ABSTRACT

Title of Dissertation: TOWARDS A SELF-CONTAINED ANALYSIS AND REANALYSIS SYSTEM FOR HURRICANES

Joseph Alan Knisely
Doctor of Philosophy, 2026

Dissertation Directed by: Professor Jonathan Poterjoy
Department of Atmospheric and Oceanic Sciences

Tropical cyclone (TC) prediction is one of the most challenging problems in atmospheric science. While track forecasts have improved dramatically over the past several decades, intensity forecasting continues to lag behind, particularly in cases of rapid intensification. This disparity reflects fundamental challenges: TC intensity is governed by complex interactions between inner-core dynamics, boundary layer processes, air-sea fluxes, and broader environmental factors. Numerical models must achieve very high resolution to resolve the eyewall region while simultaneously covering a domain large enough to capture steering flow and synoptic-scale interactions. TCs develop predominantly over observation-sparse oceans and thus rely heavily on satellite measurement systems that may introduce significant biases. Modern TC-focused systems, such as NOAA's Hurricane Analysis and Forecast System (HAFS), address some of these challenges but rely heavily on global models for initial and boundary conditions, and bias correction parameters. This dependence limits our ability to diagnose deficiencies within the model itself and impedes progress toward a fully mature, self-contained, ensemble-based TC forecasting

system. This dissertation advances hurricane forecasting through three complementary studies. First, we perform idealized experiments within a low-dimensional chaotic dynamical model to examine fundamental limitations of observation bias correction in the presence of model bias. We find that anchor observation-based methodologies can successfully separate biases only in proximity to anchor observations, but that without sufficient coverage, bias reinforcement and drift can occur. Ultimately, diagnosing and addressing model bias is the best strategy for mitigating bias reinforcement. Second, we demonstrate that HAFS can operate in a fully-cycled, self-contained mode with online satellite radiance bias correction. Experiments show that training bias correction coefficients on the native HAFS grid produces improved TC forecasts compared to adopting coefficients from external global models. Third, we extend the self-contained framework to a basin-scale ensemble data assimilation and forecasting system. Comparing different resolution configurations reveals substantial differences in cloud coverage and composition, generating significant domain-wide biases with consequent impacts on TC forecasts. Furthermore, we demonstrate the benefit of an ensemble-based framework. While obstacles remain, clear patterns emerge that inform future model development. Taken together, this work establishes a foundation for a self-contained, continuously-cycled, probabilistic TC analysis system, essential for diagnosing model deficiencies and for developing an analysis of record to support forecast verification and machine learning applications.

TOWARDS A SELF-CONTAINED ANALYSIS AND REANALYSIS SYSTEM FOR HURRICANES

by

Joseph Alan Knisely

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:
Professor Jonathan Poterjoy, Chair
Professor Kayo Ide
Professor Timothy Canty
Dr. Daryl Kleist
Dr. Elizabeth Satterfield
Professor Michael Evans, Dean's Representative

© Copyright by
Joseph Alan Knisely
2026

Dedication

This dissertation is dedicated to all workers and oppressed peoples of the world.

Acknowledgments

First, I would like to express my deepest gratitude to my advisor, Prof. Jonathan Poterjoy, for his guidance, feedback, and encouragement throughout my Ph.D. Entering the field of atmospheric sciences was a significant transition for me, and Jon was always patient in his mentorship and available for any question, brainstorming session, or discussion about my work. Jon is a brilliant scientist, and I am lucky to have been his student.

I would like to thank Dr. Elizabeth Satterfield and Dr. William "Bill" Campbell for serving as my mentors during my internship at NRL Monterey, and for their continued mentorship to this day. Your guidance has helped me grow as both a researcher and a scientist.

I acknowledge the significance of the WINGS fellowship in funding much of my Ph.D. and for providing me with opportunities to travel and present at workshops and conferences. I would like to thank the fellowship administrators, as well as Dr. Daryl Kleist, my mentor through this program, for his career guidance and assistance with NWP-related questions.

I would also like to thank Profs. Brian Hunt, Kayo Ide, Tim Canty, and Michael Evans for serving on my committee and providing valuable feedback on my dissertation and defense. I extend my thanks to the faculty and staff of the Department of AOSC. In particular, I would like to thank my fellow graduate students for the camaraderie we shared together.

I have profound gratitude for my family and friends outside of academia, who kept me grounded throughout this journey. I extend particular thanks to my parents, who were always incredibly supportive and patient, and to my brothers, who motivated me to pursue a Ph.D. in the first place.

I would like to thank all my comrades in the Graduate Labor Union. The solidarity of our union carried me through the toughest moments of my Ph.D., and the organizing work we did together gave my life purpose beyond research. I wish the union continued strength in the fight for recognition of our

essential labor in the years to come.

Last but certainly not least, my girls: my two beautiful cats, Maya and Maggi, who have filled my heart with boundless love; and my partner, Akanksha. Words cannot express how grateful I am to have you in my life. I would not have completed this Ph.D. without you.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Figures vii

List of Abbreviations xiii

Chapter 1: Introduction 1
- 1.1 Tropical Cyclone Prediction . . . 1
- 1.2 Data Assimilation for NWP . . . 3
  - 1.2.1 Variational Methods . . . 4
  - 1.2.2 Ensemble Kalman Filter . . . 5
  - 1.2.3 Particle Filters . . . 6
  - 1.2.4 Hybrid DA . . . 7
- 1.3 Satellite Radiance Observations . . . 8
- 1.4 Observation Bias Correction . . . 9
- 1.5 Motivation and Dissertation Overview . . . 11

Chapter 2: Limitations of Observation Bias Correction Investigated with a Low-Dimensional Dynamical Model 13
- 2.1 Introduction . . . 13
- 2.2 Estimating Observation and Model Error . . . 16
  - 2.2.1 Definitions . . . 16
  - 2.2.2 Accumulated Bias . . . 18
  - 2.2.3 Bias Estimation . . . 22
  - 2.2.4 Linear Bias Correction . . . 24
  - 2.2.5 Model Bias Correction . . . 28
- 2.3 Methodology . . . 30
  - 2.3.1 Model Description and Bias Simulation . . . 30
  - 2.3.2 Data Assimilation Configuration . . . 32
  - 2.3.3 Linear Bias Correction Methodologies . . . 32
- 2.4 Results . . . 35
  - 2.4.1 Observation Bias Correction Results . . . 35
  - 2.4.2 Model Correction Results . . . 47
- 2.5 Summary and Conclusions . . . 52

Chapter 3: Implications of Self-Contained Radiance Bias Correction for Data Assimilation within the Hurricane Analysis and Forecasting System (HAFS) 56
- 3.1 Introduction . . . 56

3.2 Radiance Bias Correction . . . 59
3.3 Model Configuration . . . 62
3.3.1 Forecast and Analysis . . . 62
3.3.2 Model Physics . . . 64
3.3.3 Bias Correction Configurations . . . 65
3.3.4 Observations and Data Assimilation Schedule . . . 66
3.4 Results . . . 67
3.4.1 Forecast Error Verification . . . 67
3.4.2 Domain-Averaged Error Verification . . . 72
3.4.3 Temperature and Humidity Fields . . . 75
3.4.4 Predictor Evolution and Verification . . . 79
3.4.5 Weighted Predictor Fields . . . 83
3.4.6 Case Study . . . 85
3.5 Discussion and Conclusions . . . 90

Chapter 4: Obstacles for High-Resolution, Basin-Scale Probabilistic Hurricane Forecasting 94
4.1 Introduction . . . 94
4.2 Model Configuration . . . 98
4.2.1 Static Basin-wide Analysis Workflow . . . 100
4.2.2 Storm-centric Ensemble Forecast Workflow . . . 103
4.2.3 Storm-centric Parallel Workflow . . . 104
4.2.4 Model Physics . . . 105
4.2.5 Observations and Data Assimilation Schedule . . . 106
4.3 Results . . . 107
4.3.1 Cloud Fraction and Hydrometeor Concentration . . . 107
4.3.2 Domain-wide Analysis Verifications . . . 111
4.3.3 Heat Fluxes and Feedbacks . . . 118
4.3.4 Domain-wide Forecast Verification . . . 119
4.3.5 Storm-centric Forecast TC-Metrics . . . 124
4.4 Conclusions . . . 131

Chapter 5: Summary, Future Work, and Conclusions 133
5.1 Chapter Summaries . . . 133
5.2 Future Work . . . 135
5.3 Conclusions . . . 137

Bibliography 139

List of Figures

2.1 Straightforward demonstration of bias reinforcement in bias correction schemes that do not consider the effects of model bias. In blue, the biased model trajectory is regularly updated with observations which are themselves biased. For the "Naive" bias correction, observations are corrected to a model background. Since the background itself is biased, the bias correction reinforces bias and leads to model drift, shown in red. 22
2.2 Lorenz 2005 Model III sample solution state. The truth state, Z, can be decomposed into a large-scale, X, and a small-scale, Y. Small-scale activity is driven by large-scale activity through a coupling term, such that regions where Y is active tend to follow local maximums/minimums of X. 30
2.3 Percent improvement in RMSE (top row) and absolute bias (bottom row) for four bias correction methods relative to experiments with no bias correction. From left to right, columns show: scalar mean correction, scalar mean correction with inflation, vector correction, and vector correction with inflation. Each heatmap displays results averaged over the final 100 assimilation cycles across 25 bias configurations, with rows indicating imposed model bias and columns indicating imposed observation bias. Blue shading indicates improvement over control, while red shading indicates degradation. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. 36
2.4 Percent improvement in RMSE (top row) and absolute bias (bottom row) for four bias correction methods relative cycles with anchor observations only. From left to right, columns show: scalar mean correction, scalar mean correction with inflation, vector correction, and vector correction with inflation. Each heatmap displays results averaged over the final 100 assimilation cycles across 25 bias configurations, with rows indicating imposed model bias and columns indicating imposed observation bias. Blue shading indicates improvement over control, while red shading indicates degradation. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. 37
2.5 Percent improvement in time-averaged RMSE relative to experiments with no bias correction for each bias correction methodology, aggregated across 25 model/observation bias configurations. Rows correspond to standard observation frequency (every 1 or 2 grid points), columns to anchor observation frequency (every 8, 16, or 32 grid points). Blue shading indicates improvement over control, while red shading indicates degradation. Cell annotations denote percent change. 39

2.6 Percent improvement in time-averaged RMSE relative to the spin-up period, in which only anchor observations are assimilated, for each bias correction methodology, aggregated across 25 model/observation bias configurations. Rows correspond to standard observation frequency (every 1 or 2 grid points), columns to anchor observation frequency (every 8, 16, or 32 grid points). Blue shading indicates improvement over control, while red shading indicates degradation. Cell annotations denote percent change. 40
2.7 Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction. 41
2.8 Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 32 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction. 42
2.9 Temporal evolution of RMSE & absolute bias stratified by observation coverage. a) gives standard observation RMSE, b) gives anchor observation RMSE, c) gives standard observation absolute bias, and d) gives anchor observation absolute bias. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. Results are averaged across 25 bias configurations. 43
2.10 Temporal evolution of RMSE & absolute bias stratified by observation coverage. a) gives standard observation RMSE, b) gives anchor observation RMSE, c) gives standard observation absolute bias, and d) gives anchor observation absolute bias. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 32 grid points, respectively. Results are averaged across 25 bias configurations. . 44
2.11 Time-averaged RMSE as a function of (a) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (b) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (c) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points, and (d) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points. Each line represents a different bias correction method. 45

2.12 Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction. 49
2.13 Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 32 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction. 50
2.14 Time-averaged RMSE as a function of (a) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (b) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (c) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points, and (d) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points. Each line represents a different bias correction method. 51

3.1 In blue, the entire HAFS-SAR model domain. In orange, a smaller subdomain selected for error verification (see section 3.4). 63
3.2 Forecast error verification of control and online experiments. Panel (a) reflects error for all TCs over the experimental period, while panel (b) omits TCs without aircraft reconnaissance data. These errors are verified against NHC best track and intensity data. 68
3.3 Error verification of forecast absolute error differences between control and online experiments (online-control). The median difference is given in blue, while the inner and outer quintiles are given in purple and red, respectively. Errors are verified against NHC best track and intensity data. 71
3.4 Panels (a) and (b) display bias and RMSE verification of temperature, specific humidity, u- and v- winds with ERA5 data, respectively. Data is time- and domain-averaged and displayed as a vertical profile. 73
3.5 Temperature, specific humidity, u- and v-winds RMSE verification with ERA5 data. Data is domain- and column-averaged, and taken at 6-hour lead-time increments up to 102 hours. Columns span vertically from 900 to 250 mb for temperature and wind errors, and from 900 to 500 mb for specific humidity. 74
3.6 Vertical profiles of domain- and time-averaged temperature (a) and specific humidity (b) differences between the online and control experiment (online - control). Differences over-land, over-sea, and total differences are all considered. 76

3.7 Time-averaged temperature differences between the online and control experiment at 950, 850, 750, 600, 450, and 300 mb (online - control). Note that a Gaussian smoothing filter has been applied. . . . 77
3.8 Time-averaged specific humidity differences between the online and control experiment at 950, 850, 750, 600, 450, and 300 mb (online - control). Likewise, a Gaussian smoothing filter has been applied. . . . 77
3.9 (a) Bias predictor coefficient evolution over the course of the 5-week experiment period the GDAS BC (red) and HAFS BC (blue) for channel 4 of the AMSU-A N15 satellite instrument. (b) Weighted bias predictor value evolution over the same period, in units of BT (K). . . . 79
3.10 These panels detail time- and domain-averaged weighted predictor values for select instruments and channels for each experiment. The x-axis gives each weighted predictor, while the red, blue, and black colored bars give their time- and domain-averaged values for the control experiment, online experiment, and their difference, respectively. They are: (a) MHS MetOp-B, channel 1; (b) AMSU-A N15, channel 4; (c) MHS N19, channel 1; (d) AMSU-A N18, channel 11; (e) AMSU-A N18, channel 11; (f) AMSU-A MetOp-A, channel 11. . . . 82
3.11 Time-averaged weighted predictor differences between online and control experiments (online-control) for channel 4 of the AMSU-A N15 instrument, in units of BT (K). . . . 83
3.12 On the left, the time-average temperature field difference between control and online at 950 mb (online-control). On the right, the difference in the emissivity predictor between control and online for channel 4 on the AMSU-A N15 instrument. Note that the channel 4 weighting function peaks near the surface. . . . 84
3.13 Track Error Verification, Control (a) vs Online (b), for AL20, Hurricane Teddy. The blue lines represent forecasts generated early in the model time span, transitioning to red as forecast time progresses. Black represents the best track estimate. . . . 85
3.14 Same as Fig. 3.13 but for 10-m maximum wind speed. . . . 86
3.15 Forecast error verification of control and online experiments for AL20, Teddy. These errors are verified against NHC best track and intensity data. . . . 87
3.16 (a) displays a time series of model field differences between online and control (online-control) for 4 pertinent model variables: temperature, specific humidity, relative humidity, and geopotential height. The verification domain for these variables forms an annulus around the TC, extending from 200 to 2000 km around the TC. The values are column averaged from 950 to 500 mb. (b) displays a time-averaged vertical profile of the same variables over the same verification domain. . . . 88
3.17 (a) displays the 048 hours lead-time 250 mb geopotential height model field generated at Sept. 21st 00z for the control experiment. (b) gives the geopotential height difference between online and control at the same forecast time, lead-time, and pressure level (online-control). In both cases, a Gaussian smoothing filter has been applied. A black "X" marks the NHC best estimate location for Hurricane Teddy at Sept. 23rd 00z. . . . 88

4.1 In green, the static basin-wide analysis domain. In blue, an example of a storm-centric forecast domain with a high-resolution moving nest in red. . . . 98

4.2 In the blue, the diagram details the static basin-wide analysis workflow. Notice both the deterministic variational and ensemble analyses/forecasts, and the augmentation of the 40-member high-resolution HAFS ensembles with the 80-member GDAS ensemble. In red, we show how 10 ensemble members from the basin-wide forecasts are borrowed and used as IC for long storm-centric forecasts, elsewhere in this chapter also referred to as "self-contained" ensemble forecasts. In green, we show the "parallel" workflow, which closely resembles the operational HAFS workflow, but with borrowed ensembles from the basin-wide analysis workflow for enhanced background covariance specification. 99
4.3 Time- and zonal-mean of cloud fraction, hydrometeor variables (liquid water, ice water, rain water, snow water, and graupel) and specific humidity for 3 and 6 km basin-wide forecasts, and their difference, respectively. Notice the difference in y-axis scale between cloud fraction, hydrometeors, and specific humidity. 108
4.4 Time- and domain-mean column-mean values over time for (a) cloud fraction, hydrometeor variables ((b) liquid water, (c) ice water, (d) rain water, (e) snow water, and (f) graupel), and (g) specific humidity for 3 and 6 km basin-wide forecasts. Notice the difference in y-axis scale between cloud fraction, the hydrometeors, and specific humidity 109
4.5 Time- and domain-mean differences over time for (a) downward short- and (b) long-wave radiative fluxes at the surface, along with (c) surface temperature for 3 and 6 km basin-wide forecasts. Positive values indicate 3 km is greater, negative values indicate 6 km is greater. Differences over-land, over-ocean, and total differences are all considered. Note the difference in y-axis flux values. 112
4.6 Vertical profile of time- and domain-mean differences of observations minus 6 hour forecasts. Panels give (a) temperature RMSE,(b) temperature bias, (c) humidity RMSE, and (d) humidity bias. Note that only sondes were used as observations for this verification, given their high accuracy and precision. 113
4.7 Panels (a), (b), (c), (d), and (e) show vertical profiles of time- and domain-mean RMSE for temperature, relative humidity, specific humidity, and the u- and v-winds, respectively, for 3 km and 6 km basin-wide forecasts verified against ERA5. Panels (f), (g), (h), (i), and (j) show vertical profiles of time- and domain-mean bias for the same variables. 114
4.8 Time- and domain-mean differences over time for (a) sensible and (b) latent fluxes at the surface, along with surface temperature for 3 and 6 km basin-wide forecasts. Positive values indicate 3 km is greater, negative values indicate 6 km is greater. Differences over-land, over-ocean, and total differences are all considered. Note the difference in y-axis flux values. 116
4.9 Time- and domain-mean differences over time for (a) column total precipitation, (b) convective precipitation, (c) non-convective precipitation, and (d) CAPE for 3 and 6 km basin-wide forecasts. Positive values indicate 3 km is greater, negative values indicate 6 km is greater. Differences over-land, over-ocean, and total differences are all considered. Note the difference in y-axis precip. values. 117
4.10 Domain- and time-mean RMSE for (a) temperature, (b) sp. humidity, (c) u-wind, and (d) v-wind for storm-centric forecasts initialized from 3 km IC, 6 km IC, and operational HAFS IC. Verified against ERA5. 120

4.11 Time- and lower tropospheric column-mean temperature and relative humidity RMSE, verified against ERA5, for (a) and (d) operational HAFS-B IC, (b) and (e) 3 km HAFS IC, and (c) and (f) 6 km HAFS IC, respectively. Column-mean is taken from 950 to 700 mb. 122
4.12 Time- and lower tropospheric column-mean temperature and relative humidity RMSE, verified against ERA5, for (a) and (d) operational HAFS-B Forecast, (b) and (e) Self-Contained Forecast from 3 km Analysis, and (c) and (f) Self-Contained Forecast from 6 km Analysis, respectively. Column-mean is taken from 950 to 700 mb. 123
4.13 Sea surface temperature (SST) bias (HAFS - RTOFS) as a function of forecast lead-time for 3 km and 6 km self-contained workflows and operational HAFS-B. (a) shows the inner core region (0 to RMW), (b) shows the outer core region (50 to 100 nm). Bias is computed as the spatial mean SST within each annular region, averaged over all initialization times and storms. 123
4.14 TC forecast error verification of both 3 km and 6 km self-contained and parallel workflows, and operational HAFS-B. Panels give RMSE and bias for (a) track, (b)(e) SLP, (c)(f) max wind, and (d)(g) max wind radius, respectively. Case count ranges from 61 to 59. 125
4.15 TC forecast error verification of 3 km and 6 km self-contained workflows and operational HAFS-B. Panels give RMSE and bias for (a) track, (b)(e) SLP, (c)(f) max wind, and (d)(g) max wind radius, respectively. Both the deterministic forecast and the 10 member forecast ensemble mean included for self-contained workflows. Case count ranges from 61 to 59. 126
4.16 Domain-mean values over time for (a)(d) total, (b)(e) convective, and (c)(f) non-convective precipitation for Danielle and Earl, respectively. These values are taken at 6 hr forecast over storm-centric parent domain. 127
4.17 TC forecast error verification of 6 km parallel workflow and operational HAFS-B, both with and without vortex initialization. Panels give RMSE and bias for (a) track, (b)(e) SLP, (c)(f) max wind, and (d)(g) max wind radius, respectively. Case count ranges from 61 to 59. 129

5.1 Forecast RMSE verification of track, SLP, max wind, and max wind radius for 10 member forecast ensemble mean for 3 km and 6 km self-contained workflows, with corresponding ensemble spread. Case count ranges from 61 to 59. 136

List of Abbreviations

AMSU	Advanced Microwave Sounding Unit
BC	Bias Correction
BT	Brightness Temperature
CLW	Cloud Liquid Water
CMEPS	Community Mediator for Earth Prediction Systems
CRTM	Community Radiative Transfer Model
DA	Data assimilation
ECMWF	European Centre for Medium-Range Weather Forecasts
EDMF	Eddy Diffusion Mass-Flux
FGAT	First Guess at the Appropriate Time
GDAS	Global Data Assimilation System
GEFS	Global Ensemble Forecast System
GFDL	The Geophysical Fluid Dynamics Lab
GFS	Global Forecast System
GSI	Gridpoint Statistical Interpolation
HAFS	Hurricane Analysis and Forecasting System
HFIP	The Hurricane Forecast Improvement Project
HWRF	Hurricane Weather Research and Forecasting model
HYCOM	Hybrid Coordinate Ocean Model
IC	Initial Conditions
JCSDA	Joint Center for Satellite Data Assimilation
LAM	Limited-Area Model
LBC	Lateral Boundary Conditions
LPF	Local Particle Filter
NOAA	National Oceanic and Atmospheric Administration
Noah-MP LSM	Noah-Multiparameterization Land Surface Model
NWP	Numerical Weather Prediction
MHS	Microwave Humidity Sounder
ML	Machine Learning
NCEP	National Centers for Environmental Prediction
NHC	National Hurricane Center
PBL	Planetary Boundary Layer
PF	Particle Filter
RI	Rapid Intensification
RMSE	Root Mean Square Error
RRTMG	Rapid Radiative Transfer Model for GCMs
RTOFS	Real Time Ocean Forecast System
SAR	Stand-Alone Regional
sa-SAS	Scale-Aware Simplified Arakawa-Schubert

SLP	Sea Level Pressure
TC	Tropical Cyclone
TKE	Turbulence Kinetic Energy
US	United States
UTC	Coordinated Universal Time
VI	Vortex Initialization

Chapter 1: Introduction

1.1 Tropical Cyclone Prediction

Tropical cyclones (TCs) are among the most destructive and deadly severe weather events, causing widespread loss of life, infrastructure and ecosystem damage, and economic disruption in coastal regions worldwide. TCs are some of the deadliest natural disasters in human history, with the most devastating being the 1970 Bhola cyclone, which killed nearly half a million people and countless other species in eastern India and present-day Bangladesh [Frank and Husain, 1971]. In the United States alone, TCs and hurricanes have caused over $1.3 trillion in damages and over 6800 deaths since 1980, including nearly 3000 killed in Puerto Rico in 2017 due to Hurricane Maria [Kishore et al., 2018, Smith and Katz, 2013]. The improvement in US death toll numbers can in large part be attributed to early warnings from TC prediction centers and government preparedness. Early warning systems allow emergency services and the general public to take preventive measures to safeguard homes and lives. However, TC prediction remains one of the most challenging problems in the field of atmospheric prediction.

Over the past several decades, the numerical weather prediction (NWP) capabilities to simulate and predict TCs have improved considerably. TC track forecasts have improved dramatically, with errors reduced by about half since the mid 1990s, while intensity forecasting skill has lagged behind, albeit with notable progress since 2010 [Cangialosi et al., 2020, Gall et al., 2013]. In particular, inadequate prediction of rapid intensification (RI) is a serious problem. To add to the issue, the number of rapid intensification events occurring near coasts has increased significantly in recent decades due to global warming, reducing warning lead times and compounding forecast difficulties [Li et al., 2023].

The slow-paced improvements in intensity forecasting reflect fundamental challenges. TC intensity is governed by complex interactions between inner-core dynamics, boundary layer processes,

air-sea fluxes, and environmental factors such as vertical wind shear and sea surface temperature [Emanuel and Zhang, 2016b, Hendricks et al., 2010]. Numerical models must have very high resolutions to resolve the eye-wall region, yet simultaneously cover a domain large enough to capture steering flow and other synoptic-scale interactions that govern track [Kaplan and DeMaria, 2003, Zhang et al., 2015]. At these resolutions, model physics should resolve convection explicitly, already a significant task, while also handling boundary layer and air-sea interactions at high speeds [Bryan et al., 2003, Emanuel and Zhang, 2016b, Hazelton et al., 2024, Marks et al., 1998, Wang et al., 2023]. Furthermore, accurate analyses require a combination of sparse but valuable in-situ inner-core observations as well as remotely sensed observations, such as from satellite radiance measurement systems. There are also fundamental challenges for convective-scale and coupled assimilation that are relevant for TC initialization. [Aksoy et al., 2012, Liu et al., 2012, Sippel et al., 2022, Zhang et al., 2009].

We perform experiments in the Hurricane Analysis and Forecast System (HAFS), National Oceanic and Atmospheric Administration's (NOAA) flagship TC prediction model. The operational HAFS is a limited-area model that relies heavily on the Global Data Assimilation System (GDAS) for initial conditions (ICs), lateral boundary conditions (LBCs), and key parameters such as radiance bias correction coefficients [Hazelton et al., 2023, Ramstrom et al., 2024]. HAFS is partially-cycled, meaning it assimilates mainly inner-core observations over a small storm-centric nest only, and features ad-hoc methodologies such as vortex initialization (VI) [Poyer et al., 2025]. These design choices restrict the ability to employ statistical methodologies such as multi-scale flow-dependent error covariances and native domain-trained observation bias estimation [Knisely and Poterjoy, 2023, Poterjoy et al., 2021]. Furthermore, this reliance on external models limits our ability to diagnose and correct deficiencies within the regional hurricane model itself, and impedes progress toward a fully mature, self-contained, probabilistic TC forecasting system.

The Hurricane Forecast Improvement Project (HFIP), introduced by NOAA in 2008, has highlighted and emphasized the development of enhanced computational infrastructure, improved observational strategies, advanced data assimilation (DA), and expanded application of high-resolution and ensemble-based numerical modeling systems [Gall et al., 2013]. In this dissertation, we seek to address the latter three: improving observational strategies for satellite radiances, advancing DA methodologies, and developing high-resolution ensemble-based modeling capabilities for HAFS.

We also recognize the increasing significance of machine-learning (ML)-based methods in replacing components of NWP systems, including subgrid-scale parameterizations, measurement operators, and even the model forecast itself. These data-driven models can produce fast and accurate weather forecasts, including TC forecasts [Alet et al., 2025]. While such systems often excel at reducing track errors compared to traditional dynamical models, they struggle with intensity forecasting [DeMaria et al., 2024]. This limitation likely stems in part from the coarse resolution of reanalysis training data which cannot resolve TC inner-core structure, highlighting the potential benefit of basin-scale TC-specific reanalyses at convective-permitting resolutions.

1.2 Data Assimilation for NWP

Free-running NWP forecast models will accumulate errors over time until their forecast is no longer useful. DA updates model trajectories with observations to constrain errors and produce accurate and physically-consistent state estimates, which then serve as ICs for subsequent forecasts [Bonavita and Lean, 2021, Kalnay, 2003]. The core challenge of DA is in dealing with errors in models and observations, both whose error distributions may be unknown and difficult to determine [Carrassi et al., 2018]. In other words, the analysis must merge imperfect observations with a forecast background that

includes both model error from the current cycle as well as accumulated errors from all previous cycles, all while preserving basic dynamical balances.

To accomplish this task, we use statistical methods that consider both the model and observations and their associated uncertainties. These methods fall into broadly three categories: (1) variational methods, such as three-dimensional variational (3DVar) or four-dimensional variational (4DVar); (2) ensemble Kalman filters (EnKF); and (3) Pure Bayesian methods, such as the particle filter (PF) [Bannister, 2017, Evensen, 1994, Leeuwen, 2009, Lorenc, 1986]. Additionally, these methods can be combined into hybrid methods, namely ensemble-variational (EnVar).

1.2.1 Variational Methods

Variational DA seeks to determine an optimal deterministic analysis state $\mathbf{x}^a$ that minimizes a cost function measuring the fit to both the background forecast and observations [Courtier et al., 1994, Lorenc, 1986]. The 3DVar cost function is given by

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}(\mathbf{y} - H(\mathbf{x}))^T \mathbf{R}^{-1}(\mathbf{y} - H(\mathbf{x})), \tag{1.1}$$

where $\mathbf{x}^b$ is the background state, $\mathbf{y}$ is the observation vector, H is the observation operator that maps the model state to observation space, $\mathbf{B}$ is the background error covariance matrix, and $\mathbf{R}$ is the observation error covariance matrix. The first term fits deviations from the background, while the second fits deviations from the observations.

4DVar extends this framework by incorporating observations distributed over a time window and propagating the model state forward using the forecast model $\mathcal{M}$. The 4DVar cost function can be

written as

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_0^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_0^b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - H_i(\mathbf{x}_i))^T \mathbf{R}_i^{-1}(\mathbf{y}_i - H_i(\mathbf{x}_i)), \tag{1.2}$$

where $\mathbf{x}_0$ is the control state at the beginning of the assimilation window, $\mathbf{x}_i = \mathcal{M}_{0\to i}(\mathbf{x}_0)$ is the model state propagated to time t_i, and the sum runs over all observation times within the window. Minimization of Eq. 1.2 requires the adjoint of both the observation operator and the forecast model [Talagrand and Courtier, 1987, Zupanski, 1997].

1.2.2 Ensemble Kalman Filter

Ensemble-based methods estimate the background error covariance from an ensemble of forecasts, compared to variational methods which use a prescribed static covariance [Evensen, 1994, Houtekamer and Mitchell, 1998, Whitaker and Hamill, 2002]. In EnKF, an ensemble of N background forecasts $\mathbf{x}_i^b$ (for $i = 1, \ldots, N$) is propagated forward by the forecast model, and the covariance matrix of these forecasts provides a flow-dependent estimate of background uncertainty. The ensemble mean background is given by

$$\overline{\mathbf{x}}^b = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^b, \tag{1.3}$$

and the ensemble perturbation matrix is $\mathbf{X}^b = (\mathbf{x}_1^b - \overline{\mathbf{x}}^b, \ldots, \mathbf{x}_N^b - \overline{\mathbf{x}}^b)$. The background error covariance is then approximated as

$$\mathbf{P}^b \approx \frac{1}{N-1}\mathbf{X}^b(\mathbf{X}^b)^T. \tag{1.4}$$

The analysis update follows the Kalman filter equations [Kalman, 1960]. For a linear observation operator $\mathbf{H}$, the Kalman gain is

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{1.5}$$

and the analysis mean is

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}}^b\right). \tag{1.6}$$

The innovation $\mathbf{y} - \mathbf{H}\overline{\mathbf{x}}^b$ gives the difference between observations and the background projected into observation space. Ensemble perturbations are updated to reflect the reduction in uncertainty from the observations, with specific implementations differing in how this update is performed [Burgers et al., 1998, Whitaker and Hamill, 2002]. A key advantage of the EnKF is that the background error covariance is flow-dependent, allowing the filter to capture "errors-of-the-day," or state-dependent error structures that static covariances cannot represent.

1.2.3 Particle Filters

While the EnKF provides flow-dependent error estimates, its analysis update is optimal only when errors are Gaussian and the measurement operator is linear. For strongly nonlinear systems or non-Gaussian error distributions, PFs offer a fully Bayesian alternative that does not rely on these assumptions [Leeuwen, 2009].

From Bayes' theorem, the posterior probability density function (PDF) of the model state is proportional to the product of the prior PDF and the likelihood:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x}). \tag{1.7}$$

Given N particles, the PF approximates this PDF as a sum of Dirac delta functions, such that the posterior PDF is approximated by the sum of weighted delta functions:

$$p(\mathbf{x}|\mathbf{y}) \approx \sum_{i=1}^{N} w_i \delta(\mathbf{x} - \mathbf{x}_i) \tag{1.8}$$

where the weights are proportional to the likelihood of particles given observations:

$$w_i = \frac{p(\mathbf{y}|\mathbf{x}_i)}{\sum_{m=1}^{N} p(\mathbf{y}|\mathbf{x}_m)}. \tag{1.9}$$

Unlike the EnKF, this update makes no assumption about the shape of the prior or posterior distributions.

A major challenge for particle filters in high-dimensional systems is filter collapse, wherein a single particle accumulates nearly all the weight after assimilating many observations [Snyder et al., 2008, van Leeuwen et al., 2019]. The local particle filter (LPF) addresses this by performing the weight update independently within localized regions of the domain, preventing any single particle from dominating globally [Poterjoy et al., 2019]. This localization strategy enables particle filters to be applied to geophysical systems with state dimensions far exceeding the ensemble size. The LPF has shown particular promise for systems with nonlinear dynamics and observation operators, such as the assimilation of satellite radiances for convective-scale prediction [Emanuel and Zhang, 2016b, Kurosawa and Poterjoy, 2021, Poterjoy et al., 2019].

1.2.4 Hybrid DA

In addition to pure variational and ensemble methods, operational centers frequently employ hybrid methods, which combine the strengths of both variational and ensemble-based approaches.

Hybrid methods adopt a variational framework but replace or blend the static background error covariance $\mathbf{B}$ with a flow-dependent covariance $\mathbf{P}^b$ derived from ensembles [Bannister, 2017, Hamill and Snyder, 2000]. In doing so, DA systems benefit from the ability to capture coherent "errors-of-the-day" while minimizing sensitivity to the sampling error that is common in pure ensemble methods [Clayton et al., 2013]. This hybrid EnVar framework forms the basis of the DA system used in Chapters 3 and 4 of this dissertation. Like the EnKF, the LPF can also be combined with variational methods to form a hybrid LPF-Var, discussed in the "Future Works" section of Chapter 5.

1.3 Satellite Radiance Observations

Satellite radiances are passive measurements of upwelling radiation emitted or reflected from the Earth's surface and atmosphere. They provide a wealth of information for characterizing temperature and moisture fields at a very high spatial-and temporal-resolution [Harris and Kelly, 2001]. These measurements are expressed as brightness temperatures, which are related to the measured radiance via the Planck function. The vertical sensitivity of a specific radiometer channel is defined by its weighting function, which is determined by the atmospheric absorption profile and the viewing geometry [Eyre, 1992]. Infrared channels offer high spectral and vertical resolution but cannot penetrate clouds, limiting their use to clear-sky conditions. Microwave channels have coarser spectral resolution but can penetrate most cloud cover, though challenges remain in heavy precipitation and for surface-sensitive channels. By utilizing these varying spectral sensitivities, NWP systems can infer detailed atmospheric temperature and moisture profiles.

Since their introduction, satellite radiances have become a critical source of observational data in modern NWP, providing critical information in data-sparse regions such as the open oceans and the

Southern Hemisphere [Cardinali, 2009, Eyre et al., 2022]. Radiance measurements are particularly critical for TC prediction, given their development over the open ocean.

For these observations to be assimilated, they must be compared against a simulated first-guess model state projected in observation-space via a radiative transfer model, which means they are not direct observations of state variables within NWP models [Derber and Wu, 1998]. Operational centers have historically seen greater success in assimilating clear-sky radiances, while the assimilation of cloudy-sky measurements can be challenging due to the significant nonlinearities and microphysical uncertainties inherent in cloud-affected observations [Geer and Bauer, 2011b, McNally et al., 2000, Zhu et al., 2016]. Even for clear-sky radiances, the physical processes that govern these interactions remain extraordinarily complex. As a result, these radiative transfer operations require simplified physics and approximations that tend to introduce strong biases. Furthermore, biases can arise from the calibration and operation of the satellite instrument itself, as well as from errors in the model background state [Auligné et al., 2007]. Consequently, bias estimation and correction is a particularly important issue for these remotely-sensed observations [Dee and Uppala, 2009, Eyre, 1992].

1.4 Observation Bias Correction

Fundamental to the formulation of DA is that observations and backgrounds are unbiased estimates of the truth [Burgers et al., 1998]. However, in practice, this is rarely the case, and it is instead quite common for both the model and observations to experience systemic errors, or biases [Dee, 2005]. Untreated biases in the model and/or observations lead to suboptimal updates, and can systematically pull the analysis away from the truth and towards a biased solution [Lorenz, 2005].

As such, estimating and correcting observation bias is a top priority for NWP, but the task is

complicated by the difficulty of separating observation error from model error in the absence of a known truth. Instead, researchers and operational centers must rely on statistics gathered over many cycles of DA to separate and estimate individual bias contributions. A more thorough literature review of commonly used methodologies is provided in Chapter 2. For now, we focus on bias correction strategies specifically relevant to HAFS and TC modeling systems more generally.

Operational NOAA modeling systems such as GDAS employ variational bias correction (VarBC) for satellite radiances, in which bias correction coefficients are estimated as control variables within the DA cost function minimization, using atmospheric predictors such as layer thickness, skin temperature, and total column water vapor [Harris and Kelly, 2001, Zhu et al., 2014]. These methodologies can be very effective for dealing with complex state-dependent errors, but their success is complicated by the fact that they correct to a model background, which itself contains accumulated errors from many successive cycles, and may be strongly biased as a result. When model bias is incorrectly attributed to the observations, subsequent corrections can reinforce rather than remove the underlying error, leading to ”bias reinforcement”.

One strategy to separate observation bias from model bias is to exploit ”anchor” observations, such as surface station observations, radiosondes, dropsondes, and radio occultations [Kleist et al., 2023]. These highly accurate measurements with negligible systematic biases can serve as references for the true atmospheric state, and reduce the possibility of background bias contamination. The main limitation of this strategy is that anchor observations tend to be spatially- and temporally-sparse, and they can only help diagnose background errors where they are present. Regions far from anchors remain vulnerable to error contamination [Francis et al., 2023].

Limited-area models (LAMs) like HWRF and HAFS typically cannot perform online bias correction because they operate on restricted domains with partially-cycled DA, resulting in insufficient

sample sizes to estimate and update bias correction coefficients. Instead, these LAMs adopt bias coefficients trained from global models such as GDAS, which introduces external model bias into the regional system. Furthermore, bias coefficients trained on the global model may not be appropriate since GDAS is tuned for all-sky radiances, while HAFS currently uses clear-sky radiances only. Knisely and Poterjoy [2023] show that training the VarBC bias model on the native HAFS model produces more accurate estimates of bias, on top of reducing external biases which makes assessing internal model bias more feasible.

1.5 Motivation and Dissertation Overview

The overarching goal of this dissertation is to advance HAFS toward a fully mature, self-cycled, probabilistic TC system. This requires isolating HAFS from external model influence, improving radiance bias correction, and building a robust ensemble DA infrastructure. The development of a self-contained ensemble framework for analyses and reanalyses is essential both because it enables diagnosis of HAFS-specific model errors that are otherwise masked by global model influence and because it permits the development of an analysis of record for verification and machine learning applications.

Chapter 2 uses a low-dimensional chaotic dynamical system to examine limitations of observation bias correction in the presence of model bias. In order to identify fundamental sources of analysis error, we impose static global biases and employ bias correction methodologies that exploit low-bias anchor observations to separate observation and model biases. Although these experiments are idealized, we extend findings to highly complex measurement systems in high-dimensional NWP models

Chapter 3 conducts experiments with a basin-scale HAFS configuration focused on testing the

feasibility of bias correction coefficients trained online within the native modeling grid. In doing so, we demonstrate that HAFS can operate in a fully-cycled, self-contained mode with online radiance bias correction, removing one major source of global model influence as well as improving TC-error metrics.

Chapter 4 extends the self-cycled HAFS framework to a high-resolution ensemble analysis and forecast system. We discuss obstacles to basin-scale probabilistic forecasting while maintaining resolutions comparable to those in high-resolution convection-permitting storm-centric nests. We characterize major biases in the HAFS modeling system, and verify results against reanalyses, National Hurricane Center (NHC) best tracks, and forecasts from configurations that very closely resemble the operational HAFS. Altogether, we demonstrate the utility of a self-contained ensemble-based basin-scale HAFS configuration.

Chapter 5 summarizes the previous chapters, discusses ongoing research and future work, and provides broad conclusions and implications of this research.

Chapter 2: Limitations of Observation Bias Correction Investigated with a Low-Dimensional Dynamical Model

2.1 Introduction

Estimation and treatment of observation bias is a high priority for NWP, but problems remain significant [Auligné et al., 2007, Danforth and Kalnay, 2008, Dee, 2005, Dee and Da Silva, 1998, Eyre, 2016, Harris and Kelly, 2001, Zhu et al., 2014]. In the absence of a known truth, NWP centers estimate and remove bias using statistics gathered over repeated DA cycles, namely observation innovations (observations minus background) and analysis increments [Bell et al., 2004, Bhargava et al., 2018, Campbell et al., 2010, Chandramouli et al., 2022, Crawford et al., 2020, Zhu et al., 2014]. However, this strategy is complicated by the fact that model, observation, and background errors are difficult to separate. The background state contains both model error from the current cycle forecast as well as accumulated errors from previous cycles. Additionally, both model states and observations are subject to errors of representation arising from the discrete, finite-resolution characterization of a continuous atmospheric system. This makes it challenging to attribute bias definitively to either the observations or the model. In the absence of accurate attribution and separation, NWP models can experience a type of drift known as "bias reinforcement," in which background bias is interpreted as a bias in the observations or measurement operator, and subsequent corrections ultimately lead to a positive feedback and drift [Eyre, 1992].

One common strategy is to incorporate "anchor" observations, which are highly accurate measurements with negligible systematic biases that can serve as reference points for the true atmospheric state [Buehner et al., 2015, Eyre et al., 2022]. Examples include radiosondes, dropsondes, and GPS radio occultation [Aparicio and Laroche, 2015, Kleist et al., 2023]. While anchor observations

can be extremely effective for understanding local error profiles and verifying other observation systems, their sparse spatial and temporal coverage limits their ability to constrain domain-wide bias [Auligné et al., 2007].

The methods explored in this study are motivated by several foundational works on bias estimation and correction. Dee and Da Silva [1998] developed a sequential bias estimation and correction algorithm based on an unbiased subset of the observing system using a two-stage estimation approach, where bias parameters are estimated before state estimation. Eyre [1992] introduced an early static bias correction scheme for satellite brightness temperatures, separating bias into scan-dependent and air-mass-dependent components. Harris and Kelly [2001] refined this approach by replacing observed radiance predictors with model-derived quantities such as layer thickness, surface skin temperature, and total column water vapor. This change reframes the air-mass correction as a correction to the simulated radiances rather than to the observations themselves, since the forward model is the primary source of air-mass-dependent errors. Concurrently, Derber and Wu [1998] pioneered variational bias correction (VarBC), wherein bias correction coefficients are estimated as control variables within the cost function minimization, enabling adaptive online estimation of measurement biases. In Zhu et al. [2014], this VarBC scheme was updated by introducing additional predictors, adaptive background error variances for bias predictor coefficients, and by consolidating the bias correction into a single variational step within the Gridpoint Statistical Interpolation (GSI) data assimilation system.

For the current study, we consider Buehner et al. [2015], a model bias-aware observation correction system implemented with large-scale NWP in the Environment Canada Global Deterministic Prediction System. This system uses 4DEnVar-based DA for its primary analysis but includes an extra preliminary 3DVar analysis in which only low-bias anchor observations are assimilated. This extra analysis is then used as the reference background state for estimating bias correction coefficients for the primary

analysis [Buehner et al., 2015]. Buehner et al. [2015] note the importance of low-bias observing systems with extensive coverage, such as GPS radio occultation and radiosondes, for this scheme. This aspect is further emphasized in Eyre [2016] and Francis et al. [2023], which demonstrate that anchor observations work best when they observe the same state variables as bias-corrected observations, and that regions with sparse anchors are more susceptible to model bias contamination. We also find inspiration in Bowler et al. [2017] and Crawford et al. [2020], wherein the authors demonstrate the possibility of analysis correction-based additive inflation (ACAI) for the purpose of model bias correction for atmospheric models.

In this study, we examine how observation bias correction schemes can be contaminated by model process error, and how strategies that fail to account for this error accumulate biases over successive cycles. We explore how anchor observations can be exploited to verify high-resolution but strongly biased observation systems, such as satellite radiances, and to separate measurement biases from accumulated background bias. We demonstrate that the effectiveness of such methods is strongly limited by the sparsity and proximity of anchor observations to biased observations. We show that even among schemes that do consider model process error, it is nearly impossible to solve for bias explicitly in regimes unobserved by anchor observations. These strategies can be rectified if model bias is addressed directly with direct background state corrections, but we note that such corrections may be impractical for high-dimensional NWP models.

To isolate fundamental sources of analysis error, we impose constant, global biases across all state variables and observations, and use static linear methodologies to estimate and remove these imposed biases. While operational NWP centers employ far more sophisticated techniques, such as VarBC, for notably more complex error structures, the fundamental problem remains: how to accurately estimate observation bias in a modeling system with undiagnosed model process error.

In section 2.2, we introduce a basic equation for accumulated analysis error within a linear framework and demonstrate the theoretical limitations for observation bias correction in a modeling system with undiagnosed model process error. In section 2.3, we describe a series of bias correction methodologies tested with the Lorenz [2005] Model III, a 1-D chaotic dynamical model that is commonly used for idealized DA demonstrations. In section 2.4, we present experimental results, including RMSE and mean bias verifications as well as verifications of observation and model bias estimation. In section 2.5, we discuss conclusions and implications for further research.

2.2 Estimating Observation and Model Error

2.2.1 Definitions

Let $\mathbf{x}_i^t$ be a vector of state variables with length p that describes the true discretized atmospheric state and $\mathbf{y}_i$ be a vector of observations with length k, both valid at time i. $\mathbf{x}_i$ and $\mathbf{y}_i$ are given by:

$$\mathbf{x}_i^t = M(\mathbf{x}_{i-1}^t) + \eta_i \tag{2.1}$$

$$\mathbf{y}_i = H(\mathbf{x}_i^t) + \epsilon_i \tag{2.2}$$

where M is a model operator that pushes the state forward in time, H is a measurement operator that maps the model state onto observations, η_i is the model process error, and ϵ_i is the observation error. The probability density for these errors is typically modeled using normal distributions with zero mean:

$$\eta \sim N(0, \mathbf{Q}), \tag{2.3}$$

$$\epsilon \sim N(0, \mathbf{R}), \tag{2.4}$$

where $\mathbf{Q}$ and $\mathbf{R}$ are the model error and observation error covariances, respectively. Errors are generally assumed to be random with zero mean. However, this is not realistic for NWP, as errors often have strong systemic errors, or biases. Instead, we can represent model and observation biases by sampling errors from Gaussian distributions with non-zero means,

$$\eta \sim N(\overline{\eta}, \mathbf{Q}), \tag{2.5}$$

$$\epsilon \sim N(\overline{\epsilon}, \mathbf{R}), \tag{2.6}$$

where $\overline{\eta} = \langle\eta\rangle$and $\overline{\epsilon} = \langle\epsilon\rangle$ are the mean model and observation error (or bias), respectively. During DA, the forecast, or background state, $\mathbf{x}_i^b$ is updated sequentially to reflect observations collected over a time window to produce an analysis, denoted $\mathbf{x}_i^a$. For a linear model and observation process, an equation for the analysis is given perfectly by the Kalman filter:

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \mathbf{K}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^b), \tag{2.7}$$

$$\mathbf{dx}_i = \mathbf{K}\mathbf{dy}_i, \tag{2.8}$$

where $\mathbf{K}$ is the gain matrix, $\mathbf{H}$ is the linear measurement operator in matrix form, and we define the analysis increment $\mathbf{dx}_i$ and observation innovation $\mathbf{dy}_i$ as:

$$\mathbf{dx}_i = \mathbf{x}_i^a - \mathbf{x}_i^b, \tag{2.9}$$

$$\mathbf{dy}_i = \mathbf{y}_i - \mathbf{H}\mathbf{x}_i^b. \tag{2.10}$$

We can further define analysis error and background error as

$$\mathbf{e}_i^a = \mathbf{x}_i^a - \mathbf{x}_i^t, \tag{2.11}$$

$$\mathbf{e}_i^b = \mathbf{x}_i^b - \mathbf{x}_i^t. \tag{2.12}$$

We similarly note that analysis and background bias can be defined as $\overline{\mathbf{e}_a} = \langle \mathbf{e}_a \rangle$ and $\overline{\mathbf{e}_b} = \langle \mathbf{e}_b \rangle$.

2.2.2 Accumulated Bias

We adopt a simplified version of the analysis equation similar to that of Eyre [2016] where the $p \times p$ matrix $\mathbf{W} = \mathbf{I} - \mathbf{KH}$ designates the weight assigned to the background during analysis, and $\mathbf{K}$ (dimensions $p \times k$) weights the observation innovation. In the Kalman filter, $\mathbf{K}$ is a function of both background and observation error covariances, such that high confidence in the background yields a small $\mathbf{K}$ and therefore a large $\mathbf{W}$, meaning the observation innovation is minimal. In the presence of model bias, a large $\mathbf{W}$ means the analysis will inherit substantial background bias each cycle. $\mathbf{K}$ can only partially correct this bias through observations, and if those observations are themselves biased, each successive cycle compounds both sources of error.

To isolate the fundamental sources of analysis error, we assume constant observation coverage and static background error covariance, such that these weight terms are constant across cycles. As such, the first cycle background and analysis states are given by:

$$\mathbf{x}_1^b = \mathbf{x}_1^t + \mathbf{e}_1^b, \tag{2.13}$$

$$\begin{aligned} \mathbf{x}_1^a &= \mathbf{W}\mathbf{x}_1^b + \mathbf{K}\mathbf{y}_1 \\ &= \mathbf{W}(\mathbf{x}_1^t + \mathbf{e}_1^b) + \mathbf{K}(\mathbf{H}\mathbf{x}_1^t + \epsilon_1). \end{aligned} \tag{2.14}$$

Using our previous definition of background and analysis error and eliminating $\mathbf{x}^t$ terms leaves us with an equation for analysis error:

$$\mathbf{e}_1^a = \mathbf{W}\mathbf{e}_1^b + \mathbf{K}\epsilon_1. \tag{2.15}$$

The subsequent cycle background and analysis error is given by

$$\begin{aligned}
\mathbf{e}_2^b &= \mathbf{M}_2\mathbf{e}_1^a + \eta_2 \\
&= \mathbf{M}_2\mathbf{W}\mathbf{e}_1^b + \mathbf{M}_2\mathbf{K}\epsilon_1 + \eta_2, \qquad (2.16) \\
\mathbf{e}_2^a &= \mathbf{W}\mathbf{e}_2^b + \mathbf{K}\epsilon_2 \\
&= \mathbf{W}(\mathbf{M}_2\mathbf{W}\mathbf{e}_1^b + \mathbf{M}_2\mathbf{K}\epsilon_1 + \eta_2) + \mathbf{K}\epsilon_2 \\
&= \mathbf{W}\mathbf{M}_2\mathbf{W}\mathbf{e}_1^b + \mathbf{W}\eta_2 + \mathbf{W}\mathbf{M}_2\mathbf{K}\epsilon_1 + \mathbf{K}\epsilon_2. \qquad (2.17)
\end{aligned}$$

where $\mathbf{M}_i$ is the linear model operator in matrix form. Iterating this process forward in time to the ith time step results in the following analysis error:

$$\begin{aligned}
\mathbf{e}_i^a = &\ \mathbf{W}\mathbf{M}_i \cdots \mathbf{W}\mathbf{M}_2\mathbf{W}\mathbf{e}_1^b \\
&+ \sum_{m=2}^{i-1} \mathbf{W}\mathbf{M}_i \cdots \mathbf{W}\mathbf{M}_{m+1}\mathbf{W}\eta_m + \mathbf{W}\eta_i \qquad (2.18) \\
&+ \sum_{m=1}^{i-1} \mathbf{W}\mathbf{M}_i \cdots \mathbf{W}\mathbf{M}_{m+1}\mathbf{K}\epsilon_m + \mathbf{K}\epsilon_i,
\end{aligned}$$

which we represent with a more compact notation using ordered operator products:

$$\begin{aligned}\mathbf{e}_i^a = &\left(\prod_{j=0}^{i-2}\mathbf{W}\mathbf{M}_{i-j}\right)\mathbf{W}\mathbf{e}_1^b \\ &+\sum_{m=2}^{i-1}\left(\prod_{j=0}^{i-m-1}\mathbf{W}\mathbf{M}_{i-j}\right)\mathbf{W}\eta_m + \mathbf{W}\eta_i \\ &+\sum_{m=1}^{i-1}\left(\prod_{j=0}^{i-m-1}\mathbf{W}\mathbf{M}_{i-j}\right)\mathbf{K}\epsilon_m + \mathbf{K}\epsilon_i.\end{aligned} \tag{2.19}$$

The first term describes IC error, starting from the initial background state, the second term describes accumulated model process error from successive model integrations, and the final line describes accumulated error from assimilating all past observations up to the current set. We note that while $\mathbf{H}$ does not appear in the final form, it is implied in $\mathbf{W} = \mathbf{I} - \mathbf{K}\mathbf{H}$.

This equation describes the accumulated error up to the present time, but we can simplify it further by assuming that DA has progressed over a sufficiently large number of cycles so that the first term is much smaller than the second two, and take the expectation:

$$\begin{aligned}\langle\mathbf{e}_i^a\rangle = &\sum_{m=2}^{i-1}\left(\prod_{j=0}^{i-m-1}\mathbf{W}\mathbf{M}_{i-j}\right)\mathbf{W}\overline{\eta} + \mathbf{W}\overline{\eta} \\ &+\sum_{m=1}^{i-1}\left(\prod_{j=0}^{i-m-1}\mathbf{W}\mathbf{M}_{i-j}\right)\mathbf{K}\overline{\epsilon} + \mathbf{K}\overline{\epsilon}.\end{aligned} \tag{2.20}$$

We assert that, for a stable DA system, terms from cycles in the distant past contribute negligibly to the current analysis error. We do so by examining the product term, $\sum_{m=2}^{i-1}\left(\prod_{j=0}^{i-m-1}\mathbf{W}\mathbf{M}_{i-j}\right)\mathbf{W}\overline{\eta}$, and noticing that $\mathbf{W}\mathbf{M}_j$ is applied repeatedly. $\mathbf{W}$ is contractive, and while the forecast model $\mathbf{M}$ may amplify perturbations, the analysis step applies a correction that pulls the state towards observations. Assuming we update frequently enough, the net effect must be contractive for the DA system to remain

stable, or else perturbations would amplify and cause the system to diverge. Therefore, a functioning DA system requires that this product decays over many time steps and our error equation converge to a finite steady-state, and likewise for the product term corresponding to observation error. Equivalently, we note that as the length of our time step $\Delta t \to 0$, the linearized model operator $\mathbf{M} \to \mathbf{I}$. In other words, frequent updates prevent the model dynamics from substantially evolving perturbations between cycles, ensuring the DA system remains stable and past errors decay rapidly.

Under this assumption, we can write:

$$\langle \mathbf{e}_i^a \rangle \approx \widehat{\mathbf{W}}\overline{\eta} + \widehat{\mathbf{K}}\overline{\epsilon}, \tag{2.21}$$

where $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{K}}$ represent the converged sums of matrix products weighting the contributions of model and observation bias, respectively. These converged sums would likely be dominated by the most recent updates, i.e.

$$\begin{aligned} \widehat{\mathbf{K}} &= \cdots + \mathbf{W}\mathbf{M}_i\mathbf{K} + \mathbf{K} \\ &\approx \mathbf{W}\mathbf{M}_i\mathbf{K} + \mathbf{K}, \end{aligned} \tag{2.22}$$

$$\begin{aligned} \widehat{\mathbf{W}} &= \cdots + \mathbf{W}\mathbf{M}_i\mathbf{W} + \mathbf{W} \\ &\approx \mathbf{W}\mathbf{M}_i\mathbf{W} + \mathbf{W}. \end{aligned} \tag{2.23}$$

This derivation requires a fair number of assumptions, namely that observation sparsity and bias remain relatively stable over time and that model bias is likewise consistent in character from cycle to cycle. Furthermore, this derivation and its assumptions would break down under a case with state-dependent biases and possibly correlated observation and background errors. Nevertheless, we

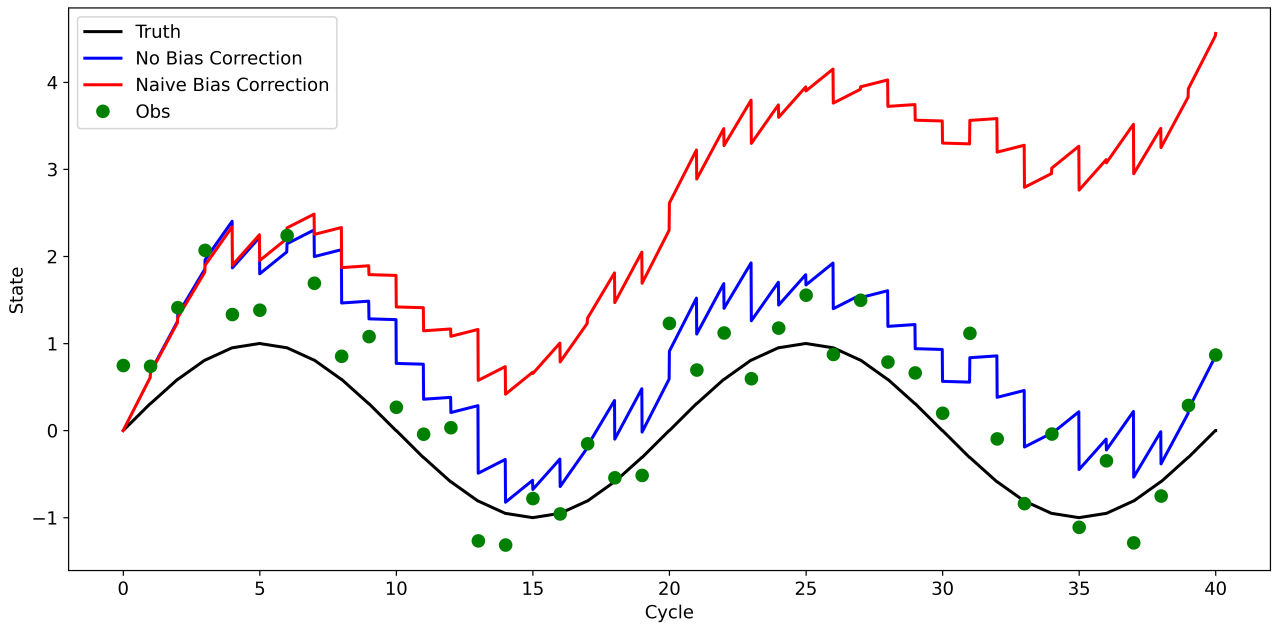


Figure 2.1: Straightforward demonstration of bias reinforcement in bias correction schemes that do not consider the effects of model bias. In blue, the biased model trajectory is regularly updated with observations which are themselves biased. For the "Naive" bias correction, observations are corrected to a model background. Since the background itself is biased, the bias correction reinforces bias and leads to model drift, shown in red.

will use it as a reference when interpreting findings from DA experiments performed with a nonlinear model.

One intuitive conclusion that follows from equation 2.21 is that bias accumulation depends greatly on the location and type of observations. This is because $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{K}}$ are functions of $\mathbf{K}$, which depends on observation coverage and error characteristics. If observations were to become more sparse over time or during specific cycles, $\mathbf{K}$ would decrease, causing $\widehat{\mathbf{W}}$ to increase in magnitude and the background bias term $\widehat{\mathbf{W}}\overline{\eta}$ to grow accordingly (and vice versa).

2.2.3 Bias Estimation

Next, we examine how model and observation bias can be estimated from statistics collected over many cycles of DA, and how we can exploit highly accurate but sparse "anchor" observations

to estimate bias in a different set of observations. Many bias correction systems implicitly use the background as a truth state for correcting observations. Suppose we take a time-average of observation innovations over a long window and use it to correct $\mathbf{Hx}$:

$$\mathbf{x}_i^a = \mathbf{x}_i^b + \mathbf{K}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^b - \overline{(\mathbf{y} - \mathbf{H}\mathbf{x}^b)}), \tag{2.24}$$

where

$$\overline{(\mathbf{y} - \mathbf{H}\mathbf{x}^b)} = \overline{\epsilon} - \mathbf{H}\overline{\mathbf{e}^b}. \tag{2.25}$$

This is a rather naive approach for correcting observations, given that we are making no attempt to account for background bias in our estimate of observation bias. If $\langle \mathbf{e}_i^a \rangle$ is the expected analysis error after performing the observation bias correction, then equation 2.21 would become:

$$\begin{aligned} \langle \mathbf{e}_i^a \rangle &= \widehat{\mathbf{W}}\overline{\eta} + \widehat{\mathbf{K}}\overline{\epsilon} - \mathbf{K}\left(\overline{\epsilon} - \mathbf{H}\overline{\mathbf{e}_b}\right) \\ &= \widehat{\mathbf{W}}\overline{\eta} + (\widehat{\mathbf{K}} - \mathbf{K})\overline{\epsilon} + \mathbf{K}\mathbf{H}\overline{\mathbf{e}_b} \\ &\approx \widehat{\mathbf{W}}\overline{\eta} + \mathbf{K}\mathbf{H}\overline{\mathbf{e}_b}. \end{aligned} \tag{2.26}$$

Taking $\widehat{\mathbf{K}} \approx \mathbf{K}$ (which itself is not a particularly reasonable approximation), this method successfully minimizes observation error but augments the error with an extra background bias term, which may ultimately cause the analysis error to grow larger than the background error, as shown in Fig. 2.1.

Many systems exploit "anchor" observations, which are highly accurate but sparse. Introducing these observations changes our converged-sum solution. First, we split our observations into two sets

of observation type, standard observations $\mathbf{y}$ and anchor observations $\mathbf{y}^*$, so that our new expected analysis error is

$$\begin{aligned} \langle \mathbf{e}_i^a \rangle &\approx \widehat{\mathbf{W}}\overline{\eta} + \mathbf{W}^*\overline{\eta} + \widehat{\mathbf{K}}\overline{\epsilon} + \mathbf{K}^*\overline{\epsilon^*} \\ &= \widehat{\mathbf{W}}\overline{\eta} + \mathbf{W}^*\overline{\eta} + \widehat{\mathbf{K}}\overline{\epsilon}, \end{aligned} \tag{2.27}$$

where anchor observation bias near zero such that $\overline{\epsilon^*} \approx 0$. Note that we introduced new converged weight terms for the standard and anchor updates, and that $\widehat{\mathbf{W}} + \widehat{\mathbf{K}} = \mathbf{I}$ and $\mathbf{W}^* + \mathbf{K}^* = \mathbf{I}$.

2.2.4 Linear Bias Correction

Following Buehner et al. [2015], we separate our update into two steps according to observation type, and use the first analysis increment to adjust observation bias correction parameters. We do so by first assimilating anchor observations alone, over many cycles, such that our preliminary analysis error would be

$$\langle \mathbf{e}_i^{a*} \rangle = \mathbf{W}^*\overline{\eta}, \tag{2.28}$$

where $\langle \mathbf{e}_i^{a*} \rangle$ is the expected analysis error of this preliminary update. In a perfect world, we would have near-complete anchor observation coverage across the entire domain. In this hypothetical scenario, we could perfectly fit each model grid point to these zero-bias observations, such that $\mathbf{K}^* \to 1$ and $\mathbf{W}^* \to 0$ and the expected analysis error of this preliminary update would be zero:

$$\langle \mathbf{e}_i^{a*} \rangle \approx 0. \tag{2.29}$$

We could then take the expectation of this analysis increment, i.e. the difference between these analyses and backgrounds over many cycles of this preliminary analysis, such that we have a perfect estimate of accumulated background bias:

$$\begin{aligned} \langle \mathbf{x}^{a*} - \mathbf{x}^{b} \rangle &= \overline{\mathbf{e}^{a*} - \mathbf{e}^{b}} \\ &= -\overline{\mathbf{e}}^{b}. \end{aligned} \tag{2.30}$$

Thus, we are able to perfectly estimate the effect of background bias from equation 2.26. We can use this estimate of model bias to augment our observation bias correction and remove the extra model bias term that accumulates through time:

$$\begin{aligned} \overline{(\mathbf{y} - \mathbf{H}\mathbf{x}^{b})} - \mathbf{H}\langle \mathbf{x}^{a*} - \mathbf{x}^{b} \rangle &= \overline{\epsilon} - \mathbf{H}\overline{\mathbf{e}_{b}} + \mathbf{H}\overline{\mathbf{e}_{b}} \\ &= \overline{\epsilon}. \end{aligned} \tag{2.31}$$

In other words, given a comprehensive enough coverage of anchoring obs, we could perfectly estimate accumulated background bias across the domain and remove it from observation bias estimates, giving us a perfect correction for the biased obs.

In practice, however, there are two major problems with this approach. First, fitting state variables exactly to observations threatens to violate the basic principle that analyses should optimally weight both observations and background according to their uncertainties. All observations contain measurement error, even anchor observations, and treating them as perfect introduces noise into the analysis. Even if they were exactly unbiased, errors of representation mean that observations cannot be treated as a perfect reference for the model state. More importantly, the background satisfies basic dynamical balances

(e.g. geostrophic, hydrostatic, etc. for NWP models), while observations sample discrete points and may introduce spatial discontinuities which disrupt these balances, though appropriate background error covariance estimation should mitigate this to some degree. Furthermore, for state variables not directly observed, the analysis relies on background error correlations to spread observation information. By inflating background errors to fit exactly to observations, we ignore these correlations and leave unobserved variables uncorrected.

Realistically, anchor observation coverage tends to be limited spatially and temporally. While it may be possible to remove model bias explicitly for state variables in close proximity to anchoring observations, this does not extend to state variables far from anchor observations. This relationship is explicitly represented in our framework through the Kalman gain matrix $\mathbf{K}^*$. For state-variables directly observed by anchor observations, $\mathbf{K}^*$ is large and model bias can be effectively removed, but for unobserved state variables, $\mathbf{K}^*$ is small or zero, leaving the background unaffected. Even in the case of strong error correlations in $\mathbf{K}^*$, some level of background error will still contaminate the estimate [Francis et al., 2023]. While it may be possible to extrapolate model bias characteristics over unobserved areas that share similar dynamical regimes (e.g. tropical Atlantic vs. mid-latitude Atlantic), such extrapolation is simplistic and ad-hoc at best, and highly susceptible to errors-of-the-day at worst.

Instead, with a more realistic approach with sparse anchor observations coverage, the preliminary analysis error would be given by equation 2.28 and the expected analysis increment would be given by

$$\begin{aligned} \langle \mathbf{x}^{a*} - \mathbf{x}^b \rangle &= \overline{\mathbf{e}^{a*} - \mathbf{e}^b} \\ &= \mathbf{W}^* \overline{\eta} - \overline{\mathbf{e}}^b, \end{aligned} \tag{2.32}$$

and the correction innovation for standard observations

$$\begin{aligned}\overline{(\mathbf{y}-\mathbf{H}\mathbf{x}^b)}-\mathbf{H}^*(\overline{\mathbf{e}^{a*}-\mathbf{e}^b}) &= \overline{\epsilon}-\mathbf{H}\overline{\mathbf{e}^b}-\mathbf{H}^*\mathbf{W}^*\overline{\eta}+\mathbf{H}^*\overline{\mathbf{e}^b}\\ &\approx \overline{\epsilon}-\mathbf{H}^*\mathbf{W}^*\overline{\eta}. \end{aligned} \tag{2.33}$$

Since this model bias term is modulated by $\mathbf{W}^*$, we can limit this residual by increasing the frequency and/or reducing the error variance of our anchor observations, but it will always remain at some level for state variables far from anchor observations. On the other hand, if the model bias is significantly large, or our anchor observation network is sufficiently sparse, such that $\mathbf{W}^*$ is weighted heavily, then this methodology may actually enhance analysis error and consequently lead to bias reinforcement.

To demonstrate this more explicitly, we can derive the expected analysis increment from the preliminary anchor update. If we assume the same assimilation schedule as above (i.e. a preliminary update with anchor observations only followed by a standard update with biased observations), we can adopt equations 2.21 and 2.23, noting that the degree to which we truncate these sums is irrelevant, such that the expectations for analysis and background error are given by:

$$\langle \mathbf{e}_i^{b*} \rangle \approx \mathbf{M}_i\mathbf{W}^*\overline{\eta}+\overline{\eta}, \tag{2.34}$$

$$\langle \mathbf{e}_i^{a*} \rangle \approx \mathbf{W}^*\mathbf{M}_i\mathbf{W}^*\overline{\eta}+\mathbf{W}^*\overline{\eta}. \tag{2.35}$$

From here, our expected analysis increment is

$$\begin{aligned}
\langle \delta \mathbf{x}_i^* \rangle &= \langle \mathbf{e}_i^{a*} \rangle - \langle \mathbf{e}_i^{b*} \rangle \\
&= (\mathbf{W}^* - \mathbf{I})(\mathbf{M}_i \mathbf{W}^* + \mathbf{I})\overline{\eta} \\
&= -\mathbf{K}^* \mathbf{H}^* (\mathbf{M}_i \mathbf{W}^* + \mathbf{I})\overline{\eta} \\
&= -\mathbf{K}^* \mathbf{H}^* (\mathbf{M}_i (\mathbf{I} - \mathbf{K}^* \mathbf{H}^*) + \mathbf{I})\overline{\eta}. \qquad (2.36)
\end{aligned}$$

This expression shows more cleanly that the expected analysis increment is essentially the accumulated bias over DA cycles, regressed from observation space onto the full state. The analysis increment is filtered through $\mathbf{K}^*\mathbf{H}^*$, which projects accumulated model bias from anchor observations back onto the state. When anchor observations are sparse, $\mathbf{H}^*$ captures only a partial view of the full bias. Likewise, when $\mathbf{K}^*$ contains weak error covariances, bias information will not propagate effectively to unobserved regions.

2.2.5 Model Bias Correction

Lastly, we implement a strategy for bias-correcting the background state itself before DA. Following Bowler et al. [2017] and Crawford et al. [2020], this methodology assumes that time-averaged analysis increments provide a suitable approximation to the model bias. With this assumption, we add $\langle \delta \mathbf{x}_i^* \rangle$ to the model background before performing DA and skip any accounting of model bias during the subsequent steps for estimating observation bias. Combined with a preliminary update with anchoring observations only, which, represent a close albeit noisy approximation of the truth, we can potentially mitigate major issues related to bias accumulation and model bias contamination in

observation bias estimates.

In this case, the model bias estimate would resemble equation 2.32. After performing a model state correction and assuming that the mean of analysis increments approximates the model bias, we arrive at the equation for expected analysis error

$$\begin{aligned} \mathbf{e}_i^a &= \mathbf{e}_i^b + (\mathbf{W}^*\overline{\eta} - \overline{\mathbf{e}}^b) + \mathbf{K}\left(\epsilon_i - \mathbf{H}(\mathbf{e}_i^b + \mathbf{W}^*\overline{\eta} - \overline{\mathbf{e}}^b)\right) \\ &\approx \mathbf{W}^*\overline{\eta} + \mathbf{K}\left(\epsilon_i - \mathbf{H}\mathbf{W}^*\overline{\eta}\right) \qquad (2.37) \\ &= \mathbf{W}\mathbf{W}^*\overline{\eta} + \mathbf{K}\epsilon_i. \qquad (2.38) \end{aligned}$$

The final expression shows that the residual model bias in the analysis is now modulated by $\mathbf{W}\mathbf{W}^*$, which is the product of the background weights for both the standard and anchor updates.

While some portions of model bias still remain, the key advantage to this method is that it reduces the effects of model bias directly during state estimation, instead of for observation bias estimation alone. The primary drawback of such a method is that it assumes a time-constant model bias, which may be inappropriate for circumstances when the bias exhibits strong state or regime dependence. This approach is particularly risky for high-resolution applications such as tropical cyclone forecasting, where time-mean increments may fail to capture the dynamical balances inherent to small-scale features like vortex structure—as these storms are translated by the mean flow. Furthermore, like previously discussed methodologies, the correction is limited by anchor observation coverage, leaving model bias largely uncorrected in poorly observed regions.

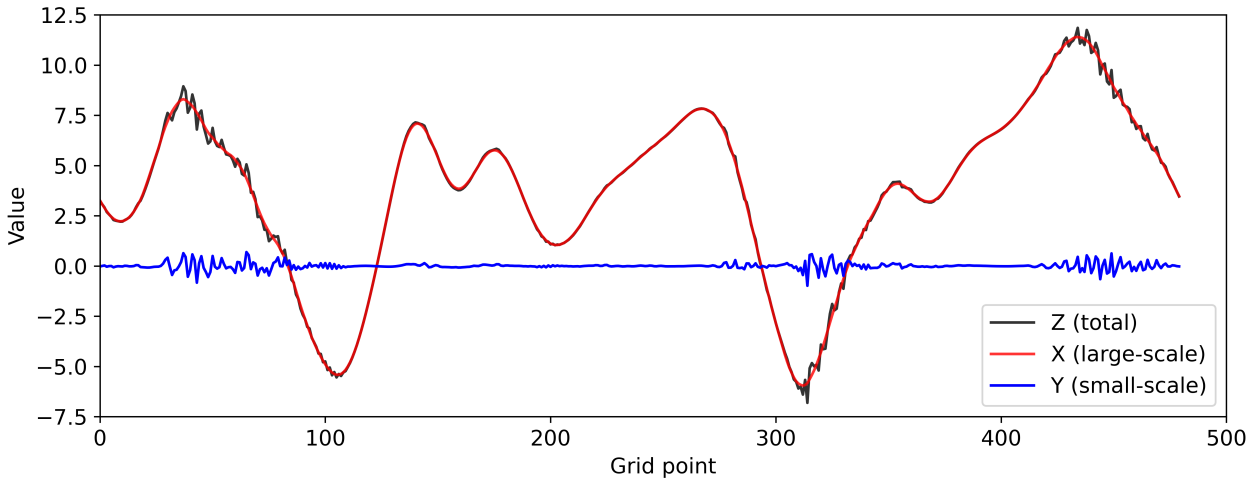


Figure 2.2: Lorenz 2005 Model III sample solution state. The truth state, Z, can be decomposed into a large-scale, X, and a small-scale, Y. Small-scale activity is driven by large-scale activity through a coupling term, such that regions where Y is active tend to follow local maximums/minimums of X.

2.3 Methodology

2.3.1 Model Description and Bias Simulation

We explore new methodologies using Model III of Lorenz (2005), an idealized one-dimensional chaotic dynamical model that simulates a hypothetical scalar quantity at one latitude and height level, and has a solution that resembles smooth planetary waves with small-scale activity superposed (see Fig. 2.2). The system features a set of coupled variables X and Y, both defined on the same ring of N grid points, which represent large-scale and small-scale waves, respectively. These variables are combined to give a single set, Z, the governing equation for which is given by:

$$dZ_n/dt = [X, X]_{K,n} + b^2[Y, Y]_{1,n} + c[Y, X]_{1,n} - X_n - bY_n + F, \tag{2.39}$$

where n is any given point on the ring, b is a damping parameter, c is a coupling parameter, F is a forcing parameter and K is a parameter that controls correlation length for X. The terms in brackets are inner products as defined in Lorenz [2005]. For all parameters, we use the default "perfect" parameter values, namely $N = 480$, $K = 32$, $F = 15$, $b = 10$, and $c = 2.5$, in addition to a final parameter $I = 12$ which sets the scale ratio between the large-scale and small-scale components.

We integrate the model using an explicit Runge-Kutta (4,5) formula with a unitless time step of 0.05, equivalent to approximately 6 hours of atmospheric error growth. After discarding the first 200 cycles (approximately 50 days) as spin-up, the subsequent T=600 cycles serve as the 'truth' state for data assimilation experiments.

We produce standard biased observations and unbiased anchoring observations by passing the truth state through measurement operators and adding noise sampled from a Gaussian with a standard deviation of 0.5 and 0.1, respectively. Depending on the experiment, we vary observation distribution and impose static, global observation biases, ranging from -2 to 2. Likewise, we impose static, global model bias on the forecast background before data assimilation, ranging from -0.5 to 0.5.

Across all methodologies, we begin each experiment with a 200-cycle spin-up period in which only unbiased anchoring observations are assimilated across the domain. After the spin-up period, we introduce biased "standard observations", which are then bias corrected based on the strategies discussed below

For comparison, we perform two control experiments. The first is a "no bias correction" experiment, in which all observations are assimilated without any bias correction at all. The second is a "naive" bias correction methodology, in which observations are corrected to the model background with no consideration of the effects of model bias. We also consider the spin-up period, wherein only anchor observations are assimilated, to be a control period.

2.3.2 Data Assimilation Configuration

We assimilate measurements using the Ensemble Square-Root Filter (EnSRF) described by Whitaker and Hamill [2002]. The ensemble size was selected to be 350 members, as filtering skill measured by RMSE and spread/skill score exhibited only marginal improvements beyond this number. This ensemble size is larger than typically feasible in operational NWP, and therefore represents a "best case" DA configuration.

We use relaxation to prior spread (RTPS) inflation with a factor ranging from 0.3 to 0.9. We also apply Gaussian localization with a standard deviation ranging from 0.78 to 0.1, where these quantities represent distance as a fraction of the total number of grid points in the period domain. These parameters are tuned for each configuration of model bias and observation frequency to achieve an optimal spread-skill relationship.

2.3.3 Linear Bias Correction Methodologies

Next we discuss the application of the linear additive bias correction methodologies described in theory in section 2.2. To minimize the influence of accumulated background and observation error, each approach splits the analysis into two steps: the first uses anchor alone and informs how to bias-correct the standard observations, while the second step assimilates the network of standard observations. Since this preliminary update assimilates unbiased anchoring observations alone, it can be assumed that the resulting analysis has a much smaller expected analysis error stemming from accumulated model bias than what you would find in the background state. From here, we examine several different options for bias correction strategy:

2.3.3.1 Preliminary observation-space update with inflated background errors

In addition to the preliminary state-space update with anchor observations mentioned above, we can perform an additional preliminary update in standard observation-space with highly inflated background errors. We use anchor observations projected into standard observation-space and inflate background errors such that $\mathbf{K}^* \rightarrow \mathbf{I}$, effectively fitting the analysis exactly to anchor observations at standard observation locations. This produces a highly accurate estimate of background bias in standard observation-space, which we then use to correct standard (biased) observations for the full state-space update.

We compute the model bias estimate as the difference between this analysis with inflated errors and the forecast background, both in standard observation-space. This estimate is then used to adjust our observation bias correction for the standard observations. The inflated-error analysis itself is discarded and not used to update the model state. Instead, a separate preliminary update without inflation provides the background for the final state-space update with (now corrected) standard observations. We note that this approach assumes a linear or quasi-linear measurement operator, and that highly nonlinear operators would require the extra preliminary update to be performed in state-space.

2.3.3.2 Scalar vs. spatially-varying bias correction

As acknowledged in subsection 2.2.4, analysis increments can be effective for isolating and estimating model bias, but only for variables that exhibit large background error covariance with observed quantities. For unobserved state variables, such as those far from anchoring obs, analysis increments provide little to no reduction in model bias, and observation bias estimation strategies over these areas are certain to suffer from bias contamination, assuming imperfect knowledge of

background and observation error covariances. If we can assume similar bias across the entire domain, then we can take the mean of model bias estimations over state variables that are observed by anchoring obs, and determine a global model bias estimate. While such a scenario is highly idealized, it is nevertheless worthwhile to examine within our experiments, as the prescribed bias is indeed spatially homogeneous. Therefore, we compare experiments in which analysis increments are averaged over observed state variables vs. using a spatially-varying analysis vector to estimate and minimize model bias contamination from our estimate of observation bias.

2.3.3.3 Direct Model Bias Correction

Following subsection 2.2.5, it is possible to directly estimate and remove bias from the background state before DA. In this case, the background is corrected before assimilating anchor observations, where increments from the original background provide a spatially heterogeneous estimate for model bias. From here, biased observations are corrected against the preliminary analysis, which theoretically should feature limited accumulated background bias as a result of the correction. Much like the previous strategy, we have the option to correct the background based on a mean estimate over observed state variables vs. using a spatially-varying bias estimate. The former strategy can be considered a ”best case” scenario, given our highly idealized static global model bias, while the latter is limited by factors such as the density and type of anchor observations.

While this method is an interesting test case as the absolutely ”best case” scenario, it may be difficult to implement in practice, since the spatial structure of bias will depend on factors such as the degree to which observation networks change over time. For this reason, we include limited results, and instead prioritize methodologies that seek to limit the influence of model bias on observation bias

estimation alone.

2.4 Results

2.4.1 Observation Bias Correction Results

In this section, we quantify the skill of the four observation correction methodologies described in section 2.3.3 against the truth. The methods are also compared to three controls: (1) experiments with no bias correction, (2) experiments with naive bias correction, and (3) experiments with anchor observations only. The four observation correction methodologies will be identified with the following naming conventions:

1. Scalar Corr.: analysis increments are averaged over observed state variables and a scalar model bias correction is applied to observation estimate.

2. Scalar Corr. w/ infl.: Extra preliminary update is performed with inflated background errors. Analysis increments from this update are averaged over observed state variables and a scalar model bias correction is applied to observation estimate.

3. Vector Corr.: analysis increments are kept as spatially-varying vector for model bias correction to observation estimate.

4. Vector Corr. w/ infl: Extra preliminary update is performed with inflated background errors. Analysis increments from this update are kept as spatially-varying vector for model bias correction to observation estimate.

In each case, the first 200 cycles serve as a spin-up in which each experiment assimilates anchor

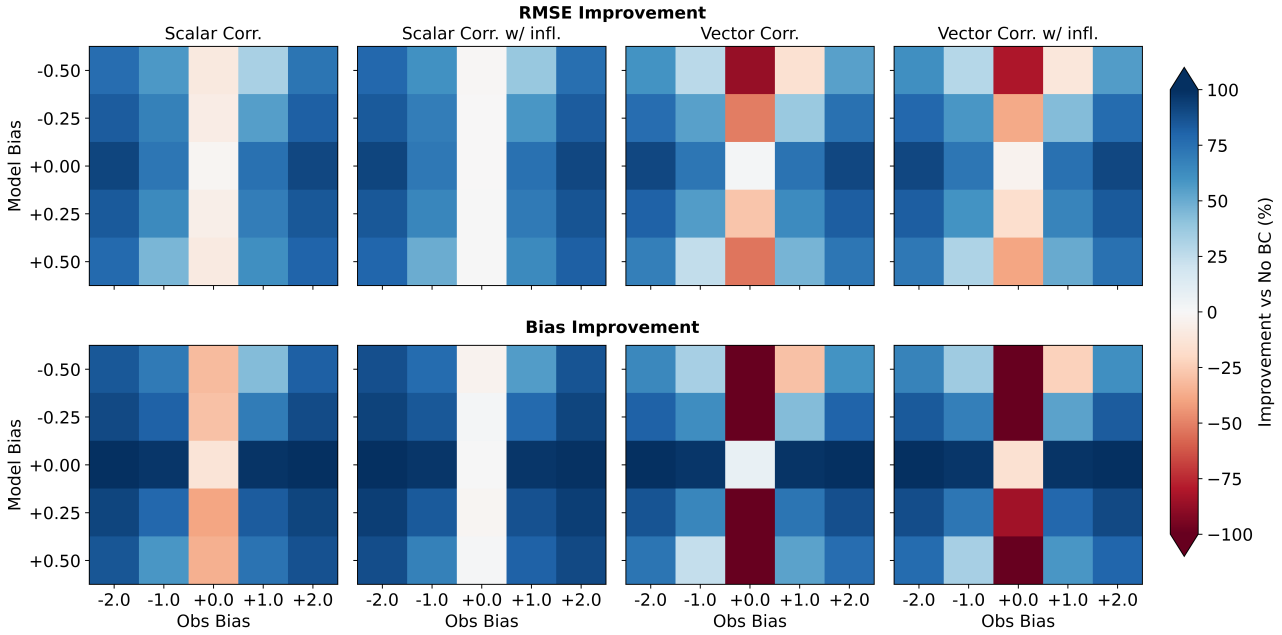


Figure 2.3: Percent improvement in RMSE (top row) and absolute bias (bottom row) for four bias correction methods relative to experiments with no bias correction. From left to right, columns show: scalar mean correction, scalar mean correction with inflation, vector correction, and vector correction with inflation. Each heatmap displays results averaged over the final 100 assimilation cycles across 25 bias configurations, with rows indicating imposed model bias and columns indicating imposed observation bias. Blue shading indicates improvement over control, while red shading indicates degradation. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively.

observations only. Afterwards, each method uses a moving window of data from the previous 100 cycles to estimate model and observation bias.

First, we examine how these correction strategies fare against a variety of imposed bias configurations for a fixed standard and anchor observation frequency and fixed control. In Fig. 2.3, we verify against experiments with no bias correction for a network comprised of standard and anchor observations at spatial frequencies of 1 and 8, respectively. Across all methodologies, we notice degraded results when observation bias is zero and increasingly improved results as imposed observation bias increases. In particular, the vector correction experiments degrade RMSE and bias quite substantially, although it should be noted that RMSE and bias for the control experiments would

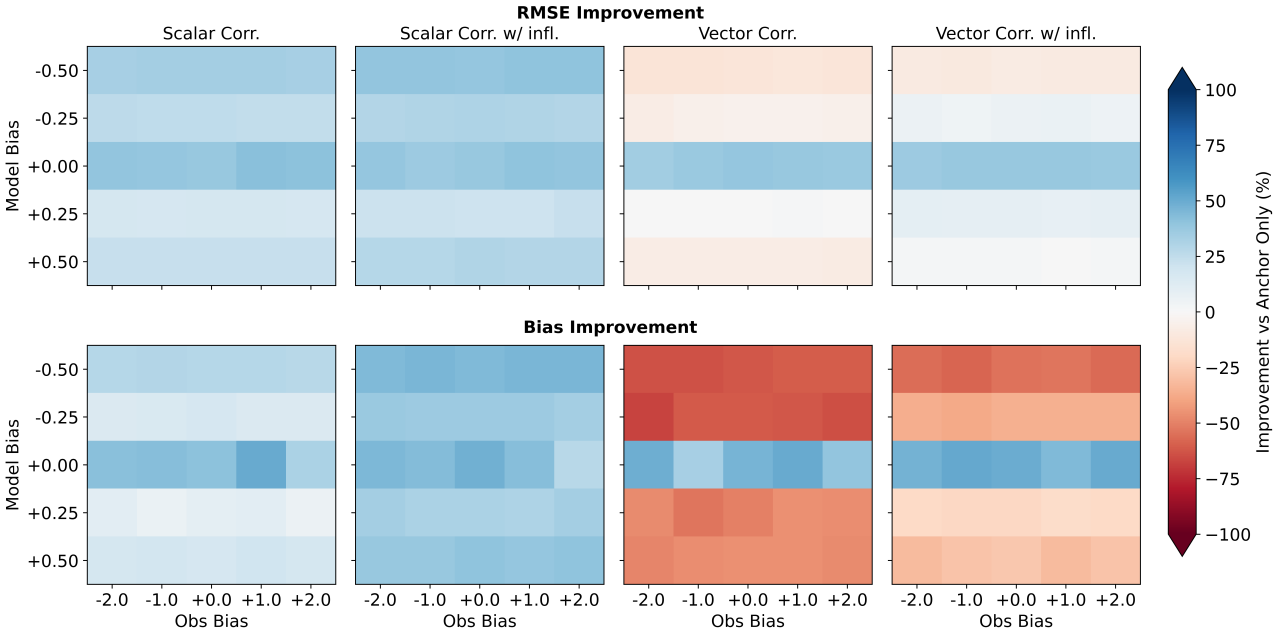


Figure 2.4: Percent improvement in RMSE (top row) and absolute bias (bottom row) for four bias correction methods relative cycles with anchor observations only. From left to right, columns show: scalar mean correction, scalar mean correction with inflation, vector correction, and vector correction with inflation. Each heatmap displays results averaged over the final 100 assimilation cycles across 25 bias configurations, with rows indicating imposed model bias and columns indicating imposed observation bias. Blue shading indicates improvement over control, while red shading indicates degradation. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively.

already be quite low for this network. Furthermore, we note that all methods exhibit larger errors as model bias increases. This result is expected, since all methodologies leave some residual model bias, and thus their performance degrades as model bias and consequently model bias contamination increase.

In Fig. 2.4, we verify vs. spin-up cycles of the same experiments, where only anchor observations are assimilated. In doing so, we assess the viability of assimilating biased observations altogether. We find that the scalar correction experiments, particularly with inflated errors, improve RMSEs and biases across the board. Meanwhile, the experiments which estimate spatially heterogeneous biases perform worse as the imposed model bias increases. While RMSEs are only degraded slightly, likely because of the increased noise associated with sparse observation networks, the total bias degradation is severe, indicating that model bias contamination is a significant detriment even for a configuration with a relatively dense network of anchor observations. Interestingly, methodology performance is largely unaffected by imposed observation bias, indicating that the primary driver of degraded performance is instead imposed model bias.

To quantify the skill of each method across the parameter space explored in numerical experiments, we summarize the performance of each method in Figs. 2.5 and 2.6. These figures give percent improvement in RMSE and absolute bias for each bias correction method relative to a control configuration (no bias correction and anchor observations only, respectively), aggregated across 25 model/observation bias combinations. The 2×3 grid stratifies results by standard observation frequency (rows) and anchor observation frequency (columns). Thus, we are able to quantify the sensitivity of bias correction performance to observation network density.

Fig. 2.5 shows that while the scalar methods depend very little on observation frequency, the spatially-varying vector experiments' performances degrade rapidly with decreasing anchor observation density. This result is unsurprising, since for these experiments, the model bias estimate is only accurate

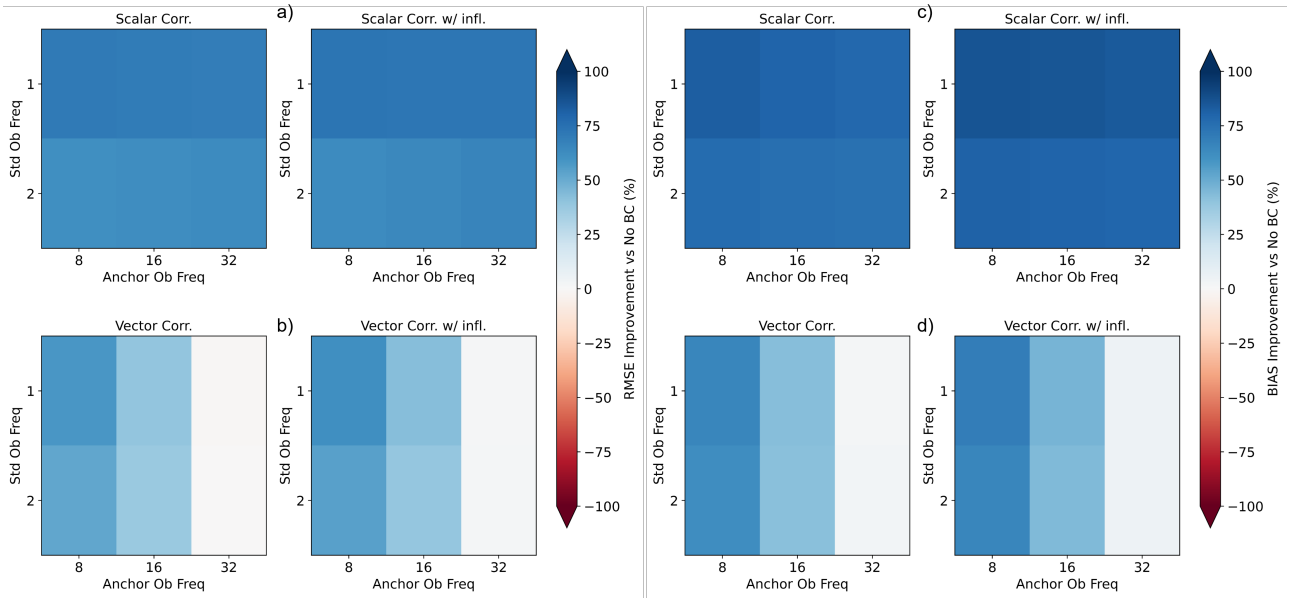


Figure 2.5: Percent improvement in time-averaged RMSE relative to experiments with no bias correction for each bias correction methodology, aggregated across 25 model/observation bias configurations. Rows correspond to standard observation frequency (every 1 or 2 grid points), columns to anchor observation frequency (every 8, 16, or 32 grid points). Blue shading indicates improvement over control, while red shading indicates degradation. Cell annotations denote percent change.

for grid points directly observed by or in close proximity to anchor observations. Extending these estimates to unobserved regions relies on background error correlations in $\mathbf{K}^*$ to spread information from anchor locations, but when anchors are sparse, these correlations are insufficient to constrain model bias across the domain. Meanwhile, the scalar experiments only consider state variables directly observed by anchoring observations, and therefore produces accurate model bias estimates when imposed model bias is global and static. Figure 2.6 confirms this pattern, and shows how absolute bias is strongly degraded for sparse anchor observation networks. In such cases, where biases are large and we cannot assume a homogeneous model bias across the domain, we have shown that it is actually more preferable to simply ignore these biased observations altogether.

We find that experiments that use an extra preliminary update with inflated background errors always perform slightly better than those without. Additionally, we note that across all experiments

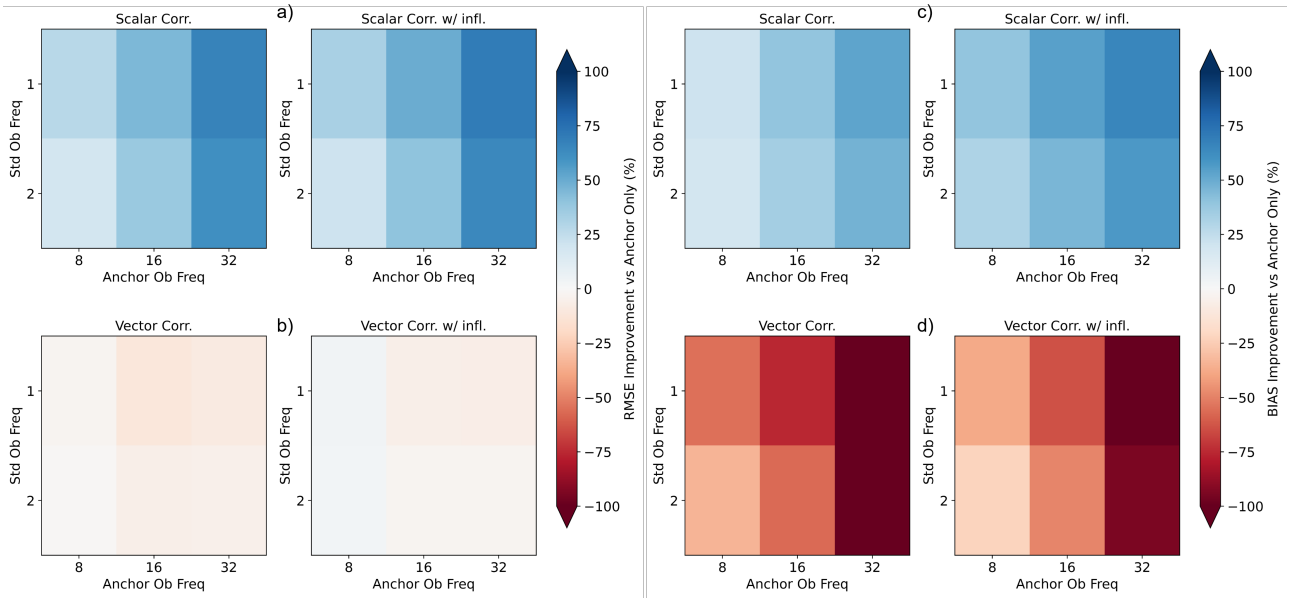


Figure 2.6: Percent improvement in time-averaged RMSE relative to the spin-up period, in which only anchor observations are assimilated, for each bias correction methodology, aggregated across 25 model/observation bias configurations. Rows correspond to standard observation frequency (every 1 or 2 grid points), columns to anchor observation frequency (every 8, 16, or 32 grid points). Blue shading indicates improvement over control, while red shading indicates degradation. Cell annotations denote percent change.

and controls, decreasing the standard observation frequency reduces the influence of bias correction in total, whether that influence is positive or negative. Since standard observation frequency has such little effect on whether the methodology ultimately improves or degrades performance, we primarily examine networks with standard observation frequencies of 1 for the remainder of these results.

Figs. 2.7 and 2.8 present the temporal evolution of RMSE, error in the model bias estimates, and errors in the observation bias estimates for each correction method, averaged across 25 bias configurations. In doing so, we reveal potential bias reinforcement and subsequent drift in the estimated bias parameters.

Across both observation networks, we notice that the naive correction experiments begin drifting immediately and very severely, eclipsing even the control ”no correction” experiment when anchor observations are sparse. The spatially-varying vector experiments stabilize quickly for the case in which

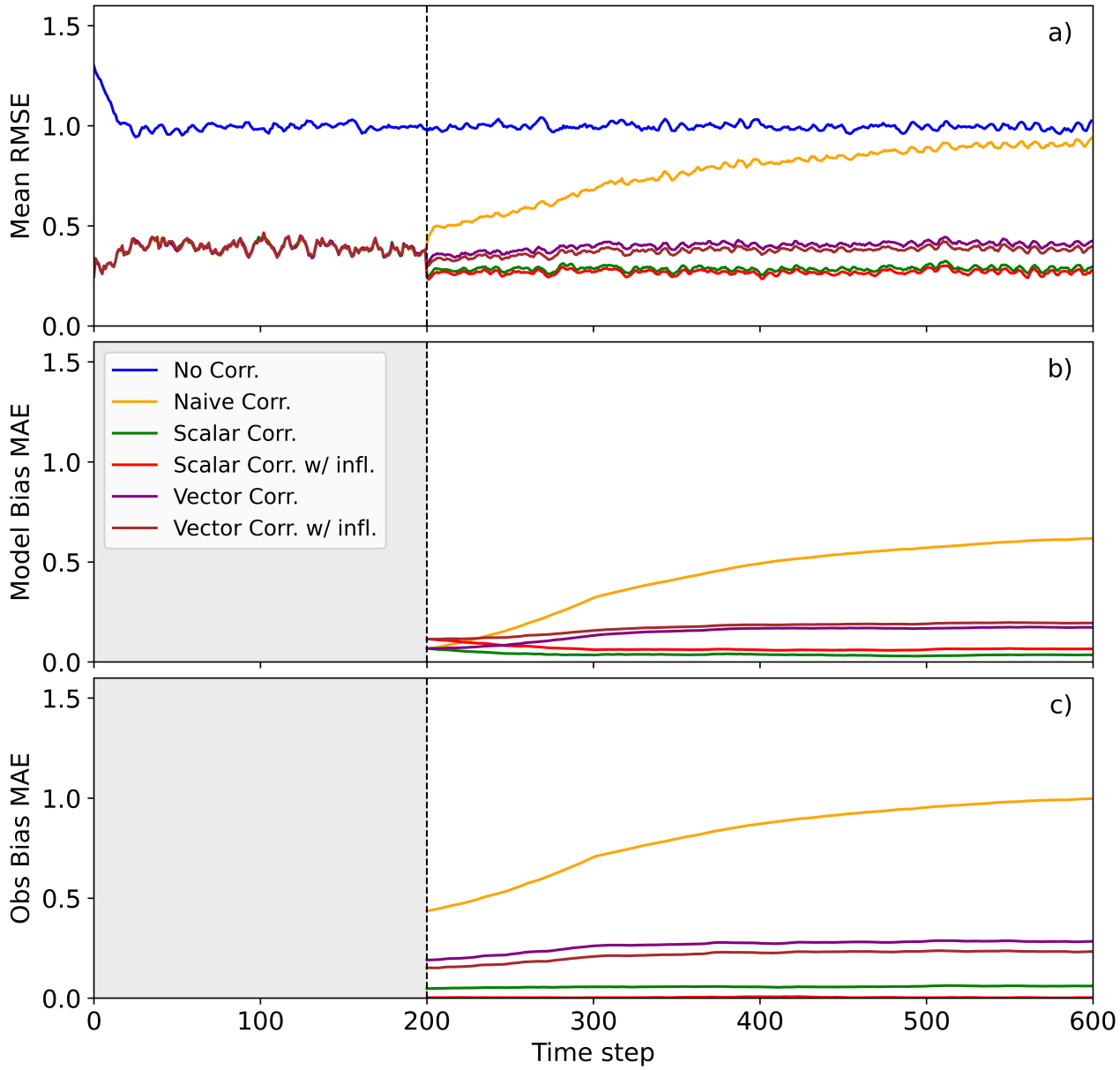


Figure 2.7: Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction.

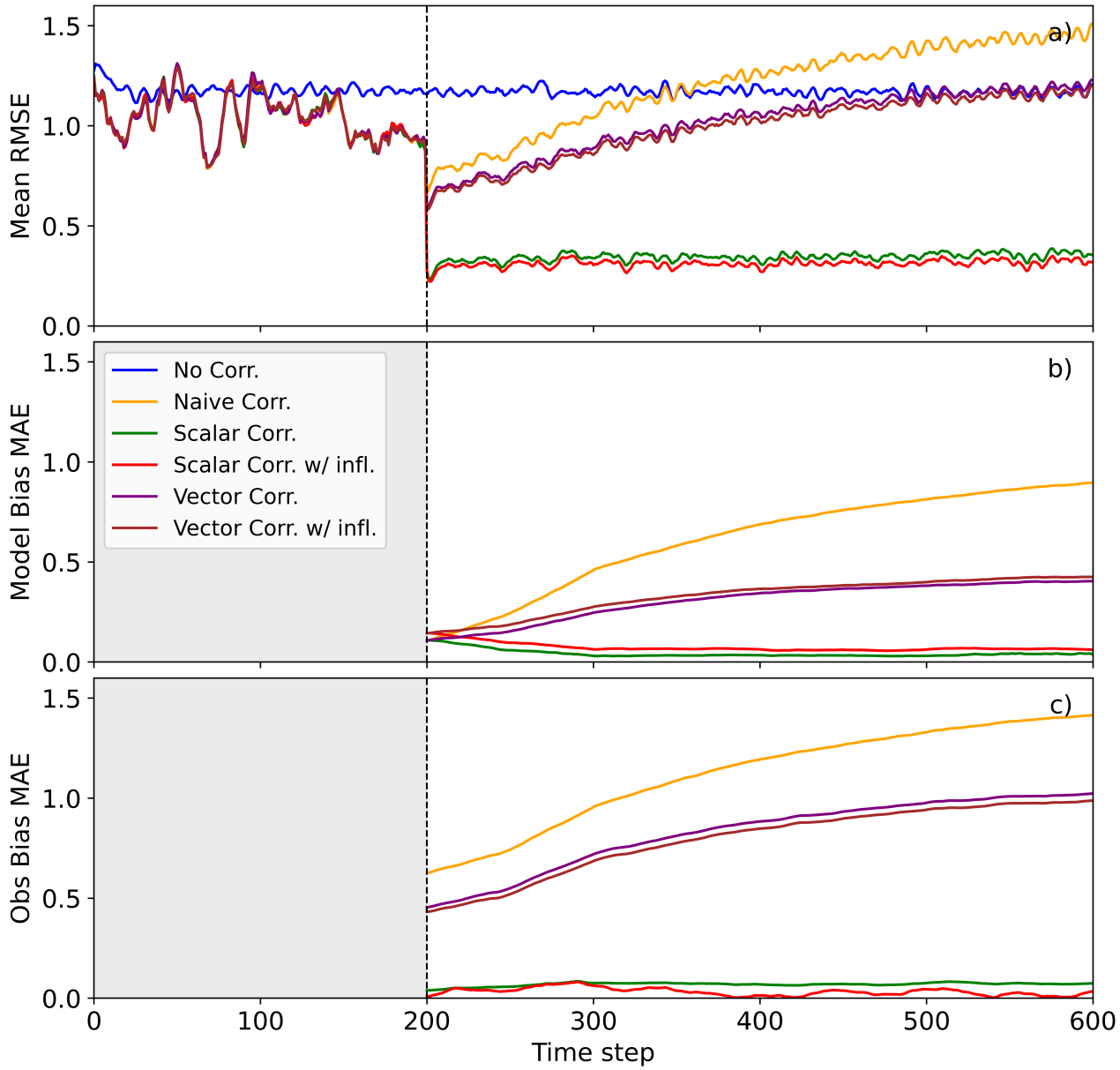


Figure 2.8: Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 32 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction.

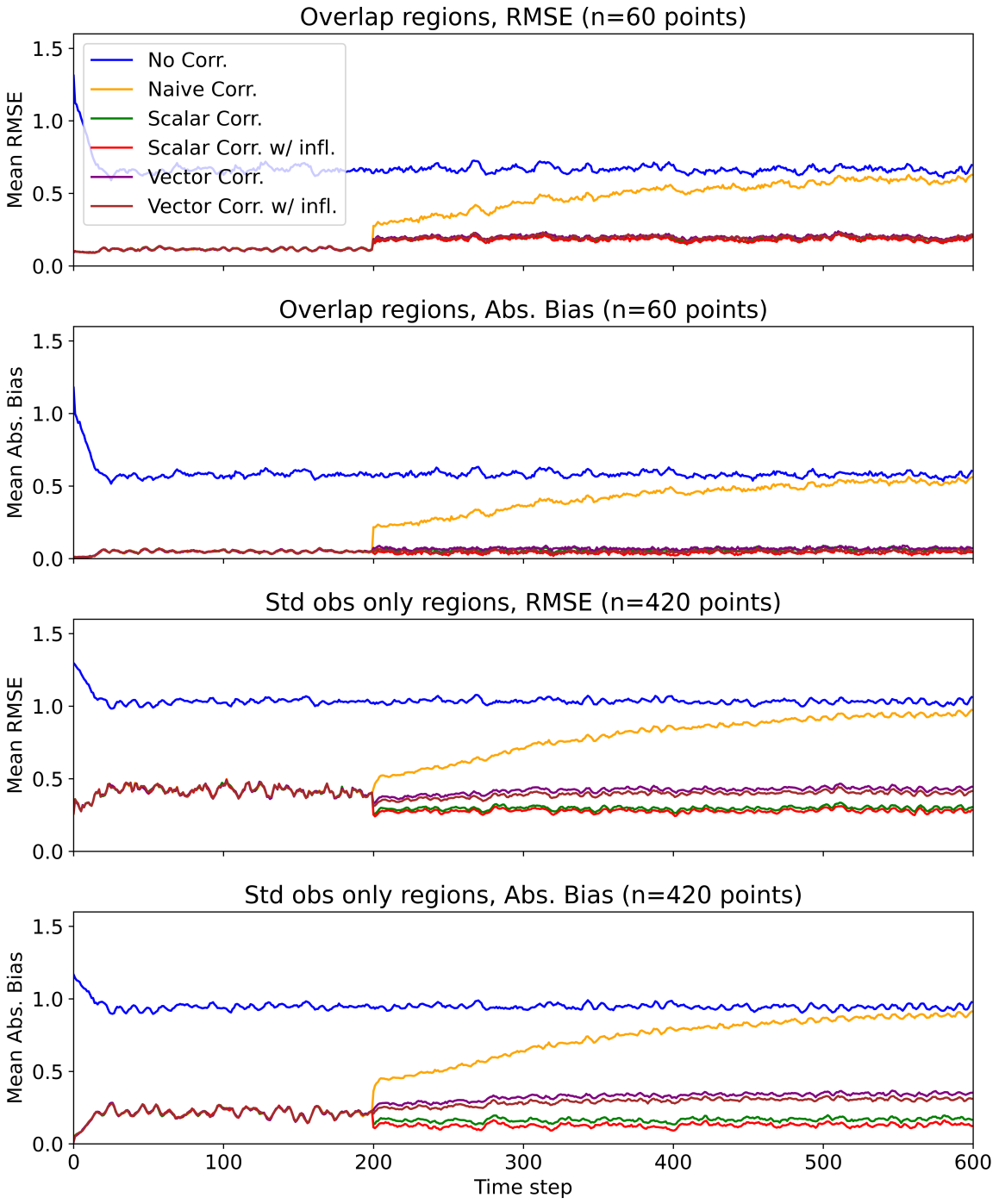


Figure 2.9: Temporal evolution of RMSE & absolute bias stratified by observation coverage. a) gives standard observation RMSE, b) gives anchor observation RMSE, c) gives standard observation absolute bias, and d) gives anchor observation absolute bias. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. Results are averaged across 25 bias configurations.

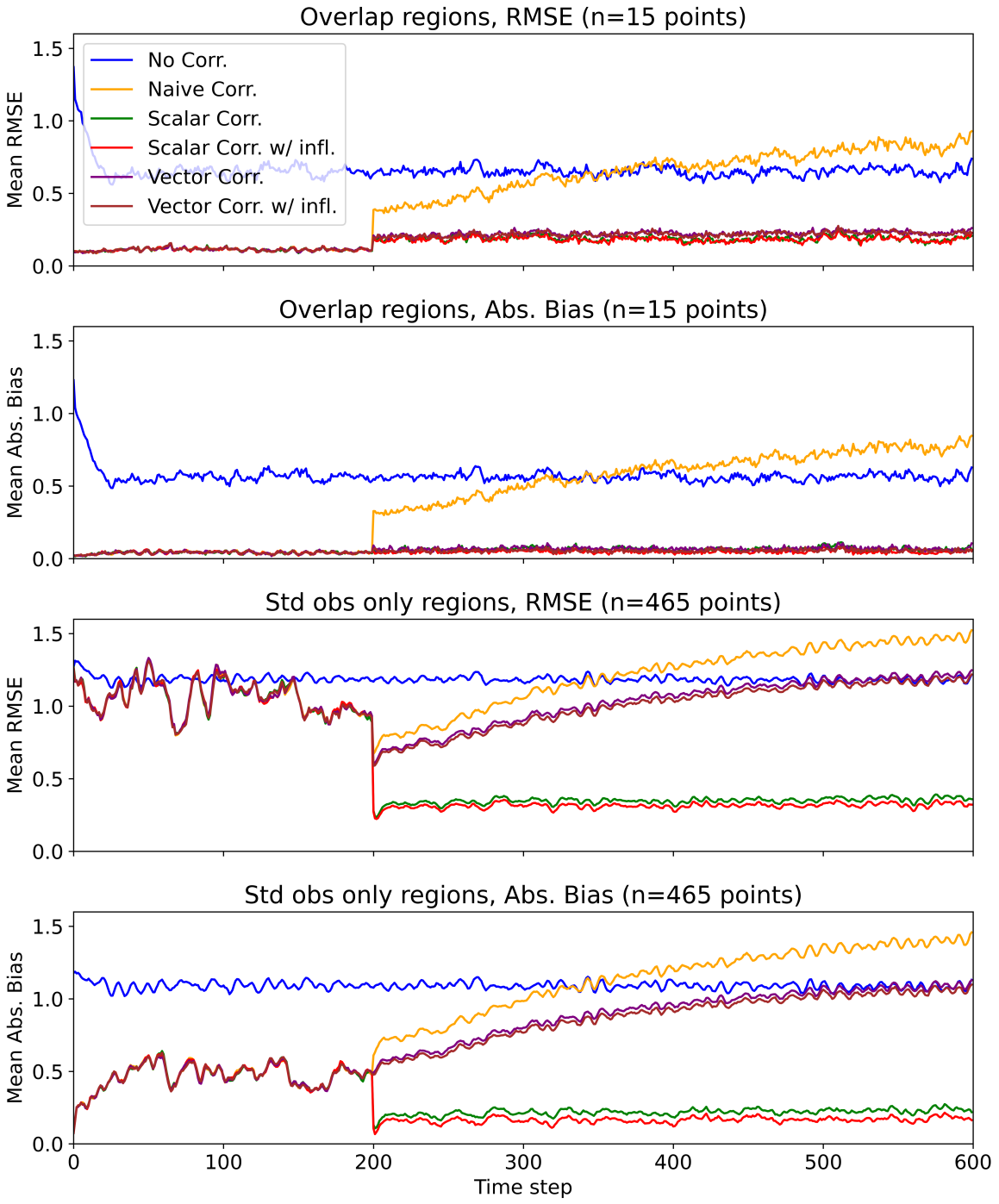


Figure 2.10: Temporal evolution of RMSE & absolute bias stratified by observation coverage. a) gives standard observation RMSE, b) gives anchor observation RMSE, c) gives standard observation absolute bias, and d) gives anchor observation absolute bias. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 32 grid points, respectively. Results are averaged across 25 bias configurations.

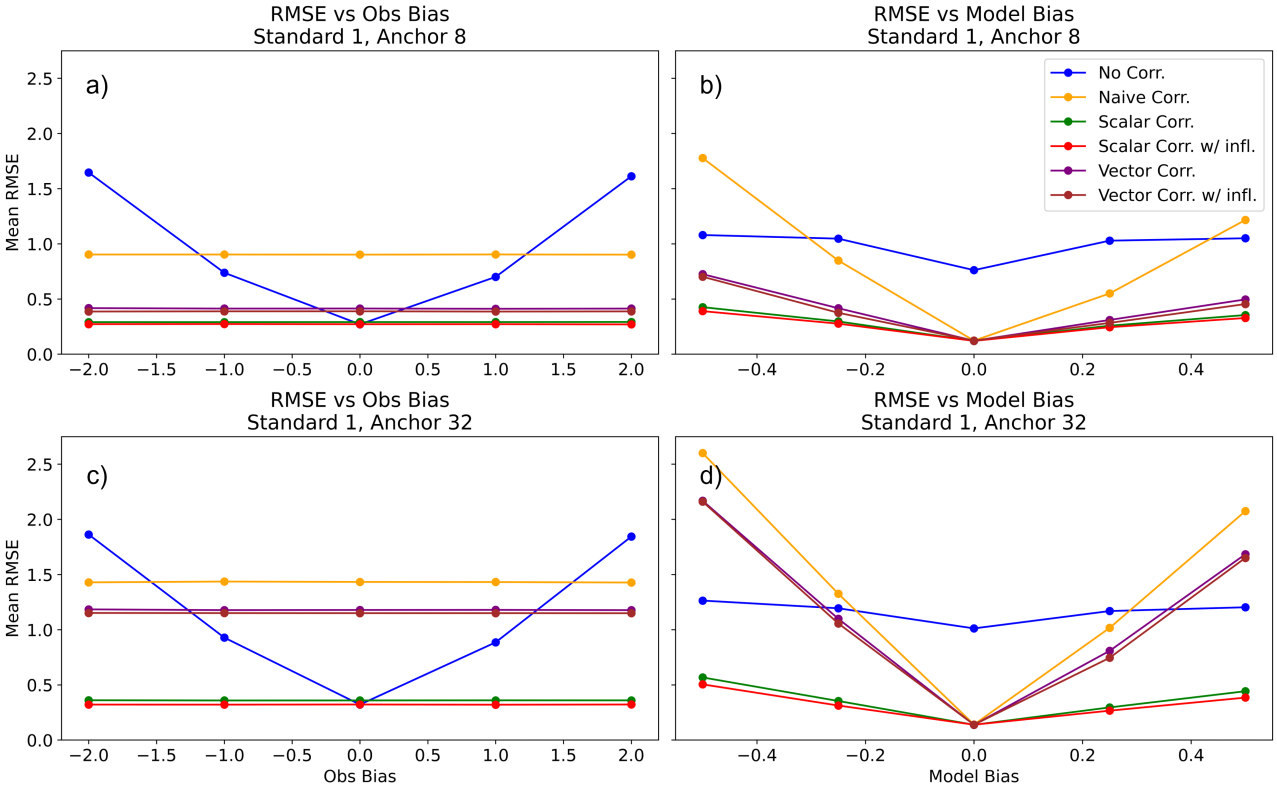


Figure 2.11: Time-averaged RMSE as a function of (a) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (b) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (c) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points, and (d) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points. Each line represents a different bias correction method.

anchor observations are more frequent, but drift severely when anchoring observations are sparse, and eventually result in larger RMSEs than the no correction experiments. For these experiments, we find that model bias estimates and observation bias estimates grow and stabilize at roughly the same rate. Meanwhile, the scalar methodologies perform the best and drift very minimally, if at all, throughout the experimental period.

Interestingly, we find that while the experiments with an extra preliminary error-inflated update have slightly improved RMSEs and observation bias estimates, they have slightly worse model bias estimates. This suggests these methodologies shift the residual model bias off of the observation

bias estimate and onto the model bias estimate instead, which results in a more accurate estimate of observation bias and therefore slightly lower RMSEs. We further note that in the case of sparse anchoring observations, the spin-up period produces lower RMSEs compared to the final cycles of the vector methodologies, again suggesting that in such scenarios it may actually be more beneficial to ignore these biased observations altogether.

Figs. 2.9 and 2.10 decompose RMSE and absolute bias into contributions from grid points co-located with anchor observations (hereafter referred to as "overlap" regions") versus grid points observed only by standard observations. In doing so, we demonstrate the limitations of these bias correction schemes in their ability to propagate model bias estimates from anchor-observed locations to unobserved regions via background error correlations.

Again, we notice a severe drift in the naive correction results across both scenarios and across all observation types. In Fig. 2.9, where anchor observations are relatively frequent, we see that all 4 model-bias-aware methodologies increase RMSE slightly over overlap regions, likely due to increased noise, but keep absolute bias relatively static. Meanwhile, the scalar methodology reduces both RMSE and absolute bias over standard observation areas, while the spatially-varying vector methodologies slightly reduce RMSEs but increase absolute biases.

The situation is very different for Fig. 2.10, where anchor observations are sparse, particularly over standard observation-only regions. For overlap regions, the same pattern holds. Over standard observation regions, the scalar method drastically reduces both RMSEs and absolute biases. The vector method demonstrates larger RMSE and absolute bias drift over standard observation regions, with error magnitudes eventually passing the spin-up period and roughly equaling the experiments with zero bias correction. These results indicate that when anchor observations are sparse, background error correlations in $\mathbf{K}^*$ are insufficient to propagate model bias estimates across distant grid points. In other

words, these spatially-varying correction methods are ineffective at constraining bias over much of the domain, and consequently do more harm than good with their corrections.

In Fig. 2.11, we assess method performance as a function of either imposed observation bias or model bias. We demonstrate this by taking the mean of RMSE results across all imposed model biases or observation biases, respectively, and plot against the opposing bias. This diagnostic assesses the robustness of each bias correction method under increasing bias magnitudes and identifies thresholds beyond which methods fail to adequately correct for systematic errors.

Immediately, we notice in Figs. 2.11a and 2.11c that the performance of these methods is largely independent of the imposed observation bias and is instead a function of imposed model bias. Furthermore, we show in Figs. 2.11b and 2.11d that each method has the same minimum RMSE at zero model bias, from which all RMSEs increase as model bias increases outward. Conversely, when no correction is applied, the RMSEs are much more sensitive to imposed observation bias, fluctuating very little when model bias is increased. We show that the spatially-varying vector methodologies perform poorly across all observation biases when anchoring observations are sparse. The relationship with imposed model bias is more complex. These results show that a vector estimation approach sufficiently removes bias when imposed model bias is low and/or when anchor observation frequency is high.

2.4.2 Model Correction Results

Lastly, we examine the potential value in correcting model bias directly in the background state, along side the observation bias correction. The figures below resemble Figs. 2.7, 2.8, and 2.11, but include two additional experiments: ”Scalar Model Correct,” in which the bias estimates are taken to be the mean of statistics for state variables directly observed by anchor observations; and ”Vector Model

Correct," in which bias estimates are taken as a spatially-varying vector over the entire state.

Immediately, we notice that both direct model correction methodologies perform better than their respective observation correction-only equivalents. In fact, the scalar model correction methodology produces the smallest RMSEs across all experiments, and is perhaps the only method that actually continuously improves RMSEs over the experimental period, although it noticeably does take extra time to converge to an accurate observation bias estimate, as seen in Figs. 2.12c and 2.13c.

Likewise, these model correction methods perform much better than their equivalents when anchor observation density is reduced to every 32 state variables. While the observation vector correction method produces significant drift, the state vector correction method continues to perform reasonably well, producing smaller RMSEs than both the no correction experiments as well as the anchor observation only spin-up period.

In Fig. 2.14, we come to largely the same conclusions as earlier in this study. Analysis RMSEs are mostly independent of observation bias magnitude across experiments. Furthermore, while the model correction methods are sensitive to imposed model bias magnitude, they nevertheless tend to outperform the observation correction-only methods. However, we find that the vector model correction method is the only bias correction method that degrades performance when imposed model bias zero. This may occur because, in the absence of model bias, the analysis increments reflect observation bias residuals rather than true model error. The method may be misinterpreting this signal as model bias and consequently applying a spurious correction to the background state, introducing error rather than removing it.

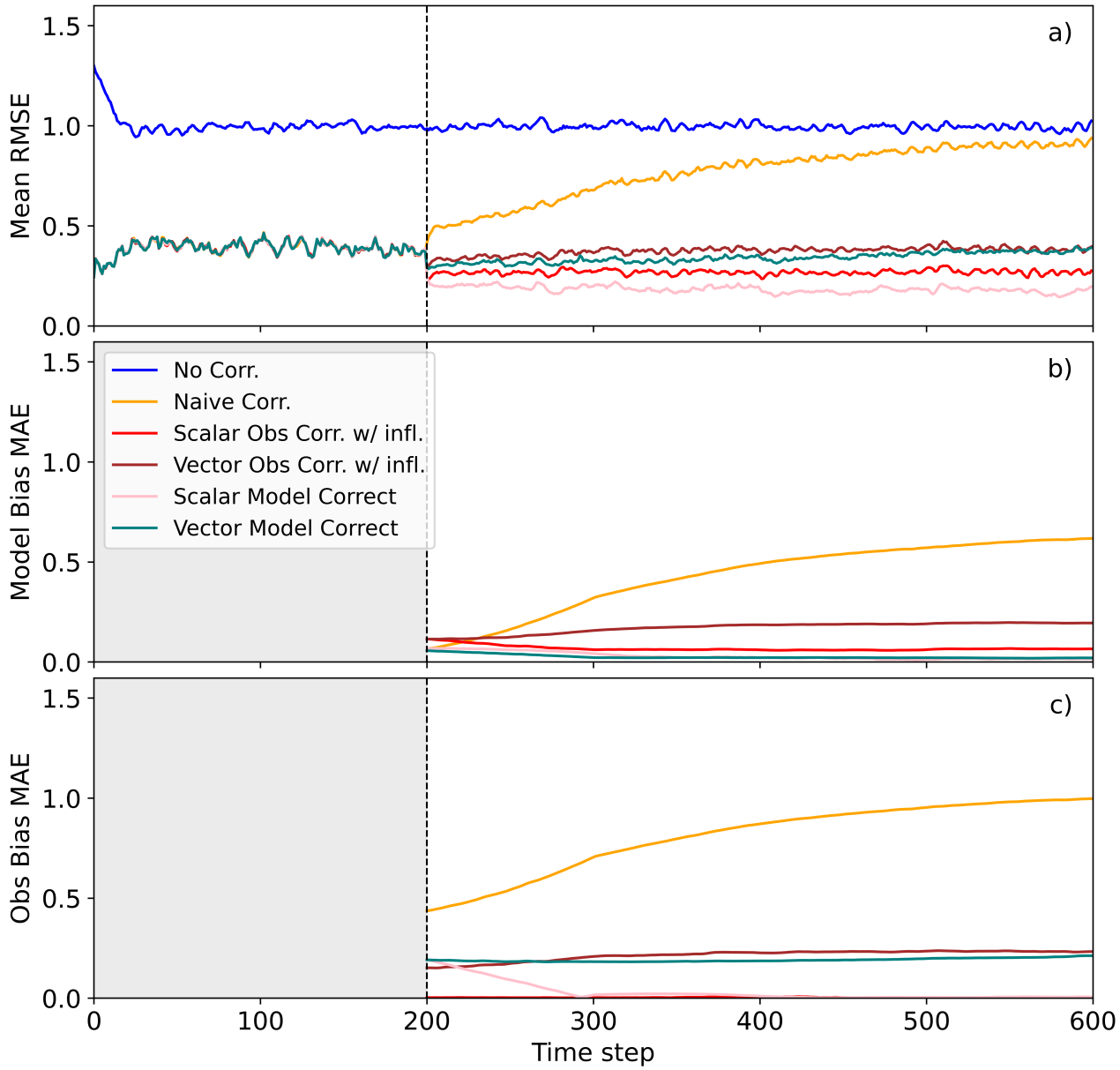


Figure 2.12: Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 8 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction.

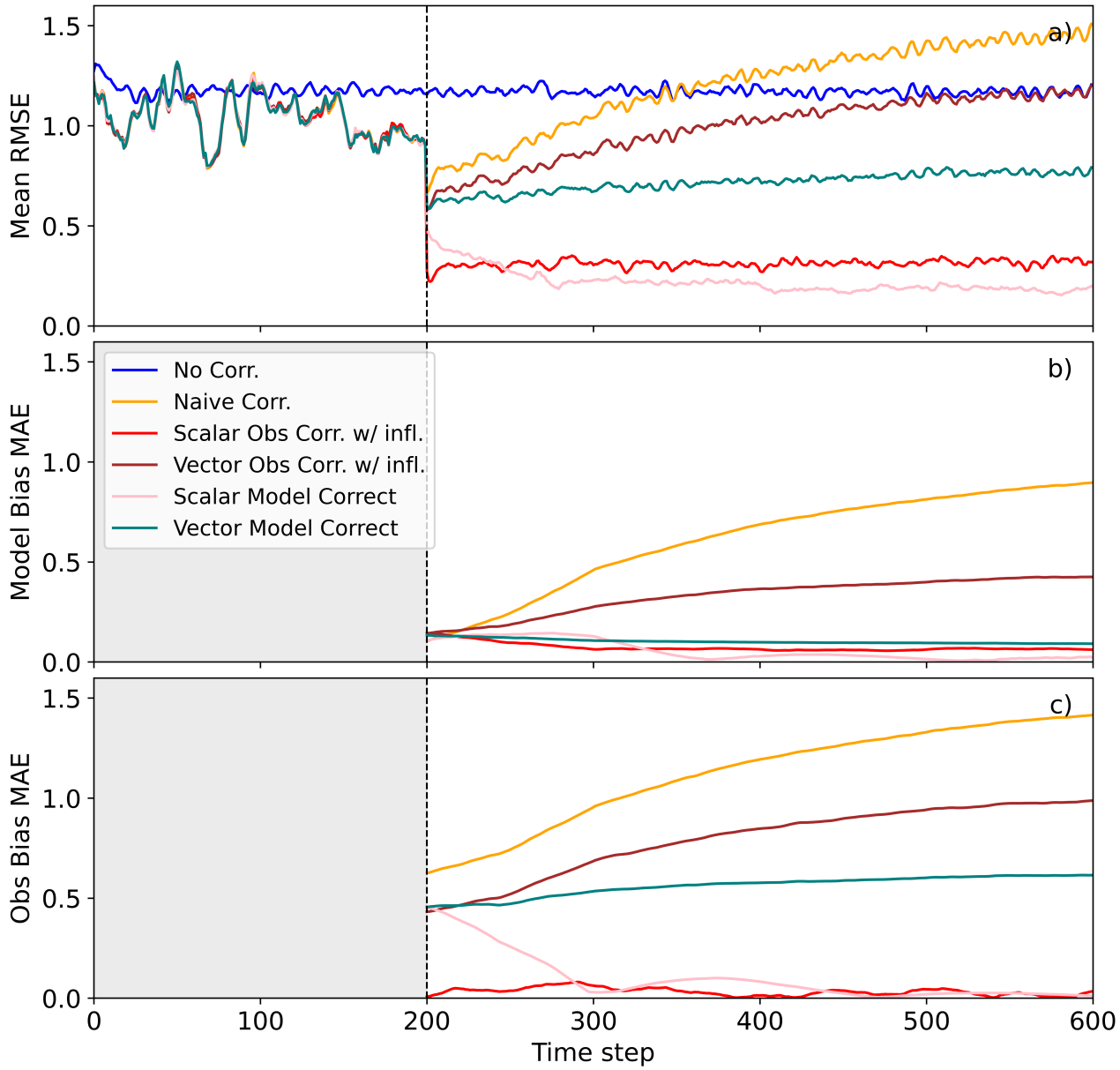


Figure 2.13: Temporal evolution of (a) RMSE vs. truth, (b) model bias estimation error, and (c) observation bias estimation error for each bias correction method, averaged across 25 bias configurations. State variables are observed by standard biased observations and unbiased anchoring observations every 1 and 32 grid points, respectively. Bias estimation error is computed as the absolute difference between estimated and imposed bias values. Note that only anchor observations are assimilated prior to cycle 200, when bias correction begins, except for "No Corr." which does not feature spin-up or bias correction.

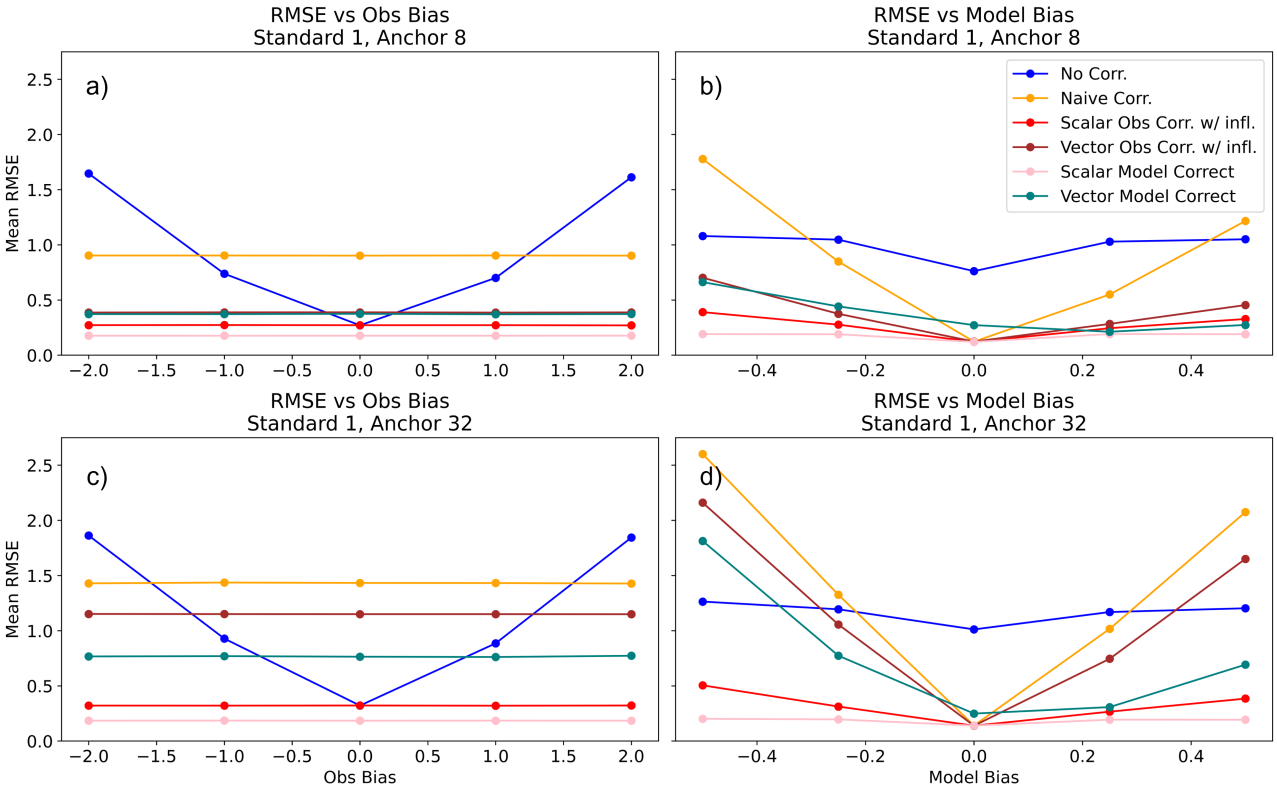


Figure 2.14: Time-averaged RMSE as a function of (a) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (b) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 8 grid points, (c) observation bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points, and (d) model bias magnitude w/ standard biased observations and unbiased anchoring observations every 1 and 32 grid points. Each line represents a different bias correction method.

2.5 Summary and Conclusions

When correcting for observations bias, we must consider both model process error as well as other sources of error that account for the accumulation of biases over many cycles. Failing to consider the influence of these sources of error can lead to bias reinforcement, and subsequently drift in our model state over many cycles. Without complete anchor observation coverage, it is impossible to solve for observation bias explicitly. Running a preliminary update with anchor observations can provide valuable information about state variables in close proximity, but the usefulness of these analysis increments depends on the error correlations in $\mathbf{K}^*$. When anchor observations are sparse or background error covariances are weak, analysis increments fail to constrain bias in unobserved regions.

Ultimately, these methods require either that we (a) have extensive anchor observation coverage, (b) inflate background errors to unrealistic levels during preliminary steps to estimate background bias in the space of anchor observations, and/or (c) extrapolate bias estimates to unobserved regimes. Alternatively, we can correct the model state itself using an estimate of model bias. However, this strategy is even more unrealistic for high-dimensional NWP systems, as time-mean analysis increments cannot account for errors-of-the-day and may break dynamical balances.

In this study, we examine the effects of various bias correction strategies on the Lorenz [2005] Model III, a dynamical chaotic model commonly used as a testbed for DA and bias correction experiments. We impose spatially homogeneous sources of bias, and use simple static bias-correction strategies to estimate the imposed model and observation bias. In doing so, we identify fundamental sources of analysis bias to demonstrate the limitations of observation bias correction strategies that fail to account for the effects of model process error as well as accumulated model and observation bias. We construct several linear methods that take advantage of unbiased "anchoring" observations

to estimate and remove the effects of background bias on observation bias estimates. We examine scalar corrections, in which the mean of model bias is taken over state variables directly observed by anchor observations, and compare vs. vector correction methods, wherein model bias estimates are spatially-varying based on anchor observation location and frequency. We then assess their performance vs. a "naive" correction strategy that does not consider the effects of background bias, as well as vs. experiments where we do not assimilate biased observations at all.

We find that naive bias correction very quickly leads to severe model state drift, and that strategies that account for background bias can potentially prevent this drift. Scalar methods represent the best observation correction strategy that can be achieved, as we rely only on quantities that are directly observed rather than inferred through the Kalman gain. Spatially-varying vector methods are more realistic but less effective, and degrade results very severely when anchor observations are sparse and when error covariances in $\mathbf{K}^*$ are weak.

Across all methods we find that performing an extra preliminary update with inflated background errors produces lowers RMSEs. This is consistent with our theoretical framework from section 2.2.4, where the preliminary update completely removes expected analysis error in observation space under specific choices where $\mathbf{K}^* \rightarrow 1$ and $\mathbf{W}^* \rightarrow 0$. We also find that the performance of all these methodologies depends primarily on model bias magnitude, as opposed to observation bias magnitude. Likewise, this is consistent within our framework, where observation bias correction successfully removes $\bar{\epsilon}$, but leaves the term $\mathbf{W}^*\overline{\eta}$ which scales with model bias magnitude.

Finally, in sparse anchor observation scenarios, we find that it may actually be preferable to ignore biased observations altogether. When $\mathbf{H}^*$ samples too few locations and $\mathbf{K}^*$ contains weak error correlations, the resulting bias estimates are incomplete and potentially harmful. Applying corrections based on these flawed estimates can actively degrade results compared to experiments in which only

anchor observations are assimilated

These results reiterate that anchor observation density is critical for effective bias correction, particularly for strongly biased observation systems such as satellite radiances, but suggest that current methods may indeed be worse than doing no correction at all. In operational NWP, we have access to a variety of anchor observations, but they each possess limiting factors: radiosondes and surface stations are concentrated over land and the Northern Hemisphere, leaving oceans and Southern Hemisphere under-sampled; radio occultations have global coverage and good vertical resolution but have limited sampling in the lower troposphere; aircraft observations are dense along flight corridors but sparse elsewhere, and their spatial-temporal availability is highly variable [Kleist et al., 2023]. Thus, while radiance bias correction may be reliable over well-anchored regions, such as over Northern Hemispheric land, it may suffer from model bias contamination over anchor observation-sparse regions, such as over ocean and in the Southern Hemisphere. These findings align with similar studies on the topic, such as Francis et al. [2023].

Of course, this study has a number of limitations that vastly simplify the issue at hand. Namely, our bias characteristics are global and static, while satellite radiances have highly variable biases with complex state-dependent interactions. In practice, the measurement operator for radiance observations takes the form of a radiative transfer model that is inherently nonlinear and state-dependent. This introduces an additional source of error that violates the assumption that observation error is independent of the model state. In this case, the observation bias would become a function of the atmospheric state, with two important implications for our findings. First, time-mean innovations would conflate true observation bias with state-dependent error, degrading bias estimates even under ideal anchor observation coverage. Second, spatially-varying state-dependent bias would require adaptive corrections that evolve with the model state itself. Nevertheless, this work illustrates challenges for this problem

even for a very simple application. Ultimately, we find the fundamental issue still remains: how to assess observation bias in the presence of model bias without a proper reference state.

Future work should focus on scaling model bias-aware strategies to operational models within existing observation bias correction frameworks, such as VarBC. Efforts should also be taken to consider how these strategies affect ensemble spread and covariance structure, and how correlations between state variables can be better exploited to expand the effectiveness of local model bias estimates.

Extending these insights to a full NWP system, the following chapter examines whether HAFS can operate in a fully-cycled, self-contained mode with online satellite radiance bias correction. In particular, we assess whether training bias correction coefficients on the native HAFS grid reduces bias and improves forecasts over the conventional approach of adopting coefficients from GDAS, which can introduce undiagnosed external bias into the estimation process.

Chapter 3: Implications of Self-Contained Radiance Bias Correction for Data Assimilation within the Hurricane Analysis and Forecasting System (HAFS)

3.1 Introduction

The accurate prediction of TC track and intensity has always presented a unique problem. The extreme weather event, which lies somewhere in between the traditionally established spatial-temporal classifications of synoptic scale and mesoscale, develops predominately over the observation-sparse ocean and produces storm conditions too intense for common surface-level observation systems to properly measure. For modern numerical weather prediction (NWP) systems, these issues remain significant obstacles [Emanuel and Zhang, 2016a]. In order to depict the TC inner-core, limited-area hurricane-focused NWP systems, such as the current operational Hurricane Weather Research and Forecasting (HWRF) model, require a high-resolution grid and ample inner-core observations, such as those collected during aircraft reconnaissance missions [Aksoy et al., 2012, Sippel et al., 2022, Zhang et al., 2009, 2013]. However, given the size of the storm and its interaction with broad synoptic-scale features, these limited-area models (LAMs) must also be able to accurately characterize synoptic-scale flow over a large domain, particularly for the purpose of predicting changes in TC structure as well as track and intensification [Carr and Elsberry, 1997, Hanley et al., 2001, Hendricks et al., 2010, Jones et al., 2018, Poterjoy and Zhang, 2014, 2016, Zeng et al., 2007].

Given the sparsity of in-situ measurements over the open ocean, TC-focused NWP systems rely heavily on satellite radiance measurements, which have a high spatial and temporal resolution over the entire model domain [Prasad et al., 1998]. These observations primarily provide information on atmospheric brightness temperature (BT), from which atmospheric properties such as temperature,

moisture, and wind can be determined [Deng et al., 2009]. Studies have shown improvements in TC prediction from assimilating satellite radiance both from infrared and microwave channels [Li and Liu, 2009, Liu et al., 2012, McNally et al., 2000, Schwartz et al., 2012, Xu et al., 2013, Zapotocny et al., 2008, Zou et al., 2013]. In this study, we concentrate on clear-sky radiances, meaning observations over cloudless or thin cloud regions, vs cloudy- or all-sky radiance measurements, which describe observations over thick clouds or precipitating clouds. Although all-sky radiance observations represent the most comprehensive description of atmospheric state available, evaluating BT over cloudy regions is complicated by nonlinearity in the observation operator [Geer and Bauer, 2011a, Minamide and Zhang, 2017, Poterjoy et al., 2017, Zhu et al., 2016]. Furthermore, complex cloud microphysical processes are difficult to parameterize and modern radiative transfer models are imperfect, which both result in degraded model representations of cloudy BTs. For these reasons, cloudy-sky radiances are often removed or used sparingly for DA purposes [Reale et al., 2018]. On the other hand, clear-sky radiances are much more easily evaluated and provide valuable temperature and moisture field data across large areas. For these reasons, the use of clear-air radiance measurements for numerical weather prediction has matured more rapidly than cloudy-sky measurements at operational modeling centers [Collard and McNally, 2009, McNally et al., 2007].

Despite the known benefits, clear-sky radiance observations have been shown to be systematically biased, and thus require a bias correction (BC) scheme. Bias can arise from several different sources, namely: from the radiative transfer model, from the calibration of the satellite instrument, as well as errors in the background model state [Derber and Wu, 1998, Liu et al., 2018, Reale et al., 2018, Zhu et al., 2014]. With a method described in detail in the following section, the Hurricane Analysis and Forecast System (HAFS) currently uses the Gridpoint Statistical Interpolation (GSI) variational DA scheme, which performs bias correction twice during the analysis, once for each outer loop [Hu

et al., 2016]. The bias correction value is added to the observation space prior and consists of a sum of weighted bias predictors. The predictors are a function of atmospheric state, while the corresponding weights are determined from analysis innovation statistics during the minimization of the variational cost function [Benáček and Mile, 2019]. These statistics are collected continuously, meaning the DA system must be fully cycled, and cover a sufficiently large domain. For this reason, it is common for this technique to be employed in global systems but it is rarely used for LAMs given the restricted domain size as well as other common restrictions, such as the adoption of partially-cycled DA [Lin et al., 2017, Poterjoy et al., 2021].

Instead, LAMs import predictor weight values from externally trained models. For example, HWRF is a partially-cycled system that regularly adopts the GDAS analysis for the parent domain [Mehra et al., 2018, Zou et al., 2013], thus requires predictor weights from GDAS. This strategy is not ideal, given that the bias correction method does not differentiate between various sources of bias and thus would not be able to separate bias imposed by the externally trained model [Poterjoy et al., 2021, Rizzi and Matricardi, 1998].

The HAFS is NOAA's next-generation FV3-based TC forecasting system. While early operational implementations of HAFS will strongly resemble the DA methodology used for HWRF, research versions of HAFS can operate within a flexible DA framework that allows users to avoid many of the heuristic design choices that are common for operational models. For example, HAFS has the ability to perform fully-cycled DA across its entire domain, allowing for the continuous updating of BC predictor weights during each analysis cycle. This capability opens the door not just to online satellite radiance bias correction, but also many other experimental techniques that require a fully-cycling system [Poterjoy et al., 2021]. For this paper, we will focus more narrowly on the effects of the online BC technique, and compare it to model results wherein BC parameters are adopted from GDAS.

In Section 3.2, we detail the radiance bias correction methodology and include relevant equations. Section 3.3 describes the data assimilation framework, model physics, and relevant observation details. Section 3.4 presents results in the form of domain- and time-averaged error verification, model field comparisons, as well as a case study of a notable TC that developed over the model time period. Section 3.5 gives a summary and concluding remarks.

3.2 Radiance Bias Correction

In order to properly use satellite radiance BT measurements, we must correct biases between the observed radiances and first-guess model values. The one-step "enhanced" bias correction scheme is performed within GSI, and is described in detail by Zhu et al. [2014]. For this study, we provide an abbreviated description of how this method operates within GSI.

The observation operator $\tilde{h}$ is described by

$$\tilde{h}(\mathbf{x}, \boldsymbol{\beta}) = h(\mathbf{x}) + b(\mathbf{x}, \boldsymbol{\beta}), \tag{3.1}$$

where $\mathbf{x}$ is the model state, $\boldsymbol{\beta}$ represents predictor coefficients, and $h(\mathbf{x})$ represents the radiative transfer model. HAFS incorporates the Joint Center for Satellite Data Assimilation (JCSDA) Community Radiative Transfer Model (CRTM) as a forward radiative transfer model to compute model-derived BT for various instruments and channels [Weng et al., 2006].

The function $b(\mathbf{x}, \boldsymbol{\beta})$ is represented by a linear combination of N_p predictor variables $p_k(\mathbf{x})$, where

$k = 0, 1, 2, ...N_p$, with corresponding predictor coefficients β_k. Thus we have

$$b(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \sum_{k=1}^{N_p} \beta_k p_k(\mathbf{x}). \tag{3.2}$$

Choices of predictors include four polynomials for scan angle; global offset; emissivity predictor; cloud liquid water (CLW) predictor; temperature lapse rate predictor; and square of the temperature lapse rate predictor. The emissivity predictor tends to be quite large over land for near-surface channels and was formulated to account for differences in emissivity over land vs. sea. On the other hand, the CLW predictor is applied to microwave radiances over oceans only. This predictor tends to be small and only applies to clear-sky radiance measurements. Although the quality control procedure is intended to screen data affected by thick clouds and precipitation, some cloud contamination is missed, and this predictor is included to fix such biases. Finally, the lapse rate and lapse rate squared terms connect changes in transmittance between vertical layers with changes in temperature, and ultimately BT bias [Zhu et al., 2014]. These coefficients are estimated during the minimization of the cost function:

$$\begin{aligned} J(\mathbf{x}, \boldsymbol{\beta}) = &\frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}_\mathbf{x}^{-1} (\mathbf{x} - \mathbf{x}_b) \\ &+ \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta_b})^T \mathbf{B}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta_b}) \\ &+ \frac{1}{2}[\mathbf{y} - \tilde{h}(\mathbf{x}, \boldsymbol{\beta})]^T \mathbf{R}^{-1} [\mathbf{y} - \tilde{h}(\mathbf{x}, \boldsymbol{\beta})], \end{aligned} \tag{3.3}$$

where $\mathbf{y}$ represents observations, $\mathbf{x}_b$ and $\boldsymbol{\beta_b}$ are first-guess estimates for $\mathbf{x}$ and $\boldsymbol{\beta}$, respectively, and $\mathbf{B}_\mathbf{x}$ is the background-error covariance for $\mathbf{x}$. $\mathbf{B}_\beta$ is the block-diagonal background-error covariance for

predictor coefficients, given by

$$\mathbf{B}_\beta = \begin{bmatrix} \mathbf{B}_\beta^{(1)} & & 0 \\ & \ddots & \\ 0 & & \mathbf{B}_\beta^{(J)} \end{bmatrix} \tag{3.4}$$

where $\mathbf{B}_\beta^{(j)}$ is a diagonal matrix of error standard deviations for channel j ($j = 1, 2, ..., N_c$), specified as 10.0 for all coefficients of all channels [Zhu et al., 2014].

To minimize the cost function, we can combine the control vectors $\mathbf{x}$ and $\boldsymbol{\beta}$ into a single control vector $\mathbf{z}$, such that

$$\mathbf{z}^\top = [\mathbf{x}^\top, \boldsymbol{\beta}^\top]. \tag{3.5}$$

with background of $\mathbf{z}^T = [\mathbf{x}_b^T, \boldsymbol{\beta_b}^T]$. Assuming no cross-covariance between $\mathbf{x}$ and $\boldsymbol{\beta}$, we write the error covariance matrix $\mathbf{B}$ as

$$\mathbf{B} = \begin{bmatrix} \mathbf{B_x} & 0 \\ 0 & \mathbf{B}_\beta \end{bmatrix}. \tag{3.6}$$

The cost function J is then given by

$$\begin{aligned} J(\mathbf{z}) = \frac{1}{2}(\mathbf{z} - \mathbf{z}_b)^T \mathbf{B}^{-1} (\mathbf{z} - \mathbf{z}_b) \\ + \frac{1}{2}[\mathbf{y} - \tilde{h}(\mathbf{z})]^T \mathbf{R}^{-1} [\mathbf{y} - \tilde{h}(\mathbf{z})]. \end{aligned} \tag{3.7}$$

In addition, several channels are designated as ”passive” channels, meaning these channels are read into GSI but do not impact the analysis. These channels may be monitored for research or quality control purposes, and are bias-corrected in a manner similar to assimilated observations.

We differentiate our experiments by the technique through which we initialize the first-guess predictor coefficients at each cycle during data assimilation. Our control experiment retrieves GDAS-

generated bias correction data for each cycle to serve as first-guess coefficients, while the ”online” experiment uses the above variational (VarBC) method to generate the predictor coefficients during each analysis step. The coefficients are updated during each outer-loop calculation, thus twice for each analysis [Hu et al., 2016].

3.3 Model Configuration

3.3.1 Forecast and Analysis

HAFS is the next-generation operational analysis and prediction system for hurricane forecasting at NOAA. The current options for HAFS feature a uniform global model with a high-resolution nest (HAFS-global-nest) and a stand-alone high-resolution regional model (HAFS-SAR). For our research, we are utilizing the HAFS-SAR domain configuration, a version of which will become operational in summer 2023. Henceforth we will refer to this configuration of the model simply as HAFS. Like the operational HAFS, we use FV3-SAR and ensemble-based variational data assimilation with the ensemble pulled from GDAS for background uncertainty specification. Unlike the operational HAFS, our domain remains in a fixed location over multiple storms and does not make use of high-resolution nests or vortex relocation and uses 3D-ensemble-Var (3DEnVar) in place of 4D-ensemble-var, which is adopted operationally. We choose to omit vortex relocation not only because we believe it will complicate the interpretation of our results, but also because we do not see it being used in the long term for TC data assimilation. Finally, we perform our own estimation of predictors for the bias correction model, as explained in more detail in section 3.3.3 [Dong et al., 2020, Gopalakishnan et al., 2021, Hazelton et al., 2021].

The FV3 dynamical core for HAFS is based on a non-hydrostatic finite volume solver using a

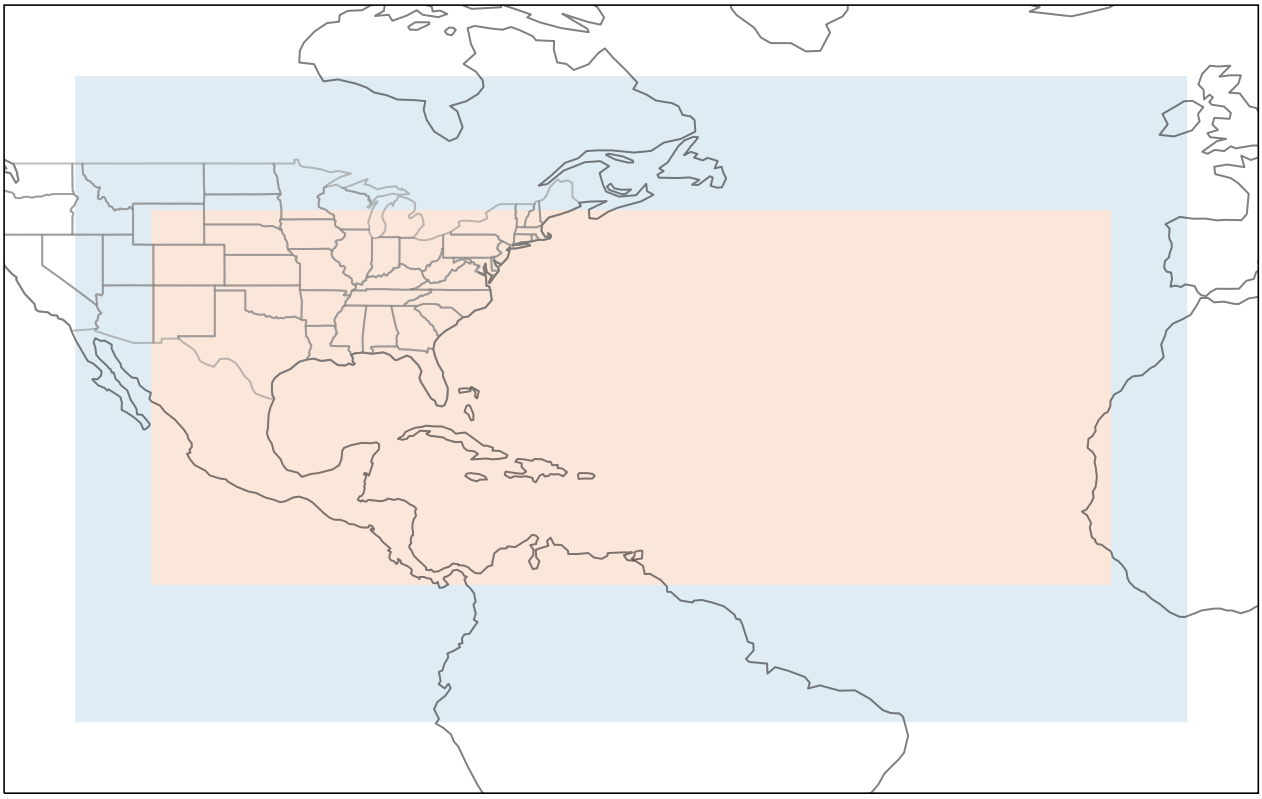


Figure 3.1: In blue, the entire HAFS-SAR model domain. In orange, a smaller subdomain selected for error verification (see section 3.4).

Lagrangian vertical coordinate [Hazelton et al., 2018]. The model domain adopted for our study uses a single grid of 1440 $\times$ 1080 horizontal grid cells, with an approximate 6-km grid spacing, and 81 vertical levels in sigma-pressure hybrid coordinates. The lowest model level is about 25 m above the surface and the top isobaric level is 2 hPa. The domain is centered at 25° N, 60° W, with a 109.5° span for longitude and a 68.2° span for latitude, as seen in Fig. 3.1. For our experiments, boundary conditions and the first set of initial conditions are interpolated from the 2020 operational global FV3-based Global Forecast System (GFS) ($\sim$13 km horizontal spacing) onto the 6-km HAFS domain. Boundary conditions are provided every 6 h from the same global GFS forecasts [Dong et al., 2020].

As previously mentioned, we use a GSI-based 3DEnVar data assimilation system. The data assimilation updates a single deterministic model solution every 6 h, using background error covariance

estimated entirely from the 80-member 6-h Global Ensemble Forecast System (GEFS) forecasts used by the operational GDAS [Bannister, 2017]—and no static error covariance. In addition, we employ the first guess at the appropriate time (FGAT) technique over the 6-hour DA window to account for differences between actual observation times and background times.

Consistent with Poterjoy et al. [2021], our configuration of HAFS does not make use of vortex relocation or finer-resolution nests, nor does it regularly re-center the model state onto GDAS solutions, which is a technique employed by operational LAMs. The partial-cycling strategy is often used to prevent model drift over time, which is often due to deficiencies in model physics, data assimilation, and boundary conditions. Nevertheless, we find our model to exhibit only a limited systematic change in background errors, which were determined from a careful inspection of observation-space statistics accumulated during the DA experiments.

3.3.2 Model Physics

We use a model physics configuration that resembles the operational National Centers for Environmental Prediction (NCEP) GFS version 16 (GFSv16), but with modifications that reflect boundary layer processes that are uniquely important for TCs. The physics suites employed include: the GFS Rapid Radiative Transfer Model for GCMs (RRTMG) shortwave/longwave radiation scheme [Mlawer et al., 1997]; the GFS Scale-Aware Simplified Arakawa-Schubert (sa-SAS) Deep Convection Scheme [Han et al., 2017a]; the Noah-Multiparameterization Land Surface Model (Noah-MP LSM) [Niu et al., 2011]; the GFS scale-aware TKE-based Moist Eddy-Diffusion Mass-Flux (EDMF) PBL and free atmospheric turbulence scheme [Han and Bretherton, 2019]; the GFS surface layer scheme with HWRF exchange coefficients [Zheng et al., 2017]; and the GFS orographic gravity wave drag scheme

[Kim and Arakawa, 1995];

The model configuration also features Community Mediator for Earth Prediction Systems (CMEPS) -based ocean coupling with the Hybrid Coordinate Ocean Model (HYCOM), which is described in detail in Chassignet et al. [2007, 2009]. The HYCOM system features a 1/12-degree resolution North Atlantic basin domain with 41 vertical levels, initial conditions from NCEP's Real Time Ocean Forecast System (RTOFS), persistent oceanic lateral boundary conditions, and atmospheric forcing from GFSv16 for non-overlapping areas.

3.3.3 Bias Correction Configurations

In this section, we outline a pair of experiments, hereby referred to as "control" or "GDAS BC" vs. "online" or "HAFS BC," that examine the effectiveness of online varBC for our LAM-configuration of HAFS. The "control" experiment borrows bias correction coefficients from GDAS, importing them for each outer loop of the analysis. These GDAS-based coefficients are obtained directly from the realtime GDAS, which were updated via the same GSI-based VarBC method detailed in section 3.2 over an equivalent training period.

The "online" experiment initializes BC coefficients from GDAS values for the first analysis cycle, and then transitions to self-cycling these coefficients within HAFS, updating them during each outer loop and using the resulting output coefficients as the first guess for the subsequent cycle [Hu et al., 2016]. In this way, the cycling of coefficients allows bias to accumulate during the data assimilation cycles, limiting potential system biases arising from the borrowing of GDAS coefficients. In other words, the configuration limits potential influence from an outside model and allows for better identification and understanding of biases arising from the satellite instruments and measurement operator alone.

3.3.4 Observations and Data Assimilation Schedule

The experiments begin 00 UTC August 18 and end at 00 UTC on September 22 2020, covering nine storms from TC Josephine to Hurricane Sally. In particular, this time span is chosen to include Hurricane Laura, a Category 4 hurricane which was the strongest TC of the season by maximum sustained winds and among the most powerful hurricanes on record to make US landfall [NOAA, 2021]. Our HAFS experiments assimilate measurements at 00, 06, 12, and 18 UTC each day, and generate medium-range (102-h) deterministic predictions on 00 and 12 UTC cycles. We also omit the first 7 days of data assimilation to allow HAFS to become independent from the GDAS initial conditions that are used to initialize experiments.

The suite of assimilated measurements includes all measurements assimilated by GFS except proprietary measurements, namely Aircraft Communications Addressing and Reporting System data and cloudy radiances. In addition, HAFS assimilates tail Dopplar radar and TC inner core radar data as collected on aircraft reconnaissance flights, which are not assimilated by the GDAS. In addition, HAFS assimilates inner core data, including tail Dopplar radar and dropsonde observations, which are not assimilated by the GDAS.

The Geophysical Fluid Dynamics Lab (GFDL) vortex tracker is used to generate the TC track information [Biswas et al., 2018]. TC track, maximum 10-m surface wind, and minimum surface pressure from model forecasts are verified against "best track" data from the National Hurricane Center [Hazelton et al., 2021].

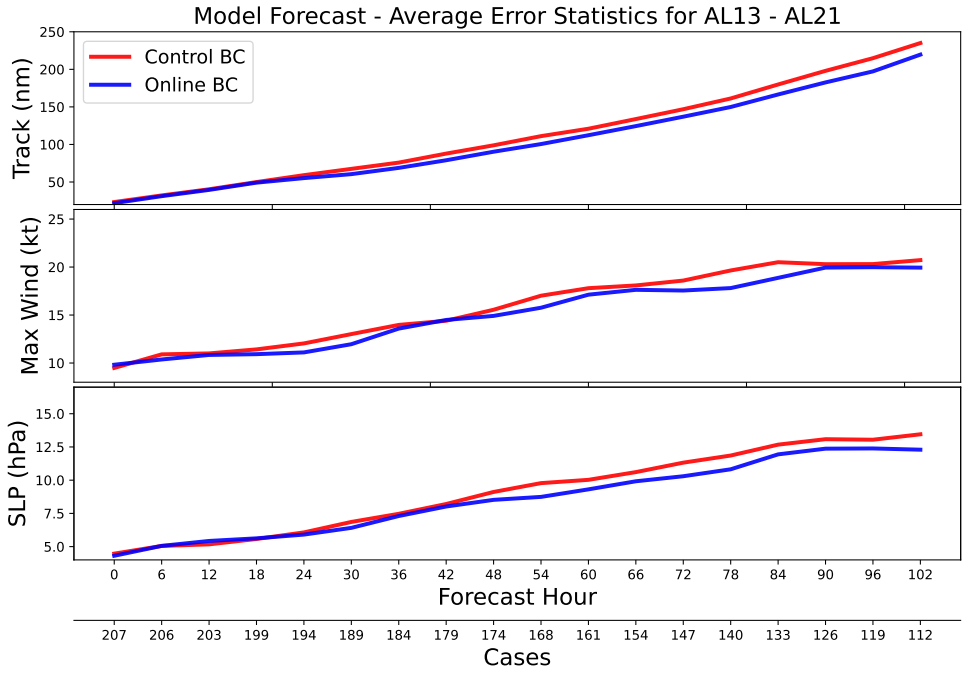
3.4 Results

The pair of experiments demonstrate substantial differences in domain-wide error characteristics as well as forecast track and intensity errors, which we describe in this section. Moreover, we will show that many of the differences can be easily interpreted by analyzing the diverging evolution of predictor weights. More precisely, changes in weighted predictor fields over the domain amount to explainable differences in model variable fields, such as temperature and humidity, by changing the bias applied to the observation space background. For this purpose, we demonstrate how the online configuration affects weighted predictor fields and drives spatially-coherent differences in thermodynamic variables between the experiments. From this analysis, we can determine how this configuration affects error characteristics and consequently improves TC forecasts. In general, we focus on variables that are directly related to design choices in our data assimilation system, choosing to omit variables that are more complex to attribute, such as storm surge and wave generation. Although these hazards are vitally important to consider when designing operational modeling systems, we limit the scope of this study to variables with simple interpretability.

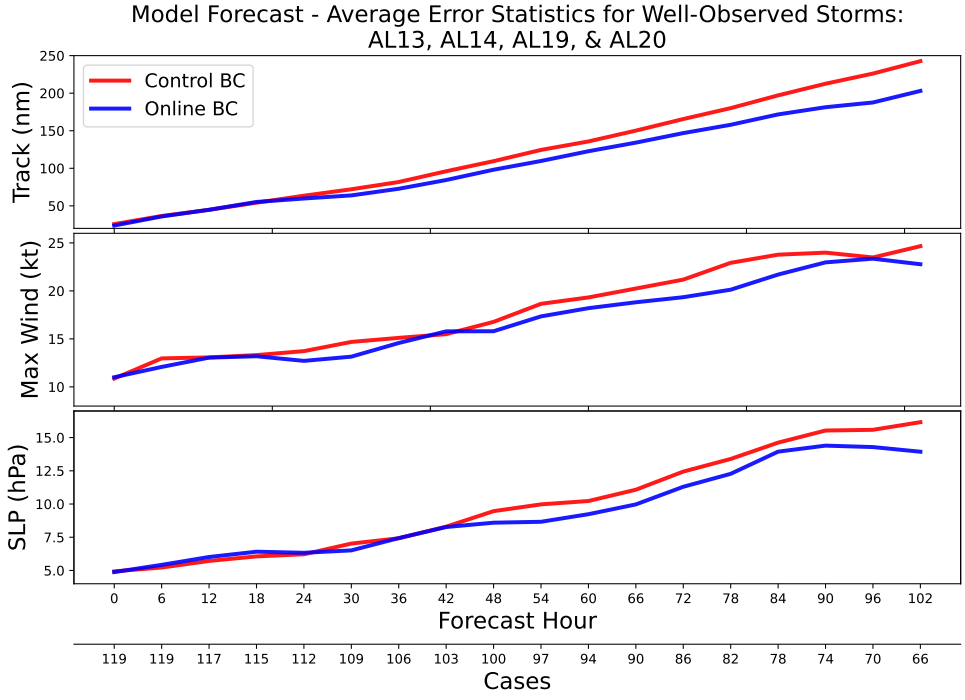
3.4.1 Forecast Error Verification

The following subsection provides a verification of TC forecast error statistics, specifically: track, 10-m maximum winds, and minimum sea level pressure (SLP). These statistics are summarized for all storms in Fig 3.2a, where forecast mean errors are calculated from a sample of storm forecasts over the 6-week experimental period and plotted as a function of forecast lead-time. The number of cases for each time is also given on the secondary axis.

For each variable analyzed, error statistics up to 18-24 hours are nearly identical across both



(a)



(b)

Figure 3.2: Forecast error verification of control and online experiments. Panel (a) reflects error for all TCs over the experimental period, while panel (b) omits TCs without aircraft reconnaissance data. These errors are verified against NHC best track and intensity data.

experiments, after which differences begin to emerge. In regards to track performance, the online configuration provides consistently better results after 18 hours of lead time, while the intensity forecasts provide generally improved results over the same range.

Several previous studies have demonstrated that considerable forecast improvements can be obtained from the assimilation of flight level radar and dropsonde observations from reconnaissance flight missions [Aksoy et al., 2012, Hartman et al., 2021, Lu et al., 2017, Sippel et al., 2022, Zhang et al., 2009]. Specifically, the assimilation of inner-core observations has been shown to significantly improve intensity forecasts even without the help of ad-hoc vortex relocation techniques. For this reason, we further stratify the results to include only storms that had multiple reconnaissance flight missions requested by the National Hurricane Center, hereby referred to as "well-observed" storms (NHC; Fig. 3.2b). These storms are Hurricanes Laura, Marco, Sally, and Teddy. Since land-threatening storms are the most likely to receive NHC reconnaissance flight authorization, the accurate forecasting of these storms is particularly important for NWP centers.

Unsurprisingly, the forecast error statistics for "well-observed" storms are quite similar to the cumulative results for all storms, but a few key differences stand out. In regards to track forecasts, both experiments exhibit similar errors up to about 24 hours, after which the online configuration exhibits markedly better performance. The intensity forecast results are mixed up until 42 hours, after which the online experiment reports consistently lower errors. Note that the total intensity errors for these well-observed storms are generally higher than for the all-storms data set; this is because these well-observed storms are, on average, more intense than the excluded storms, leading to proportionally larger errors.

It is likely that the similar performance and relatively low intensity errors of these experiments at early lead times is a reflection of the influence of reconnaissance observations. The divergence of

these configurations and subsequent superiority of the online experiment is perhaps a sign that the HAFS BC configuration benefits significantly from the online clear-sky satellite bias correction when combined with valuable inner-core observations. Since the assimilated clear-sky radiances are unlikely to include measurements near the TC vortex due to extensive cloud cover, it instead seems likely the clear-sky radiance bias correction results in an improved characterization of the surrounding synoptic environment while these inner-core observations "fill in the gaps" near the storm center. Furthermore, if these inner-core observations can more precisely place the storm vortex location within the domain, and large-scale features are more accurate as a result of the online bias correction scheme, improved intensity forecasts will follow as a result of a more accurate environmental characterization [Aksoy et al., 2012, Hartman et al., 2021, Lu et al., 2017].

As a final verification of forecast performance and variability, we summarize model forecast differences between the control and online experiment with Fig. 3.3. We examine the same error statistics as above but take the difference in absolute errors between experiments at each lead time and plot the median as well as inner and outer quintiles. This test essentially shows us the frequency in which one experiment outperforms the other at each lead time. In regards to track performance, the online experiment generally provides improved error statistics more frequently than the control after 24 hours lead time, as indicated by the median and inner quintiles. Conversely, the median and inner quintiles very consistently hover around zero for max wind and SLP. Comparing these results to Fig. 3.2a would suggest that while the frequency of improved error statistics is relatively the same across experiments, the errors themselves tend to be smaller for the online BC experiment. We note that we performed a similar analysis for "well-observed" storms only (not shown) and the results were marginally improved in expected ways; i.e., all quintiles are slightly elevated in a manner that would be expected based on Fig. 3.2b.

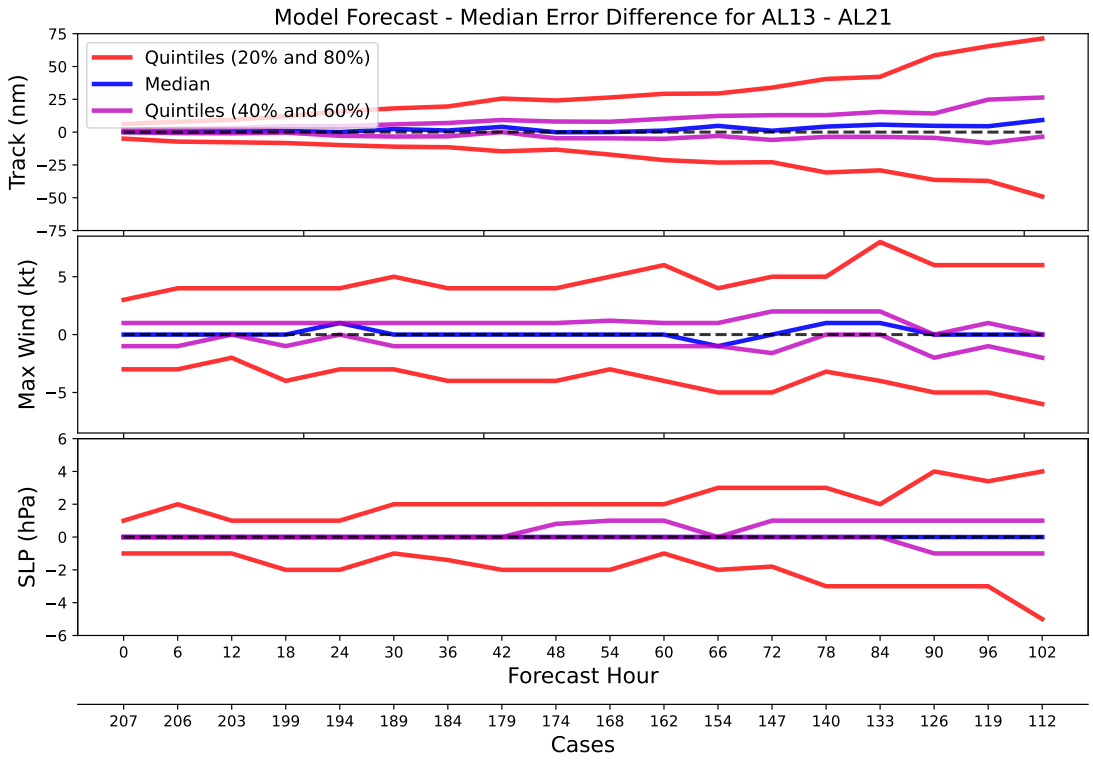


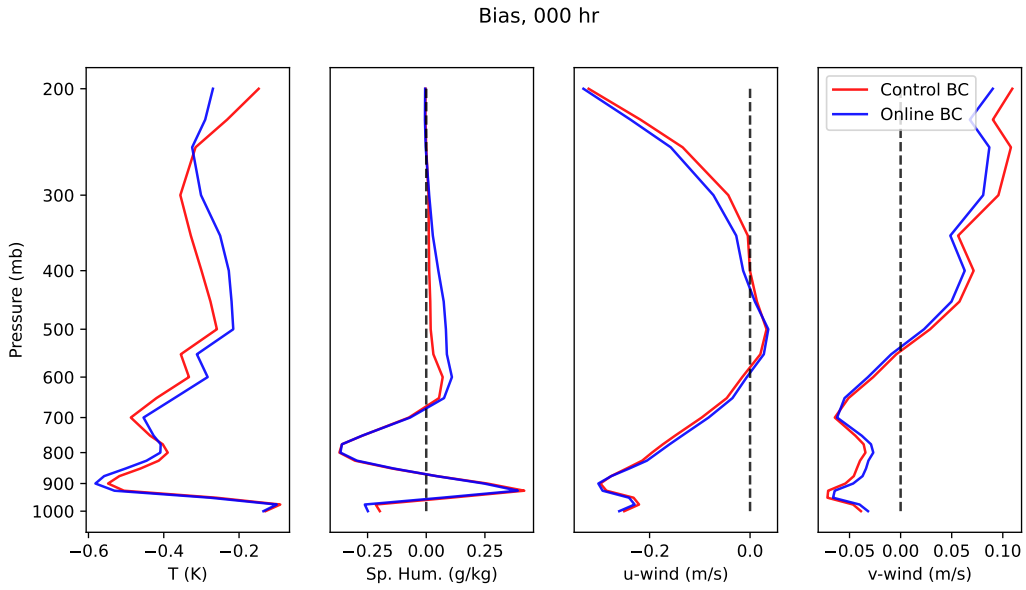
Figure 3.3: Error verification of forecast absolute error differences between control and online experiments (online-control). The median difference is given in blue, while the inner and outer quintiles are given in purple and red, respectively. Errors are verified against NHC best track and intensity data.

3.4.2 Domain-Averaged Error Verification

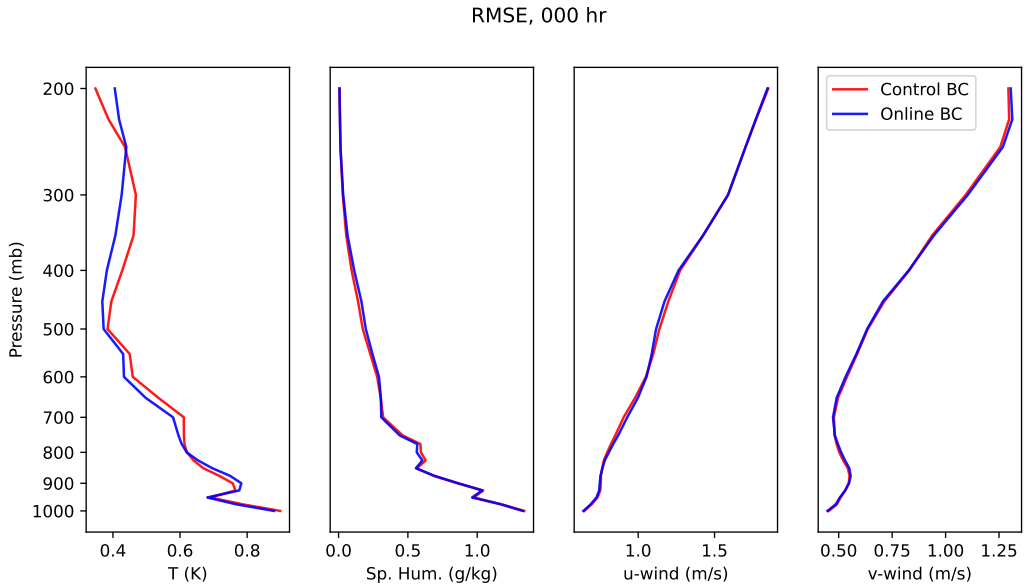
Next, we compare domain-wide variables, such as temperature, humidity, and wind speeds. We verify the data vs. ECMWF Reanalysis v5 (ERA5) data, which uses 4DVar data assimilation and comprehensive observation data sets including all-sky radiances; it is also relatively high-resolution compared to other global reanalysis data sets [Hersbach et al., 2020].

Fig. 3.4 displays vertical profiles of horizontally averaged bias and root mean square error (RMSE) values for pertinent variables, i.e. temperature, specific humidity, and u- and v- winds, for both control and online experiments at 00 UTC analysis time. Note that the investigated domain is slightly restricted compared to the model analysis domain (see Fig. 3.1) in order to minimize the effects of boundary conditions. In addition, the vertical profiles are capped at 200 mb, above which the effects of the model-top "sponge layer" produce unpredictable and difficult-to-characterize results.

The most prominent differences between control and online experiment error profiles can be found in temperature and humidity profiles, which is unsurprising given that these variables share a strong error dependence with BT. Both temperature bias profiles in Fig. 3.4a reflect a domain-wide negative-temperature bias present in this configuration of HAFS, particularly in the lower-troposphere, however, there are marked differences in temperature bias and RMSE profiles in the mid- and upper-troposphere. The online experiment exhibits a slightly positive temperature bias from 950 to 750 mb, a more negative bias from 750 to 250 mb, and a significantly more positive bias above 250 mb. The specific humidity bias profile is largely similar between the two experiments, however, the control configuration exhibits slightly more negative biases at the surface and through the mid-troposphere, from approximately 650 to 300 mb. Results are mixed for u- and v-winds, with only very slight differences in RMSE profiles as shown in Fig. 3.4b. We note that since these statistics are domain-averaged, regional biases of opposing



(a)



(b)

Figure 3.4: Panels (a) and (b) display bias and RMSE verification of temperature, specific humidity, u- and v- winds with ERA5 data, respectively. Data is time- and domain-averaged and displayed as a vertical profile.

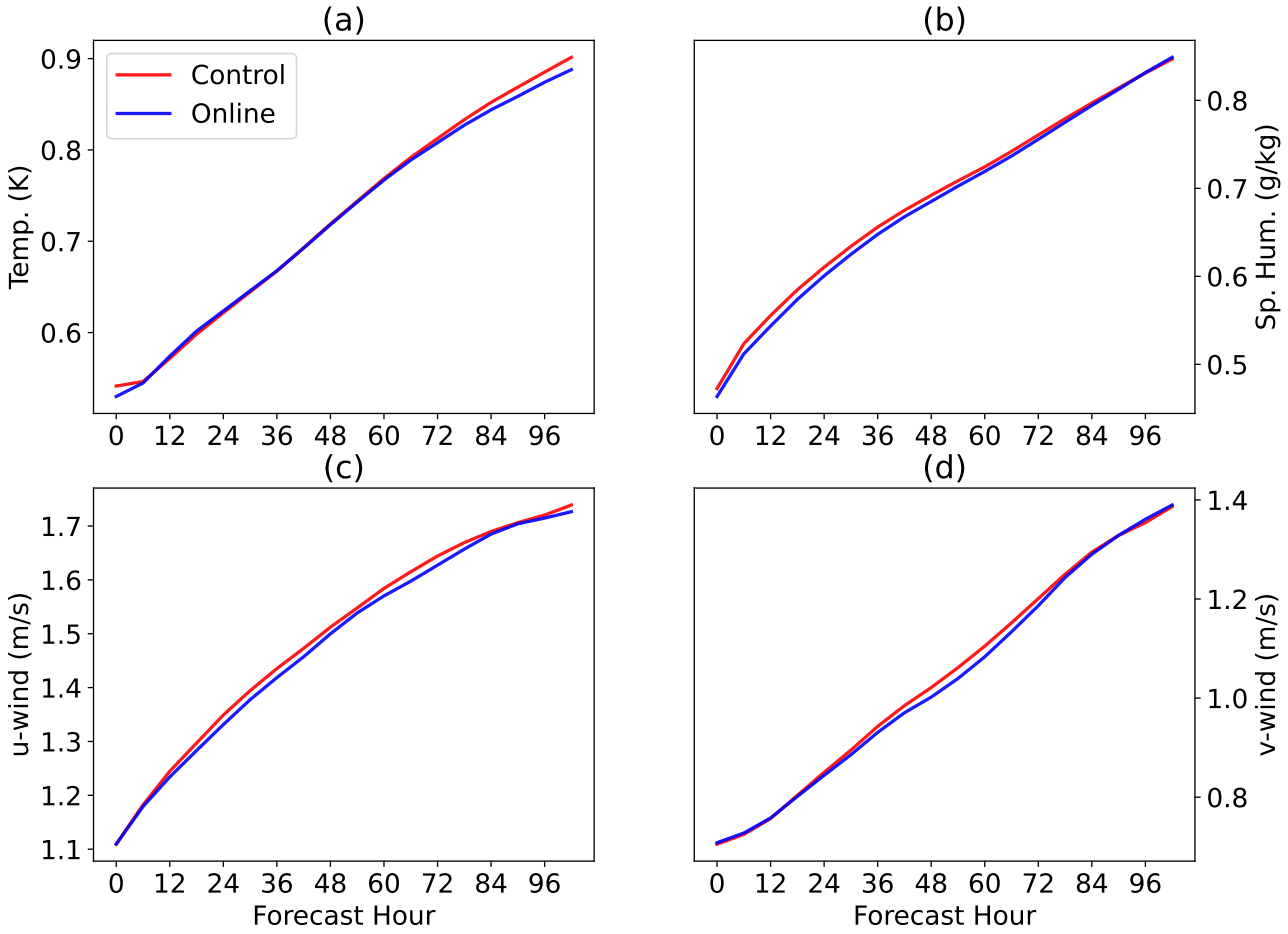


Figure 3.5: Temperature, specific humidity, u- and v-winds RMSE verification with ERA5 data. Data is domain- and column-averaged, and taken at 6-hour lead-time increments up to 102 hours. Columns span vertically from 900 to 250 mb for temperature and wind errors, and from 900 to 500 mb for specific humidity.

sign can cancel each other out, a fact that is explored further in the next section.

We also examine the temperature, specific humidity, u- and v- wind RMSE profiles across forecast lead times, where the online configuration tends to exhibit more substantial improvements in error profiles. Figure 3.5 shows time series of horizontal- and vertical column averages of RMSE for these variables taken from 0 to 102 hours. Like the calculations for Fig. 3.4, the verification domain is restricted to a subset of points within the total HAFS domain (see Fig. 3.1). Each variable is column averaged from 900 to 250 mb so as to minimize the influence of surface and model top boundary

conditions, except for specific humidity which was averaged up to 500 mb given the small mixing ratios in the upper troposphere. Temperature errors show notable improvements over early and late lead times, albeit with little difference between control and online from 6 to 84 hours lead time. Conversely, across specific humidity, u- and v- winds, the difference between control and online is small at early lead times but consistently demonstrates improved error statistics at longer lead times, beyond 6-18 hours. These improvements mostly trail off at lead times beyond 84 hours. The improved domain-wide error statistics for the online configuration at longer lead times shown in Fig. 3.5 mirrors the improvements we see in storm track and intensity errors at longer lead times, as referenced in Fig. 3.2.

3.4.3 Temperature and Humidity Fields

In this section, we discuss major differences in the temperature and humidity fields between the online and control experiment. We highlight temperature and humidity in our discussion, as profiles of these variables are the most directly related to radiance. Furthermore, we emphasize discrepancies in experiments at different height layers and horizontally across the domain, particularly over-land vs. over-ocean. In doing so, we illustrate how the online bias correction strategy directly influences our state estimates, which in turn explain different forecast results between experiments.

In figure 3.6, we examine vertical profiles of domain- and time-averaged differences in temperature and specific humidity between online and control experiments. We also differentiate between total, over-land, and over-sea differences. Given that the domain is mostly ocean, it is unsurprising that total difference profiles for both temperature and humidity are quite close to their respective profiles over the ocean. However, stark differences exist over land versus ocean, which can be clearly distinguished in Fig. 3.7 and 3.8.

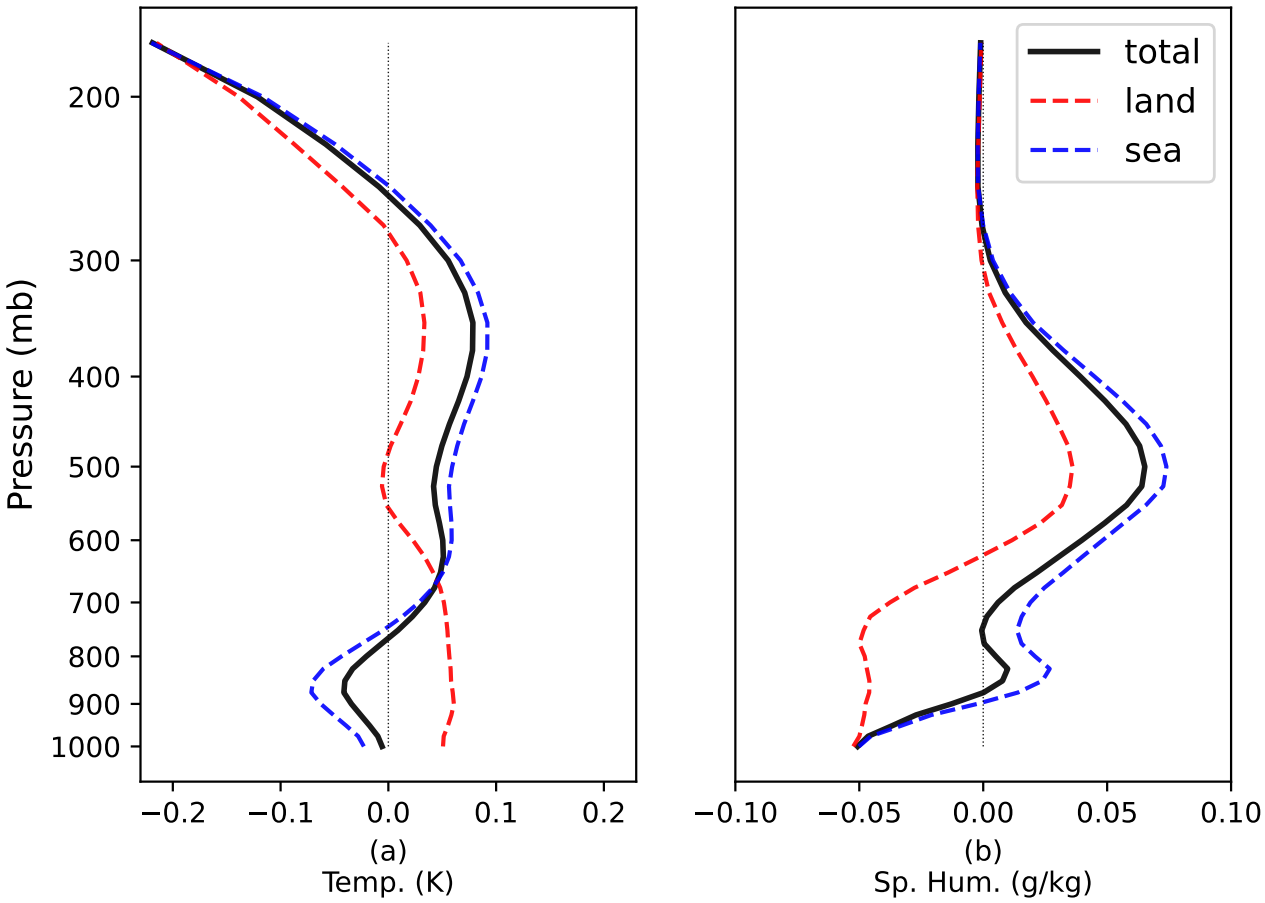


Figure 3.6: Vertical profiles of domain- and time-averaged temperature (a) and specific humidity (b) differences between the online and control experiment (online - control). Differences over-land, over-sea, and total differences are all considered.

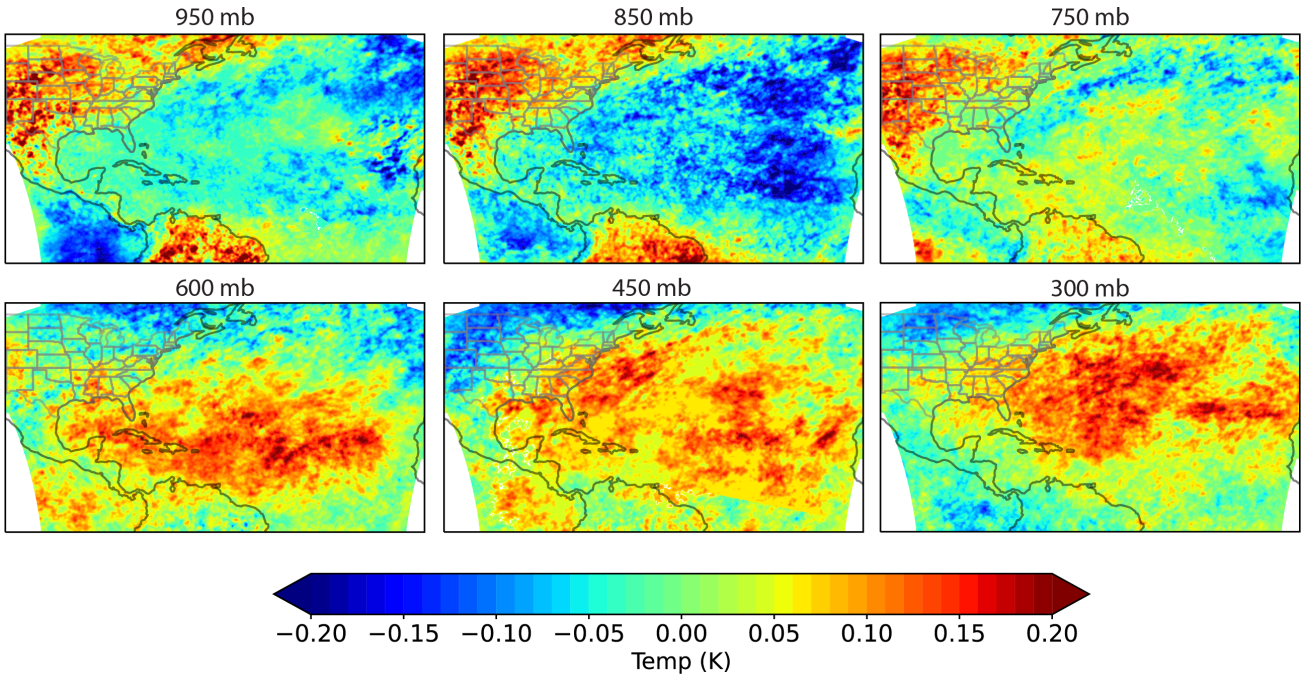


Figure 3.7: Time-averaged temperature differences between the online and control experiment at 950, 850, 750, 600, 450, and 300 mb (online - control). Note that a Gaussian smoothing filter has been applied.

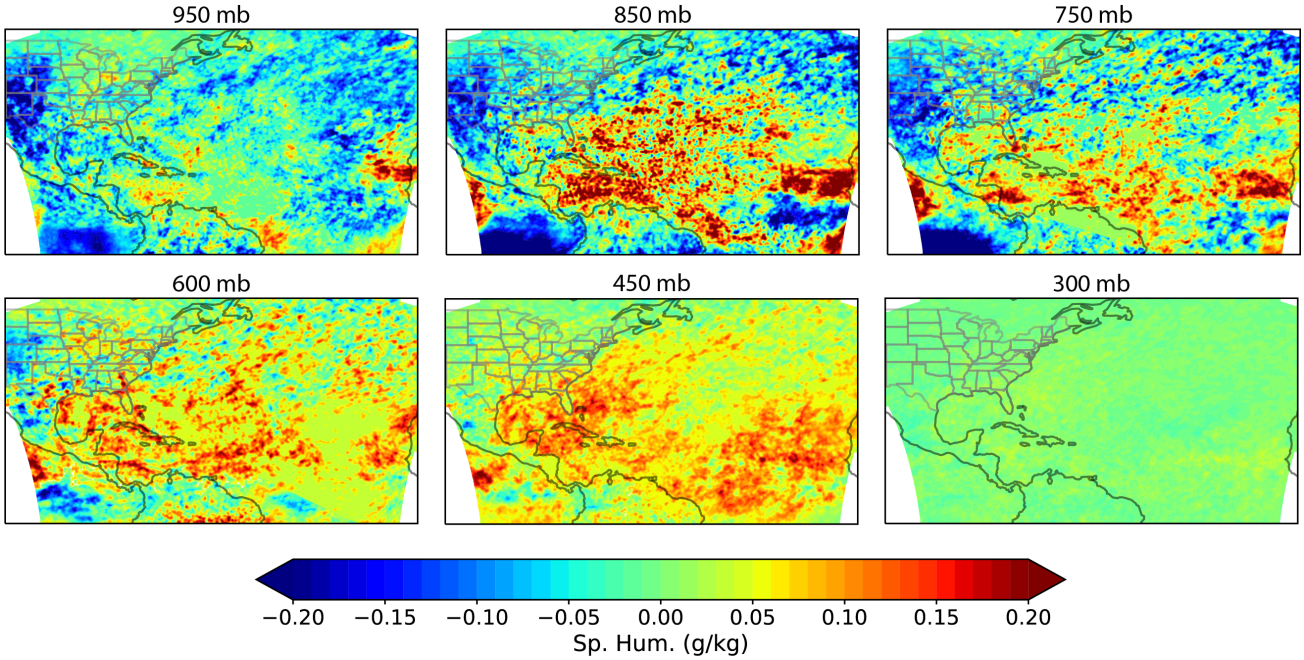
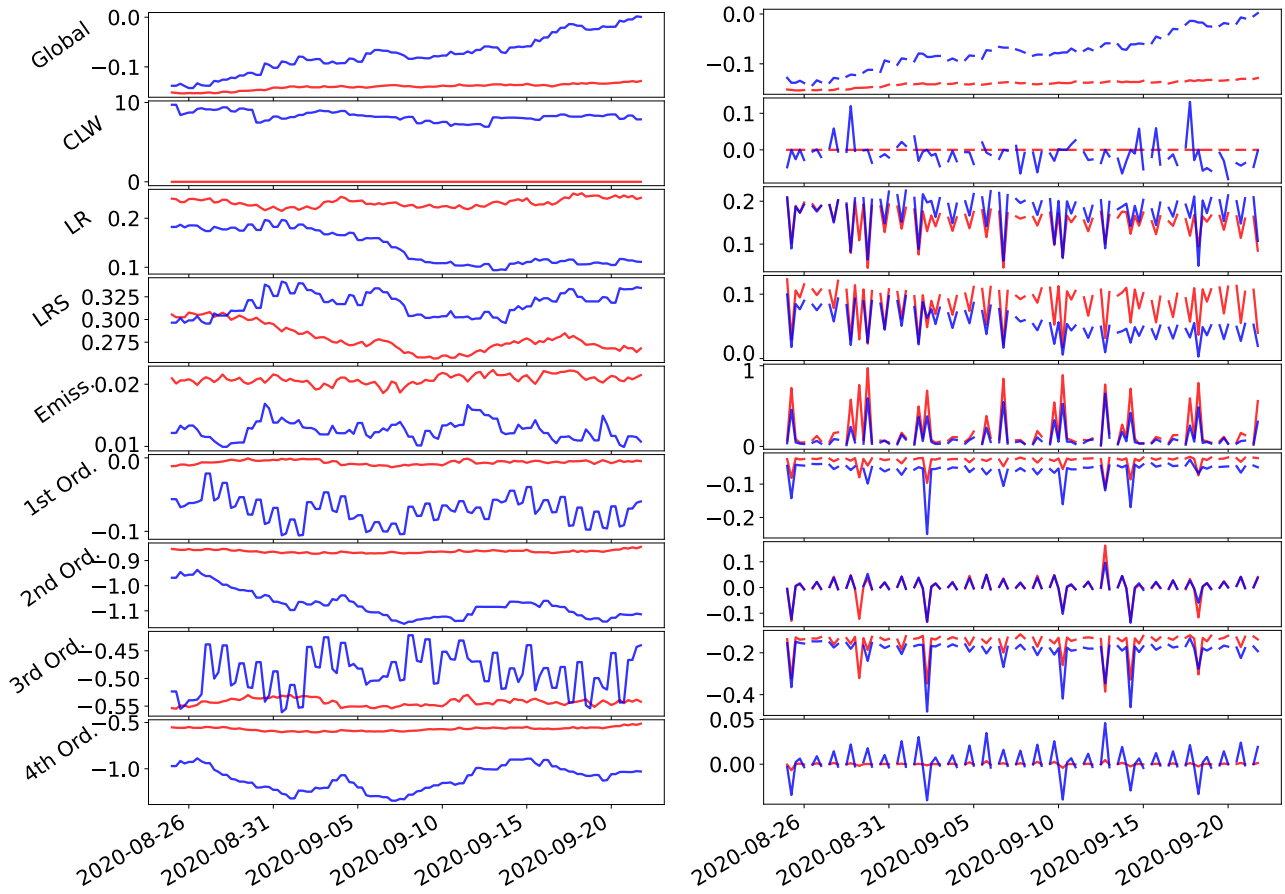


Figure 3.8: Time-averaged specific humidity differences between the online and control experiment at 950, 850, 750, 600, 450, and 300 mb (online - control). Likewise, a Gaussian smoothing filter has been applied.

Relative to the control, we find that online configuration produces cooler surface and lower-tropospheric temperature fields in total and over-sea locations, with a local minimum of around 850 mb, before becoming comparatively warmer into the mid-troposphere, with local maxima around 650 and 350 mb. The warmest areas exist over the central Atlantic, particularly where most storms track during our experiments, as shown in Fig. 3.7. Conversely, this configuration produces slightly warmer temperature fields over land throughout the lower troposphere, with local maxima of around 900 mb. Throughout the mid-troposphere, differences over land are slightly positive or near zero, although throughout western and northern North America these differences are strongly negative. Above 300 mb, all three profiles take a sharply negative turn.

Specific humidity exhibits less coherent layer-wise patterns, as shown in Fig. 3.8, but examining its domain-averaged vertical profile in Fig. 3.6 reveals some notable differences over-land vs. over-ocean. For example, while the online experiment is uniformly drier at the surface, the total and over-sea specific humidity difference is positive from 850 to 300 mb, with local maxima at 825 and 500 mb. Meanwhile, the over-land profile remains dry throughout the lower troposphere, with a local minima at 775 mb, before switching signs at 650 mb and remaining positive throughout the rest of the mid-troposphere, and reaching a local maximum at 525 mb. Above 300 mb, all three profiles approach zero, reflecting the low humidity in the upper troposphere.

As we will demonstrate in the next subsection, these differences over land and sea may arise from a number of factors, the most obvious of which are different radiative characteristics overland vs. water. In addition, high-quality in-situ observing systems are far more prevalent over land, which may lead to different bias characteristics. Alternatively, the analysis over-land is less dependant on satellite radiances and thus fewer differences exist between the two configurations, explaining the overall less dramatic difference in temperature over-land.



(a) (b)

Figure 3.9: (a) Bias predictor coefficient evolution over the course of the 5-week experiment period the GDAS BC (red) and HAFS BC (blue) for channel 4 of the AMSU-A N15 satellite instrument. (b) Weighted bias predictor value evolution over the same period, in units of BT (K).

3.4.4 Predictor Evolution and Verification

In order to understand why the online configuration results in improved forecast errors and notable differences in model fields, it is first necessary to evaluate how the online configuration affects the evolution of bias correction predictor coefficients as well as the weighted predictors themselves. It would be difficult to evaluate the individual effects of each and every instrument- and channel-specific predictor field, and so instead we narrow our analysis to specific channels with qualities that are

beneficial for our investigation.

With this in mind, we have chosen to focus on channel 4 of the Advanced Microwave Sounding Unit-A (AMSU-A) N15 satellite instrument. This channel measures radiance over a frequency of 52.8 GHz, has a weighting function that peaks near the surface, and has a relatively wide spread of weighted predictor values, making it convenient for our analysis [Deng et al., 2009, Goldberg et al., 2001, Liu et al., 2012, Zhu et al., 2016]. Furthermore, the AMSU-A instruments onboard NOAA polar-orbiting satellites are often indicated by ensemble and adjoint sensitivity analysis to yield large impacts on synoptic-scale forecasts [Ota et al., 2013]. Moreover, the instrument has a high volume of observations ingested after quality control and relatively large differences in channel-specific weighted predictor values between control and online experiments.

Figures 3.9a and 3.9b show the evolution of channel 4 AMSU-A N15 coefficients and weighted predictors over the 5 weeks of cycled DA analyzed for this study. Note that the magnitude of the unit-less coefficients in Fig. 3.9a vary widely and that these coefficients are multiplied by bias predictors whose magnitude varies considerably themselves. Thus, the values of the coefficients are not as significant as the weighted predictors themselves in 3.9b which have units of brightness temperature and therefore are comparable.

Instead, Fig. 3.9a is notable because it shows the rate at which the online coefficients depart from values estimated by GDAS over the course of our experiments. While some coefficients seem to reach a quasi-steady estimate in 1-3 weeks, others, such as the global offset coefficient, appear to diverge up until the end of the experiment. Figure 3.9a also suggests that the HAFS online coefficients are considerably more noisy than their GDAS counterparts. Given that these HAFS coefficients are derived over a limited-area domain and are influenced by boundary conditions, this behavior is not surprising. Also note that since the CLW predictor is not used by GDAS, the value of its predictor is exactly zero

for the entirety of the control experiment.

Figure 3.9b displays the evolution of the weighted predictors themselves, that is, the product of domain-wide predictor variables and the coefficient weights as featured in Fig. 3.9a. These values have units of BT (K), and combined constitute the total bias applied to the observation-space background as described in section 3.2. Note the strong diurnal pattern in weighted predictor value, and the relative similarity in values between experiments for the polynomial predictors. Instead, differences in the global offset, CLW, lapse rate, lapse rate squared, and emissivity predictors are predominate. Furthermore, much like in Fig. 3.9a, it can be seen that certain predictors begin to diverge 1-2 weeks into the experimental period, with some continuing to diverge even at the end of the 5-week period.

Although we have chosen to focus on channel 4 of the AMSU-A N15 satellite instrument, there are many more significant channels across multiple instruments that fulfill similar criteria. Figure 3.10 details a number of these channels. Each bar plot displays a different instrument's time- and domain-average weighted predictor value for all 9 predictors for the control and online experiment, as well as their difference.

Notice that for low-numbered channels, which tend to peak at or near the surface, the emissivity-weighted predictors tend to produce large disparities between the control and online configuration. Alternatively, for high numbered channels that peak in the upper-troposphere and stratosphere, the predictor of interest tends to be the lapse rate and lapse rate squared predictors. It is also noteworthy that the distributions of weighted predictor values across instruments are very similar, e.g. channel 11 has similar predictor values and differences across multiple instruments. This suggests that these differences in bias are dominated by model bias. We speculate that model physics may be the major contributor to such biases, but ultimately emphasize this as an example of the importance of eliminating outside model bias, such as model bias from GDAS.

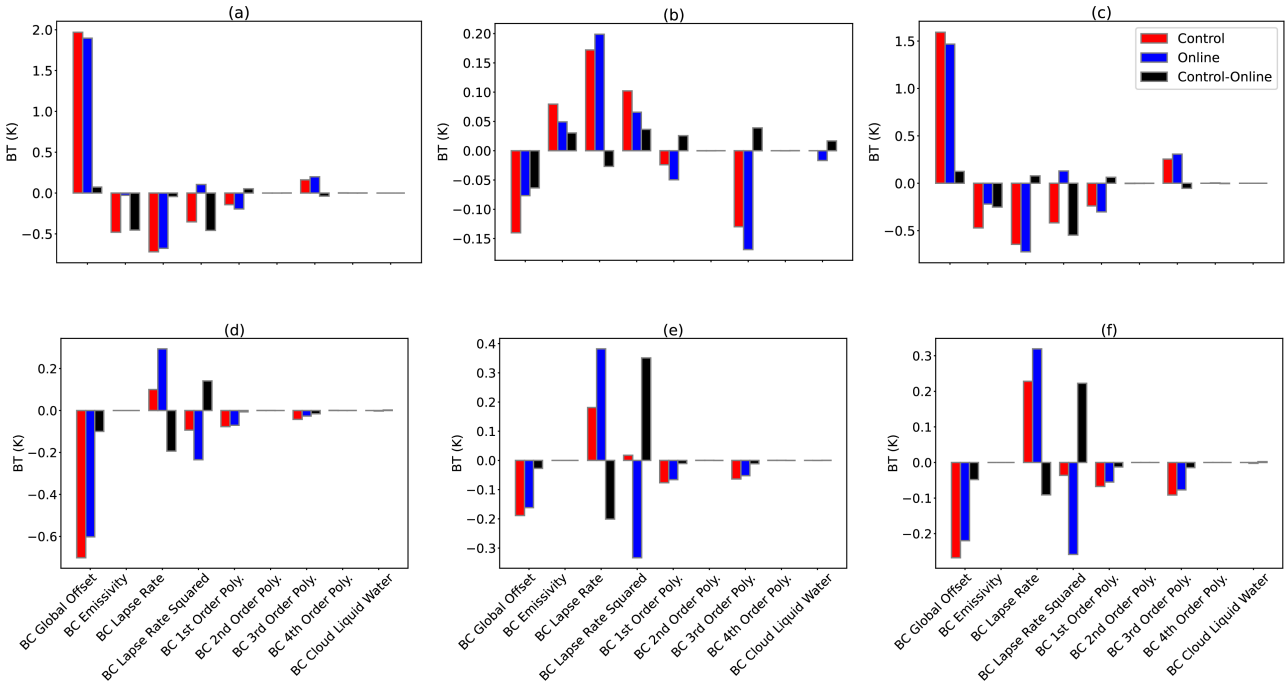


Figure 3.10: These panels detail time- and domain-averaged weighted predictor values for select instruments and channels for each experiment. The x-axis gives each weighted predictor, while the red, blue, and black colored bars give their time- and domain-averaged values for the control experiment, online experiment, and their difference, respectively. They are: (a) MHS MetOp-B, channel 1; (b) AMSU-A N15, channel 4; (c) MHS N19, channel 1; (d) AMSU-A N18, channel 11; (e) AMSU-A N18, channel 11; (f) AMSU-A MetOp-A, channel 11.

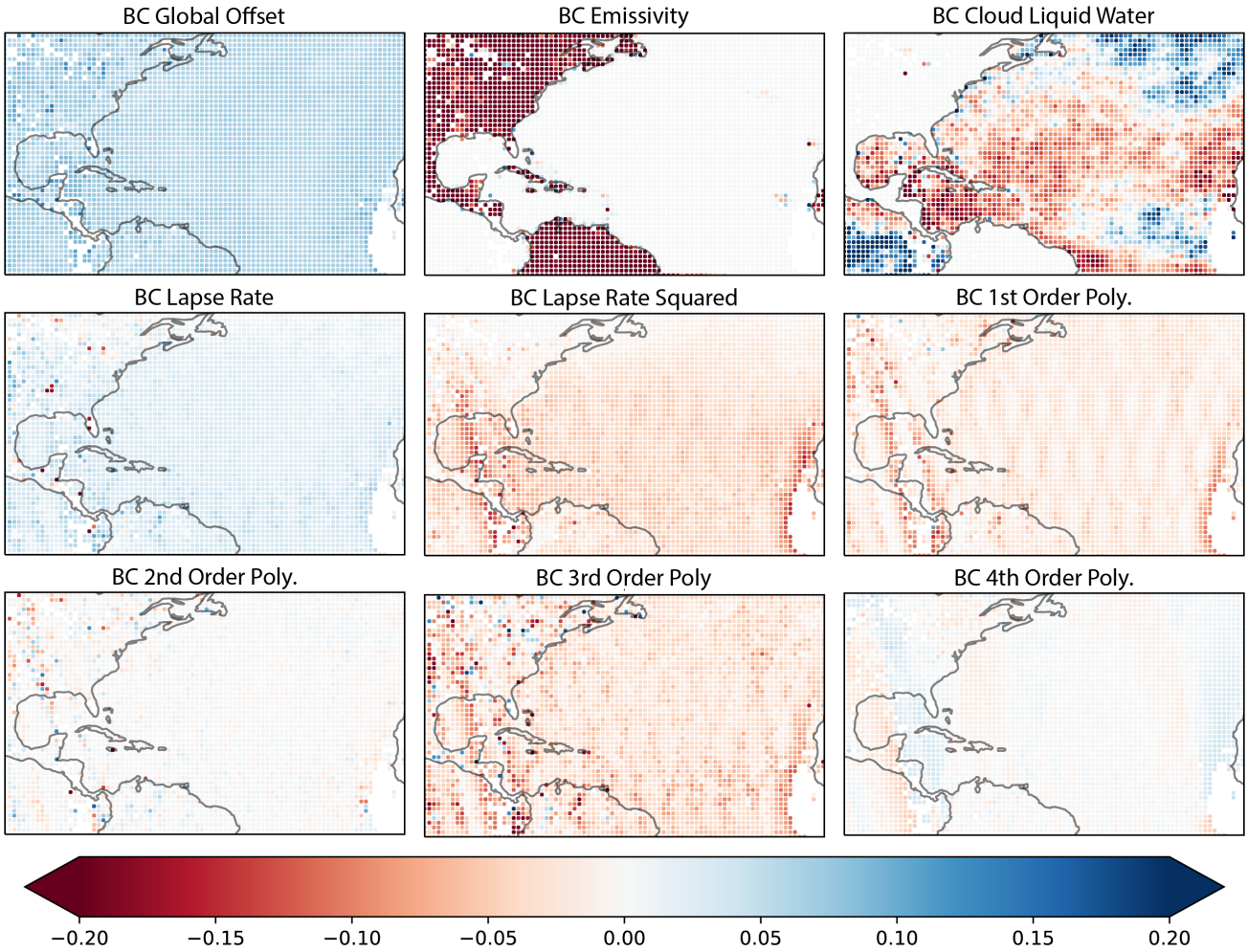


Figure 3.11: Time-averaged weighted predictor differences between online and control experiments (online-control) for channel 4 of the AMSU-A N15 instrument, in units of BT (K).

3.4.5 Weighted Predictor Fields

As an example of how the online varBC method may affect model fields, it is useful to examine the difference in the weighted predictor fields between experiments for one specific channel and instrument, and then compare these differences to differences in model field variables. For the same reasons mentioned earlier, we examine the difference in weighted predictor fields between the online and control experiment for channel 4 of AMSU-A on the NOAA-15 satellite.

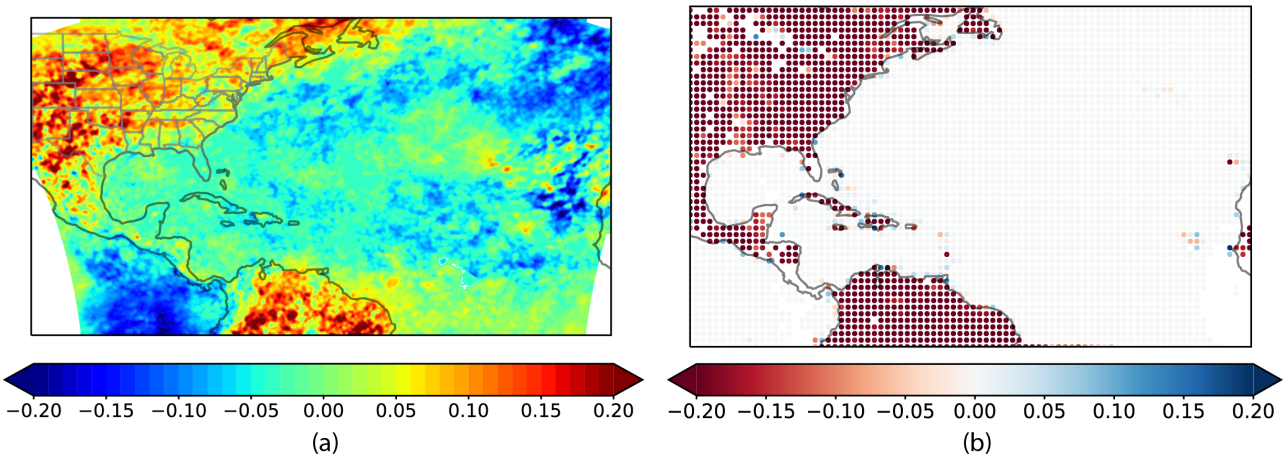


Figure 3.12: On the left, the time-average temperature field difference between control and online at 950 mb (online-control). On the right, the difference in the emissivity predictor between control and online for channel 4 on the AMSU-A N15 instrument. Note that the channel 4 weighting function peaks near the surface.

Figure 3.11 shows the time-averaged online-control differences in weighted predictor field contributions for each predictor for the AMSU-A channel 4. Although many of the weighted predictor fields show little coherence, a few display notable patterns. The global offset term tends to contribute to a uniform negative bias estimate across the entire domain, with a few exceptions that tend to congregate around coastlines, while the emissivity predictor is strongly positive across most of the North and South American landmass and mostly zero over-ocean. It is unsurprising that the emissivity predictor produces the starkest differences given that this channel peaks near the surface, and, as previously mentioned, this predictor tends to dominate for low-altitude peaking channels. The cloud liquid water predictor is conversely zero across most of the landmass, but exhibits varied patterns over water, having positive values over the Caribbean and mid-Atlantic basin vs. negative values in the northern Atlantic region and in the eastern Pacific. The lapse rate and lapse rate squared predictors tend to produce a quasi-uniform pattern, albeit with the magnitude of the predictors increasing slightly towards the equator, while the polynomial predictors display less coherence across the domain.

In Fig. 3.12, we consider the influence of a single predictor by comparing time-averaged online-

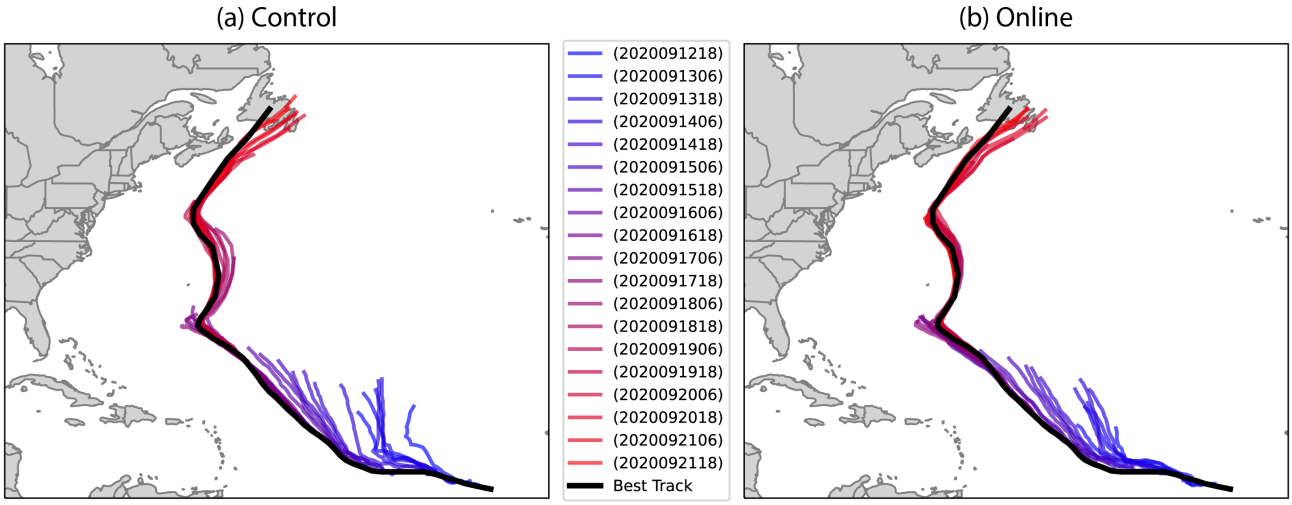


Figure 3.13: Track Error Verification, Control (a) vs Online (b), for AL20, Hurricane Teddy. The blue lines represent forecasts generated early in the model time span, transitioning to red as forecast time progresses. Black represents the best track estimate.

control temperature differences at 950 mb to weighted predictor field differences for the emissivity predictor. A correlation can be seen between the positive contribution of the emissivity predictor bias term and the warmer temperatures present over-land in the online experiment's analysis. Since channel 4 peaks near the surface, it can be assumed the bias term has directly or indirectly affected the model field, but given the magnitude of the bias term it is likely this relationship falls under the former category. Of course, it is difficult to fully attribute these temperature field changes without fully considering the effects of each predictor over every channel and instrument, a task that is outside the scope of this paper. Instead, we focus on the macro-effects of the self-cycling bias correction technique and its effects on TC track and intensity forecasting.

3.4.6 Case Study

To give an example of how the online varBC alters the track and intensity forecasts compared to the control experiment, we examine both configurations' forecasts of Hurricane Teddy. Teddy, the

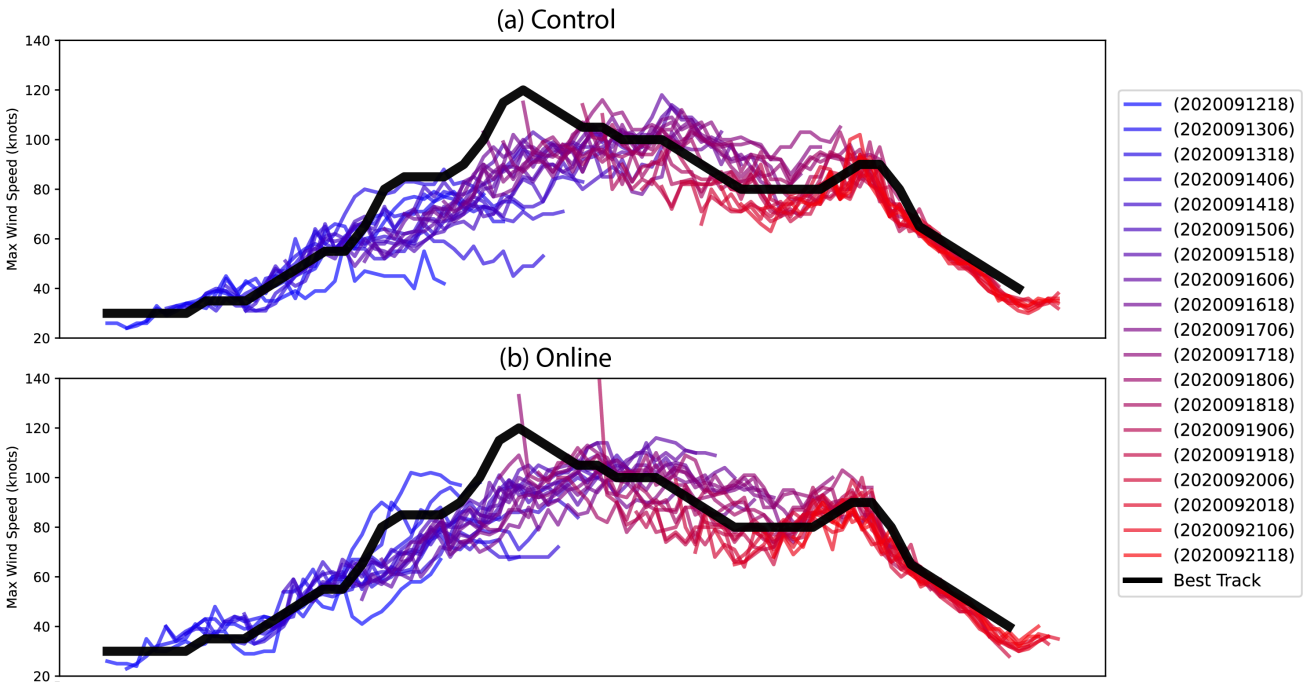


Figure 3.14: Same as Fig. 3.13 but for 10-m maximum wind speed.

19th TC of the 2020 season, was a Category 4 Cape Verde hurricane with a particularly long track, and multiple rapid changes in intensity during its lifetime. Teddy developed into a TC on Sept. 14th and a Category 1 hurricane on Sept. 16th, subsequently reaching its peak intensity on Sept. 17th. The storm weakened over the next few days before it re-intensified after interacting with a mid-latitude upper-level trough. It later weakened and dissipated after making landfall over Nova Scotia. Teddy's track was almost entirely over the observation-sparse ocean, and, due to its significant land-threat, featured multiple NHC reconnaissance mission data sets. It also featured some of the largest and strongest Atlantic hurricane wind fields on record [NOAA, 2021]. For these reasons, we determined Teddy would be an excellent choice to examine as a case study for our satellite radiance bias correction experiments.

Figure 3.13 shows Teddy's NHC best track estimate, along with the evolution of both the control and online experiments' track forecasts, and Fig. 3.14 displays similar information but for maximum

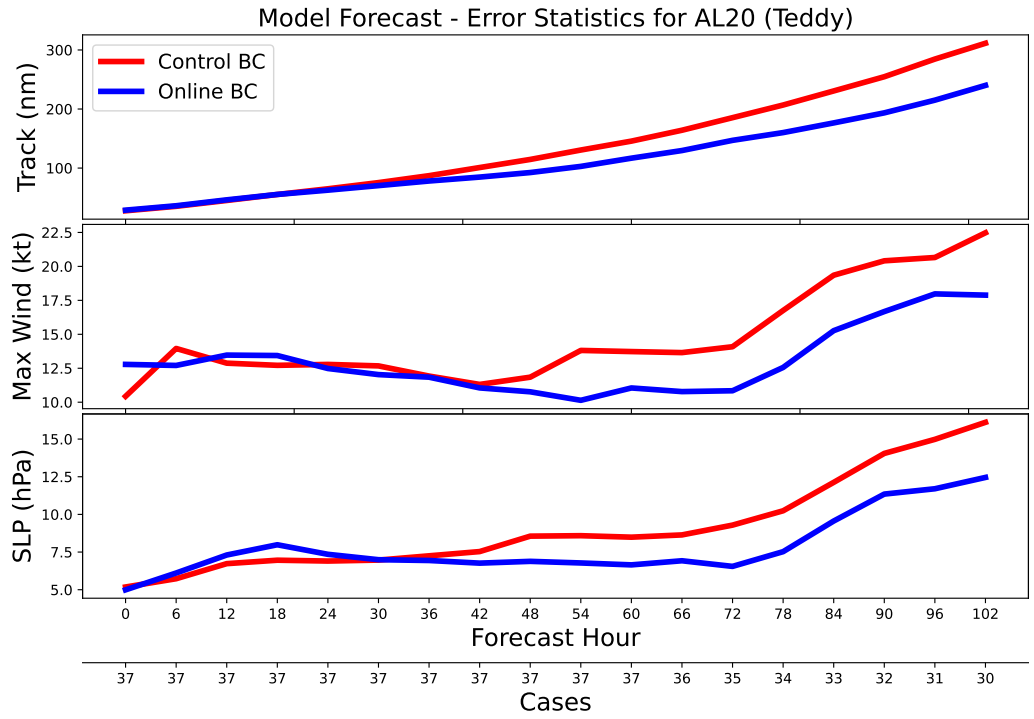


Figure 3.15: Forecast error verification of control and online experiments for AL20, Teddy. These errors are verified against NHC best track and intensity data.

wind speed error. The improved track forecasting skill of the online experiment, especially at longer forecast lead times, is evident. Furthermore, we find that the online configuration results in noticeably improved track and intensity errors around the initial TC intensification time (approx. 00 UTC Sept. 14), and likewise for track error during Teddy's re-intensification (approx. 12 UTC Sept. 22).

These conclusions are further supported in Fig. 3.15, where track and intensity error statistics are plotted as a function of lead time. Although the intensity forecast results are mixed at early lead times, the online configuration features substantially improved intensity forecasts after 42 hours.

In order to explain how the online configuration produces more accurate track and intensity forecasts, we provide a more detailed analysis of model field variables at these two pivotal moments in Teddy's evolution: initial intensification and re-intensification. At initial intensification, we find that the online configuration produces a TC vortex that is deeper, tighter, and intensifies more quickly than for

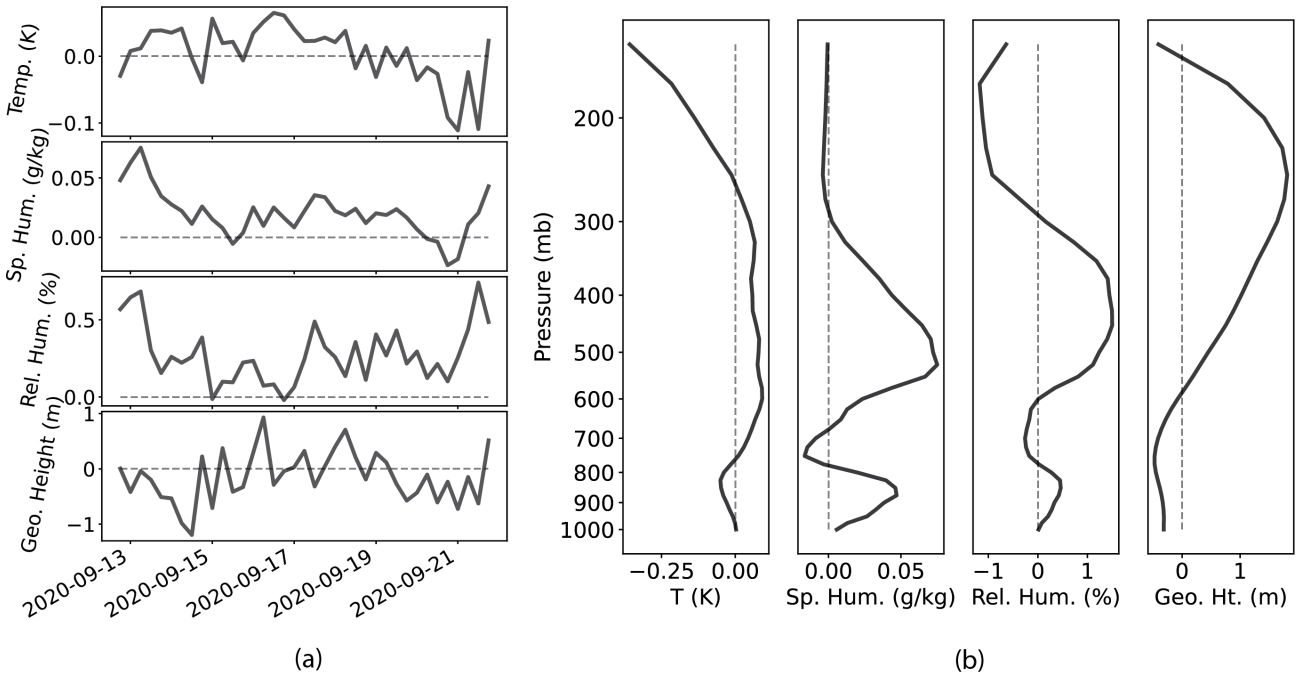


Figure 3.16: (a) displays a time series of model field differences between online and control (online-control) for 4 pertinent model variables: temperature, specific humidity, relative humidity, and geopotential height. The verification domain for these variables forms an annulus around the TC, extending from 200 to 2000 km around the TC. The values are column averaged from 950 to 500 mb. (b) displays a time-averaged vertical profile of the same variables over the same verification domain.

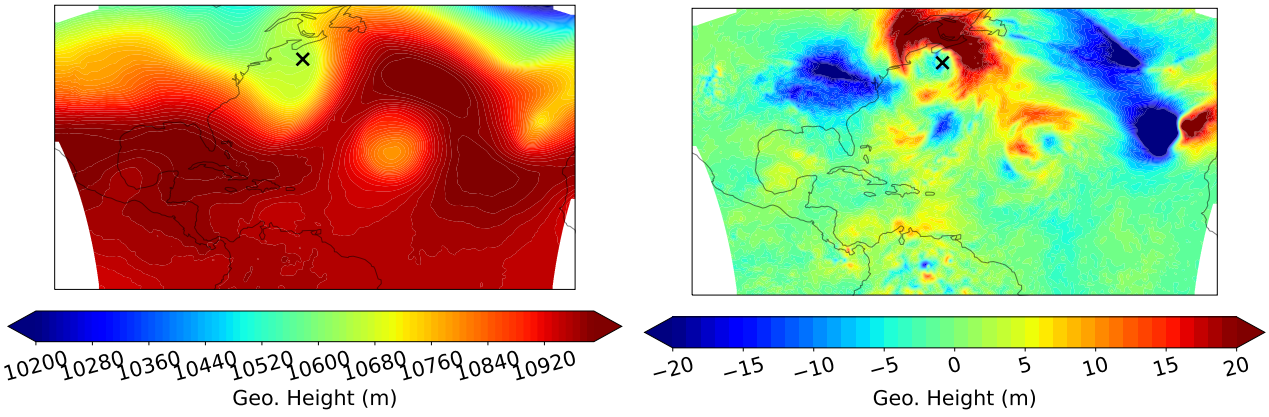


Figure 3.17: (a) displays the 048 hours lead-time 250 mb geopotential height model field generated at Sept. 21st 00z for the control experiment. (b) gives the geopotential height difference between online and control at the same forecast time, lead-time, and pressure level (online-control). In both cases, a Gaussian smoothing filter has been applied. A black "X" marks the NHC best estimate location for Hurricane Teddy at Sept. 23rd 00z.

the control experiment. We examine a special verification domain that forms a ring or annulus around the vortex, extending from 200 to 2000 km from the predicted vortex center. In Fig. 3.16a, we display a time series of model field differences between online and control (online-control) for 4 pertinent model variables: temperature, specific humidity, relative humidity, and geopotential height. Consequently, we find that the online configuration produces environmental conditions that are moister at low altitudes surrounding the storm, leading to conditions that are more conducive to rapid intensification.

Following Teddy's initial intensification and subsequent decay, the storm drifted northwards and interacted with an upper-level trough on Sept. 22, as seen in Fig. 3.17a. A secondary intensification occurred, resulting in a Category 2 hurricane, which was accompanied by an altered north-northeastward track, as seen in Fig. 3.13 and 3.14. Although both configurations provided skillful predictions for intensity at this point, the control produced track forecasts that extended too far east initially, while the online experiment produced much more accurate track forecasts. Analysis of the model fields at extended lead times (48+ hours) revealed that the control configuration predicted the upper-tropospheric trough to have propagated farther east than observed, as suggested in Fig. 3.13a., while the online configuration predicted a farther westward trough location, as can be seen in Fig. 3.17b. Due to this slightly more inaccurate environmental characterization from the control experiment, the forecasts initialized from these analyses brought the TC too far to the east.

To summarize, we find that online bias correction method produces improvements in track and intensity forecasting in this case, although we caution against the generalization of these results alone. Instead, we highlight that this case study demonstrates how a more accurate characterization of large-scale environment can lead to improved forecast results, particularly when coupled with inner-core observational data sets. Combined with the forecast error verification over a range of TCs presented in section 3.4.1, particularly those with ample inner-core observations, we hypothesize that these results

could extend to other TC forecasts as well.

3.5 Discussion and Conclusions

For many limited-area models, the decision to operate on a partially-cycling basis often comes down to concerns of model drift and an insufficient sampling of innovations for estimating bias for satellite radiance measurements. For modeling systems that are able to avoid these challenges, the ability to run continuous, fully cycling, domain-wide DA can prove extremely valuable. One such benefit comes from the ability to self-cycle information such as innovations and model fields needed to estimate key parameters for specifying measurement bias, namely predictor weights or "coefficients" that are used for statistical bias models. The decision for regional models to adopt bias correction coefficients from global models is usually one of practicality, given that partially-cycling systems are unable to adopt an online configuration. However, even in cases where the system is fully-cycling, these coefficients are still often adopted from lower-resolution global models. This decision is made because regional domain sizes are assumed to be too small, and global model data benefits from more thorough model tuning that could produce better bias correction configurations.

The current study investigates the potential of developing a fully-cycled prediction system for the NOAA Hurricane Analysis and Forecast System (HAFS) that performs its own self-contained bias correction scheme for satellite radiance measurements. A number of key conclusions can be drawn from comparisons of a fully cycled HAFS system with and without this feature. First, introducing online estimates of bias correction parameters can produce improved TC forecasts, even in limited-area models such as HAFS. We postulate that the fully online bias correction leads to improved characterization of the surrounding synoptic environment during DA, leading to more accurate steering flow associated

with storms. In particular, we find the impact of online bias correction to be most noticeable for storms with extensive inner-core observations, as these observations can more precisely characterize TC vortex location within the domain by ”filling in the gaps” near the storm center. We explore this finding in more detail for a case study of Hurricane Teddy, where the online configuration produced an environment that was more conducive to rapid intensification and consequently did a better job of forecasting Teddy's initial intensification and development. Likewise, Teddy's later interaction with an upper-level trough and the resulting re-intensification was more faithfully represented by the online experiment.

We also establish that fully online satellite bias correction can produce lower domain-averaged errors for key variables such as temperature and humidity, although these relationships are highly dependant on qualities such as lead time and height level and thus are arguably less robust than the TC-focused error statistics presented is section 4a.

Furthermore, we can draw some conclusions about major differences in mean bias over land vs. over sea, and how these differences can be tied to how biases are specified in satellite radiance measurements. For example, we find that the online method generally produces warmer and wetter model states throughout the mid-atmosphere both over land and over sea, but that the relationship is more complicated near the surface. Clearly this is a reflection of different radiative characteristics of land vs. sea, and we demonstrate that these differences are a direct result of weighted predictor field specification. We also note that these differences in mean bias are largest over sea, where in-situ measurements are sparse and accurate specification of satellite radiance measurements becomes particularly important.

This study also shows that modeling systems that ingest only clear-sky radiances should not adopt configurations that are tuned for ingesting all-sky radiances. In this case, the GDAS model has tuned its bias predictor coefficient values towards all-sky radiances, e.g. setting the cloud liquid water predictor

to zero across the entire domain. This is problematic for the control HAFS system considering said system only assimilates clear-sky radiances, and thus would benefit from the cloud liquid water bias correction term. A similar conclusion can be drawn for models that borrow bias information, among other things, from global models with different model tops. HAFS has a maximum pressure level of 2 hPa, whereas GDAS extends well up into the upper stratosphere. This disparity results in a strong difference in model variables, such as temperature, at high altitudes. It also results in peculiar and unexplainable values for weighted predictors in channels that peak at high altitudes, particularly the global offset, lapse rate, and lapse rate squared predictors. These disparities make it difficult to assess the effectiveness of these bias predictors for these channels.

We also note that experiments performed in the current study adopt a single domain with 6-km grid spacing, which differs from the operational configuration of HAFS that uses 2-km nest. While our results demonstrate that adopting a fully online bias correction strategy can produce notable differences in forecast skill, particularly mean errors, we speculate that these differences may be even larger if our study adopted the same convective-permitting 2-km grid spacing that is used for the operational HAFS. Doing so will further reduce any similarities in model process error that exist between resolved scales and choices of physical parameterization schemes between regional and global configurations of FV3 (HAFS vs. GFS), and further exacerbate the issues raised in this study for quantifying bias in radiance measurements.

In concluding, we also note several directions in which this research could be continued or expanded. Given that several predictor coefficients and weighted predictor differences continue to diverge even at the end of the 6 weeks of cycling DA, it may be useful to extend this study for entire hurricane seasons. It would also be beneficial to reproduce the online bias correction experiment starting with zero values for coefficients, as opposed to starting with coefficients adopted from GDAS. In this

way, we could eliminate bias from the GDAS model more quickly, and it may further improve forecasts at earlier cycles in the experiment period. Eventually, this research may find value for the expansion of radiance assimilation from just clear-sky assimilation to all-sky assimilation. Furthermore, there may be a need to explore alternate additional predictors specific to the TC problem. Lastly, future research will focus on additional requirements needed to maintain a fully-cycled ensemble DA system for operational hurricane prediction. This effort parallels similar efforts underway for other regional modeling applications, such as severe convective storms [e.g., Schwartz et al., 2022], and sets the stage for the basin-scale ensemble framework introduced in the following chapter.

Chapter 4: Obstacles for High-Resolution, Basin-Scale Probabilistic Hurricane Forecasting

4.1 Introduction

Tropical Cyclone (TC) focused numerical weather prediction (NWP) systems have made significant improvements in recent years, in large part due to contributions from the Hurricane Forecast Improvement Project (HFIP). Hurricane track errors have been greatly reduced, and intensity errors have seen recent improvements with additional in-situ and satellite observation systems; improvements to model dynamics and physics; land, aerosol, ocean, and wave coupling; and the integration of advanced data assimilation and machine learning methodologies [Cangialosi et al., 2020, Hazelton et al., 2023, Peng et al., 2024, Worthen et al., 2025]. In addition, model developers are increasingly paying closer attention to non-traditional forecast metrics and associated hazards, such as wind radii, storm surge, and rainfall [Cangialosi et al., 2020, Hazelton et al., 2023].

Nevertheless, TC NWP modeling systems continue to struggle with a number of unique challenges. TCs straddle the line between synoptic and mesoscale weather systems, and thus require modeling systems that are both high-resolution and can accurately represent large-scale atmospheric motions over the oceans, where in situ measurements are sparse [Knisely and Poterjoy, 2023]. TC modeling systems require planetary boundary layer (PBL) and microphysics schemes that are specifically tuned to handle intense fluxes and hydrometeor transport around the inner-core while still taking into account multiscale interactions [Hazelton et al., 2024, Wang et al., 2024]. Furthermore, the accurate simulation of inner-core features is difficult due to gaps in scientific knowledge, and thus requires highly accurate, well-placed measurements such as dropsondes and Tail Doppler Radar (TDR) observations [Rogers et al., 2013].

Another major issue with NOAA's modern TC-forecasting efforts is the lack of an "analysis of record" or reanalysis of past events. Without a comprehensive, basin-wide analysis of record, it is significantly more difficult to verify the performance of operational forecasting systems [Torn and Snyder]. For example, while we can verify some TC-specific metrics, such as track and max wind speed, it is not possible to verify how these models simulate broader environmental conditions across the entire Atlantic basin and their associated uncertainties, which is imperative to understand in order to properly predict cyclone genesis and rapid intensification (RI) [Kaplan and DeMaria, 2003]. Additionally, the increasing prevalence of data-driven machine learning (ML)-based models and parameterizations highlights the need for comprehensive high-quality training datasets [Bouallègue et al., 2024]. Building an ideal analysis of record would require a self-contained model configuration, continuously cycled, ensemble-based for flow-dependent error covariances, and a large domain that covers entire basins [Poterjoy et al., 2021].

The operational Hurricane Analysis and Forecast System (HAFS) system was first introduced for real-time experiments in 2022 [Hazelton et al., 2023]. The system features two main model configurations, mainly differentiated by their choice of physics suites: HAFS-A and HAFS-B. While the system runs very quickly and produces relatively accurate model forecasts, it also suffers from a number of potential deficiencies that threaten a hard ceiling in terms of forecast performance improvement. HAFS is only initiated when a storm is present and only cycles model states within a relatively small region that encompasses the core of the storm itself, primarily relying on initial conditions from the Global Data Assimilation System (GDAS) elsewhere, and uses ad-hoc methodologies such as vortex initialization [Hazelton et al., 2023, Poyer et al., 2025, Ramstrom et al., 2024]. These GDAS-provided ICs are much lower in resolution (13 km), and the fact that the model is not continuously cycled means that it cannot benefit from methodologies such as multi-scale flow-dependent error covariances—

between variables in the storm core and environment—and native domain-trained bias estimations for observations. As previously stated, these limitations restrict our ability to establish an analysis of record.

Modern TC forecast systems typically have a forecast resolution of 2-6 km, with deterministic and/or ensemble analyses of the same or slightly lesser resolution, compared to global modeling systems in the U.S. such as the Global Forecast System (GFS) and Global Ensemble Forecast System (GEFS), which have horizontal resolutions of 13 km, and 26 km, respectively. This resolution gap is quite significant, as high-resolution models such as HAFS can resolve (or partially resolve) complex physical processes that are typically parameterized at coarser resolutions [Wang et al., 2024]. While HAFS often adopts similar physics suites as the GFS, these schemes are typically modified or tuned for the small-domain around the storm itself to produce the most accurate predictions of TC intensity metrics, such as max wind speed [Zhang et al., 2015]. However, tuning these physics parameters for the environment immediately surrounding the storms may not be ideal for the remainder of the basin [Marks et al., 1998]. Thus, producing a self-contained modeling configuration that covers the entire Atlantic basin—while maintaining resolutions comparable to high-resolution nests—will require careful consideration of how physics are parameterized across the domain. Additionally, explicitly resolving these physical processes, such as convection and cloud processes, introduce an additional complexity for traditional DA methodologies such as 4DVar or EnKF as they typically require tangent linear models and/or Gaussian background error assumptions. These approximations break down and perform poorly when attempting to update highly nonlinear physical processes or measurements with highly nonlinear measurement operators, such as cloudy-sky radiances [Geer and Bauer, 2011b, Kurosawa and Poterjoy, 2021, 2023, Poterjoy, 2022a].

In this chapter, we introduce a series of experiments using a configuration of HAFS that features

a static continuously-cycled basin-wide analysis, from which an ensemble of storm-centric forecasts is initialized. While the forecast domain configuration closely resembles that of the operational HAFS system, the analysis domain configuration is experimental and research-oriented in nature, and features methodologies such as online clear-sky radiance bias correction and augmented ensemble perturbations, which seek to minimize errors during DA [Kurosawa and Poterjoy, 2025, Poterjoy, 2022a, Poterjoy et al., 2021]. We investigate how this expanded analysis domain configuration performs at two different resolutions, 3 km and 6 km, and verify against observations, best track data, and reanalysis data. We pay particular attention to model physics, and examine how significantly different model states and error characteristics can be attributed to major differences in cloud coverage and composition. From here, we assess the performance of these configurations in forecasting two TC events during the 2022 season, Hurricanes Danielle and Earl, and relate their skill to the model-state differences and error characteristics discussed earlier. We compare these self-contained ensemble forecasts with forecast results from a configuration that resembles the operational HAFS-B workflow, as well as a "parallel" workflow which resembles the HAFS-B workflow but uses high-resolution ensembles from the basin-wide analysis for uncertainty specification.

In section 4.2, we describe the domain and workflow configuration, as well as the data assimilation framework, model physics, and observation details. In section 4.3, we verify both the basin-wide analyses as well as storm-centric forecasts initiated from these analyses. These results are presented as domain- and time-mean error verifications against ERA5, model field comparisons, and TC-specific error metrics against National Hurricane Center (NHC) best track data. In section 4.4, we summarize and discuss broader conclusions, as well as future work.

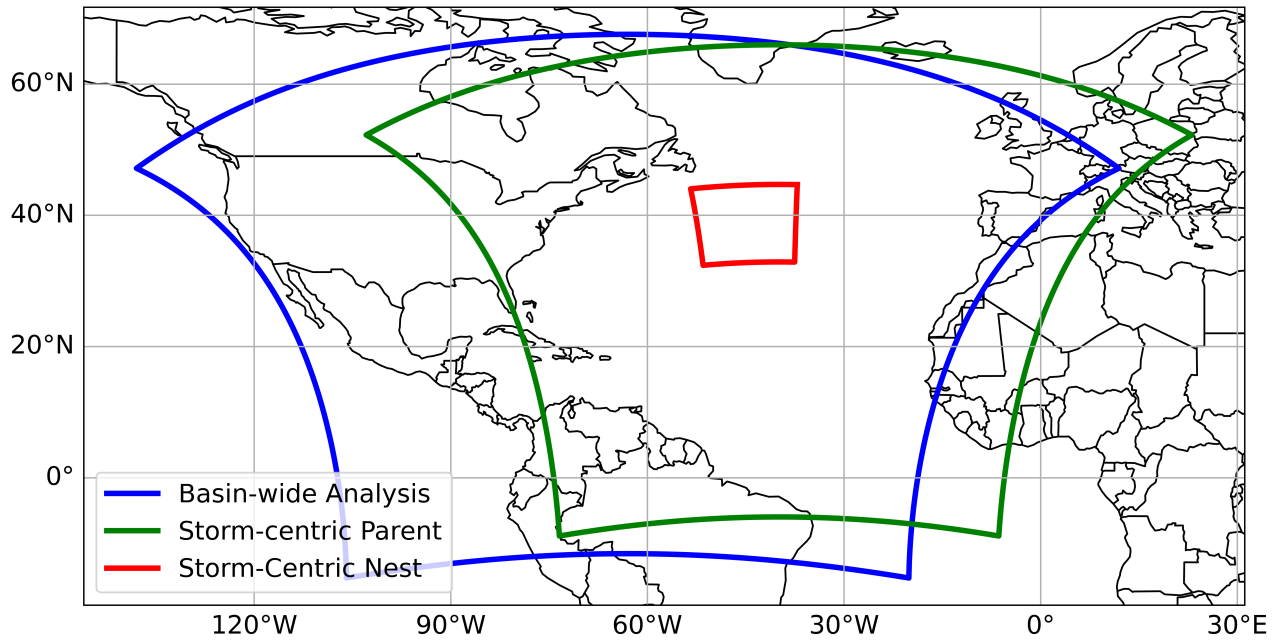


Figure 4.1: In green, the static basin-wide analysis domain. In blue, an example of a storm-centric forecast domain with a high-resolution moving nest in red.

4.2 Model Configuration

HAFS is NOAA's current generation flagship TC model, and the TC-focused component of the multi-application Unified Forecast System (UFS) [Jacobs, 2021]. Our experiments feature multiple workflow configurations, including both a stand-alone static basin-wide analysis and forecast workflow for producing initial conditions and flow-dependent error covariances as well as a storm-centric multi-scale ensemble forecast configuration for producing probabilistic forecasts of individual storm events. We also include a workflow configuration that very closely mirrors the operational workflow, but substitutes high resolution HAFS ensembles for background error covariance specification. In this section, we explain each workflow configuration in detail, and will refer to Fig. 4.2 frequently during the descriptions. We will also discuss the various methodologies and model physics used for the version of HAFS adopted for our study.

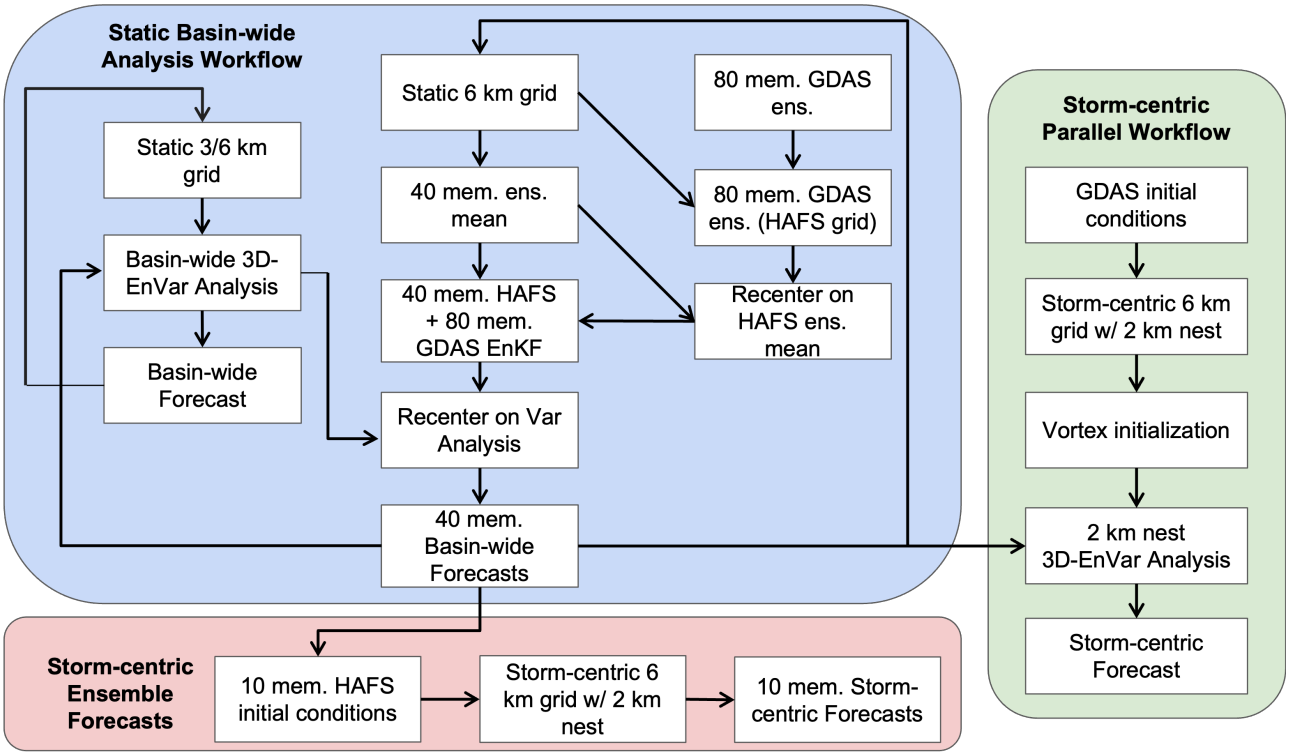


Figure 4.2: In the blue, the diagram details the static basin-wide analysis workflow. Notice both the deterministic variational and ensemble analyses/forecasts, and the augmentation of the 40-member high-resolution HAFS ensembles with the 80-member GDAS ensemble. In red, we show how 10 ensemble members from the basin-wide forecasts are borrowed and used as IC for long storm-centric forecasts, elsewhere in this chapter also referred to as ”self-contained” ensemble forecasts. In green, we show the ”parallel” workflow, which closely resembles the operational HAFS workflow, but with borrowed ensembles from the basin-wide analysis workflow for enhanced background covariance specification.

4.2.1 Static Basin-wide Analysis Workflow

The primary model workflow used for these experiments, shown in blue in Fig. 4.2, has a large static analysis and forecast domain which spans the contiguous USA, Central America, parts of Canada and South America, and the entire Atlantic basin, as seen in Fig. 4.1. This workflow is cycled continuously and assimilates observations across the whole domain every 6 h using an ensemble-based 3D variational (3DEnVar) data assimilation system. Our configuration features a high-resolution 40-member ensemble augmented with 80 GDAS members to produce flow-dependent error covariances, as described in section 4.2.1.3. Boundary conditions and the first set of initial conditions are interpolated from the 2022 operational global FV3-based GFS (∼13 km horizontal spacing) onto our HAFS domain. Subsequent cycle boundary conditions are provided every 6 h from the same global GFS forecasts [Dong et al., 2020], but IC and error covariance matrices come from the previous cycle forecast and forecast ensembles, respectively. In addition, we employ the first guess at the appropriate time (FGAT) technique over the 6-hour DA window to account for differences between actual observation times and background times.

4.2.1.1 Model Resolution

The FV3 dynamical core for the HAFS forecast is based on a non-hydrostatic finite volume solver using a Lagrangian vertical coordinate [Hazelton et al., 2018]. The adopted model domain has 81 vertical levels in sigma-pressure hybrid coordinates, where the lowest model level is about 25 m above the surface and the top isobaric level is 2 hPa. The domain uses an extended Schmidt gnomonic grid (ESG) projection, and is centered at 28° N, 63° W, with a 109.5° span for longitude and a 81.6° span for latitude [Purser et al., 2023].

We examine results from basin-wide analysis workflows with two different resolutions: 3 km and 6 km. In both cases, the high-resolution HAFS ensembles have a 6 km horizontal resolution, with grid specifications the same as the 6 km deterministic grid. In this study, we will examine how this difference in grid resolution affects domain-wide error statistics as well as TC-specific metrics such as track and max wind speed. In particular, we will highlight the effect that this differing resolution has on model physics, which seems to drive many of the differences we see in model states and error statistics.

4.2.1.2 Online Radiance Bias Correction

For TC-focused applications, satellite radiance observations are particularly important, as the analysis domain stretches predominantly over the ocean, where accurate in-situ observations are sparse. Satellite radiance measurements can be strongly biased, however, and the bias correction algorithms featured in TC NWP models, such as variational bias correction (VarBC), rely heavily on accurate and precise in-situ obs, or ”anchor obs,” to train bias models and subsequently estimate and correct ob bias. While it is common operationally to train bias correction models on external models, we have found that training the GSI VarBC bias model on the native HAFS modeling grid produces more accurate estimations of bias, and therefore more accurate model states and TC forecasts [Knisely and Poterjoy, 2023].

To further complicate the matter, cloud microphysics can significantly affect both clear-sky and all-sky satellite radiance assimilation. For the former, cloud coverage can affect how many observations the system throws out due to cloud contamination, and for the latter, cloud composition and thickness can affect observation innovations through the radiative transfer model measurement operator. This aspect is particularly important to our study as we examine how changes in model resolution affect

model physics, and consequently the evolution of model error and bias.

For our ”online” experiments, we initialize BC coefficients from GDAS values for the first analysis cycle but then transitions to self-cycling these coefficients within HAFS, meaning the resulting output coefficients are used as the first guess for the subsequent cycle [Hu et al., 2016]. In this way, the cycling of coefficients allows bias stemming from the model and data assimilation to accumulate during successive cycles, but limits potential system biases arising from the borrowing of GDAS coefficients. In other words, the configuration limits potential influence from an outside model and allows for better identification and understanding of biases arising from the satellite instruments and measurement operator alone. At the same time, the online estimation of predictor coefficients in HAFS is more prone to sampling error than those computed by GDAS, owing to the limited regional extent of the model domain. This limitation was found in Poterjoy et al. [2021] and Knisely and Poterjoy [2023] to be small compared to the benefits of using the native model to train a bias model for observations.

4.2.1.3 Augmented Ensemble Perturbations

DA requires careful estimation of uncertainty, especially background uncertainty, but current operational regional NWP systems often use relatively small ensemble sizes (e.g. 40 members in our configuration of HAFS) which can lead to sampling deficiencies. Following Kurosawa and Poterjoy [2025], the revised HAFS experiments introduce 80 additional ensemble perturbations from the GDAS system into background uncertainty estimates. The methodology resembles a hybrid DA methodology, in which a weighted sum of two background error covariances are used to mitigate sampling deficiencies. The primary difference is that instead of climatologically derived statistics, we instead use flow-dependent perturbations from the global system. These 80 low-res GDAS ensemble

perturbations are re-centered on the HAFS ensemble mean and then added to the 40 high-res HAFS ensemble perturbations, from which a combined Kalman gain is estimated during the EnKF update step [Kurosawa and Poterjoy, 2025].

This methodology has been shown to significantly reduce sampling error, thus providing a practical means of achieving basin-wide continuously-cycled ensemble data assimilation with an affordable ensemble size. Furthermore, it's reliance on flow-dependent GDAS perturbations makes it particularly attractive for the application of extreme weather, where historically derived error estimates may grossly misrepresent the current true atmospheric state [Kurosawa and Poterjoy, 2025]. The implications of using high-resolution cycled ensemble members vs. course-resolution GDAS members for background error covariance specification in the vicinity of TCs is explored later in a later section.

4.2.2 Storm-centric Ensemble Forecast Workflow

In addition to the short basin-wide forecasts produced from the previous workflow, we also produce an ensemble of 11 long storm-centric forecasts which mimic the operational HAFS forecast domain configuration. These forecasts adopt a 6 km parent domain with a high-resolution 2 km moving nest, and are initialized from the deterministic basin-wide analysis described in the previous section, along with the first 10 ensembles from the same system [Hurricane Analysis and Forecast System, 2025, Poyer et al., 2025]. We include this workflow configuration specifically because it mirrors the operational HAFS workflow, and thus allows for a direct comparison against operational forecast results. Furthermore, the inclusion of a forecast ensemble allows us to examine ensemble means and spread for domain-wide and TC-specific error metrics, formulating a hurricane prediction product that can be verified probabilistically. Altogether, the combination of this self-contained basin-wide analysis and

ensemble forecast workflow provides a fully probabilistic DA framework for testing and verifying the feasibility of a potential high-resolution basin-wide prediction system built around HAFS.

4.2.3 Storm-centric Parallel Workflow

We also test a series of experiments that very closely resemble the HAFS 2023 operational HAFS-B workflow. Like the self-contained ensemble forecast workflow, this configuration adopts a 6 km parent domain with a high-resolution 2 km moving nest for the forecast, but is instead largely initialized from GDAS, as opposed to from the self-contained analysis. This parallel workflow features vortex initialization and limited data assimilation, namely around the core of the storm itself, which primarily focuses on capturing and assimilating valuable in-situ inner-core measurements such as TDRs and dropsondes which are not necessarily assimilated by GDAS [Ramstrom et al., 2024].

Like our self-contained analysis workflow, this framework employs a 3DEnVar system with FGAT, but with notable changes to the construction of our background error covariance, or $\mathbf{B}$, matrix. While the operational HAFS configuration borrows the 80 member ensemble from GDAS for background error covariance specification, our experimental parallel workflow uses the same augmented ensemble method as described in section 4.2.1.3 to combine the high-resolution 40 member HAFS ensemble from our basin-wide analysis configuration with the 80 member ensemble from GDAS. Thus, with this method, we can evaluate how the inclusion of high-resolution HAFS ensemble perturbations affects our estimation of uncertainty, and subsequently if such a configuration can improve the prediction of TC events.

For the sake of comparison, we run a final configuration which is exactly the same as the 2023 HAFS-B operational configuration, and thus resembles our parallel workflow in every way except that

it uses just the 80 member ensemble from GDAS for background error covariance specification. From hence forth, we refer to this final configuration as ”operational HAFS-B” to distinguish it from the parallel workflows, and use the results of this experiment as a ”control” for several verification metrics.

4.2.4 Model Physics

We use a model physics configuration that resembles the operational NCEP GFS version 16 (GFSv16), but with modifications that reflect boundary layer processes that are uniquely important for TCs. For more in-depth detail, we refer to Wang et al. [2024]. Major components of the physics package include: the GFS Rapid Radiative Transfer Model for GCMs (RRTMG) shortwave/longwave radiation scheme [Mlawer et al., 1997]; the Noah-Multiparameterization Land Surface Model (Noah-MP LSM) [Niu et al., 2011]; the GFS scale-aware TKE-based Moist Eddy-Diffusion Mass-Flux (EDMF) PBL and free atmospheric turbulence scheme [Han and Bretherton, 2019]; the GFS surface layer scheme with HWRF exchange coefficients [Zheng et al., 2017]; and the GFS orographic gravity wave drag scheme [Kim and Arakawa, 1995].

In addition, we draw closer attention to aspects of HAFS that we suspect have the most direct impact on findings presented in future sections. HAFS uses the simplified Arakawa-Schubert (sa-SAS) deep convection scheme is based on Pan and Wu [1995], with updates to introduce a shallow convection scheme using a mass flux approach in Han and Pan [2011] and to add scale- and aerosol-aware parameterizations in Han et al. [2017b]. Cloud microphysics are represented using the Thompson scheme, which is a partial double-moment bulk microphysical parameterization that predicts the mixing ratios of cloud water, rain, cloud ice, snow, and graupel, as well as the number concentration of cloud water and ice Thompson and Eidhammer [2014], Thompson et al. [2008], Wang et al. [2024]. The

scheme is aerosol-aware, incorporating the activation of aerosols as cloud condensation and ice nuclei [Thompson and Eidhammer, 2014]. Our version of HAFS also features Community Mediator for Earth Prediction Systems (CMEPS) -based ocean coupling with the Hybrid Coordinate Ocean Model (HYCOM), which is described in detail in Chassignet et al. [2007, 2009]. The HYCOM system features a 1/12-degree resolution North Atlantic basin domain with 41 vertical levels, initial conditions from NCEP's Real Time Ocean Forecast System (RTOFS), persistent oceanic lateral boundary conditions, and atmospheric forcing from GFSv16 for non-overlapping areas [Hazelton et al., 2023].

4.2.5 Observations and Data Assimilation Schedule

The experiments begin 06 UTC August 23 and end at 06 UTC on September 11 2022, covering two and half weeks from the 2022 season and two named storms, Danielle and Earl. Our basin-wide analysis assimilates measurements at 00, 06, 12, and 18 UTC each day, and generate short-range (9-h) deterministic forecasts. From these forecasts, we initialize the ensemble of long-range (126-h) storm-centric forecasts. For many verification, we omit the first 5 days of data assimilation to allow HAFS to become independent from the GDAS initial conditions that are used to initialize experiments.

The suite of assimilated measurements includes all measurements assimilated by GDAS except proprietary or novel measurements, such as cloudy radiances. In addition, HAFS assimilates inner core data, including TDR velocities, dropsonde measurements, and flight-level wind observations, which are not assimilated by the GDAS.

The Geophysical Fluid Dynamics Lab (GFDL) vortex tracker is used to generate the TC track information [Biswas et al., 2018]. TC track, maximum 10-m surface wind, and minimum surface pressure from model forecasts are verified against "best track" data from the National Hurricane Center

[Hazelton et al., 2021].

4.3 Results

In the following section, we detail systematic differences between the model states generated by our basin-wide experiments, and investigate how these differences contribute to forecast skill. Namely, we highlight how choices of forecast and deterministic analysis resolution, namely across the "gray-zone" [Yano et al., 2018] of 6 km to 3 km grid spacing results in major differences in the placement and concentration of hydrometeors (cloud water, ice, rain, snow, and graupel) as well as cloud fraction and specific humidity. We also track how differences in model physics affect surface variables, such as precipitation and surface fluxes, as well as domain-wide air mass biases. Finally, we assess how these differences in analyses, and subsequently IC for storm-centric forecasts, affect storm forecast outcomes.

4.3.1 Cloud Fraction and Hydrometeor Concentration

We begin by detailing major differences in cloud composition and placement between the 3 km and 6 km basin-wide experiments. We highlight overall differences in cloud fraction and concentration and dissecting individual prognostic hydrometeor variables (cloud water, ice, rain, snow, and graupel). In many cases, we take spatial- and temporal-means over the whole domain or individual pressure levels and across the experimental period.

In Fig. 4.3, we examine time- and zonal-means of 6 h forecast cloud fraction, hydrometeors, and specific humidity, respectively, and the difference between our two experiments. The comparisons show a greater fraction of cloud coverage from the surface to 600 mb in the 3 km experiment, peaking between 700 and 650 mb, mixed differences from 600 to 400, and again greater fractions of coverage in

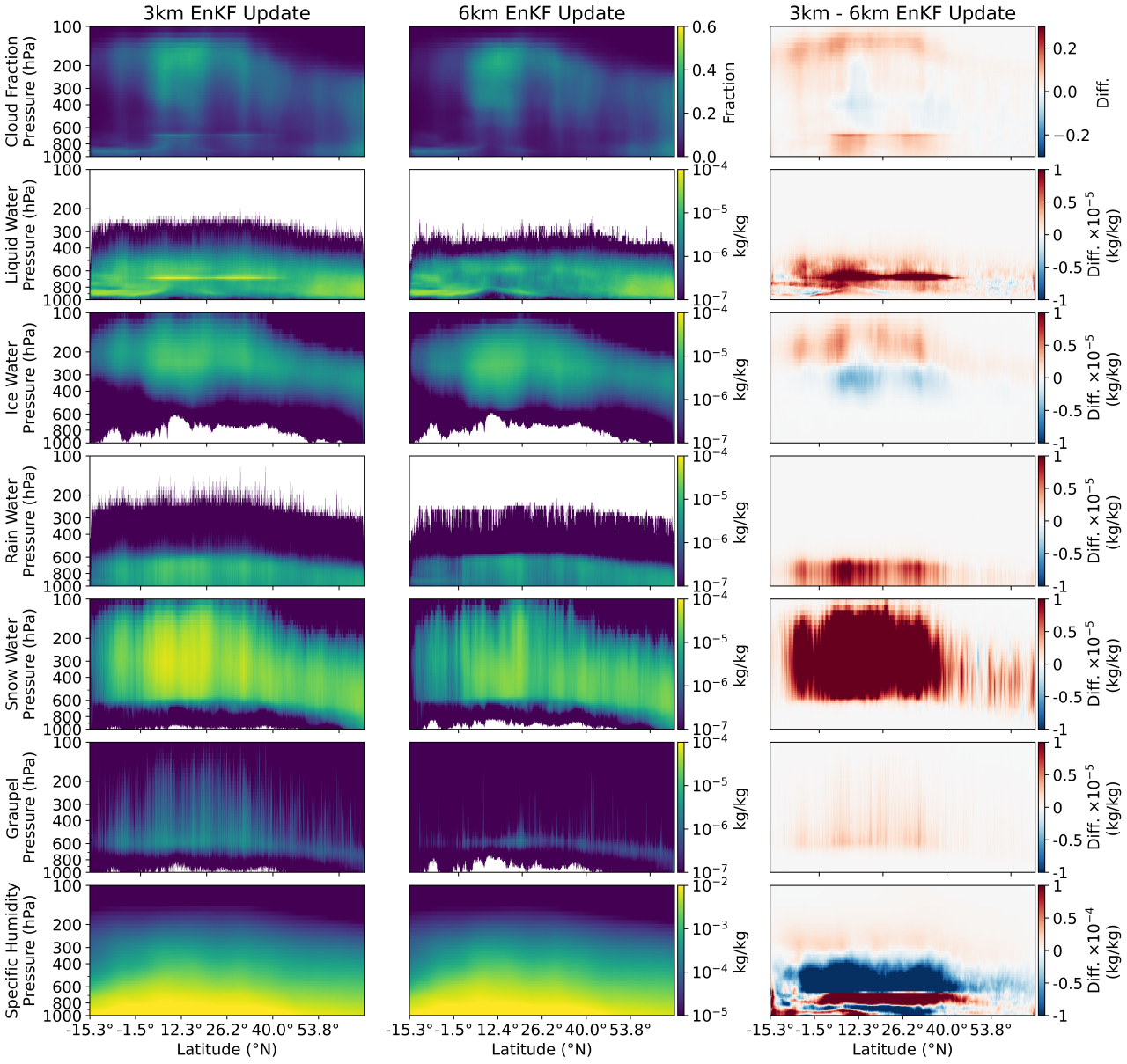


Figure 4.3: Time- and zonal-mean of cloud fraction, hydrometeor variables (liquid water, ice water, rain water, snow water, and graupel) and specific humidity for 3 and 6 km basin-wide forecasts, and their difference, respectively. Notice the difference in y-axis scale between cloud fraction, hydrometeors, and specific humidity.

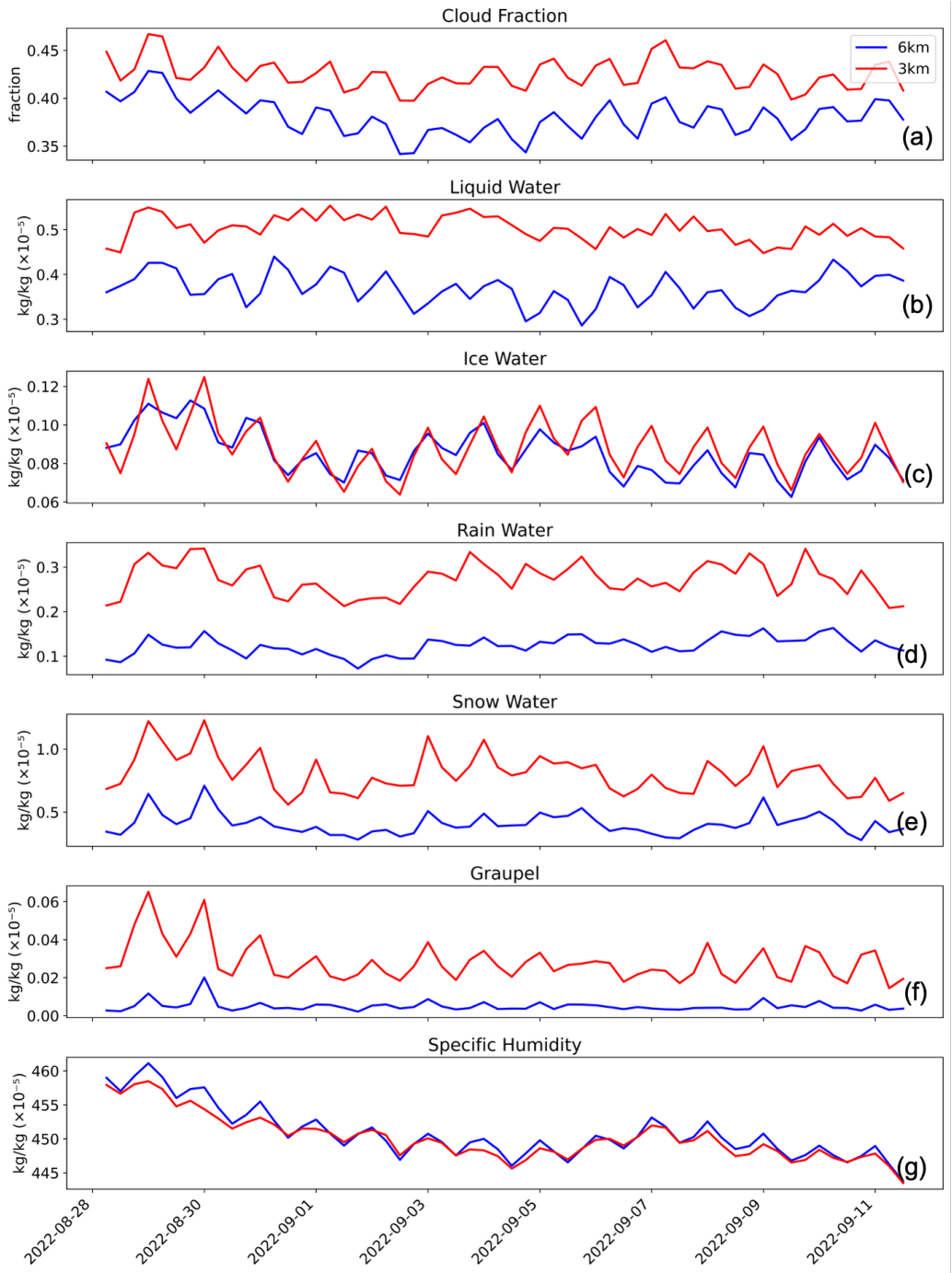


Figure 4.4: Time- and domain-mean column-mean values over time for (a) cloud fraction, hydrometeor variables ((b) liquid water, (c) ice water, (d) rain water, (e) snow water, and (f) graupel), and (g) specific humidity for 3 and 6 km basin-wide forecasts. Notice the difference in y-axis scale between cloud fraction, the hydrometeors, and specific humidity

the 3 km experiment above 400 mb.

Evidently, the increase in cloud fraction throughout the lower troposphere in the 3 km experiment is a result of increased concentrations of cloud liquid water and rain water. It further corresponds to a large positive difference in specific humidity from about 800 to 700 mb, although below this layer results are more mixed and depend on latitude. These results suggest a greater presence of hazy stratus/nimbostratus clouds throughout the 3 km experimental results, and the sharp contrast in specific humidity from below to above 700 mb suggests a difference in static stability between our experiments.

Above this layer, Fig. 4.3 shows markedly greater specific humidity throughout the 6 km model states, up to about 400 mb. We find larger cloud ice mixing ratios throughout the layer in the 6 km experiments, whereas the 3 km experiments show greater snow and graupel mixing ratios. This result is noteworthy because snow and graupel are typically associated with strong convective updrafts, yet the convective-activity and error metrics presented in the next section indicate generally stronger convection in the 6 km experiments (see Fig. 4.9 in subsection 4.3.2).

In the upper troposphere, we find a cold cloud bias for the 3 km results, wherein greater snow and ice water hydrometeor concentrations lead to greater cloud fractions above 300 mb. We note that most of the differences in hydrometeors across experiments taper off at higher latitudes (above 50°N). Fig. 4.4 further confirms many findings from above, and shows that these differences are largely consistent over time. It also shows how specific humidity is very similar across both experiments across all date-times, despite the fact that cloud fraction and most hydrometeor concentrations are noticeably greater for the 3 km experiment.

4.3.2 Domain-wide Analysis Verifications

Next, we examine how differences in cloud composition and coverage affect surface properties that are important for modeling both TCs and their environment, such as surface radiative fluxes, temperature, and heat fluxes. Our investigation of air mass differences is further supplemented by comparing short-term (0 – 6 h forecasts) to in-situ observations of temperature and humidity. We also verify the data vs. ECMWF Reanalysis v5 (ERA5) data, which uses 4DVar data assimilation and comprehensive observation data sets including all-sky radiances; it is also relatively high-resolution compared to other global reanalysis data sets [Hersbach et al., 2020].

In Fig. 4.5, we examine differences in the downward short-wave radiation flux, the downward long-wave radiation flux, and the surface temperature, respectively, between the 3 km and 6 km experiments. We note that positive values mean that the 3 km results have higher values, and vice versa, and segregate the variables by land, ocean, and total values. The comparison shows that downward short-wave flux differences are strongly negative across nearly all date-times and across the entire domain, which, in conjunction with findings in the previous section, suggests that the thick, low-level clouds being produced more heavily in the 3 km experiment are blocking more incoming solar radiation compared to the 6 km experiment. Likewise, we examine downward long-wave flux in panel (b), and notice that these flux differences are more mixed but tend to slightly more positive, meaning greater values for the 3 km experiment, especially while both Earl and Danielle are of tropical storm strength or greater. While we should note that these flux difference values are an order of magnitude smaller than the incoming radiative fluxes in panel (a), these results further bolster the conclusion that the 3 km configuration is producing far more cloud cover than the 6 km configuration.

The net impact of the identified changes in surface energy balance are summarized in Fig. 4.5c,

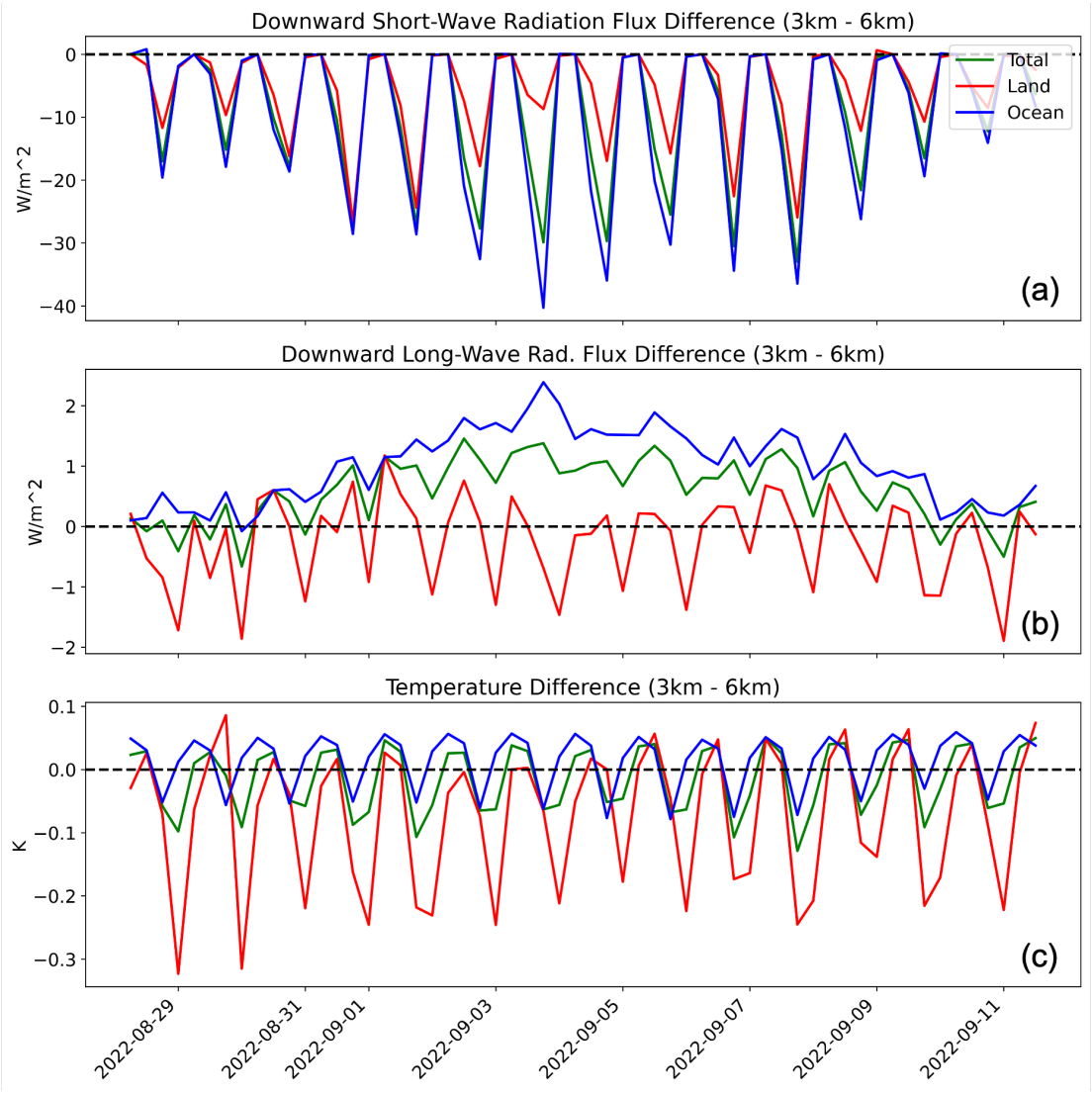


Figure 4.5: Time- and domain-mean differences over time for (a) downward short- and (b) long-wave radiative fluxes at the surface, along with (c) surface temperature for 3 and 6 km basin-wide forecasts. Positive values indicate 3 km is greater, negative values indicate 6 km is greater. Differences over-land, over-ocean, and total differences are all considered. Note the difference in y-axis flux values.

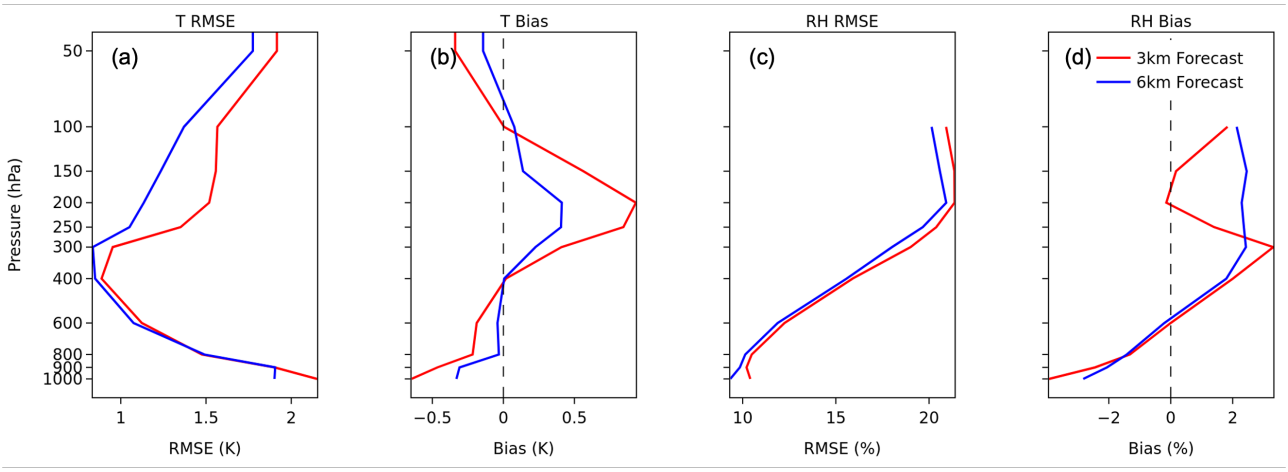


Figure 4.6: Vertical profile of time- and domain-mean differences of observations minus 6 hour forecasts. Panels give (a) temperature RMSE,(b) temperature bias, (c) humidity RMSE, and (d) humidity bias. Note that only sondes were used as observations for this verification, given their high accuracy and precision.

which shows surface temperature differences. Here, we find that the 6 km experiment is generally warmer over land, while over the open ocean the differences are more mixed and depend on the time of day. Across the entire domain, the 6 km results tend to be warmer during the day while the 3 km model state surface tends to be warmer or neutral at night. In the next section, we provide a more detailed examination of causes for this result, along with implications for bulk factors, such as air mass bias across the entire HAFS domain.

To provide additional context to the noted cloud differences (and resulting radiative response) across the two experiments, we examine domain-wide time-average forecast departures from observations to quantify time- and domain-average bias and root mean square error for temperatures and relative humidity; Fig. 4.6 shows both metrics plotted as a function of altitude. We first notice that the 3 km experiment results in greater RMSEs for both variables throughout almost the entire profile, indicating that systematic differences in hydrometeor quantities are likely due to elevated errors in this experiment. Air mass bias is more complicated: the 3 km experiment tends to be more biased

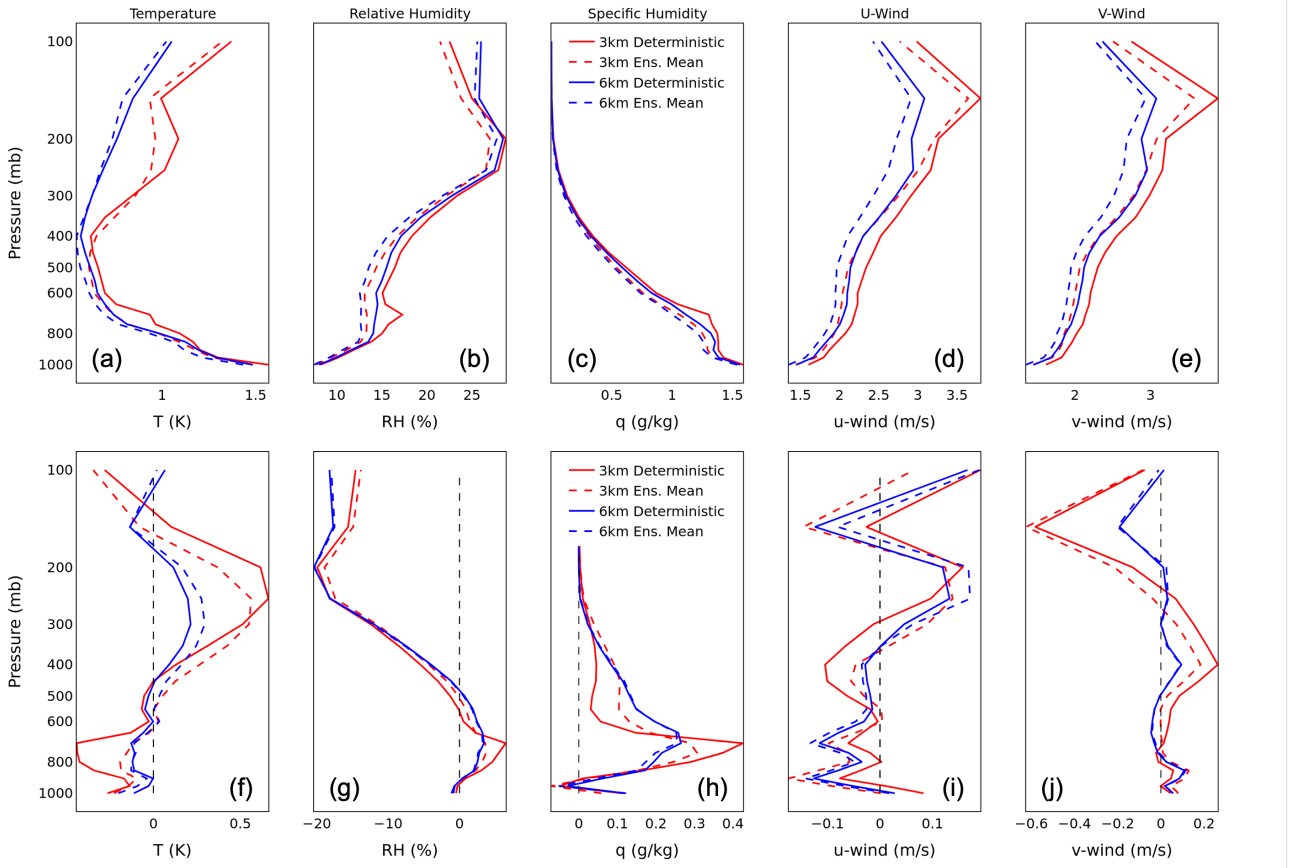


Figure 4.7: Panels (a), (b), (c), (d), and (e) show vertical profiles of time- and domain-mean RMSE for temperature, relative humidity, specific humidity, and the u- and v-winds, respectively, for 3 km and 6 km basin-wide forecasts verified against ERA5. Panels (f), (g), (h), (i), and (j) show vertical profiles of time- and domain-mean bias for the same variables.

throughout the profile, first having a stronger negative temperature bias from the surface up to about 400 mb, and then having a stronger positive bias from 400 to 100 mb. The 3 km experiment also has a larger dry bias near the surface, with more mixed results throughout the rest of the column.

We note that the observation-space verifications are performed using sounding measurements only, which are highly accurate but relatively sparse over the ocean. Therefore, we supplement this verification with domain- and time-average forecast departures from ERA5 in Fig. 4.7, which relies on a much larger number of environmental measurements and 4D-Var data assimilation over a long window. Here, we examine 006 hr deterministic and ens. mean forecast RMSEs in panel (a) and bias in (b) for temperature, relative and specific humidity, and u- and v- winds. We find that the temperature RMSE and bias profiles are very similar to that shown in Fig. 4.6, whereas the the relative humidity profile show some differences, particularly throughout the mid- to upper-troposphere with a stronger negative RH bias for both 3 and 6 km experiments. We further note that the 3 km experiment relative and specific humidity bias profiles show a sharp spike between 800 and 600 mb, which corroborates findings from Fig. 4.3. Finally, we note that while u- and v-winds have relatively high RMSEs, their biases throughout the column tend to be near zero. Ens. mean forecast results are largely similar to their deterministic counterparts, albeit producing slightly smaller RMSEs for temperature, humidity, and wind. Since the ens. mean for the 3 km basin-wide analyses experiments is performed at 6 km but recentered on the 3 km deterministic analysis, we notice the ens mean. bias profile often sits in-between the 3 km deterministic and 6 km deterministic/ens. mean profiles.

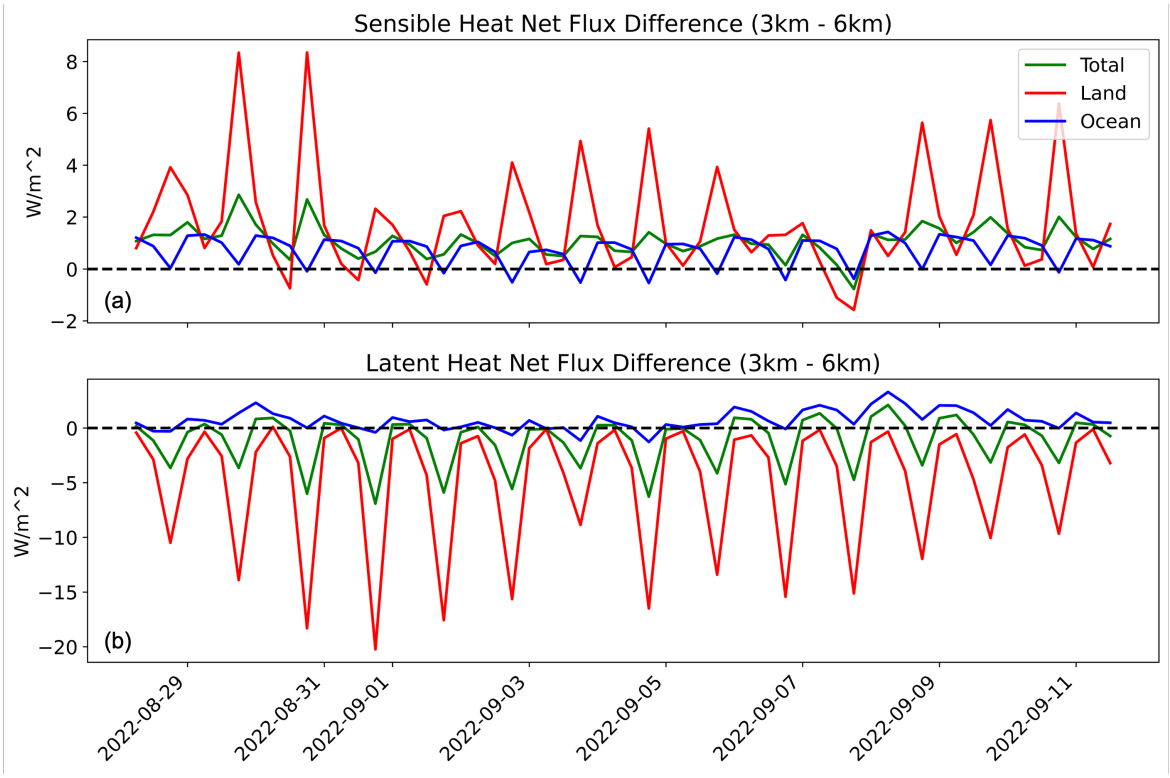


Figure 4.8: Time- and domain-mean differences over time for (a) sensible and (b) latent fluxes at the surface, along with surface temperature for 3 and 6 km basin-wide forecasts. Positive values indicate 3 km is greater, negative values indicate 6 km is greater. Differences over-land, over-ocean, and total differences are all considered. Note the difference in y-axis flux values.

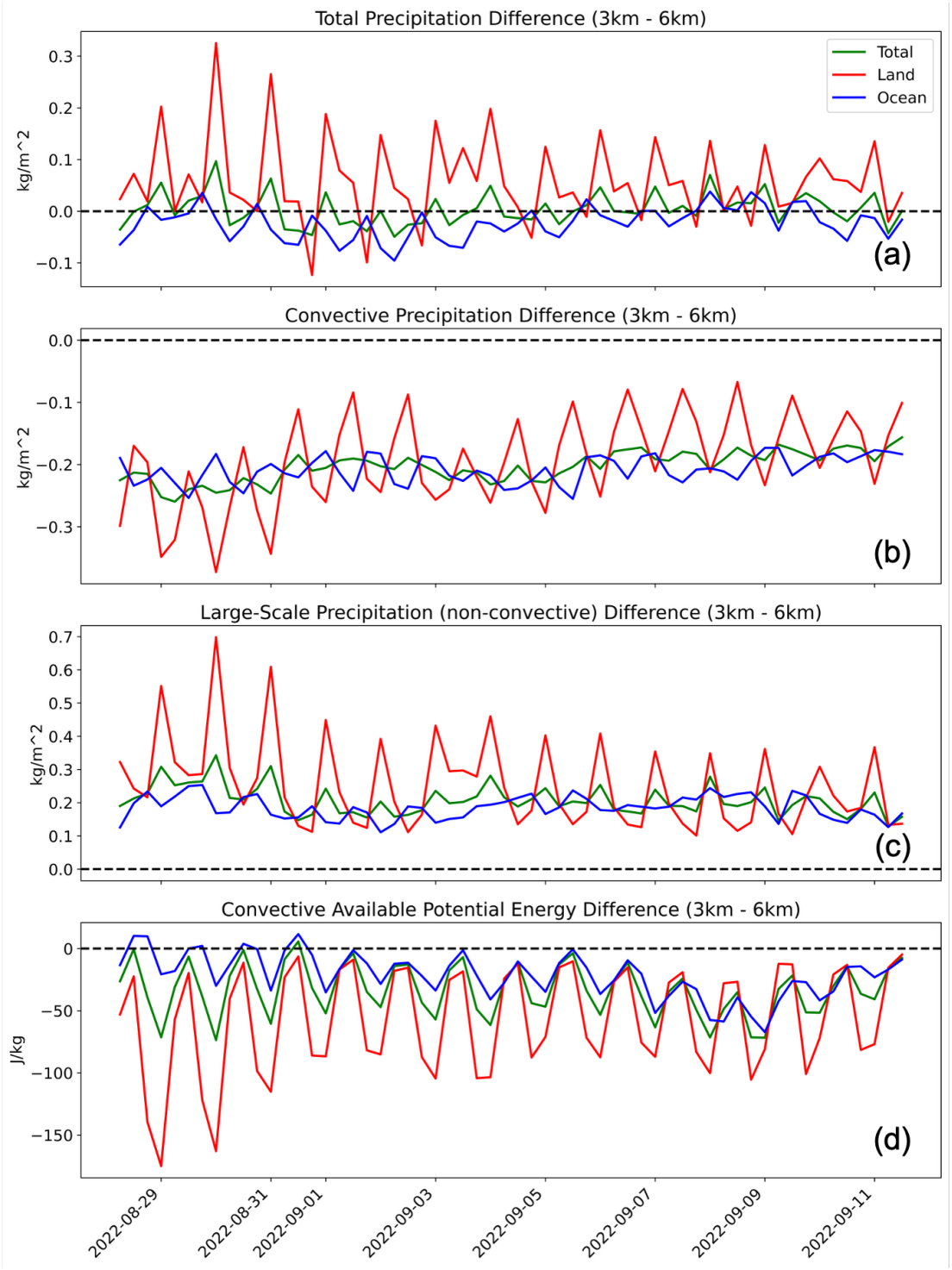


Figure 4.9: Time- and domain-mean differences over time for (a) column total precipitation, (b) convective precipitation, (c) non-convective precipitation, and (d) CAPE for 3 and 6 km basin-wide forecasts. Positive values indicate 3 km is greater, negative values indicate 6 km is greater. Differences over-land, over-ocean, and total differences are all considered. Note the difference in y-axis precip. values.

4.3.3 Heat Fluxes and Feedbacks

Next, we examine differences in latent and sensible heat flux during the same period. In Fig. 4.8b, we see that latent heat flux from the surface to the atmosphere is mostly biased towards the 6 km experiment over the entire time period, particularly over land, and that it peaks during the mid afternoon. This suggests that the increased solar radiation reaching the surface is having a strong effect heating the surface, leading to increased evaporation and latent heat flux, and consequently leading to a warmer and more accurate low level air mass temperature bias (Fig. 4.6). Meanwhile, sensible heat fluxes are mostly stronger in the 3 km experiment, which makes sense as a cool air mass bias in the lower troposphere would encourage additional heat fluxes from the surface. The warmer temperatures of the surface of the 3 km experiment, as shown in Fig. 4.5c, are the result of less evaporative cooling deficit during the day, and the retention of more of its thermal energy. The strong positive air mass bias aloft in the 3 km experiment is more difficult to explain, but may be a result of enhanced high-altitude ice/snow cloud formation as shown in fig. 4.3 and the latent heat released by cloud condensation/deposition.

Altogether, these differences in surface fluxes and air mass biases result in significantly different precipitation patterns and static stability across the domain. In Fig. 4.9, we examine differences in total, convective, and non-convective accumulated precipitation, along with CAPE, respectively. Immediately we notice that while total precipitation differences are generally near zero, being only slightly biased towards the 3 km experiments over land and vice versa for the 6 km results over ocean, there are large differences in convective vs non-convective precipitation patterns. Namely, that convective precipitation values are greater for 6 km results across the entire domain and over all time steps, and vice versa for non-convective precipitation for the 3 km experiment. Unsurprisingly, our verification of CAPE in panel (d) shows a greater values across the entire domain and across most date times for the 6 km experiment.

The stronger convective activity in the 6 km experiment can be clearly tied back to the strong latent heat fluxes and more positive low-level temperature and humidity bias, as shown in Figs. 4.8 and 4.6, respectively, which themselves are a consequence of differences in incoming short-wave radiation flux at the surface due to significantly different cloud compositions and location throughout domain. The 3 km experiment's enhanced low-level clouds suppress surface heating, which results in reduced latent heat flux and a negative low-level air mass bias. This effect, combined with a strong positive air mass bias aloft, results in enhanced static stability and depressed deep convective activity across the entire domain.

These results suggest a nonlinear feedback wherein resolution-dependent differences in cloud formation become self-reinforcing through many modeling cycles. While both configurations use identical physics parameterizations, the observed differences in convective vs. non-convective precipitation and cloud composition suggest that these schemes parameterize or resolve convective processes differently across the two resolutions. Because each analysis cycle inherits the thermodynamic environment from the previous forecast, these biases compound over time, ultimately resulting in significantly different analyses and subsequently IC for storm-centric forecasts.

4.3.4 Domain-wide Forecast Verification

In the following subsection, we verify 126-h forecasts, as detailed in section 4.2, and compare both the self-contained storm-centric forecasts as well as forecast initialized from the "parallel-workflow" configurations that use model data from 3 km and 6 km analyses. We differentiate between the "deterministic" self-contained forecasts, i.e. the storm-centric forecast initialized from the 3DEnVar analysis, vs. the ensemble mean of the 10-member storm-centric forecast ensemble. We contrast

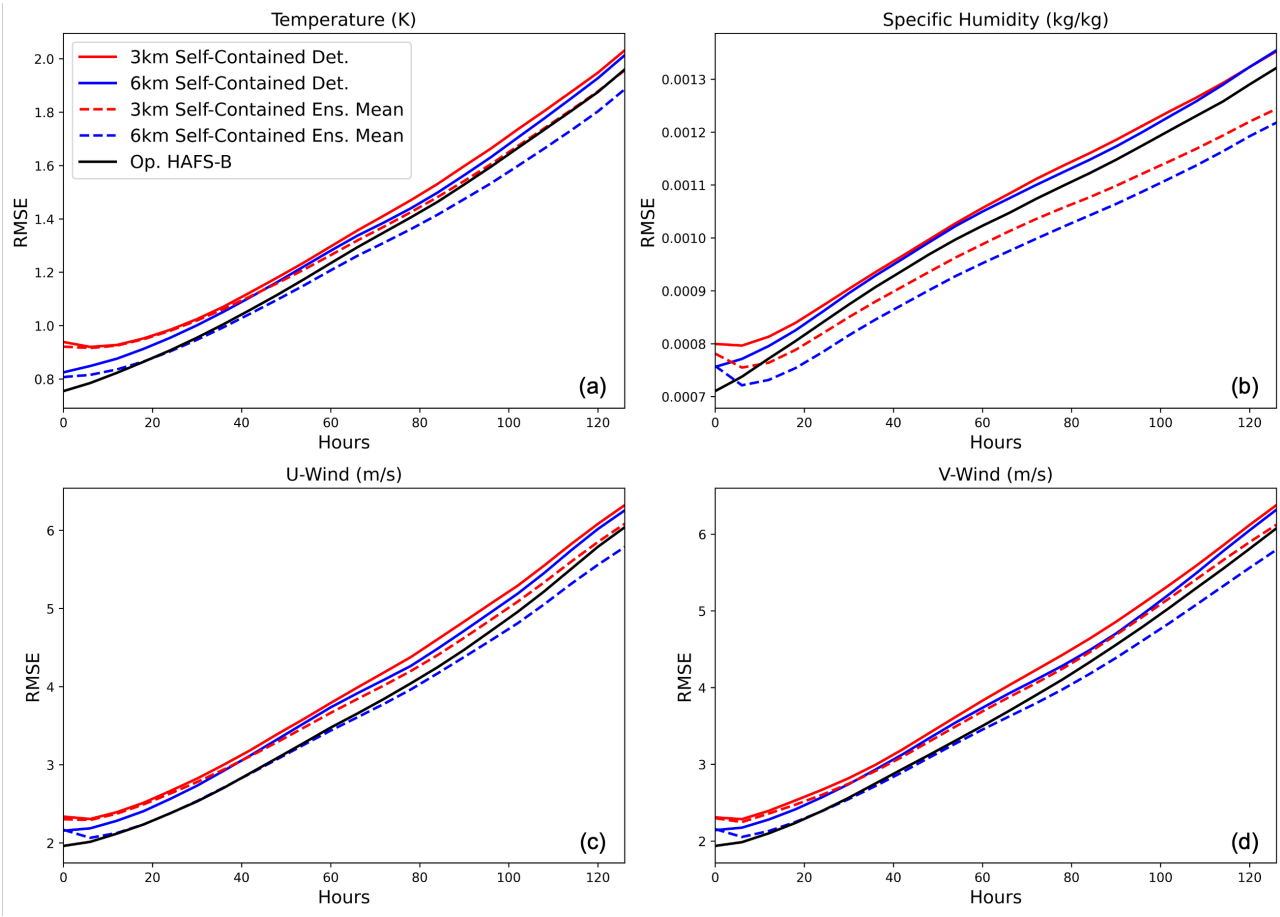


Figure 4.10: Domain- and time-mean RMSE for (a) temperature, (b) sp. humidity, (c) u-wind, and (d) v-wind for storm-centric forecasts initialized from 3 km IC, 6 km IC, and operational HAFS IC. Verified against ERA5.

experiments against results from the operational HAFS-B configuration.

We produce and compare domain-wide error verifications over long lead-times for state variables such as temperature, humidity, and wind speed. We also examine limited surface variables, such as sea surface temperature (SST) and precipitation. We show how differences in initial conditions lead to different domain error characteristics at longer lead-times, and consequently affect TC forecast performance.

We begin with a general verification of storm-centric deterministic and ens. mean forecasts initiated from the 3 and 6 km basin-wide analyses. In Fig. 4.10, we verify time- and domain-wide

departures of temperature, humidity, and u- and v-winds against ERA5 data from 0 to 126 hour lead-time. We find that the errors for all variables grow at a roughly similar rate, with the largest differences in errors across forecasts to be in the 0 hour lead-time and the smallest differences at the end at 126 hours. Nevertheless, the operational HAFS-B outperforms the 3 and 6 km deterministic forecasts across all variables and lead-times. The 6 km deterministic forecast tends to outperform the 3 km deterministic at early (under 30 hours) and late (over 60-72 hours) lead-times, and perform approximately equally at lead-times in-between. Since the only difference in these forecasts is coming from the initialization, and therefore the data assimilation system that produces these IC, it's clear that our HAFS system still requires tuning to produce comparably skillful IC as the GDAS.

We highlight that while the operational HAFS-B performs better than both deterministic self-contained forecasts across all lead-times, the ens. mean forecasts do produce some improved RMSEs over the operational model, likely because ensemble averaging reduces random error. The 6 km self-contained ens. mean forecast shows lower RMSEs for temperature, u-, and v- wind over late lead-times (over 60 hour), while both the 3 and 6 km ens. mean forecasts show lower specific humidity errors across nearly all forecast lead-times (over 12 hour).

Next, we verify time- and lower-tropospheric column-mean (950 to 700 mb) values for temperature and humidity against ERA5, and examine spatial patterns over the Atlantic basin. In Fig. 4.11, we verify analyses from the 3 and 6 km basin-wide framework, which serve as IC for self-contained forecasts, along with IC for the operational HAFS-B forecast. We omit the ens. mean analyses, as they are identical to the deterministic analyses due to ensemble recentering. We show that the 3 km analysis has a noticeable cold bias and strong positive relative humidity bias over nearly the entire domain, consistent with findings from earlier in Fig. 4.6. The 6 km analysis looks much more similar to the operational HAFS IC, albeit with slightly warmer biases through the tropics, the Gulf of

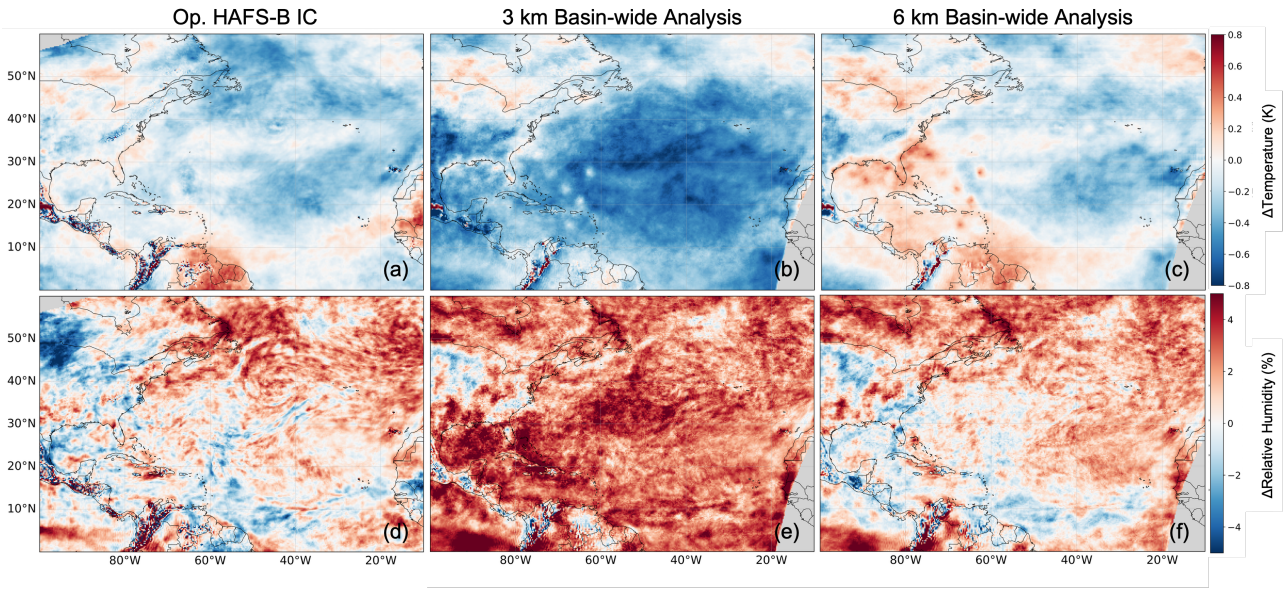


Figure 4.11: Time- and lower tropospheric column-mean temperature and relative humidity RMSE, verified against ERA5, for (a) and (d) operational HAFS-B IC, (b) and (e) 3 km HAFS IC, and (c) and (f) 6 km HAFS IC, respectively. Column-mean is taken from 950 to 700 mb.

Mexico, and eastern Atlantic basin. The 6 km analysis also appears to have slightly higher moisture biases over the same regions.

We perform the same verification of column-mean variables for operational and self-contained forecasts at their longest lead-time, 126 hours, and show the results in Fig. 4.12. In doing so, we reveal biases in the forecast model itself, and differentiate them from biases in the analysis states. We find that all three long forecast error fields look extremely similar, and while we do not include ens. mean forecasts, we note that they also closely resemble the error fields in Fig. 4.12. The forecasts have strong negative temperature biases throughout the Arctic regions and the mid-latitude Atlantic basin, save for a small warm bias over the Gulf and the Azores, while having strong positive biases over most of South America, Midwestern and Western US, and over western and eastern Canada. Relative humidity verification is more mixed, but we do notice patterns of strong relative humidity bias throughout most of the Atlantic basin.

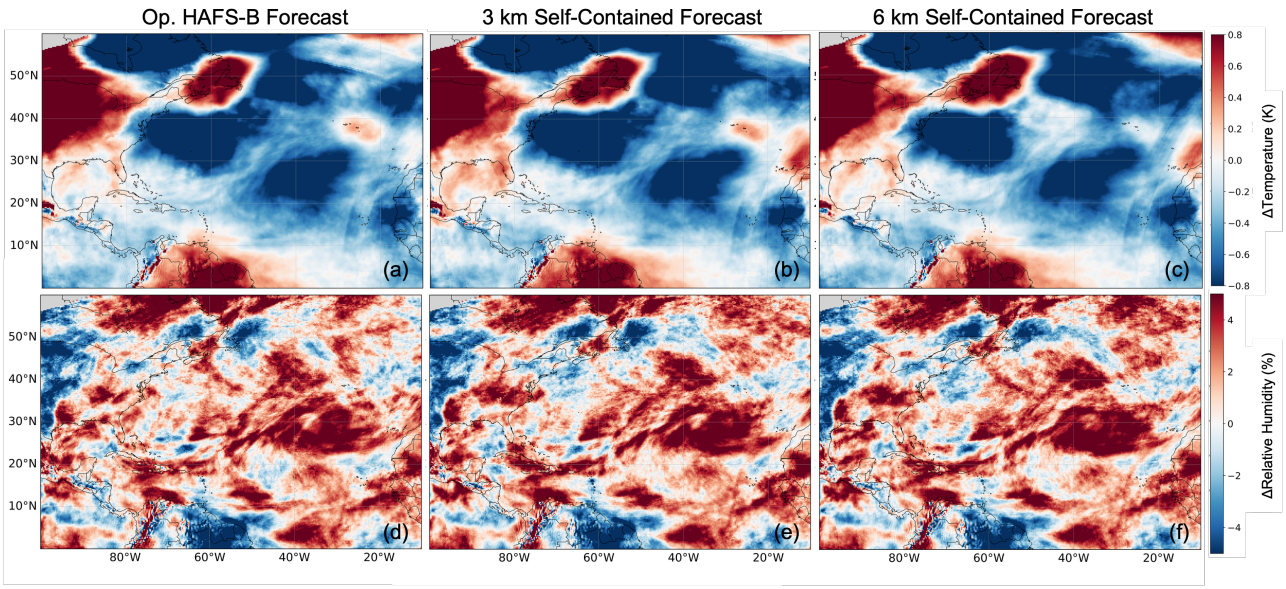


Figure 4.12: Time- and lower tropospheric column-mean temperature and relative humidity RMSE, verified against ERA5, for (a) and (d) operational HAFS-B Forecast, (b) and (e) Self-Contained Forecast from 3 km Analysis, and (c) and (f) Self-Contained Forecast from 6 km Analysis, respectively. Column-mean is taken from 950 to 700 mb.

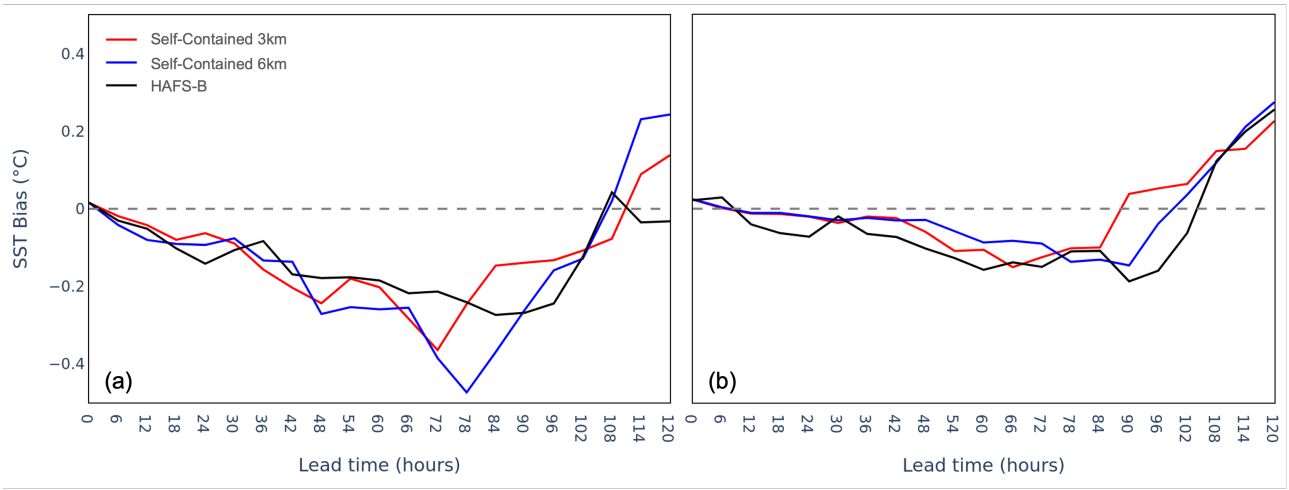


Figure 4.13: Sea surface temperature (SST) bias (HAFS - RTOFS) as a function of forecast lead-time for 3 km and 6 km self-contained workflows and operational HAFS-B. (a) shows the inner core region (0 to RMW), (b) shows the outer core region (50 to 100 nm). Bias is computed as the spatial mean SST within each annular region, averaged over all initialization times and storms.

The strong negative temperature biases at long lead times may be related to air-sea interactions. To explore this, we verify HYCOM SSTs against RTOFS analyses, which provide initial conditions for HYCOM in HAFS [Hazelton et al., 2023]. Figure 4.13 shows domain-mean SST bias as a function of lead time, split into inner core (center to RMW) and outer core (50–100 nm) regions. Across all configurations, a cold SST bias develops with increasing lead time before transitioning to a warm bias after 90–108 h. In the inner core (panel (a)), the 3 km and 6 km self-contained forecasts show stronger cold biases than operational HAFS-B, potentially due to stronger storm intensities in these configurations driving greater ocean mixing and upwelling (discussed further in the following subsection). Outer core biases (panel (b)) are smaller overall and more similar across configurations, with operational HAFS-B showing slightly stronger cold biases.

4.3.5 Storm-centric Forecast TC-Metrics

In the following subsection, we provide a series of verifications of TC forecast errors statistics, specifically: track, 10-m maximum winds, minimum sea level pressure (SLP), and radius max wind (RMW). These statistics are summarized for 3 and 6 km self-contained forecasts, 3 and 6 km parallel workflows, and HAFS-B operational workflow in Fig. 4.14. Forecast RMSEs and mean bias verified against NHC best track data are collected across both storms in the experimental period (Danielle and Earl) and plotted as a function of forecast lead-time, with case counts ranging from 61 to 59.

Track errors across all experiments are nearly identical up to about 42 hours, after which the self-contained forecasts begin to diverge and worsen in track performance, especially with the 3 km forecast. As seen in the SLP and max wind verifications, the self-contained forecasts spin up considerably more intense cyclones. While all configurations produce strong negative SLP biases at long lead-times, the

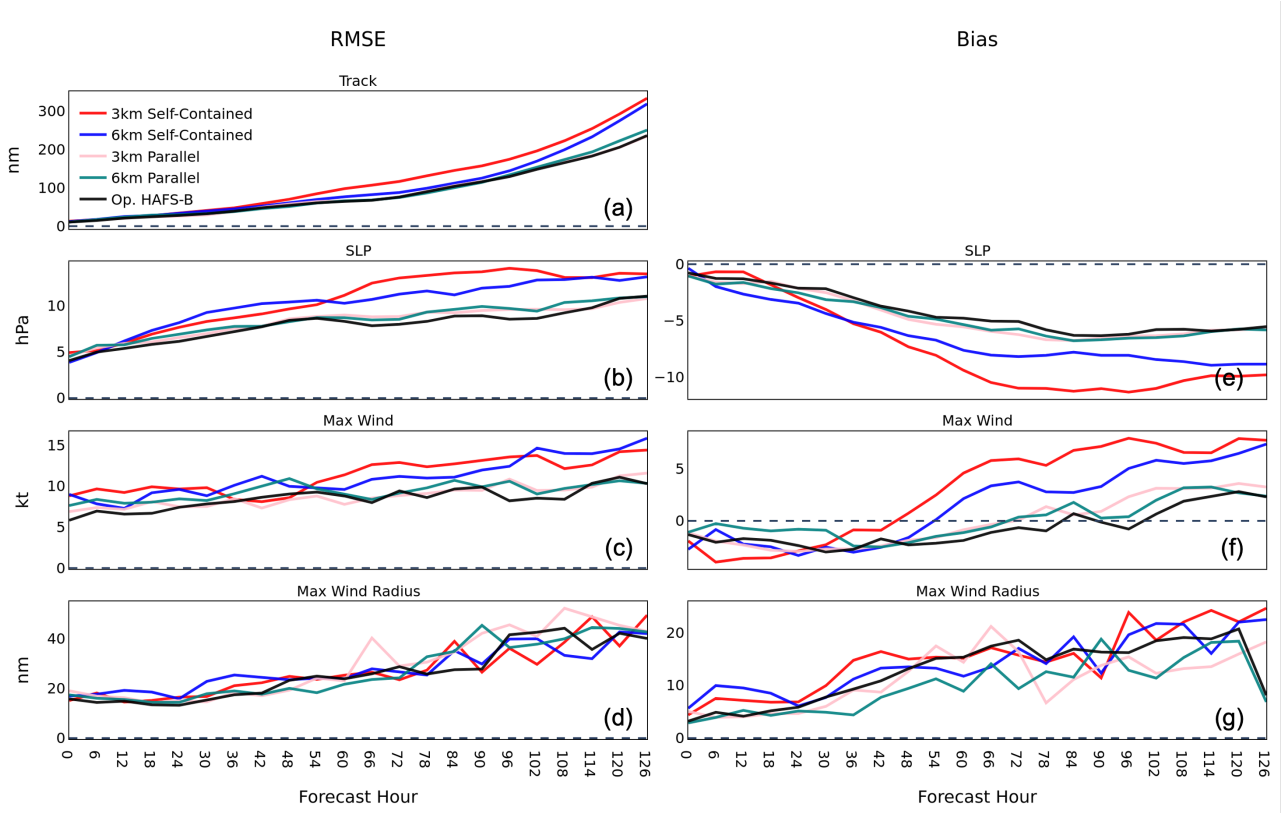


Figure 4.14: TC forecast error verification of both 3 km and 6 km self-contained and parallel workflows, and operational HAFS-B. Panels give RMSE and bias for (a) track, (b)(e) SLP, (c)(f) max wind, and (d)(g) max wind radius, respectively. Case count ranges from 61 to 59.

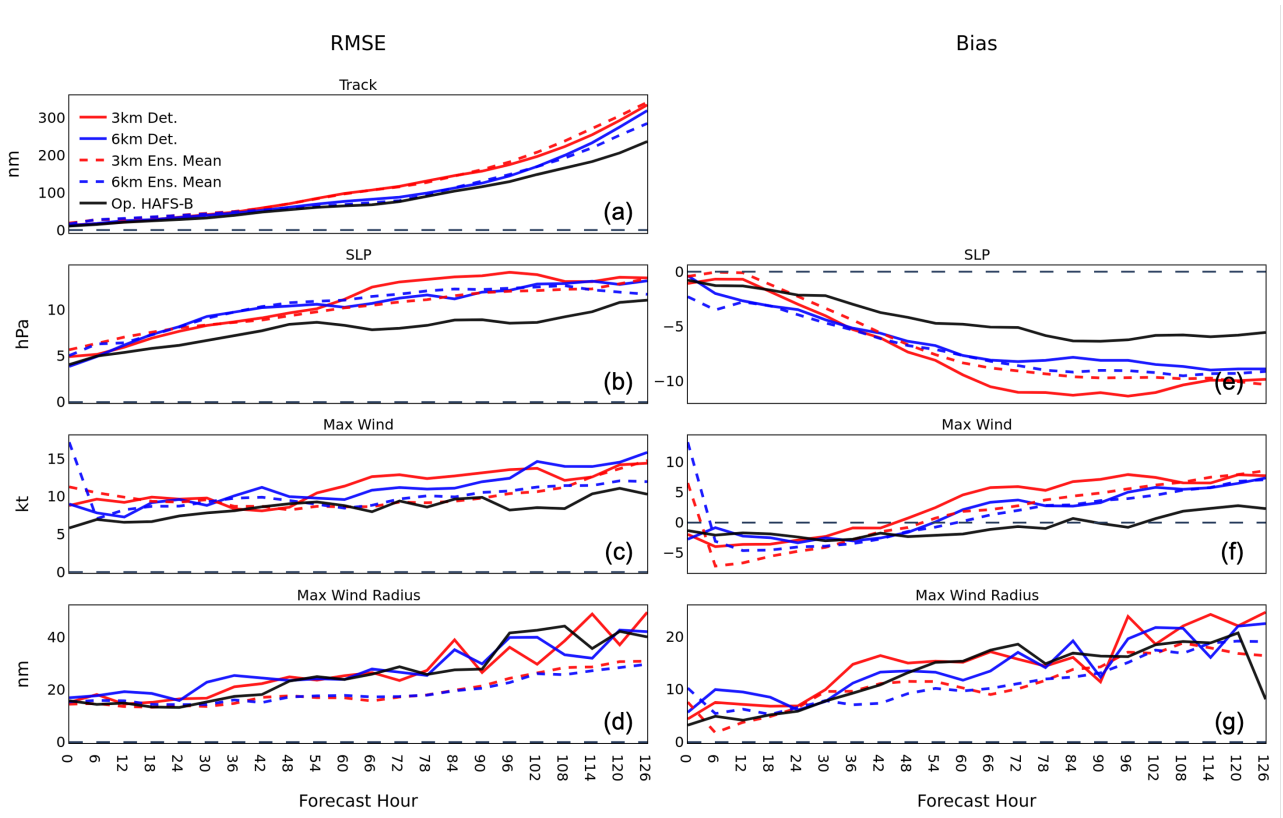


Figure 4.15: TC forecast error verification of 3 km and 6 km self-contained workflows and operational HAFS-B. Panels give RMSE and bias for (a) track, (b)(e) SLP, (c)(f) max wind, and (d)(g) max wind radius, respectively. Both the deterministic forecast and the 10 member forecast ensemble mean included for self-contained workflows. Case count ranges from 61 to 59.

3 km and 6 km self-contained forecasts have SLP biases of about -10 and -8 hPa, respectively, vs. approximately -5 hPa for the operational and parallel workflows. All 5 configurations have slightly negative max wind biases up to about 48 hours. After 48 hours, the 3 and 6 km max wind diverge to about 6 or 7 kt, while the operational and parallel forecasts remain near zero out to 84-90 hours, after which their biases become slightly positive. On the other hand, RMW results are mixed, as no single forecast experiment has noticeably lower RMSEs. All forecast configurations have positive RMW biases, although the parallel configurations tend to have slightly smaller biases than the rest.

To summarize, while the 3 and 6 km self-contained forecasts perform reasonably well with track

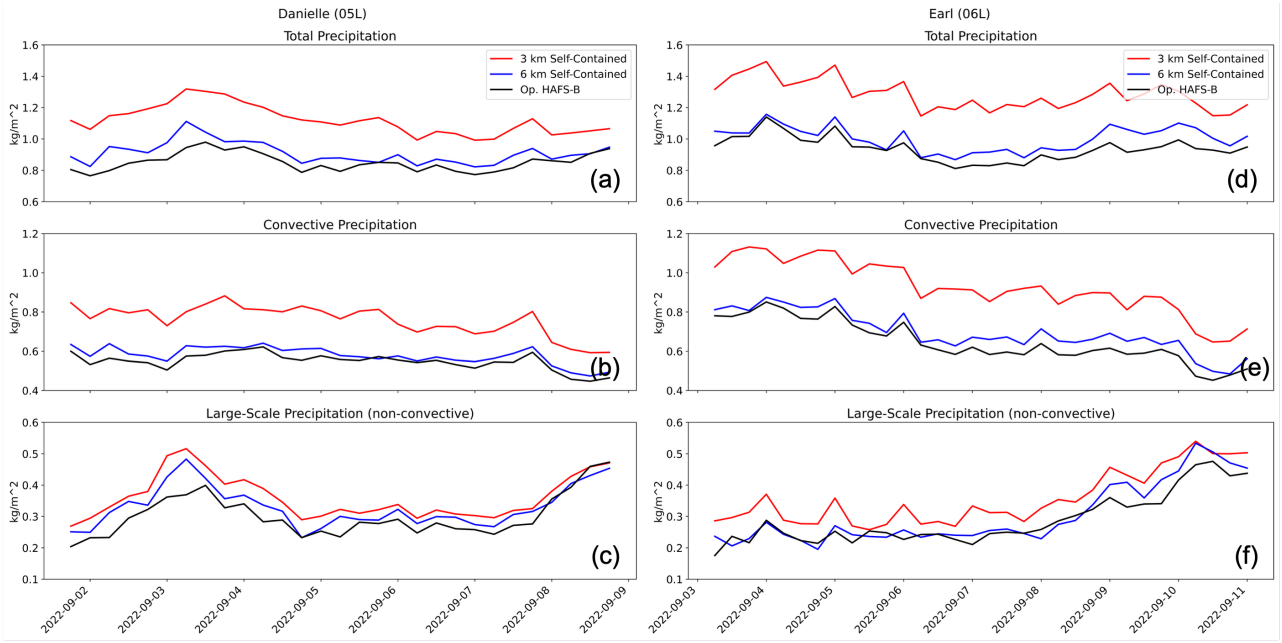


Figure 4.16: Domain-mean values over time for (a)(d) total, (b)(e) convective, and (c)(f) non-convective precipitation for Danielle and Earl, respectively. These values are taken at 6 hr forecast over storm-centric parent domain.

up to about 42 hours, their intensity forecasts tend to be considerably more intense than the best track data. The 3 km self-contained forecast in particular predicts deep and powerful storms 3 to 4 days out, although the 6 km self-contained forecast catches up by the end of the experimental period.

As further evidence, we point to Fig. 4.15, where we show the same verifications and include ens. mean forecast metrics for the 3 and 6 km self-contained frameworks. Save for a slightly smaller RMSEs and bias in RMW for the ens. mean forecasts, the results are largely the same: these forecasts produce storms that are larger and more intense than best track data. We note that the max wind spikes at analysis time are primarily the result of recentering the ensemble on the variational solution, which introduces localized wind anomalies near the vortex core. These anomalies are not representative of the broader vortex intensity and adjust rapidly.

These results suggest that deepening SLP drives the max wind bias, given that SLP bias drops below zero almost immediately while max wind lags behind by dozens of hours. To investigate

this further, we again examine precipitation values, but this time over the storm-centric forecast parent domains at 6 hour lead-times. Fig. 4.16 shows precipitation, convective precipitation, and non-convective precipitation for the self-contained and operational HAFS forecasts for both Danielle and Earl. Contrary to earlier results (Fig. 4.9), we now find that noticeably higher precipitation and convective precipitation rates in the 3 km self-contained forecast over both storms. The 6 km self-contained forecast also gives higher precipitation values than the operational HAFS, although to a lesser extent. We note that these patterns hold over the 2 km high-resolution nest domain as well, although they are not shown.

This increase in precipitation, particularly convective precipitation, may explain the enhanced intensity seen in Figs. 4.14 and 4.15. As shown in Fig. 4.11, the self-contained analyses produce enhanced low-to-mid level humidity, particularly for the 3 km configuration. The 3 km analysis favors shallow convection, leading to an accumulation of moisture and cloud liquid water around 700 mb (Fig. 4.4). Mid-tropospheric moisture is a well-established driver of deep convection in tropical cyclones, so these initial conditions provide a favorable environment for intense storm development [Gray et al., 1975]. Crucially, while the 3 km self-contained framework uses a 3 km resolution analysis, its storm-centric forecast uses a 6 km parent domain. As shown earlier, the 6 km configuration produces less statically stable conditions that favor deep convection (Figs. 4.7). It may be the case that this resolution transition from 3 km analysis to 6 km forecast releases the accumulated mid-level moisture through enhanced deep convection, increasing precipitation efficiency and diabatic heating, which deepens the TC warm core and drives more intense storms.

We note that the 6 km self-contained forecasts also exhibit enhanced mid-level humidity compared to operational HAFS initial conditions (Fig. 4.11), and correspondingly produce more intense storms, though to a lesser extent than the 3 km configuration. This suggests that the basin-scale self-cycling

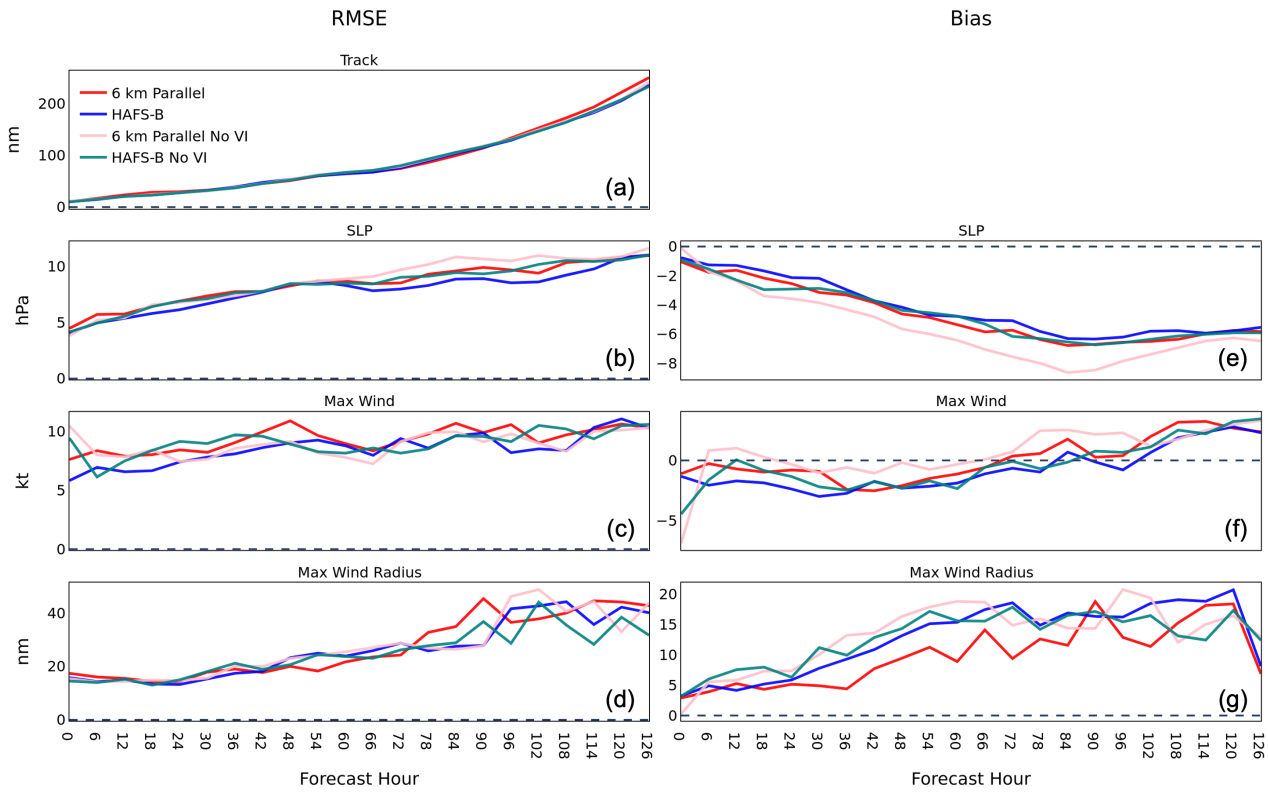


Figure 4.17: TC forecast error verification of 6 km parallel workflow and operational HAFS-B, both with and without vortex initialization. Panels give RMSE and bias for (a) track, (b)(e) SLP, (c)(f) max wind, and (d)(g) max wind radius, respectively. Case count ranges from 61 to 59.

framework itself may precondition the environment for enhanced convection, with the 3 km resolution amplifying this effect through its tendency to favor shallow convection and mid-level moisture accumulation. To investigate further, we should examine vertical profiles of diabatic heating and warm core structure, as well as secondary circulation strength.

Having examined the self-contained forecasts, we now turn to the parallel and operational configurations. Circling back to Fig. 4.14, we show the parallel and operational configurations produce very similar metrics. They both spin up storms that are too intense, albeit considerably less than the self-contained forecasts, and too large. The parallel configurations do have slightly larger intensity biases and slightly lower biases for RMW, but ultimately it may be too similar to draw firm conclusions.

The similarity of these results is not surprising given the similarity of their framework, differing only in background covariance and bias correction coefficient data. Given the heavy influence of GDAS IC on the operational/parallel configuration and the additional impact of VI, it is difficult to tease out how these differences in uncertainty quantification and bias correction, respectively, may lead to these slightly different TC-error metrics.

To investigate this further, we perform a limited set of experiments that resemble the 6 km parallel and operational HAFS forecasts but omit VI. Verification results are shown in Fig. 4.17, where we compare the 6 km parallel and operational HAFS results, both with and without VI. All four experiments have nearly identical track results, whereas intensities deviate slightly. The 6 km parallel with no VI produces a storm that is slightly more intense, with a more negative SLP bias and higher max wind bias, although these differences are small and taper off at later lead times. Meanwhile, the 6 km parallel with VI largely follows the operational forecasts, albeit still having the lowest RMW bias. Interestingly, omitting VI results in slightly stronger storms, though whether this reflects VI dampening intensity or other factors remains unclear given the limited sample size.

Both operational forecast configurations are extremely similar, such that no conclusions can be drawn. The two-storm sample may be too small to differentiate the effects of these configuration changes. Future work should focus on expanding the storm sample size, as well as examining how the different analysis updates (with and without additional high-resolution HAFS ensemble information) affect the inner-core of the storms themselves.

4.4 Conclusions

The experiments presented here demonstrate that HAFS can operate within a fully self-contained, basin-wide analysis framework that features high-resolution, self-cycled ensembles, without ad-hoc methodologies such as vortex initialization. This framework enables probabilistic ensemble-based data assimilation and long ensemble forecasts initialized from EnKF-based analyses. This framework is essential for establishing an analysis of record for Atlantic TC events. It is valuable both as a verification tool and as a training dataset for data-driven ML models. At the same time, these experiments underscore the challenges that emerge when physics suites tuned specifically for TC applications are expanded to cover the entire Atlantic basin at convection-permitting scales.

We find that the transition from 6 km to 3 km resolution produces substantial differences in cloud physics, with notable changes in the placement and concentration of hydrometeors, humidity, and cloud fraction. The 3 km configuration favors enhanced low-level cloudiness and shallow convection, while the 6 km configuration produces conditions more favorable for deep convection. These differences in cloud coverage and composition translate into systematic changes in surface radiative fluxes, which, in turn, drive air mass biases throughout the atmospheric column. Because each analysis cycle inherits the thermodynamic environment from the previous forecast, these biases compound over time through a nonlinear feedback loop: cloud differences alter surface fluxes, which affect stability, which then feeds back into cloud formation. While both configurations use identical physics parameterizations, the observed differences suggest that the schemes behave differently at the two resolutions. Isolating the precise mechanisms remains a subject for future work.

These differences in analysis have meaningful consequences for TC prediction. For Hurricanes Danielle and Earl, the self-contained forecasts produce storms that are considerably more intense than

best track observations. We suggest that enhanced mid-level humidity in the self-contained analyses preconditions the environment for deep convection, leading to overly intense storms. This effect is most substantial for the 3 km configuration. However, it is also present in the 6 km self-contained forecasts, suggesting that the basin-scale cycling framework itself contributes to this preconditioning.

A key motivation for this work is to isolate deficiencies in the HAFS modeling system that would otherwise be difficult to diagnose given the heavy influence of external models on operational configurations. The self-contained framework removes dependence on GDAS initial conditions and externally trained bias correction coefficients, allowing HAFS-specific biases to emerge clearly. The results presented here provide actionable information for tuning future HAFS configurations, particularly regarding resolution-dependent differences in cloud physics and their effects on TC forecast performance. We emphasize that the goal is not necessarily to outperform the operational system, which benefits from well-tuned GDAS initial conditions, but rather to understand and address fundamental limitations in the regional LAM. We note that this framework is computationally expensive and complex, and addressing the biases identified here will require significant resources and sustained effort.

Future work should focus on diagnosing and tuning model physics to address the cloud and air mass biases that emerge at convection-permitting scales across the basin. In particular, understanding how the scale-aware convection and microphysics schemes behave differently at 3 km vs. 6 km resolution is essential for improving both the self-contained framework and operational HAFS configurations. Finally, this framework provides the foundation for an analysis of record for Atlantic TC events, which would support verification of operational forecasts, retrospective analysis of past events, and the development of training datasets for machine learning applications. While challenges remain, this study demonstrates that a self-contained, probabilistic HAFS framework is achievable and offers a path toward improved TC prediction.

Chapter 5: Summary, Future Work, and Conclusions

5.1 Chapter Summaries

In this dissertation, we aim to explore research that advances TC and hurricane forecasting towards a more mature, self-contained, and probabilistic modeling framework. In particular, we focus on improving observation assimilation through understanding fundamental bias correction limitations, as well as building and testing stand-alone ensemble modeling applications.

In Chapter 1, we discuss current challenges in TC modeling, particularly in limited-area partially cycled modeling frameworks. We introduce and explain the purpose of data assimilation and provide a brief overview of the DA methodologies used throughout this dissertation research. Next, we discuss the assimilation of satellite radiance measurements into NWP models, highlighting their strengths as an observation network but also emphasizing their strongly biased nature. This naturally leads into the necessity of bias correction, and the challenges modelers face when attempting to estimate and separate different sources of bias from observation bias.

Chapter 2 is directed towards the improvement of observation bias correction methodologies, particularly when the addition of undiagnosed model bias complicates bias estimation. We approach the problem through experiments with the Lorenz [2005] Model III, an idealized chaotic dynamical model that produces model states that roughly resemble planetary waves. In order to identify fundamental limitations, we impose static, global biases and employ straightforward bias correction methodologies that exploit low-bias anchoring observations to separate observation and model process error. We find that anchor observation-based methodologies can successfully separate biases near anchor observations, but that without sufficient coverage, these methods fail, resulting in bias reinforcement and analysis drift

over many cycles. In sparse-anchor regimes, ignoring biased observations altogether may be preferable to flawed corrections. Ultimately, this research emphasizes the essential nature of both dense anchor observation networks and the reduction of model bias for successful observation bias estimation and assimilation.

Chapter 3 focuses on bias correction of satellite radiance measurements within the HAFS system. Specifically, we compare the results of online bias correction in which bias correction coefficients are estimated on the native HAFS modeling grid, vs. adopting bias coefficients from an external global modeling system, the GDAS. we perform fully-cycled HAFS experiments over a 6-week period during the 2020 hurricane season, and find that online bias correction improves TC forecasts, particularly for storms with extensive inner-core anchoring observations. This research has two important implications. First, it shows that self-contained online bias correction within a limited-area TC model is feasible and can be beneficial. Second, performing online bias correction removes one more source of external model influence, allowing us to assess and fix biases within the HAFS system itself more precisely.

In Chapter 4, we extend this fully-cycled HAFS framework to include basin-wide ensemble DA and forecasting. The goal here is to explore the obstacles to high-resolution basin-wide DA for TC forecasting. Given that limited-area TC models are usually tuned to characterize near-storm environments, extending the analysis and forecast domain to basin-wide scales while maintaining resolutions comparable to those in high-resolution storm-centric nests poses significant challenges for model physics and dynamics. We compare 3 km and 6 km basin-wide analysis model states, highlighting differences in physics and significant domain-wide biases. From these analyses, we generate storm-centric ensemble forecasts for individual storms and compare their performance with the operational HAFS-B model configuration. We also examine the performance of configurations that closely resemble the operational system but that include covariance information from ensembles

cycled within the basin-wide framework. Changes in resolution produce substantial differences in cloud coverage/composition, which in turn produce major domain-wide air mass biases. Altogether, these differences in forecasted initial conditions have substantial impacts on TC forecasts. While challenges remain for basin-wide high resolution TC modeling, this study has shown that there are clear patterns in model dynamics and physics that should be addressed to improve both current operational HAFS and advance self-cycled basin-scale TC modeling. This work also demonstrates the feasibility of basin-wide ensemble DA, a necessary step toward establishing an analysis of record for Atlantic TC events.

5.2 Future Work

In this section, we focus on future work within the self-cycled basin-scale HAFS framework established in Chapters 3 and 4. In additional to future work focused on diagnosing and tuning model physics to address the cloud and air mass biases that emerge at convection-permitting scales across the basin, future research must address underdispersed spread in the ensemble DA and forecast systems.

Fig. 5.1 shows TC forecast-specific error metrics verified against best track data, specifically: track, minimum SLP, 10-m maximum winds, and RMW. These statistics are summarized for both Danielle and Earl, where mean forecast RMSEs and bias are calculated over a sample of cases and plotted as a function of forecast lead-time. In this example, we plot 3 km and 6 km self-contained ens. mean forecast results along with their respective ensemble spread. We find that the ensemble forecasts are dramatically underdispersed, particularly with regards to track and SLP. This has practical consequences for both DA and probabilistic severe weather forecasting. In cycling DA systems, persistent underdispersion can lead to filter divergence, where the ensemble becomes so narrow that observations are underweighted and the analysis drifts from the true atmospheric state. For severe

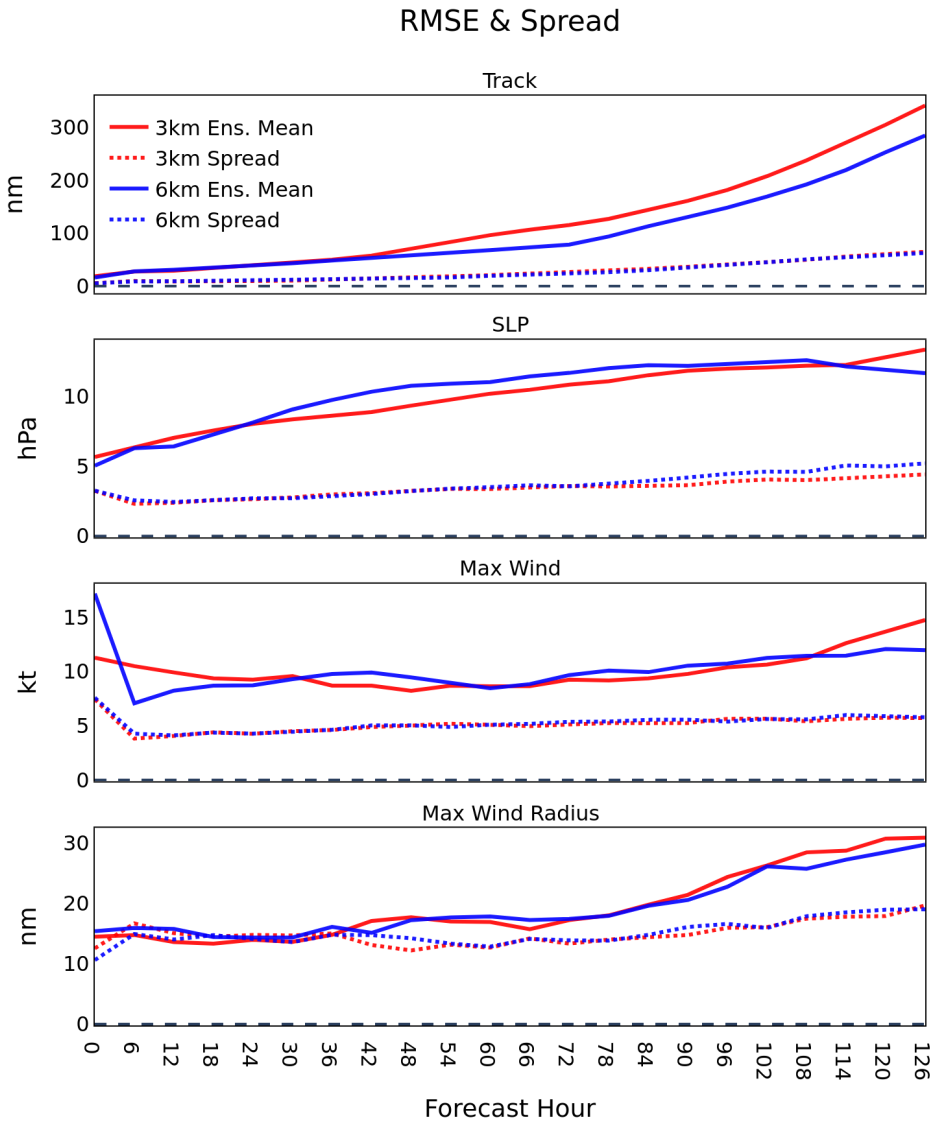


Figure 5.1: Forecast RMSE verification of track, SLP, max wind, and max wind radius for 10 member forecast ensemble mean for 3 km and 6 km self-contained workflows, with corresponding ensemble spread. Case count ranges from 61 to 59.

weather forecasting, underdispersed forecasts can produce overconfident predictions that can potentially underestimate or misconstrue the probability of high-impact scenarios.

One promising approach to address underdispersion is the LPF. Unlike the EnKF, the LPF does not rely solely on covariances to describe background uncertainty, making it better suited for nonlinear model dynamics and measurement operators [Kurosawa and Poterjoy, 2021, Poterjoy, 2022b]. This is particularly beneficial for TC applications, where near-storm dynamics are strongly nonlinear and satellite radiance observations, which have nonlinear measurement operators, are critical [Emanuel and Zhang, 2016a]. Prior studies have shown that replacing EnKF-based systems with LPF-based frameworks can significantly reduce sampling deficiency [Kurosawa and Poterjoy, 2025, Poterjoy, 2022a]. In ongoing work, we have implemented the LPF within the basin-scale HAFS analysis system from Chapter 4, replacing the EnKF for updating flow-dependent error covariances in the hybrid variational-ensemble DA system. Preliminary results are consistent with prior findings.

5.3 Conclusions

This research advances three interconnected components of TC model development: improving observational strategies for satellite radiance assimilation; isolating and diagnosing sources of bias within the HAFS system; and building toward a self-contained, probabilistic TC analysis framework. Taken together, the findings across all three studies converge on a central theme: that diagnosing and reducing model bias is a prerequisite for meaningful progress in any of these areas.

The theoretical framework developed in Chapter 2 makes this point most explicitly. No observation bias correction strategy, however simple or sophisticated, can fully compensate for undiagnosed model process error. This conclusion carries direct practical weight. In operational

TC forecasting, radiance bias correction is performed continuously over data-sparse ocean regions where anchor observations are sparse and model bias goes largely undetected. The results here suggest that in such regimes, current correction strategies may actively degrade analyses rather than improve them. Chapter 3 reinforces this by demonstrating that when HAFS trains its own bias correction coefficients natively, rather than inheriting them from GDAS, the system can better isolate and correct for its own biases rather than absorbing those of an external model, with subsequent improvements in forecast skill.

This progression from idealized theory to operational application also highlights a broader challenge for the field: the lack of an Atlantic TC analysis of record. Without a consistent, high-quality reanalysis dataset, it is difficult to verify whether forecast improvements reflect genuine advances in model skill. The self-contained, continuously-cycled ensemble framework demonstrated in Chapter 4 represents a step toward filling this gap. In addition to forecast verification, such a reanalysis would provide the high-resolution, physically consistent training data increasingly demanded by AI/ML-based prediction systems. This need will only grow as data-driven approaches become more relevant to operational forecasting.

Ultimately, the path toward a mature, probabilistic TC forecasting system requires denser and more accurate observation networks, better-constrained model physics across wider domains, and ensemble DA methodologies capable of characterizing flow-dependent covariance structures unique to the problem of TC evolution. This dissertation establishes that such a system is achievable, and identifies the scientific and practical obstacles that must be overcome to get there.

Bibliography

A. Aksoy, S. Lorsolo, T. Vukicevic, K. J. Sellwood, S. D. Aberson, and F. Zhang. The hwrf hurricane ensemble data assimilation system (hedas) for high-resolution data: The impact of airborne doppler radar observations in an osse. *Monthly Weather Review*, 140:1843–1862, 2012.

F. Alet, I. Price, A. El-Kadi, D. Masters, S. Markou, T. R. Andersson, J. Stott, R. Lam, M. Willson, A. Sanchez-Gonzalez, and P. Battaglia. Skillful joint probabilistic weather forecasting from marginals, 2025.

J. M. Aparicio and S. Laroche. Estimation of the added value of the absolute calibration of GPS radio occultation data for numerical weather prediction. 143(4):1259–1274, 2015. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-14-00153.1. URL https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00153.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

T. Auligné, A. P. McNally, and D. P. Dee. Adaptive bias correction for satellite data in a numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society*, 133(624): 631–642, 2007. doi: 10.1002/qj.56. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.56.

R. N. Bannister. A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143:607–633, 2017.

M. J. Bell, M. J. Martin, and N. K. Nichols. Assimilation of data into an ocean model with systematic errors near the equator. 130(598):873–893, 2004. ISSN 1477-870X. doi: 10.1256/qj.02.109. URL https://onlinelibrary.wiley.com/doi/abs/10.1256/qj.02.109.

P. Benáček and M. Mile. Satellite bias correction in the regional model aladin/cz: Comparison of different varbc approaches. *Monthly Weather Review*, 147(9):3223 – 3239, 2019. doi: 10.1175/MWR-D-18-0359.1. URL https://journals.ametsoc.org/view/journals/mwre/147/9/mwr-d-18-0359.1.xml.

K. Bhargava, E. Kalnay, J. A. Carton, and F. Yang. Estimation of systematic errors in the GFS using analysis increments. 123(3):1626–1637, 2018. ISSN 2169-897X, 2169-8996. doi: 10.1002/2017JD027423. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/2017JD027423.

M. K. Biswas, D. Stark, and L. Carson. *GFDL Vortex Tracker*, 2018.

M. Bonavita and P. Lean. 4d-var for numerical weather prediction. 76(2):65–66, 2021. ISSN 1477-8696. doi: 10.1002/wea.3862. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wea.3862. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/wea.3862.

Z. B. Bouallègue, M. C. A. Clare, L. Magnusson, E. Gascón, M. Maier-Gerber, M. Janoušek, M. Rodwell, F. Pinault, J. S. Dramsch, S. T. K. Lang, B. Raoult, F. Rabier, M. Chevallier, I. Sandu, P. Dueben, M. Chantry, and F. Pappenberger. The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-23-0162.1. URL https://journals.ametsoc.org/view/journals/bams/105/6/BAMS-D-23-0162.1.xml. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

N. E. Bowler, A. M. Clayton, M. Jardak, E. Lee, A. C. Lorenc, C. Piccolo, S. R. Pring, M. A. Wlasak, D. M. Barker, G. W. Inverarity, and R. Swinbank. Inflation and localization tests in the development of an ensemble of 4d-ensemble variational assimilations. *Quarterly Journal of the Royal Meteorological Society*, 143(704):1280–1302, 2017. doi: 10.1002/qj.3004. URL https://onlinelibrary.wiley.com/doi/10.1002/qj.3004.

G. H. Bryan, J. C. Wyngaard, and J. M. Fritsch. Resolution Requirements for the Simulation of Deep Moist Convection. *Monthly Weather Review*, 131(10):2394–2416, Oct. 2003. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(2003)131⟨2394:RRFTSO⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/131/10/1520-0493_2003_131_2394_rrftso_2.0.co_2.xml.

M. Buehner, R. McTaggart-Cowan, A. Beaulne, C. Charette, L. Garand, S. Heilliette, E. Lapalme, S. Laroche, S. R. Macpherson, J. Morneau, and A. Zadra. Implementation of Deterministic Weather Forecasting Systems Based on Ensemble–Variational Data Assimilation at Environment Canada. Part I: The Global System. *Monthly Weather Review*, 143(7):2532–2559, July 2015. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-14-00354.1. URL https://journals.ametsoc.org/view/journals/mwre/143/7/mwr-d-14-00354.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

G. Burgers, P. J. v. Leeuwen, and G. Evensen. Analysis scheme in the ensemble kalman filter. 126 (6):1719–1724, 1998. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1998)126⟨1719:ASITEK⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/126/6/1520-0493_1998_126_1719_asitek_2.0.co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

W. F. Campbell, C. H. Bishop, and D. Hodyss. Vertical covariance localization for satellite radiances in ensemble kalman filters. *Monthly Weather Review*, 138(1):282–290, Jan 2010. doi: 10.1175/2009MWR3017.1. URL https://journals.ametsoc.org/view/journals/mwre/138/1/2009mwr3017.1.xml.

J. P. Cangialosi, E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada. Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather and Forecasting*, 35(5):1913–1922, 2020. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-20-0059.1. URL https://journals.ametsoc.org/view/journals/wefo/35/5/wafD200059.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

C. Cardinali. Monitoring the observation impact on the short-range forecast. 135(638):239–250, 2009. ISSN 1477-870X. doi: 10.1002/qj.366. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.366. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.366.

L. E. Carr and R. L. Elsberry. Models of tropical cyclone wind distribution and beta-effect propagation for application to tropical cyclone track forecasting. *Monthly Weather Review*, 125(12):3190 – 3209, 1997. doi: 10.1175/1520-0493(1997)125⟨3190:MOTCWD⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/125/12/1520-0493_1997_125_3190_motcwd_2.0.co_2.xml.

A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. 9(5):e535, 2018. ISSN 1757-7799. doi: 10.1002/wcc.535. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.535.

K. Chandramouli, X. Wang, A. Johnson, and J. Otkin. Online Nonlinear Bias Correction in Ensemble Kalman Filter to Assimilate GOES-R All-Sky Radiances for the Analysis and Prediction of Rapidly Developing Supercells. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002711, 2022. ISSN 1942-2466. doi: 10.1029/2021MS002711. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002711. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002711.

E. P. Chassignet, H. E. Hurlburt, O. M. Smedstad, G. R. Halliwell, P. J. Hogan, A. J. Wallcraft, R. Baraille, and R. Bleck. The HYCOM (HYbrid Coordinate Ocean Model) data assimilative system. *Journal of Marine Systems*, 65:60–83, 2007.

E. P. Chassignet, H. E. Hurlburt, E. J. Metzger, O. M. Smedstad, J. A. Cummings, G. R. Halliwell, R. Bleck, R. Baraille, A. J. Wallcraft, C. Lozano, H. L. Tolman, A. Srinivasan, S. Hankin, P. Cornillon, R. Weisberg, A. Barth, R. He, F. Werner, and J. Wilkin. Us godae: Global ocean prediction with the hybrid coordinate ocean model (hycom). *Oceanography*, 22(2):64–75, 2009. ISSN 10428275, 2377617X. URL http://www.jstor.org/stable/24860960.

A. M. Clayton, A. C. Lorenc, and D. M. Barker. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quarterly Journal of the Royal Meteorological Society*, 139(675):1445–1461, 2013. ISSN 1477-870X. doi: 10.1002/qj.2054. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2054. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2054.

A. D. Collard and A. P. McNally. The assimilation of infrared atmospheric sounding interferometer radiances at ecmwf. *Quarterly Journal of the Royal Meteorological Society*, 135:1044–1058, 2009.

P. Courtier, J.-N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994. ISSN 1477-870X. doi: 10.1002/qj.49712051912. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712051912. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49712051912.

W. Crawford, S. Frolov, J. McLay, C. A. Reynolds, N. Barton, B. Ruston, and C. H. Bishop. Using analysis corrections to address model error in atmospheric forecasts. *Monthly Weather Review*, 148(9): 3729–3745, Sep 2020. doi: 10.1175/MWR-D-20-0008.1. URL http://journals.ametsoc.org/doi/10.1175/MWR-D-20-0008.1.

C. M. Danforth and E. Kalnay. Impact of online empirical model correction on nonlinear error growth. 35(24), 2008. ISSN 1944-8007. doi: 10.1029/2008GL036239. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2008GL036239. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008GL036239.

D. P. Dee. Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613): 3323–3343, 2005. doi: 10.1256/qj.05.137. URL https://onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.137.

D. P. Dee and A. M. Da Silva. Data assimilation in the presence of forecast bias. *Quarterly Journal of the Royal Meteorological Society*, 124(545):269–295, 1998. doi: 10.1002/qj.49712454512. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712454512.

D. P. Dee and S. Uppala. Variational bias correction of satellite radiance data in the era-interim reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 135(644):1830–1841, 2009. doi: 10.1002/qj.493. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.493.

M. DeMaria, J. L. Franklin, G. Chirokova, J. Radford, R. DeMaria, K. D. Musgrave, and I. Ebert-Uphoff. Evaluation of Tropical Cyclone Track and Intensity Forecasts from Artificial Intelligence Weather Prediction (AIWP) Models, Sept. 2024. URL http://arxiv.org/abs/2409.06735. arXiv:2409.06735 [physics].

G. Deng, D. Zhang, T. Zhu, and A. Wang. Use of the advanced microwave sounding unit data to improve typhoon prediction. *Progress in Natural Science*, 19:369–376, 2009.

J. C. Derber and W.-S. Wu. The Use of TOVS Cloud-Cleared Radiances in the NCEP SSI Analysis System. *Monthly Weather Review*, 126:2287–2299, 1998.

J. Dong, B. Liu, Z. Zhang, W. Wang, A. Mehra, A. T. Hazelton, H. R. Winterbottom, L. Zhu, K. Wu, C. Zhang, V. Tallapragada, X. Zhang, S. Gopalakrishnan, and F. Marks. The Evaluation of Real-Time Hurricane Analysis and Forecast System (HAFS) Stand-Alone Regional (SAR) Model Performance for the 2019 Atlantic Hurricane Season. *Atmosphere*, 11:617, 2020.

K. Emanuel and F. Zhang. On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences*, 73(9):3739 – 3747, 2016a. doi: 10.1175/JAS-D-16-0100.1. URL https://journals.ametsoc.org/view/journals/atsc/73/9/jas-d-16-0100.1.xml.

K. Emanuel and F. Zhang. On the Predictability and Error Sources of Tropical Cyclone Intensity Forecasts. *Journal of the Atmospheric Sciences*, 73(9):3739–3747, Sept. 2016b. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-16-0100.1. URL https://journals.ametsoc.org/view/journals/atsc/73/9/jas-d-16-0100.1.xml.

G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. 99:10143–10162, 1994. ISSN 2156-2202. doi: 10.1029/94JC00572. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/94JC00572.

J. Eyre. A bias correction scheme for simulated TOVS brightness temperatures, Oct. 1992. URL https://www.ecmwf.int/node/9330. Issue: 186 Pages: 35 Place: Shinfield Park, Reading Publication Title: ECMWF Technical Memoranda.

J. R. Eyre. Observation bias correction schemes in data assimilation systems: a theoretical study of some of their properties. *Quarterly Journal of the Royal Meteorological Society*, 142(699):2284–2291, 2016. doi: 10.1002/qj.2819. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2819.

J. R. Eyre, W. Bell, J. Cotton, S. J. English, M. Forsythe, S. B. Healy, and E. G. Pavelin. Assimilation of satellite data in numerical weather prediction. part II: Recent years. 148(743):521–556, 2022. ISSN 1477-870X. doi: 10.1002/qj.4228. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4228. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4228.

D. Francis, A. Fowler, A. Lawless, J. Eyre, and S. Migliorini. The effective use of anchor observations in variational bias correction in the presence of model bias. *Quarterly Journal of the Royal Meteorological Society*, 149, May 2023. doi: 10.1002/qj.4482.

N. L. Frank and S. A. Husain. the deadliest tropical cyclone in history? *Bulletin of the American Meteorological Society*, 52(6):438–445, June 1971. ISSN 0003-0007, 1520-0477. doi: 10.1175/1520-0477(1971)052⟨0438:TDTCIH⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/bams/52/6/1520-0477_1971_052_0438_tdtcih_2_0_co_2.xml.

R. Gall, J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer. The Hurricane Forecast Improvement Project. *Bulletin of the American Meteorological Society*, 94(3):329–343, Mar. 2013. doi: 10.1175/BAMS-D-12-00071.1. URL https://journals.ametsoc.org/view/journals/bams/94/3/bams-d-12-00071.1.xml.

A. J. Geer and P. Bauer. Observation errors in all-sky data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 137:2024–2037, 2011a.

A. J. Geer and P. Bauer. Observation errors in all-sky data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 137(661):2024–2037, 2011b. ISSN 1477-870X. doi: 10.1002/qj.830. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.830. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.830.

M. D. Goldberg, D. S. Crosby, and L. Zhou. The limb adjustment of amsu-a observations: Methodology and validation. *Journal of Applied Meteorology*, 40(1): 70 – 83, 2001. doi: 10.1175/1520-0450(2001)040⟨0070:TLAOAA⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/40/1/1520-0450_2001_040_0070_tlaoaa_2.0.co_2.xml.

S. Gopalakishnan, S. Upadhayay, Y. Jung, F. Marks, V. Tallapragada, A. Mehra, M. Brennan, M. DeMaria, A. Aksoy, G. J. Alaka, C. Alexander, M. Bender, L. Bernardet, M. Biswas, J. Cangialosi, J. Dong, R. Dunlap, M. Ek, J. L. Franklin, L. Gramer, L. Harris, A. Hazelton, J. S. Hilderbrand, E. Kalina, H.-S. Kim, D. Koch, P. Kucera, B. Liu, P. McCaslin, T. Marchok, J. Moskaitis, K. Musgrave, L. Nance, K. Newman, M. Onderlinde, A. Penny, W. Ramstrom, E. N. Rappaport, J. Sims, J. Sippel, R. Torn, X. Wang, W. Wang, Y. Weng, J. Whitaker, H. Winterbottom, B. C. Zachry, D. A. Zelinsky, C. Zhang, X. Zhang, Z. Zhang, and L. Zhu. 2020 HFIP R&D Activities Summary: Recent Results and Operational Implementation. Technical report, United States. National Oceanic and Atmospheric Administration. Office of Oceanic and Atmospheric Research. and National Weather Service (U.S.), 2021. URL https://repository.library.noaa.gov/view/noaa/31443.

W. M. Gray et al. Tropical cyclone genesis. 1975.

T. M. Hamill and C. Snyder. A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. *Monthly Weather Review*, 128(8):2905–2919, Aug. 2000. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(2000)128⟨2905:AHEKFV⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/128/8/1520-0493_2000_128_2905_ahekfv_2.0.co_2.xml.

J. Han and C. S. Bretherton. Tke-based moist eddy-diffusivity mass-flux (edmf) parameterization for vertical turbulent mixing. *Weather and Forecasting*, 34(4):869 – 886, 2019. doi: 10.1175/WAF-D-18-0146.1. URL https://journals.ametsoc.org/view/journals/wefo/34/4/waf-d-18-0146_1.xml.

J. Han and H.-L. Pan. Revision of convection and vertical diffusion schemes in the NCEP global forecast system. *Weather and Forecasting*, 26(4):520–533, 2011. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-10-05038.1. URL https://journals.ametsoc.org/view/journals/wefo/26/4/waf-d-10-05038_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

J. Han, W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang. Updates in the ncep gfs cumulus convection schemes with scale and aerosol awareness. *Weather and Forecasting*, 32(5):2005 – 2017, 2017a. doi: 10.1175/WAF-D-17-0046.1. URL https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0046_1.xml.

J. Han, W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang. Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Weather and Forecasting*, 32(5):2005–2017, 2017b. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-17-0046.1. URL https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0046_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

D. Hanley, J. Molinari, and D. Keyser. A composite study of the interactions between tropical cyclones and upper-tropospheric troughs. *Monthly Weather Review*, 129(10):2570 – 2584, 2001. doi: 10.1175/1520-0493(2001)129⟨2570:ACSOTI⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/129/10/1520-0493_2001_129_2570_acsoti_2.0.co_2.xml.

B. A. Harris and G. Kelly. A satellite radiance-bias correction scheme for data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 127(574):1453–1468, 2001. doi: 10.1002/qj.49712757418. URL http://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712757418.

C. M. Hartman, X. Chen, E. E. Clothiaux, and M.-Y. Chan. Improving the analysis and forecast of hurricane dorian (2019) with simultaneous assimilation of goes-16 all-sky infrared brightness temperatures and tail doppler radar radial velocities. *Monthly Weather Review*, 149(7):2193 – 2212, 2021. doi: 10.1175/MWR-D-20-0338.1. URL https://journals.ametsoc.org/view/journals/mwre/149/7/MWR-D-20-0338.1.xml.

A. Hazelton, Z. Zhang, B. Liu, J. Dong, G. Alaka, W. Wang, T. Marchok, A. Mehra, S. Gopalakrishnan, X. Zhang, M. Bender, V. Tallapragada, and F. Marks. 2019 Atlantic Hurricane Forecasts from the Global-Nested Hurricane Analysis and Forecast System: Composite Statistics and Key Events. *Weather and Forecasting*, 36:519–538, 2021.

A. Hazelton, G. J. Alaka, L. Gramer, W. Ramstrom, S. Ditchek, X. Chen, B. Liu, Z. Zhang, L. Zhu, W. Wang, B. Thomas, J. Shin, C.-K. Wang, H.-S. Kim, X. Zhang, A. Mehra, F. Marks, and S. Gopalakrishnan. 2022 real-time hurricane forecasts from an experimental version of the hurricane analysis and forecast system (HAFSV0.3s). *Frontiers in Earth Science*, 11, 2023. ISSN 2296-6463. doi: 10.3389/feart.2023.1264969. URL https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2023.1264969/full. Publisher: Frontiers.

A. Hazelton, X. Chen, G. J. Alaka, G. R. Alvey, S. Gopalakrishnan, and F. Marks. Sensitivity of HAFS-b tropical cyclone forecasts to planetary boundary layer and microphysics parameterizations. *Weather and Forecasting*, 39(4):655–678, 2024. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-23-0124.1. URL https://journals.ametsoc.org/view/journals/wefo/39/4/WAF-D-23-0124.1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

A. T. Hazelton, L. Harris, and S.-J. Lin. Evaluation of tropical cyclone structure forecasts in a high-resolution version of the multiscale gfdl fvgfs model. *Weather and Forecasting*, 33(2):419 – 442, 2018. doi: 10.1175/WAF-D-17-0140.1. URL https://journals.ametsoc.org/view/journals/wefo/33/2/waf-d-17-0140_1.xml.

E. A. Hendricks, M. S. Peng, B. Fu, and T. Li. Quantifying environmental control on tropical cyclone intensity change. *Monthly Weather Review*, 138(8):3243 – 3271, 2010. doi: 10.1175/2010MWR3185.1. URL https://journals.ametsoc.org/view/journals/mwre/138/8/2010mwr3185.1.xml.

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume,

and J.-N. Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi-org.proxy-um.researchport.umd.edu/10.1002/qj.3803. URL https://rmets-onlinelibrary-wiley-com.proxy-um.researchport.umd.edu/doi/abs/10.1002/qj.3803.

P. L. Houtekamer and H. L. Mitchell. Data Assimilation Using an Ensemble Kalman Filter Technique. *Monthly Weather Review*, 126(3):796–811, Mar. 1998. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1998)126⟨0796:DAUAEK⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/126/3/1520-0493_1998_126_0796_dauaek_2.0.co_2.xml.

M. Hu, C. Zhou, H. Shao, D. Stark, and K. Newman. *GSI Advanced User's Guide*, 2016.

Hurricane Analysis and Forecast System, 2025. Hurricane analysis and forecast system, 2025. URL https://emc.ncep.noaa.gov/hurricane/HFSB/about.php?branch=impl.

N. A. Jacobs. Open Innovation and the Case for Community Model Development. *Bulletin of the American Meteorological Society*, 102(10):E2002–E2011, Oct. 2021. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-21-0030.1. URL https://journals.ametsoc.org/view/journals/bams/102/10/BAMS-D-21-0030.1.xml.

T. A. Jones, X. Wang, P. Skinner, A. Johnson, and Y. Wang. Assimilation of goes-13 imager clear-sky water vapor (6.5 μm) radiances into a warn-on-forecast system. *Monthly Weather Review*, 146(4): 1077 – 1107, 2018. doi: 10.1175/MWR-D-17-0280.1. URL https://journals.ametsoc.org/view/journals/mwre/146/4/mwr-d-17-0280.1.xml.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, Mar. 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003. ISBN 978-0-521-79179-3.

J. Kaplan and M. DeMaria. Large-scale characteristics of rapidly intensifying tropical cyclones in the north atlantic basin. *Weather and Forecasting*, 18(6):1093–1108, 2003. ISSN 1520-0434, 0882-8156. doi: 10.1175/1520-0434(2003)018⟨1093:LCORIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/18/6/1520-0434_2003_018_1093_lcorit_2_0_co_2.xml.

Y.-J. Kim and A. Arakawa. Improvement of orographic gravity wave parameterization using a mesoscale gravity wave model. *Journal of Atmospheric Sciences*, 52(11): 1875 – 1902, 1995. doi: 10.1175/1520-0469(1995)052⟨1875:IOOGWP⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/52/11/1520-0469_1995_052_1875_ioogwp_2_0_co_2.xml.

N. Kishore, D. Marqués, A. Mahmud, M. V. Kiang, I. Rodriguez, A. Fuller, P. Ebner, C. Sorensen, F. Racy, J. Lemery, L. Maas, J. Leaning, R. A. Irizarry, S. Balsari, and C. O. Buckee. Mortality in Puerto Rico after Hurricane Maria. *New England Journal of Medicine*, 379(2):162–170, July 2018.

ISSN 0028-4793. doi: 10.1056/NEJMsa1803972. URL https://www.nejm.org/doi/full/10.1056/NEJMsa1803972. _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMsa1803972.

D. Kleist, J. R. Carley, A. Collard, E. Liu, S. Liu, C. R. Martin, C. Thomas, R. Treadon, and G. Vernieres. Current state of data assimilation capabilities at NCEP's environmental modeling center, 2023. URL https://repository.library.noaa.gov.

J. Knisely and J. Poterjoy. Implications of Self-Contained Radiance Bias Correction for Data Assimilation within the Hurricane Analysis and Forecasting System (HAFS). *Weather and Forecasting*, 38(9):1719–1738, Aug. 2023. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-23-0027.1. URL https://journals.ametsoc.org/view/journals/wefo/38/9/WAF-D-23-0027.1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

K. Kurosawa and J. Poterjoy. Data assimilation challenges posed by nonlinear operators: A comparative study of ensemble and variational filters and smoothers. *Monthly Weather Review*, 149(7):2369–2389, 2021. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-20-0368.1. URL https://journals.ametsoc.org/view/journals/mwre/149/7/MWR-D-20-0368.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

K. Kurosawa and J. Poterjoy. A statistical hypothesis testing strategy for adaptively blending particle filters and ensemble kalman filters for data assimilation. *Monthly Weather Review*, 151(1):105–125, 2023. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-22-0108.1. URL https://journals.ametsoc.org/view/journals/mwre/151/1/MWR-D-22-0108.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

K. Kurosawa and J. Poterjoy. Augmented flow-dependent perturbations to mitigate sampling errors: Experiments for a regional application of the NOAA unified forecast system. *Weather and Forecasting*, 40(7):1147–1161, 2025. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-24-0147.1. URL https://journals.ametsoc.org/view/journals/wefo/40/7/WAF-D-24-0147.1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

P. J. v. Leeuwen. Particle Filtering in Geophysical Systems. *Monthly Weather Review*, 137(12):4089–4114, Dec. 2009. ISSN 1520-0493, 0027-0644. doi: 10.1175/2009MWR2835.1. URL https://journals.ametsoc.org/view/journals/mwre/137/12/2009mwr2835.1.xml.

J. Li and H. Liu. Improved hurricane track and intensity forecast using single field-of-view advanced ir sounding measurements. *Geophysical Research Letters*, 36, 2009.

Y. Li, Y. Tang, S. Wang, R. Toumi, X. Song, and Q. Wang. Recent increases in tropical cyclone rapid intensification events in global offshore regions. *Nature Communications*, 14(1):5167, Aug. 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40605-2. URL https://www.nature.com/articles/s41467-023-40605-2.

H. Lin, S. S. Weygandt, A. H. N. Lim, M. Hu, J. M. Brown, and S. G. Benjamin. Radiance preprocessing for assimilation in the hourly updating rapid refresh mesoscale model: A study using airs data. *Weather and Forecasting*, 32(5):1781 – 1800, 2017. doi: 10.1175/WAF-D-17-0028.1. URL https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0028_1.xml.

Y.-A. Liu, H.-L. A. Huang, A. H. N. Lim, and W. Gao. Adaptive bias correction of advanced infrared sounding radiance assimilation in a regional model and its impact on typhoon forecast. *Journal of Applied Remote Sensing*, 12, 2018.

Z. Liu, C. S. Schwartz, C. Snyder, and S.-Y. Ha. Impact of assimilating amsu-a radiances on forecasts of 2008 atlantic tropical cyclones initialized with a limited-area ensemble kalman filter. *Monthly Weather Review*, 140(12):4017 – 4034, 2012. doi: 10.1175/MWR-D-12-00083.1. URL https://journals.ametsoc.org/view/journals/mwre/140/12/mwr-d-12-00083.1.xml.

A. C. Lorenc. Analysis methods for numerical weather prediction. 112(474):1177–1194, 1986. ISSN 1477-870X. doi: 10.1002/qj.49711247414. URL https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.49711247414. Publisher: John Wiley & Sons, Ltd.

E. N. Lorenz. Designing chaotic models. *Journal of the Atmospheric Sciences*, 62(5):1574–1587, May 2005. doi: 10.1175/JAS3430.1. URL https://journals.ametsoc.org/view/journals/atsc/62/5/jas3430.1.xml.

X. Lu, X. Wang, M. Tong, and V. Tallapragada. Gsi-based, continuously cycled, dual-resolution hybrid ensemble–variational data assimilation system for hwrf: System description and experiments with edouard (2014). *Monthly Weather Review*, 145(12):4877 – 4898, 2017. doi: 10.1175/MWR-D-17-0068.1. URL https://journals.ametsoc.org/view/journals/mwre/145/12/mwr-d-17-0068.1.xml.

F. D. Marks, L. K. Shay, and Pdt-5. Landfalling tropical cyclones: Forecast problems and associated research opportunities. *Bulletin of the American Meteorological Society*, 79(2):305–323, 1998. ISSN 0003-0007. URL https://www.jstor.org/stable/26215049. Publisher: American Meteorological Society.

A. P. McNally, J. C. Derber, W. Wu, and B. B. Katz. The use of tovs level-1b radiances in the ncep ssi analysis system. *Quarterly Journal of the Royal Meteorological Society*, 126:689–724, 2000.

A. P. McNally, P. D. Watts, J. A. Smith, R. Engelen, G. A. Kelly, J. N. Thepaut, and M. Matricardi. The assimilation of airs radiance data at ecmwf. *Quarterly Journal of the Royal Meteorological Society*, 132:935–957, 2007.

A. Mehra, V. Tallapragada, Z. Zhang, B. Liu, L. Zhu, W. Wang, and H.-S. Kim. Advancing the state of the art in operational tropical cyclone forecasting at ncep. *Tropical Cyclone Research and Review*, 7:51–56, 2018.

M. Minamide and F. Zhang. Adaptive observation error inflation for assimilating all-sky satellite radiance. *Monthly Weather Review*, 145(3):1063 – 1081, 2017. doi: 10.1175/

MWR-D-16-0257.1. URL https://journals.ametsoc.org/view/journals/mwre/145/3/mwr-d-16-0257.1.xml.

E. J. Mlawer, S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough. Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, 102(D14):16663–16682, 1997. doi: https://doi.org/10.1029/97JD00237. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JD00237.

G.-Y. Niu, Z.-L. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, A. Kumar, K. Manning, D. Niyogi, E. Rosero, M. Tewari, and Y. Xia. The community noah land surface model with multiparameterization options (noah-mp): 1. model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, 116(D12), 2011. doi: https://doi.org/10.1029/2010JD015139. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JD015139.

NOAA. Monthly tropical cyclones report for annual 2020. Technical report, NOAA National Centers for Environmental Information, 2021.

Y. Ota, J. C. Derber, E. Kalnay, and T. Miyoshi. Ensemble-based observation impact estimates using the ncep gfs. *Tellus A: Dynamic Meteorology and Oceanography*, 65:20038, 2013.

H.-L. Pan and W.-S. Wu. Implementing a mass flux convection parameterization package for the NMC medium-range forecast model. 1995. URL https://repository.library.noaa.gov.

J. Peng, Z. Zhang, W. Wang, R. Panda, B. Liu, Y. Weng, A. Mehra, V. Tallapragada, X. Zhang, S. Gopalakrishnan, W. Komaromi, J. Anderson, and A. Poyer. HAFS ensemble forecast in AWS cloud. *Frontiers in Earth Science*, 12, 2024. ISSN 2296-6463. doi: 10.3389/feart.2024.1396612. URL https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2024.1396612/full. Publisher: Frontiers.

J. Poterjoy. Implications of multivariate non-gaussian data assimilation for multiscale weather prediction. *Monthly Weather Review*, 150(6):1475–1493, 2022a. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-21-0228.1. URL https://journals.ametsoc.org/view/journals/mwre/150/6/MWR-D-21-0228.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

J. Poterjoy. Regularization and tempering for a moment-matching localized particle filter. *Quarterly Journal of the Royal Meteorological Society*, 148(747):2631–2651, 2022b. ISSN 1477-870X. doi: 10.1002/qj.4328. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4328. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4328.

J. Poterjoy and F. Zhang. Inter-comparison and coupling of ensemble and variational data assimilation approaches for for the analysi s and forecasting of Hurricane Karl (2010). *Mon. Wea. Rev.*, 142:3347–3364, 2014.

J. Poterjoy and F. Zhang. Comparison of Hybrid Four-Dimensional Data Assimilation Methods with and without the Tangent Linear and Adjoint Models for Predicting the Life Cycle of Hurricane Karl (2010). *Mon. Wea. Rev.*, 144:1449–1468, 2016.

J. Poterjoy, R. A. Sobash, and J. L. Anderson. Convective-scale data assimilation for the weather research and forecasting model using the lo cal particle filter. *Mon. Wea. Rev.*, 145:1897–1918, 2017.

J. Poterjoy, L. Wicker, and M. Buehner. Progress toward the application of a localized particle filter for numerical weather prediction. *Monthly Weather Review*, 147(4):1107–1126, 2019. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-17-0344.1. URL https://journals.ametsoc.org/view/journals/mwre/147/4/mwr-d-17-0344.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

J. Poterjoy, G. J. Alaka Jr., and H. R. Winterbottom. The irreplaceable utility of sequential data assimilation for numerical weather prediction system development: Lessons learned from an experimental hwrf system. *Weather and Forecasting*, 36:661–677, 2021.

A. Poyer, W. Komaromi, J. Anderson, S. Gopalakrishnan, G. Alaka, J. Dale, A. Hazelton, W. Hogsett, F. Marks, A. Mehra, K. Newman, J. Rhome, V. Tallapragada, X. Zhang, and Z. Zhang. Hurricane forecast improvement program (HFIP) annual report 2023. 2025. URL https://repository.library.noaa.gov/view/noaa/67184.

V. S. Prasad, K. J. Ramesh, A. K. Bohra, and R. C. Bhatia. Assessment of quality and impact of full resolution tovs temperature profile data on the operational global data assimilation-forecast system of india. *Meteorology and Atmospheric Physics*, 68:197–212, 1998.

R. J. Purser, D. Jovic, J. Dong, L. Technologies, G. Ketefian, and J. Beck. The motivation, specification, and optimization of the extended schmidt gnomonic (ESG) grid for limited area applications. 2023.

W. Ramstrom, X. Zhang, K. Ahern, and S. Gopalakrishnan. Implementation of storm-following nest for the next-generation hurricane analysis and forecast system (HAFS). *Frontiers in Earth Science*, 12, 2024. ISSN 2296-6463. doi: 10.3389/feart.2024.1419233. URL https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2024.1419233/full. Publisher: Frontiers.

O. Reale, E. L. McGrath-Spangler, W. McCarty, D. Holdaway, and R. Gelaro. Impact of adaptively thinned airs cloud-cleared radiances on tropical cyclone representation in a global data assimilation and forecast system. *Weather and Forecasting*, 33(4):909 – 931, 2018. doi: 10.1175/WAF-D-17-0175.1. URL https://journals.ametsoc.org/view/journals/wefo/33/4/waf-d-17-0175_1.xml.

R. Rizzi and M. Matricardi. The use of tovs clear radiances for numerical weather prediction using an updated forward model. *Quarterly Journal of the Royal Meteorological Society*, 124(548):1293–1312, 1998. doi: https://doi-org.proxy-um.researchport.umd.edu/10.1002/qj.49712454813. URL https://rmets-onlinelibrary-wiley-com.proxy-um.researchport.umd.edu/doi/abs/10.1002/qj.49712454813.

R. Rogers, S. Aberson, A. Aksoy, B. Annane, M. Black, J. Cione, N. Dorst, J. Dunion, J. Gamache, S. Goldenberg, S. Gopalakrishnan, J. Kaplan, B. Klotz, S. Lorsolo, F. Marks, S. Murillo, M. Powell, P. Reasor, K. Sellwood, E. Uhlhorn, T. Vukicevic, J. Zhang, and X. Zhang. NOAA'S Hurricane Intensity Forecasting Experiment: A Progress Report. *Bulletin of the American Meteorological Society*, 94(6):859–882, June 2013. doi: 10.

1175/BAMS-D-12-00089.1. URL https://journals.ametsoc.org/view/journals/bams/94/6/bams-d-12-00089.1.xml.

C. S. Schwartz, Z. Liu, Y. Chen, and X.-Y. Huang. Impact of assimilating microwave radiances with a limited-area ensemble data assimilation system on forecasts of typhoon morakot. *Weather and Forecasting*, 27(2):424 – 437, 2012. doi: 10.1175/WAF-D-11-00033.1. URL https://journals.ametsoc.org/view/journals/wefo/27/2/waf-d-11-00033_1.xml.

C. S. Schwartz, J. Poterjoy, J. R. Carley, D. C. Dowell, G. S. Romine, and K. I. de. Comparing partial and continuously cycling ensemble kalman filter data assimilation systems for convection-all owing ensemble forecast initialization. *Weather and Forecasting*, 37(1):85 – 112, 2022. doi: 10.1175/WAF-D-21-0069.1. URL https://journals.ametsoc.org/view/journals/wefo/37/1/WAF-D-21-0069.1.xml.

J. A. Sippel, X. Wu, S. D. Ditchek, V. Tallapragada, and D. Kleist. Impacts of assimilating additional reconnaissance data on operational gfs tropical cyclone forecasts. *Weather and Forecasting*, 2022.

A. B. Smith and R. W. Katz. US billion-dollar weather and climate disasters: data sources, trends, accuracy and biases. *Natural Hazards*, 67(2):387–410, June 2013. ISSN 0921-030X, 1573-0840. doi: 10.1007/s11069-013-0566-5. URL http://link.springer.com/10.1007/s11069-013-0566-5.

C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson. Obstacles to High-Dimensional Particle Filtering. *Monthly Weather Review*, 136(12):4629–4640, Dec. 2008. ISSN 1520-0493, 0027-0644. doi: 10.1175/2008MWR2529.1. URL https://journals.ametsoc.org/view/journals/mwre/136/12/2008mwr2529.1.xml.

O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory. 113(478):1311–1328, 1987. ISSN 1477-870X. doi: 10.1002/qj.49711347812. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.49711347812. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49711347812.

G. Thompson and T. Eidhammer. A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *Journal of the Atmospheric Sciences*, 71(10):3636–3658, 2014. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-13-0305.1. URL https://journals.ametsoc.org/view/journals/atsc/71/10/jas-d-13-0305.1.xml. Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

G. Thompson, P. R. Field, R. M. Rasmussen, and W. D. Hall. Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. part II: Implementation of a new snow parameterization. *Monthly Weather Review*, 136(12):5095–5115, 2008. ISSN 1520-0493, 0027-0644. doi: 10.1175/2008MWR2387.1. URL https://journals.ametsoc.org/view/journals/mwre/136/12/2008mwr2387.1.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

R. D. Torn and C. Snyder. Uncertainty of tropical cyclone best-track information. *Weather and Forecasting*, 27(3):715–729. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-11-00085.1. URL https://journals.ametsoc.org/view/journals/wefo/

27/3/waf-d-11-00085_1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

P. J. van Leeuwen, H. R. Künsch, L. Nerger, R. Potthast, and S. Reich. Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, 145(723):2335–2365, 2019. ISSN 1477-870X. doi: 10.1002/qj.3551. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3551. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3551.

W. Wang, J. Han, F. Yang, J. Steffen, B. Liu, Z. Zhang, A. Mehra, and V. Tallapragada. Improving the intensity forecast of tropical cyclones in the hurricane analysis and forecast system. *Weather and Forecasting*, 38(10):2057–2075, 2023. ISSN 1520-0434, 0882-8156. doi: 10.1175/WAF-D-23-0041.1. URL https://journals.ametsoc.org/view/journals/wefo/38/10/WAF-D-23-0041.1.xml. Publisher: American Meteorological Society Section: Weather and Forecasting.

W. Wang, J. Han, J. Shin, X. Chen, A. Hazelton, L. Zhu, H.-S. Kim, X. Li, B. Liu, Q. Liu, J. Steffen, R. Sun, W. Zheng, Z. Zhang, and F. Yang. Physics schemes in the first version of NCEP operational hurricane analysis and forecast system (HAFS). *Frontiers in Earth Science*, 12, 2024. ISSN 2296-6463. doi: 10.3389/feart.2024.1379069. URL https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2024.1379069/full. Publisher: Frontiers.

F. Weng, Y. Han, P. van Delst, Q. Liu, T. Kleespies, B. Yan, and J. L. Marshall. JCSDA Community Radiative Transfer Model (CRTM). Technical report, Joint Center for Satellite Data Assimilation, 2006.

J. S. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, Jul 2002. doi: 10.1175/1520-0493(2002)130⟨1913:EDAWPO⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/130/7/1520-0493_2002_130_1913_edawpo_2.0.co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

D. Worthen, J. Wang, R. Montuoro, D. Heinzeller, B. Li, G. Theurich, U. Turuncoglu, D. Rosen, D. Jovic, B. Curtis, R. Mahajan, H. Lei, A. Richert, A. Chawla, J. Wang, J. Meixner, A. Abdolali, M. Masarik, L. Pan, M. Barlage, B. Liu, M. Vertenstein, T. Craig, R. Benson, T. Robinson, T. Clune, W. Jiang, N. Barton, G. Vandenberghe, M. Potts, J. Kim, N. Perlin, C. Book, L. Bernardet, F. Yang, S. Sun, H.-S. Kim, B. Baker, J. Huang, C.-H. Jeon, and I. Stajner. Coupling infrastructure capability in UFS weather model. 2025. URL https://repository.library.noaa.gov/view/noaa/61453.

D. Xu, Z. Liu, X.-Y. Huang, J. Min, and H. Wang. Impact of assimilating iasi radiance observations on forecasts of two tropical cyclones. *Meteorology and Atmospheric Physics*, 122:1 – 18, 2013.

J.-I. Yano, M. Z. Ziemiański, M. Cullen, P. Termonia, J. Onvlee, L. Bengtsson, A. Carrassi, R. Davy, A. Deluca, S. L. Gray, V. Homar, M. Köhler, S. Krichak, S. Michaelides, V. T. J. Phillips, P. M. M. Soares, and A. A. Wyszogrodzki. Scientific Challenges of Convective-Scale Numerical Weather

Prediction. *Bulletin of the American Meteorological Society*, 99(4):699–710, Apr. 2018. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-17-0125.1. URL https://journals.ametsoc.org/view/journals/bams/99/4/bams-d-17-0125.1.xml.

T. H. Zapotocny, J. A. Jung, J. F. L. Marshall, and R. E. Treadon. A two-season impact study of four satellite data types and rawinsonde data in the ncep global data assimilation system. *Weather and Forecasting*, 23(1):80 – 100, 2008. doi: 10.1175/2007WAF2007010.1. URL https://journals.ametsoc.org/view/journals/wefo/23/1/2007waf2007010_1.xml.

Z. Zeng, Y. Wang, and C.-C. Wu. Environmental dynamical control of tropical cyclone intensity—an observational study. *Monthly Weather Review*, 135(1):38 – 59, 2007. doi: 10.1175/MWR3278.1. URL https://journals.ametsoc.org/view/journals/mwre/135/1/mwr3278.1.xml.

F. Zhang, Y. Weng, J. A. Sippel, Z. Meng, and C. H. Bishop. Cloud-resolving hurricane initialization and prediction through assimilation of doppler radar observations with an ensemble kalman filter. *Monthly Weather Review*, 137:2105–2125, 2009.

J. A. Zhang, D. S. Nolan, R. F. Rogers, and V. Tallapragada. Evaluating the impact of improvements in the boundary layer parameterization on hurricane intensity and structure forecasts in HWRF. *Monthly Weather Review*, 143(8):3136–3155, 2015. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-14-00339.1. URL https://journals.ametsoc.org/view/journals/mwre/143/8/mwr-d-14-00339.1.xml.

M. Zhang, M. Zupanski, M.-J. Kim, and J. A. Knaff. Assimilating amsu-a radiances in the tc core area with noaa operational hwrf (2011) and a hybrid data assimilation system: Danielle (2010). *Monthly Weather Review*, 141(11):3889 – 3907, 2013. doi: 10.1175/MWR-D-12-00340.1. URL https://journals.ametsoc.org/view/journals/mwre/141/11/mwr-d-12-00340.1.xml.

W. Zheng, M. Ek, K. Mitchell, H. Wei, and J. Meng. Improving the stable surface layer in the ncep global forecast system. *Monthly Weather Review*, 145(10):3969 – 3987, 2017. doi: 10.1175/MWR-D-16-0438.1. URL https://journals.ametsoc.org/view/journals/mwre/145/10/mwr-d-16-0438.1.xml.

Y. Zhu, J. Derber, A. Collard, D. Dee, R. Treadon, G. Gayno, and J. A. Jung. Enhanced radiance bias correction in the National Centers for Environmental Prediction's Gridpoint Statistical Interpolation data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 140:1479–1492, 2014.

Y. Zhu, E. Liu, R. Mahajan, C. Thomas, D. Groff, P. Van Delst, A. Collard, D. Kleist, R. Treadon, and J. C. Derber. All-sky microwave radiance assimilation in ncep's gsi analysis system. *Monthly Weather Review*, 144:4709–4735, 2016.

X. Zou, F. Weng, B. Zhang, L. Lin, Z. Qin, and V. Tallapragada. Impacts of assimilation of atms data in hwrf on track and intensity forecasts of 2012 four landfall hurricanes. *Journal of Geophysical Research: Atmospheres*, 118:11,558–11,576, 2013.

D. Zupanski. A General Weak Constraint Applicable to Operational 4DVAR Data Assimilation Systems. *Monthly Weather Review*, 125(9):2274–2292, Sept. 1997. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1997)125⟨2274:AGWCAT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/125/9/1520-0493_1997_125_2274_agwcat_2.0.co_2.xml.