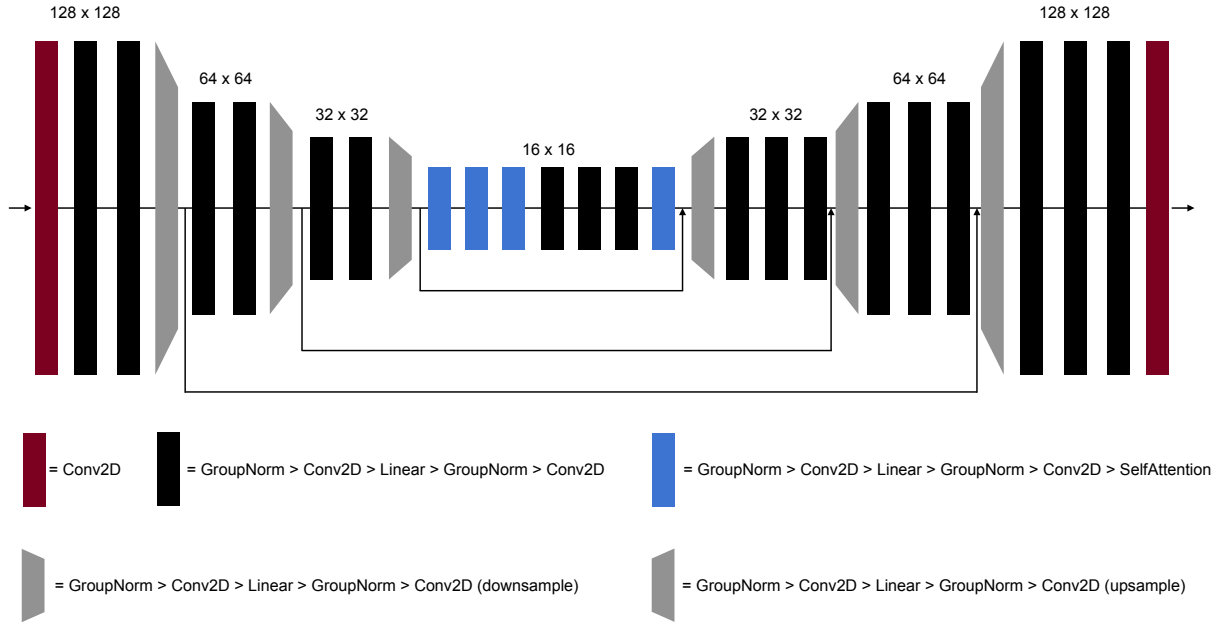


Figure B.1: Schematic of the UNet neural network architecture used in both the GenDA diffusion prior and UNet Regression baseline.

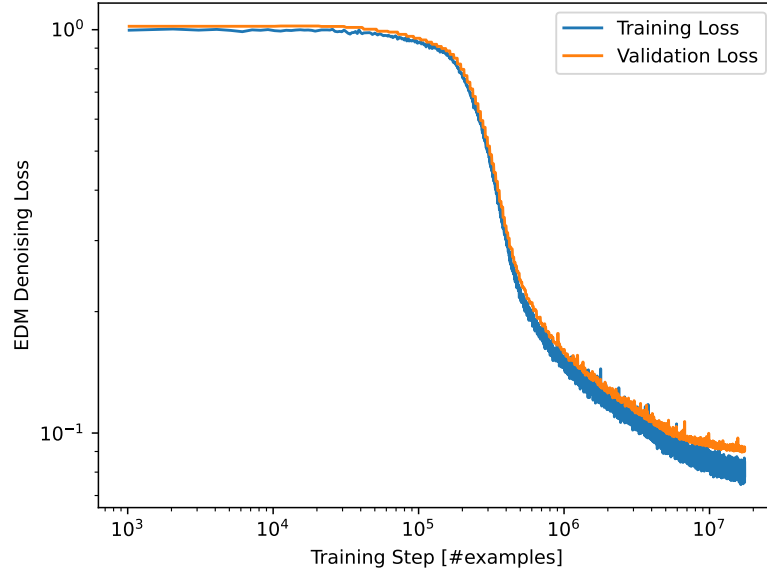


Figure B.2: Learning curves for training the GenDA score-based diffusion prior to de-noise GLORYS 12 surface ocean states. The EDM de-noising loss function (Equation B.11) is shown as a function of training step calculated both on the training (blue) and cross-validation (orange) datasets.

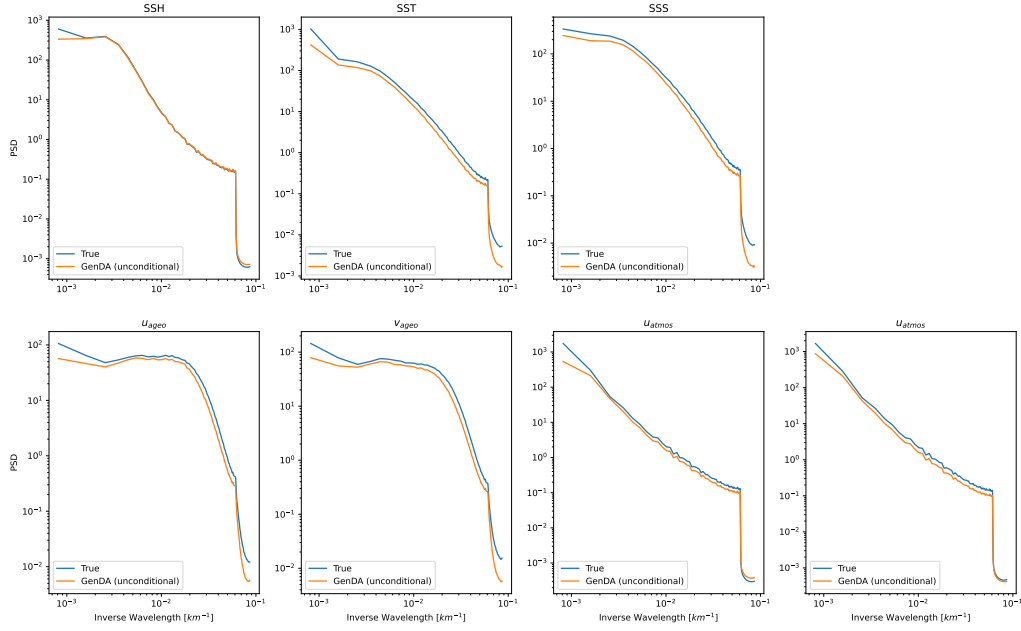


Figure B.3: Spectra of each variable when generated unconditionally (with no assimilation of observations) by the GenDA score-based diffusion prior (orange) compared to the GLORYS ground truth (blue). (a) SSH, (b) SST, (c) SSS, (d) u_{ageo} , (e) v_{ageo} , (f) u_{atmos} , and (g) v_{atmos} .

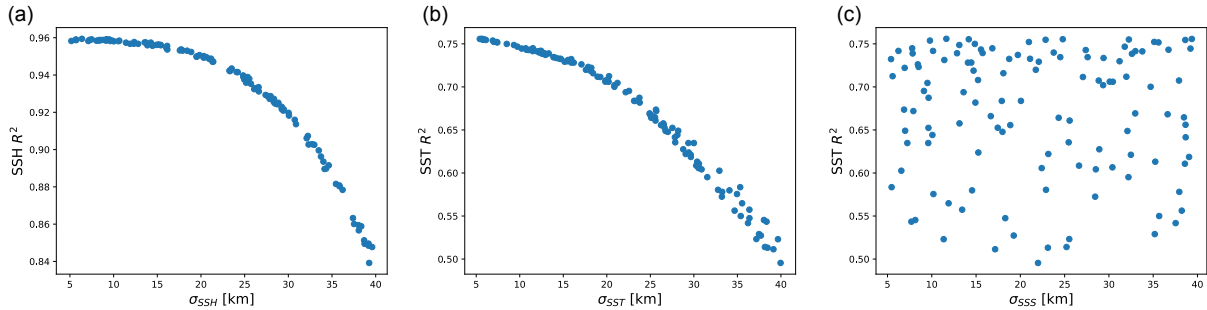


Figure B.4: Effect of varying coarse-graining scales on GenDA OSE state estimate R^2 against independent L3 SSH and SST observations. (a) SSH R^2 with varying l_{SSH} , (b) SST R^2 with varying l_{SST} , and (c) SST R^2 with varying l_{SSS} . No clear co-variance appears between SST R^2 and σ_{SSS} . We don't show them here, but co-variances between l_{SSS} and SSH R^2 , between l_{SST} and SSH R^2 , and between l_{SSH} and SST R^2 are also not apparent with the plots looking like panel (c).

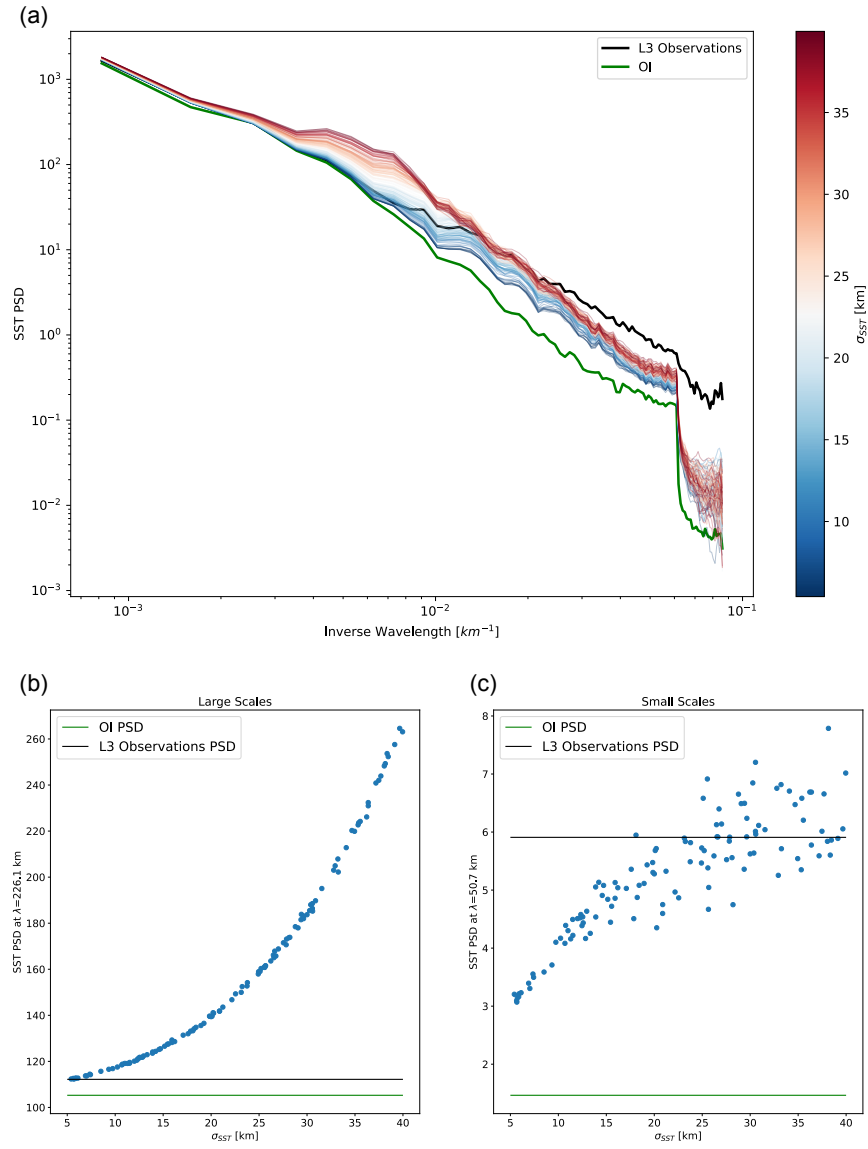


Figure B.5: Spectral response of GenDA SST to varying l_{SST} on a near cloud-free day of the tuning dataset, 2019-09-01. (a) SST spectra for GenDA with varying l_{SST} (blue to red colormap), for L3 observations with the few cloudy pixels filled with the domain mean (black), and for the low-resolution L4 OI satellite product (green). The GenDA SST spectrum values at (b) a large wavelength ($\lambda = 226.1$ km) and (c) a small wavelength ($\lambda = 50.7$ km) are scattered against l_{SST} and compared to L3 observations (black) and the low-resolution L4 OI satellite product (green).

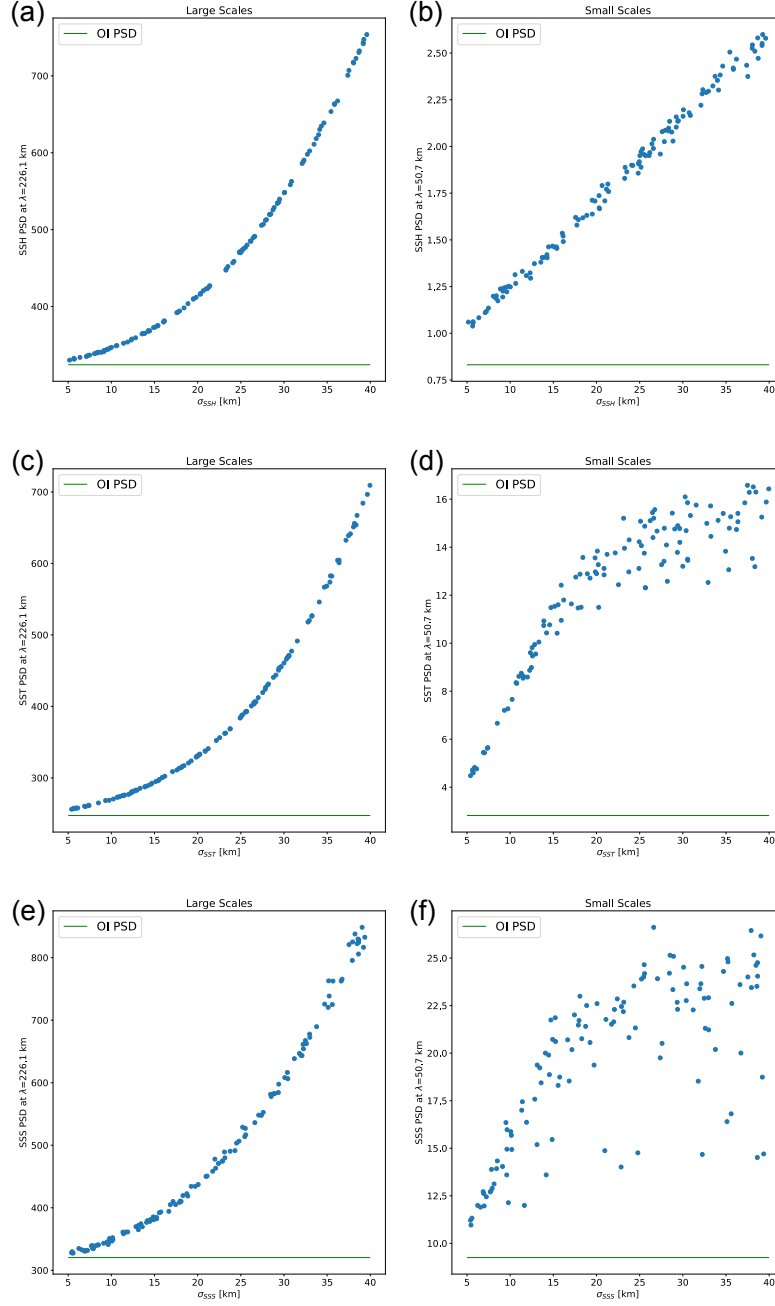


Figure B.6: Spectral response of GenDA at large ($\lambda = 226.1$ km) and small ($\lambda = 50.7$ km) scales to varying coarse-graining scales averaged over all 12 days of the tuning dataset. (a) & (b) Large-scale and small-scale response of GenDA SSH to l_{SSH} respectively. (c) & (d) Large-scale and small-scale response of GenDA SST to l_{SST} respectively. (e) & (f) Large-scale and small-scale response of GenDA SSS to l_{SSS} respectively. The low-resolution L4 OI satellite product is shown in green in all panels.

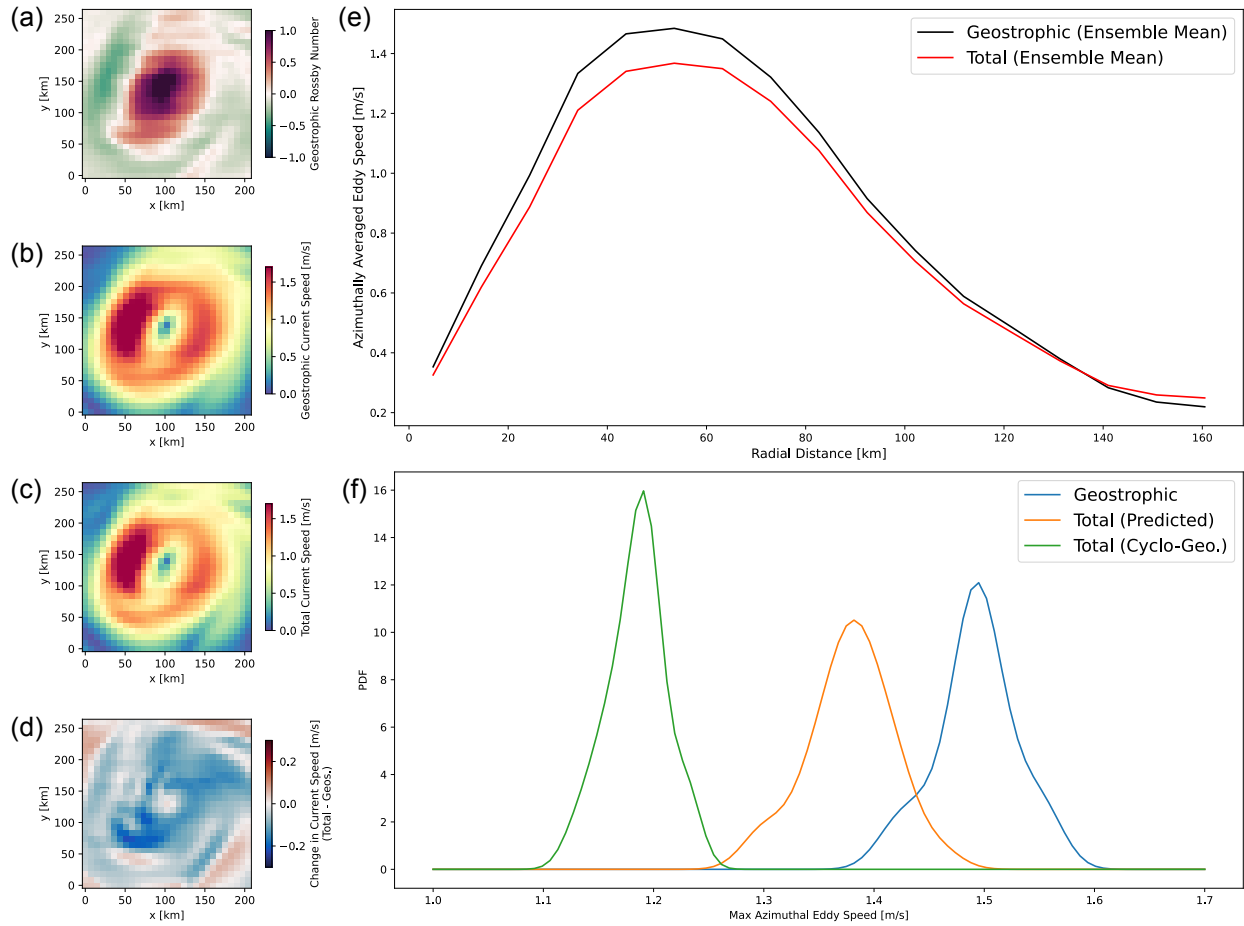


Figure B.7: Comparing GenDA ageostrophic currents to cyclo-geostrophy for eddy reconstructed from real-world observations (cyclonic eddy at 62°W 36°N in Figure 13). Zoomed in view of the eddy in one of the GenDA ensemble members showing: (a) geostrophic vorticity, (b) geostrophic current speed, (c) total current speed (GenDA geostrophic plus GenDA ageostrophic), and (d) the change in speed between total and geostrophic currents. (e) Azimuthally averaged velocity profiles for geostrophic and total current speed, we show here the ensemble mean but each ensemble member looks qualitatively similar. (f) PDF over GenDA ensemble members of the eddy's characteristic speed for geostrophic currents, total currents, and solving cyclo-geostrophic balance using the geostrophic currents.

Chapter 4

Observational constraints on the submesoscale ocean kinetic energy cascade from generative data assimilation

This chapter is in preparation for submission to a journal.

4.1 Introduction

The ocean receives energy at scales far removed from those at which it can dissipate, and the cascade of energy between these limits is a defining, yet poorly observed, feature of the ocean circulation and its variability. Energy is injected primarily at large scales over thousands of kilometers by wind forcing and is ultimately dissipated by small-scale turbulent mixing at millimeter scales [Ferrari and Wunsch, 2009]. Between these limits lies a hierarchy of motions from geostrophic mesoscale eddies hundreds of kilometers across to submesoscale fronts, filaments, and eddies at one to tens of kilometers that mediate the transfer of energy across scales [Klein et al., 2019]. At (sub)mesoscales this transfer is bidirectional — energy can cascade both upscale and downscale depending on the dynamical regime — meaning that smaller scales can feed back and impact larger scales. Since existing eddy parameterizations used in global climate models do not account for this bidirectional energy transfer [Gent and McWilliams, 1990; Fox-Kemper et al., 2008],

quantifying the (sub)mesoscale cascade and its variability is essential both for understanding the extent to which small-scale processes drive variability at larger scales [Sasaki et al., 2014; Sérazin et al., 2018] and for informing the next generation of climate model parameterizations.

High-resolution numerical simulations have given an increasingly detailed picture of the (sub)mesoscale kinetic energy cascade. At mesoscales, the predominance of coherent balanced eddies is expected to drive a strong upscale cascade, feeding energy back to the largest scales [Salmon, 1980; Scott and Arbic, 2007; Tulloch et al., 2011; Aluie et al., 2018]. At submesoscales, coherent eddies coexist with fronts, filaments, and unbalanced waves, with these different features driving competing cascades in opposite directions. Coherent submesoscale eddies drive an upscale cascade through eddy-eddy interactions [Capet et al., 2008b; Sasaki et al., 2014; Qiu et al., 2014; Schubert et al., 2020; Dong et al., 2020; Ajayi et al., 2021; Wang et al., 2022b], while the more three-dimensional circulation at fronts and filaments, and potentially eddy-wave interactions, drive a downscale cascade toward dissipation scales [D’asaro et al., 2011; Barkan et al., 2017; Schubert et al., 2020; Srinivasan et al., 2023]. The net cascade thus reflects the relative dominance of these competing processes, a balance that likely varies geographically and seasonally. Submesoscale dynamics are known to be strongly seasonal, with wintertime mixed layer deepening driving instabilities that inject kinetic energy at submesoscales [Callies et al., 2015; Buckingham et al., 2016; Taylor and Thompson, 2023]. High-resolution simulations suggest part of this energy progressively cascades to larger scales throughout the winter and spring, contributing significantly to the seasonal cycle of mesoscale eddy energy and, through it, the broader variability of the ocean circulation [Sasaki et al., 2014; Qiu et al., 2014]. Despite these insights from simulations, observational constraints on the (sub)mesoscale cascade remain limited, and whether high-resolution simulations faithfully reproduce the cascade in the real ocean remains an open question.

Observing the (sub)mesoscale kinetic energy cascade is fundamentally challenging due to its multiscale nature. To constrain the cascade, observations must simultaneously have fine spatial resolution to resolve submesoscale features at the kilometer-scale, whilst covering large domains hundreds of kilometers across to constrain the mesoscale state. Furthermore, since submesoscale features evolve on daily time-scales, necessitating temporally concurrent observations. Surface geostrophic currents inferred from gridded satellite altimetry provided early observational confirmation of the mesoscale upscale cascade [Scott and Wang, 2005], but conventional gridded altimetry smooths all but the largest mesoscale variability [Ballarotta et al.,

2019], which induces systematic biases in the diagnosed cascade [Arbic et al., 2013]. A strong, seasonal upscale cascade at mesoscales, with energy progressively cascading to larger scales throughout the winter and spring, has been identified indirectly in along-track altimetry observations [Steinberg et al., 2022; Lawrence and Callies, 2022; Schubert et al., 2023] and directly in high-resolution gridded observational datasets derived from data assimilation [Storer et al., 2023] and deep learning [Martin et al., 2024]. While these studies confirm that the upscale cascade plays a significant role in driving mesoscale seasonality, they lack the resolution to probe submesoscales and likely under-estimate the strength of the cascade even at mesoscales. In situ studies using drifter [Balwada et al., 2022], shipboard ADCP [Qiu et al., 2022; Chai et al., 2026], and mooring [Naveira Garabato et al., 2022] observations have provided valuable but temporally and spatially sparse constraints on the submesoscale cascade, pointing to a bi-directional and seasonally-varying energy transfer. The recent development of airborne sensors capable of directly measuring submesoscale surface currents offers a promising route to observing the submesoscale cascade [Freilich et al., 2023; Farrar et al., 2025], though current deployments have been limited to the California Current region, leaving open the question of how representative these findings are of other dynamical regimes. Satellite observations offer the synoptic spatial coverage and temporal continuity that in situ and airborne platforms cannot, but extracting submesoscale-resolving state estimates from space has remained an unsolved methodological challenge.

The launch of the Surface Water and Ocean Topography (SWOT) mission in 2022 marked a fundamental advance in observing submesoscale ocean dynamics from space, providing wide-swath sea surface height (SSH) observations at spatial resolutions down to 250 m with global coverage every 21 days [Morrow et al., 2019; Fu et al., 2024; Archer et al., 2025]. However, diagnosing the kinetic energy cascade from SWOT SSH alone [Qiu and Chen, 2025] is hampered by the limited 120 km swath width, the difficulty of inverting surface currents from SSH snapshots, and contamination by unbalanced internal wave signals [Gaultier et al., 2016; Klein et al., 2019]. Deep learning methods have recently demonstrated the ability to reconstruct high-resolution ocean state estimates by synergizing a variety of different satellite observables [Martin et al., 2024; Fablet et al., 2024; Beauchamp et al., 2025; Lenain et al., 2025]. Generative data assimilation has shown particular promise for this problem by combining multiple satellite observables into physically plausible, high-resolution estimates of the multi-variate surface ocean state [Martin et al., 2025]. The method trains a generative model on high-resolution ocean simulations to learn how surface variables

such as SSH, SST, salinity, and surface currents co-vary and uses this learned prior to constrain state estimation from sparse, multi-modal satellite observations. Since the learned prior encodes complex, non-linear dependencies between variables, this allows reconstruction of unobserved quantities such as ageostrophic surface currents from their learned relationships to satellite observables [Martin et al., 2025]. Synergizing different observables such as SSH and SST in a unified framework allows the method to fully exploit the global satellite observing system in a way that single-observable approaches cannot.

Here we refine generative data assimilation for submesoscale-resolving surface ocean state estimation from a combination of SWOT SSH, SST, satellite salinity, and along-track altimetry observations, and apply this framework to provide novel observational constraints on the strength and seasonality of the submesoscale kinetic energy cascade. We find a strongly seasonal upscale cascade extending down to $\mathcal{O}(10)$ km and a downscale cascade localized at submesoscale fronts and filaments, both of which were essentially missing in prior satellite estimates. This finding is in broad agreement with high-resolution numerical simulations, providing direct observational evidence for a strong, seasonal upscale (sub)mesoscale kinetic energy cascade. We will demonstrate that generative data assimilation is a powerful new tool for constraining submesoscale ocean dynamics from satellite observations, providing a window on a scale range unresolved in typical global simulations whose role in larger-scale ocean variability remains to be fully quantified.

4.2 Methods

4.2.1 Score-based data assimilation

Score-based data assimilation [Rozet and Louppe, 2023a] is a method for solving inverse problems using a data-driven prior learned from simulation data, implemented in practice as an unconditional diffusion model [Ho et al., 2020; Song et al., 2020; Karras et al., 2022]. We seek to infer the complete surface dynamical state, x , from sparse satellite observations, y , related through an observation operator $\mathcal{A}(x)$. The core idea is to first train a diffusion model on high-resolution simulation output to learn a prior $p(x)$ over physically consistent surface ocean states, and then use this prior within a Bayesian framework to find the state maximizing the posterior, $p(x|y)$.

The diffusion process in diffusion models defines a continuous transformation along a monotonically

increasing ‘time’ coordinate, τ , from the data distribution at $\tau = 0$, in our case the distribution of plausible surface ocean states in the simulation, to pure Gaussian noise at $\tau = T$ by progressively adding noise to the state. This process can be reversed by integrating a stochastic differential equation that depends only on the ‘score function’, $\nabla_x \log p(x(\tau))$, meaning new samples can be drawn from the data distribution by starting from random noise and integrating backwards along τ . In practice, the score function is learned by training a neural network to remove additive Gaussian noise at a wide range of noise levels [Song et al., 2020; Karras et al., 2022].

To perform state estimation from observations, we wish to sample from the posterior distribution rather than the prior. In score-based data assimilation, to sample from the posterior distribution given observations, the posterior score, $\nabla_x \log p(x(\tau)|y)$, is approximated by augmenting the learned unconditional score with an approximation of the observation likelihood [Rozet and Louppe, 2023a] which can be evaluated given only the unconditional score network and the observation operator:

$$\nabla_x \log p(x(\tau)|y) = \nabla_x \log p(x(\tau)) + \nabla_x \log p(y|x(\tau)) \quad (4.1)$$

$$\approx \nabla_x \log p(x(\tau)) + \nabla_x \log \mathcal{N}(y|\mathcal{A}(\hat{x}(\tau)), \Sigma_y(\tau)), \quad (4.2)$$

where $\hat{x}(\tau)$ is the denoised state predicted by the network, and $\Sigma_y(\tau)$ is a heuristic variance that increases with τ [Rozet and Louppe, 2023a; Manshausen et al., 2025; Martin et al., 2025].

State estimation is thus achieved by reverse-time integration from random noise, guided jointly by the learned prior and the observations. This process yields a physically plausible state consistent with y . Ensemble state estimates can be generated by starting from different random noise initial conditions. Rozet and Louppe [2023a], Manshausen et al. [2025], and Martin et al. [2025] provide full details of the approximated observation likelihood term, $\nabla_x \log p(y|x(\tau))$, and prior applications; below we describe the refinements beyond the method of Martin et al. [2025] made for submesoscale-resolving state estimation.

4.2.2 Learning surface ocean dynamics through video diffusion

In Martin et al. [2025], we used score-based data assimilation for surface ocean state estimation, training a diffusion prior to generate a state vector, x , where x corresponds to a single time snapshot of SSH, SST, sea surface salinity (SSS), and the zonal and meridional components of surface ocean currents. While this

allows learning relationships between variables, the lack of any explicit handling of time evolution means that the method draws independent samples each time-step and thus cannot generate coherent dynamical evolution where there are large gaps between observations. This becomes a key limitation for submesoscale-resolving state estimation as observations at a single time give only sparse, localized coverage. The width of the SWOT swath is 120 km, which sets too restrictive a domain size for robustly characterizing the submesoscale cascade. In addition, satellite observations of SST are missed due to the presence of clouds. By integrating observations over several consecutive time-steps, the coverage can be greatly expanded by exploiting the precessing orbit of SWOT and the time evolution of the clouds that block SST observations. To assimilate observations over multiple time-steps, a spatiotemporal prior is required to model the evolution of submesoscale eddies, fronts, and filaments between successive observations.

In this study, we thus train a diffusion prior to generate a state vector which captures a number of sequential time snapshots. This way our model can implicitly learn the *dynamics* of the surface ocean, with rich phenomena like eddy merging and splitting, interacting eddy dipoles, and strain-driven deformation around eddy cores that are all captured in the high-resolution simulation training data. In practice, we train a diffusion prior to generate five consecutive time-steps which we will demonstrate is sufficient to learn physically-realistic eddy evolution and to provide broad spatial coverage when observations are assimilated over the full time-series. We model the time dimension explicitly by utilizing a video diffusion architecture [Ho et al., 2022]. Specifically, in the neural network architecture of our diffusion model, we apply the same 2D spatial convolutional filters to each time-step independently, before applying pixel-wise temporal attention to model the time dynamics. Applying temporal attention at different depths in the backbone allows the model to capture multi-scale dynamics. This architecture is inspired by that used in the global atmosphere foundation model Climate in a Bottle [Brenowitz et al., 2025] (Appendix C1). This design allows assimilation of observations over multiple time-steps. Employing a prior that captures time evolution also provides a significantly stronger physical constraint on the reconstructed states than a purely spatial prior.

4.2.3 Submesoscale-permitting ocean model simulation training data

To provide a large training dataset of physically plausible realizations of (sub)mesoscale ocean dynamics, we train our diffusion prior on surface data from the NASA LLC4320 simulation, a global $1/48^\circ$ MITgcm simulation [Rocha et al., 2016; Su et al., 2018]. The LLC4320 simulation is the highest-resolution global ocean simulation run with realistic forcing for a long enough period to cover a full annual cycle. More details about the configuration of LLC4320 can be found in Rocha et al. [2016]. The simulation covers fourteen months of output at $1/48^\circ$ resolution, spun up using increasingly high resolution simulations which were initially initialized using ECCO initial conditions to ensure a large-scale ocean state consistent with real observations. We discard the first two months and train our prior on the final full annual cycle of output to mitigate the effects of spin-up when LLC4320 is adjusting to the increase in model resolution from its low-resolution parent simulation. We focus our study on the region connecting the Atlantic and Indian ocean basins encompassing the Agulhas Retroflection, Agulhas Return Current, and extend the domain to the Indo-Atlantic sector of the Antarctica Circumpolar Current, covering 5° to 40° E, and from 35° to 55° S (Figure 4.1). We focus on this region since it features energetic (sub)mesoscale eddy turbulence and allows us to avoid training on the full range of global ocean dynamics while maintaining a large training domain to facilitate training on a diverse training dataset.

We train our diffusion prior to generate SSH, SST, SSS, and both zonal and meridional surface currents from LLC4320. The neural network processes patches of spatial size 256 by 256 pixels where the data are presented on the native LLC4320 grid, which in this region has a typical spacing of 1.5-2 km. We train a diffusion model to generate a series of 5 time-steps, where each time-step is a 12-hour average of the hourly simulation output. These choices were made to make the network’s field of view sufficient to capture a range of (sub)mesoscale dynamics while staying within reasonable computational bounds. For SST and SSS, we remove the patch mean so that the network learns to generate SST and SSS anomalies, while for SSH we subtract a time-varying linear plane fit to remove both the large-scale background and the effects of barotropic tides and the inverse barometer effect which both impact SSH in LLC4320 but are filtered out from real world altimeter observations. For surface currents, we take the current at 15 m depth to reduce the impacts of surface winds and further pre-process by subtracting the patch mean to remove large-scale signals which can be easily inferred in the real world setting from existing methods and data products. We generate

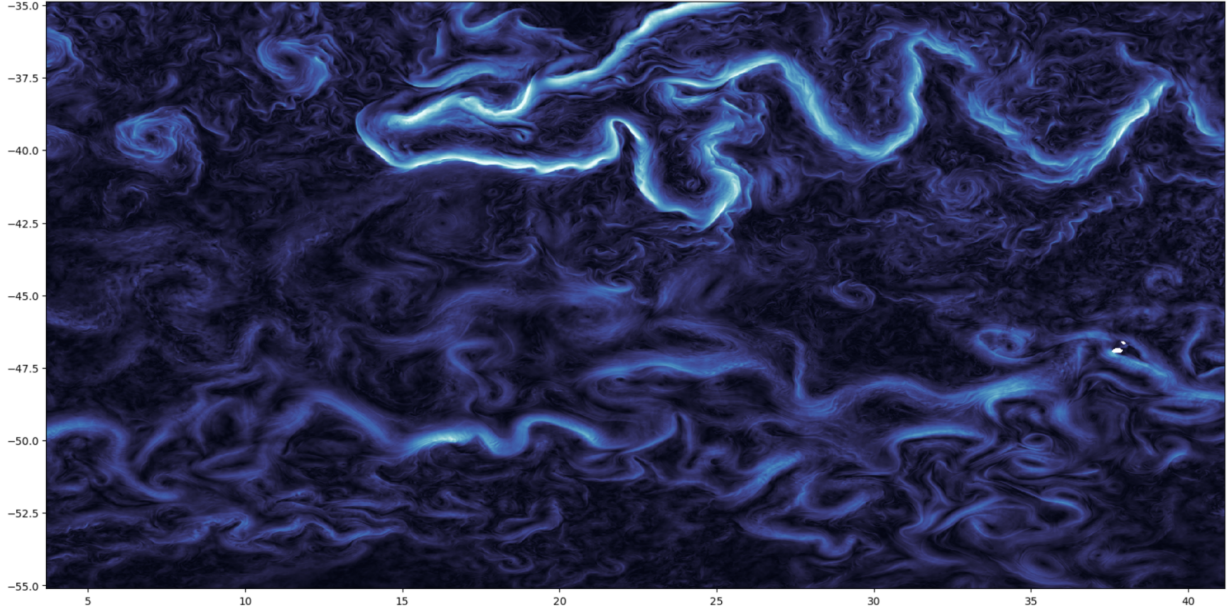


Figure 4.1: Surface current speed from the LLC4320 simulation showing the training domain used in this study.

training examples by sub-sampling from our full training domain and excluding any examples containing land or sea ice in any pixel.

Since LLC4320 is known to have overly energetic internal tides which project strongly onto SSH and surface currents [Savage et al., 2017; Yu et al., 2019; Luecke et al., 2020; Arbic et al., 2022], we apply a sub-inertial filter to the SSH and surface current fields to remove the internal tides before taking the 12-hourly average. We apply a low-pass temporal filter to the data removing signals with frequency above $1.1f$, where f is the local inertial frequency. While more sophisticated approaches have been proposed in previous work to more cleanly isolate the signature of internal tides [Torres et al., 2018; Jones et al., 2023], this approach is a compromise between efficacy and ease of implementation. While this step allows us to train a diffusion prior to generate ocean states with the internal tides (mostly) filtered out, when we assimilating real world SWOT SSH observations, sub-inertial unbalanced wave signals will be present in the data. Developing a method to remove internal tides from SWOT observations is an active area of research [Lguensat et al., 2020; Wang et al., 2022a; Gao et al., 2024; Lyu et al., 2024; Wang et al., 2025]. Contamination from unbalanced internal tides may thus be present after applying our generative data assimilation approach to real SWOT observations. We perform additional experiments to assess the impact of this on the quality

of the reconstructions. The contamination of unbalanced internal tides is expected to be most significant in local summertime when the mixed layer is shallow, but in our study region Torres et al. [2018] found balanced motions to be the leading order contributor to SSH and surface current variance in the 10-100 km scale range throughout the year.

We create a large training dataset of patches from the study region that sample a full annual cycle of LLC4320 output. This exposes the diffusion model to a diverse range of (sub)mesoscale surface ocean dynamics. Since the model learns local dynamics and is given no input information about geographical location, the learned prior will capture a wide range of physically plausible dynamics. We expect that the diversity of states seen during training will help the model to generalize across dynamical regimes if supported by the observations during state estimation.

4.2.4 Observation operator for submesoscale surface ocean state estimation

Building upon the method presented in Martin et al. [2025], we design an observation operator to exploit multi-modal, multi-resolution satellite observations of SSH, SST, and SSS. For SSH, we combine three observation sources: submesoscale-resolving wide-swath SWOT observations, along-track nadir altimeter observations, and coarse gridded SSH estimates from NeurOST [Martin et al., 2024]. For the sparse SWOT and nadir altimeter observations, the observation operator sub-samples the predicted SSH at the grid points observed by satellites, while to assimilate the coarse NeurOST SSH we first coarse-grain the prediction with a filter chosen to reflect the effective resolution of NeurOST. Assimilating coarse satellite products in this way allows us to leverage the accuracy of these products at large scales without penalizing the diffusion prior for generating variance beneath the coarse products’ effective resolution [Martin et al., 2025]. For SST, we leverage geostationary observations from SEVIRI [OSISAF, 2017] to obtain high-resolution, cloud-occluded SST measurements with consistent sampling and instrument errors owing to being sampled by a single platform and further assimilate coarse SST estimates from objective analysis [Remote Sensing Systems, 2017]. Since the resolution of SEVIRI SST is $\sim 0.05^\circ$ in our study region, we first coarse-grain the predicted SST to this resolution before masking out pixels where the SEVIRI measurements are occluded by clouds. For SSS, we assimilate only coarse satellite estimates from objective analysis [E.U. Copernicus Marine Service Information (CMEMS), 2024g] as high-resolution satellite SSS observations

are not available, and for surface currents we assimilate only large-scale NeurOST geostrophic currents, since satellites do not observe small-scale surface currents and the dynamics at large scales are mostly geostrophic. Details of the precise coarse-graining scales and specific satellite products are in Appendix C1. Taken together, the observation operator allows us to exploit high-resolution but gappy observations of SSH and SST to constrain the submesoscale state. Then, the multi-variate learned prior transfers these submesoscale observational constraints to the more coarsely observed SSS and surface current fields.

Hereafter, we refer to the generative data assimilation method used here using the shorthand ‘Gen-LLC4320’.

4.3 Experiments

We assess the state estimation capabilities of GenLLC4320 from simulated satellite observations in an observing system simulation experiment before applying it to real world satellite observations to constrain the submesoscale kinetic energy cascade.

4.3.1 Observing system simulation experiment

To assess the state estimation capabilities of GenLLC4320 in a observing system simulation experiment (OSSE) where the full ground truth is known, we evaluate it first using simulated satellite observations sampled from the simulation. Synthetic altimetry observations are generated by sampling the LLC4320 fields with observation masks taken from satellite observations for the year 2024. For SST, we apply cloud masks taken from the SEVIRI satellite observations for the year 2024 and further coarse-grain the LLC4320 SST to the resolution of the SEVIRI data product, with nominal 0.05° resolution. Simulated coarse gridded satellite SSH, SST, SSS, and surface geostrophic currents are generated by coarse-graining the simulation in space and time to reflect the effective resolution of these various products (Appendix C1).

Throughout this study we focus on two study regions in which we explore the strength and seasonality of the submesoscale cascade which were chosen to reflect distinct background oceanographic conditions. We refer to these regions as the Agulhas Return Current (25.5° to 40.0°E , 35.7° to 45.9°S) and the Agulhas Ring Path (-7.8° to 6.8°E , 26.8° to 38.1°S). The Return Current is a strongly energetic region bisected by the meandering jet of the Agulhas Return Current, while the Ring Path is a relatively quiescent open ocean

region with weak background currents and strong mesoscale eddies shed from the Agulhas retroflection, these Agulhas Rings periodically transit through the Ring Path region. The Return Current lies within the larger domain used to train the diffusion prior while the Ring Path is outside the training domain. This requires the prior to extrapolate to potentially distinct regional dynamics. The quantitative evaluation of the state estimates in Section 4.4.1 for the Return Current region is conducted only on the withheld two-month spin-up phase of LLC4320 to avoid any contamination from the training data.

To ensure that the method is applied in windows where observations of SSH and SST provide strong constraints on the submesoscale surface ocean state, we sub-sample the two study regions described above in space and time to select the windows with good observational coverage of both SSH and SST. We use all SWOT and SEVIRI observations available during the year 2024 to select windows corresponding to the dimensions of x by requiring a minimum of 15% and 30% pixels to be observed by SWOT and SEVIRI respectively. For SWOT this corresponds roughly to at least three SWOT passes through the domain during the 2.5 day sequence reconstructed by the prior, and for SEVIRI corresponds to moderate but not severe cloud cover. We select as many windows as possible subject to these constraints while also minimizing overlap between selected windows and maximizing coverage of the full annual cycle. After this procedure we are left with 191 and 204 selected windows in the Ring Path and Return Current respectively. By reconstructing small, non-uniformly distributed patches from our larger study regions some noise and bias will be introduced in the estimate of the submesoscale cascade compared to what would be found by calculating it on the full domain. In Appendix C2 we demonstrate that estimating the cascade in these small windows provides a consistent result with that calculated from the full domain for scales below $\mathcal{O}(100)\text{km}$.

We evaluate the point-wise accuracy of the state estimates using the Continuous Ranked Probability Score [CRPS; Gneiting and Raftery, 2007; Hersbach, 2000], a scoring rule that jointly assesses the accuracy and calibration of probabilistic predictions by measuring the distance between the predicted ensemble distribution and the observed value. The CRPS decomposes into the mean absolute error of the ensemble mean minus a term rewarding ensemble spread that is proportional to forecast uncertainty, penalizing overconfident predictions where ensemble spread is low but errors are large. For a deterministic prediction, CRPS reduces to the mean absolute error, providing a unified framework for comparing GenLLC4320 ensemble predictions against deterministic products such as coarse gridded satellite products.

4.3.2 Application to real world satellite observations

We apply GenLLC4320 to real world surface ocean state estimation to provide observational constraints on the strength of the cascade. We apply the method in the same well-observed windows from both the Return Current and Ring Path regions using observations from 2024. To assimilate real observations we simply replace the simulated satellite observations generated from the simulation with SWOT and nadir altimetry and SEVIRI SST observations, and replace the simulated coarse gridded products with coarse satellite products of SSH, SST, SSS, and surface geostrophic currents.

To enable independent validation of the real world state estimates, we apply additional cloud masks to the SEVIRI SST observations and withhold SWOT observations on the second and fourth time-step of the reconstructed state. This provides independent high-resolution SSH and SST observations with which we can validate the GenLLC4320 reconstructed SSH and SST. Since in the validation experiment we withhold a significant fraction of available observations from the assimilation, the reconstruction skill will be a lower bound on the performance of GenLLC4320 when all available observations are assimilated. To avoid degrading the observations too drastically, we withhold the SSH and SST observations individually in two separate experiments, so we use degraded SSH with full SST and vice versa. We do not directly validate small-scale SSS or surface currents in the real world setting due to the sparsity of in situ observations and the lack of high-resolution ground truth from satellites.

4.4 Results

4.4.1 Generative data assimilation enables submesoscale-resolving surface ocean state estimation

We begin by evaluating GenLLC4320 in the OSSE setting (Section 4.3.1), where comparison against the known LLC4320 ground truth provides a controlled assessment of state estimation, presenting results for the Return Current region here. Comparison of the GenLLC4320 reconstructed fields to the unseen LLC4320 data shows that GenLLC4320 accurately captures much of the (sub)mesoscale variability in SSH, SST, SSS, and surface currents, illustrated here through the vorticity field (Figure 4.2). The close correspondence between prediction and ground truth for SSH and SST is perhaps not surprising, since SWOT SSH

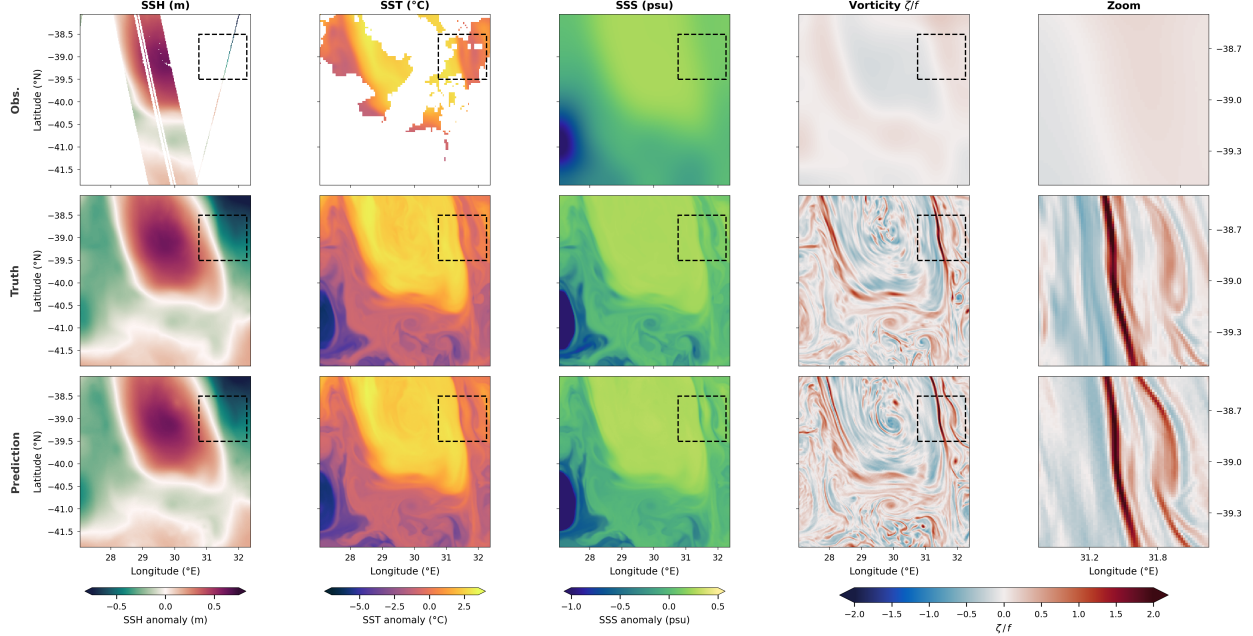


Figure 4.2: Example GenLLC4320 reconstruction from simulated satellite observations compared to LLC4320 ground truth. The panels from left to right show SSH, SST, SSS, vorticity, and a zoomed in regional crop of the vorticity fields with the simulated observations in the top row, the LLC4320 ground truth in the middle row, and the GenLLC4320 state estimate in the bottom row.

and SEVIRI SST observations from the two 12-hour steps before and after the presented snapshot are assimilated. This data provides broad spatial coverage of SSH and SST, although the diffusion prior must still leverage the learned time dynamics to assimilate these observations over multiple timesteps. For SSS, only a coarse gridded estimate was assimilated and GenLLC4320 downscales SSS to predict much of the submesoscale SSS variability including sharp fronts and filaments, as well as a coherent eddy $\sim 10\text{-}20$ km across at (30°E , 41°S) which was absent in the low-resolution SSS observations assimilated. The fidelity of the reconstructed SSS shows that coarse satellite SSS combined with high-resolution SST provides strong constraints on submesoscale SSS variance since the two are closely connected dynamically through density. Close inspection reveals some discrepancies at the smallest scales, likely reflecting processes that modulate SSS independently of SST and are therefore not constrained by observations. Small-scale surface ocean currents, that are not observed by satellites, are captured to a remarkable degree by GenLLC4320, capturing the strong vorticity signatures of submesoscale fronts and eddies by learning their relation to the well-observed SSH and SST fields. In the North East of the domain, a series of SST fronts with gradients of alternating

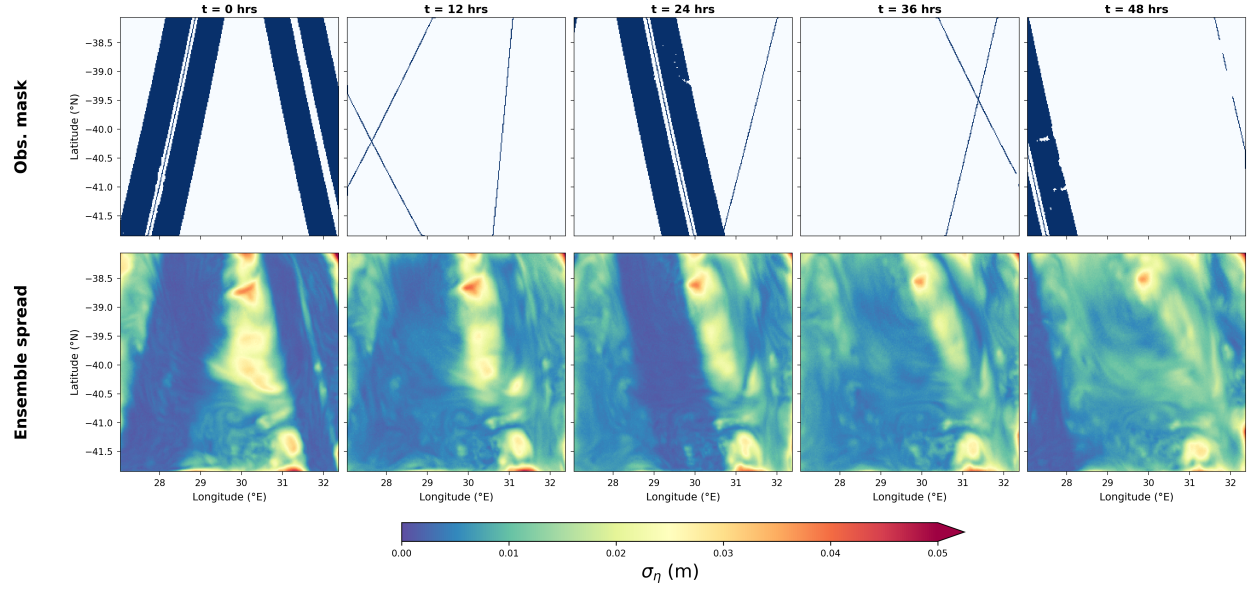


Figure 4.3: GenLLC4320 ensemble standard deviation for the reconstructed SSH field from a 20-member ensemble for the window shown in Figure 4.2. The top row shows the locations of the satellite altimetry observation assimilated.

sign is correctly translated into intense vorticity bands with Rossby number exceeding two in some locations, underlining their ageostrophic submesoscale dynamics. In some locations, such as at (30°E, 38.5°S), GenLLC4320 generates submesoscale eddies that have no corresponding feature in the ground truth. Since this location was observed in neither SSH nor SST, GenLLC4320 generates a variety of different physically plausible submesoscale realizations while remaining consistent with observational constraints. The ensemble spread of a 20-member GenLLC4320 ensemble is observed to be higher in this location reflecting the model’s enhanced uncertainty between satellite retrievals (Figure 4.3). Tracking the ensemble spread over time (Figure 4.3) illustrates that the video diffusion prior can effectively propagate observation constraints to nearby time-steps, evidenced by suppressed ensemble spread under each SWOT passover persisting for multiple time-steps.

To quantitatively evaluate reconstruction accuracy beyond a single snapshot, we evaluate reconstructions over 22 well-observed windows in the Return Current, generating a 20-member ensemble of state estimates for each. We compare the performance of GenLLC4320 to the coarse gridded satellite products (Figure 4.4). The spectra for the coarse satellite products for all variables strongly under-estimates small-scale variance when compared to the LLC4320 ground truth, while the GenLLC4320 ensemble members each exhibit

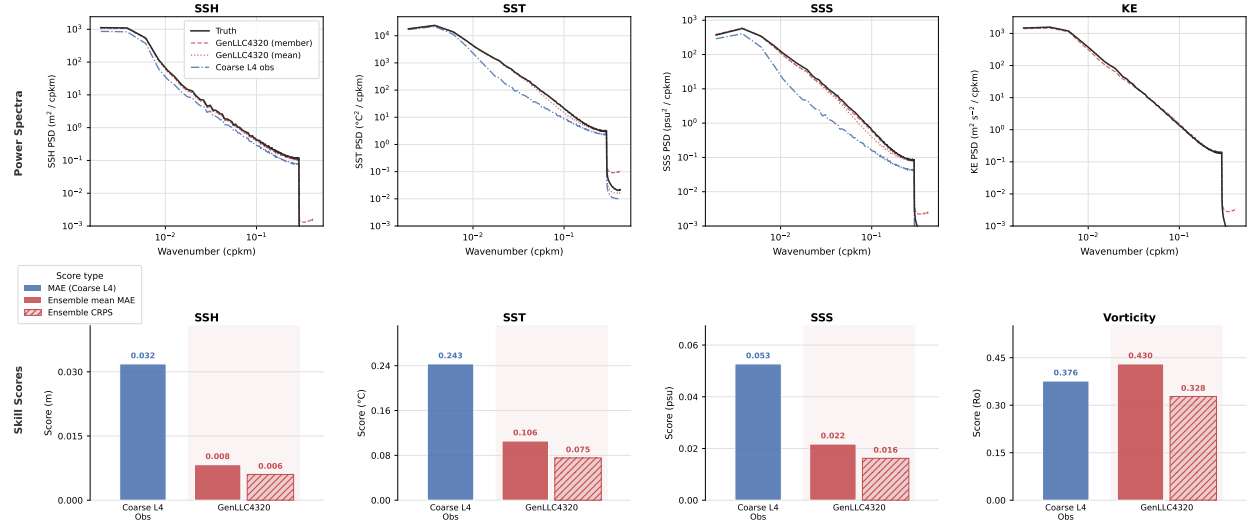


Figure 4.4: State estimation evaluation metrics for GenLLC4320 in the Return Current for the observing system simulation experiment, aggregated over 22 observation windows each with a 20-member ensemble. Top row: comparison of the power spectra for SSH, SST, SSS, and kinetic energy for the LLC4320 ground truth (black), GenLLC4320 ensemble members (red dashed), GenLLC4320 ensemble mean (red dotted), and coarse gridded satellite products (blue dot dashed). Bottom row: point-wise accuracy comparison comparing the mean absolute error of the coarse gridded satellite products (blue) to the mean absolute error of the GenLLC4320 ensemble mean (red solid), and the CRPS of the GenLLC4320 ensemble members (red hatched) for, from left to right, SSH, SST, SSS, and vorticity.

physically realistic spectra which closely track the ground truth down to the grid scale. The GenLLC4320 ensemble mean is smoother at the smallest scales, reflecting the under-constrained nature of small-scale variability in gaps between observations. Deterministic methods would suppress these small-scale features entirely and regress to the ensemble mean. Across all variables, the GenLLC4320 ensemble state estimates outperform the coarse gridded satellite products in point-wise accuracy, as shown by their lower CRPS values. For SSH and SST, the significantly lower CRPS values are in part due to the the assimilation of temporally adjacent observations, while SSS reflects the inherently low-resolution nature of satellite SSS observations. GenLLC4320 exploits high-resolution SST to recover much of the submesoscale SSS signal. For vorticity, the GenLLC4320 ensemble members have a higher mean absolute error than is obtained applying geostrophic balance to coarse gridded SSH, but lower CRPS confirms that GenLLC4320 generates a well-calibrated probabilistic prediction whose ensemble spread appropriately reflects reconstruction uncertainty.

When applied to real world satellite observations, GenLLC4320 state estimates retain their physical

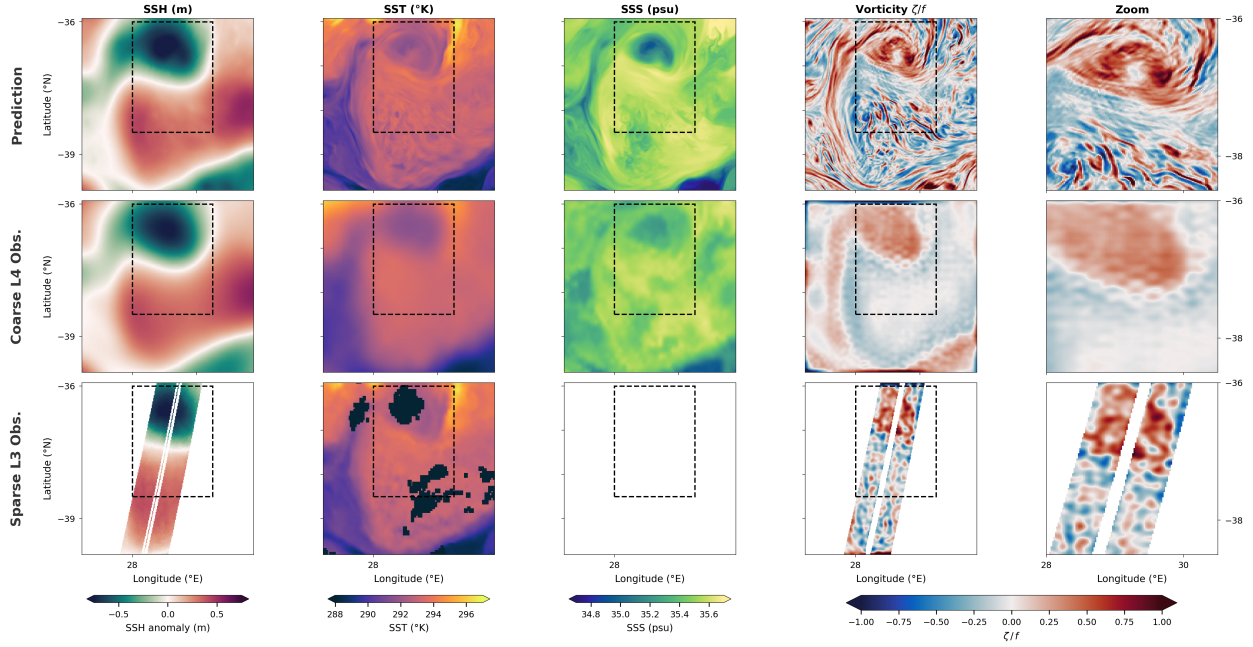


Figure 4.5: Example GenLLC4320 reconstruction from real satellite observations compared to operational gridded satellite products. The panels from left to right show SSH, SST, SSS, vorticity, and a zoomed in regional crop of the vorticity fields with the GenLLC4320 state estimate in the top row, operational gridded satellite products in the middle row, and sparse high-resolution satellite observations in the bottom row. The SWOT pass shown was not assimilated in the state estimation, and the vorticity in the bottom row is computed from the SWOT SSH assuming geostrophy after applying a 6 km Gaussian low pass filter to suppress noise from the instrument.

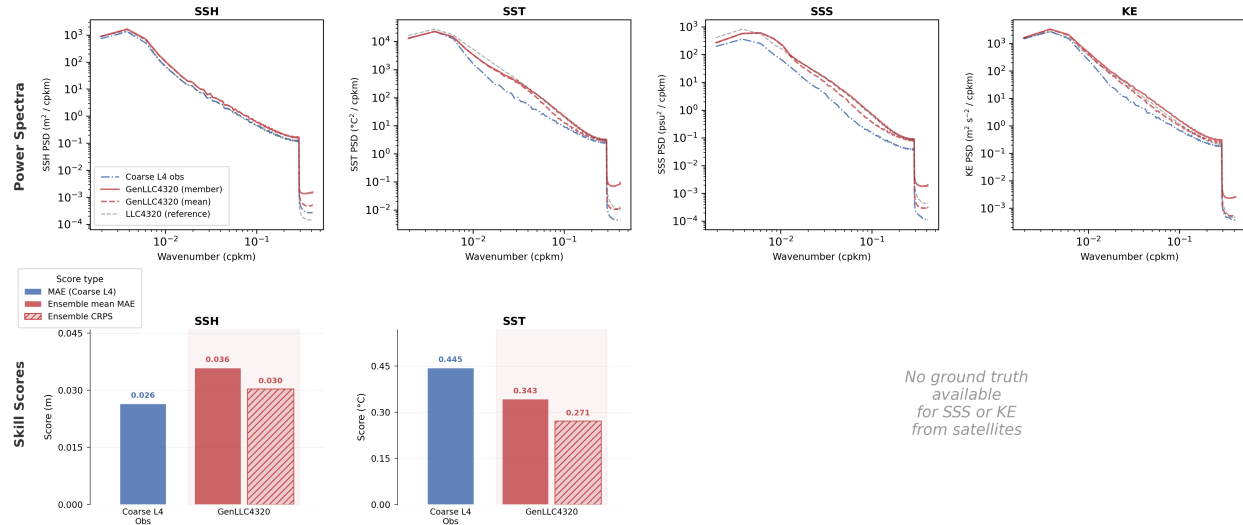


Figure 4.6: State estimation evaluation metrics for GenLLC4320 in the Return Current for the experiment with real world observations, aggregated over 20 observation windows each with a 10-member ensemble. Top row: comparison of the power spectra for SSH, SST, SSS, and kinetic energy for the coarse gridded satellite products (blue dot dashed), GenLLC4320 ensemble member (red solid), GenLLC4320 ensemble mean (red dashed), and LLC4320 is shown for reference (dashed gray) though in the real world context this is no longer the ground truth. Bottom row: point-wise accuracy comparison comparing the mean absolute error of the coarse gridded satellite products (blue) to the mean absolute error of the GenLLC4320 ensemble mean (red solid), and the CRPS of the GenLLC4320 ensemble members (red hatched) for SSH and SST using the withheld satellite observations. Point-wise accuracy for SSS and surface currents in the real world setting is not evaluated here due to the lack of satellite observations of these quantities.

plausibility, exhibiting rich mesoscale and submesoscale structure unresolved in operational gridded satellite products (Figure 4.5). Submesoscale eddy signatures visible in SEVIRI SST are reflected in the reconstructed SSS and surface vorticity fields without any high-resolution observations of these quantities. Direct validation of the reconstructed submesoscale vorticity is challenging given the absence of synoptic current observations, but applying geostrophic balance to an independent withheld SWOT pass reveals submesoscale eddy-like structures coinciding with the aggregation of submesoscale features generated by GenLLC4320 in the center of the domain (Figure 4.5). The stronger mesoscale eddy signature in the northern part of the domain reconstructed by GenLLC4320 relative to NeurOST is corroborated by the geostrophic vorticity from withheld SWOT SSH. Quantitatively, the spectra of the real world GenLLC4320 state estimates show significantly stronger variability at small scales than gridded satellite products for SSH, SST, and SSS, and for surface currents when compared to geostrophic currents from NeurOST (Figure 4.6). The CRPS and mean absolute error for SST are both significantly lower than that of the gridded satellite product when evaluated against withheld SEVIRI observations. In contrast, the SSH mean absolute error and CRPS appear marginally higher in GenLLC4320 when compared against withheld SWOT observations than for NeurOST. This likely reflects the limited spatial coverage when a full SWOT pass is withheld, leaving substantial portions of the domain unobserved at any timestep which the diffusion prior does not fully overcome. Nonetheless, the reconstruction errors for SSH are only marginally higher than NeurOST. Since cascade diagnostics are computed using all available SWOT observations with none withheld for validation, the observational coverage in the cascade analysis is more spatially comprehensive, covering most of the domain in a 2.5 day window.

4.4.2 State estimates capture the strength and seasonality of the submesoscale cascade in LLC4320

Despite assimilating no high-resolution observations of surface currents, GenLLC4320 reconstructs submesoscale surface ocean dynamics from simulated satellite observations with high fidelity; there is a striking similarity between snapshots of the predicted vorticity, strain, and divergence fields and the corresponding LLC4320 ground truth (Figures 4.7 & 4.8). GenLLC4320 captures the strength and geometry of mesoscale eddies and the Agulhas Return Current jet, while generating submesoscale eddies, fronts, and filaments with

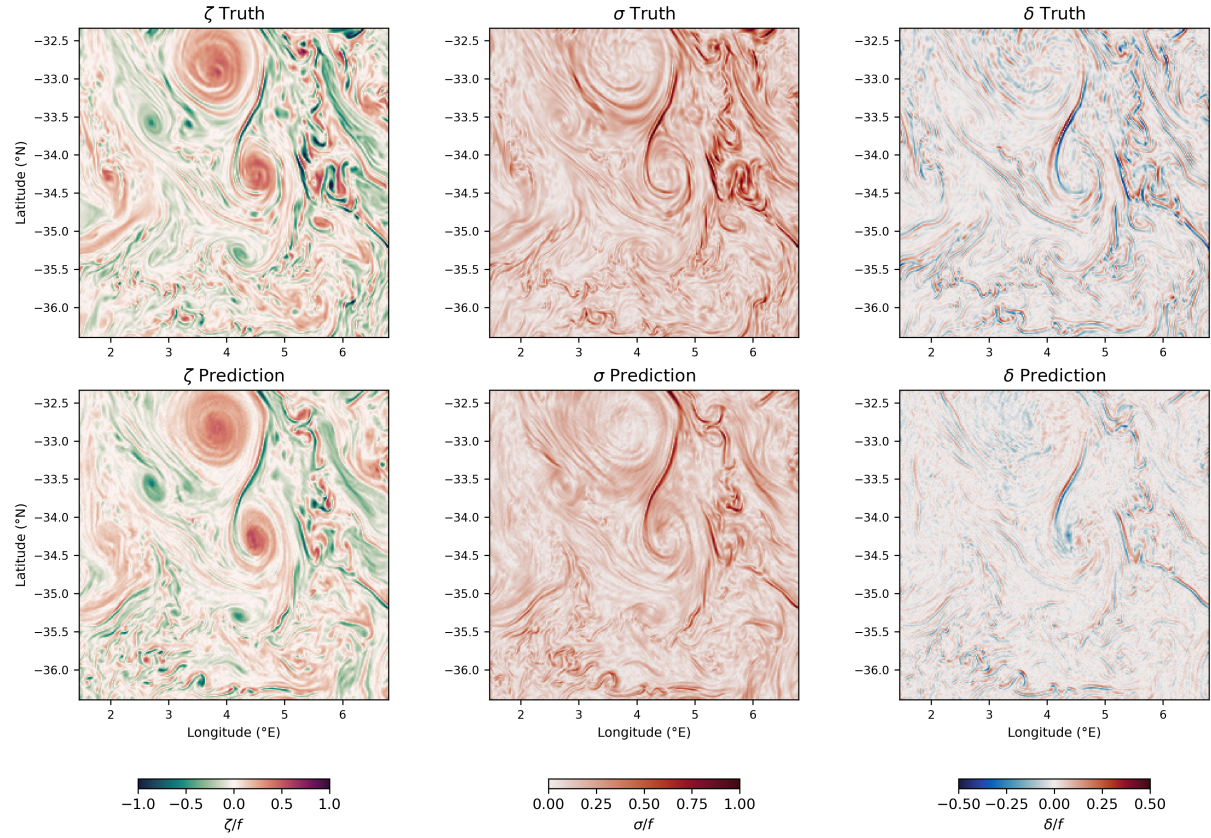


Figure 4.7: Vorticity, strain, and divergence for a snapshot in the Agulhas Ring Path comparing LLC4320 ground truth (top row) to GenLLC4320 prediction (bottom row).

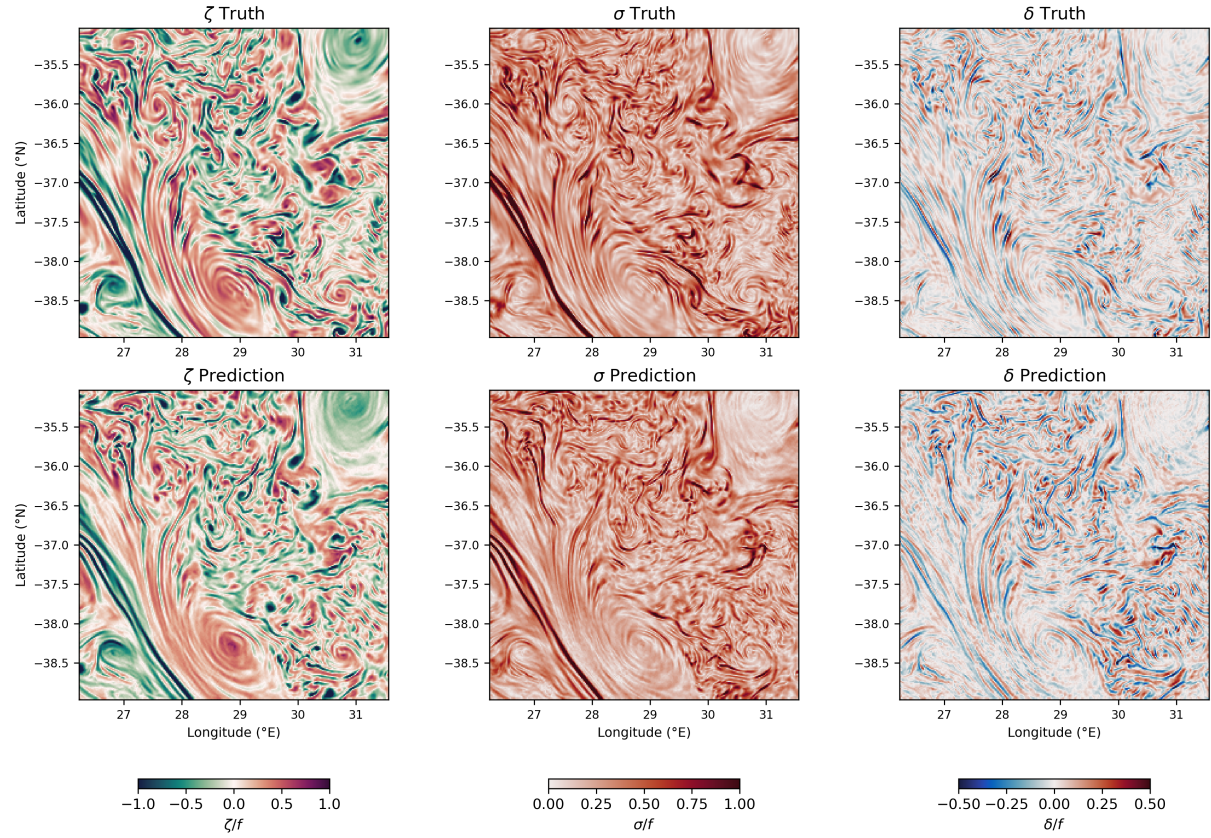
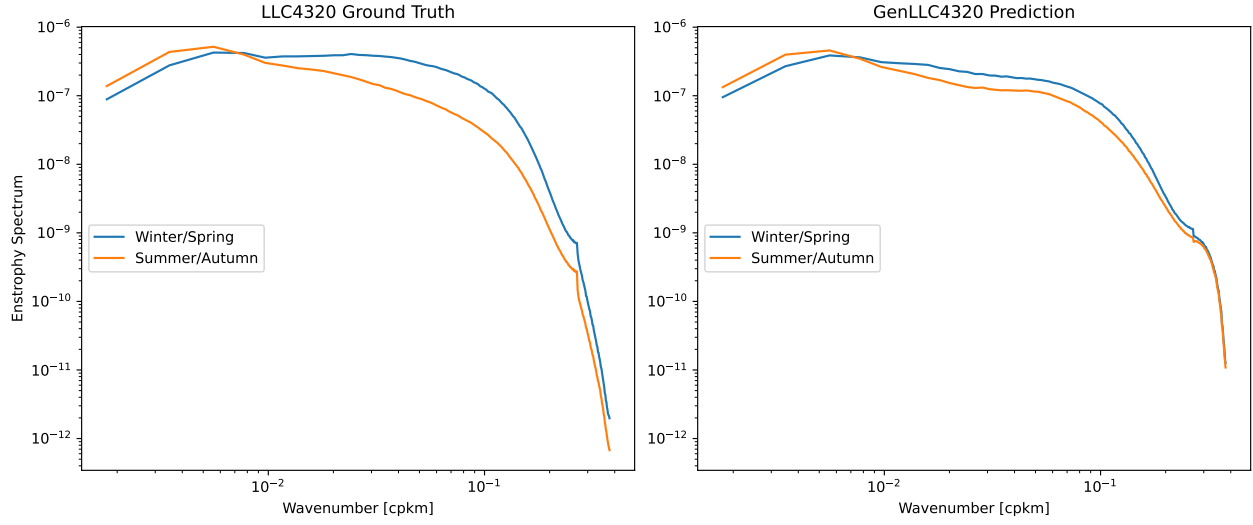
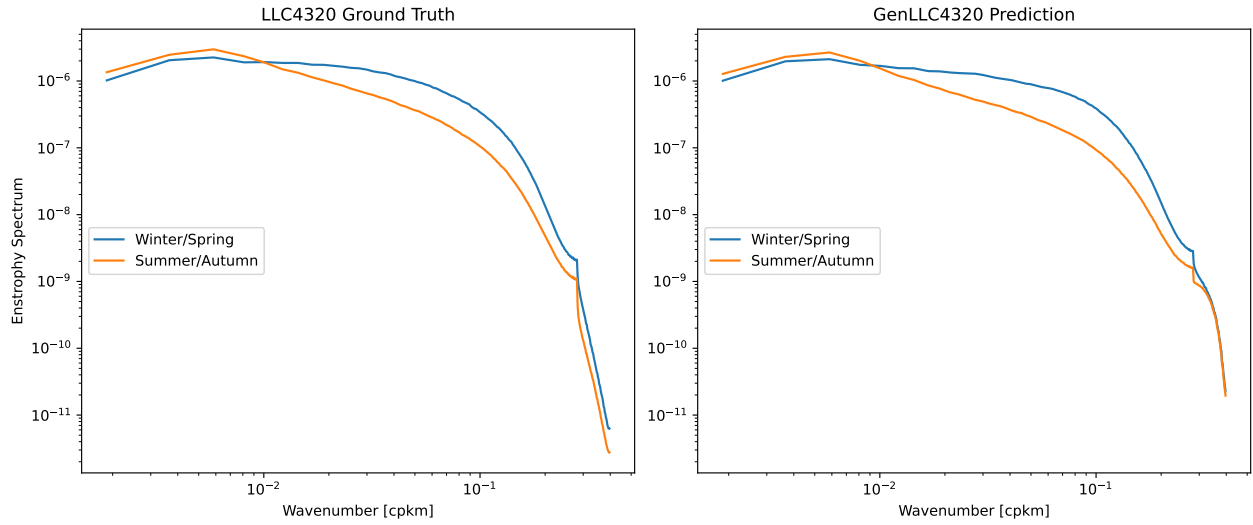


Figure 4.8: Vorticity, strain, and divergence for a snapshot in the Agulhas Return Current comparing LLC4320 ground truth (top row) to GenLLC4320 prediction (bottom row).

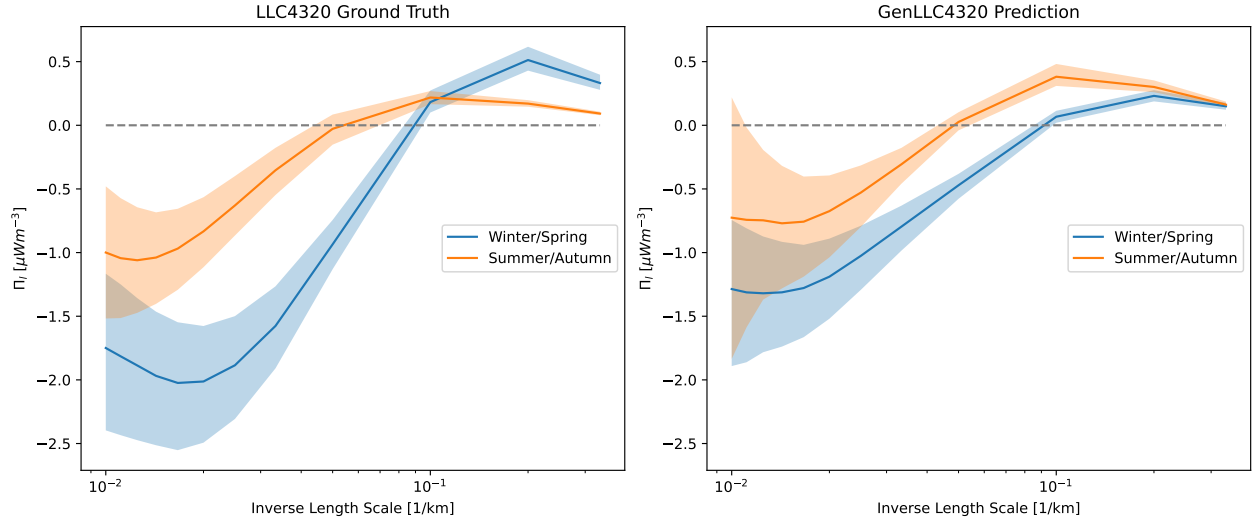


(a) Ring Path

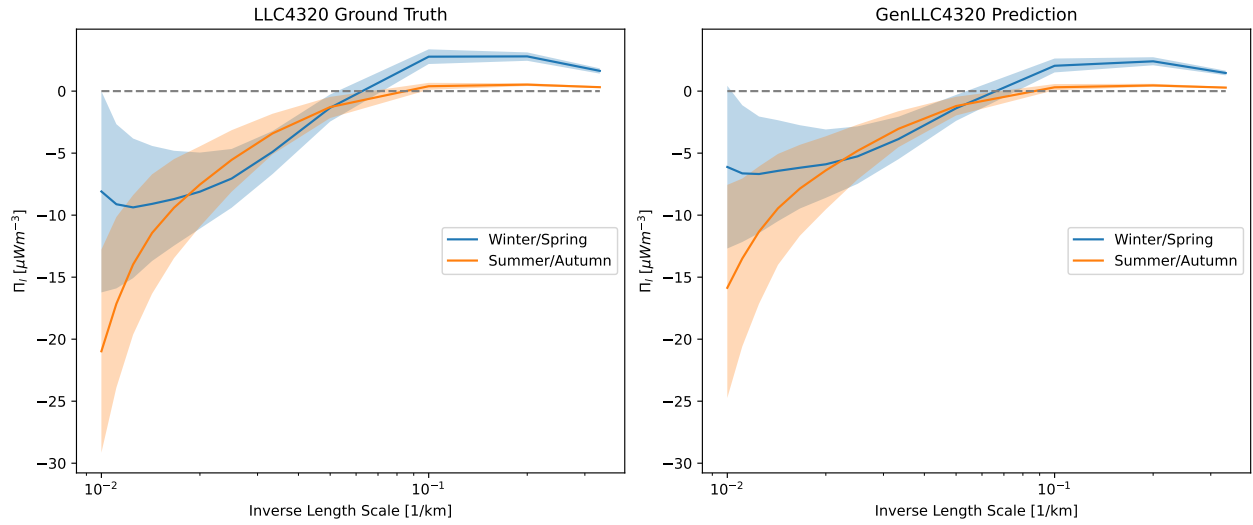


(b) Return Current

Figure 4.9: Enstrophy spectra comparing GenLLC4320 to LLC4320 ground truth in OSSE setting for the Agulhas Ring Path and Agulhas Return Current regions.



(a) Ring Path



(b) Return Current

Figure 4.10: KE cascade comparing GenLLC4320 to LLC4320 ground truth in OSSE setting for the Agulhas Ring Path and Agulhas Return Current regions.

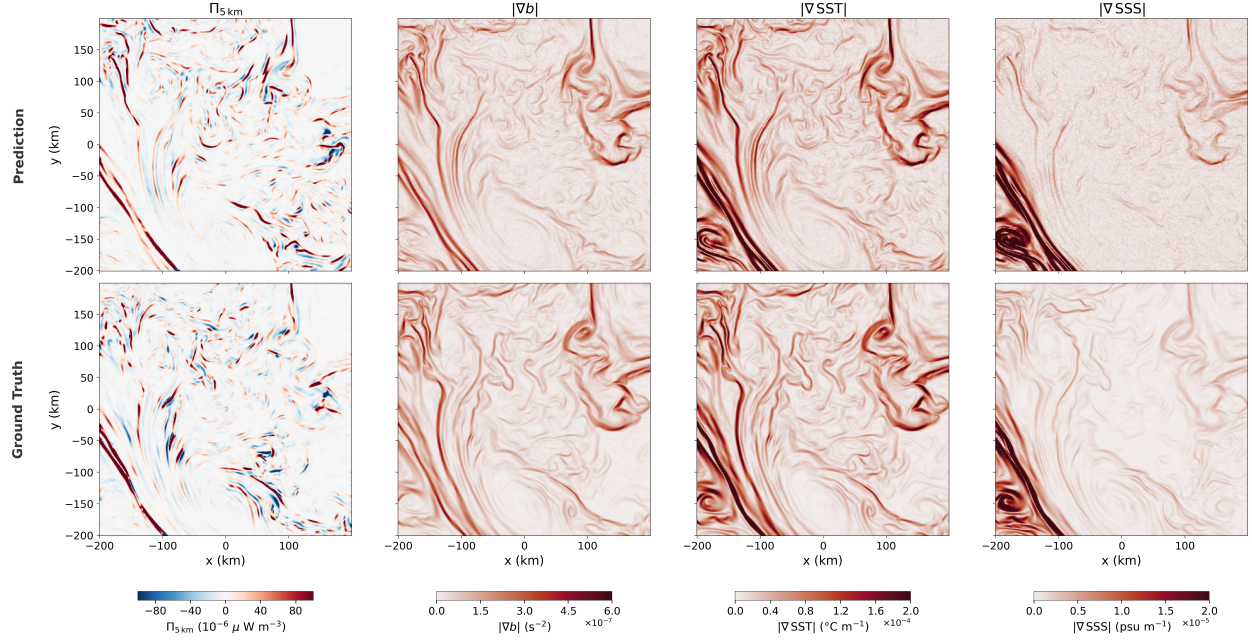


Figure 4.11: Snapshot in the Return Current showing the co-location of the downscale cascade at small scales with strong submesoscale buoyancy fronts. From left to right, spectral energy flux at 5 km scale, surface buoyancy gradient magnitude, SST gradient magnitude, and SSS gradient magnitude for the GenLLC4320 prediction (top row) and the LLC4320 ground truth (bottom row).

$\mathcal{O}(1)$ Rossby number, indicating that these features lie firmly in the submesoscale dynamical regime where ageostrophic effects are significant. The ageostrophic nature of the dynamics can be seen in strong frontal features in the reconstructed divergence field. The divergence would be uniformly zero if the prior were simply inferring surface currents using geostrophic balance from SSH. A clear seasonality in the submesoscale dynamics is also apparent in the enstrophy spectra (Figure 4.9), with shallower spectra in winter/spring reflecting the wintertime energization of submesoscale turbulence [Sasaki et al., 2014; Callies et al., 2015], captured by GenLLC4320 in both regions albeit with some underestimation of the seasonal contrast in the Ring Path.

The capability of GenLLC4320 to generate state estimates with physically plausible eddy dynamics gives an improved estimation of the strength and directionality of the kinetic energy cascade, diagnosed here using the coarse-graining framework [Aluie et al., 2018; Storer and Aluie, 2023; Storer et al., 2023]. The strength and sign of the spectral energy flux, Π_l , which quantifies the transfer of energy from scales larger than a given length-scale, l , with a positive (negative) sign indicating downscale (upscale) energy

transfer, is set by a complex tensor product between filtered and unfiltered velocities and velocity gradients, making it highly sensitive to accurately reconstructing velocity across scales and capturing the geometry of eddies [Aluie et al., 2018]. Despite this challenge, GenLLC4320 shows a remarkable ability to capture the cascade across scales in both the Ring Path and Return Current regions when compared to the LLC4320 simulation ground truth (Figure 4.10). In the Ring Path, the reconstructed cascade correctly captures the $\mathcal{O}(10)$ km scale at which the cascade transitions from an upscale to a downscale energy transfer, and reproduces the pronounced wintertime intensification of the upscale cascade, with the LLC4320 ground truth reaching $-2 \mu\text{Wm}^{-3}$ in winter/spring compared to $-1 \mu\text{Wm}^{-3}$ in summer/autumn. The Return Current cascade is stronger in both seasons, with the LLC4320 ground truth reaching an order of magnitude larger than that found in the Ring Path and with no clearly identifiable seasonality, reflecting the intense year-round (sub)mesoscale dynamics driven by the Agulhas Return Current jet in contrast to the relatively quiescent Ring Path. In both regions and seasons, the true cascade values lie mostly within the range of GenLLC4320 predictions, with the largest discrepancy being an up to $\sim 40\%$ underestimation of the wintertime upscale cascade in the Ring Path, consistent with the moderate underestimation of submesoscale vorticity identified shown in Figure 4.9. The strong downscale cascade at scales below $\mathcal{O}(10)$ km appears to be episodic and localized at strong submesoscale buoyancy fronts and filaments (Figure 4.11), consistent with prior studies [Schubert et al., 2020; Srinivasan et al., 2023]. Having established that GenLLC4320 can recover the strength, seasonality, and scale-dependence of the submesoscale kinetic energy cascade from realistic synthetic observations, we now apply the same framework to real satellite observations to provide observational constraints on the cascade.

4.4.3 Real world observations support the existence of a strong, seasonal upscale cascade at submesoscales

When applied to real world satellite observations, GenLLC4320 state estimates show strong signatures of submesoscale dynamics (Figures 4.12 & 4.13). The GenLLC4320 reconstructions include coherent submesoscale eddies with $\mathcal{O}(1)$ Rossby number as small as 10-20 km in size combined with strong frontal features with corresponding regions of elevated strain and divergence. This is in contrast to what is retrievable from state-of-the-art gridded altimetry (i.e. NeurOST), which captures only the vorticity and strain

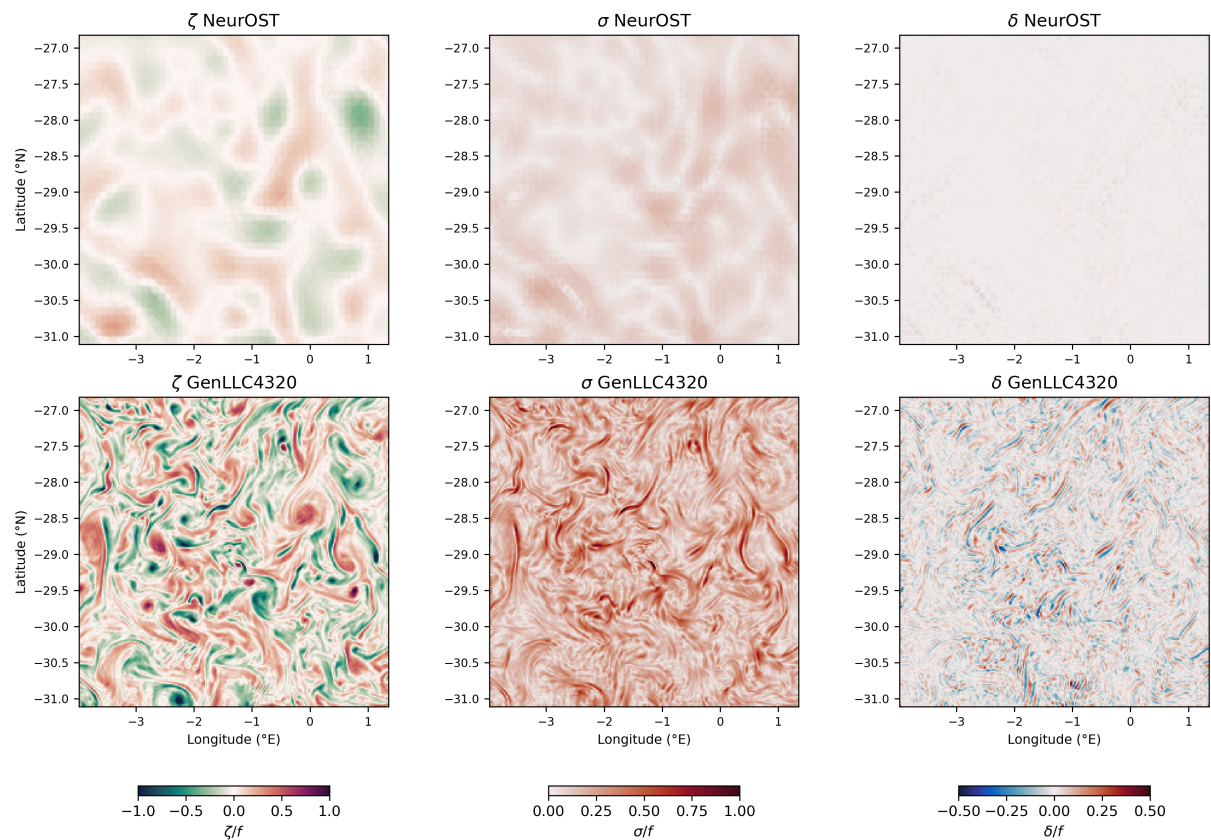


Figure 4.12: Vorticity, strain, and divergence for a snapshot in the Agulhas Ring Path for state estimates from real world satellite observations, comparing NeurOST gridded altimetry (top row) to GenLLC4320 prediction (bottom row).

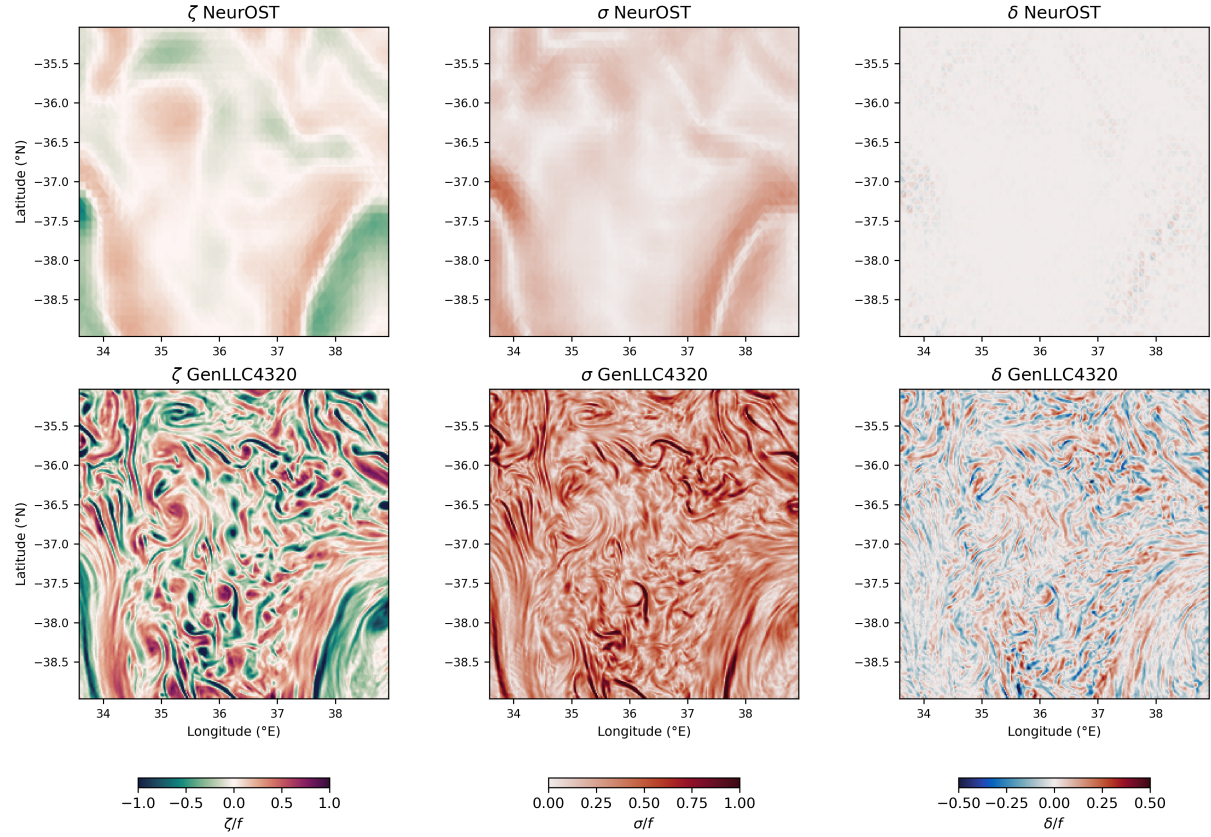
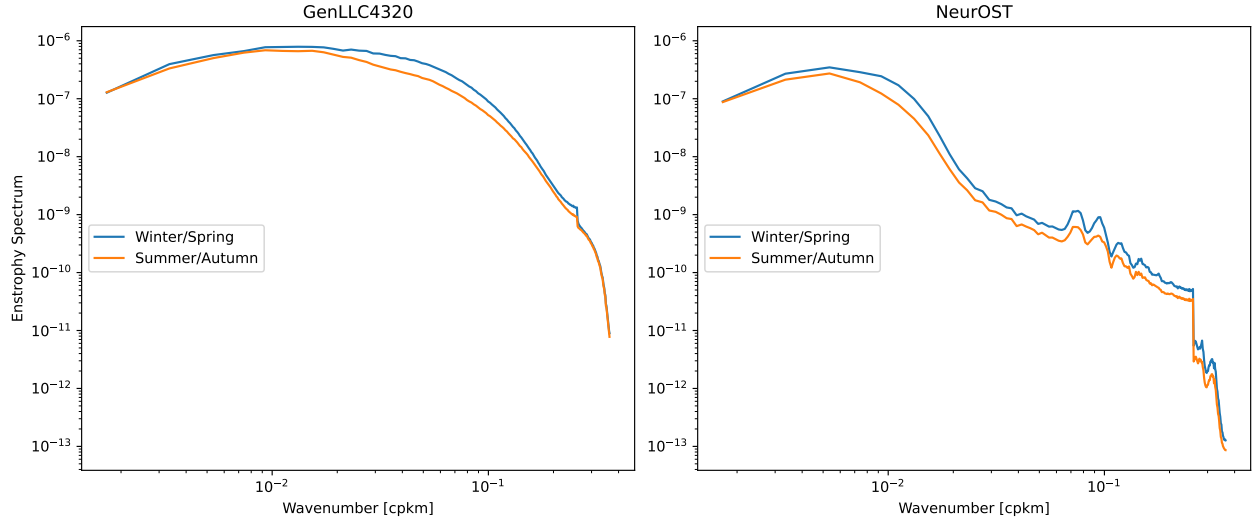
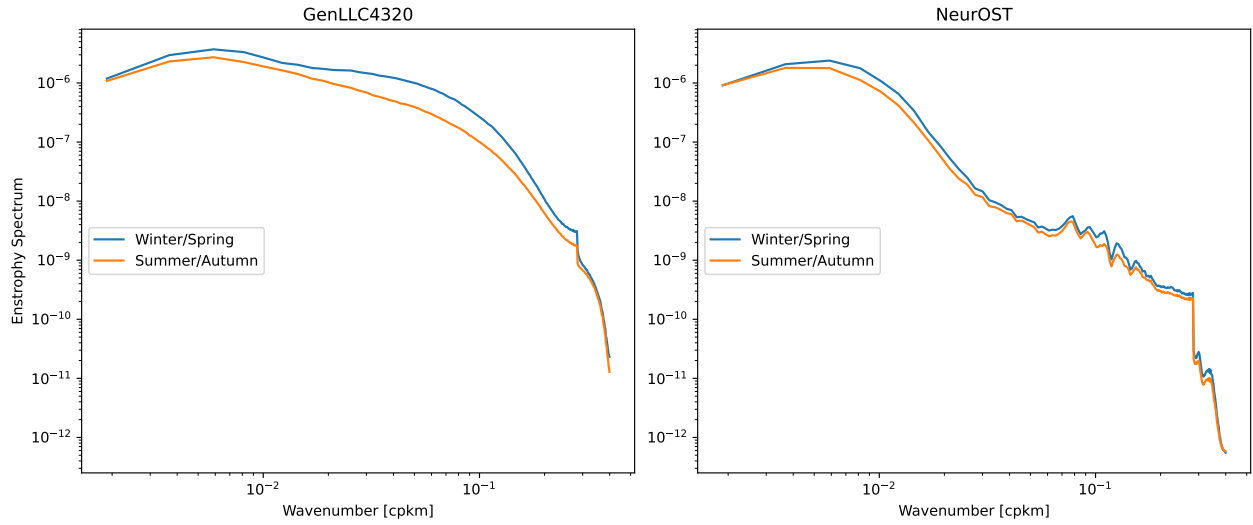


Figure 4.13: Vorticity, strain, and divergence for a snapshot in the Agulhas Return Current for state estimates from real world satellite observations, comparing NeurOST gridded altimetry (top row) to GenLLC4320 prediction (bottom row).



(a) Ring Path



(b) Return Current

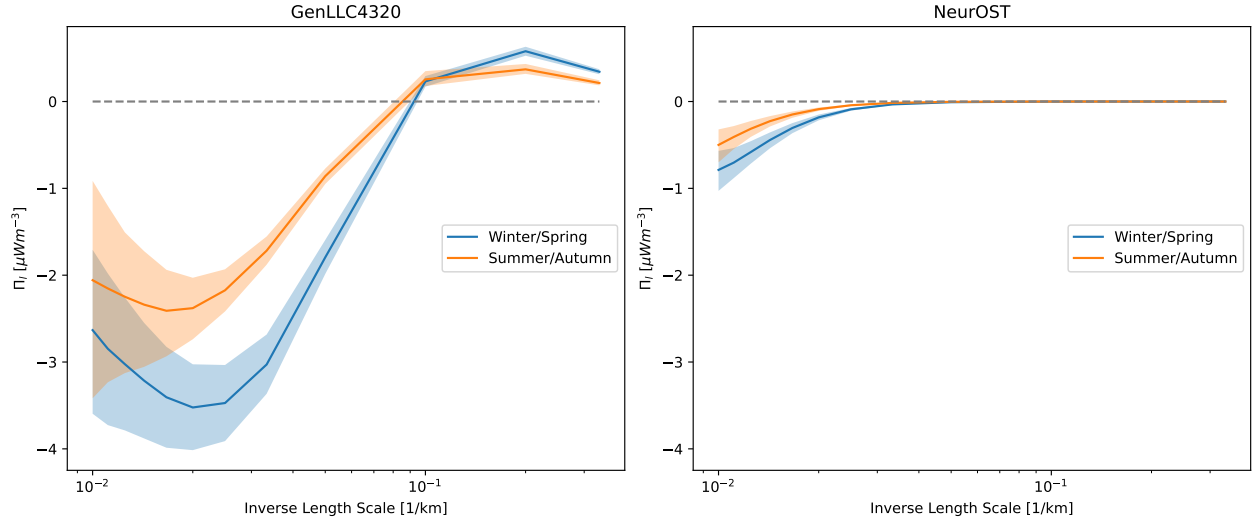
Figure 4.14: Enstrophy spectra for state estimates from real world satellite observations, comparing GenLLC4320 to NeurOST gridded altimetry for the Agulhas Ring Path and Agulhas Return Current regions.

signatures of larger mesoscale features and has zero divergence by construction, as geostrophic currents are divergence-free. The wintertime energization of submesoscales noted in LLC4320 is also apparent in the real world GenLLC4320 reconstructions, with a shallower enstrophy spectral slope in winter/spring than summer/autumn, whereas NeurOST only captures a faint seasonal contrast at scales above 100 km (Figure 4.14). GenLLC4320 thus provides a new observational window on seasonally-varying submesoscale surface ocean dynamics.

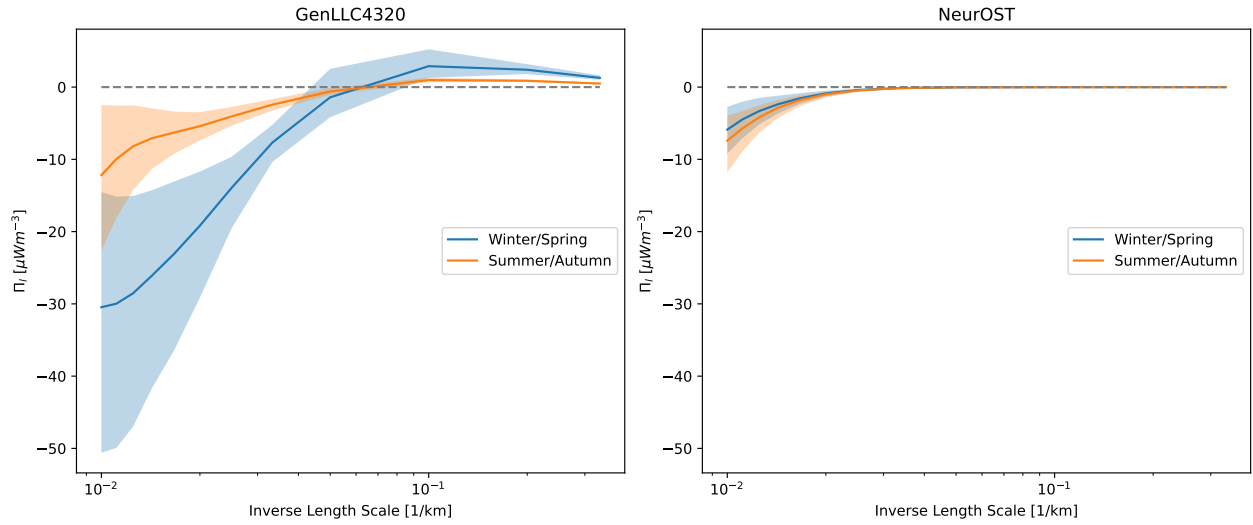
GenLLC4320 reveals a strong upscale kinetic energy cascade extending down to ~ 10 km in both regions, with a marked seasonal intensification that is entirely absent from currently available gridded satellite products (Figure 4.15). NeurOST, which resolves only mesoscale variability above ~ 100 km, yields a cascade signal that is essentially indistinguishable from zero across the submesoscale range in both regions and seasons, capturing neither the strength nor the seasonality of the submesoscale cascade. GenLLC4320, in contrast, reveals a strong upscale cascade in the Ring Path peaking at $-3.5 \mu\text{Wm}^{-3}$ in winter and spring which is 40% stronger than that found in summer and autumn. The cascade reverses sign to a downscale energy transfer at scales below ~ 10 km. The Return Current cascade, as in the LLC4320 simulation, is significantly stronger than in the Ring Path, reaching $\sim -30 \mu\text{Wm}^{-3}$ in winter and spring and $\sim -10 \mu\text{Wm}^{-3}$ in summer/autumn, a roughly threefold seasonal contrast. While the LLC4320 simulation exhibits no seasonality in the cascade in the Return Current (Section 4.4.2), the real world GenLLC4320 reconstructions show a pronounced winter and spring intensification similar to that found in the Ring Path. This analysis is confined to only a single seasonal cycle but this region exhibits significant non-seasonal variability driven by the turbulent jet dynamics of the Return Current, this discrepancy may reflect natural interannual variability rather than a systematic bias in the simulation. In both regions, the magnitudes of the observed cascade exceed those in the LLC4320 simulation by ~ 75 -100%.

4.5 Discussion

The submesoscale kinetic energy cascade from our satellite observation-constrained generative state estimates emerges as a seasonally-intensified feature of the upper ocean, with an upscale cascade extending to $\mathcal{O}(10)$ km that is not found in state-of-the-art gridded altimetry products. The presence of a strong upscale submesoscale cascade in both energetic and quiescent open ocean regions suggests that submesoscale



(a) Ring Path



(b) Return Current

Figure 4.15: KE cascade estimated from real world satellite observations using GenLLC4320 and NeurOST in the Ring Path and Return Current regions.

dynamics, with spatial scales well below the resolution of global climate models, systematically energize larger-scale mesoscale variability through an upscale cascade that is not represented in existing eddy parameterizations [Gent and McWilliams, 1990; Fox-Kemper et al., 2008]. The wintertime intensification of this upscale cascade is broadly consistent with the proposed mechanism of submesoscale eddy absorption by mesoscale eddies demonstrated in high-resolution simulations [Sasaki et al., 2014; Qiu et al., 2014; Schubert et al., 2020]. Although the precise formation mechanisms driving this wintertime submesoscale energization are likely heterogeneous [Lapeyre and Klein, 2006b; Boccaletti et al., 2007; Thomas et al., 2008; Taylor and Thompson, 2023], the preferential generation of submesoscale eddies in winter is well established in both simulations and observations [Callies et al., 2015]. The results presented here confirm that much of this energy cascades upscale to energize the mesoscale. The contrasting seasonality between the Ring Path, where the cascade follows a clear wintertime submesoscale cycle, and the Return Current, where year-round jet dynamics appear to drive a strong cascade independently of mixed layer forcing, suggests that the drivers of submesoscale cascade variability are regime-dependent, with the seasonal imprint on mesoscale energetics of submesoscale eddy absorption likely most pronounced in quiescent open ocean regions away from strong background currents. The downscale cascade below $\mathcal{O}(10)$ km found in our state estimates coincides with submesoscale fronts and filaments, consistent with results from numerical simulations suggesting that fronts play an out-sized role in cascading energy down to dissipation scales [Schubert et al., 2020; Srinivasan et al., 2023]. Thus, our results are qualitatively consistent with prior observational constraints from in situ studies pointing to a bi-directional, seasonally-varying submesoscale cascade [Balwada et al., 2022; Naveira Garabato et al., 2022]. The use of satellite observations opens the possibility of extending to regions without in situ campaigns. By assimilating SWOT observations across successive time-steps, we can resolve a wider range of scales than studies confined to individual SWOT swaths [Qiu and Chen, 2025], with SST providing an additional submesoscale constraint between SWOT passes.

The assimilation of satellite observations leads to a 75 – 100% increase in the strength of the submesoscale kinetic energy cascade compared to that in the LLC4320 training data. This result should be treated with some caution given that our analysis spans only a single seasonal cycle and we have not examined the impact of natural interannual variability. Nonetheless, the magnitude of the increase is substantial and in the Ring Path the GenLLC4320 state estimates have a cascade magnitude that closely tracks, in both seasons,

results obtained from a multi-year $1/60^\circ$ resolution regional simulation [Schubert et al., 2020], supporting the interpretation that this discrepancy reflects a genuine bias in LLC4320 rather than a sampling artifact. An under-estimation of the strength of the submesoscale cascade in LLC4320 is consistent with its $1/48^\circ$ grid resolution, which only partly resolves submesoscale mixed layer dynamics and likely generates systematically under-energetic submesoscales. Indeed, studies comparing the model to real world SWOT SSH [Archer et al., 2025] and in situ observations [Luecke et al., 2020] draw similar conclusions, highlighting that state-of-the-art high-resolution global ocean models still under-resolve submesoscale dynamics. A second, and not mutually exclusive, explanation for the discrepancy could come from the limited duration of the simulation that covers fourteen months. If the submesoscale cascade plays a meaningful role in setting the seasonality of mesoscale eddies, as has been hypothesized [Sasaki et al., 2014], then it is likely that a submesoscale-resolving model would need to be integrated for multiple years to get a realistic, equilibrated view of mesoscale-submesoscale ocean scale interactions. This highlights the need for multi-year global submesoscale-resolving integrations. These simulations will require significantly faster models [Silvestri et al., 2025].

While our diffusion prior was trained to generate SSH and surface currents with super-inertial internal gravity wave signatures filtered out, the real SWOT SSH observations assimilated in Section 4.4.3 contain some internal wave contamination. Given the temporal sparsity of SWOT sampling, these signals will be difficult to remove, although our study region was chosen to partially mitigate this contamination. In this region strong balanced eddy dynamics are large relative to internal waves. Prior studies show that balanced eddy turbulence is the leading order driver of SSH variance in the 10-100 km wavelength range and with the contamination by internal waves most significant in the summertime [Torres et al., 2018]. In Appendix C3 we present an observing system simulation experiment to demonstrate that submesoscale dynamics inferred using GenLLC4320 are not substantially degraded by assimilating SWOT observations containing super-inertial SSH signals, internal wave contamination nonetheless remains a limitation of this study, and robust methods for separating balanced and unbalanced signals in SWOT SSH remain an active area of development [Lguensat et al., 2020; Wang et al., 2022a; Gao et al., 2024; Lyu et al., 2024; Wang et al., 2025]. A key limitation of generative data assimilation is the reliance on numerical simulation data to train the diffusion prior, which potentially encodes model biases into the resulting state estimates. As a result,

dynamical processes which are poorly represented in the LLC4320 training data — for instance symmetric instability, eddy-wave interactions, or summertime mixed layer instability — will likely be systematically under-represented in the generative state estimates. In cases where submesoscale features are directly observed, such as in SWOT SSH or SEVIRI SST, the assimilation process can override statistical biases in the prior, producing state estimates that reflect the observed statistics of submesoscale eddies and fronts rather than the simulation statistics. Fundamentally, generative data assimilation results should be interpreted through a similar lens to those from physics-based data assimilation: the state estimates are observationally constrained but carry the biases of the underlying GCM, and where observations provide strong constraints the estimates will adapt, but in observation gaps the prior dominates. The development of diffusion priors that can either be updated or learned entirely from partial observations could reduce the reliance of generative data assimilation on simulation training data [Rozet et al., 2024].

The results of this study establish generative data assimilation as a valuable new approach to surface ocean state estimation that can resolve the submesoscale. Conventional data assimilation systems rely on physics-based dynamical models as priors which must close three-dimensional mass and energy budgets, which in practice limits state estimation to resolutions coarser than the submesoscale. Regression-based deep learning approaches [Fablet et al., 2024; Lenain et al., 2025] can achieve high resolution but produce deterministic point estimates that provide no uncertainty quantification and in large observation gaps regress toward the ensemble mean, which is not itself a physically consistent ocean state. By employing a generative prior trained on high-resolution physics-based simulations, generative data assimilation enables sampling from a distribution of states that are consistent with available observations but which are each plausible under the learned prior. By targeting surface-only state estimation and training on output from free-running high-resolution simulations, generative data assimilation sidesteps the computational burden of full three-dimensional state estimation while achieving submesoscale resolution that conventional systems cannot. As high-resolution ocean simulations and the global satellite observing system continue to improve, generative data assimilation offers a scalable route to surface ocean state estimation that advances with both.

Several natural extensions and opportunities for future research follow directly from this study. As the length of the SWOT observation record continues to grow and since the framework presented here could be re-trained to target other ocean basins, generative data assimilation can provide statistically robust multi-year

constraints on the seasonality of the cascade, explore its interannual variability, and provide an increasingly detailed picture of its geographical variability. While our study focuses exclusively on surface ocean dynamics, submesoscale fronts have been found to drive strong vertical transports in the ocean interior with consequences for global climate and marine ecosystems [Klein and Lapeyre, 2009; Mahadevan and Tandon, 2006; Su et al., 2018; Siegelman et al., 2020a]. Since submesoscale fronts typically have strong surface expressions, developing three-dimensional generative data assimilation systems or applying idealized methods for inferring interior dynamics from high-resolution surface fields [Wang et al., 2013; Qiu et al., 2020; Carli et al., 2024] could help in future to quantify the strength and variability of the submesoscale vertical pump. While here we leveraged the global LLC4320 simulation, a promising direction for training high-fidelity diffusion priors that fully resolve submesoscale dynamics would be to train on higher-resolution regional simulations run under a range of background conditions and forcing regimes, which could be efficiently generated using GPU-accelerated regional ocean modelling frameworks [Wagner et al., 2025; Klion et al., 2025]. Together, these directions offer a route to the global, multi-year observational constraints on submesoscale dynamics that ocean circulation theory and climate model development have long required.

Appendix C

Supporting Information for Chapter 4

C1 Extended Methods

C1.1 Video diffusion prior architecture and training details

We employ a UNet backbone neural network architecture with the augmentations widely used in the diffusion model literature [Song et al., 2020]: self attention near the bottleneck, noise level conditioning by performing a sinusoidal embedding of the scalar noise level to create a conditioning image tensor with multiple frequency channels, and replacing batch normalization with layer normalization to improve training stability when batches contain samples with very different noise levels. To process the time dimension in our architecture, we apply pixel-wise temporal attention at a number of resolution layers within the encoder-decoder path, folding time along the batch dimension for all other operations to process each timeframe independently using the same weights. We follow the Elucidated Diffusion Modeling (EDM) framework [Karras et al., 2022] for training our prior, training a denoiser neural network with additive Gaussian noise applied at noise levels between σ_{min} and σ_{max} . We follow the arguments of Brenowitz et al. [2025] when selecting these hyperparameters for geophysical data (see Table C.1) and replace the log-normal noise distribution used in Karras et al. [2022] with the log-uniform distribution used in Brenowitz et al. [2025]. We additionally condition our diffusion prior on time of day, and day of year through additional sinusoidal embeddings [Brenowitz et al., 2025] to provide context on the diurnal and seasonal cycle both of which impact submesoscale surface ocean dynamics.

During model development we found that training a single diffusion backbone over the full range of noise levels, σ , led to unsatisfactory performance, underlining the challenge of training a single model to jointly handle a wide range of dynamics from mesoscales down to submesoscales. We thus instead pursued a ‘mixture of experts’ approach, where separate models were trained for low and high noise levels respectively, allowing each to focus on a distinct scale range, since noise levels in diffusion models are loosely equivalent to spatial scales with higher noise levels corresponding to larger scales. This strategy is similar to that employed to prevent overfitting at large scales in Brenowitz et al. [2025] except here we train separate models from scratch on the different noise ranges rather than training a single model and picking different checkpoints for different noise ranges. The small and large σ experts are trained with overlapping σ ranges, and at inference we switch which model is called at a prescribed handover σ at the center of the overlap range. During training, we also conditioned the model on patch-averaged surface wind stress, surface heat flux, and surface freshwater flux, however at inference we found these fluxes to be largely ignored by the model and all results presented in the main text used placeholder values of zero for these fluxes for simplicity of implementation.

Hyperparameter	Value
σ_{min}	10^{-3}
σ_{max}	200
Resolution levels for self attention	32, 16
Resolution levels for temporal attention	64, 32, 16
Number of channels in first UNet block	128
Channel multipliers per resolution	1, 2, 2, 3, 4
UNet blocks per resolution level	3
Learning rate	10^{-4}
Batch size	48
Training resources	3 days on 8 A100 GPUs for each of the 2 experts
Small σ expert noise range	$[10^{-3}, 1]$
Large σ expert noise range	$[10^{-1}, 200]$
Handover σ between experts	0.5

Table C.1: Training and architecture hyperparameters

C1.2 Score-based data assimilation hyperparameters

At inference, we use the score-based data assimilation (SDA) framework [Rozet and Louppe, 2023a] to guide the generation of the diffusion prior with sparse observations. Since the SDA algorithm was developed for the variance-preserving stochastic differential equation (VPSDE) sampler, as opposed to the probability flow sampler used in EDM [Karras et al., 2022], we translate our EDM diffusion prior to the VPSDE formulation using the formula derived in the appendix of Manshausen et al. [2025]. Table C.2 details the values used in this study for the SDA hyperparameters which are described in Rozet and Louppe [2023a]. These values were arrived at through explorations carried out on simulated observations from the training partition of the LLC4320 data. We found the fidelity of reconstructions to improve as Γ was reduced, but with increasing instances of numerical instability below the value used here, and even with this value there were a small minority of ensemble members that went numerically unstable at inference and we prune before the analyses presented in the main text since they are easily identifiable by their pixelated, un-physical predictions. We found applying Langevin Monte Carlo corrector steps only at low noise levels beneath $t = 0.5$ helped to stabilize sampling while retaining the benefits of corrector steps; with no corrector steps the generated states tended to smooth out variance at submesoscales.

Hyperparameter	Value
Γ	0.05
η	10^{-4}
Number of steps	196
Number of corrector steps above $t = 0.5$	0
Number of corrector steps below $t = 0.5$	5
τ	0.2

Table C.2: SDA inference hyperparameters

C1.3 Coarse-graining computation of kinetic energy cascade using FlowSieve

We use *FlowSieve*, a parallelized coarse-graining and spectral energy flux computation code [Storer and Aluie, 2023], to compute the strength and direction of the kinetic energy cascade which is quantified through the spectral energy flux, Π_l . The mathematical underpinnings of this framework and its comparison to traditional Fourier methods for estimating the spectral energy flux are outlined in Aluie et al. [2018] and the

sensitivity of the cascade to strain and divergence are further explored in Srinivasan et al. [2023]. We re-grid all data onto a regular local Cartesian grid with 1 km resolution before all coarse-graining calculations using the ENU projection. For coarse gridded products, high-frequency interpolation artifacts appear after regridding to 1 km which we remove before spectral flux computations by applying a Gaussian low-pass filter with 3-pixel filter width.

C1.4 Details of the satellite data products used in this study

We use real satellite observations from the year 2024 in this study, and in the observing system simulation experiment take the sampling masks from this year of observations and apply it to the simulation.

The sparse, high-resolution SST observations are taken from the SEVIRI instrument onboard the Meteosat Second Generation (MSG) satellites. These satellites are on a geostationary orbit, providing continuous in time observations, albeit with coverage of only a limited geographical domain. We use the Level 3 collated product produced by OSI SAF which aggregates the observations onto a 0.05° grid with hourly resolution, applies numerous atmospheric corrections, applies a cloud mask, and provides quality flags for each data point. Since our diffusion prior is trained to produce 12-hourly means, we composite the SEVIRI observations within a 12-hour window, discarding measurements not flagged "good" or "very good" and discarding grid points with fewer than 3 observations within the 12-hour window. We note that the cloud masking algorithm used in the production of this dataset appears to sometimes artificially mask out some oceanographic features like mesoscale eddies and submesoscale fronts which could in future be alleviated by using lower-level data products as was done in Lenain et al. [2025], though we prefer here to retain the atmospheric and sensor corrections made at Level 3.

The sparse, high-resolution SSH observations are taken from both conventional nadir altimeters and SWOT. The nadir altimeter data is taken from the CMEMS Level 3 product where numerous tidal and atmospheric corrections have been made as well as calibration between the different altimeter missions. For SWOT, we take wide-swath observations from the AVISO Level 3 2.5 km Expert product (v2.0) where tidal corrections, inter-calibration with conventional altimeters, atmospheric corrections, orbit roll correction, and quality flagging have all been implemented. Note, we use here the 'unfiltered' SSH anomaly variable, preferring not to rely upon the experimental convolutional neural network denoising method implemented in

the product which could obscure real signals of submesoscale dynamics. We use the 2.5 km product rather than the 250 m product since the grid resolution of LLC4320, and hence our diffusion prior, is considerably larger at 1.5-2 km and this mitigates potential contamination from surface waves.

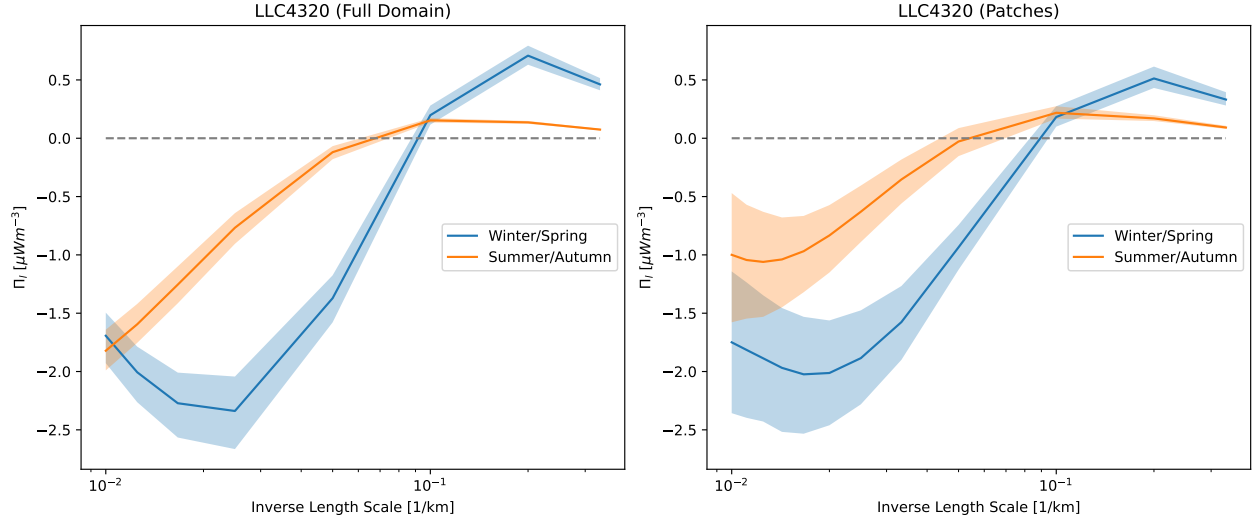
In addition, we leverage coarse gridded satellite products for SSH, SST, and SSS both as a large-scale constraint in the assimilation and to provide a comparison to GenLLC4320 state estimates. For SSH, we use the NeurOST product which is produced by synthesizing along-track nadir altimeter observations with gridded SST from the NASA MUR product using a neural network trained on past observations. The grid resolution is $1/10^\circ$, though the effective resolution is 90-120 km in wavelength in the region considered here. We also use surface geostrophic currents from the NeurOST product which have been validated against surface drifters [Martin et al., 2024], though we coarse-grain these with a 100 km Gaussian filter before assimilation to ensure we only assimilate geostrophic currents at larger scales where geostrophy is valid and where the NeurOST estimates will be most reliable. For SST, we use the REMSS MW OI gridded product which uses optimal interpolation to merge measurements from a number of missions equipped with microwave SST sensors into a $1/4^\circ$ product. While microwave observations are lower resolution than infrared, they can penetrate clouds, meaning this product appears to have a temporally consistent effective spatial resolution which is important for this study since we assimilate the product as a large-scale constraint. For SSS, we use the CMEMS Multi Observation Global Ocean Sea Surface Salinity and Sea Surface Density Level 4 product which uses a multivariate optimal interpolation algorithm combining satellite and in situ salinity observations with additional constraints from satellite SST to produce a $1/8^\circ$ product. When assimilating these coarse gridded satellite products in the generative data assimilation process we first coarse-grain the state estimate using prescribed coarse-graining scales chosen to approximate the effective resolution of the above products. These scales are detailed in Table C.3.

Hyperparameter	Value
SSH	30 km
SST	15 km
SSS	30 km
u/v	100 km
σ_{time} for all variables	2.5 days

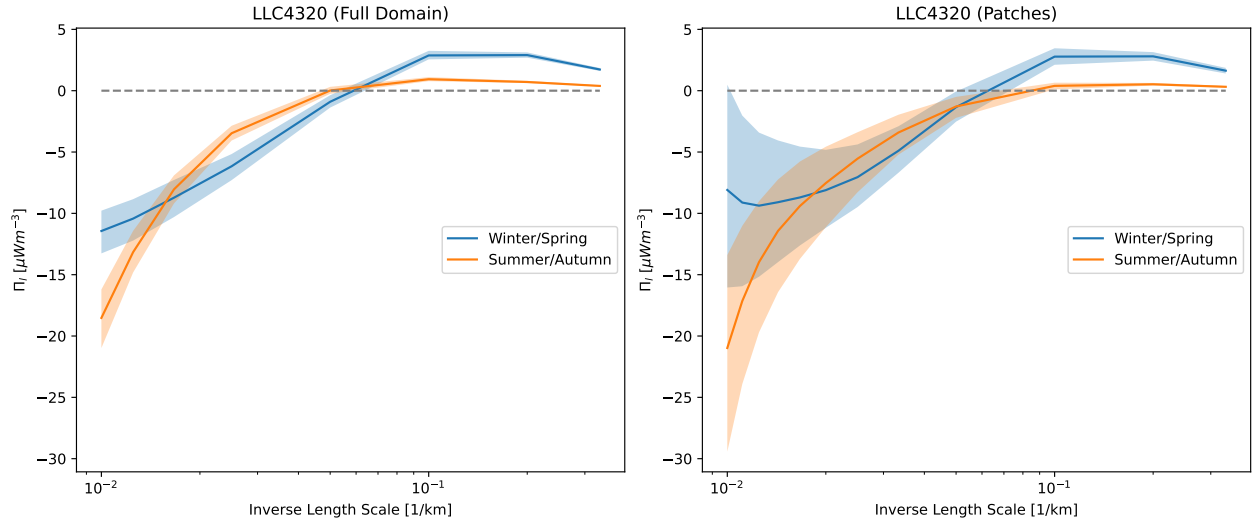
Table C.3: Coarse-graining scales (Gaussian filter width, σ) applied to each variable before assimilating coarse gridded satellite products.

C2 Validity of estimating the submesoscale kinetic energy cascade in small spatio-temporal windows

Since our diffusion prior is restricted to generating a relatively restricted domain of 256x256 LLC4320 grid points (~ 400 km) and relies upon good observational coverage both from SWOT SSH and SEVIRI SST, in the main text we diagnose the strength and seasonality of the kinetic energy cascade by applying the coarse-graining framework [Aluie et al., 2018; Storer and Aluie, 2023] within a set of well-observed windows and averaging over many such windows to get seasonal composite of the cascade in our larger study regions (the Return Current and the Ring Path). This limited sampling introduces an extra potential source of noise and error in the cascade estimates which we assess here. We compute the seasonal composites of the cascade from the LLC4320 ground truth from the full underlying fields on all time-steps in our study regions (which are ~ 1000 km in size) and compare the result to that estimated from the small well-observed windows used in the main text (Figure C.1). This comparison shows that in both regions the computation on the full domain results in the same qualitative conclusions about the cascade below 100 km, with the estimations in good agreement across the scale range considered in this study. The noise and bias from our limited sampling becomes more pronounced at larger scales approaching 100 km, which is reflected in a corresponding broadening of the bootstrapped confidence intervals. The increased uncertainty at larger scales stems from excluding a border region with width that of the coarse-graining filter scale before computing the domain-averaged spectral flux to avoid contamination from edge effects. The close correspondence between the small patch and full domain computations increases confidence that the results presented in the main text are not influenced of sampling artifacts and provide a robust estimate of the kinetic energy cascade below 100 km.



(a) Ring Path



(b) Agulhas Return Current

Figure C.1: Comparison of the kinetic energy cascade for the LLC4320 ground truth inferred from the small well-observed windows used in the main text compared to computation on the full fields for the study region showing the validity of estimating KE cascade in small, well-observed patches.

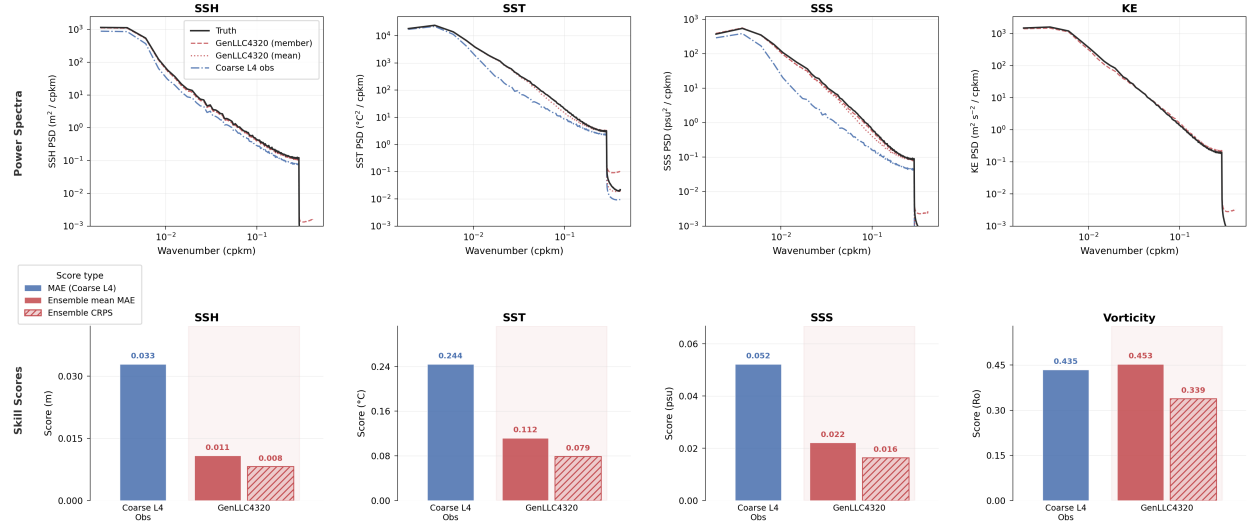


Figure C.2: Same as Figure 4.4 but assimilating the total, unfiltered LLC4320 SSH observations.

C3 Additional evaluations of GenLLC4320 surface state estimation

C3.1 Testing the effect of super-inertial SSH observation contamination on GenLLC4320 state estimates

real world SWOT SSH observations will contain a degree of signal contamination from sub-inertial internal gravity waves whereas our diffusion prior was trained to generate and assimilate SSH with sub-inertial internal tides filtered as a pre-processing step. To assess the potential effect of assimilating SSH observations with the super-inertial SSH signals contaminating the SSH we repeat the evaluation shown in Figure 4.4 using simulated satellite observations from LLC4320 but this time assimilate the unfiltered, total SSH, taking a randomized hourly time-step from within the prior's 12-hour step to reflect that SWOT observes instantaneous SSH rather than a 12-hour average (Figure C.2). The target variable for surface currents remains the sub-inertial filtered data, so that we can assess the bias introduced in the prediction by super-inertial signals in the SSH observations. The GenLLC4320 state estimation metrics remain largely unchanged from those presented in the main text, with a minor degradation in vorticity CRPS the only apparent sign of degraded state estimation due to super-inertial SSH contamination. Looking at the reconstructed kinetic energy spectra, there is no clear sign of enhanced variance at smaller scales that might be expected if the reconstructions were being strongly affected by the super-inertial SSH signals. While the effect of real world super-inertial

SSH contamination on GenLLC4320 state estimates remains an open question, and there is a clear need for novel methods to filter these signals from SWOT observations, the relatively minor state estimation degradation here increases confidence in the robustness of our observational constraints on the submesoscale cascade.

Chapter 5

Conclusions

5.1 Summary

This thesis has sought to provide observational constraints on kinetic energy transfers between scales in the ocean mediated by (sub)mesoscale eddy turbulence. In pursuit of this scientific goal, we have developed novel deep learning methods for surface ocean state estimation from sparse satellite observations that have significantly improved our ability to estimate surface ocean currents from space.

In Chapter 1, we developed a deep learning method for mapping SSH at high resolution from sparse, along-track satellite altimetry and satellite SST observations. We demonstrated that neural networks can learn to synergize SSH and SST observations to fill in the large gaps in SSH observations while retaining small-scale signal content smoothed out by conventional objective analysis approaches. Crucially, the smaller scales in SSH that were previously smoothed out contain rich information about mesoscale eddy dynamics, with quantities such as geostrophic vorticity and strain being severely under-estimated in conventional gridded SSH products. By developing a method that can be trained exclusively on real-world satellite observations, we ensured that our deep learning SSH reconstructions are fully observation-driven and can thus be used to provide independent observational validation of ocean model simulations.

In Chapter 2, we built upon the promise of the method proposed in Chapter 1 by developing a global gridded SSH data product, NeurOST [OSTST, 2024], and applied this dataset to provide observational constraints on the strength and seasonality of the mesoscale kinetic energy cascade. By refining and scaling

our deep learning method for global training and inference, we developed, to our knowledge, the first deep learning-driven global satellite oceanography dataset. When validated against independent satellite altimetry observations, surface drifters, and in an observing system simulation experiment, NeurOST significantly outperforms the previous operational state-of-the-art global SSH product in terms of both accuracy and resolution. Beyond quantitative improvements in SSH mapping metrics, NeurOST revealed *qualitatively* different eddy dynamics to that inferred from conventional SSH products, with stronger non-linear eddy interactions and enhanced seasonality at small scales. Crucially for the scientific objectives of this thesis, we found that the improved resolution of NeurOST led to a significant re-estimation of the strength of the mesoscale kinetic energy cascade, finding that the artificial smoothing of SSH in prior data products led to a significant under-estimation of the strength of the upscale kinetic energy cascade. By resolving for the first time in gridded altimetry the smaller end of the mesoscale range, we found direct observational evidence of the widely-hypothesized seasonal upscale cascade of kinetic energy, with energy progressively cascading to larger scales as spring progresses. The resolution limits of prior SSH data products prevented us from observing the scale range at which this seasonal cascade operates. Finally, while NeurOST confirms the existence of a seasonal upscale cascade in the mesoscale range which peaks in springtime, the dataset is still of insufficient resolution to test whether this energy ultimately originates at submesoscales in wintertime.

In Chapter 3, we developed a deep learning method for multi-variate surface ocean state estimation from satellite observations using generative data assimilation. In contrast to the method from Chapters 1 & 2, this method leverages high-resolution ocean simulations to learn a ‘prior’ for what physically-realistic surface ocean states should look like which can then be utilized in a Bayesian state estimation framework to generate physically-plausible surface ocean states that are consistent with available, sparse observations. In practice, this is achieved by guiding an unconditional diffusion model, which was trained on simulation data, using sparse observations to perform diffusion posterior sampling. While this approach removes the methodological independence from ocean models, the simulation-informed prior allows us to target scales and applications that would not be possible with the approach from Chapters 1 & 2. Specifically, by learning a joint prior over multiple surface ocean variables (SSH, SST, SSS, and surface currents) generative data assimilation can effectively synergize information between variables and we demonstrate this allows to improve the resolution of SSS estimates from SST observations, and infer ageostrophic surface currents by

synergizing SSH and SST observations. Crucially for our scientific objective of inferring kinetic energy cascades, this generative approach ensures the reconstructed states retain physically plausible fine-scale information which regression approaches like that in Chapters 1 & 2 smooth out, leading to un-physical spectral roll-off at small scales.

Finally, in Chapter 4 we refine the generative data assimilation method proposed in Chapter 3 and apply it to obtain observational constraints on the strength and seasonality of the submesoscale kinetic energy cascade. While NeurOST revealed a strong seasonal upscale cascade at mesoscales across a range of regions, questions remain about whether the ultimate source of this energy lies at smaller submesoscales in wintertime or is generated at mesoscales through seasonally modulated mesoscale baroclinic instability [Qiu et al., 2014]. To tackle this problem, we trained a diffusion prior on a submesoscale-permitting ocean simulation and used generative data assimilation to obtain high-resolution surface ocean state estimates from satellite observations. By applying this framework to real observations, we obtained first-of-their-kind observation-constrained surface ocean state estimates that resolve scales well below those resolved in NeurOST. By probing the submesoscale cascade in different regions we find a strong seasonal upscale cascade at submesoscales extending down to $\mathcal{O}(10 \text{ km})$ to be a recurring feature of surface ocean dynamics between regions. Below 10 km, the cascade reverses sign due to strong downscale spectral fluxes localized at strong surface buoyancy fronts. The significance of this seasonal cascade of energy from submesoscales in driving larger mesoscale variability varies between regions, with the seasonality of the cascade most pronounced away from strong background currents which can modulate mesoscale variability and drive year round submesoscale dynamics. While similar conclusions have been drawn from ocean simulations, by assimilating real-world satellite observations our study lends crucial observational support to the view that submesoscale ocean dynamics can feed back to larger scales and play an active role in the ocean’s broader internal variability.

5.2 Discussion & Outlook

5.2.1 NeurOST: Broader Implications, Future Applications, & Refinements

Perhaps the most important contribution of this thesis was the development of a new global gridded SSH and surface geostrophic currents dataset, NeurOST [OSTST, 2024]. At the time of writing, this dataset spans from 2010 through to late 2024, providing a rich dataset for exploring mesoscale eddy dynamics throughout the global ocean. While in Chapter 2 we applied NeurOST to constrain the mesoscale kinetic energy cascade, the dataset opens a broader range of questions about mesoscale eddy dynamics across the global ocean. Since the dataset was published through NASA PO.DAAC, it has begun to be applied in a diverse range of downstream physical oceanography studies on topics including deep submesoscale fronts [Siegelman et al., 2025], internal tides [Zaron et al., 2026], and eddy detection [Perez-Corona et al., 2026]. One striking feature of eddy dynamics in NeurOST compared to the widely-used DUACS product found, but perhaps under-emphasized, in Martin et al. [2024] was that throughout the subtropics the ubiquitous westward propagation of eddies observed in DUACS appears to be an artifact of the objective analysis method, with NeurOST instead revealing a complex web of non-linearly interacting eddies below the resolution limit of DUACS. Since the westward propagation apparent in DUACS was widely attributed to the beta drift effect [Chelton et al., 2011] and has been hypothesized to play a key role in the broader ocean energy cycle [Zhai et al., 2010], future studies re-visiting these findings using the NeurOST maps would be beneficial. Finally, a growing literature has sought to identify the impacts of mesoscale eddies on biogeochemistry by co-locating in situ observations with satellite-tracked eddies derived from DUACS SSH maps [Huang and Xu, 2018; Frenger et al., 2018; Zhang et al., 2019; Cornec et al., 2021; Keppler et al., 2024]. Since we have found striking differences between the mapped mesoscale eddy fields in NeurOST and DUACS, especially in the subtropics and western boundary currents, we believe NeurOST could be of significant value to future studies of the impact of eddies on biogeochemistry, indeed preliminary co-locations of satellite chlorophyll measurements with NeurOST surface currents suggest a much cleaner delineation of the boundaries of coherent eddies trapping plankton in the Gulf Stream than from DUACS [Lang, 2026].

A number of future extensions and refinements to NeurOST can be envisaged and work is under way to explore some of these avenues. First, while NeurOST presently only extends back to 2010, satellite al-

timetry observations from at least two concurrent nadir altimeters are available back to 1993, so extending NeurOST backwards in time to cover the full altimetry era would be useful for studies of long-term changes in ocean eddy dynamics [Martínez-Moreno et al., 2021]. NeurOST was trained on observations starting in 2010 due to the high density of concurrent altimeter missions, so care should be taken to check that the method generalizes to prior years which may exhibit dynamical regimes unseen in the training data or whether it must be re-trained on a wider range of years. To establish robust constraints on long-term trends and inter-annual variability of the mesoscale eddy field it would be beneficial to also produce a version of NeurOST that uses only a two-satellite altimeter constellation since this removes any spurious signals from the improvements in the observing system over time [Barceló-Llull et al., 2024]. While limiting the input to two satellites would lead to a degradation in resolution, Martin et al. [2024] found that NeurOST with a two-satellite input constellation could still map smaller scale signals than DUACS achieves using six. Finally, the launch of SWOT presents exciting new opportunities for increasing the resolution of NeurOST by incorporating high-resolution, wide swath observations in the SSH mapping. Preliminary experiments found that incorporating SWOT in the NeurOST input with no re-training led, counter-intuitively, to a degradation in SSH mapping, emphasizing the need for further training to learn to process the wide swath inputs which differ significantly from along-track altimetry. The growing number of years of available SWOT observations, concurrent with a large constellation of nadir altimeters, should make adapting NeurOST in future to process SWOT relatively straightforward.

5.2.2 Generative Data Assimilation

The generative data assimilation method explored in Chapters 3 & 4 shows great promise for ocean state estimation, and generative modeling is an exciting frontier in the application of deep learning to ocean dynamics. The ability to generate physically realistic states and the flexibility to guide this generation to achieve a range of objectives post-training make such approaches a compelling alternative to the end-to-end regression framework used in Chapters 1 & 2 and explored widely in the literature. Another appealing property of generative methods is their ability to quantify uncertainty by generating an ensemble of states.

While the results of Chapters 3 & 4 are encouraging, some limitations of generative data assimilation should be borne in mind for future applications, especially for increasingly ambitious ocean state estimation

tasks like 3D reconstructions. First, by training the prior on an ocean model simulation the method will likely build in biases stemming from poorly-resolved or parameterized ocean physics in the simulation training data. For applications to the submesoscale range, where observational validation of models has been lacking, this is a key limitation to bear in mind when interpreting the results. Methods to train generative priors from sparse observations alone have been proposed [Rozet et al., 2024], and could be especially valuable for submesoscale oceanography where satellite observations now have significantly finer resolution than global simulations. Notwithstanding these limitations, simulation-trained generative data assimilation methods will likely drive significant advances in ocean state estimation as training data from high-resolution simulations provide a significantly improved prior compared to the low-resolution ocean models used in physics-based data assimilation frameworks. While in this thesis we trained the diffusion prior on output from global simulations like LLC4320, a promising avenue in future would be to instead train on a library of regional simulations run at resolutions beyond those possible in global runs, allowing to learn a prior which fully resolves submesoscale processes. Applications of diffusion models to high-resolution atmospheric modeling suggest the locality of small-scale dynamics means surprisingly small datasets can be used during training [Brenowitz et al., 2025] and the advent of GPU-accelerated ocean modeling should make generating such training datasets straightforward in future [Wagner et al., 2025; Klion et al., 2025].

5.2.3 Ocean Scale Interactions

The main scientific finding of this thesis has been to confirm that the strongly seasonal upscale cascade at (sub)mesoscales observed in high-resolution ocean models appears to be present in the real-world ocean. While this result was foreshadowed by extensive explorations of high-resolution ocean model simulations, the myriad ways in which models cannot represent the full complexity of real-world ocean dynamics make observational constraints on these processes crucial. It is worth emphasizing that the findings from Chapter 2, confirming the existence of a strongly seasonal upscale cascade at mesoscales throughout the subtropics, are grounded entirely in real-world satellite observations, with no reference to an ocean simulation during training. The findings from Chapter 4, that there is a strong submesoscale upscale cascade peaking in wintertime across a range of regions, should be viewed with more circumspection, as the generative data assimilation method uses a prior that was trained on LLC4320. Nonetheless, the observation guidance

used in generative data assimilation should prevent the generation of states that are completely at odds with observations and the cascade strength was found to increase significantly with real-world observations compared to synthetic observations from the simulation. Chapter 4 thus suggests that satellite observations support the existence of the seasonal upscale submesoscale cascade in the real-world ocean.

Despite the progress made in this thesis toward observational constraints on (sub)mesoscale ocean scale interactions, there remain numerous open questions. Firstly, while the results of Chapter 4, in addition to a number of prior studies [Balwada et al., 2022; Naveira Garabato et al., 2022; Qiu et al., 2022; Qiu and Chen, 2025; Chai et al., 2026], point to a strongly seasonal upscale submesoscale kinetic energy cascade, a complex picture emerges with competing upscale and downscale cascades which vary in their relative strengths between regions and seasons. There is thus a need for granular, global monitoring of submesoscale energy cascades to build a more complete picture of the oceanic submesoscale’s upscale impacts and the extent to which they need to be represented in coarse ocean models. Submesoscale-resolving satellite missions like SWOT will be crucial to this effort. Interactions between balanced (sub)mesoscale eddies and unbalanced waves could represent an important energy pathway between the mesoscale and submesoscale [Barkan et al., 2017; Rocha et al., 2018], but constraining this from satellites alone is made challenging by the long orbital repeat time of SWOT and the lack of direct observations of surface currents. Finally, a more granular view of the spatiotemporal variability of the submesoscale cascade and its interactions with mesoscales would allow to evaluate the extent to which explicitly resolving submesoscales in global ocean model simulations impacts larger-scale variability.

5.2.4 Final Thoughts

By developing deep learning methods for ocean state estimation and leveraging submesoscale-resolving satellite missions, this thesis has sought to better constrain the dynamical interactions between submesoscale and mesoscale ocean turbulence. The convergence of high-resolution simulations, novel satellite observations, and maturing deep learning methodologies now makes routine global monitoring of submesoscale surface ocean currents a realistic prospect and with it, a more mechanistic and predictive understanding of how energy moves between scales across the global ocean.

Bibliography

- Saleh Abdalla, Abdolnabi Abdeh Kolahchi, Michaël Ablain, Susheel Adusumilli, Suchandra Aich Bhowmick, Eva Alou-Font, Laiba Amarouche, Ole Baltazar Andersen, Helena Antich, Lotfi Aouf, et al. 2021. Altimetry for the future: Building on 25 years of progress. *Advances in Space Research*, 68(2):319–363.
- Ryan Patrick Abernathey and J Marshall. 2013. Global surface eddy diffusivities derived from satellite altimetry. *Journal of Geophysical Research: Oceans*, 118(2):901–916.
- Angelina Agabin, J. Xavier Prochaska, Peter C. Cornillon, and Christian E. Buckingham. 2024. Mitigating masked pixels in a climate-critical ocean dataset. *Remote Sensing*, 16(13).
- Adekunle Ajayi, Julien Le Sommer, Eric Chassignet, Jean-Marc Molines, Xiaobiao Xu, Aurelie Albert, and Emmanuel Cosme. 2020. Spatial and temporal variability of the north atlantic eddy field from two kilometric-resolution ocean models. *Journal of Geophysical Research: Oceans*, 125(5):e2019JC015827.
- Adekunle Ajayi, Julien Le Sommer, Eric P Chassignet, Jean-Marc Molines, Xiaobiao Xu, Aurelie Albert, and William Dewar. 2021. Diagnosing cross-scale kinetic energy exchanges from two submesoscale permitting ocean models. *Journal of Advances in Modeling Earth Systems*, 13(6):e2019MS001923.
- Hussein Aluie, Matthew Hecht, and Geoffrey K Vallis. 2018. Mapping the energy cascade in the North Atlantic Ocean: The coarse-graining approach. *Journal of Physical Oceanography*, 48(2):225–244.
- Angel Amores, Gabriel Jordà, Thomas Arsouze, and Julien Le Sommer. 2018. Up to what extent can we characterize ocean eddies using present-day gridded altimetric products? *Journal of Geophysical Research: Oceans*, 123(10):7220–7236.

- Brian K Arbic, Shane Elipot, Jonathan M Brasch, Dimitris Menemenlis, Aurelien L Ponte, Jay F Shriver, Xiaolong Yu, Edward D Zaron, Matthew H Alford, Maarten C Buijsman, et al. 2022. Frequency dependence of near-surface oceanic kinetic energy from drifter observations and global high-resolution models. *arXiv preprint arXiv:2202.08877*.
- Brian K Arbic, Kurt L Polzin, Robert B Scott, James G Richman, and Jay F Shriver. 2013. On eddy viscosity, energy cascades, and the horizontal resolution of gridded satellite altimeter products. *Journal of Physical Oceanography*, 43(2):283–300.
- Brian K Arbic, Robert B Scott, Dudley B Chelton, James G Richman, and Jay F Shriver. 2012. Effects of stencil width on surface ocean geostrophic velocity and vorticity estimation from gridded satellite altimeter data. *Journal of Geophysical Research: Oceans*, 117(C3).
- Théo Archambault, Arthur Filoche, Anastase Charantonis, and Dominique Béréziat. 2024a. Pre-training and fine-tuning attention based encoder decoder improves sea surface height multi-variate inpainting. In *VISAPP*.
- Théo Archambault, Arthur Filoche, Anastase Charantonis, Dominique Béréziat, and Sylvie Thiria. 2024b. Learning sea surface height interpolation from multi-variate simulated satellite observations. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004047.
- Théo Archambault, Arthur Filoche, Anastase Charantonis, and Dominique Béréziat. 2023. Multimodal unsupervised spatio-temporal interpolation of satellite ocean altimetry maps. In *VISAPP*.
- Matthew Archer, Jinbo Wang, Patrice Klein, Gerald Dibarbouré, and Lee-Lueng Fu. 2025. Wide-swath satellite altimetry unveils global submesoscale ocean dynamics. *Nature*, 640(8059):691–696.
- AVISO. 2021. Data challenge ose - 2021a SSH mapping OSE [Dataset]. <https://doi.org/10.24400/527896/a01-2021.005>.
- AVISO/DUACS. 2024. SWOT level-3 KaRIn low rate SSH expert (v2.0) [Dataset]. Accessed on 02-11-2024.

- Maxime Ballarotta, Sammy Metref, Aurelie Albery, Emmanuel Cosme, Maxime Beauchamp, and Florian Le Guillou. 2021. Ocean Data Challenges: 2021a SSH mapping OSE. https://github.com/ocean-data-challenges/2021a_SSH_mapping_OSE.
- Maxime Ballarotta, Clément Ubelmann, Marie-Isabelle Pujol, Guillaume Taburet, Florent Fournier, Jean-François Legeais, Yannice Faugère, Antoine Delepouille, Dudley Chelton, Gérald Dibarboure, et al. 2019. On the resolutions of ocean altimetry maps. *Ocean Science*, 15(4):1091–1109.
- Maxime Ballarotta, Clément Ubelmann, Marine Rogé, Florent Fournier, Yannice Faugère, Gérald Dibarboure, Rosemary Morrow, and Nicolat Picot. 2020. Dynamic mapping of along-track ocean altimetry: Performance from real observations. *Journal of Atmospheric and Oceanic Technology*, 37(9):1593–1601.
- Maxime Ballarotta, Clément Ubelmann, Pierre Veillard, Pierre Prandi, Hélène Etienne, Sandrine Mulet, Yannice Faugère, Gérald Dibarboure, Rosemary Morrow, and Nicolas Picot. 2023. Improved global sea surface height and current maps from remote sensing and in situ observations. *Earth System Science Data*, 15(1):295–315.
- Dhruv Balwada, Jin-Han Xie, Raffaele Marino, and Fabio Feraco. 2022. Direct observational evidence of an oceanic dual kinetic energy cascade and its seasonality. *Science Advances*, 8(41):eabq2566.
- Bàrbara Barceló-Llull, Pere Rosselló, Vincent Combes, Antonio Sánchez-Román, M Isabelle Pujol, and Ananda Pascual. 2024. Robust intensification of global ocean eddy kinetic energy from three decades of satellite altimetry observations. *arXiv preprint arXiv:2406.08014*.
- Roy Barkan, Kraig B Winters, and James C McWilliams. 2017. Stimulated imbalance and the enhancement of eddy kinetic energy dissipation by internal waves. *Journal of Physical Oceanography*, 47(1):181–198.
- Alexander Barth, Aida Alvera-Azcárate, Matjaz Licer, and Jean-Marie Beckers. 2020. DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. *Geoscientific Model Development*, 13(3):1609–1622.
- Alexander Barth, Aida Alvera-Azcárate, Charles Troupin, and Jean-Marie Beckers. 2022. DINCAE 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations. *Geoscientific Model Development*, 15(5):2183–2196.

- Maxime Beauchamp, Ronan Fablet, Clément Ubelmann, Maxime Ballarotta, and Bertrand Chapron. 2020. Intercomparison of data-driven and learning-based interpolations of along-track nadir and wide-swath SWOT altimetry observations. *Remote Sensing*, 12(22):3806.
- Maxime Beauchamp, Quentin Febvre, Hugo Georgenthum, and Ronan Fablet. 2022. 4DVarNet-SSH: end-to-end learning of variational interpolation schemes for nadir and wide-swath satellite altimetry. *Geoscientific Model Development Discussions*, 2022:1–37.
- Maxime Beauchamp, Ioanna Karagali, Guisella Gacitúa, Jacob L Høyer, Maxime Ballarotta, and Ronan Fablet. 2025. Multiscale neural assimilation scheme for high-resolution sea surface temperature reconstruction from satellite observations. *Sci. Rep.*, 15(1):39928.
- Giulio Boccaletti, Raffaele Ferrari, and Baylor Fox-Kemper. 2007. Mixed layer instabilities and restratification. *Journal of Physical Oceanography*, 37(9):2228–2250.
- Melissa M Bowen, William J Emery, John L Wilkin, Paul C Tildesley, Ian J Barton, and Rebecca Knewton. 2002. Extracting multiyear surface currents from sequential thermal imagery using the maximum cross-correlation technique. *Journal of Atmospheric and Oceanic Technology*, 19(10):1665–1676.
- Noah D Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S Pritchard. 2025. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*.
- Francis P Bretherton, Russ E Davis, and CB Fandry. 1976. A technique for objective analysis and design of oceanographic experiments applied to MODE-73. In *Deep Sea Research and Oceanographic Abstracts*, volume 23, pages 559–582. Elsevier.
- Christian E Buckingham, Alberto C Naveira Garabato, Andrew F Thompson, Liam Brannigan, Ayah Lazar, David P Marshall, AJ George Nurser, Gillian Damerell, Karen J Heywood, and Stephen E Belcher. 2016. Seasonality of submesoscale flows in the ocean surface boundary layer. *Geophysical Research Letters*, 43(5):2118–2126.

- Bruno Buongiorno Nardelli, Davide Cavaliere, Elodie Charles, and Daniele Ciani. 2022. Super-resolving ocean dynamics from space with computer vision algorithms. *Remote Sensing*, 14(5):1159.
- Michele Buzzicotti. 2023. Data reconstruction for complex flows using ai: Recent progress, obstacles, and perspectives. *Europhysics Letters*, 142(2):23001.
- Jared L Callahan, Kazuki Maeda, and Steven L Brunton. 2019. Robust flow reconstruction from limited measurements via sparse representation. *Physical Review Fluids*, 4(10):103907.
- Jörn Callies, Raffaele Ferrari, Jody M Klymak, and Jonathan Gula. 2015. Seasonality in submesoscale turbulence. *Nature communications*, 6(1):6862.
- Yuhan Cao, Changming Dong, Alexandre Stegner, Brandon J Bethel, Chunyan Li, Jihai Dong, Haibin Lü, and Jingsong Yang. 2023. Global sea surface cyclogeostrophic currents derived from satellite altimetry data. *Journal of Geophysical Research: Oceans*, 128(1):e2022JC019357.
- Xavier Capet, JAMES C McWilliams, M Jeroen Molemaker, and Alexander F Shchepetkin. 2008a. Mesoscale to submesoscale transition in the california current system. part iii: Energy balance and flux. *Journal of Physical Oceanography*, 38(10):2256–2269.
- Xavier Capet, JAMES C McWilliams, M Jeroen Molemaker, and Alexander F Shchepetkin. 2008b. Mesoscale to submesoscale transition in the California Current System. Part III: Energy balance and flux. *Journal of Physical Oceanography*, 38(10):2256–2269.
- Elisa Carli, Rosemary Morrow, Oscar Vergara, Robin Chevrier, and Lionel Renault. 2023. Ocean 2D eddy energy fluxes from small mesoscale processes with SWOT. *Ocean Science*, 19(5):1413–1435.
- Elisa Carli, Lia Siegelman, Rosemary Morrow, and Oscar Vergara. 2024. Surface quasi geostrophic reconstruction of vertical velocities and vertical heat fluxes in the southern ocean: Perspectives for swot. *Journal of Geophysical Research: Oceans*, 129(9):e2024JC021216.
- Junwei Chai, Jian Zhao, and Xiaopei Lin. 2026. Characteristics of energy cascade in iceland basin and irminger sea. *Journal of Physical Oceanography*, 56(2):409–430.

- Dudley B Chelton, John C Ries, Bruce J Haines, Lee-Lueng Fu, and Philip S Callahan. 2001. Satellite altimetry. In *International Geophysics*, volume 69, pages 1–ii. Elsevier.
- Dudley B Chelton, Michael G Schlax, and Roger M Samelson. 2011. Global observations of nonlinear mesoscale eddies. *Progress in oceanography*, 91(2):167–216.
- Wei Chen, Richard P Mied, and Colin Y Shen. 2008. Near-surface ocean velocity from infrared images: Global optimal solution to an inverse model. *Journal of Geophysical Research: Oceans*, 113(C10).
- Toshio Michael Chin, Jorge Vazquez-Cuervo, and Edward M Armstrong. 2017. A multi-scale high-resolution analysis of global sea surface temperature. *Remote sensing of environment*, 200:154–169.
- Daniele Ciani, Claudia Fanelli, and Bruno Buongiorno Nardelli. 2024. Estimating ocean currents from the joint reconstruction of absolute dynamic topography and sea surface temperature through deep learning algorithms. *EGUsphere*, 2024:1–25.
- CMEMS. 2023a. Copernicus marine in situ - global ocean-delayed mode in situ observations of surface (drifters, HFR) and sub-surface (vessel-mounted ADCPs) water velocity [Dataset]. <https://doi.org/10.17882/86236>.
- CMEMS. 2023b. Global ocean along track 1 3 sea surface heights reprocessed 1993 ongoing tailored for data assimilation [Dataset]. <https://doi.org/10.48670/moi-00146>.
- CMEMS. 2023c. Global ocean gridded 1 4 sea surface heights and derived variables reprocessed 1993 ongoing [Dataset]. <https://doi.org/10.48670/moi-00148>.
- Copernicus Climate Change Service (C3S). 2024. Era5 hourly data on single levels from 1940 to present [Dataset]. Accessed on 09-24-2024.
- Copernicus Marine in situ TAC. 2024. Copernicus Marine In Situ - Global Ocean-Delayed Mode in situ Observations of surface (drifters, HFR) and sub-surface (vessel-mounted ADCPs) water velocity. [Dataset]. Accessed on 07-04-2024.
- Marin Cornec, Rémi Laxenaire, Sabrina Speich, and Hervé Claustre. 2021. Impact of mesoscale eddies on deep chlorophyll maxima. *Geophysical Research Letters*, 48(15):e2021GL093470.

- Peter C Cornillon, Eric Firing, Andrew Thompson, Leonid Ivanov, Igor Kamenkovich, Christian E Buckingham, and Yakov Afanasyev. 2019. Oceans. In Boris Galperin and Peter Read, editors, *Zonal Jets Phenomenology, Genesis, and Physics*, chapter 3, pages 46 – 71. Cambridge University Press.
- Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869.
- Benoit Cushman-Roisin and Jean-Marie Beckers. 2011. *Introduction to geophysical fluid dynamics: physical and numerical aspects*. Academic press.
- Jihai Dong, Baylor Fox-Kemper, Hong Zhang, and Changming Dong. 2020. The seasonality of submesoscale energy production, content, and cascade. *Geophysical Research Letters*, 47(6):e2020GL087388.
- Claire Dufau, Marion Orszynowicz, G  rald Dibarboure, Rosemary Morrow, and Pierre-Yves Le Traon. 2016. Mesoscale resolution capability of altimetry: Present and future. *Journal of Geophysical Research: Oceans*, 121(7):4910–4927.
- Eric D’asaro, Craig Lee, Luc Rainville, Ramsey Harcourt, and Leif Thomas. 2011. Enhanced turbulence and energy dissipation at ocean fronts. *science*, 332(6027):318–322.
- William J Emery, AC Thomas, MJ Collins, William R Crawford, and DL Mackas. 1986. An objective method for computing advective surface velocities from sequential infrared satellite images. *Journal of Geophysical Research: Oceans*, 91(C11):12865–12878.
- E.U. Copernicus Marine Service Information (CMEMS). 2024a. Global high resolution ODYSSEA sea surface temperature multi-sensor L3 [Dataset]. Accessed on 09-24-2024.
- E.U. Copernicus Marine Service Information (CMEMS). 2024b. Global ocean along track l 3 sea surface heights NRT [Dataset]. Accessed on 07-04-2024.
- E.U. Copernicus Marine Service Information (CMEMS). 2024c. Global ocean along track l 3 sea surface heights reprocessed 1993 ongoing tailored for data assimilation [Dataset]. Accessed on 07-04-2024.
- E.U. Copernicus Marine Service Information (CMEMS). 2024d. Global ocean along track L3 sea surface heights reprocessed 1993 ongoing tailored for data assimilation [Dataset]. Accessed on 09-24-2024.

- E.U. Copernicus Marine Service Information (CMEMS). 2024e. Global ocean gridded normalized measurement noise of sea level anomalies [Dataset]. Accessed on 09-24-2024.
- E.U. Copernicus Marine Service Information (CMEMS). 2024f. Global ocean physics reanalysis [Dataset]. Accessed on 09-24-2024.
- E.U. Copernicus Marine Service Information (CMEMS). 2024g. Multi observation global ocean sea surface salinity and sea surface density [Dataset]. Accessed on 09-24-2024.
- R. Fablet, B. Chapron, J. Le Sommer, and F. Sévellec. 2024. Inversion of sea surface currents from satellite-derived sst-ssh synergies with 4dvarnets. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS003609.
- Ronan Fablet, Mohamed Mahmoud Amar, Quentin Febvre, Maxime Beauchamp, and Bertrand Chapron. 2021a. End-to-end physics-informed representation learning for satellite ocean remote sensing data: Applications to satellite altimetry and sea surface currents. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, (3).
- Ronan Fablet and Bertrand Chapron. 2022. Multimodal learning-based inversion models for the space-time reconstruction of satellite-derived geophysical fields. *arXiv preprint arXiv:2203.10640*.
- Ronan Fablet, Bertrand Chapron, Lucas Drumetz, Etienne Mémin, Olivier Pannekoucke, and François Rousseau. 2021b. Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, 13(10):e2021MS002572.
- Ronan Fablet, Quentin Febvre, and Bertrand Chapron. 2022. Multimodal 4DVarNets for the reconstruction of sea surface dynamics from SST-SSH synergies. *arXiv preprint arXiv:2207.01372*.
- Ronan Fablet, Quentin Febvre, and Bertrand Chapron. 2023. Multimodal 4DVarNets for the reconstruction of sea surface dynamics from SST-SSH synergies. *IEEE Transactions on Geoscience and Remote Sensing*.
- Claudia Fanelli, Daniele Ciani, Andrea Pisano, and Bruno Buongiorno Nardelli. 2024. Deep learning for super-resolution of mediterranean sea surface temperature fields. *EGUsphere*, 2024:1–18.

- J Thomas Farrar, Eric D’Asaro, Ernesto Rodríguez, Andrey Shcherbina, Luc Lenain, Melissa Omand, Alex Wineteer, Paban Bhuyan, Fred Bingham, AB Villas Boas, et al. 2025. S-mode: the sub-mesoscale ocean dynamics experiment. *Bulletin of the American Meteorological Society*, 106(4):E657–E677.
- Q. Febvre, J. Le Sommer, C. Ubelmann, and R. Fablet. 2024a. Training neural mapping schemes for satellite altimetry with simulation data. *Journal of Advances in Modeling Earth Systems*, 16(7):e2023MS003959.
- Quentin Febvre, Ronan Fablet, Julien Le Sommer, and Clément Ubelmann. 2022. Joint calibration and mapping of satellite altimetry data using trainable variational models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1536–1540. IEEE.
- Quentin Febvre, Julien Le Sommer, Clément Ubelmann, and Ronan Fablet. 2024b. Training neural mapping schemes for satellite altimetry with simulation data. *Journal of Advances in Modeling Earth Systems*, 16(7):e2023MS003959.
- Raffaele Ferrari and Carl Wunsch. 2009. Ocean circulation kinetic energy: Reservoirs, sources, and sinks. *Annual Review of Fluid Mechanics*, 41:253–282.
- Baylor Fox-Kemper, Raffaele Ferrari, and Robert Hallberg. 2008. Parameterization of mixed layer eddies. part i: Theory and diagnosis. *Journal of Physical Oceanography*, 38(6):1145–1165.
- Mara Freilich, Luc Lenain, and Sarah T Gille. 2023. Characterizing the role of non-linear interactions in the transition to submesoscale dynamics at a dense filament. *Geophysical Research Letters*, 50(15):e2023GL103745.
- Ivy Frenger, Matthias Münnich, and Nicolas Gruber. 2018. Imprint of southern ocean mesoscale eddies on chlorophyll. *Biogeosciences*, 15(15):4781–4798.
- Lee-Lueng Fu, Dudley B Chelton, Pierre-Yves Le Traon, and Rosemary Morrow. 2010. Eddy dynamics from satellite altimetry. *Oceanography*, 23(4):14–25.
- Lee-Lueng Fu, Tamlin Pavelsky, Jean-Francois Cretaux, Rosemary Morrow, J Thomas Farrar, Parag Vaze, Pierre Sengenès, Nadya Vinogradova-Shiffer, Annick Sylvestre-Baron, Nicolas Picot, et al. 2024. The

- Surface Water and Ocean Topography mission: A breakthrough in radar remote sensing of the ocean and land surface water. *Geophysical Research Letters*, 51(4):e2023GL107652.
- Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. 2022. SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3170–3180.
- Zhanwen Gao, Bertrand Chapron, Chunyong Ma, Ronan Fablet, Quentin Febvre, Wenxia Zhao, and Ge Chen. 2024. A deep learning approach to extract balanced motions from sea surface height snapshot. *Geophysical Research Letters*, 51(7):e2023GL106623.
- Pierre Garcia, Inès Larroche, Amélie Pesnec, Hannah Bull, Théo Archambault, Evangelos Moschos, Alexandre Stegner, Anastase Charantonis, and Dominique Béréziat. 2025. Orcast: Operational high-resolution current forecasts. *arXiv preprint arXiv:2501.12054*.
- Lucile Gaultier, Clément Ubelmann, and Lee-Lueng Fu. 2016. The challenge of using future SWOT data for oceanic field reconstruction. *Journal of Atmospheric and Oceanic Technology*, 33(1):119–126.
- Peter R Gent and James C McWilliams. 1990. Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, 20(1):150–155.
- Tom M George, Georgy E Manucharyan, and Andrew F Thompson. 2021. Deep learning to infer eddy heat fluxes from sea surface height patterns of mesoscale turbulence. *Nature communications*, 12(1):800.
- Subhankar Ghosh, Arun Sharma, Jayant Gupta, Aneesh Subramanian, and Shashi Shekhar. 2024. Towards kriging-informed conditional diffusion for regional sea-level data downscaling: A summary of results. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 372–383.
- Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
- Edwin Goh, Alice R Yepremyan, Jinbo Wang, and Brian Wilson. 2023. MAESSTRO: Masked autoencoders for sea surface temperature reconstruction under occlusion. *EGUsphere [pre-print]*, 2023:1–20.

- C González-Haro and J Isern-Fontanet. 2014. Global ocean current reconstruction from altimetric and microwave SST measurements. *Journal of Geophysical Research: Oceans*, 119(6):3378–3391.
- Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Yiming Guo, Scott Bachman, Frank Bryan, and Stuart Bishop. 2022. Increasing trends in oceanic surface poleward eddy heat flux observed over the past three decades. *Geophysical Research Letters*, 49(16):e2022GL099362.
- Qiuchang Han, Xingliang Jiang, Yang Zhao, Xudong Wang, Zhijin Li, and Renhe Zhang. 2024. Generative diffusion model-based downscaling of observed sea surface height over kuroshio extension since 2000. *arXiv preprint arXiv:2408.12632*.
- Ute Hausmann and Arnaud Czaja. 2012. The observed signature of mesoscale eddies in sea surface temperature and the associated heat transport. *Deep Sea Research Part I: Oceanographic Research Papers*, 70:60–72.
- Jing He and Amala Mahadevan. 2024. Vertical velocity diagnosed from surface data with machine learning. *Geophysical Research Letters*, 51(6):e2023GL104835.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Hans Hersbach. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.
- Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.
- Helene T Hewitt, Malcolm Roberts, Pierre Mathiot, Arne Biastoch, Ed Blockley, Eric P Chassignet, Baylor

- Fox-Kemper, Pat Hyder, David P Marshall, Ekaterina Popova, et al. 2020. Resolving and parameterising the ocean mesoscale in earth system models. *Current Climate Change Reports*, 6:137–152.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.
- Brian J Hoskins. 1982. The mathematical theory of frontogenesis. *Annual Review of Fluid Mechanics*, 14(1):131–151.
- BL Hua and P Klein. 1998. An exact criterion for the stirring properties of nearly two-dimensional turbulence. *Physica D: Nonlinear Phenomena*, 113(1):98–110.
- BL Hua, J.C. McWilliams, and P. Klein. 1998. Lagrangian accelerations in geostrophic turbulence. *Journal of Fluid Mechanics*, 366:87–108.
- Jie Huang and Fanghua Xu. 2018. Observational evidence of subsurface chlorophyll response to mesoscale eddies in the north pacific. *Geophysical Research Letters*, 45(16):8462–8470.
- Jordi Isern-Fontanet, Bertrand Chapron, Guillaume Lapeyre, and Patrice Klein. 2006. Potential use of microwave sea surface temperatures for the estimation of ocean currents. *Geophysical Research Letters*, 33(24).
- Jordi Isern-Fontanet, Guillaume Lapeyre, Patrice Klein, Bertrand Chapron, and Matthew W Hecht. 2008. Three-dimensional reconstruction of oceanic mesoscale currents from surface information. *Journal of Geophysical Research: Oceans*, 113(C9).
- Jordi Isern-Fontanet, Mahesh Shinde, and Cristina González-Haro. 2014. On the transfer function between surface fields and the geostrophic stream function in the mediterranean sea. *Journal of physical oceanography*, 44(5):1406–1423.

- Andrei Yu Ivanov and Anna I Ginzburg. 2002. Oceanic eddies in synthetic aperture radar images. *Journal of Earth System Science*, 111:281–295.
- Steven R Jayne and Jochem Marotzke. 2002. The oceanic eddy heat transport. *Journal of Physical Oceanography*, 32(12):3328–3345.
- C Spencer Jones, Qiyu Xiao, Ryan P Abernathey, and K Shafer Smith. 2023. Using lagrangian filtering to remove waves from the ocean surface velocity field. *Journal of Advances in Modeling Earth Systems*, 15(4):e2022MS003220.
- JPL MUR MEaSUREs Project. 2015. GHRSSST level 4 MUR global foundation sea surface temperature analysis. ver. 4.1 [Dataset]. Accessed on 07-04-2024.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.
- Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577.
- Kathryn A Kelly. 1989. An inverse model for near-surface velocity from infrared images. *Journal of Physical Oceanography*, 19(12):1845–1864.
- Lydia Keppler, Yassir A Eddebbar, Sarah T Gille, Nicola Guisewhite, Matthew R Mazloff, Veronica Tamsitt, Ariane Verdy, and Lynne D Talley. 2024. Effects of mesoscale eddies on southern ocean biogeochemistry. *Agu Advances*, 5(6):e2024AV001355.
- Patrice Klein and Guillaume Lapeyre. 2009. The oceanic vertical pump induced by mesoscale and submesoscale turbulence. *Annual review of marine science*, 1(1):351–375.
- Patrice Klein, Guillaume Lapeyre, Lia Siegelman, Bo Qiu, Lee-Lueng Fu, Hector Torres, Zhan Su, Dimitris Menemenlis, and Sylvie Le Gentil. 2019. Ocean-scale interactions from space. *Earth and Space Science*, 6(5):795–817.

- Hannah Klion, Robert Hetland, Jean Sexton, Ann Almgren, Iulian Grindeanu, Kyle Hinson, and Vijay Mahadevan. 2025. Remora: Regional modeling of oceans refined adaptively (built on amrex). *Journal of Open Source Software*, 10(110):7958.
- Alisa Kugusheva, Hannah Bull, Evangelos Moschos, Artemis Ioannou, Briac Le Vu, and Alexandre Stegner. 2024. Ocean satellite data fusion for high-resolution surface current maps. *Remote Sensing*, 16(7):1182.
- JH LaCasce and A Mahadevan. 2006. Estimating subsurface horizontal and vertical velocities from sea-surface temperature. *Journal of Marine Research*, 64(5):695–721.
- Gary SE Lagerloef, Gary T Mitchum, Roger B Lukas, and Pearn P Niiler. 1999a. Tropical Pacific near-surface currents estimated from altimeter, wind, and drifter data. *Journal of Geophysical Research: Oceans*, 104(C10):23313–23326.
- Gary SE Lagerloef, Gary T Mitchum, Roger B Lukas, and Pearn P Niiler. 1999b. Tropical pacific near-surface currents estimated from altimeter, wind, and drifter data. *Journal of Geophysical Research: Oceans*, 104(C10):23313–23326.
- Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. 2023. Learning skillful medium-range global weather forecasting. *Science*, page eadi2336.
- Sarah Lang. 2026. personal communication.
- G Lapeyre. 2009. What mesoscale signal does the altimeter see? On the decomposition in baroclinic modes and the role of the surface boundary condition. *J. Phys. Oceanogr*, 39:2857–2874.
- G Lapeyre, P Klein, and BL Hua. 1999. Does the tracer gradient vector align with the strain eigenvectors in 2D turbulence? *Physics of fluids*, 11(12):3729–3737.
- GUILLAUME Lapeyre and PATRICE Klein. 2006a. Dynamics of the upper oceanic layers in terms of surface quasigeostrophy theory. *Journal of physical oceanography*, 36(2):165–176.
- GUILLAUME Lapeyre and PATRICE Klein. 2006b. Dynamics of the upper oceanic layers in terms of surface quasigeostrophy theory. *Journal of physical oceanography*, 36(2):165–176.

- WG Large and S Pond. 1981. Open ocean momentum flux measurements in moderate to strong winds. *Journal of physical oceanography*, 11(3):324–336.
- Albion Lawrence and Jörn Callies. 2022. Seasonality and spatial dependence of mesoscale and sub-mesoscale ocean currents from along-track satellite altimetry. *Journal of Physical Oceanography*, 52(9):2069–2089.
- Florian Le Guillou, Bertrand Chapron, and Marie-Hélène Rio. 2024. Vardyn: Dynamical joint-reconstructions of sea surface height and temperature from multi-sensor satellite observations. *Authorea Preprints*.
- Florian Le Guillou, Lucile Gaultier, Maxime Ballarotta, Sammy Metref, Clément Ubelmann, Emmanuel Cosme, and Marie-Hélène Rio. 2023. Regional mapping of energetic short mesoscale ocean dynamics from altimetry: performances from real observations. *Ocean Science*, 19(5):1517–1527.
- Florian Le Guillou, Sammy Metref, Emmanuel Cosme, Clément Ubelmann, Maxime Ballarotta, Julien Le Sommer, and Jacques Verron. 2021. Mapping altimetry in the forthcoming SWOT era by back-and-forth nudging a one-layer quasigeostrophic model. *Journal of Atmospheric and Oceanic Technology*, 38(4):697–710.
- PY Le Traon, F Nadal, and N Ducet. 1998. An improved mapping method of multisatellite altimeter data. *Journal of Atmospheric and Oceanic Technology*, 15(2):522–534.
- Jean-Michelle Lellouche, Eric Greiner, Romain Bourdallé-Badie, Gilles Garric, Angélique Melet, Marie Drévillon, Clément Bricaud, Mathieu Hamon, Olivier Le Galloudec, Charly Regnier, Tony Candela, Charles-Emmanuel Testut, Florent Gasparin, Giovanni Ruggiero, Mounir Benkiran, Yann Drillet, and Pierre-Yves Le Traon. 2021. The copernicus global 1/12 oceanic and sea ice GLORYS12 reanalysis. *Frontiers in Earth Science*, 9:698876.
- Luc Lenain, Kaushik Srinivasan, Roy Barkan, and Nick Pizzo. 2025. An unprecedented view of ocean currents from geostationary satellites. *Research Square Pre-Print*.
- Marina Lévy, Peter JS Franks, and K Shafer Smith. 2018. The role of submesoscale currents in structuring marine ecosystems. *Nature communications*, 9(1):4758.

- Redouane Lguensat, Ronan Fablet, Julien Le Sommer, Sammy Metref, Emmanuel Cosme, Kaouthier Oueniche, Lucas Drumetz, and Jonathan Gula. 2020. Filtering internal tides from wide-swath altimeter data using convolutional neural networks. In *2020 IEEE Int. Geosci. Remote Sens. Symp.*, pages 3904–3907. IEEE.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.
- Conrad A Luecke, Brian K Arbic, James G Richman, Jay F Shriver, Matthew H Alford, Joseph K Ansong, Steven L Bassette, Maarten C Buijsman, Dimitris Menemenlis, Robert B Scott, et al. 2020. Statistical comparisons of temperature variance and kinetic energy in global ocean models and observations: Results from mesoscale to internal wave frequencies. *Journal of Geophysical Research: Oceans*, 125(5):e2019JC015306.
- Jingwen Lyu, Yue Wang, Christian Pedersen, Spencer Jones, and Dhruv Balwada. 2024. Multi-scale decomposition of sea surface height snapshots using machine learning. *arXiv preprint arXiv:2409.17354*.
- Amala Mahadevan and Amit Tandon. 2006. An analysis of mechanisms for submesoscale vertical motion at ocean fronts. *Ocean Modelling*, 14(3-4):241–256.
- Peter Manshausen, Yair Cohen, Peter Harrington, Jaideep Pathak, Mike Pritchard, Piyush Garg, Morteza Mardani, Karthik Kashinath, Simon Byrne, and Noah Brenowitz. 2025. Generative data assimilation of sparse weather station observations at kilometer scales. *J. Adv. Model. Earth Syst.*, 17(10):e2024MS004505.
- Georgy E Manucharyan, Lia Siegelman, and Patrice Klein. 2021. A deep learning approach to spatiotemporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence. *Journal of Advances in Modeling Earth Systems*, 13(1):e2019MS001965.
- Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Karthik Kashinath, Jan Kautz, and Mike Pritchard. 2024. Residual diffusion modeling for km-scale atmospheric downscaling. *PREPRINT available at Research Square*.

- Scott A Martin, Georgy E Manucharyan, and Patrice Klein. 2023. Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies. *Journal of Advances in Modeling Earth Systems*, 15(5):e2022MS003589.
- Scott A Martin, Georgy E Manucharyan, and Patrice Klein. 2024. Deep learning improves global satellite observations of ocean eddy dynamics. *Geophysical Research Letters*, 51(17):e2024GL110059.
- Scott A Martin, Georgy E Manucharyan, and Patrice Klein. 2025. Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. *J. Adv. Model. Earth Syst.*, 17(8):e2025MS005063.
- Josué Martínez-Moreno, Andrew McC Hogg, Matthew H England, Navid C Constantinou, Andrew E Kiss, and Adele K Morrison. 2021. Global changes in oceanic mesoscale currents over the satellite altimetry record. *Nature Climate Change*, 11(5):397–403.
- Josué Martínez-Moreno, Andrew McC Hogg, Andrew E Kiss, Navid C Constantinou, and Adele K Morrison. 2019. Kinetic energy of eddy-like features from sea surface altimetry. *Journal of Advances in Modeling Earth Systems*, 11(10):3090–3105.
- James C McWilliams. 1984. The emergence of isolated coherent vortices in turbulent flow. *Journal of Fluid Mechanics*, 146:21–43.
- Sammy Metref and Maxime Ballarotta. 2023. Ocean Data Challenges: 2023a SSH mapping OSE. https://github.com/ocean-data-challenges/2023a_SSH_mapping_OSE.
- Sammy Metref, Maxime Ballarotta, Julien Le Sommer, Emmanuel Cosme, Aurelie Albert, Maxime Beauchamp, Florian Le Guillou, Clement Ubelmann, and Quentin Febvre. 2023. Ocean data challenges. <https://ocean-data-challenges.github.io/>.
- Rosemary Morrow, Lee-Lueng Fu, Fabrice Ardhuin, Mounir Benkiran, Bertrand Chapron, Emmanuel Cosme, Francesco d’Ovidio, J Thomas Farrar, Sarah T Gille, Guillaume Lapeyre, et al. 2019. Global observations of fine-scale ocean surface topography with the surface water and ocean topography (SWOT) mission. *Frontiers in Marine Science*, 6:232.

- NASA PODAAC. 2023. GHRSSST level 4 MUR global foundation sea surface temperature analysis (v4.1) [Dataset]. <https://doi.org/10.5067/GHGMR-4FJ04>.
- Alberto C Naveira Garabato, Xiaolong Yu, Jörn Callies, Roy Barkan, Kurt L Polzin, Eleanor E Frajka-Williams, Christian E Buckingham, and Stephen M Griffies. 2022. Kinetic energy transfers between mesoscale and submesoscale motions in the open ocean’s upper layers. *Journal of Physical Oceanography*, 52(1):75–97.
- Akira Okubo. 1970. 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.
- OSISAF. 2017. Ghrsst level 3c indian-ocean (io) sub-skin sea surface temperature from the spinning enhanced visible and infrared imager (seviri) on msg1 in gds2 format produced by osisaf.
- Alexander Ostrovskii and Leonid Piterbarg. 1995. Inversion for heat anomaly transport from sea surface temperature time series in the northwest Pacific. *Journal of Geophysical Research: Oceans*, 100(C3):4845–4865.
- OSTST. 2024. Daily NeurOST L4 sea surface height and surface geostrophic currents [Dataset]. Accessed on 07-08-2024.
- Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.
- Pierrick Penven, Issufo Halo, Stéphane Pous, and Louis Marié. 2014. Cyclogeostrophic balance in the Mozambique Channel. *Journal of Geophysical Research: Oceans*, 119(2):1054–1067.
- Yuritzy Perez-Corona, Hector Torres, and Karina Ramos-Musalem. 2026. Detection and characterization of mesoscale eddies in the gulf of california using high-resolution satellite altimetry. *Remote Sensing*, 18(3):434.

- Bo Qiu and Shuiming Chen. 2025. Fine-scale upper-ocean variability in the kuroshio extension region from the wide-swath swot measurements. *Journal of Physical Oceanography*, 55(11):2229–2242.
- Bo Qiu, Shuiming Chen, Patrice Klein, Hideharu Sasaki, and Yoshikazu Sasai. 2014. Seasonal mesoscale and submesoscale eddy variability along the north pacific subtropical countercurrent. *Journal of Physical Oceanography*, 44(12):3079–3098.
- Bo Qiu, Shuiming Chen, Patrice Klein, Hector Torres, Jinbo Wang, Lee-Lueng Fu, and Dimitris Menemenlis. 2020. Reconstructing upper-ocean vertical velocity field from sea surface height in the presence of unbalanced motion. *Journal of Physical Oceanography*, 50(1):55–79.
- Bo Qiu, Toshiya Nakano, Shuiming Chen, and Patrice Klein. 2022. Bi-directional energy cascades in the pacific ocean from equator to subarctic gyre. *Geophysical Research Letters*, 49(8):e2022GL097713.
- Shikhar Rai, Matthew Hecht, Matthew Maltrud, and Hussein Aluie. 2021. Scale of oceanic eddy killing by wind from global satellite observations. *Science Advances*, 7(28):eabf4920.
- Remote Sensing Systems. 2017. Ghrsst level 4 mw-oi global foundation sea surface temperature analysis version 5.0 from REMSS [Dataset]. <https://doi.org/10.5067/GHMWO-4FR05>.
- M-H Rio, S Mulet, and N Picot. 2014. Beyond GOCE for the ocean circulation estimate: Synergetic use of altimetry, gravimetry, and in situ data provides new insight into geostrophic and ekman currents. *Geophysical Research Letters*, 41(24):8918–8925.
- M-H Rio, R Santoleri, R Bourdalle-Badie, A Griffa, L Piterbarg, and G Taburet. 2016. Improving the altimeter-derived surface currents using high-resolution sea surface temperature data: a feasibility study based on model outputs. *Journal of Atmospheric and Oceanic Technology*, 33(12):2769–2784.
- Cesar B Rocha, Teresa K Chereskin, Sarah T Gille, and Dimitris Menemenlis. 2016. Mesoscale to submesoscale wavenumber spectra in drake passage. *Journal of Physical Oceanography*, 46(2):601–620.
- Cesar B Rocha, Gregory L Wagner, and William R Young. 2018. Stimulated generation: Extraction of energy from balanced flow by near-inertial waves. *Journal of Fluid Mechanics*, 847:417–451.

- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.
- François Rozet, G r me Andry, Fran ois Lanusse, and Gilles Louppe. 2024. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682.
- Fran ois Rozet and Gilles Louppe. 2023a. Score-based data assimilation. *Advances in Neural Information Processing Systems*, 36:40521–40541.
- Fran ois Rozet and Gilles Louppe. 2023b. Score-based data assimilation for a two-layer quasi-geostrophic model. *arXiv preprint arXiv:2310.01853*.
- Daniel L Rudnick and Raffaele Ferrari. 1999. Compensation of horizontal temperature and salinity gradients in the ocean mixed layer. *Science*, 283(5401):526–529.
- Rick Salmon. 1980. Baroclinic instability and geostrophic turbulence. *Geophysical & Astrophysical Fluid Dynamics*, 15(1):167–211.
- Hideharu Sasaki, Patrice Klein, Bo Qiu, and Yoshikazu Sasai. 2014. Impact of oceanic-scale interactions on the seasonal modulation of ocean dynamics by the atmosphere. *Nature Communications*, 5(1):5636.
- Anna C Savage, Brian K Arbic, Matthew H Alford, Joseph K Ansong, J Thomas Farrar, Dimitris Menemenlis, Amanda K O’Rourke, James G Richman, Jay F Shriver, Gunnar Voet, et al. 2017. Spectral decomposition of internal gravity wave sea surface height in global models. *Journal of Geophysical Research: Oceans*, 122(10):7803–7821.
- Ren  Schubert, Jonathan Gula, Richard J Greatbatch, Burkard Baschek, and Arne Biastoch. 2020. The submesoscale kinetic energy cascade: Mesoscale absorption of submesoscale mixed layer eddies and frontal downscale fluxes. *Journal of Physical Oceanography*, 50(9):2573–2589.

- René Schubert, Oscar Vergara, and Jonathan Gula. 2023. The open ocean kinetic energy cascade is strongest in late winter and spring. *Communications Earth & Environment*, 4(1):450.
- Robert B Scott and Brian K Arbic. 2007. Spectral energy fluxes in geostrophic turbulence: Implications for ocean energetics. *Journal of physical oceanography*, 37(3):673–688.
- Robert B Scott and Faming Wang. 2005. Direct evidence of an oceanic inverse kinetic energy cascade from satellite altimetry. *Journal of Physical Oceanography*, 35(9):1650–1666.
- Hyodae Seo, Larry W O’Neill, Mark A Bourassa, Arnaud Czaja, Kyla Drushka, James B Edson, Baylor Fox-Kemper, Ivy Frenger, Sarah T Gille, Benjamin P Kirtman, et al. 2023. Ocean mesoscale and frontal-scale ocean–atmosphere interactions and influence on large-scale climate: A review. *Journal of climate*, 36(7):1981–2013.
- Guillaume Sérazin, Thierry Penduff, Bernard Barnier, Jean-Marc Molines, Brian K Arbic, Malte Müller, and Laurent Terray. 2018. Inverse cascades of kinetic energy as a source of intrinsic variability: A global ogcm study. *Journal of Physical Oceanography*, 48(6):1385–1408.
- Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.
- Lia Siegelman, Patrice Klein, Pascal Rivière, Andrew F Thompson, Hector S Torres, Mar Flexas, and Dimitris Menemenlis. 2020a. Enhanced upward heat transport at deep submesoscale ocean fronts. *Nature Geoscience*, 13(1):50–55.
- Lia Siegelman, Patrice Klein, Andrew F Thompson, Hector S Torres, and Dimitris Menemenlis. 2020b. Altimetry-based diagnosis of deep-reaching sub-mesoscale ocean fronts. *Fluids*, 5(3):145.
- Lia Siegelman, Baptiste Picard, and Christophe Guinet. 2025. Ubiquity and seasonality of deep submesoscales in the southern ocean revealed by elephant seals. *Research Square Pre-Print*.
- Simone Silvestri, Gregory L Wagner, Navid C Constantinou, Christopher N Hill, Jean-Michel Campin, Andre N Souza, Siddhartha Bishnu, Valentin Churavy, John Marshall, and Raffaele Ferrari. 2025. A gpu-

- based ocean dynamical core for routine mesoscale-resolving climate simulations. *Journal of Advances in Modeling Earth Systems*, 17(4):e2024MS004465.
- Anirban Sinha and Ryan Abernathey. 2021. Estimating ocean surface currents with machine learning. *Frontiers in Marine Science*, 8:672477.
- K Shafer Smith and Geoffrey K Vallis. 2001. The scales and equilibration of midocean eddies: Freely evolving flow. *Journal of Physical Oceanography*, 31(2):554–571.
- Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Maike Sonnewald, Redouane Lguensat, Daniel C Jones, Peter Dueben, Julien Brajard, and Venkatramani Balaji. 2021. Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*.
- Kaushik Srinivasan, Roy Barkan, and James C McWilliams. 2023. A forward energy flux at submesoscales driven by frontogenesis. *Journal of Physical Oceanography*, 53(1):287–305.
- Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using LSTMs. In *International conference on machine learning*, pages 843–852. PMLR.
- Jacob M Steinberg, Sylvia T Cole, Kyla Drushka, and Ryan P Abernathey. 2022. Seasonality of the mesoscale inverse cascade as inferred from global scale-dependent eddy energy observations. *Journal of Physical Oceanography*, 52(8):1677–1691.
- Benjamin A Storer and Hussein Aluie. 2023. FlowSieve: A coarse-graining utility for geophysical flows on the sphere. *Journal of Open Source Software*, 8(84):4277.
- Benjamin A Storer, Michele Buzzicotti, Hemant Khatri, Stephen M Griffies, and Hussein Aluie. 2022. Global energy spectrum of the general oceanic circulation. *Nature Communications*, 13(1):5314.

- Benjamin A Storer, Michele Buzzicotti, Hemant Khatri, Stephen M Griffies, and Hussein Aluie. 2023. Global cascade of kinetic energy in the ocean and the atmospheric imprint. *Science Advances*, 9(51):eadi7420.
- Ehud Strobach, Patrice Klein, Andrea Molod, Abdullah A Fahad, Atanas Trayanov, Dimitris Menemenlis, and Hector Torres. 2022. Local air-sea interactions at ocean mesoscale and submesoscale in a western boundary current. *Geophysical Research Letters*, 49(7):e2021GL097003.
- Zhan Su, Jinbo Wang, Patrice Klein, Andrew F Thompson, and Dimitris Menemenlis. 2018. Ocean submesoscales as a key component of the global heat budget. *Nature communications*, 9(1):775.
- Adam Subel, Yifei Guan, Ashesh Chattopadhyay, and Pedram Hassanzadeh. 2022. Explaining the physics of transfer learning a data-driven subgrid-scale closure to a different turbulent flow. *arXiv preprint arXiv:2206.03198*.
- Ziheng Sun, Laura Sandoval, Robert Crystal-Ornelas, S Mostafa Mousavi, Jinbo Wang, Cindy Lin, Nicoleta Cristea, Daniel Tong, Wendy Hawley Carande, Xiaogang Ma, et al. 2022. A review of earth artificial intelligence. *Computers & Geosciences*, page 105034.
- Guillaume Taburet, Antonio Sanchez-Roman, Maxime Ballarotta, Marie-Isabelle Pujol, Jean-François Legéais, Florent Fournier, Yannice Faugere, and Gerald Dibarboure. 2019. DUACS DT2018: 25 years of reprocessed sea level altimetry products. *Ocean Science*, 15(5):1207–1224.
- Olivier Talagrand. 1999. Evaluation of probabilistic prediction systems. In *Workshop Proceedings" Workshop on Predictability", 20-22 October 1997, ECMWF, Reading, UK*.
- Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z Li. 2022. SimVP: Towards simple yet powerful spatiotemporal predictive learning. *arXiv preprint arXiv:2211.12509*.
- John R Taylor and Andrew F Thompson. 2023. Submesoscale dynamics in the upper ocean. *Annual Review of Fluid Mechanics*, 55:103–127.
- Sylvie Thiria, Charles Sorrow, Theo Archambault, Anastase Charantonis, Dominique Bereziat, Carlos Mejia,

- Jean-Marc Molines, and Michel Crépon. 2023. Downscaling of ocean fields by fusion of heterogeneous observations using deep learning algorithms. *Ocean Modelling*, 182:102174.
- Leif N Thomas, Amit Tandon, and Amala Mahadevan. 2008. Submesoscale processes and dynamics. *Geophysical Monograph Series*, 177.
- Andrew F Thompson, Ayah Lazar, Christian Buckingham, Alberto C Naveira Garabato, Gillian M Damerell, and Karen J Heywood. 2016. Open-ocean submesoscale motions: A full seasonal cycle of mixed layer instabilities from gliders. *Journal of Physical Oceanography*, 46(4):1285–1307.
- Prasad G Thoppil, James G Richman, and Patrick J Hogan. 2011. Energetics of a global ocean circulation model compared to observations. *Geophysical Research Letters*, 38(15).
- Hector Torres, Alexander Wineteer, Patrice Klein, Tong Lee, Jinbo Wang, Ernesto Rodriguez, Dimitris Menemenlis, and Hong Zhang. 2023. Anticipated capabilities of the odyssey wind and current mission concept to estimate wind work at the air–sea interface. *Remote Sensing*, 15(13):3337.
- Hector S Torres, Patrice Klein, Dimitris Menemenlis, Bo Qiu, Zhan Su, Jinbo Wang, Shuiming Chen, and Lee-Lueng Fu. 2018. Partitioning ocean motions into balanced motions and internal gravity waves: A modeling study in anticipation of future space missions. *Journal of Geophysical Research: Oceans*, 123(11):8084–8105.
- Ross Tulloch, John Marshall, Chris Hill, and K Shafer Smith. 2011. Scales, growth rates, and spectral fluxes of baroclinic instability in the ocean. *Journal of Physical Oceanography*, 41(6):1057–1076.
- Clément Ubelmann, Loren Carrere, Chloé Durand, Gérald Dibarboure, Yannice Faugère, Maxime Ballarotta, Frédéric Briol, and Florent Lyard. 2022. Simultaneous estimation of ocean mesoscale and coherent internal tide sea surface height signatures from the global altimetry record. *Ocean Science*, 18(2):469–481.
- Clement Ubelmann, Bruce Cornuelle, and Lee-Lueng Fu. 2016. Dynamic mapping of along-track ocean altimetry: Method and performance from observing system simulation experiments. *Journal of Atmospheric and Oceanic Technology*, 33(8):1691–1699.

- Clement Ubelmann, Gérald Dibarboure, Lucile Gaultier, Aurélien Ponte, Fabrice Ardhuin, Maxime Ballarotta, and Yannice Faugère. 2021. Reconstructing ocean surface current combining altimetry and future spaceborne Doppler data. *Journal of Geophysical Research: Oceans*, 126(3):e2020JC016560.
- Clement Ubelmann, Patrice Klein, and Lee-Lueng Fu. 2015. Dynamic interpolation of sea surface height and potential applications for future high-resolution altimetry mapping. *Journal of Atmospheric and Oceanic Technology*, 32(1):177–184.
- Takaya Uchida, Ryan Abernathey, and Shafer Smith. 2017. Seasonality of eddy kinetic energy in an eddy permitting global climate model. *Ocean Modelling*, 118:41–58.
- Geoffrey K Vallis. 2017. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press.
- Gregory L Wagner, Simone Silvestri, Navid C Constantinou, Ali Ramadhan, Jean-Michel Campin, Chris Hill, Tomas Chor, Jago Strong-Wright, Xin Kai Lee, Francis Poulin, et al. 2025. High-level, high-resolution ocean modeling at all scales with oceananigans. *arXiv preprint arXiv:2502.14148*.
- Han Wang, Nicolas Grisouard, Hesam Salehipour, Alice Nuz, Michael Poon, and Aurélien L Ponte. 2022a. A deep learning approach to extract internal tides scattered by geostrophic turbulence. *Geophysical Research Letters*, 49(11):e2022GL099400.
- Jinbo Wang, Glenn R Flierl, Joseph H LaCasce, Julie L McClean, and Amala Mahadevan. 2013. Reconstructing the ocean’s interior from surface data. *Journal of Physical Oceanography*, 43(8):1611–1626.
- Shihong Wang, Zhenya Song, Weidong Ma, Qi Shu, and Fangli Qiao. 2022b. Mesoscale and submesoscale turbulence in the Northwest Pacific Ocean revealed by numerical simulations. *Deep Sea Research Part II: Topical Studies in Oceanography*, 206:105221.
- Shuo Wang, Xiaoyan Li, Xueming Zhu, Jiandong Li, and Shaojing Guo. 2024. Spatial downscaling of sea surface temperature using diffusion model. *Remote Sensing*, 16(20):3843.
- Yue Wang, Jingwen Lyu, Tatsu Monkman, C Spencer Jones, Christian Pedersen, and Dhruv Balwada. 2025. A multi-scale probabilistic machine learning model for balanced and unbalanced sea surface height decomposition. *Authorea Preprints*.

- John Weiss. 1991. The dynamics of enstrophy transfer in two-dimensional hydrodynamics. *Physica D: Nonlinear Phenomena*, 48(2-3):273–294.
- Carl Wunsch. 1999. Where do ocean eddy heat fluxes matter? *Journal of Geophysical Research: Oceans*, 104(C6):13235–13249.
- Qiyu Xiao, Dhruv Balwada, C Spencer Jones, Mario Herrero-González, K Shafer Smith, and Ryan Abernathey. 2023. Reconstruction of surface kinematics from sea surface height using neural networks. *Journal of Advances in Modeling Earth Systems*, 15(10):e2023MS003709.
- Chi Xu, Xiao-Dong Shang, and Rui Xin Huang. 2014. Horizontal eddy energy flux in the world oceans diagnosed from altimetry data. *Scientific Reports*, 4(1):1–7.
- Jang Gon Yoo, Sung Yong Kim, and Hyeon Seong Kim. 2018. Spectral descriptions of submesoscale surface circulation in a coastal region. *Journal of Geophysical Research: Oceans*, 123(6):4224–4249.
- Xiaolong Yu, Aurélien L Ponte, Shane Elipot, Dimitris Menemenlis, Edward D Zaron, and Ryan Abernathey. 2019. Surface kinetic energy distributions in the global oceans from a high-resolution numerical model and surface drifter observations. *Geophysical Research Letters*, 46(16):9757–9766.
- Edward D Zaron, Amy F Waterhouse, Jinbo Wang, Uwe Send, and Andrew J Lucas. 2026. Observations of baroclinic tides in the california current from the swot satellite altimeter and in situ data. *Journal of Geophysical Research: Oceans*, 131(2):e2025JC023119.
- Xiaoming Zhai, Helen L Johnson, and David P Marshall. 2010. Significant sink of ocean-eddy energy near western boundaries. *Nature Geoscience*, 3(9):608–612.
- Zhengguang Zhang, Bo Qiu, Patrice Klein, and Seth Travis. 2019. The influence of geostrophic strain on oceanic ageostrophic motion and surface chlorophyll. *Nature Communications*, 10(1):2838.
- Zhengguang Zhang, Wei Wang, and Bo Qiu. 2014. Oceanic mass transport by mesoscale eddies. *Science*, 345(6194):322–324.
- Ruichen Zhu, Yanqin Li, Zhaohui Chen, Tianshi Du, Yueqi Zhang, Zhuoran Li, Zhiyou Jing, Haiyuan

Yang, Zhao Jing, and Lixin Wu. 2023. Deep learning improves reconstruction of ocean vertical velocity. *Geophysical Research Letters*, 50(19):e2023GL104889.