Title
Improved Earth System Prediction Using Large Ensembles and Machine Learning

Author
Chapman, William

Publication Date
2021

UNIVERSITY OF CALIFORNIA SAN DIEGO

Improved Earth System Prediction Using Large Ensembles and Machine Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy

in

Oceanography

by

William E. Chapman

Committee in charge:

Fred Martin Ralph, Co-Chair
Shang-Ping Xie, Co-Chair
Ian Eisenman
Amato T. Evan
Jan Kleissl
Aneesh C. Subramanian

2021

The Dissertation of William E. Chapman is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To Endre & Liam

TABLE OF CONTENTS

Dissertation Approval Page iii
Dedication iv
Table of Contents v
List of Figures viii
List of Tables xii
Acknowledgments xiii
Curriculum Vitae xvii
Abstract of Dissertation xix
Chapter 1 - Introduction 1
1.1 Motivation and History 1
1.2 Dissertation Overview 6
1.2.1 Machine Learning & Statistical Methods for Numerical Weather Prediction Post-Processing 6
1.2.2 Subseasonal Predictability and Forecast Windows of Opportunity 8
1.3 Quick Dissertation Guide 11
Chapter 2 - Improving Atmospheric River Forecasts with Machine Learning 13
2.1 Overview 13
2.2 Data and Methodology 16
2.2.1 Forecasts 16
2.2.2 Ground-truth 17
2.2.3 Experimental Design 17
2.2.4 Convolutional Neural Networks 17
2.2.5 ARcnn Example 19
2.3 Verification Metrics 20
2.3.1 GFS error 21
2.4 Results 21
2.5 Summary and Conclusions 24
2.6 Acknowledgements 25
2.7 Support Information 31
2.7.1 ARcnn Architecture and Methodology 31
2.7.2 Land Falling AR Statistics 33
Chapter 3 - Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning 43
3.1 Overview 43
3.2 Data, Methods, and Metrics 48

3.2.1 Region of Interest 48
3.2.2 Ground Truth 48
3.2.3 Dynamic Model For Deterministic Forecast 49
3.2.4 Machine Learning Generated Forecast Uncertainty 50

3.3 Results 58
3.3.1 Deterministic Predictions 59
3.3.2 Diebold-Mariano Test 60
3.3.3 Brier skill score and CRPS 62
3.3.4 Spread/Skill 63
3.3.5 Stratified Rank Histogram 64
3.3.6 Variable Importance 66
3.3.7 Length of Training 67

3.4 Discussion and Conclusion 68

3.5 Evaluation Metrics 71
3.5.1 Evaluation Metrics 72

3.6 Acknowledgement 76

3.7 Supporting Information 89
3.7.1 Hyperparameter Search 89
3.7.1.1 Neural Network 89
3.7.1.2 Convolutional Neural Network 90
3.7.1.3 Brier Score Decomposition 94
Reliability Diagram Explanation 94

Chapter 4 - Monthly Modulations of ENSO Teleconnections: Implications for Potential Predictability in North America 108

4.1 Overview 109

4.2 Data and Methods 113
4.2.1 AGCM experiments 113
4.2.2 Definition of ENSO and Compositing 114
4.2.3 Variable Selection 115
4.2.4 Observations 116
4.2.5 Rossby Wave Source & Wave Activity Flux 117
4.2.6 Variance Patterns 118
4.2.7 Signal-to-Noise & Potential Predictability 118
4.2.8 Kullback-Leibler divergence 119

4.3 Atmospheric Response to ENSO 120
4.3.1 Rossby Wave Source 122
4.3.2 Wave Activity Flux 124
4.3.3 Additional Sources of ENSO Forced Extratropical Waves 126

4.4 Signal vs. Noise 127
4.4.1 ENSO modulation of Internal Variance 128
4.4.2 Signal-to-Noise Ratio 130

4.5 ENSO Potential Predictability & Observations 132

4.6 Summary and Discussion 136

4.7 Acknowledgement 138

4.8 Supporting Information 154

Chapter 5 - Subseasonal PNA Forecast Skill and Tropical PNA Drivers in the ECMWF 20th Century Hindcast 157

5.1 Introduction 158

5.2 Data 161

5.2.1 Observations 161

5.2.2 ECMWF 20th Century Hindcast Model 161

5.2.3 ENSO Index 162

5.2.4 Model Drift 162

5.2.5 PNA Index 163

5.2.6 Rossby Wave Source 164

5.2.7 MJO Calculation 165

5.2.8 MJO Calculation 166

5.3 PNA Forecast Skill 168

5.4 Variability and Drivers of S2S PNA Predictability 172

5.4.1 4-week Averaged Response 173

5.4.1.1 Tropical SSTs 173

5.4.2 Weekly Averaged Response 177

5.4.2.1 Tropical SSTs 177

5.4.2.1.1 Mean 179

5.4.2.1.2 Spread 180

5.4.2.2 Madden Julian Oscillation Teleconnection 181

5.4.2.2.1 Mean 184

5.4.2.2.2 Spread 185

5.5 Discussion 188

5.6 Acknowledgement 190

5.7 Supporting Information 207

Chapter 6 - Conclusion 212

6.1 Summary of Contributions 212

6.2 Future Directions 216

Chapter 7 - References 220

LIST OF FIGURES

Figure 2.1. Forecasts and analysis valid for IVT […] 27

Figure 2.2. Spatial distribution of 96-hour forecast […] 28

Figure 2.3. ROI average temporal evolution […] 29

Figure 2.4. Spatial distribution of percent improvement of 96-hour IVT forecast after ARcnn postprocessing for […] 30

Figure 2.1S. ARcnn model architecture 36

Figure 2.2S. Coastal (as defined by the shaded swath in panel (c)) […] 37

Figure 2.3S. Coastal (as defined by the shaded swath in panel (c)) (a) RMSE, CRMSE, Bias, and […]. 38

Figure 2.4S. Spatial distribution of percent improvement of 96-hour IVT forecast […] 39

Figure 2.5S. Background Error Fields of GFS vs AARcnn […] 40

Figure 3.1. Coastal evaluation locations and climatological (December-March 1984-2019) Integrated Vapor Transport (colorfill) 81

Figure 3.2. Root mean squared error […] 82

Figure 3.3. Two-sided Diebold-Mariano test using twCRPS (threshold = [250 kg m^{-1} s^{-1}]) […] 83

Figure 3.4. Brier skill score at thresholds of 250 (a), 350 (b), and 500+ (c) [kg m^{-1} s^{-1}] forecasted units of IVT […] 84

Figure 3.5. Binned spread-skill plots for forecasts from 0-48 hours (a-f) and 56-120 hours (g-l) […] 85

Figure 3.6. Stratified Rank Histograms. […] 86

Figure 3.7. Relative permutation importance for the input predictors (defined in Table 3.1) in the CNN (black) […] 87

Figure 3.8. Forecasted 48-hour threshold weighted (250 kg m^{-1} s^{-1}) continuous ranked probability skill score […] 88

Figure. 3.1S. network architecture for the leveraged Neural Network (a; NN) and convolutional neural network (b; CNN) […] 95

Figure. 3.2S. As in Figure 3.4, but for threshold weighted continuous ranked probability score at thresholds of […] 96

Figure. 3.3S. The mean values of the weights across every landfall locations received by each predictor as a […] 97

Figure. 3.4S.. Diebold-Mariano test for […] 98

Figure. 3.5S. Diebold-Mariano test for […] 99

Figure. 3.6S. As in Figure 3.4. but for Brier decomposed Resolution at thresholds of 250 (a), 350 (b), and 500+ (c) [kg m^{-1} s^{-1}] forecasted units of IVT[…] 100

Figure 3.7S. As in Figure 3.4. but for Brier decomposed Reliability at thresholds of […] ... 101

Figure 3.8S. reliability diagrams for the CNN (a), AnEn (b), NN (c), and GEFS$_{nn}$ (d) forecast systems for an exceedance […] 102

Figure 3.9S. Reliability diagrams for […] 103

Figure 3.10S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.8S but for an exceedance threshold of 500 [kg m^{-1} s^{-1}] units of IVT......... 104

Figure 3.11S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.8S but for an exceedance […] 105

Figure 3.12S. Reliability diagrams for […] 106

Figure 3.13S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.11S but for an exceedance threshold of 500 [kg m^{-1} s^{-1}] units of IVT....... 107

Figure 4.1 February and March ensemble mean monthly response to (column I) El Niño and (column II) La Niña: […] 140

Figure 4.2 Ensemble mean monthly (December–April) response to (column I) El Niño and (column II) La Niña: […] 141

Figure 4.3 (a) Vortex stretching (VS). (b) Absolute vorticity advection by divergent flow (AVA). (c) Rossby wave source (RWS) anomaly for El Niño (red line), neutral (black line) […] 142

Figure 4.4 Ensemble mean monthly (DJFMA) response to (column I) El Niño and (column II) La Niña: […] 143

Figure 4.5 (column I) Internal variability and (column II) forced leading EOF mode of 200-hPa atmospheric geopotential height variability, calculated for each month […] ... 144

Figure 4.6 (a) Yearly development of composite RMS signal (dotted line) and noise (solid line) area averaged in (b) […] 145

Figure 4.7 February and March RMS noise response to (column I) El Niño and (column II) […]. 146

Figure 4.8 Area averaged (25°–70°N, 155°E–60°W) signal-to-noise ratio for (a) 200-hPa geopotential height (b) 2-m temperature (values over land only) […] 147

Figure 4.9. Monthly (DJFMA) signal-to-noise relationship for temperature (colorfill) […] 148

Figure 4.10. As in Fig. 4.9, but for North American precipitation (colorfill) and 200-hPa geopotential height (contour) 149

Figure 4.11. Monthly d4PDF (a),(b) PNA index and (c),(d) AII index mean and 5th and 95th percentiles across 100 ensemble members […] 150

Figure 4.12. (column I) El Niño and (column II) La Niña monthly composite of observed precipitation (colorfill) and 200-hPa geopotential height (contour; negative dashed). […] 151

Figure 4.13. (a)–(j) Observed monthly (DJFMA) T2m Kullback–Leibler divergence (KL) […]. 152

Figure 4.1S. Ensemble mean monthly (Dec-Apr) response to El Niño (column I) and La Niña (column II) […] 154

Figure 4.2S. RMS noise (Dec-Apr) response to El Niño (column I) and La Niña (column II): composite map of anomalous RMS noise 200-hPa […] 155

Figure 4.3S. El Niño (column I) and La Niña (column II) monthly composite of observed temperature (colorfill) and 200-hPa geopotential height (contour, negative dashed) […]. 156

Figure 5.1 Leading modes of observed variability [10N,80N], [140E, 60W] in the area weighted 200-hPa CERA20-C […] 192

Figure 5.2 PNA RMSE (CERA-20C vs. CSF-20C) and ensemble spread as a function of forecast lead at daily time-scales a) November initialization and b) February initializations in the CSF-20C. 193

Figure 5.3 PNA correlation as a function of lead time for different time scales in the November (a) and February (b) forecast initializations and correlation of the initial averaged […] 194

Figure 5.4 The ROC skill score for PNA events at different thresholds, for the 4-weeks averaged forecast, centered on forecast day 28. […] 195

Figure 5.5 Scatter diagrams of PNA Magnitude (a,b), Tropical 200-hPa Divergence (c,d), Midlatitude 200-hPa Divergence (e,f), and 200-hPa Rossby Wave Source anomaly (g,h). […] 196

Figure 5.6 Percent of modulation to atmospheric variability as a function of 1 standard deviation change in […] 197

Figure 5.7 7-14 day averaged lagged response of Z200a (contour, black 10m interval, positive: solid; negative: dashed) […] 198

Figure 5.8 Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble mean […] 199

Figure 5.9 Anomalous sea surface temperature and zonal and meridional wind 850-hPa regressed on the extended EOF of SST […] 200

Figure 5.10 7-14 day averaged lagged response of Z200a (contour, black 10m interval, positive: solid; negative: dashed) […] 201

Figure 5.11 0-7 day averaged lagged four components of the Rossby Wave Source anomaly (a,c) and the divergence (b,d) in the eight MJO phases in region: [20N,40N];[140E,170W] […] 202

Figure 5.12 Left (contour; black at 10m intervals) and right […]..203

Figure 5.13 Anomalous velocity potential and zonal and meridional wind 200-hPa regressed on the extended EOF of SST (week 0-5, region: [-15S,15N],[0W, 360W]) […]....204

Figure 5.14 Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble spread weekly averaged Velocity Potential anomalies […]..205

Figure 5.15 Correlation as a function of forecast lead time of the anomalous spread […]....206

Figure 5.1S The ROC skill score for PNA events at different thresholds, for the 4-weeks averaged forecast, centered on forecast day 28. […]..207

Figure 5.2S Composite Rossby wave source (colorfill), irrotational components of the 200-hPa wind (vector) […]..208

Figure 5.3S Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble spread weekly averaged Velocity Potential anomalies […]..209

Figure 5.4S Composite of anomalous VP at 200-hPa and 200-hPa irrotational components of the wind anomalies (vector) for each phase of the MJO in the […].......................210

Figure 5.5S Bivariate correlation of the CSF-20C MJO and CERA-20C MJO observations by forecast […]...211

LIST OF TABLES

Table 2.1S. Training, Validation and Testing Datasets 41

Table 2.2S. ARcnn: Hyperparameters & Architecture. 42

Table 3.1. Abbreviations and descriptions of all input variables. 78

Table 3.2. CNN/NN Parameters by category […] 79

Table 3.3. Stratified rank histogram reliability index by method and for forecasts aggregated from lead times 0-48 hours and 54 – 120 hours. 80

Table 4.1. The d4PDF defined ENSO states by year. 153

ACKNOWLEDGEMENTS

This is the last section that I will pen before handing this dissertation to my committee, so they can judged the worth of my work over the last 5 years. This section is written by a man who is exceedingly tired and all together sick of this task. This exhaustion is transient -- I am sure, and I would like to pay earnest tribute to those who helped me to get to where I am. However, this section will never be perfect, as I am not eloquent enough to express how much you all have influenced me.

First and foremost, I would like to thank my academic advisors Dr. Marty Ralph and Dr. Shang-Ping Xie. You simultaneously gave me the freedom to explore my own questions, and the guidance to keep my eyes pointed in the correct direction. You challenged me to look at my science questions (and the field as a whole) critically, and I have become a better scientist for it. You have been giant advocates for my career, and the next steps of my journey are largely thanks to you. Thank you for giving me access to the vibrant Scripps community, and your research groups -- it has been an honor and a thrill.

Thanks to my professors, mentors, co-authors and confidants. I walked away from every conversation a combination of inspired, confused, and probably hungry (the last two were my fault though). Thank you all for your thoughtful advice. Particularly, I thank my committee members: Ian Eisenman, Amato Evan, Jan Kleissl, and Aneesh Subramanian for their time, effort, and conversation throughout this process. It is an honor to have access to scientists of your caliber, thank you for what you have invested in me.

Thank you to my extended family. You are too numerous, and your personality traits are too large to list. Also, I genuinely believe that the order in which your names appeared would end up as a point of contention (or competition) at a future Thanksgiving. You will not and should not read this dissertation; it is a boring.

To my friends/mentors in the UCSD community: Gabe, Blake, Chase, Hugh, Reuben, Ludo, Anna, Julie, Meredith, Kara, Carly, Chad, Tom, Nick, Max, and many more, thank you for the conversation over coffee/food or in the water or in the sun. To the Yak Pen: Luke, Margaret, Not-Curly Mike, Jessica, Jacob, Julia, Theresa, and Ruths, thank you all for being a light and a shoulder. I hope this thesis is below and within enough.

Thank you, Luca, for indulging in my science, teaching me, and challenging me. The ideas, seeded by our conversations, are scattered all over this dissertation. Thank you to Sue Haupt and Stefano Alessandrini, for indulging me in conversation despite your busy calendars, and for opening the doors to the NCAR community. I thank my mentor and friend, Aneesh Subramanian, you sit at the center of the Venn diagram of kindness and intelligence, and I am so fortunate to get to pick your brain. Thank you for always approaching our conversations with curiosity before skepticism. To my friend Tony Jakubisin, I would be an entirely different and worse person had I not met you. You give your hard-earned knowledge so selflessly, and I have learned so much from you. I thank my friend Mike Sierks, we have been on a journey together for 12 years now and it looks like finishing this dissertation means finally splitting up the band, and that is so bittersweet. I am deeply indebted to you for your friendship, conversation, perspective, humor, and kindness. Jack and Allie are lucky to have to you. Fiona, thank you for all of your unwavering support and love. I am going to continue to borrow your moral compass, if that is ok. I am so excited for the next leg of our trek, I think it is going to be both silly and serious; tedious and astonishing, but undoubtably fun. Grady, Erik, and Eric: ditch boys will live forever, unless they fall in the ditch.

I thank the lifelong friends that have stuck with me from when I was a smelly high school photography rat until now. Cristen, Adam, Luke, Megan, Jenna, Megan, Shelby, Grant, Katey, Jon, Kaitie, Michael, Bonnie, Billy, Molly, Thomas (and your gaggle of children) – every time I see you all, I am reminded about how fundamental our beginnings are to how things eventuate. Thank you for being the ones that set my foundational definition of friendship. Thanks Kevin for helping me stay sane and grounded during these last five years, I could not have got this done without your support.

I have dedicated this work to Endre and Liam, my nephews. You were both born while I struggled through this research. It has been a privilege to watch you grow. You two are curious about the world in the best ways, and I think that is one of the strongest traits to have. I owe a special thanks to my sister, who has been my editor and advocate since day one. Thanks Laura. I thank my brother Gisle for letting me watch him throw a frisbee - you are so unathletic. I thank my father for propagating the scientific tradition and the love for interrogating data on to me. I thank my mom for balancing me out, and showing me the importance of a cup of coffee, friends, a book, and a nap. Thank you Lincoln for checking me when I get too lofty, you are wise beyond your years.

Very few of the folks I mention in this acknowledgment will ever read these words, and that is great. It has been a wonderful way to reflect on all the good that has profoundly affected the last ten odd years of my life, and my life's trajectory for the next 100 (I plan on being the first human to live to 130).

Chapter 2, in full, is a reprint of the material as it appears in Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving

Atmospheric River Forecasts with Machine Learning. Geophysical Research Letters. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full is a reprint of the material as it appears in Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.P., Lerch, S., Hayatbini, N. (2021) Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning. Monthly Weather Review. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Chapman, W. E., Subramanian, A. C., Xie, S. P., Sierks, M. D., Ralph, F. M., & Kamae, Y. (2021). Monthly Modulations of ENSO Teleconnections: Implications for Potential Predictability in North America. Journal of Climate. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in part is currently being prepared for submission for publication of the material. Chapman W. E., Subramanian, A. C., Xie S. P., Ralph, F. M.. The dissertation author was the primary investigator and author of this paper.

VITA

2012	Bachelor of Science in Engineering, University of California San Diego
2014-2016	Master of Science in Civil & Environmental Engineering, Stanford University
2021	Doctor of Philosophy in Oceanography, University of California San Diego

PUBLICATIONS

W. Chapman, L. Delle Monache, S. Alessandrini, A. Subramanian, F. Ralph, S. Xie, S. Lerch, and N. Hayatbini, “Probabilistic predictions from deterministic atmospheric river forecasts with deep learning”, Monthly Weather Review, *In Press*, 20 Sept. 2021.

W. Chapman, A. C. Subramanian, S.-P. Xie, M. D. Sierks, F. M. Ralph, and Y. Kamae, “Monthly modulations of enso teleconnections: Implications for potential predictability in north america”, Journal of Climate, pp. 1–71, 3 Mar. 2021.

P. B. Gibson, **W. Chapman**, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser, “Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts”, Nature - Communications Earth & Environment, vol. 2, no. 1, p. 159, Aug. 2021, ISSN: 2662-4435.

S. E. Haupt, **W. Chapman**, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, “Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the oxford 2019 workshop”, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 379, no. 2194, p. 20 200 091, 2021

S. Meech, S. Alessandrini, **W. Chapman**, and L. Delle Monache, “Post-processing rainfall in a high-resolution simulation of the 1994 piedmont flood”, Bulletin of Atmospheric Science and Technology, Jan. 2021, ISSN: 2662-1509.

G. Schamberg, **W. Chapman**, S.-P. Xie, and T. P. Coleman, “Direct and indirect effects—an information theoretic perspective”, Entropy, vol. 22, no. 8, p. 854, 2020.

A. M. Wilson, **W. Chapman**, A. Payne, A. M. Ramos, C. Boehm, D. Campos, J. Cordeira, R. Garreaud, I. V. Gorodetskaya, J. J. Rutz, et al., “Training the next generation of researchers in the science and application of atmospheric rivers”, Bulletin of the American Meteorological Society, vol. 101, no. 6, E738–E743, 2020.

W. Chapman, A. Subramanian, L. Delle Monache, S. Xie, and F. Ralph, “Improving atmospheric river forecasts with machine learning”, Geophysical Research Letters, vol. 46, no. 17-18, pp. 10 627–10 635, 2019

ABSTRACT OF THE DISSERTATION

Improved Earth System Prediction Using Large Ensembles and Machine Learning

by

William E. Chapman

Doctor of Philosophy in Oceanography

University of California San Diego, 2021

F. Martin Ralph, Co-Chair
Shang-Ping Xie, Co-Chair

The purpose of this thesis is to examine and advance North American weather predictability from weather to subseasonal time-scales. Specifically, it focuses on 1) developing machine learning/deep learning methods and models to improve predictability through numerical weather prediction (NWP) post-processing on weather time-scales (0-7 days) and 2)

examining the physical mechanisms which govern the evolution of the predictable components and noise components of teleconnection modes on subseasonal time-scales (7 days - 1 month).

NWP deficiencies (e.g., sub-grid parameterization approximations), nonlinear error growth associated with the chaotic nature of the atmosphere, and initial condition uncertainty lead initial small forecast errors to eventually result in weather predictions which are as skillful as random forecasts. A portion of these forecast errors are inherent to the NWP models alone, systematic biases. The first two chapters develop cutting-edge vision-based deep-learning algorithms to advance the current state-of-the-art NWP post-processing and correct these systematic biases. Using dynamic forecasts of North Pacific integrated vapor transport (IVT) as a test case, we develop post-processing systems which are spatially aware, readily encode non-linear predictor interaction, easily ingest ancillary weather variables, and have state of the art training methods that systematically prevent model overfitting. Further, we outline a framework to quantify uncertainty in single-point (deterministic) forecasts using neural networks. The uncertainty is shown to be probabilistically rigorous, leading to calibrated probabilistic forecasts which outperform or compete with calibrated dynamic NWP ensemble systems for IVT under atmospheric river conditions.

The second half of this thesis shifts focus to subseasonal time scales and examines predictability in the Pacific North American (PNA) sector in boreal winter. Particularly, it investigates the physical mechanisms involved in the intraseasonal modulation of atmospheric Signal-to-Noise (SN), and how it is affected by slowly varying climate modes (ENSO and MJO). These mechanisms are further explored using a fully-coupled hindcast of the 20th century, showing that the increased SN leads to high model forecast skill at subseasonal timescales in

particular forecast windows of opportunity. Additionally, we reveal the MJO as the largest growing mode of tropical forecast uncertainty which directly influences PNA forecast certainty.

Chapter 1

Introduction

1.1 Motivation and History

There is a reason people curse the weatherman and the Old Farmer's Almanac. When lives and livelihoods depend upon accurate prediction of atmospheric conditions, a false weather report rankles, or worse. Yet, out of necessity and hope, humans have persevered in the quest to accurately unravel the atmosphere's mysteries through short- and long-term weather prediction. The history of weather prediction is, in truth, a dramatic story of human observation, dire necessity, emerging technologies, and exponentially increasing complexity, beginning with perhaps a single data point ("*red sky at night, sailor's delight*"), later thousands of data points (data sets of observed weather conditions shared through telegraph technology), eventuating in modern day numerical weather prediction and climate modeling. Each developmental leap in weather prediction, from the earliest folklore to modern data assimilation, is a stepping stone on the complex and ever-widening path to improved earth system modeling. This dissertation represents a small contribution on this path.

Early efforts to observe and predict local and regional weather relied on human observation of natural phenomena, including the appearance of the sky, moon, and behavior of animals. Long before the Almanac and other predicting efforts became widespread, folklore often informed farmers', sailors', and travelers' decisions. For example, cirrostratus clouds causing the appearance of a halo around the moon was said to predict a coming storm: "*If the moon shows like a silver shield, Be not afraid to reap your field; But if she rises haloed round, Soon we'll tread on deluged ground*." Other examples include: "*Birds flying low, expect rain* *and a blow*" and "*Catchy drawer and sticky door, Coming rain will pour and pour.*" or "*When windows won't open, and the salt clogs the shaker, the weather will favour the umbrella maker!*"

In 1885, Cleveland Abbe, the architect of the U.S. Weather Bureau wrote in the journal *Science*, "Meteorology has received enthusiastic support by our own and all other nations. We are now doing about all that can be done by the mere utilization of the telegraph and weather map and the cautious application of general average rules, but we are still powerless in the presence of any unusual movement of the atmosphere" (Abbe, 1895). Abbe was commenting on the state of weather forecasting which, at the time, was merely an interconnected series of telegraph machines reporting the local weather and a series of historical "look-up tables" (synoptic charts) developed from those observations. Forecasters would find past weather states which were analogous to the current observations and use them to "predict" the future state of the atmosphere based on how those analogous events developed. Abbe was disappointed in this simple system of observing, recording, and averaging. His statement in *Science* was an informal call for the development of the theoretical equations of atmospheric motion. This call was answered by Villhelm Bjerknes with his development of circulation theory (Bjerknes et al., 1898) and the papers that followed. These papers set out the seven equations of motion which still form the basis of our understanding of atmospheric dynamics (Bjerknes, 1900, 1904). As an aside, Lorenz (1969a) estimated that nearly 140 years' worth of upper-level observational data would be required to understand the growth of small errors through the use of analog matching. Van den Dool (1994), again examining analog matching, found that nearly 10^{30} years of observations would be needed in order to find analogs that could match within current observation error of 500-hPa heights over the whole of the Northern Hemisphere. Without the development of the physics of atmospheric motion, weather prediction would have surely ground to a halt.

Bjerknes' equations were meant to eliminate the need for continued simple observation, as they developed a framework for fundamental accurate weather prediction based on the physics of the atmosphere. Despite having a fundamental understanding of atmospheric motion, in testing these theories, it became apparent that to support these theories, more atmospheric observations were required rather than fewer. Bjerknes set out collecting observations and developing analytical methods to improve prediction. Concurrently, Richard Lewis Fry was developing methods to numerically integrate Bjerknes's tendency equations and thus forecast the weather. In May of 1910, Bjerknes led a detailed upper atmosphere intensive observational campaign utilizing over 150 weather balloons (both manned and unmanned) and 35 kites (Lynch, 2006). The observations were organized into a series of synoptic charts, with 10 vertical levels at 17 locations across Europe [Aside: The observational campaign intentionally coincided with the passing of Halley's comet, which at the time, was theorized to affect the weather] (Bjerknes, 1910). Richardson used these observations to publish "Weather Prediction by Arithmetic Finite Differences" which eventually became "Weather Prediction by Numerical Processes" (L. F. Richardson, 2007). Computers did not exist at that time; thus, the complex calculations of forecast integration were done via pen and paper. It took Richardson ~2 years to complete one six-hour forecast. The predicted forecast failed to resemble the observed state of the atmosphere, but this did not deter the meteorologists of the day.

The advent of the computers increased the speed at which the calculations were performed. Aided by the fundamental work of Jules Charney, John von Neumann, Norman Phillips, Fred Shuman, Joseph Smagorinsky, Syukoro Manabe (and many others), the age of computer driven Numerical weather prediction (NWP) and climate modeling was born. Subgrid physics and model integration schemes grew in complexity and forecast accuracy increased. The modern age of atmospheric science and weather predictability is now dependent on the computer.

Throughout NWP development, the importance of accurate observations cannot be overstated. It became increasingly obvious that the success of weather predictions hinged upon accurate and dense observation networks. The two fields (observational based analysis and numerical prediction) grew in parallel, feeding off each other. During this growth, increasingly complex and successful methods to combine the fields more effectively were developed.

In the recent era, observations have been used to further improve predictability through statistical correction of NWP model forecast error. Model deficiencies are now discovered and corrected through careful comparison of forecast to observation. These techniques were first pioneered by Harry Glahn, and his model output statistics schemes (MOS, Glahn & Lowry, 1972), but have dramatically grown in complexity and efficacy in recent times. At its core, model post-processing aims to improve the quality and utility of forecasts. Born out of the desire to prove forecast improvements, entire fields of study have been dedicated to deterministic and probabilistic model verification (see Richardson (2000) and Wilks (2011) for a review).

The first two chapters of this dissertation develop a series of new methods that fit within the effort to correct NWP bias with long observational records on weather time-scales. These methods are developed from modern computer-vision based machine learning techniques and represent the state-of-the-art in linking NWP forecast error and modern observations to improve predictability. Chapter 2 pioneers the use of Convolutional Neural Networks (CNN) in NWP post-processing. These algorithms can readily and flexibly encode spatial information, enabling the use of large, gridded input datasets, which post-processing methods typically are unable to do. Chapter 3 uses CNNs to develop and quantify forecast uncertainty. We further show that the probabilistic forecasts are sharp and reliable.

During the boom of the NWP computational age, Lorenz, (1969b) described the inherent nonlinearities in the tendency equations that led to an upscale transfer of energy (and therefore error) from smaller to larger scales. His work theorized a fundamental limit on skillful NWP on the order of 12 days. However, fundamental research showed the existence of distinct dynamical structures associated with low-frequency atmospheric variability. This research suggested that this variability is attributable to the interaction of planetary scale-waves and is fed by synoptic scale weather events. Additionally, this low-frequency variability is embedded in and is inherently more predictable than synoptic-scale weather events. While much of this variability is a result of the chaotic climate system, a portion of these low frequency modes are associated with anomalous boundary forcing conditions in the equatorial Pacific (i.e. anomalous sea surface temperatures i.e. El Niño/Southern Oscillation (ENSO)) or via planetary scale tropical waves (i.e. the Madden-Julian Oscillation, (MJO, Madden & Julian, 1971)). Bjerknes (1969); Gill (1980); Horel & Wallace (1981); Hoskins & Karoly (1981); Simmons et al. (1983) demonstrated both observationally and mechanistically that boundary conditions force preferred modes of atmospheric extratropical variability. These anomalous boundary forcings drive upper atmospheric divergence and act as a source for barotropic and baroclinic anomalies, driving a Rossby wave response which affects weather globally. The atmospheric science field has termed this remote forcing phenomenon, “teleconnections”. Further, it has shown that, if the boundary conditions can be accurately predicted, the forced midlatitude teleconnection variability can be predicted in tandem.

The final two chapters of this dissertation focus on how the coupling of observations with long-running NWP and climate models can inform and extend predictability by developing "windows of forecast opportunity" for the dominant teleconnection affecting North American weather, the Pacific North American (PNA) pattern (Wallace & Gutzler, 1981b). These chapters discuss the physical mechanisms which actively force the PNA and examine the intraseasonal development of predictability across a boreal winter season associated with these forcing mechanisms.

1.2 Dissertation Overview

1.2.1 Machine Learning & Statistical Methods for Numerical Weather Prediction Post-Processing

Forecast post-processing acts to correct systematic errors in an NWP system by creating a linking function that relates a response variable of interest to a set of predictor variables. Each post processing method develops an individual linking function, but the goal of every system is to nudge a given forecast towards a set of 'ground-truth' observations. Deterministic methods include multiple linear regression approaches (i.e. model output statistics, e.g., Glahn and Lowry, 1972; Carter, Dallavalle, and Glahn, 1989; Wilks and Hamill, 2007), running mean techniques (e.g., Stensrud and Skindlov, 2002; Stensrud and Yussouf, 2003), and algorithms based on Kalman Filtering (e.g. Homleid, 1995; Roeger et al., 2003). Two prolific approaches for probabilistic forecasts, Bayesian model averaging (Raftery et al., 2005) and non-homogeneous regression (EMOS, Gneiting et al., 2005), rely on parametric forecast distributions. A predictive distribution is specified, and the linking function estimates its parameters, for example the mean and the standard deviation in the case of a Gaussian distribution. In the EMOS framework, the distribution parameters are connected to summary statistics of the ensemble predictions through suitable link functions which are estimated by minimizing a probabilistic loss function over a training dataset.

NWP post-processing, essentially, a supervised machine learning (ML) task. In supervised learning, ML algorithms are optimized by examining input/output pairs of defined model training data. The goal of model training is to create an algorithm that when fed an input, predicts an accurate output. Here, the input data is the NWP predicted field (i.e., forecasted temperature), and the output is the observed field (observed temperature). The algorithms task is to learn the forecast model error conditioned on the predictor state and correct for this error. A recent explosion of successes in a subclass of supervised ML, aptly named deep learning (DL), motivated the work in the next two chapters. A more fundamental understanding of DL is provided in chapters 2 and 3, but the topic is briefly introduced here.

In its most naive description, the term "deep learning" is simply a regression task (think, $y_1 = b_1 + \begin{bmatrix} m_n \\ \cdots \\ m_N \end{bmatrix} \cdot \begin{bmatrix} x_n \\ \cdots \\ x_N \end{bmatrix}$) that is forced through multiple, sequential, matrix operations ("nodal layers") rather than a single matrix operation (i.e., y_1 is used as predictor variables in y_2). The term "deep" refers only to the algorithmic architecture, in that, there are multiple layers of matrix operations. Each layer consists of a set of nodes which is constituted by a weighted sum of the nodes from previous layers (analogous to, $[\boldsymbol{m}]$) plus a bias term (analogous to, b). The first layer is an input of the defined predictor variables, and the last layer is the desired predictand. The layers in between the first and the last are termed "hidden layers". The output of each hidden layer is activated by prescribed non-linear functions. The non-nonlinear activations allow the DL model to approximate non-linear relationships between the inputs and the desired output. In the context of NWP post-processing, the input layers are the dynamic model forecast variables and the output layers are the corrected forecasts.

The model “learns” to correct a forecast by observing a myriad of forecast/observation pairs, determining some model gradient of a defined ‘loss’ between the pairs (examples being mean squared error for regression tasks or cross-entropy for categorical tasks) and stochastically walking down this gradient field to some local loss minimum. DL methods typically have thousands of parameters and rely on stable model predictor/observation pairs and long running datasets to be successful. In the following two chapters of this dissertation, we examine 11 years of one deterministic forecast and a 34 year hindcast of a stable NWP model. These data sets are crucial for developing the statistics necessary to accurately characterize and correct model error.

This work sits on the forefront of DL for NWP post-processing in a few regards. To my knowledge, the second chapter is the first example of applying convolutional neural networks (CNN) to spatial model post-processing. CNNs are similar the deep learning networks described above, but account for spatial data rather than single point data. Most post-processing methods rely solely on point-based information to correct forecasts. This new method can develop complex relationships between spatial forecast points (see chapter 2 for more detail). In the third chapter we develop a framework to quantify uncertainty in single-point (deterministic) forecasts using CNNs and distributional regression. The uncertainty is shown to be probabilistically rigorous, leading to calibrated probabilistic forecasts which outperform or compete with calibrated dynamic NWP ensemble systems.

1.2.2 Subseasonal Predictability and Forecast Windows of Opportunity

Subseasonal–to–Seasonal (S2S) predictions exist as a hybrid weather-climate scale interaction encompassing lead times of 2 weeks to 2 years and have been recognized by the academic community as a forecast period requiring additional research (Vitart et al., 2017). S2S forecasts with lead times between 2-weeks and 1-month sit beyond the limit of Lorenz's theoretical forecast window for NWP. Therefore, S2S forecasts typically exist in a veritable 'no-man's land' where atmospheric internal variability dominates the predictable component of the atmosphere. However, there are small forecast S2S periods in which slowly varying climate modes shift the probability density functions of a desired event, influencing overall predictability. Chapters 4 and 5 focus on the boundary forcing derived, physical mechanisms which drive subseasonal forecast skill over the North American sector by examining the forced and internal variability of the PNA pattern and its surface variable manifestations (i.e. temperature and precipitation).

The PNA has been long studied as the dominant midlatitude response to ENSO and MJO forcing (e.g., Horel & Wallace, 1981; Mori & Watanabe, 2008). It is one of the most prominent modes of winter time low-frequency Northern Hemisphere climate variability and has been identified as the driver of anomalous temperature and precipitation (e.g., Deser et al., 2018; Leathers et al., 1991) patterns across the whole of North America. It has been argued that the skill of the seasonal forecasting NWP system in the midlatitudes is dominantly attributable to a model's ability at predicting the PNA pattern (O'Reilly et al., 2017; Vitart, 2004). However, much less focus has been paid to the intraseasonal development of the PNA pattern and its downstream effects.

Studies which focus on a model's representation of the midlatitude jet have shown that the background atmospheric state is crucial for determining the pattern location expression and strength of the PNA (e.g., Dawson et al. 2011; Henderson et al. 2017) during boundary forcing events. Additionally, the predictability of the PNA can be significantly modulated by the internal dynamics of the midlatitude atmosphere inherent to the interaction of the large-scale, mean state jet and the synoptic-scale atmospheric processes (e.g., Palmer 1988). The background state of the jet evolves significantly across a boreal winter. Despite this, most studies examine the PNA response to boundary forcing in a seasonal mean framework. The impact of the annual cycle on the global wind-field and thus the PNA's Rossby wave guide leads to significant dynamic monthly evolution of the midlatitude response to this vorticity forcing (see Chapter 4). Therefore, studies that focus on a seasonal mean rather than accounting for the seasonal development of the background state will yield potentially misleading results by mixing the derived model skill across various degrees of forcing response (Newman & Sardeshmukh, 1998). However, likely due to the relatively short length of the observational record, much less focus has been paid to the intraseasonal development of PNA forecast skill and the tropical drivers of the PNA teleconnections when compared with seasonal forecasting. Chapters 4 and 5 expand the scientific field's current understanding on the physical mechanisms that lead to the evolution of the forced signal and the internal variability of the PNA in boreal winter.

Chapter 4 addresses the problem of a relatively short observation record by developing robust statistics from an atmospheric general circulation model, which are trusted to physically represent the examined processes. The utility of using long running ensembles is explored to inform statistics on forecast windows of opportunity that arise from within the seasonal development of ENSO. This chapter argues that the ENSO climate mode does represent a seasonal shift in teleconnection statistics, but it is not an equal and uniform shift across the boreal winter season. Within the season there are times in which the shift is more present and more dominant, and therefore more predictable (see chapter 4 for details). Chapter 4 uses a perfect prognostic framework to assess predictability and Chapter 5 examines how applicable that predictable skill is to actual weather forecasts. Again, we examine the physical mechanisms governing this shift in teleconnection statistics. Chapter 5 tests the hypotheses laid out in Chapter 4 in a coupled 110-year seasonal hindcast issued by the European Center for Medium-Range Weather Forecasts, and further explores the intraseasonal dynamics of the MJO as a second source PNA forcing and PNA forecast uncertainty.

Lastly, there has been a surge of recent interest from stakeholders and weather agencies to leverage machine learning to aid in subseasonal forecasting efforts (Webb et al., 2017). Recent work has shown that ML/DL models are as skillful as NWP at S2S lead times for predicting North American weather (e.g., Gibson et al., 2021; Hwang et al., 2019). However, ML/DL, especially in the physical sciences, is reliant on domain knowledge to be successful and to discover more model skill. Chapter 4 and 5 expands this domain knowledge and lays out fundamental aspects of the drivers of statistical model skill and can be used to systematically separate and determine ML training regimes. It is the authors hope that the physics of these systems will continue to feed into the atmospheric ML community and the fields will grow in tandem.

1.3 Quick Dissertation Guide

The work presented in this dissertation uses state-of-the-art numerical weather prediction models, climate models, and machine-learning/deep-learning methods combined with modern observational data sets to investigate sources of forecast certainty and of error growth across weather and subseasonal time scales. In Chapter 2 we develop modern deep learning algorithms to correct systematic biases in deterministic forecasts of integrated vapor transport. This work serves as motivation and an initial groundwork for Chapter 3, which uses convolutional neural networks to not only correct forecast error, but also to forecast uncertainty. This new method is tested against the current state-of-the-art forecast ensemble systems and shown to compete with or outperform these systems in every probabilistic forecast measure. This not only represents a significant skill improvement, but also huge computational resource savings.

Chapter 4 examines the physical mechanisms which govern the forced and chaotically manifested PNA from an intraseasonal perspective during El Niño/Southern Oscillation (ENSO) boreal winters. Chapter 4 develops robust statistics from an atmospheric general circulation model, which are trusted to physically represent the desired processes, but often come with their own set of errors and biases. It explores the utility of using long running ensembles to inform statistics on forecast windows of opportunity that arise from within the seasonal development of ENSO. It argues that the ENSO climate mode does represent a seasonal shift in teleconnection statistics, but it is not an equal and uniform shift across the whole season. Within the season there are times at which the shift is more present and more dominant, and therefore more predictable (see Chapter 4 for details). Chapter 5 tests the hypotheses laid out in Chapter 4 in a coupled 110-year seasonal hindcast issued by the European Center for Medium-Range Weather Forecasts. Chapter 4 used a perfect prognostic framework to assess predictability and Chapter 5 examines how applicable that predictable skill is to actual weather forecasts, while examining the tropically derived sources of predictability.

Chapter 2

Improving Atmospheric River Forecasts with Machine Learning

Abstract

This study tests the utility of convolutional neural networks (CNN) as a postprocessing framework for improving the National Center for Environmental Prediction's Global Forecast System's (GFS) integrated vapor transport (IVT) forecast field in the Eastern Pacific and Western United States. IVT is the characteristic field of atmospheric rivers, which provide over 65% of yearly precipitation at some western U.S. locations. The method reduces full field root mean squared error (RMSE) at forecast leads from 3 hours to 7 days (9-17% reduction), while increasing correlation between observations and predictions (0.5-12% increase). This represents a ~1-2-day lead time improvement in RMSE. Decomposing RMSE shows that random error and conditional biases are predominantly reduced. Systematic error is reduced up to 5-days forecast lead, but accounts for a smaller portion of RMSE. This work demonstrates CNNs potential to improve forecast skill out to 7 days for precipitation events affecting the western U.S.

2.1 Overview

Numerical weather prediction (NWP) models provide the atmospheric variables necessary to determine projected atmospheric states, based on a numerical integration of a discretized version of the Navier-Stokes equations (L. F. Richardson, 1922). However, due to uncertainty in initial conditions, numerical approximation, and model deficiencies, error increases non-linearly and NWP forecast skill decreases with model time integration (Lorenz, 1963). Statistical forecast postprocessing techniques, which utilize historical forecasts and observations to correct for error in current predictions, have been found to significantly improve forecast skill across multiple atmospheric variables. Algorithms developed to determine and correct for NWP error include: model output statistics approaches (Carter et al., 1989; Glahn & Lowry, 1972; D. S. Wilks & Hamill, 2007), running mean techniques (e.g., Hacker & Rife, 2008; Stensrud & Skindlov, 2002; Stensrud & Yussouf, 2003), algorithms based on Kalman Filtering (e.g., Delle Monache et al., 2006; Homleid, 1995; McCollor & Stull, 2008; Roeger et al., 2003), and analog-based methods which draw from past events to match designed features of the current forecast to correct it (Delle Monache et al., 2011).

The North American West Coast (NAWC) presents a challenge in water forecasting. Wintertime precipitation provides almost all the annual input to the water budget, generally within a few large horizontal vapor transport events (Dettinger et al., 2011) termed atmospheric rivers (ARs). ARs are long (>2000 km) and narrow (<1000 km) corridors of anomalous vapor transport, typically associated with a low level jet, ahead of the cold section of an extratropical cyclone (e.g., Dacre et al., 2015; Sodemann & Stohl, 2013; Warner et al., 2012), which deliver the majority of poleward vapor transport (>90 %) in less than 10 % of the zonal circumference of the extratropics (Ralph et al., 2004; Zhu & Newell, 1998). Vertically integrated vapor transport (IVT) is the characteristic metric which defines the strength of an AR (Ralph et al., 2019). IVT is a combined thermodynamic and momentum metric which integrates specific humidity and zonal and meridional components of the wind from 1000 to 300 hPa.

ARs contribute 30-65% of annual precipitation on the U.S. West Coast, and ARs contribute 60-100% of the most extreme NAWC hydrometeorological events (Gershunov et al., 2017; Lamjiri et al., 2017). Lavers et al. (2016) found that IVT evolution is dominated by synoptic scale processes, and thus has a higher predictability than precipitation, which depends more on mesoscale and microphysical processes. Therefore, at long lead times, forecasting IVT, rather than precipitation, may be more valuable to water management and hazard mitigation. However, forecasting for AR events has proved difficult. A study by Wick et al. (2013) examined the National Centers of Environmental Prediction's Global Forecast System (GFS) West Coast forecast skill over the Northeast Pacific across three cold seasons and found that average AR landfall location errors were approximately 600 km at seven days lead time.

We propose a novel NWP postprocessing technique, applied to the IVT field, that leverages a sub-class of machine learning computer vision techniques: convolutional neural networks (CNN). CNNs are able to encode features from an input field, at varying spatial scales and levels of abstraction (Bengio, 2009; Hinton, 2006), which maximize predictive skill to a specified output field. These networks are adept at processing large and complex datasets and determining meaningful relationships. CNNs have proven to be extremely successful at image recognition, semantic segmentation, image denoising, and image super resolution (Bojarski et al., 2017; Dong et al., 2014; He et al., 2016; Long et al., 2015b; K. Zhang et al., 2017). CNNs are well suited to atmospheric fields, where systems across multiple scales govern atmospheric flow.

More recently, flexible forecast prediction and postprocessing approaches based on artificial neural networks, which take advantage of increased computational power to learn from a large database of past forecasts, have been proposed (e.g., Tao et al., 2016). Neural networks reduced bias and improved ensemble 2-m temperature prediction over Germany (Rasp & Lerch, 2018). Random forests have been used for storm-based probabilistic hail forecasting (Gagne et al., 2017). When combined with the physical understanding of atmospheric processes, machine learning has been shown to aid in high impact weather decision making (McGovern et al., 2017). Specifically, CNNs are beginning to be used for scientific discovery and forecasting and have emerged as diagnostic tools for determining important atmospheric variables across scales (e.g., Kurth et al., 2018; Toms et al., 2019). CNNs have been utilized to provide forecast uncertainty estimates upon initialization (Scher & Messori, 2018). Additionally, purely CNN-based forecast methods have arisen for prediction and nowcasting applications, relying on data alone to mimic atmospheric dynamics (Dueben & Bauer, 2018; Scher, 2018; Scher & Messori, 2019; Xingjian et al., 2015). This study aims to extend the utility of CNNs as a postprocessing method to improve predictions up to 7-days ahead.

At every forecast lead time, we create a new CNN which inputs a GFS IVT magnitude forecast field and outputs a corrected IVT forecast field. The present study evaluates whether historical forecast error can be used in conjunction with CNNs as a postprocessing tool to improve short- and medium-range IVT forecasts.

2.2 Data and Methodology

2.2.1 Forecasts

GFS predictions (Moorthi et al., 2001) at a 0.5-degree horizontal spatial resolution on 64 vertical levels for daily 0000 and 1200 UTC model initializations are utilized to calculate forecasted IVT. Here, the forecasts from 3-168 hours are examined (3-hour increments for the first 12-hour period, 12-hour increment for the following day, and 24-hour increments for the remaining 168-hour forecast lead times) for the cold season (defined here as October-April) from 2006 to 2018. This includes ~5,000 data fields for every forecast lead time, or ~55,000 forecasted fields across all lead times. This study's region of interest (ROI) spans coastal North America and the Eastern Pacific from 180° W to 110° W longitude, and 10° N to 60° N latitude.

2.2.2 Ground-truth

IVT from the National Aeronautics and Space Administration's Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) reanalysis is used as ground truth to diagnose forecast error and CNN model training. MERRA-2 provides a regularly gridded record of the global atmosphere, including assimilated satellite, surface station, wind profiler, radio occultation and radiosonde observations. MERRA-2 data is resolved on a 0.625 x 0.5-degree grid and interpolated to 21-pressure levels between 1000 and 300 hPa for IVT calculations (Gelaro, McCarty, Suárez, Todling, Molod, Takacs, Randles, Darmenov, Bosilovich, Reichle, Wargan, et al., 2017; Will McCarty et al., 2016). For consistency, GFS IVT is regridded and upscaled to MERRA-2 resolution using a first- and second-order conservative remapping scheme (Schulzweida, 2019).

2.2.3 Methodology and Experimental Design

We compare four separate forecasts to examine the relative skill of the CNN postprocessing; 1) GFS is used as the dynamical NWP model and provides a deterministic forecast of future IVT states from current meteorological observations; 2) a climatological forecast (CF) created from a 21-day running mean, centered on the forecast day of interest, from MERRA-2 IVT fields spanning 1980-2018; 3) a persistence forecast (PF) created by repeating the GFS analysis at 0-hour lead for every lead time; 4) a forecast derived from postprocessing the GFS IVT forecast with a CNN (hereafter referred to as ARcnn when referencing the architecture and ARcnn-IVT for the forecast).

2.2.4 Convolutional Neural Networks and the Network Used in this Study

Neural networks are known to be able to approximate nonlinear functions (Nielsen, 2015). CNNs are a class of neural network, in which multiple layers of optimized functions map input data fields (GFS forecasts in this study) to an output (ARcnn-IVT). CNNs use convolutional kernels to propagate images from one layer to the next. Each convolutional kernel is trained to highlight important image features. Following each convolutional layer, nonlinearities are introduced, which operate on every produced feature map. ARcnn was inspired by a class of CNNs termed *denoising autoencoders* (Vincent et al., 2008).

Denoising autoencoders are trained, with coupled pairs of noisy and clean images, taking a noise corrupted image and removing that noise. Here, GFS IVT forecasts are treated as noisy images, the noise representing the prediction error, and ARcnn corrects the forecast towards a clean image, MERRA-2 ground truth. ARcnn contains no compression or pooling information layers which reduce dimensionality. Therefore, a consistent dimension (determined by the latitude and longitude points of the ROI) is retained throughout the network and in the prediction.

The optimization of the kernel filter weights occurs iteratively, in which each iteration finds the weights of the functions to minimizes the loss between the output (ARcnn-IVT) and a desired field (MERRA-2). ARcnn utilizes an Adam optimizer (Kingma et al., 2014) with a learning rate that decreased from 0.001 to $5e^{-6}$ upon validation plateaus and batch size 20. The error is determined between the network forecast and the ground truth data, and the gradient of the error field is calculated for each kernel weight of the network. The model weights update each iteration by stepping in the direction opposite of this gradient. ARcnn optimized utilizing mean squared error loss. Once trained, ARcnn produces an estimated IVT field that has learned error from previous forecasts and has the ability to correct a portion of these errors. A detailed description of CNNs and the ARcnn model architecture is in the supporting material. For further information on CNNs the reader is referred to Nielsen, (2015).

GFS forecasts were separated by date into training (October 2008 – April 2016), validation (October 2016 – April 2017), and testing (October 2017 – April 2018) datasets. Training data is shown iteratively to the neural network to optimize CNN model weights. Validation data is used to compute performance metrics during training. Testing data is unseen by the network and utilized only for evaluating the postprocessing skill. The final year of data (October 2017 -April 2018) is reserved for testing and is independent from any training data. Each lead time in the testing period consists of ~450 forecasts. Table 2.1S shows the number of samples and the frequency of ARs in the training, validation, and testing datasets. Each forecast lead is trained, validated and tested on ~5000 forecasts. A new CNN is created and trained for each forecast lead time. However, across these CNNs, there is valuable similarity in the IVT feature detection during convolution. To exploit this similarity during training, we utilized a sequential training scheme in which the model network weights from previous forecast lead times initialized network weights at subsequent forecast leads. This decreased the number of model training cycles and improved total error testing results (not shown).

Table 2.2S summarizes the model architecture and training parameters. An exhaustive number of training cycles, using common CNN model parameters, was performed to determine optimal model settings. The final parameters were selected by choosing the configuration with the lowest validation error.

2.2.5 ARcnn Example

ARcnn output valid for 29 November 2017 illustrates potential forecast improvements (Figure 2.1). The GFS forecast IVT field at 96-hour lead time (Figure 2.1b) is input into the 96-

hour ARcnn. Once the network has been trained, a postprocessed forecast is generated within milliseconds. The IVT field passes through ARcnn (Figure 2.1A), and a corrected field is produced (Figure 2.1c). The resultant field is compared against ground truth (Figure 2.1a). GFS over-predicts the magnitude of IVT, has a notable location error, and misses the primary orientation of the storm. After the GFS IVT field is processed with ARcnn, the network correctly reduces the magnitude of peak IVT, particularly at high latitudes near the Alaskan Coast, moving the dominant IVT signal southward and eastward (Figure 2.1d). Additionally, ARcnn reorients the dominant AR spatial axis to a more accurate zonal direction (Figure 2.1e vs. 2.1f), leading to a more accurate forecast. Figure 2.1 is a representative sample of the method drawn from the 95th percentile of corrected 96-hour events in the testing data set (as measured by RMSE).

2.3 Verification Metrics and GFS Error Patterns

Forecast error (*e*) is defined as the difference between the forecasted IVT field (*f*) and the ground truth (*r*) IVT field *(e = f - r*) at a given time and location. We have applied four metrics to the forecast systems: root mean squared error (RMSE), bias (Bias), centered root-mean-square error (CRMSE), and spatial Pearson Correlation (PC) coefficient. Bias and CRMSE arise from a decomposition of RMSE (Taylor, 2001). Bias represents the systematic error, defined as the mean error over the test data set (Bias = $\bar{e}$). CRMSE is the remaining random error and conditional biases, which contains the error not present from mean shift (CRMSE = $\left\{\frac{1}{N}\sum_{n=1}^{N}\left[\left(f-\bar{f}\right)-\left(r-\bar{r}\right)\right]^{2}\right\}^{0.5}$). Finally, the Pearson correlation indicates the linear relationship between the forecasted and observed time series ($PC = \frac{E(f,r)}{\sigma_f \sigma_r}$).

2.3.1 GFS error patterns

The largest sources of GFS forecast error occur predominantly in the locations with high climatological IVT, indicating that AR position, magnitude, and timing constitute a large fraction of total error. Figure 2.2 shows the 96-hour error metrics for every GFS forecast in the dataset. The RMSE field is dominated by random error and conditional bias over systematic error (as indicated by high values of CRMSE, Figure 22.b, as compared with Bias, Figure 2.2c). The AR corridor, defined here as the 200 kg m^{-1} s^{-1} IVT contour, for the 2006-2018 MERRA-2 climatology (Figure 2.2 contours), coincides with the greatest magnitude of CRMSE in the field. The model systematically under predicts IVT magnitude at high latitudes and over predicts IVT at low latitudes. The highest levels of PC occur on the southern flank of the AR corridor, in the climatological subtropical jet region (Figure 2.2d). This may be associated with the lower predictability of mesoscale frontal waves associated with ARs. Conversely, the latitudinal band of high predictability exists within the jet region and is an area of largely synoptically forced IVT processes. This latitudinal band of predictability is consistent with findings in Lavers et al. (2016).

2.4 Results

All statistics will be presented from a seasonal perspective derived from the testing data set (October '17 - April '18). The Guan and Waliser (2015) AR detection algorithm identified an AR present in 76% of the forecast periods. The AR distribution is not spatially uniform, (Figure 2.2 contour), with a skewness towards high latitude, with landfalls predominantly in Oregon, Washington, and southern British Columbia.

ARcnn-IVT performance is evaluated at 3-hourly forecast intervals out to 12 hours, a 12-hour forecast interval out to 24 hours, and in 24-hour increments from 24 hours forecast lead until the 168-hour lead (7-day). Results for each forecast system are resampled 2000 times for error metrics using a 30% split; the variance (colorshade) of the bootstrapped sample are small compared to the mean (Figure 2.3). At 3-hour lead time, GFS and ARcnn-IVT outperform persistence and climatology, with the postprocessed ARcnn-IVT further improving on Bias and CRMSE over GFS. At the fifth forecast day the correction of ARcnn-IVT bias begins to deteriorate and the bias is statistically even between GFS and ARcnn after this point (Figure 2.3a). CRMSE (Figure 2.3b) continues to improve as compared to GFS for the entire testing period. Importantly, the magnitude of CRMSE dominates Bias and therefore the RMSE is improved (Figure 2.3c). At the seventh day forecast lead GFS has a larger RMSE than climatology.

However, ARcnn remains the most skillful forecast (by total RMSE). The magnitude of RMSE error at the 7th day for ARcnn is equal to that of the GFS at the 5th day. This is due to the reduction of CRMSE by the postprocessing technique. Similarly, ARcnn has a higher correlation with the ground-truth at every lead time, with statistically significant differences starting at hour 12. The PC of the ARcnn of the 7th day is equal to that of the 6th day for the GFS forecast.

Figure 2.5S shows the spatial distribution of RMSE, CRMSE, and Bias for GFS and ARcnn for the full testing dataset. Figure 2.4 shows the spatial ARcnn-IVT metrics of performance at the 96-hour forecast lead, with cool colors indicating that ARcnn-IVT is improving the GFS forecast, conditioned on forecasted IVT values with over 250 kg m^{-1} s^{-1}, to ensure that the network is correcting for high vapor transport events. After ARcnn is applied, each pixel is assessed for RMSE, Bias, CRMSE, and PC, resampling 1000 times utilizing 50% of the available data field in order to estimate error metrics. Importantly, RMSE (Figure 2.4a) at almost every grid point is decreased, indicating forecast improvement. Additionally, PC (Figure 2.4d) is improved at most locations with very few exceptions in the spatial domain, indicating a more skillful forecast.

The Bias field (Figure 2.4c) shows the least improvement, where ARcnn-IVT systematically under predicts the magnitude of high-valued IVT. Overlaid on the figure are the ±40 contours of GFS Bias. It is clear that the dominant sources of systematic error are targeted by ARcnn (as indicated by cool colors contained inside the ±40 contours, and figure 2.5Sc and 2.5Sf). However, the strongest failures in the Bias field comes over the areas of coastal landfall. The field is almost uniformly improved for CRMSE (Figure 2.4b).

Due to a low contribution of systematic error compared to random error and conditional bias, the RMSE is still dominantly benefiting with ARcnn postprocessing. Overall, ARcnn generates an IVT field with significantly more skill than GFS. When compared to GFS, ARcnn increases correlation between ground truth and predictions at all lead times (0.5-12% increase), and the method improves RMSE at forecast leads ranging from 3 hours to 7 days (9-17 % reduction), equivalent to an increased forecast skill time horizon of 24 and 48 hour/day improvements, respectively. For context, NWP forecast systems, through model improvements and assimilation of more observational data, have historically achieved an RMSE error skill improvement of ~1 day every 10 years (Magnusson & Källén, 2013).

Interpretable CNNs are an active area of research in the machine learning community (e.g., Kuo et al., 2019), and ongoing research involves using CNNs to elucidate physical processes associated with forecast error. We speculate ARcnn-IVT improvements to CRMSE involve corrections to conditional bias. Upon exhaustive inspection of individual testing forecasts, it appears that ARcnn is recognizing common IVT structures and correcting the IVT fields in similar ways given that shape. Conditional bias, i.e., conditioned on storm shape and magnitude, is the most accurate terminology to describe this correction. IVT systems that appear similar to Figure 2.1b are similarly corrected, with a reduction in high latitude IVT and a zonal elongation of the IVT signal. Whereas with IVT fields that are zonally stunted, ARcnn reduces the total IVT and moves the IVT signal eastward, indicating GFS typically propagates this signal too slowly. CNNs are adept at modulating output based on input spatial field encodings. The strength of this method is the adaptive adjustment given a wide range of forecasted fields. This kind of correction results mostly in a CRMSE reduction rather than a Bias correction.

Importantly, coastal landfalling IVT 96-hour forecast RMSE is significantly improved for IVT forecasts greater than 250 kg m^{-1} s^{-1}. A detailed examination of coastal error (RMSE, CRMSE, Bias, and PC) can be found in Figures 2.2Sand 2.3S. The RMSE error reduction is found to be significant (90th percentile) which is important for the societal impact of landfalling ARs. Similar error reduction spatial patterns were observed for all forecast lead times (not shown). For low IVT forecasts (IVT < 250 kg m^{-1} s^{-1}) (Figures 2.3S and 2.4S), the improvement in forecast skill (as measured by RMSE, CRMSE, Bias, and PC) is even greater, with a significant improvement to RMSE, CRMSE, and PC, and no significant change to Bias.

2.5 Summary and Conclusion

This paper explored the utility of convolutional neural networks (CNNs) to improve integrated vapor transport (IVT) 0-7 day forecast skill. We have shown that CNNs can be used to improve forecast prediction of the GFS numerical weather prediction model for the North American West Coast and Eastern Pacific IVT 3-168 hour forecasts. This postprocessing is beneficial at every forecast lead time in reducing full field CRMSE and improves Bias out to 5 forecast days, leading to a full field RMSE improvement. ARcnn yields significantly higher PC between forecasted and ground-truth values at all lead times over 12 hours. ARcnn provides a forecast that has greater skill than climatology, compared to GFS that degraded below climatological skill at 7 days lead. Ongoing work involves testing this method on an ensemble system to determine the benefit on accuracy and uncertainty quantification.

CNN postprocessing was shown here to increase IVT forecast skill. Additionally, the success of a deep learning relies on the quantity of data. As forecasts are produced, CNN postprocessing techniques stand to improve as a more fully sampled distribution of AR activity is realized. CNNs continue to evolve, and model architectures are continuously under development. Opportunity exists for the weather prediction community to leverage computer vision advances. While a stand-alone machine learning weather prediction that competes with modern NWP has not been developed, combining numerical weather prediction with a data-derived CNN deep learning correction is a logical step in forecast improvement.

2.6 Acknowledgements

Datasets used to create analyses supporting this study are hosted within the University of California San Diego Library Digital Collections (https://doi.org/10.6075/J0D798R5). This study is supported by the U.S. Army Corps of Engineers (USACE)-Cooperative Ecosystem Studies Unit (CESU) as part of Forecast Informed Reservoir Operations (FIRO), grant W912HZ-15-2-0019 and the California Department of Water Resources Atmospheric River

Program, grant number 4600010378 TO#15 Am 22. The authors would like to thank Microsoft's AI for Earth project (https://www.microsoft.com/aiforearth) for contributing computational resources, making this work possible. Additionally, we thank the anonymous reviewers for their insightful comments that strengthened this work.

Chapter 2, in full, is a reprint of the material as it appears in Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving Atmospheric River Forecasts with Machine Learning. Geophysical Research Letters. The dissertation author was the primary investigator and author of this paper.

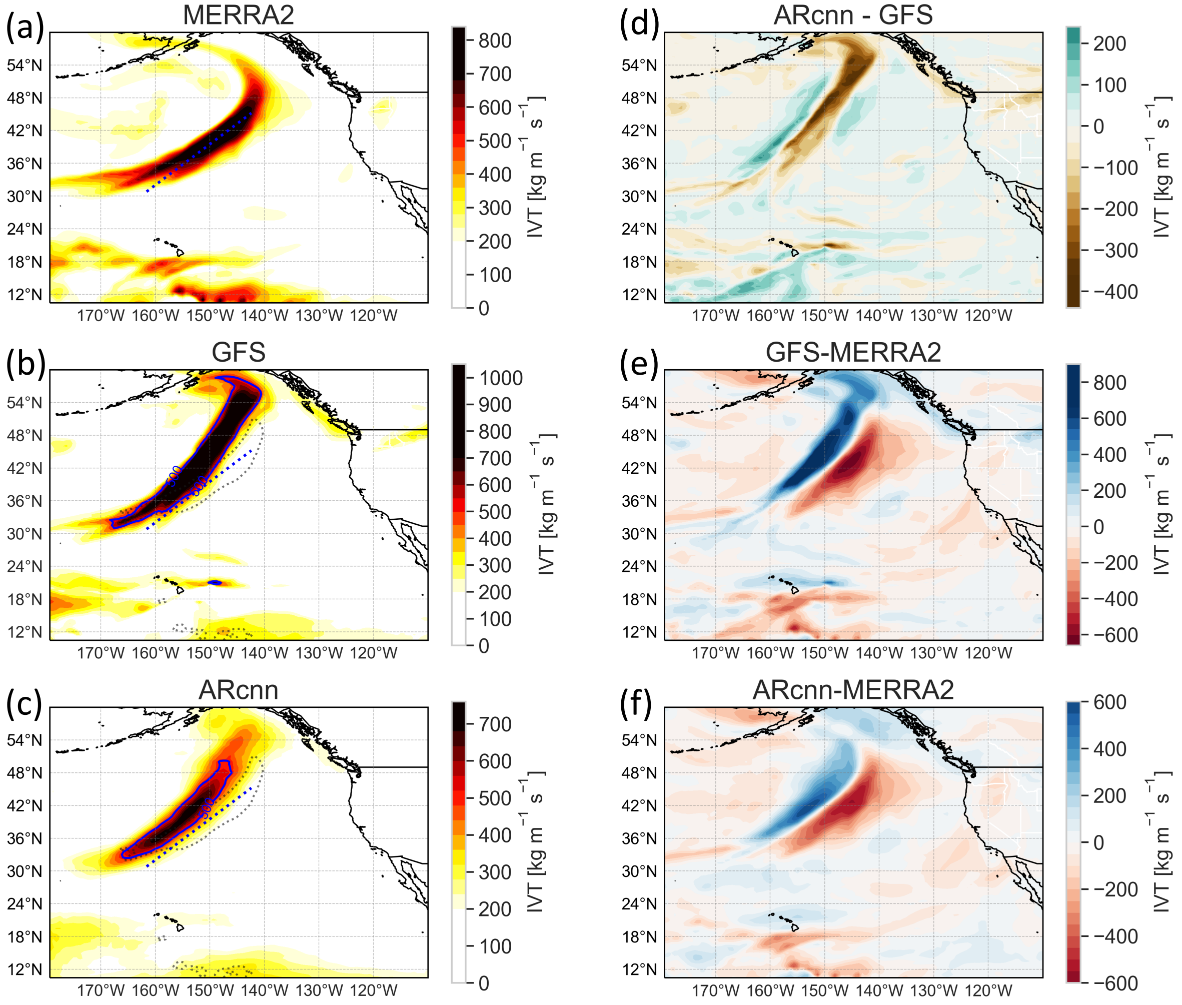

Figure 2.1. Forecasts and analysis valid for IVT fields on 29 November 2017. (**a**) MERRA-2 analysis field with the IVT = 600 kg m^{-1} s^{-1} contour (solid) and dominant storm axis (dotted) as determined by IVT > 350 kg m^{-1} s^{-1} raw image moment. (**b**) GFS 96-hour forecast with the MERRA-2 600 IVT contour and dominant storm axis. (**c**) ARcnn-IVT 96-hour forecast with the MERRA-2 600 IVT contour and dominant storm axis. (**d**) Difference between ARcnn-IVT and GFS. (**e**) Difference between GFS and MERRA-2 IVT field. (**f**) Difference between GFS and MERRA-2 IVT field.

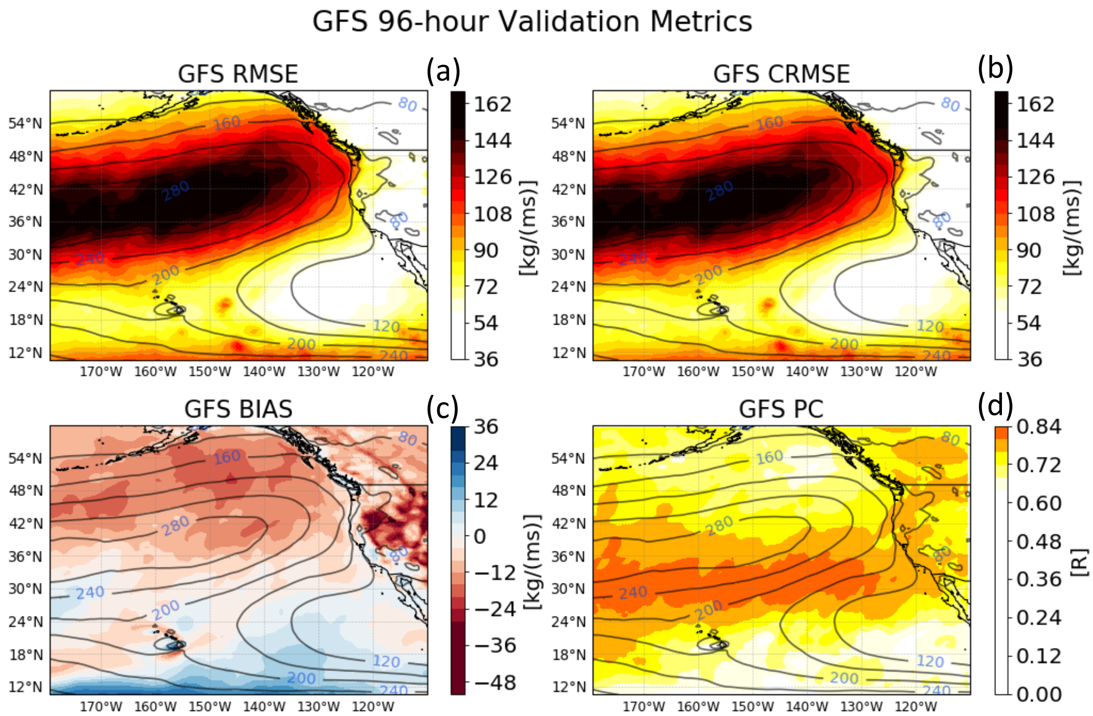

Figure 2.2. Spatial distribution of 96-hour forecast GFS (**a**) RMSE (**b**) CRMSE (**c**) GFS forecast Bias. (**d**) Pearson correlation (in color) and (**a, b, c, d**) climatological AR field (in contour). Forecast dates range from the Oct 2006 - April 2018.

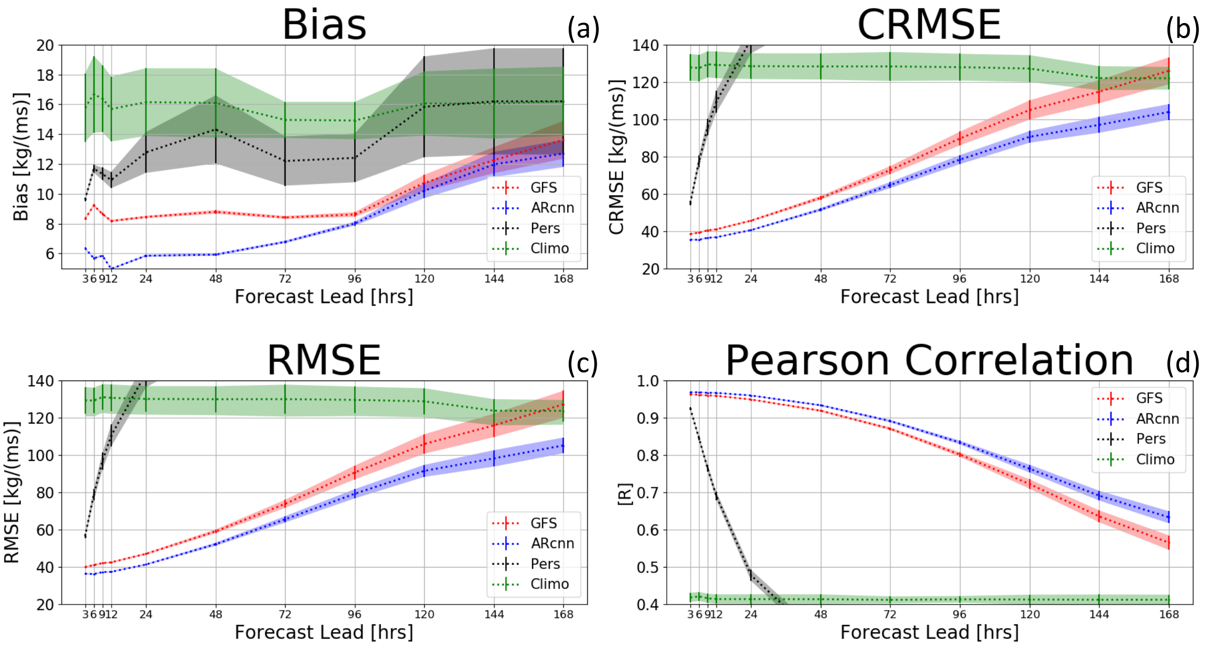

Figure 2.3. ROI average temporal evolution of (**a**) Bias, (**b**) CRMSE, (**c**) RMSE, and (**d**) PC of raw GFS, ARcnn-IVT, persistence (Pers), and climatology (Climo) forecasts. Resampled bootstrap variance intervals are shown for each forecast.

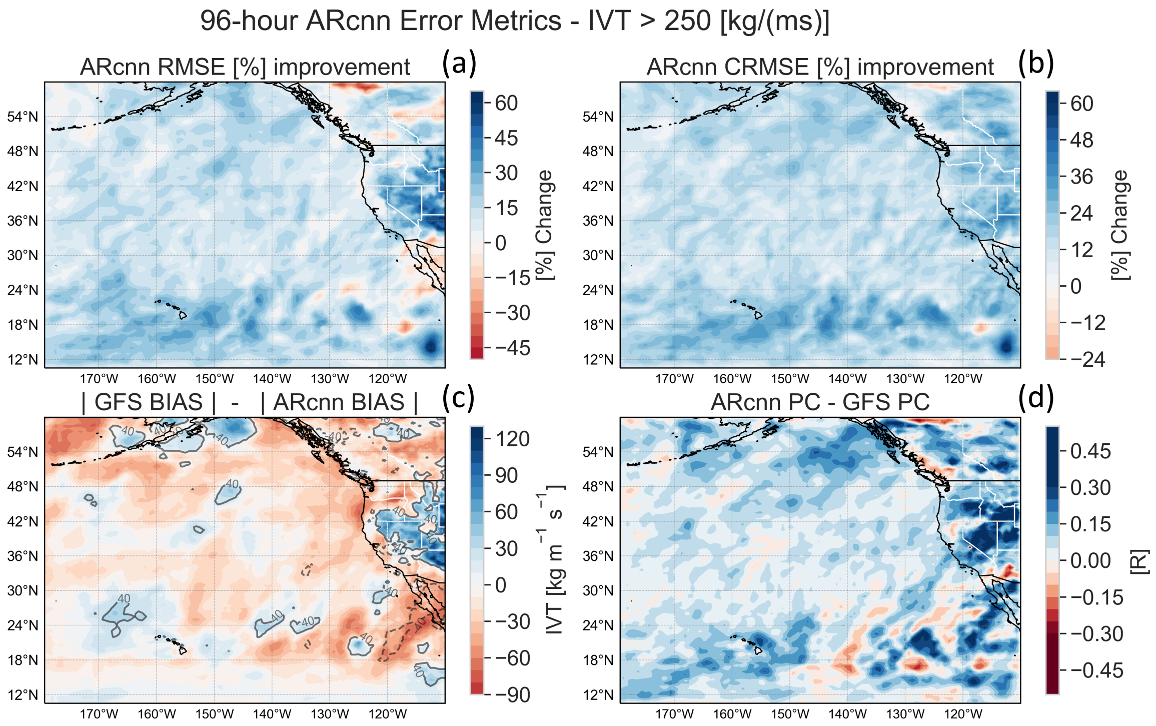

Figure 2.4. Spatial distribution of percent improvement of 96-hour IVT forecast after ARcnn postprocessing for (**a**) RMSE, and (**b**) CRMSE. Contours indicated average IVT field. Spatial distribution of the 96-hour forecast difference between GFS minus ARcnn for (**c**) Bias (Contours indicate GFS Bias fields of 40 units kg m^{-1} s^{-1}, dashed lines are negative), and (**d**) Pearson correlation. Calculated for locations when IVT forecast is over 250 kg m^{-1} s^{-1}. All dates from October 2017 - April 2018 testing dataset. In all plots (**a, b, c, d**), cool colors imply that the CNN postprocessing is improving the forecast.

2.7 Chapter 2 – Support Information

This supporting information is included for further reference for the reader. The text is dominantly a description of the specific architecture and implementation of the convolutional neural network designed for this study. Water managers and constituents are largely concerned with AR statistics at coastal landfalling points. We therefore show the same statistics as referenced in Figure 2.3 of the main text, highlighting only the coastal landfalling points. Lastly, we show the error fields (RMSE, CRMSE, and Bias) for the raw GFS and the Post-Processed forecast (ARcnn) on the testing dataset.

All forecast data for training, validating and testing this network is publicly available from the National Center for Environmental Prediction's Global Forecast System's 0.5° model.

2.7.1 ARcnn Architecture and Methodology

Figure 2.1S shows the CNN architecture of ARcnn. We build ARcnn using the Keras library (Chollet, 2015) with the Tensorflow back end (Martín Abadi et al., 2016). ARcnn contains 9 separate layers with the functions performed in those layers separated by arrows. CNN's are a class of artificial deep learning networks, in which multiple layers of optimized functions, map input data fields (GFS) to an output (ARcnn). Each layer of the CNN contains a convolutional layer (conv2d in Figure 2.1) which contains a three-dimensional array of feature mapping filters. These filters are small patches (3 x 3 matrices for all of ARcnn convolutions) of weights that are slid across the input two-dimensional IVT image and create output feature maps by summing the product of the weights and the input field. The number of filters per layer is specified by the user and shown as the number following the “conv2d” in each model layer in

Figure 2.2S. The produced feature maps are input to ‘activation layers’ (both RELU and sigmoid layers here) with non-linear functions, which enable non-linear relationships to be learned from the input field to the output. In ARcnn, the activation layer is a rectified linear unit function, this is a common choice of CNNs as RELU gradients are fast to calculate during training (1 if positive, 0 if negative) (Lecun et al. 2015), and hence train quickly. The final activation layer is a sigmoid function, which is a smooth and bounded function with real valued outputs, and is thus ideal for this application.

CNN’s typically include compression layers, which reduce the dimensionality of each layer input thus reducing the required amount of filter training parameters. The compression is achieved by increasing the stride of the sliding windows, or by instituting pooling layers, which act on small matrix patches and reduce each patch to a scalar via a specified function (averaging, maximum, etc). However, in ARcnn pooling or compression layers decreased validation accuracy during training, and were not included, resulting in a consistent feature map height and width determined by the latitude / longitude of the ROI, and a layer depth specified by the number of requested convolutional filters.

Due to the spatial averaging of convolution, deeper convolutional layers act on a larger spatial extent in the input (termed ‘receptive field’). In order to further increase the receptive field of deeper layers, dilated convolutions (Yu & Koltun, 2015) are implemented at varying depths (details shown Figure 2.2S and dilation size shown in Table 2.2S). Increasing the receptive field is logical in meteorological applications as, not only local, but remote features have influence over IVT characteristics. Batch normalization (batch norm in Figure 2.1S), which re-centers and normalizes individual layers, after each convolution was found to drastically decrease validation error and over-fitting (Ioffe & Szegedy, 2015). This network borrows from the Residual Network architecture that introduced skipped connections (represented by arrows from boxes (1) to (8); (2) to (7); (3) to (6); (4) to (5) into addition layers in Figure 2.2S), which force the network to learn residual corrections to layer inputs (He et al., 2015). The skipped connections symmetrically connect shallow layers to the deeper layers and allows information to more easily propagate from the less abstracted convolutions (shallower) to deeper sections of the network. The skipped connections aid in better back propagating the gradient as gradients are passed through both the weight blocks and the skipped connection.

Optimizing the kernel filter weights occurs iteratively on predetermined training data through gradient descent, in which each iteration finds the weights of the functions that minimizes the loss (here, mean squared error) between the output (ARcnn) and a desired field (MERRA2). Multiple gradient descent methods where tested, and an Adam optimizer (Kingma et al. 2014) with a learning rate that decreased from 0.001 to 5e^{-6} upon validation plateaus was selected. The training occurs in batches of 20 IVT forecasts, selected at random, the batches continue until the training data has been exhausted. The error loss is calculated between the network prediction and the ground truth data, and the gradient of the error field is determined for each kernel weight. The model weights update each iteration by taking a small step in the direction opposite of the gradient field.

2.7.2 Land Falling AR Statistics

Figure 2.2S shows the landfalling error metrics for 96-hour GFS and ARcnn, for points in which IVT is forecasted to be greater than 250 kg m^{-1} s^{-1}. Here, only the whole number latitudinal points within the shaded region are included in the analysis. There is significant improvement to the landfalling RMSE and CRMSE (90th percentile), and Bias is significantly degraded. However, random error dominates this field and therefore total error is reduced. Correlation does not show a significant improvement, but there is a clear trend, and based on full field statistics ARcnn is significantly improving the forecast field PC. With the addition of more latitude/longitude the PC, becomes further improved. Every other forecast lead time show similar trends (not shown).

Figure 2.3S shows the landfalling error metrics for 96-hour GFS and ARcnn, for points in which IVT is forecasted to be less than 250 kg m^{-1} s^{-1}. Here, only the whole number latitudinal points within the shaded region are included in the analysis. There is significant improvement to the landfalling RMSE and CRMSE (90th percentile). Bias is statistically unchanged but shows a trend towards degradation. However, Bias accounts for a very small fraction of total error. Correlation for low IVT forecasts is significantly improved. Every other forecast lead times show similar trends (not shown). The entire region of interest is showed for the 96hr lead of low magnitude IVT (< 250 kg $m^{-1}s^{-1}$) in Figure 2.4S. Similar magnitudes of improvement can be seen in total field RMSE, and while Bias is negatively attenuated, it does not suffer from the same large Bias degradation as seen in Figure 2.4c. For low magnitude IVT (< 250 kg $m^{-1}s^{-1}$) forecast correction ARcnn is extremely beneficial.

Figure 2.5S shows the background error field in the testing dataset of GFS and ARcnn, respectively. There is a significant improvement across the RMSE and CRMSE fields. However, the Bias term shows weak improvement or detriment depending on the spatial area examined. Although Bias across the entire field accounts for a low fraction of error, locations of relatively high Bias (e.g., the U.S. landmass and the Aleutian Islands) largely benefit from ARcnn postprocessing. Other areas suffer from slight Bias insertion (e.g., the Oregon Coast).

Because the fraction of Bias is low compared to the CRMSE, the network corrects mainly the latter, specifically conditionally biases, which are reflected in errors impacting the position and magnitude of IVT in the AR region.

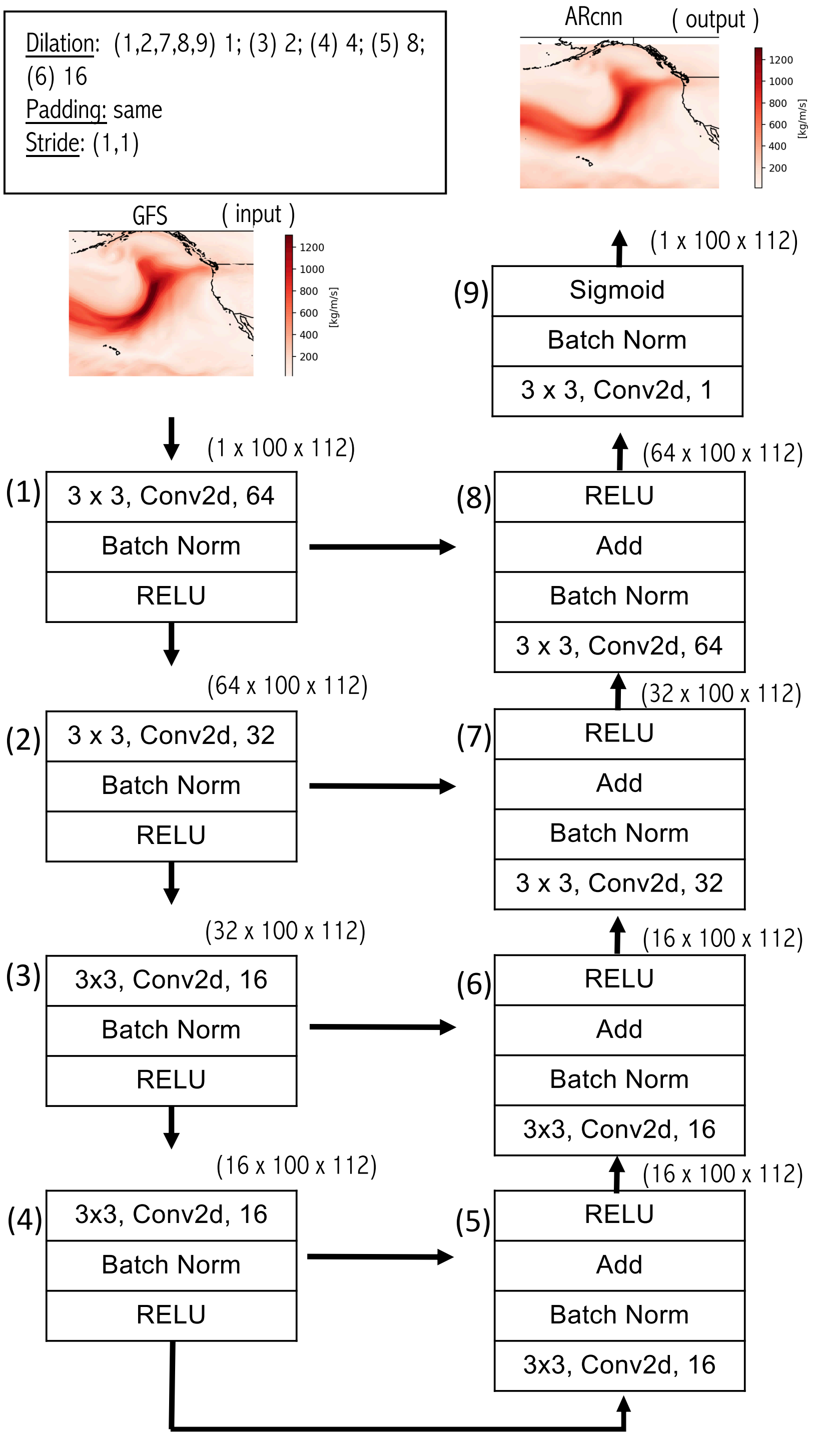

Figure 2.1S. ARcnn model architecture.

IVT > 250 Kg m^{-1} s^{-1}

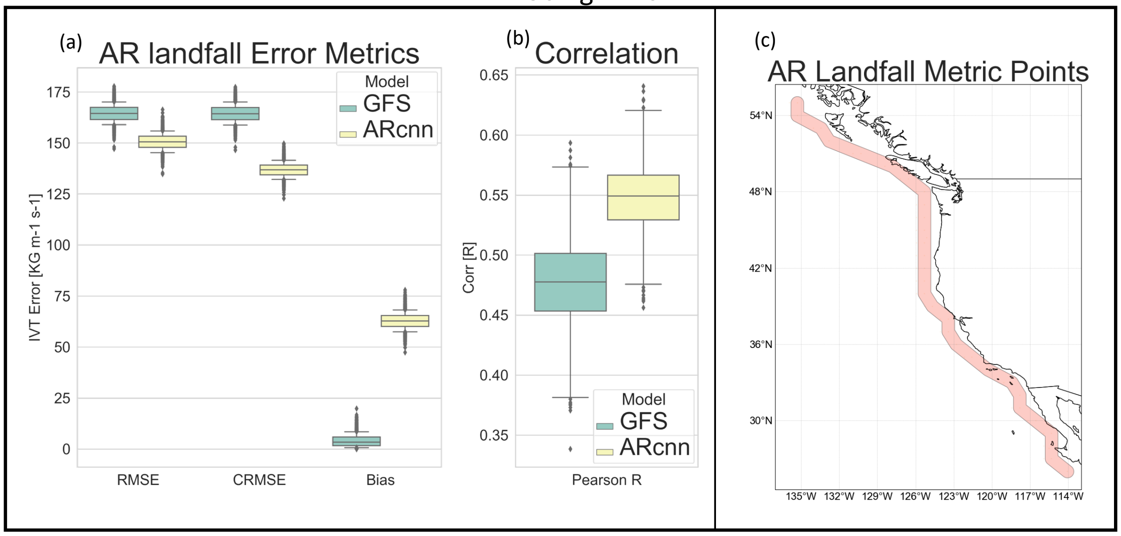

Figure 2.2S. Coastal (as defined by the shaded swath in panel (c)) (a) RMSE, CRMSE, Bias, and (b) Pearson Correlation for landfalling AR error metrics. Only IVT forecasts < 250 kg m^{-1} s^{-1} were considered.

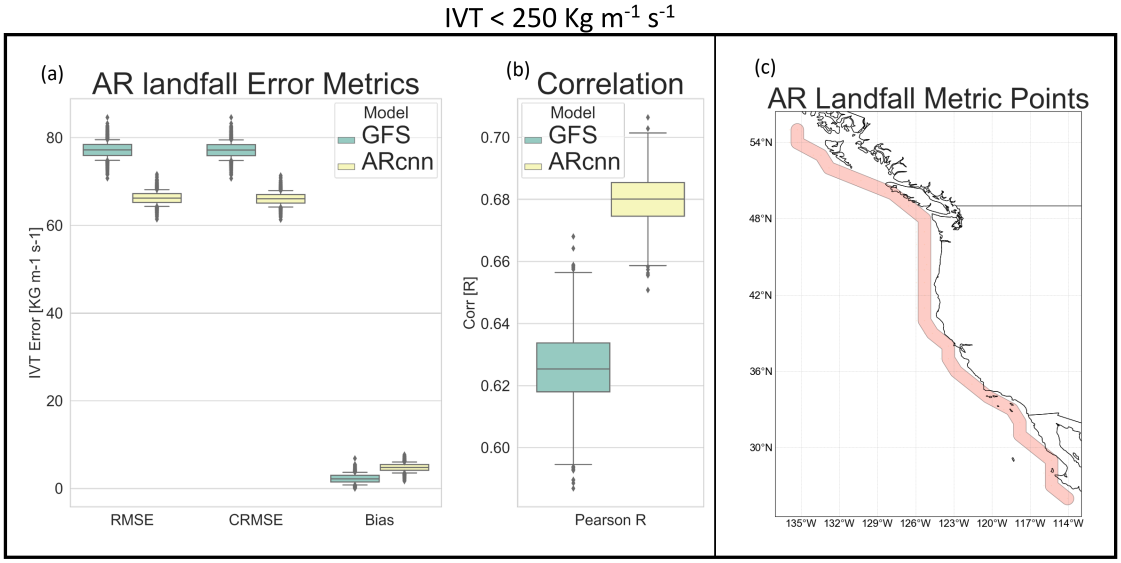

Figure 2.3S. Coastal (as defined by the shaded swath in panel (c)) (a) RMSE, CRMSE, Bias, and (b) Pearson Correlation for landfalling AR error metrics. Only IVT forecasts < 250 kg m^{-1} s^{-1} were considered.

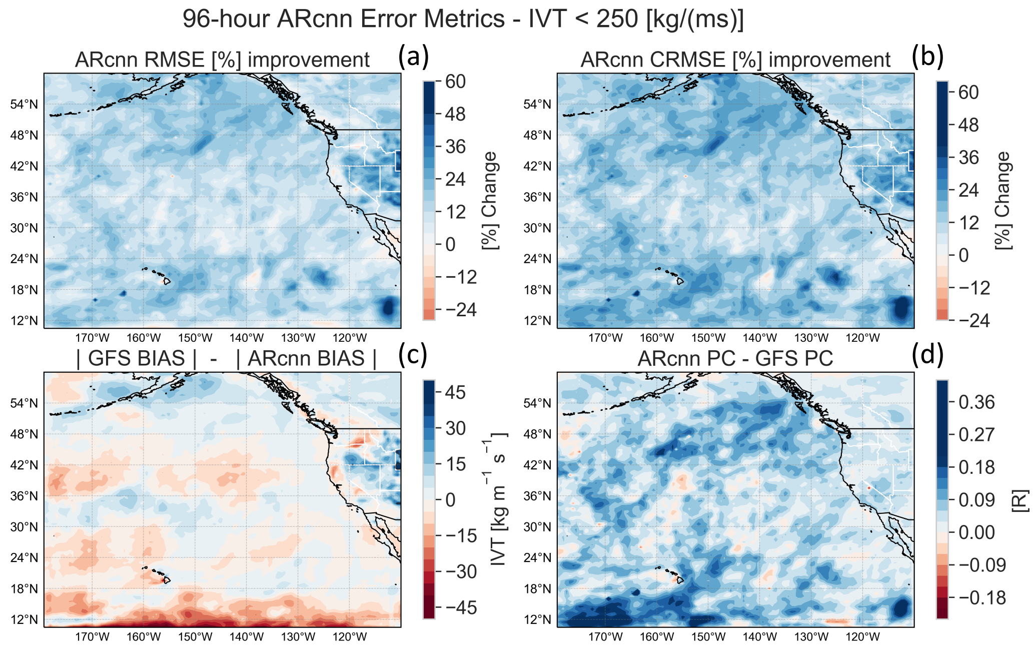

Figure 2.4S. Spatial distribution of percent improvement of 96-hour IVT forecast after ARcnn post-processing for (a) RMSE and (b) CRMSE. Contours indicated average IVT field. Spatial distribution of the 96-hour forecast difference between GFS minus ARcnn for (c) Bias, (d) Pearson correlation. Calculated for locations when IVT forecast is under 250 kg m^{-1} s^{-1}. All dates from October 2017 - April 2018 testing dataset. In all plots (a, b, c, d), cool colors imply that the CNN post processing is improving the forecast.

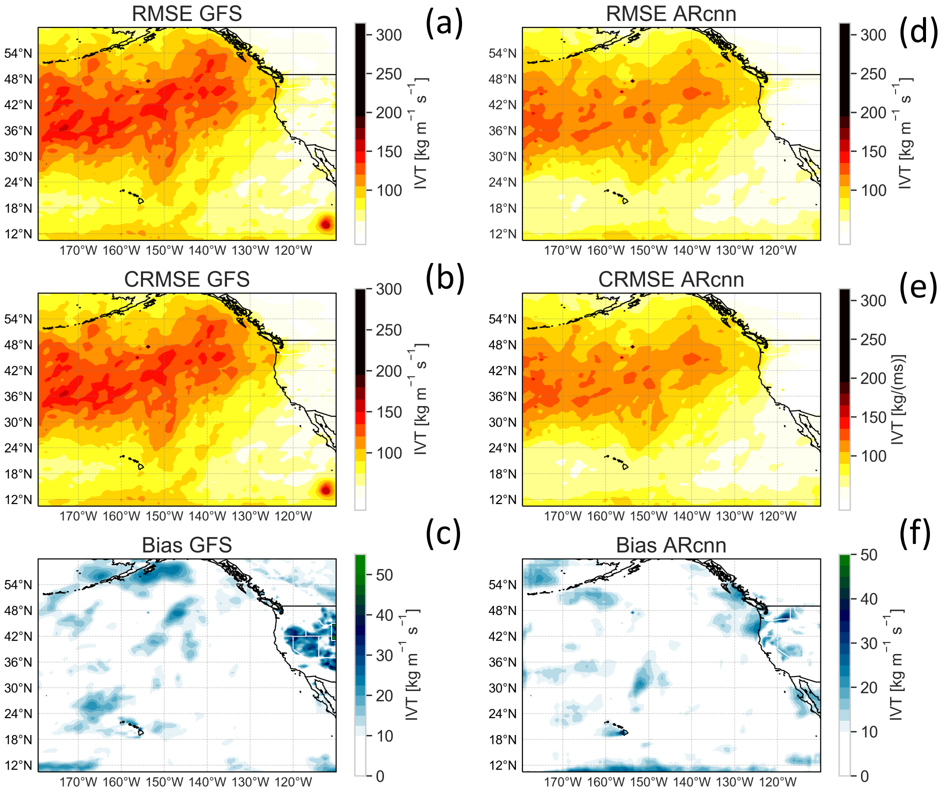

Figure 2.5S. Background Error Fields of GFS vs AARcnn respectively in the testing data set. **(a, d)** IVT RMSE. **(b, e)** IVT CRMSE & **(c, f)** IVT Bias.

Table 2.1S. Training, Validation and Testing Datasets.

	Training	**Validation**	**Testing**
Period	Oct ’08 – April ‘16	Oct ’16 – April ‘17	Oct ’17 – April ‘18
Samples Total	~ 44000	~ 4950	~ 4950
Samples Per Forecast lead	~ 4000	~ 450	~ 450
AR Field Occurrence	~ 76%	~ 76%	~ 73%

Table 2.2S. ARcnn: Hyperparameters & Architecture.

Description	Value
Number of Hidden Layers	9
Kernels Per Hidden Layer	layers (1-9): 64,32,16,16,16,16,32,64,1
Kernel Size	layers (1-9): [3 x 3]
Kernel Dilation	layers (1-9): 1,1,2,4,8,16,1,1,1
Learning Rate	0.001 – 5e-5
Batch Size	20
Convolution Padding	Same
Stride	[1,1]

Chapter 3

Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning

Abstract

Deep Learning (DL) post-processing methods are examined to obtain reliable and accurate probabilistic forecasts from single-member numerical weather predictions of integrated vapor transport (IVT). Using a 34-year reforecast, based on the Center for Western Weather and Water Extremes West-WRF mesoscale model of North American West Coast IVT, the dynamically/statistically derived 0-120 hour probabilistic forecasts for IVT under atmospheric river (AR) conditions are tested. These predictions are compared to the Global Ensemble Forecast System (GEFS) dynamic model and the GEFS calibrated with a neural network. Additionally, the DL methods are tested against an established, but more rigid, statistical-dynamical ensemble method (the Analog Ensemble). The findings show, using continuous ranked probability skill score and Brier skill score as verification metrics, that the DL methods compete with or outperform the calibrated GEFS system at lead times from 0-48 hours and again from 72-120 hours for AR vapor transport events. Additionally, the DL methods generate reliable and skillful probabilistic forecasts. The implications of varying the length of the training dataset are examined and the results show that the DL methods learn relatively quickly and ~10 years of hindcast data are required to compete with the GEFS ensemble.

3.1 Overview

Deterministic numerical weather prediction (NWP) systems are momentous forecast tools but are fatedly flawed in that they represent a single plausible realization of a possible weather future. Due to initial condition uncertainty, NWP deficiencies (e.g., sub-grid parameterization approximations), and nonlinear error growth associated with the chaotic nature of the atmosphere, initially small forecast errors eventually result in weather predictions which are as skillful as random forecasts (Lorenz 1963). Dynamic ensembles prediction systems (EPS) are utilized to represent the evolution of multiple likely weather trajectories. Though multiple methods for creating dynamic ensembles exist (e.g., Epstein 1969; Hacker et al. 2011; Kirtman et al. 2014); most modern EPS systems create ensembles by running many realizations of the atmospheric state evolution, initializing each ensemble with slightly varied starting conditions or using varied model physics (e.g., Toth & Kalnay 1993). A range of possible weather scenarios results, providing probabilistic bounds for future weather.

Ensemble systems have greatly advanced in the modern era, yet raw EPS forecasts still suffer from significant systematic model bias that must be corrected with statistical post-processing methods (Hemri et al. 2014). Often the systematic bias is particularly projected into the spread of the ensemble members, and under/over dispersive forecasts are common. This leads to a low correlation between the raw ensemble uncertainty and the forecast error; reducing the value of the model spread for forecast uncertainty quantification. Recent advances in deep learning (DL) and machine learning (ML) techniques have provided a significant step forward in calibrating statistical ensembles (e.g., Rasp and Lerch 2018).

The current study investigates ML's algorithmic ability to provide uncertainty quantification from single-member NWP model realizations, providing a valuable probabilistic measure of uncertainty, at a significantly lower real-time computational cost. This study leverages the methods developed in Rasp and Lerch (2018) (henceforth; RL2018) for ensemble calibration, but tailors them for the generation of probabilistic predictions from a historical set of single-member deterministic forecasts. Additionally, this study adds further algorithmic spatial awareness through vision-based DL methods (convolutional neural networks; CNN). Recently, there has been a surge of interest in DL-based NWP post-processing systems (see, Haupt et al. 2021 & Vannitsem et al. 2021). Similar to RL2018, Ghazvinian et al. (2021), developed a NN-based scheme that minimizes the continuous ranked probability score (CRPS) from a prescribed parametric forecast distribution (censored, shifted gamma) for rainfall prediction. Additionally, more flexible, distribution-free methods, have also been developed which leverage quantile-based probabilities transformed to a full predictive distribution (Scheuerer et al. 2020) or create direct approximations of the quantile function via regression based on Bernstein polynomials (Bremnes 2020).

Traditional ensemble model output statistics (EMOS) post-processing schemes fit parameters of prescribed distributions (Gneiting et al. 2005). Here we retain the parametric distribution prediction framework but leverage multiple NN architectures to statistically link the CRPS loss function to the NWP system and train the networks through stochastic gradient descent. NN's offer some ready advantages over more established EMOS methods. For example, EMOS is rigid with respect to feature selection and requires explicit prescription of predictor-predictand relationships in their implementation. Alternatively, NN's offer extreme flexibility in incorporating and ingesting ancillary weather variables as predictors. NN's can quickly encode spatial information through convolution (e.g., Chapman et al. 2019), and temporal information with recurrent NN's or attention-network systems (e.g., Li et al. 2020; Theocharides et al. 2020). NN's allow the post-processing system to readily encode predictor-

predictor variable interactions and capture nonlinear variable interaction (Nielsen, 2015). Additionally, Modern DL training schemes (i.e. dropout, regularization, early training stopping) have been implemented which systematically prevent algorithmic overfitting (Krogh & Hertz 1992; Srivastava et al. 2014).

Though many prominent probabilistic ensemble regression calibration methods exist (e.g., Gneiting et al. 2005; Raftery et al. 2005; Scheuerer & Hamill 2015) most leverage ensemble mean and spread characteristics rather than single-member deterministic models. Still some established post-processing methods operate solely on deterministic fields, or can be adapted to operate on deterministic hindcasts (e.g., Lerch & Thorarinsdottir 2013; Robertson et al. 2013; Scheuerer & Hamill 2015; D. S. Wilks 2009; Wu et al. 2011), though most of these methods have been tested with the mean of a dynamic ensemble.

Analog-based techniques, in which historical stores of similar forecasts are used to estimate uncertainty, have been similarly formulated to provide statistically developed uncertainty in forecasts starting from a dynamical ensemble (e.g., Hamill & Whitaker 2006), or from single-member deterministic predictions (Delle Monache et al. 2013). Here we use the latter approach, the analog ensemble, modified for optimal rare event prediction (Alessandrini et al. 2019) as a state-of-the-art baseline to assess the DL methods.

For this study, a newly developed 34-year deterministic hindcast is leveraged. This long training dataset provides near unprecedented opportunity to correct for systematic forecast error. High impact, 0-5 day (at 6-hour intervals) probabilistic integrated vapor transport (IVT) prediction for landfalling North American West Coast (NAWC) atmospheric river (AR) events is the focus. Vertically integrated IVT is the characteristic metric which defines the strength of an AR (Ralph et al., 2018). IVT is a combined thermodynamic and momentum metric which integrates specific humidity and zonal and meridional components of the wind from 1,000 to 300 hPa. Though we train our post-processing systems on every forecasted value, the study focuses on verifying IVT events above 250 [kg m^{-1} s^{-1}] (~85th percentile of observed IVT), as events below this threshold rarely result in extreme precipitation and are thus less societally impactful.

This study aims to test computationally efficient and flexible DL methods to estimate forecast uncertainty from a single-member NWP system for probabilistic, AR associated, IVT prediction. Uncertainty quantification is explored with DL methods by leveraging a distributional regression framework -- which aims to develop the conditional distribution of the weather given a deterministic set of variables. The DL methods train by optimizing CRPS - a mathematically principled loss function for probabilistic forecasts (Camporeale & Carè, 2021; Gneiting et al., 2005; Matheson & Winkler, 1976). We pit the developed statistical uncertainty methods against modern state-of-the-art dynamic ensembles (calibrated and not) to test their skill. Additionally, we use feature permutation (McGovern et al. 2019) to explore the variable importance in the NN-based systems. Finally, we test the length of training data required to develop skillful forecasts, in order to determine the length of hindcast required to train a prediction system.

The remainder of the paper is structured as follows. Section 3.2 presents the dynamic forecast systems, the statistical ensemble and dynamic ensemble post-processing methods used, and the ground truth data. Section 3.3 discusses the resulting forecast skill, examines input variable importance, and explores the required length of training data to quantify uncertainty reliably. A discussion of possible extensions follows, and we present conclusions in Section 3.4.

3.2 Data, Methods, and Metrics

3.2.1 Region of Interest

Forecasting land-falling AR events over the NAWC is crucial (Ralph et al. 2020; Wilson et al. 2020), and several contributions on AR forecast skill assessment are present in the literature (DeFlorio et al. 2018; DeHaan et al. 2021; Nardi et al. 2018; Nayak et al. 2014; Wick et al. 2013). ARs bring valuable precipitation to this drought prone region (Fish et al., 2019; Lamjiri et al., 2017) while simultaneously being the dominant driver of flooding across the NAWC (Corringham et al. 2019; Ralph et al. 2020). One forecast product is a series of ensemble-based forecast imagery that shows a forecast lead time-latitude framework spanning the west coast of North America with illustrated IVT data from the NCEP-GEFS ensemble, known as the AR landfall tool (ARLT, Cordeira et al. 2017; Cordeira & Ralph 2021) on the Center for Western Weather and Water Extremes (CW3E) web portal. ARLT shows NCEP-GEFS data in a pseudo-Hovmöller coastline-spanning framework, illustrating IVT data and providing situational awareness of the likelihood, intensity, duration, and timing of possible landfalling ARs. Due to the importance of forecasting landfalling ARs, this study examines the probabilistic forecast accuracy of landfalling grid-points in every examined/post-processed forecast system, and the two adjacent (moving westward) oceanic model points. Figure 3.1 shows the examined points in this study. All verification metrics and forecast assessment henceforth are diagnosed at these 144 landfalling locations.

3.2.2 Ground Truth

IVT from the National Aeronautics and Space Administration's Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) reanalysis is used as ground truth to diagnose forecast error and in ML training. MERRA-2 provides regularly gridded observations of the global atmosphere with assimilated satellite, upper air, remote sensing, and surface data. The MERRA-2 product is resolved on a 0.5° latitude x 0.625° longitude grid and interpolated to 21-pressure levels between 1,000 and 300 hPa for IVT calculation (Gelaro et al. 2017; McCarty et al. 2016). Every forecast field in this study is regridded to the MERRA-2 grid using a 1st and 2nd order conservative remapping scheme (Schulzweida et al. 2006) prior to ML training.

3.2.3 Dynamic Model For Deterministic Forecast

Uncertainty quantification is generated for a version of the Weather Research and Forecast model that has been tuned specifically to Western U.S. extreme precipitation (West-WRF, Martin et al. 2018). West-WRF is a near real-time model developed at CW3E that was run retrospectively to generate a 34-year (1984-2019) hindcast spanning December through March of each year. In addition to providing a long training data set, the model's consistency with the operational version provides an unprecedented opportunity for training machine learning models on historical forecasts. The model is operationally run at a 9 km resolution, but we use 1st and 2nd order conservative remapping to regrid this data to the common MERRA-2 grid, the model domain spans 25°N to 60°N and 150°W to 115°W. In this study, the December – March season is referred to as a water year (WY) with the year specified as the March of that year. For example, December 2018 – March 2019 is referred to as WY2019. West-WRF is evaluated in the last three years of the data set (WY2017, WY2018, WY2019). 3-fold cross-validation is leveraged in which the previous year is used as validation data and each of the three evaluated WYs is held out as testing data.

3.2.4 Machine Learning Generated Forecast Uncertainty

Four ML methods for uncertainty quantification are evaluated and compared against a dynamical ensemble's raw model output and a dynamical ensemble calibrated with a neural network. The computational cost of developing ML-based probabilistic predictions compared with dynamical ensembles is significantly less, both in real-time forecasting and for hindcast generation. Each method is described below. For each post-processing system, the inputs are described in Table 3.1. Multiple deep learning (DL) models are trained, with their architecture shown in Table 3.2, and model architecture diagrams shown in supplemental Figure 3.1S. An extensive, though not exhaustive, hyperparameter search was conducted on two forecast lead times (48 hours and 96 hours) to select model parameters by minimizing the model loss (described below) on the validation dataset. To aide future DL post-processing development, the hyperparameter search methodology and model architecture intuition is described in the supplemental material.

3.2.4a Neural Network with Location Embeddings

Here the neural network (NN) is described, with a focus on the architecture, hyperparameters, and training routine utilized in this study. For a more complete exploration of the topic of NN's, the reader is referred to Nielsen (2015). DL functionality is developed in python using the Tensorflow 2.0 (Abadi et al. 2016) library with the embedded Keras (Chollet et al. 2018) implementation. Following Chapman et al. (2019), an independent NN is trained for every forecast lead time. The input for this method is described in Table 3.1, the output is the mean and standard deviation of a probability distribution representing a probabilistic forecast for IVT at a specified (to the NN) coastal location.

Neural networks (NN) approximate nonlinear functions and processes (Nielsen 2015) through a series of feed forward matrix operations. NN’s pass input predictor variables through a succession of “hidden” layers, resulting in a specified output layer. Each layer is described by the number of nodal points in that layer with the initial layer being the number of input variables. Prior to input, each predictor variable is standardized using the global (every point in the examined domain) mean and standard deviation. In this work, a simple model with 2 hidden layers containing 30 and 40 nodes, respectively, is used. Nodes from adjacent layers are connected via model weights. The hidden nodal point values are determined by the sum of the product of associated model weights and the input values from the previous layer. Each nodal point is then ‘activated’ by a nonlinear function before passing the variables to the following layer. We use a Rectified Linear Unit (ReLU) activation function (Nair & Hinton 2010). The task of training a NN is to learn the optimal nodal weights, computed iteratively through backward optimization and gradient descent. In particular, each iteration seeks to minimize the cost of a specified loss function, by determining the gradient field of the weights and taking a small step in the direction opposite to this gradient. The NN leverages an Adam optimizer (Kingma et al. 2014) with a 0.005 training step that reduces by 10% on a validation plateau of 5 epochs (learning cycles). After 8 epochs of no decrease in validation error, training is ended. This typically resulted in ~40 training epochs.

The output model parameters (μ_{IVT}, σ_{IVT}) are estimated by minimizing the prescribed loss function of the continuous ranked probability score (CRPS) of a Gaussian distribution truncated at zero (as the magnitude of IVT cannot be negative). This loss function has been used in several notable EMOS post-processing studies, largely in applications of wind speed prediction (e.g., Baran & Lerch 2015; Thorarinsdottir & Gneiting 2010; Thorarinsdottir & Johnson 2012). CRPS is discussed in the *Results* section and the analytical expression of the CRPS Gaussian distribution is provided in the appendix material. Multiple CRPS loss functional families were attempted (logistic, log normal, gamma, and Gaussian) (see Jordan et al. (2019) for these formulations), and a Gaussian truncated at zero provided the best fit, as determined by evaluation of the threshold-weighted CRPS and the shape of the stratified rank histogram, evaluated on the validation dataset -- motivating the final loss function choice.

Additionally, location embeddings are used as an input to the NN. Embeddings are responsible for encoding a vectorized version of discrete information, in this case, an ID number specified for each of the 144 locations (1-144). These vectors are learned and updated during training, but do not correspond to any real variable. This allows the network to learn customized nodal weights for each lat/lon location while still benefitting from the relationships learned at every location (Guo & Berkhahn 2016; RL2018). The vector length is specified as part of the network architecture. By conducting a hyperparameter search, it was determined that 2 latent variables (vector length) provided the greatest model performance without adding additional model parameters. Thus, one NN can be trained for the entire domain and the bias specific to each location (e.g., topographically or latitudinally driven NWP biases (Gowan et al. 2018)) can be corrected.

Due to the data set spanning multiple decades, we have noticed oscillations in NWP model skill and bias. To mitigate the potential effects of secular climate change, or slowly varying decadal variability, we institute a customized training regime similar to model transfer learning (Torrey & Shavlik 2010). The model is first trained on the full 32-year training set (34 years, minus 1 year of validation and 1 year of test). Next those model weights are saved and frozen from updating, a final layer is concatenated to the network and we “fine-tune” on just 1 WY, two years prior to the testing data set (one year prior to the validation). For example, a NN that is tested on WY2019 is initially trained on WY1985-WY2017, then frozen, a new layer is concatenated, and it is then tuned on WY2017. The training schedule is exactly similar to that described above with identical criterion for ending the training. While the mean prediction is relatively unaffected by fine-tuning, this was found to significantly improve predicted spread statistics (not shown).

3.2.4b Convolutional Neural Networks

The convolutional neural network (CNN) architecture is shown in Table 3.2 and Figure 3.1s. The architecture is adapted from a U-NET (Long et al. 2015). The U-NET architecture is ideal for this task as it passes less abstracted information from shallow layers in the CNN to deep layers [see Ronneberger et al. (2015) for more detail]. Additionally, versions of this architecture have been shown to significantly reduce IVT deterministic forecast error (Chapman et al. 2019). The computational details are similar to those described in the above NN. Again, a Gaussian distribution truncated at zero provided the best skill on our cross-validated dataset and was selected as the loss function. The CNN utilizes an Adam optimizer (Kingma et al. 2014) with a 0.0001 training step that reduces by 40% on a validation plateau of 2 epochs. After 8 epochs of no decrease in the validation error, training is ended. This typically resulted in ~50 training epochs. The CNN uses identical predictors to the NN, the largest difference being that CNNs operate on images by updating weights associated with convolutional kernels which are slid across input image fields and trained to highlight salient forecast features. In the CNN, the entire spatial domain is fed to the model at training for each independent forecast rather than independent training data for each location (as in the NN). The goal of model training is thus to learn the optimal weights in the convolutional kernels which minimize CRPS by best predicting μ_{IVT} and σ_{IVT} for every pixel in the image. The network is trained to optimize predictions in the entire model domain, however, in the following analysis, the CNN is evaluated only at the aforementioned coastal locations. The implications of this choice are discussed in section 4. The reader is referred to Zhang et al., (2021) for a theoretical description of CNN's and convolutional kernel training.

The same training regimen is utilized for the CNN with 1 additional convolutional layer concatenated to the end of the network after freezing all previous layers. The CNN is then fine-tuned on the WY two years previous to the testing WY.

3.2.4c Fully Connected Distributional Regression

We include, as an additional baseline, a parametric prediction method that is conceptually similar to traditional distributional regression performed via EMOS systems. We implement a fully connected neural network (FCN) with no hidden layers, trained using CRPS estimated from a Gaussian distribution truncated at zero. The FCN, without inputting ancillary predictor fields, is conceptually equivalent to a global EMOS scheme, but differs in the parameter estimation approach. Here, as it is easily implemented and also demonstrated improvement in minimizing CRPS, we include all of the predictor variables that are supplied to the NN and CNN, and the same location embedding vector supplied only to the NN (see Table 3.2). The FCN leverages an Adam optimizer (Kingma et al. 2014) with a 0.005 training step that reduces by 10% on a validation plateau of 5 epochs (learning cycles). After 8 epochs of no decrease in validation CRPS, training is ended. In order to create as similar training conditions to the CNN and NN we fine-tune the FCN system by loading the model weights from the model trained on the 34 years of data, reducing the learning rate to 0.00001 and training again on data from 2 years prior to the testing data (thus, 1 years prior to the validation data). The FCN serves to assess the value of the nonlinear predictor-predictand relationships in both the NN and CNN. A local FCN implementation was also tested, but showed poorer forecast performance and calibration for high threshold IVT events ([250, 350, 500] kg m^{-1} s^{-1}).

3.2.4d The Analog Ensemble for Rare Events

To compare the DL-based statistical ensemble to a state-of-the-art ML-based ensemble method, an analog ensemble (AnEn; Delle Monache et al. 2013) coupled with a recent bias correction innovation for rare events (Alessandrini et al. 2019) is constructed. The AnEn generates an ensemble by exploiting an issued NWP forecast, a history of forecasts made by the same model, and the corresponding resultant observed weather. When a forecast is issued, the AnEn is tasked with searching for analogous forecasts in the historical record, it then uses the corresponding resultant observations for the analogous forecasts as the ensemble prediction.

Let $f\left(y\middle|\boldsymbol{x^f}\right)$ be the probability distribution of the observed value y of some predicted quantity given a model prediction $\boldsymbol{x^f}$. $\boldsymbol{x^f}$ is a vector of k predictor variables issued from the NWP forecast ($\boldsymbol{x^f} = x_1^f, x_2^f, x_3^f, \dots, x_k^f$) which includes the desired forecast variable (IVT) and a suite of other relevant predictor variables (Table 3.1). AnEn uses a distance function to then identify the closest analogs to $\boldsymbol{x^f}$ from a database of previously issued forecasts ($\boldsymbol{x^j}$). The ground truth observations y_j from the previously issued forecasts form the ensemble. Like the NN, the analog ensemble is a point-based method, and only forecasts at a given location are used to form this ensemble. The distance function is given by $d\left(\boldsymbol{x^f}, \boldsymbol{x^j}\right) = \sum_{i=1}^{k} \frac{w_i}{\sigma_i} \sqrt{\sum_{r=-t}^{\tilde{t}} \left(x_{i,t+r}^{f} - x_{i,t'+r}^{j}\right)^2}$ where the current NWP forecast ($\boldsymbol{x^f}$) is valid at time t at a given forecast location. $\boldsymbol{x^j}$ is the analog at the same location with the same forecast lead time but valid at a past time (t'), k defines the number of predictor variables weighted by w_i. σ_i is the standard deviation of the time series of past forecasts of a given variable at the same location and forecast lead. $\tilde{t}$ is equal to half the number of additional times over which the metric is computed. Accounting for a forecast window ($\tilde{t}$) ensures that the trend of the examined variables is considered and has been shown to be valuable for minimizing forecast error (Alessandrini et al., 2015). This temporal trend gives the AnEn a potential predictor advantage over the CNN and NN.

The AnEn is run and optimized (through w_i selection) at every location individually. By leveraging small predictor sets (k), a full brute-force optimization search can be conducted by trying every permutation of w_i, and examining mean squared error on an independent validation data set subject to $\sum_{i=1}^{k} w_i = 1$, where $w_i \in [0, .1, .2, \ldots, 1]$. Which has been shown to improve predictions in several past studies (e.g., Alessandrini et al., 2015; Junk et al., 2015). Supplemental figure 3.3S shows the distribution of predictor variable weights across every location. The predictor variable of interest (IVT) is dominantly weighted, as expected, with the remaining variables accounting for ~10% of variability each. To match the dynamic ensemble, we specify the return of 21 ensemble members. The numbers of ensembles were varied, but the results showed little sensitivity between 10 – 50 members.

The AnEn has a tendency to introduce a conditional negative bias when predicting events in the right tail of the forecast (extreme or rare events), which are the focus of this study.

To ensure that the AnEn is optimized to correctly forecast rare events (as our target is to most accurately forecast ARs), and to set the best baseline possible for the DL/ML methods, we leverage the modifications to the AnEn as presented in Alessandrini et al. (2019) for conditional bias correction. The proposed method is based on a linear regression analysis between forecast and observations performed independently at each lead time and location. Each member is adjusted by adding a factor proportional to the difference between the target forecast and the mean of the past analog forecasts multiplied by the coefficient obtained after the linear regression analysis. We refer the reader to Alessandrini et al. (2019) to examine additional details of the bias adjustment algorithm. A threshold value of 300 [kg m^{-1} s^{-1}] units of IVT (~90th percentile of station observations) is used to enact the bias correct. This value was determined by incrementing the value of the threshold from 250 – 500 (in 50-unit increments) and minimizing the CRPS on the validation datasets.

3.2.4e Raw and Calibrated Global Ensemble Forecast System

We assess the probabilistic skill of the NN-based methods (FCN, NN, and CNN) against a state-of-the-art dynamical ensemble: the operational 0.5° latitude x 0.5° longitude National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) version 11.0.0 from December 1st to March 31st of WYs ending in 2017, 2018, and 2019. The GEFS includes 21 members (20 perturbed initial conditions, 1 control member). These data were obtained from The Interactive Grand Global Ensemble (TIGGE) data portal at the European Centre for Medium-Range Weather Forecasts. WY17 and WY18 contained 70 missed forecasts in the TIGGE system so these were then calculated from 1° latitude x 1° longitude GEFS data obtained from the National Centers for Environmental Information Data Archive and were simply interpolated to the 0.5° grid spacing. We additionally apply the same

NN post-processing to the GEFS ensemble system as described in RL2018, by minimizing CRPS while using the ensemble IVT mean and standard deviation as predictors and leveraging the embedded forecast location. This algorithm was shown to outperform the best traditional post-processing methods (RL2018). The NN applied to the GEFS system ($\mathrm{GEFS_{nn}}$ henceforth) is subject to the same train/test split as described above in which the network is trained in a 3-fold cross-validation manner in which the previous year is used as validation data and each of the three evaluated WYs are held out as testing data. Table 3.1 describes the input variables. Table 3.2 describes the utilized network.

3.3 Results

In this section, we evaluate the predictive performance of the post-processing systems and raw dynamic ensemble, all based on the cross-validated testing data from WY17, WY18, WY19. For an introduction to the evaluation methods and underlying theory, see the Appendix. We use skill scores ($\mathrm{SS} = 1 - \frac{S}{S_{ref}}$, $-\infty < SS \leq 1$,) where positive/negative values are shown to be more/less skillful than the reference forecast (S_{ref}). Python code for reproducing the results and models is available online (github url).

This analysis evaluates 6 forecast systems, termed: AnEn, FCN, CNN, NN, GEFS, and $\mathrm{GEFS_{nn}}$, evaluated from 0 to 120 hours (in 6-hour intervals). The AnEn, FCN, NN and CNN systems are built from an historical data set including a single deterministic forecast (based on the dynamical model West-WRF), while the GEFS is built from the raw GEFS EPS forecast. Additionally, the results focus on high impact IVT events which are likely to cause NAWC precipitation. We threshold at 250, 350, and 500 [kg $\mathrm{m^{-1}}$ $\mathrm{s^{-1}}$] units of IVT, which guarantees local AR conditions. This represents percentile values of ~85th, ~93rd, ~97th, respectively. Right tailed events are traditionally more difficult for post-processing methods to improve upon. Though higher impact events exist (500+ threshold), their rarity prevents robust probabilistic statistical comparisons (Wilks 2010), thus the above stated thresholds are evaluated. All results shown henceforth are for the independent testing data years (WY2017, WY2018, WY2019).

The GEFS and AnEn consist in a set of ensemble members while the FCN, NN, CNN, and GEFS$_{nn}$ include the mean and standard deviation of a truncated Gaussian distribution. To ensure a fair assessment, the following verification is conducted by computing the mean and standard deviation for every individual forecast and randomly sampling from that distribution to create pseudo-ensembles. The exact ensembles for the GEFS and AnEn were also assessed, but the resulting analysis was not significantly changed.

3.3.1 Deterministic Predictions

Figure 3.2 shows the root mean squared error (RMSE) (a) and Pearson correlation (PC) (b) of the deterministic forecasts (ensemble mean) using the West-WRF raw reforecast as the reference forecast in 12-hour increments. Though the primary focus of this work is to evaluate the probabilistic skill of the ensemble forecast methods, we first demonstrate the deterministic skill of the forecast systems. For each method, this is taken as the mean of the predictive distribution/ensemble. The authors realize that this is not a direct comparison because the post-processing methods were not applied to the same forecast baseline (i.e., West-WRF vs. GEFS).

At all lead times, the GEFS ensemble mean (red) forecast is more skillful than the West-WRF deterministic model from which the AnEn, FCN, NN, and CNN are developed. The GEFS$_{nn}$ (red) and Raw GEFS (light red) systems ensemble mean performance differences are not statistically significant and the GEFS_{nn} ensemble calibration is largely just influencing the ensemble spread statistics (discussed further below) to improve the forecast skill. The CNN (white) is resulting in the largest improvements of West-WRF reforecast when compared to AnEn, FCN, and NN at all lead times in both PC and RMSE, with a stable improvement of ~10% at every lead time for RMSE while improving the correlation from 1-7% with greater improvements at the longer lead times. The NN (light blue) also improves the forecast at every lead time at ~5% for RMSE and improves correlation from 0-3% across lead times. The NN generally outperforms the FCN, showing the value of the adding nonlinear activations, with statistically significant improvement at lead times 0, 12, 36, 48, 72, and 96 hours for RMSE and at every lead time past 12 hours for PC. The NN is run locally with embedded location ID information, and therefore does not have the benefit of a global field view (like the CNN), this additional spatial feature helps to quantify the difference in mean statistics. The CNN corrections result in forecasts that significantly outperform those from the FCN and NN at every lead time for both PC and RMSE metrics. The AnEn (blue) initially negatively impacts the analysis forecast (F000) but improves the skill of the deterministic forecast from 24 hours – 120 hours with similar statistics as the NN.

3.3.2 Diebold-Mariano Test (Under AR Conditions)

For comparative model assessment, proper scoring rules are leveraged to simultaneously evaluate the calibration and sharpness of forecasts (Gneiting & Raftery, 2007). Proper scoring rules assign a numerical score to pairs of probabilistic forecasts and observations such that the expected score is optimized if the true distribution of the observation is issued as a forecast. Here, two negatively oriented (a smaller value is better) proper scoring metrics are examined, the Brier skill (BS, Brier 1950) and twCRPS (Gneiting & Ranjan 2011).

Figure 3.3 shows the results of the two-sided Diebold-Mariano (DM, Diebold & Mariano 2002) test calculated based on mean threshold weighted CRPS (twCRPS, Gneiting & Ranjan 2011) over all the samples at each lead time (0-120 hours) as the determining metric, with a threshold set to 250 [kg m^{-1} s^{-1}] units of IVT. Simultaneous interpretation of the test results across lead times requires that we account for test multiplicity. We do so by controlling the false discovery rate at $\alpha_{FDR} = 0.05$ (see Appendix for details) (Benjamini & Hochberg 1995; D. Wilks 2016). Each panel (a – e) leverages a separate reference forecast (AnEn, FCN, NN, CNN, GEFS_{nn}, and GEFS, respectively) to compare to each other forecast. The reference is shown in bold, and a gray dash-dot line is used to delineate the reference further. Red panels indicate that the reference performs underperforms the compared forecast, blue panels indicate that the reference forecast outperforms compared forecast, and white panels show that the difference is not statistically significant between the two systems. Panels with large blue swaths are better forecast systems, than the compared post-processing system.

It is apparent that when comparing individual forecast systems built from West-WRF (AnEn, FCN, NN, CNN, Fig. 3.3a, 3.3b, 3.3c, 3.3d, respectively) the forecasts from the CNN generally outperform the other systems. Forecasts from the CNN significantly outperform forecasts from the AnEn at all lead-times except from 84-96 hours in which the differences are not statistically significant. The forecasts from the CNN significantly outperform the GEFS ensemble system or the differences are statistically significant not at all lead times except the 66-hour lead forecasts. Additionally, at 0-48 and 96-120 hours the CNN is competitive with the calibrated GEFS forecast (GEFS_{nn}). The GEFS_{nn} systematically significantly outperforms the

GEFS system for all short lead times (0-60 hours) and outperforms or not significantly different from GEFS from 72-120 hours. We show similar figures for the 350 and 500 [kg m^{-1} s^{-1}] IVT thresholds in supplemental Figures 3.4S and 3.5S. While still generally outperforming each forecast system from a Brier skill score (BSS, Brier, 1950) and twCRPS perspective, the CNN struggles more to improve over the GEFS with high impact events at the longer lead times (3-5 days), this is discussed further in the discussion section. The NN is shown to significantly outperform the FCN for the 250 [kg m^{-1} s^{-1}] threshold at all lead times and is generally more skillful (though not always significantly) at the 350 and 500 [kg m^{-1} s^{-1}] thresholds (Figures 3.4S and 5S).

3.3.3 Brier skill score and CRPS

Figure 3.4 shows the Brier skill score (BSS) at three threshold levels (250, 350, 500, [kg m^{-1} s^{-1}]) of IVT for forecasts from 0 – 120 hours using the GEFS forecast BS as a reference metric. The $GEFS_{nn}$ may leave the ensemble mean forecast relatively unaffected (Fig. 3.2), but it improves the GEFS forecast by calibrating its probabilistic skill (Fig. 3.4, dark red). Within the first 36 hours the CNN outperforms or shows insignificant differences from the $GEFS_{nn}$ forecast system for every threshold value (Fig. 3.4 white vs. red). Between 96 and 120 hours the NN-based ensembles again compete with or outperform the GEFS forecast systems. The NN is able to outperform the AnEn at most lead times out to 48 hours and the two methods have similar performance from 60 to 120 hours. At lead times between 96 and 120 hours and higher impact events the AnEn and NN show similar skill to the CNN (Fig. 3.4b and 3.4c). Though the differences are not always statistically significant, the NN generally outperforms the FCN at most lead times. To complement Figure 3.3 and Figure 3.4, Figure 3.2S shows the twCRPS skill score at the same three threshold levels (250, 350, 500, [kg m^{-1} s^{-1}]) of IVT for forecasts from 0 – 120 hours using the GEFS forecast as a baseline metric. The twCRPS tells a very similar story to the BSS and DM test.

In both BSS and twCRPSS, between 60 and 84 hours, we note a drop in skill between the AnEn or the NN-derived ensembles which are built on a deterministic prediction and the dynamic ensemble system. Figure 3.2 shows that this skill is largely derived from a comparative discrepancy in deterministic forecast skill between the two forecast systems used to build these ensembles (GEFS and West-WRF). Again, we stress that each method is built from different dynamic forecast models and this does not represent a detriment added by the post-processing methods (see Fig. 3.2). It appears that this comparatively larger forecast skill difference (see Fig. 3.2 hours 0-48 vs hours 60-84) is responsible for the difference of skill in the interim forecast window.

3.3.4 Spread/Skill

Figure 3.5 shows the binned spread-skill plots of the evaluated models partitioned into the 0-48 hours and 54-120 hours for forecasts of IVT. In the first 48 hours (Fig. 3.5a-f) the GEFS model (light red) is severely overconfident (Fig. 3.5a). The AnEn faces the opposite problem and appears to overestimate values of forecast uncertainty (Fig. 3.5c). The remaining models (CNN, NN, $GEFS_{nn}$) provide statistically consistent forecasts and indicate that they are able to capture the flow-dependent forecast uncertainty because their spread dependably reflects the forecast error variance. The CNN and the NN are virtually indistinguishable and perfectly calibrated while the $GEFS_{nn}$ does reflect small conditional bias towards the highest binned events. Across all tested models, forecasts from 54-120 hours (Fig. 3.5g-l) are less calibrated, but still represent a good flow-dependent forecast uncertainty relationship. The GEFS, FCN, NN and CNN forecasts are overconfident and contain a slight low bias. The AnEn is the best calibrated forecast for the right tailed forecast error events, followed closely by the NN and GEFS$_{nn}$ forecast systems, showing that these systems capture the flow-dependent forecast uncertainty since the spread dependably reflects the forecast error variance.

3.3.5 Stratified Rank Histogram

Figure 3.6 shows the stratified rank histograms of the evaluated models partitioned into the 0-48 hours and 54-120 hours. The histograms are stratified into three categories: [250-350), [350-500), [500+] [kg m^{-1} s^{-1}]. For the analog ensemble and GEFS system we use the 21 ensemble members generated by each system. To be consistent we sample 21 random pulls from the distribution described by each individual forecast to form a pseudo ensemble and build the stratified rank histogram from those forecasts. Bröcker & Smith, (2007) and Siegert et al., (2012) demonstrated that when stratifying on the ensemble forecast mean (or other ensemble derived statistics), a uniform rank histogram distribution is not necessitated to show a calibrated forecast ensemble system. Bellier et al., (2017) offered a graphical test to check the true calibration shape through random sampling of ensemble members which serve as pseudo-observations to determine the shape of a perfectly calibrated forecast ensemble. After conducting this test for the prescribed IVT thresholds [250-350), [350-500), [500+] [kg m^{-1} s^{-1}], it was determined that a uniform stratified distribution is optimal (not shown). To aid in interpretation, Table 3.3 shows the reliability index (RI, Delle Monache et al. 2006) for the stratified rank histograms. Here, $RI = \sum_{i=1}^{K+1} \left| f_i - \frac{1}{K+1} \right|$, where f_i is the frequency of observations in the ith rank and K is the number of forecasted ensembles.

We first examine the 0-48 hour forecasts. The most apparent error is in the GEFS forecast ensemble system which is highly under-dispersive/overconfident, and a general lack of statistical consistency. Applying a neural network with location embeddings to this dynamic ensemble ($GEFS_{nn}$) results in a very well calibrated forecast for AR events (Fig. 3.6b). This confirms that $GEFS_{nn}$ is largely correcting the forecast spread while leaving the ensemble mean relatively unchanged (See Fig. 3.2). The AnEn is over-dispersive/ underconfident from 0-48 hours. Despite developing all of the spread characteristics from data alone (unlike the $GEFS_{nn}$), the NN and the CNN (Fig. 3.6d and 3.6e) both represent well calibrated probabilistic distributions. There is a small indication of under prediction for both of these systems, exacerbated further in the CNN. The FCN struggles to calibrate the right tailed events, showing a high bias. This demonstrates the important nonlinear information in the predictor fields as the NN shows a very well calibrated ensemble, with the same input predictors. The RI values in table 3.2 indicate that the NN is largely more calibrated than the CNN system though all post-processing methods outperform the raw GEFS calibration.

The 54-120 hour forecasts struggle more with statistical consistency. The GEFS ensemble again shows signs of over dispersion, coupled with a high bias (Fig. 3.6g), but is less affected by the over confidence than at shorter lead times. The $GEFS_{nn}$ acts to fix the dispersion and produces a relatively calibrated ensemble though there are signs of a low bias. The NN-based methods and AnEn again produce fairly calibrated ensembles. The CNN struggles prominently with a low bias (Fig. 3.6j and table 3.2). The AnEn and NN are relatively indistinguishable and produce a well calibrated statistical ensemble. The FCN again shows a severe high bias and offers a good contrast to the NN and CNN methods.

3.3.6 Variable Importance

To investigate rankings of input variable importance in the FCN, NN, and CNN we use a single-pass permutation-based measure introduced by Breiman (2001). The goal is to determine the level of twCRPS deterioration when the statistical link between forecast field (F_j) and the target observation (y_j) is broken by randomly permuting each F_j, one at a time, over all forecast samples. We use the mean twCRPS($\tau = 250$ [kg m^{-1} s^{-1}]) of the non-permuted input features as a relative reference baseline. The reference twCRPS baseline is recalculated at each lead time in order to prevent skewing the variable importance via dependence on model lead-time forecast skill. If performance deteriorates significantly (high values in Fig. 3.7) the variable is considered important. The single-pass permutation algorithm is described in detail in the Appendix. Figure 3.7 shows the relative variable permutation importance at forecast lead times 12 (a), 24 (b), 48 (c), 72 (d) 96 (e), and 120 (f) hours for a twCRPS with threshold $\tau =$ 250 [kg m^{-1} s^{-1}] units of IVT.

The most important variable across the three systems, at all lead times, is the NWP model output IVT. IVT's importance, relative to other variables, diminishes at longer forecast horizons. The CNN considers integrated water vapor (IWV) as the second most important variable, accounting for model degradation of 6-10% across all forecast lead times. IWV is an integrated component of specific humidity calculated at the same model levels as IVT. Its relative variable importance indicates that the CNN is learning some error dependence which is contained solely in the thermodynamic component of IVT. The CNN shows minor dependence on the remaining forecast variables. We note that the CNN does not leverage location ID as a predictor (see Table 3.2). The CNN does not show much sensitivity to the meridional or zonal components of the 500 hPa wind, though these variables have been shown to impact forecast error state in other NWP systems (Stone et al., 2020).

The FCN and the NN leverage the same input predictors and the difference in variable importance is an indication of the important nonlinear predictor-predictand relationships learned by the NN. Generally, the addition of the non-linearity, spreads predictor sensitivity more evenly across multiple variables, leading to greater importance of several variables. The second most important variable for the FCN and NN, across most lead times, is the meridional component of the wind at 500 hPa (v500). This is intuitive as modulation to the v500 variable is leveraged to diagnose storm track variability (Chang & Yu 1999; Wirth et al. 2018) and is an indication of amplification in the synoptic scale control (via large-scale troughing or ridging) over the AR system. Meridional orientated ARs tend to be stronger in magnitude (higher IVT) and result in greater precipitation (Cobb et al. 2021; Hecht & Cordeira 2017). Interestingly, the FCN and NN systems both show a sensitivity to surface pressure, which is not learned in the CNN. The location ID (input via an embedding layer) accounts for a 1-2% model degradation across all lead times and is twice as important in the NN than the FCN, indicating some nonlinear dependence on spatially dependent information.

3.3.7 Length of Training

The NN-based methods and AnEn, run in real-time, have a significantly lower cost compared to the GEFS system as only a single deterministic forecast is required to produce this probabilistic prediction. However, a longer training data set was used compared to the stable GEFS system (GEFS model version 11.0.0 was only stable for 3 years). To test the impact of the length of training (deterministic hindcast years that are required) needed to achieve comparative skill we retrain the CNN, NN, and AnEn holding out forecast years one year at a time counting backwards in time. For example, the methods are trained solely on WY2016 and skill is determined on the testing dataset. Then the methods are trained on WY2015-WY2016, and so on until the entire dataset is utilized. Figure 3.8 shows forecasted 48-hour twCRPS ($\tau = 250[kgm^{-1}s^{-1}]$) of AnEn, CNN, and NN by the number of years in the training dataset using the GEFS$_{nn}$ as a reference forecast.

The NN trains well with a single year of data and plateaus in skill quickly afterward. As one NN is trained for the entire domain, the NN is able to learn the forecasted IVT error relationship from every point in the field domain, effectively multiplying the length of the training data by the number of points used (144, though these are not necessarily independent forecast). The AnEn learns most quickly within the first 10 years but continues to learn as the length of training data is extended. This can be explained by the fact that, as more similar analogs are added with each year and the AnEn is unable to extrapolate forecast information but must rely on the past forecast record (except for the right tail of the distribution when the bias correction for rare events is applied). Though we truncate the figure at 23 years, the AnEn continues to learn for the 34-year period (though marginally; not shown). The CNN is the worst performing forecast for the first 2 years and does not significantly outperform a simple NN until 9 years of data is utilized. The associated cost of producing a hindcast, if a CNN is desired is thus high (though again, only a deterministic hindcast is required). The CNN appears to plateau after 11 years of training and only very marginal skill is added in the remaining 20 years of training data.

3.4 Discussion and Conclusion

Integrated water vapor transport (IVT) is post-processed to derive a forecast uncertainty quantification. There has been a recent surge of interest and method development of Machine Learning (ML) and Deep Learning (DL) for numerical weather prediction (NWP) post-processing (Baran & Baran 2021; Kirkwood et al. 2021; McGovern et al. 2017; Meech et al. 2021; Schulz & Lerch 2021; Vannitsem et al. 2021). The examined DL methods (NN and CNN) are flexible and easy to implement with modern DL toolboxes. The ML methods can compete well with dynamic model ensemble due to the severe under-dispersion of the GEFS ensemble, and the DL ability to adjust the deterministic (mean) score to be competitive with the Global Ensemble Forecast System (GEFS) mean (Fig. 3.2). At lead times when the GEFS mean forecast skill significantly outperforms the deterministic dynamical forecast skill, the probabilistic methods have trouble competing (compare Fig. 3.2 and Fig. 3.4). The $GEFS_{nn}$ does not adjust the GEFS ensemble mean prediction, but simply calibrates the ensemble spread. A well calibrated dynamically generated ensemble would be more difficult to outperform.

During the first 48 hours the convolutional neural networks (CNN) developed from a long running deterministic forecast system shows the best performance compared to each of the tested forecast systems, including a calibrated dynamic ensemble system. This represents a significant computational cost saving as a single deterministic model run is required as an input variable. At longer lead times (3-5 days) the CNN again is the best performing forecast system for AR conditions (250 [kg m^{-1} s^{-1}] of IVT), but struggles more than the other systems with predicting the highest impact events (350 and 500+ [kg m^{-1} s^{-1}] of IVT). From a Brier Skill Score (BSS) and Threshold Weighted Continuous Rank Probability Score (twCRPS) perspective the CNN is the best performing forecast at nearly all lead times and for every threshold when compared each other forecast system. However, the BSS can be broken down into components of reliability and resolution (Murphy 1973). Supplemental figures 3.6S & 3.7S shows these components. For the longer lead times, and high impact events it is clear that the CNN is favoring resolution at the expense of reliability. The supplemental material also contains reliability diagrams (Bröcker & Smith 2007) to demonstrate this issue, while still reliable the CNN is marginally less reliable than the other methods (Figs. 3.8S-3.13S). The CNN is clearly reliable for IVT events with magnitudes greater than 250 [kg m^{-1} s^{-1}] at all lead times but struggles slightly with reliability at the longer forecast leads for the highest impact events. Therefore, if a user wants to know if AR conditions are probable, the CNN is the best West-WRF based forecast available among the NN-based methods and AnEn and is competitive or better than the dynamic ensemble methods. For AR events greater than 500 [kg m^{-1} s^{-1}] the AnEn or NN systems are more reliable, but much less resolved (Fig. 3.6S). Our results show that a NN trains extremely quickly and with a single year of hindcast data can create a very reliable probabilistic forecast.

Challenges remain for this DL post-processing systems. The demonstrated DL methods are distinctly disadvantaged in that they fit unimodal parametric distributions, and variables that are not described well by a simple distribution will yield poor probabilistic forecast skill. Additionally, these are highly parameterized models and significant computational time was required to find the prescribed model hyperparameters. The presented neural networks do not offer a seamless forecast system, with individual networks trained at each lead time. The FCN and NN embed location information in their forecasts, which was shown to effect forecast skill, this could easily be extended to embed temporal information (by embedding representations of forecast lead) which would unify the forecast system into a single neural network rather than training individual network's at every forecast lead (e.g., Ham et al. 2021). Additionally, the stability of these networks has not been proven under changing climate scenarios, and the relative non-stationarity of the training data could affect future long hindcast projects. Though a method for addressing this issue through fine-tuning was presented in this study, more work needs to be done to see if this offers a robust solution.

Work is underway to improve the neural network based forecasts for high impact events. The shape of Figure 3.6j indicates a slight dry model bias, suggesting that simple post-processed conditional bias correction may improve CNN model skill further. Success could be found in simply developing the loss function to act on twCRPS rather than CRPS alone. Additionally, focusing loss just on the coastal landfalling points, training with a greater percentage of high IVT events in the training set, adding AR/no AR discriminator networks to the CNN, or adding metrics that specifically target calibration all have offered positive results in initial testing.

This study uses neural networks and a CNN for distributional regression to quantify the prediction uncertainty from deterministic numerical weather forecast systems. The networks compete with or outperform state-of-the-art dynamic models, even when calibrated with the most modern post-processing methods. The model's parameters are estimated by optimizing continuous ranked probability score, a standard metric in evaluating probabilistic weather forecasts, but one that is rarely used in ML communities. The models are flexible, fast, and can be readily trained with a few years of hindcast data.

3.5 Metrics

We provide a summary of the methods used for forecast evaluation. In the following we will refer to a forecast by F, to the random variable of the observation by Y and to a realization of Y by y, i.e. observed IVT. General skill metrics are referred to as S. In order to avoid the known issues in evaluating stratified forecasts (e.g, Bellier et al. 2017; Lerch et al. 2017) careful consideration is taken to ensure that skill metrics remain proper when stratified, additionally all thresholding criteria is performed on the forecasted ensemble mean (Hamill & Colucci 1997; Siegert et al. 2012).

3.5.1 Evaluation Metrics

3.5.1.1 Deterministic Metrics

While the primary goal of this work is to determine the quality of the NN-derived uncertainty quantification. However, we additionally evaluate how the post-processing methods affect the deterministic ensemble mean predictions. This is done largely because the skill of the deterministic prediction greatly affects the efficacy of the probabilistic methods. For example, CRPS reduces to mean absolute error for a deterministic forecast. We show root mean squared error (RMSE, where RMSE = $\sqrt{\frac{1}{N}\sum_{j=1}^{N}(F_j - y_j)^2}$) and Pearson correlation (PC, where PC = $\frac{\sum_{j=1}^{N}(F_j-\bar{F})(y_j-\bar{Y})}{\sqrt{\sum_{j=1}^{N}(F_j-\bar{F})^2}\sqrt{\sum_{j=1}^{N}(y_j-\bar{Y})^2}}$) skill scores. We aggregate forecasts over every location and demonstrate the methods comparative skill score (SS). Each of the above scores (S) may be converted to SS by comparison with the same metric evaluated for a reference forecast $\left(S_{ref}\right)$ through (SS $= 1 - \frac{S}{S_{ref}}$), $-\infty < SS \leq 1$, where positive/negative values are shown to be more/less skillful than the reference forecast.

3.5.1.2 Probabilistic Predictions

Brier Skill (BS, Brier 1950) is used to assess the prediction of binary events, $BS_\tau\left(F_j, y_j\right) = \left[F_j(\tau) - I\{y_j \geq \tau\}\right]^2$, where τ is a prescribed threshold value and I is the indicator (step) function which takes the value 1 if the j^{th} verifying observation exceeds τ and is 0 otherwise, and $F_j(\tau)$ is the probability of event occurrence, which is forecasted. BS is particularly useful to check how skillful probabilistic IVT forecasts are in predicting different events across various established AR thresholds (e.g., Guan & Waliser 2015; Ralph et al. 2019). Continuous ranked probability score (CRPS, Matheson & Winkler 1976) is a measure of overall predictive performance which integrates the squared difference between cumulative probability distribution functions of the forecast (F) and observation (Y), $CRPS(F_j, y_j) = \int_{-\infty}^{\infty} \left[F_j(x) - I\{y_j \leq x\}\right]^2 dx$, where I, again, is the indicator (step) function which takes the value of 1 if $x \geq y$ and 0 elsewhere. The integral in CRPS can be computed analytically for ensemble forecasts (Hersbach 2000) and a suite of continuous forecast distributions (Jordan et al. 2019). To remain proper, CRPS must be tailored to forecast extreme events; Gneiting & Ranjan (2011) defined the threshold-weighted CRPS (twCRPS), $twCPRS(F_j, y_j) = \int_{-\infty}^{\infty} w(x)\left[F_j(x) - I\{y_j \leq x\}\right]^2 dx$, where w is a nonnegative weight function and when $w = 1$ twCRPS reduces to CRPS. To examine extreme events (right tail of distribution) we can set $w(x) = I\{x \geq \tau\}$ where again τ is a prescribed threshold value. The twCRPS integral can be computed numerically and we leverage this method for our verification. Again, for this metric, we aggregate the forecasts over every station location and demonstrate the methods comparative skill.

To compare the relative performance of each scheme we evaluate the Brier skill score $\left[BSS = 1 - \left(\frac{\overline{BS}}{\overline{BS_{ref}}}\right)\right]$ and continuous ranked probability skill score $\Big[CRPSS =$

$1 - \left(\frac{\overline{CRPS}}{\overline{CRPS}_{ref}}\right)\Big]$ which is shown in Figure 3.4 and Figure 3.2S, respectively. Positive values indicate a skill improvement.

The integral in the CRPS equation can be computed analytically for ensembles, and for many continuous forecast distributions (see, Jordan et al. 2019). In this work we use the exact Gaussian CRPS solution to train our neural networks. Though this rarely occurs, as the focus of this study is on IVT events that are above the 250 [kg m^{-1} s^{-1}] threshold, we truncate predictions at 0 as negative values of integrated vapor transport are non-physical. The exact solution of the Gaussian CRPS with mean value (μ) , standard deviation (σ) and observation (y) is $CRPS(F_{\mu\sigma}, y) = \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\varphi\frac{y-\mu}{\sigma} - \frac{1}{\sqrt{\pi}}\right\}$, where Φ and φ denote the CDF and PDF of a standard Gaussian distribution, respectively (Gneiting et al., 2005).

We utilize a two-sided Diebold-Mariano (DM) test to assess whether differences in forecast performances are statistically significant (Diebold & Mariano 2002). Consider two forecasts F_1 and F_2, with respective mean scores $\bar{S}(F_i) = \frac{1}{n}\sum_{j=1}^{n} S\left(F_j^i, y_j\right)$ for $i = 1{,}2$ over a test $j = 1, \ldots, n$ where the forecast F_j^i was issued k time steps before the observation (y_j). The DM test assumes that under standard regularity conditions and the forecast cases are independent, $t_n = \sqrt{n}\,\frac{\bar{S}(F^1) - \bar{S}(F^2)}{\hat{\sigma}_n}$, where $\hat{\sigma}_n = \frac{1}{n}\sum_{j=1}^{n}(S(F_j^1, y_j^1) - S(F_j^2, y_j^2))^2$ follows the standard Gaussian distribution under the null hypotheses of equal predictive performance of two forecast sources. The null hypothesis is rejected for large values of $|t_n|$, by obtaining the corresponding $p - value$, where forecast F^1 out/under performs F^2, if t_n is negative/positive.

In order to account for test multiplicity when comparing methods across multiple forecast lead-times, we follow Wilks (2016), and use the Benjamini-Hochberg procedure (Benjamini &

Hochberg 1995) to control for the false discovery rate. Given the ordered lead-time $p - values$ $p_1, \dots, p_M$ of M hypothesis tests, a new threshold $p - value$ (p^*) is determined via $p^* = p_{(i^*)}$ where $i^* = \min\left(i = 1, \dots, M : p_{(i)} \leq \alpha_{FDR} \cdot \frac{i}{M}\right)$, and we choose $\alpha_{FDR} = 0.05$. We then reject the null hypothesis if the test $p - value$ < p^*. In Figure 3.3, M is set to the number of tested lead times while comparing the twCRPS ($M = 22$).

Rank Histograms (RH) are diagnostic tools that assess the calibration of a forecast ensemble (Hamill 2001). An ensemble is statistically consistent when ensemble members cannot be distinguished from observations. Therefore, an observation ranked among the corresponding ordered ensembles is equally likely to assume any position. If a significant amount of forecasts are assessed in this manner, a histogram of the observation ranks should show a perfectly uniform distribution (rank probability of $(1/(n + 1)$, where n = number of ensemble members). Bellier et al. (2017) proposed the use and demonstrated the consistency of forecast-based stratified rank histograms, which show calibration between given forecast thresholds, and easily enables one to assess the contribution of each stratum to the overall rank histogram uniformity.

The ability of a forecast systems to quantify uncertainty is examined with binned spread-skill plots, which compare ensemble spread (i.e., the standard deviation of the ensemble members) to RMSE of the ensemble mean over small class intervals of model spread, rather than considering the overall average spread as in the dispersion diagram (e.g., van den Dool 1989; Wang & Bishop 2003). The spread of a forecast perfectly describes the uncertainty of the system if the actual forecast error equals its spread (Leutbecher & Palmer 2008). The ability to quantify the prediction uncertainty thus requires the two metrics to match at all binned values of ensemble spread, resulting in a line that falls upon the 1:1 line.

Permutation importance has been explored for describing variable importance to DL models in multiple earth system studies (Brenowitz et al. 2020; McGovern et al. 2019; Rasp & Lerch 2018). Permutation importance is here defined as $\Delta(x^*, x)_{F^1, F^2} = \overline{twCRPS}\left(F_j^1(x), y_j(x)\right) - \overline{twCRPS}\left(F_j^2(x^*), y_j(x)\right)$ where x^* represents in input variable space with a singly randomly permutated input variable selected from the set of input features The input features have length equal to the total number of forecasts (j). Permutation importance (PI) is then set in reference against the non-permuted forecast twCRPS ($PI = \left[\frac{\Delta(x^*, x)_{F^1, F^2}}{\overline{twCRPS}\left(F_j^1(x), y_j(x)\right)}\right]$). The random permutation process is repeated for each input variable for form Figure 3.7.

3.6 Acknowledgement

We acknowledge ample code sharing and collaboration and encourage the machine learning/atmospheric science community to continue this open practice. We thank the West-WRF team led by Daniel Steinhoff for generating the deterministic forecast. This work was supported by USACE FIRO grant W912HZ1520019 and CDWR AR Program grant 4600013361. Sebastian Lerch gratefully acknowledges support by the Deutsche Forschungsgemeinschaft through SFB/TRR 165 “Waves to Weather”, and by the Vector Stiftung through the Young Investigator Group “Artificial Intelligence for Probabilistic Weather Forecasting”.

Python code for reproducing the results and models is available online (github url). West-WRF simulations are archived at the Center for Western Weather and Water Extremes and on the National Center for Atmospheric Research servers are readily available upon request. GEFS data can be retrieved through the TIGGE archive (https://www.ecmwf.int/en/research/projects/tigge). MERRA2 data can be retrieved at https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/data_access/.

Chapter 3, in full is a reprint of the material as it appears in Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.P., Lerch, S., Hayatbini, N. (2021) Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning. Monthly Weather Review. The dissertation author was the primary investigator and author of this paper.

Table 3.1. Abbreviations and descriptions of all input variables

Feature	**Description**	**Model [Input]**
IVT	Integrated Vapor Transport (IVT) [Kg m^{-1} s^{-1}]	CNN/NN/FCN/GEFSnn/AnEn
P_{sfc}	Surface Pressure [hPa]	CNN/NN/FCN/AnEn
U_{500}	500 hPa Zonal Wind [m/s]	CNN/NN/FCN/AnEn
V_{500}	500 hPa meridional Wind [m/s]	CNN/NN/FCN/AnEn
Z_{500}	500 hPa Geopotential Height [m/s]	CNN/NN/FCN/AnEn
IWV	Integrated Water Vapor [mm]	CNN/NN/FCN/AnEn
locID	Location ID number	NN/FCN/GEFSnn/AnEn

Table 3.2. CNN Parameters by category with the convention *Network Layer Component (abbreviation)*: Convolutional Layer (Conv); Max pooling (MP); Addition (Add); Concatenation (Concat); Zero Padding (Pad); Crop (Crop); Dense (Dense); Input (Input); Input Embedding layer (Input Embed); Embedding Vector (embedding). X is the batch size. N represents number of predictors (see Table 3.1).

Convolutional Neural Network				
Layer	**Parameters**	**Activation**	**Norm**	**Shape**
Input	-	-	-	[X,71,57,N]
Pad	[1,3]	-	-	[X,72,60,N]
Conv0	[3,3,16],1,1	LeakyReLU	BatchNorm	[X,72,60,16]
Conv1	[3,3,16],1,1	LeakyReLU	BatchNorm	[X,72,60,16]
MP	2	-	-	[X,36,30,16]
Conv2	[3,3,32],1,1	LeakyReLU	BatchNorm	[X,36,30,32]
Conv3	[3,3,32],1,1	LeakyReLU	BatchNorm	[X,36,30,32]
Conv4	[3,3,32],1,1	LeakyReLU	BatchNorm	[X,36,30,32]
Add	[Conv2, Conv4]	-	-	[X,36,30,32]
Conv2dT	[2,2,16],1,1	LeakyReLU	BatchNorm	[X,72,60,16]
Concat	[Conv2dT, Conv0]	-	-	[X,72,60,32]
Conv5	[3,3,16],1,1	LeakyReLU	-	[X,72,60,16]
Conv6	[3,3,1,2],1,1	Linear	-	[X,72,60,2]
Crop	[1,3]	-	-	[X,71,57,2]
Conv7	[3,3,32],1,1	LeakyReLU	-	[X,71,57,32]
Output	[3,3,1,2],1,1	Linear	-	[X,71,57,2]

Table 3.3. Stratified rank histogram reliability index by method and for forecasts aggregated from lead times 0-48 hours and 54 – 120 hours.

	AnEn	**FCN**	**NN**	**CNN**	**GEFS$_{nn}$**	**GEFS**
F00-F048	.23	.24	.07	.097	**.056**	.50
F054-F120	**.081**	.28	.084	.15	.09	.25

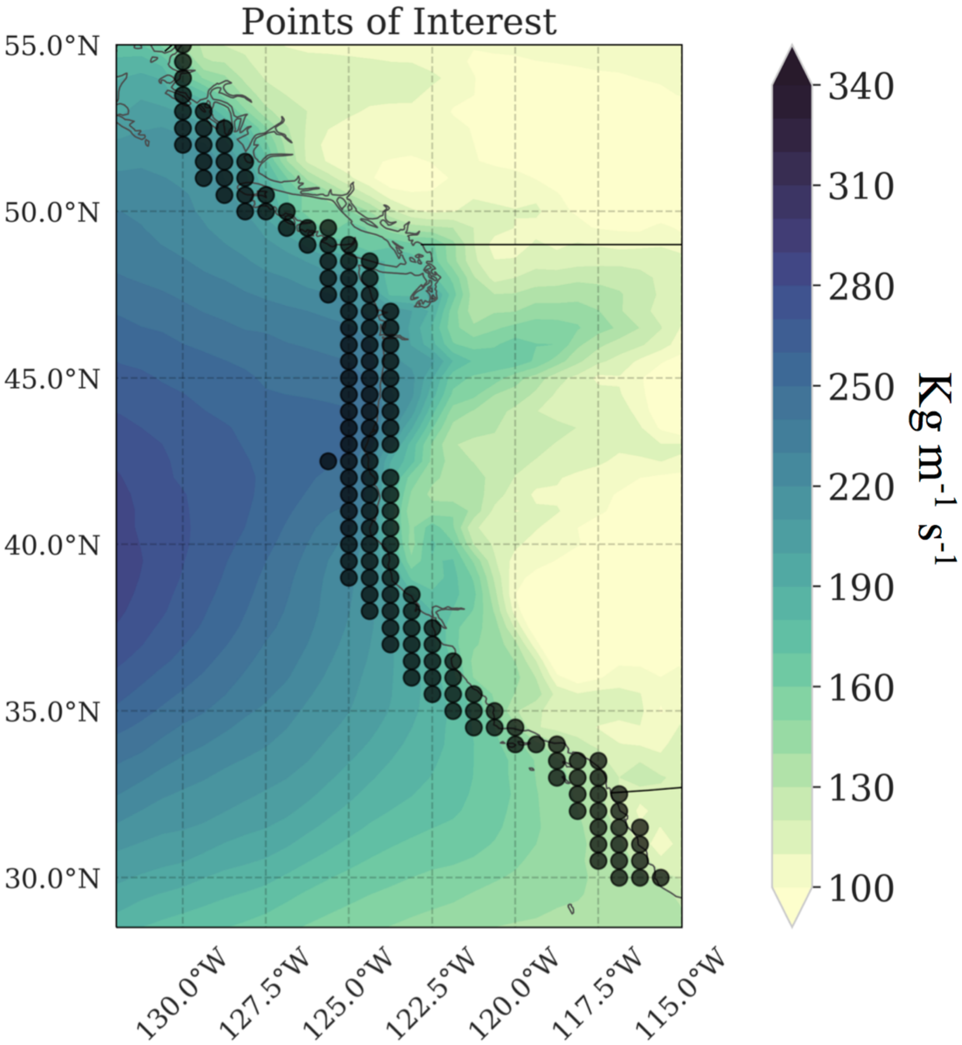

Figure 3.1. Coastal evaluation locations and climatological (December-March 1984-2019) Integrated Vapor Transport (colorfill).

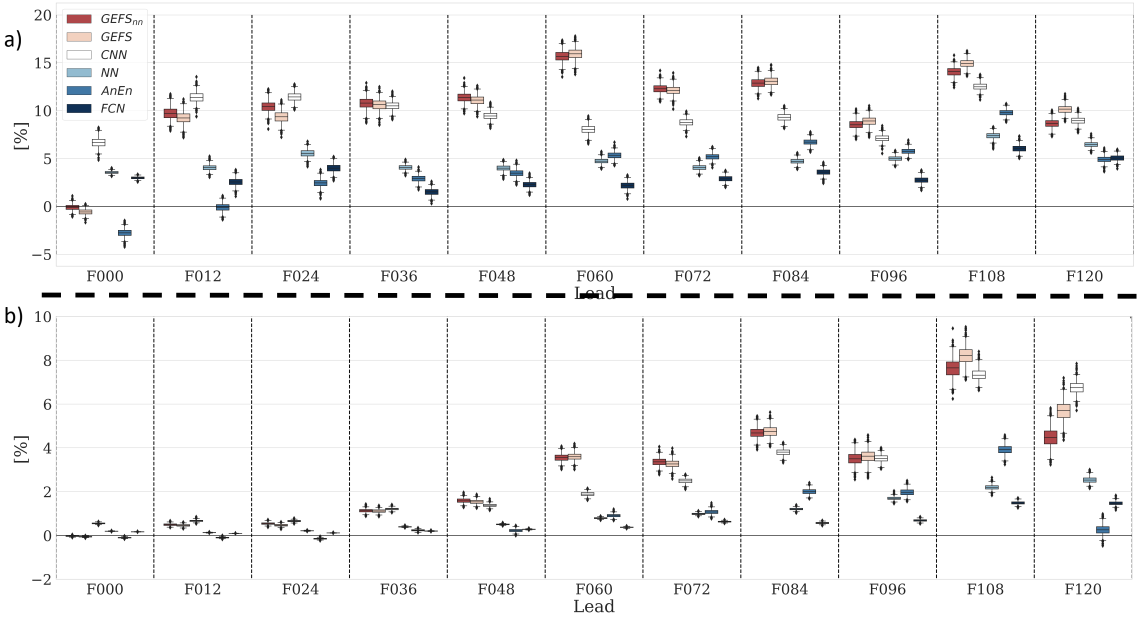

Figure 3.2. Root mean squared error (a) and Pearson Correlation (b) skill scores against the forecast lead time for the ensemble mean or predicted mean from each forecast system. The West-WRF reforecast is used as the reference forecast, with positive values showing percent improvement. Shown predictions include GEFSnn (dark red), GEFS (light red), CNN (white), NN (light blue), AnEn (blue), FCN (dark blue). The error bars indicated the 95% bootstrap confidence intervals (where n=1000).

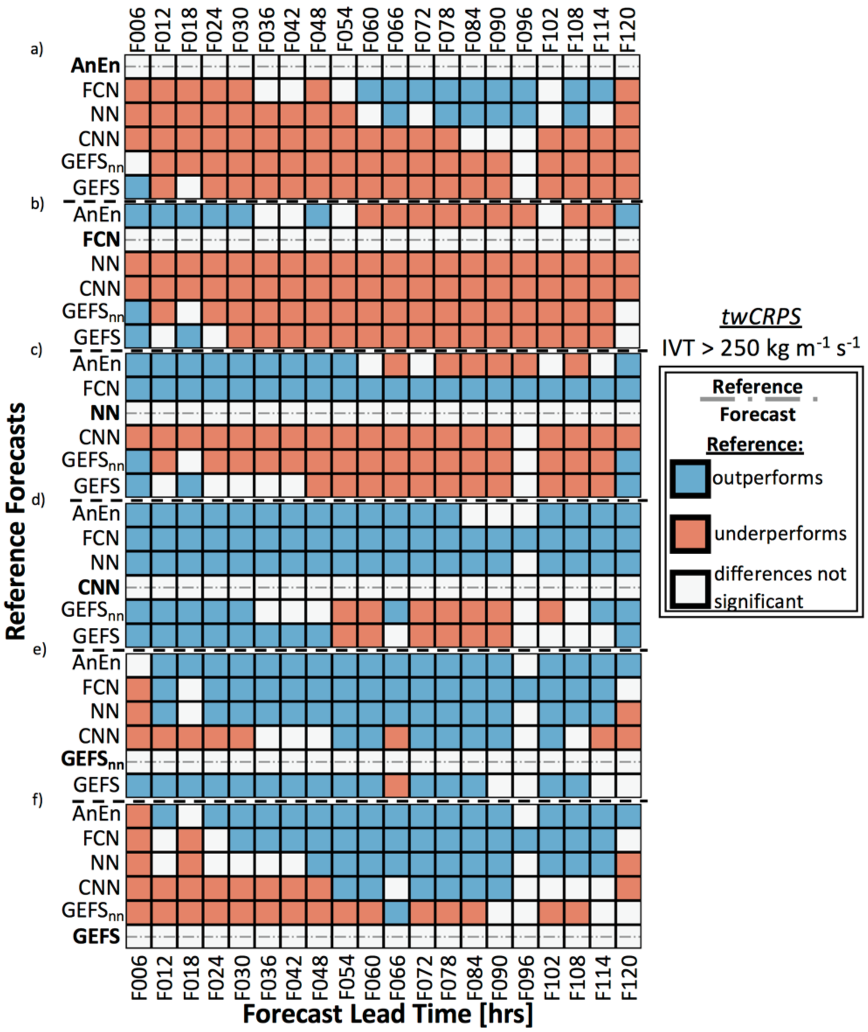

Figure 3.3. Two-sided Diebold-Mariano test using twCRPS (threshold = [250 kg m^{-1} s^{-1}]) for the five forecast systems [AnEn (a), FCN (b), NN (c), CNN (d), GEFSnn (e), and GEFS (f)]. The reference forecast is indicated with a grey dash-dot line. Blue shade indicates that the reference forecast significantly outperforms the rows forecast, red indicates the reference forecast significantly underperforms the rows forecast, and white indicates that the reference forecast and the rows forecast differences are not statistically significant. Significance is determined by examining the test p-values after controlling for the false discovery rate at the level $\alpha_{FDR} = 0.05$.

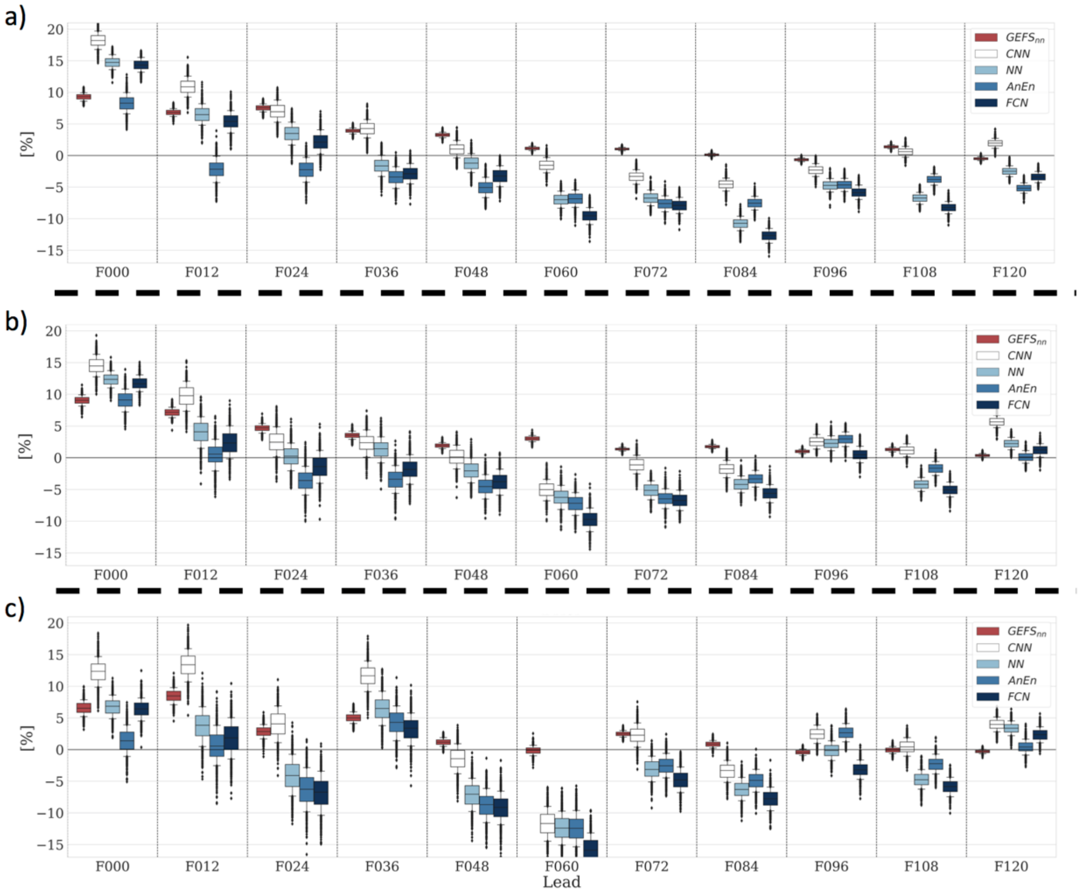

Figure 3.4. Brier skill score at thresholds of 250 (a), 350 (b), and 500+ (c) [kg m^{-1} s^{-1}] forecasted units of IVT against the forecast lead time for the ensemble mean or predicted mean from each forecast system. The GEFS ensemble is used as the reference forecast, with positive values showing percent improvement. Shown predictions include GEFSnn (dark red), GEFS (light red), CNN (white), NN (light blue), AnEn (blue), FCN (dark blue). The error bars indicated the 95% bootstrap confidence intervals (where n=1000).

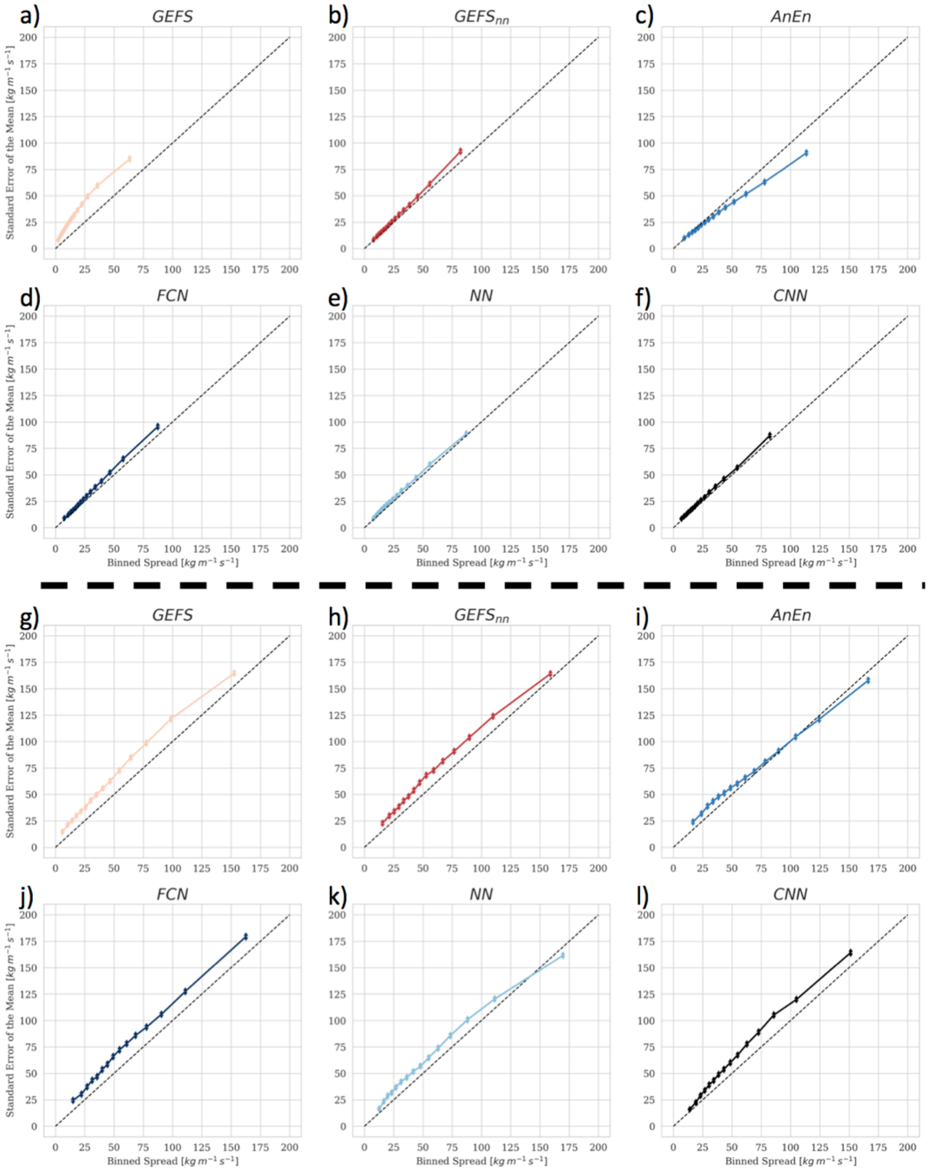

Figure 3.5. Binned spread-skill plots for forecasts from 0-48 hours (a-f) and 56-120 hours (g-l). Error bars indicate the 95% bootstrap confidence interval (n=1000), and the 1:1 dotted line indicates a perfect spread-skill line. For each plot, ensemble spread is binned into 15 equally populated class intervals. Shown prediction systems include GEFS (a, g; light red), GEFSnn (b, h; dark red), AnEn (c, i; blue), FCN (d, j; dark blue), NN (e, k; light blue), CNN (f, l; black).

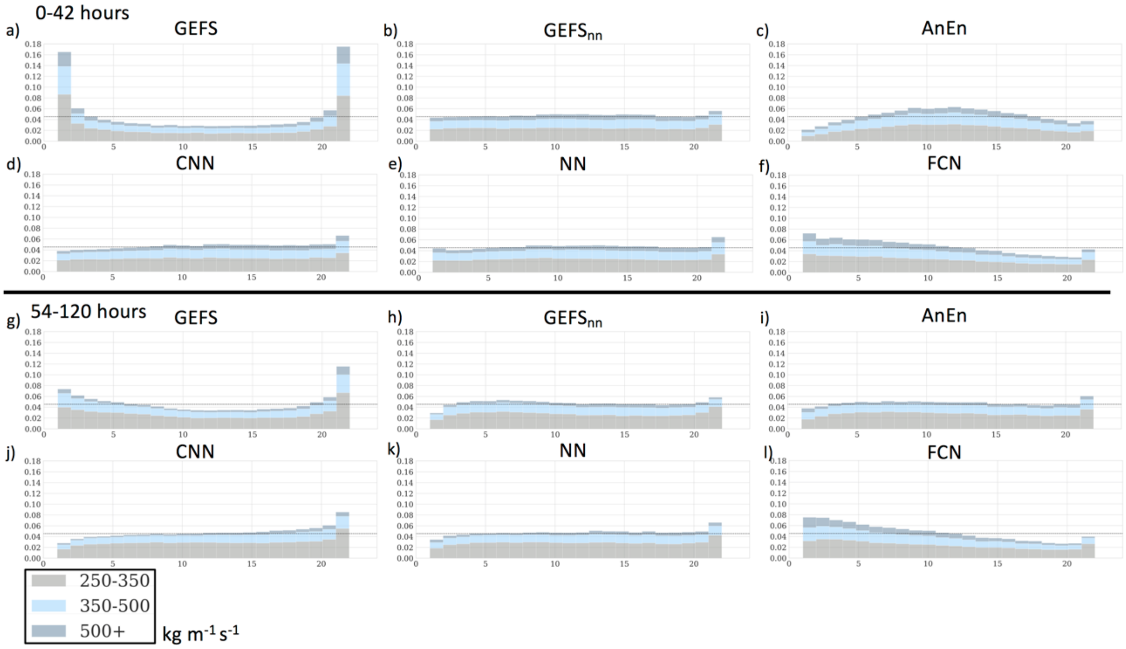

Figure 3.6. Stratified Rank Histograms. X-axis is number of ensembles +1, Y-axis is fractional occurrence of rank. A perfectly calibration forecast is uniform with amplitude at the shown horizontal dotted line. for forecasts from 0-48 hours (a-e) and 56-120 hours (f-j). Stratified on forecasts between ([250-350), [350-500), [500+]). 21 ensembles are drawn from the distribution representing the mean and gaussian spread from individual forecast systems (as labeled).).

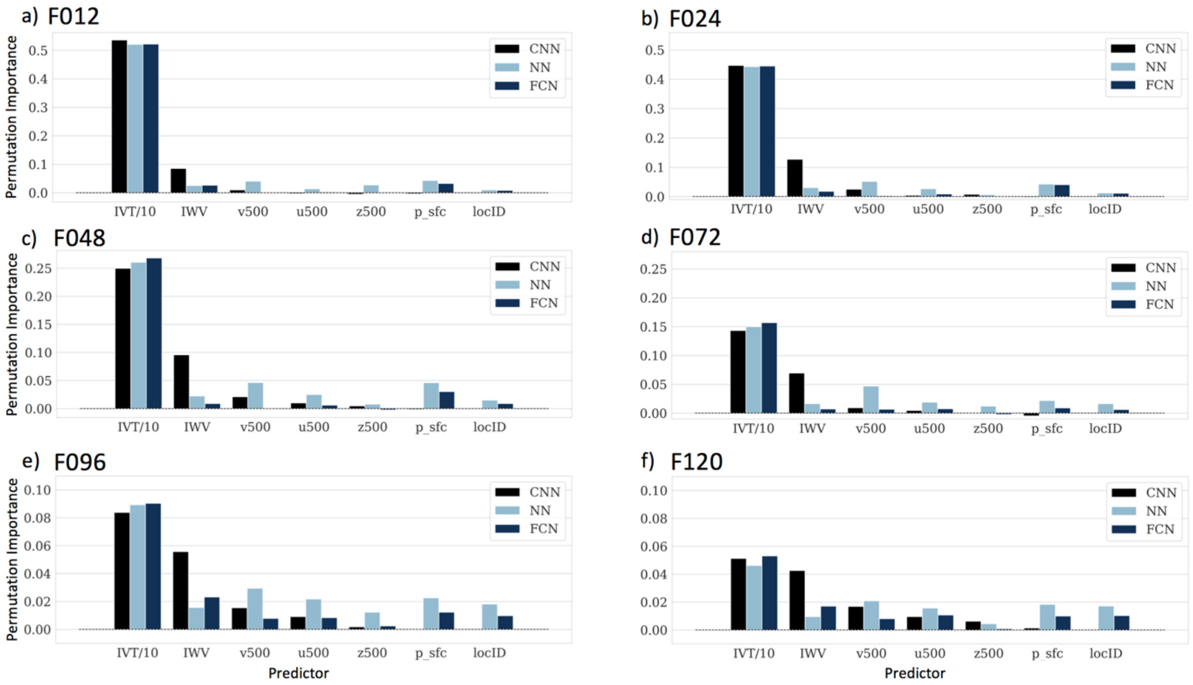

Figure 3.7. Relative permutation importance for the input predictors (defined in Table 3.1) in the CNN (black), NN (light blue), FCN (dark blue) post processing systems for lead times: 12 (a), 24 (b), 48 (c), 72 (d), 96 (e), and 120 (f), using twCRPS with threshold set to 250 kg m^{-1} s^{-1}. Note the changing scale on the y-axis. The CNN system does not leverage a locID predictor. IVT relative predictor importance is divided by 10.

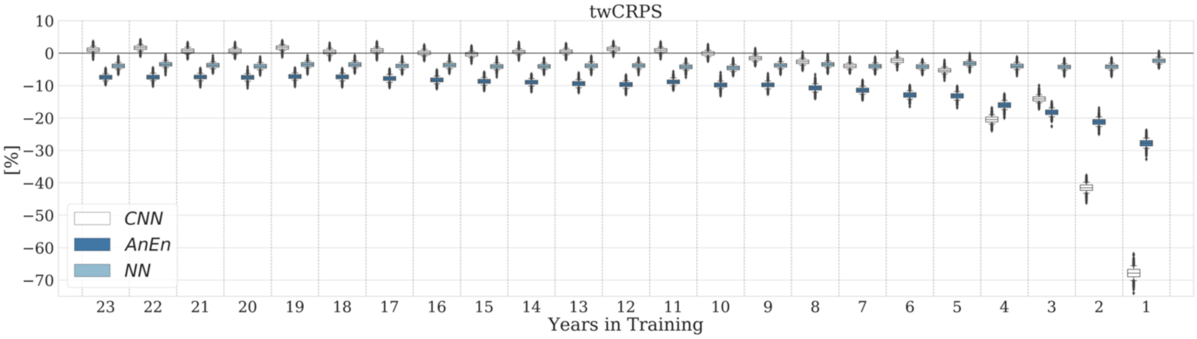

Figure 3.8. Forecasted 48-hour threshold weighted (250 kg m^{-1} s^{-1}) continuous ranked probability skill score against the number of water years included in the training dataset. GEFS$_{nn}$ is used as a reference forecast. Shown prediction systems include: CNN (white), NN (light blue), AnEn (dark blue). The error bars indicated the 95% bootstrap confidence intervals (where n=1000).

3.7 Chapter 3 – Supporting Information

3.7.1 Neural Network Model Architectures and Hyperparameter Search

Here we explain the process and choice of the model architectures used in this study. A vast library of studies have been published on the details of neural network components and architectures and thus we describe the process of parameter selection, but do not describe the parameter itself. We encourage the reader to see Nielsen, (2015) for more details on individual parameter function.

3.7.1.1 Neural Network

A hyperparameter search was conducted so select the model configuration for the described post-processing neural network (NN) system. Hyperparameters were optimized on two forecast lead-times: forecast hour 48 and forecast hour 96, representing short and long forecast horizons (relative to the lead times examined in this study). Though automatic libraries exist to aide in exploring hyperparameter configurations (Hertel et al., 2020), the authors take a manual approach based on DL system intuition and testing. Model configuration selection was determined by finding the minimum CRPS across *the validation datasets*, for three-fold cross-validated data of water years 2016, 2017, and 2018 (see main text for more details). The validation is rotated such that results are evaluated with 3 separate validation forecast years. A near, unlimited set of hyperparameters configurations exist, in order to limit the search space, a few hyperparameters were held constant. The NN always leveraged an Adam optimizer (Kingma et al., 2014) with a 0.005 training step that reduces by 10% on a validation plateau of

5 epochs (learning cycles). After 8 epochs of no decrease in validation error, training is ended. This typically resulted in ~40 training epochs. The batch size remained constant at 50 samples. The Leaky ReLU alpha parameter is held constant at 0.3. Initially, DL model sensitivity was tested to the depth (number of hidden layers) of the neural network. Little change in model CRPS (outside of what could be expected from stochastic gradient descent) was observed on networks with a depth greater than two layers deep, regardless of the other hyperparameters chosen. A depth of 2-layers was thus chosen, and the remainder of the model configurations were cycled through exhaustively according to the following lists, exploring every combination of parameter choice. The selected hyperparameter is shown in bold.

Embedding Vector Size ∈ [**2**, 4, 6, 10]

Dense Node Size ∈ [10, 20**, 30, 40**, 50, 60]

Dense Fine Tune Layer Node Size ∈ [8,16, **32**, 64]

Activation Function ∈ [**'ReLU'**, 'LeakyReLU']

Additionally, L2 regularization, batch normalization and dropout (Ioffe & Szegedy, 2015; Srivastava et al., 2014), were all tested, but the model trained quickly and was not easily subject to overfitting and these measures did not drastically effect CRPS skill. Though we chose the hyperparameters that minimized CRPS most effectively across the validation dataset, we note that the model was not highly sensitive to skill changes and many of the model configurations yielded highly similar results. The batch size remained constant at 50 samples. The final model configuration was chosen that had the lowest CRPS when averaged across the 48 hour and 96 hour forecast lead points.

3.7.1.2 Convolutional Neural Network

A hyperparameter search was conducted to select the model configuration for the described post-processing convolutional neural network (CNN) system. Hyperparameters were optimized on two forecast lead-times: forecast hour 48 and forecast hour 96, representing short and long forecast horizons (relative to the lead times examined in this study). Though automatic libraries exist to aid in exploring hyperparameter configurations (Hertel et al., 2020), the authors take a manual approach based on DL system intuition and testing. Model configuration selection was determined by finding the minimum CRPS across *the validation datasets*, for three-fold cross-validated data of water years 2016, 2017, and 2018 (see main text for more details). The validation is rotated such that results are evaluated with 3 separate validation forecast years. A near, unlimited set of hyperparameters configurations exist; in order to limit the search space, a few hyperparameters were held constant. The CNN always leveraged an Adam optimizer (Kingma et al., 2014) with a 0.005 training step that reduces by 10% on a validation plateau of 5 epochs (learning cycles). After 8 epochs of no decrease in validation error, training is ended. This typically resulted in ~40 training epochs. The batch size remained constant at 50 samples.

The CNN is a U-NET architecture (Ronneberger et al., 2015), which was also leveraged in Chapman et al., (2019) for integrated vapor transport post-processing. Though we offer a brief description here, we refer the reader for a detailed description of the U-Net architecture to Ronneberger et al. (2015). The U-Net is symmetric and consists of three main components 1) the contraction pass 2) the bottleneck layers 3) the expansion pass (see Fig. 3.1S). The contraction pass is made up of contraction blocks, where each block takes an input and applies 3x3 convolutional kernels, for a specified number of layers, followed by 2x2 max-pooling layers. These blocks thus compress the input space to half its input size with each block pass.

The number of convolutional kernels doubles after each contraction block to expand the amount of abstracted information per block. The bottleneck layers take the compressed information and pass it through a specified number of residual blocks (see, He et al., 2015 for information on residual blocks). Each expansion block takes the compressed image, passes it through a specified number of convolutional layers and expands it using a 2x2 convolutional transpose layer, doubling the size of the input matrix with each expansion block. The number of expansion blocks must equal the number of contraction blocks in order to recover the full input image field. The U-NET incorporates "skipped-connections" (see green arrows in Fig. 3.1S) which connect layers earlier in the network to layers deeper in the network allowing less abstracted information (layers early in the network) to easily pass from the input vector to the output IVT. Again, the U-NET must be symmetric to support the skipped connections from the contraction block to the expansion block.

U-net architecture allows for four main model choices which determine the larger model structure: 1) number of convolutional layers per contraction/expansion block prior to max-pooling/convolutional transposing, 2) number of model levels (number of compression/expansion blocks), 3) number of bottleneck layers, 4) number of starting convolutional filters. The following list of these choices were used to search the hyperparameter space, with the selected parameter in bold.

Number of layers per contraction/expansion block ∈ [1, **2**, 3, 4]

Number of starting convolutional filters ∈ [8, **16**, 32, 64]

Number of model levels ∈ [**1**, 2, 3, 5]

Bottleneck residual blocks ∈ [**1**, 2, 3, 4]

Activation Function ∈ ['ReLU', **'LeakyReLU'**]

Additionally, batch normalization (Ioffe & Szegedy, 2015) was applied to assist model training speed. Though we chose the hyperparameters that minimized CRPS most effectively across the validation dataset, we note that the model was not highly sensitive to skill changes once the number of model levels of the CNN was set, and all configurations provided skillful models. The batch size remained constant at 50 samples. The model configuration was chosen that had the lowest CRPS when averaged across the 48 hour and 96 hour forecast lead points.

It is the authors' belief that a good set of hyperparameters was selected, but we would not classify them as the global optimal set. A greater search area could have yielded more skill. However, we note that the network was not particularly sensitive to hyperparameter choice, and improvements to the network likely would not yield greatly improved results.

Further Component Description:

ConvX: These block represent 2d convolution. Each convolutional kernel is a 3x3 matrix.

Conv2dT: This block is a transposed 2d convolution (also called fractionally-strided convolution), used for expanding the images back to their original input size via.

Embedding Vector: This layer transforms positive indexes into dense vectors of specified size 2.

Leaky ReLU: This is a non-linear activation function. Leaky ReLU is a piecewise linear function with a slope of 1 if the input value is greater than 0 $f(x) = x \ where \ x \geq 0$, and a slope of 0.3 if less than zero $f(x) = 0.3 \cdot x \ where \ x < 0$.

Batch Normalization: Applies a transformation that ensures the output of each layer has a mean near 0 and standard deviation near 1.

Max Pooling: This is used to down sample the input vector via a 2x2 matrix which selects the maximum value in the matrix, the stride is set to 1.

3.7.1.3 Brier Decomposition

The Brier skill score, which is the RMSE of probabilistic forecasts, can be decomposed into 3 additive components: reliability, resolution and uncertainty. We refer the reader to (Murphy, 1973) for a detailed decomposition. We note that we compile forecasts into 10 probability bins ([0-0.1), [0.1-0.2), etc.) to aggregate the forecasts for this calculation. Additionally, the uncertainty, which depends solely on the sample climatology (thus only depends on observations) is not shown as it is definitionally exactly similar for every forecast system. We show the forecast resolution and reliability for thresholds of [250, 350, 500] (kg m^{-1} s^{-1}) in figures 3.4S and 3.5S.

3.7.1.4 Reliability Diagram Explanation

Following Bröcker & Smith (2007) we show reliability diagrams which assess the calibration of a forecast system using a variety of threshold values across every forecast at every station location. We compile forecasts into 10 probability bins ([0-0.1), [0.1-0.2), etc.) to aggregate the forecasts. We also examine frequency of forecast probabilities for each category which helps to assess the sharpness given a specific threshold graphically, with sharp forecasts characterized by higher frequencies for the forecasted probabilities close to either 0 or 1.

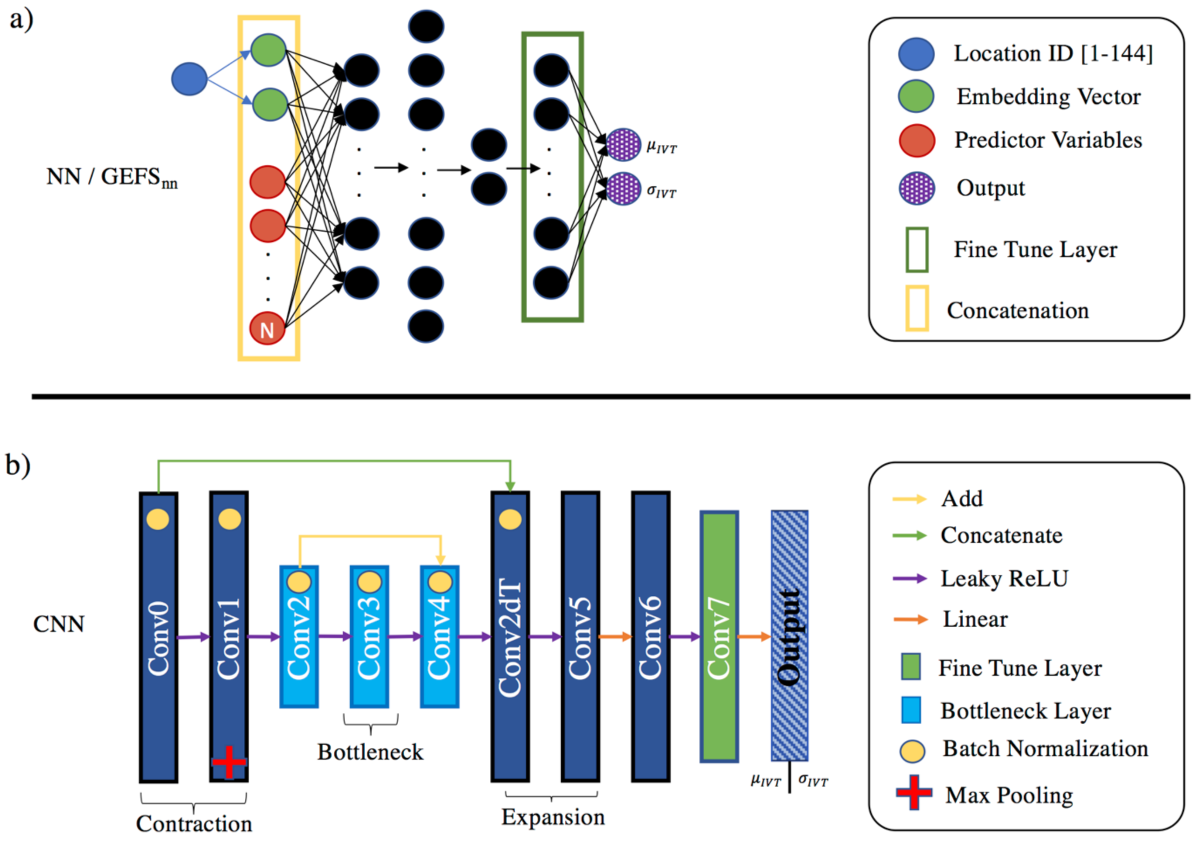

Figure. 3.1S. network architecture for the leveraged Neural Network (a; NN) and convolutional neural network (b; CNN). See Table 3.1 for input variables and Table 3.2 for layer specifications. For ease of viewing hidden layer nodes (black dots in (a)) are not shown fully connected, (a) is a fully connected NN. We include an additional component key below to aid the reader in network interpretation.

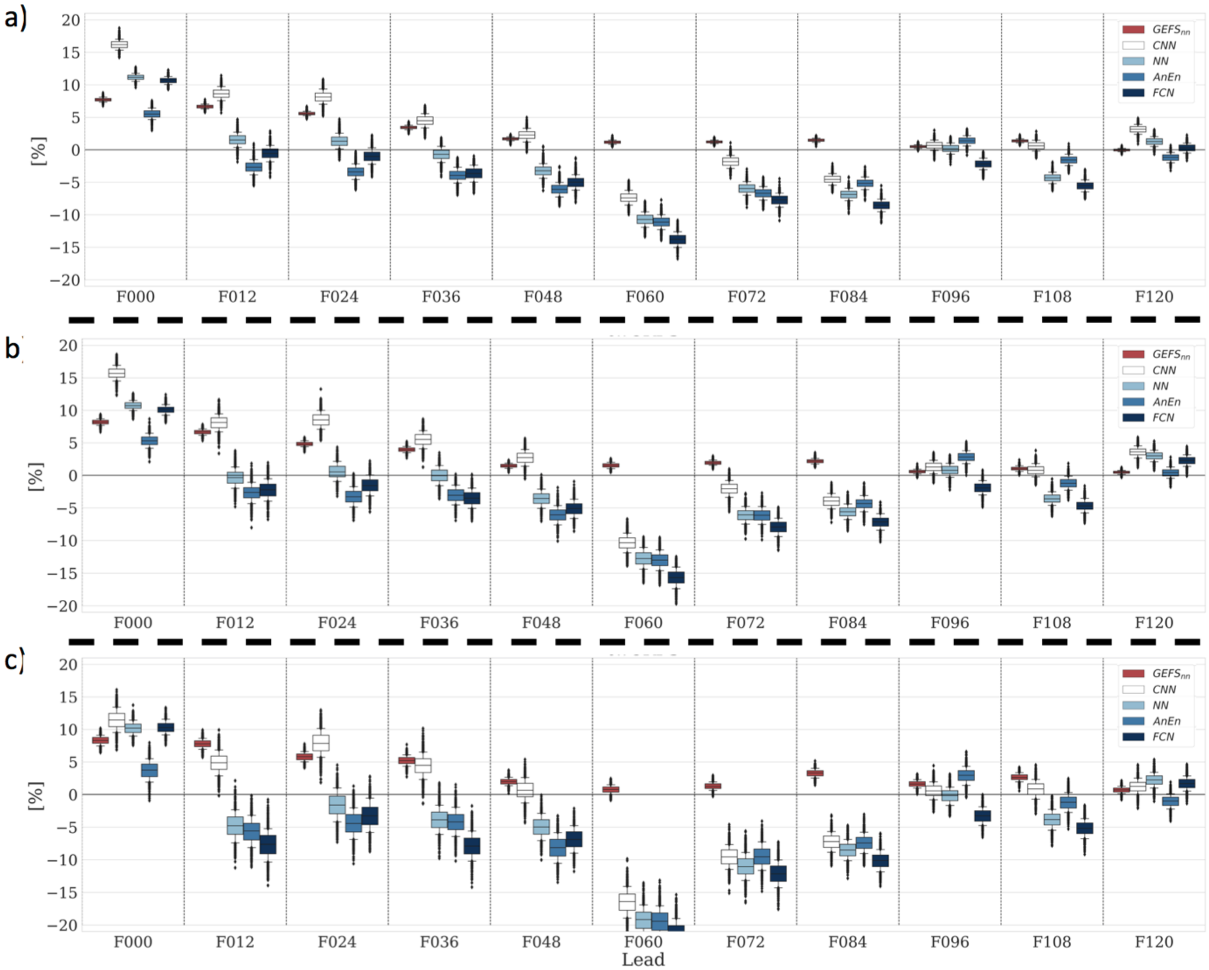

Figure. 3.2S. As in Figure 3.4, but for threshold weighted continuous ranked probability score at thresholds of 250 (a), 350 (b), and 500+ (c) [kg m^{-1} s^{-1}] forecasted units of IVT.

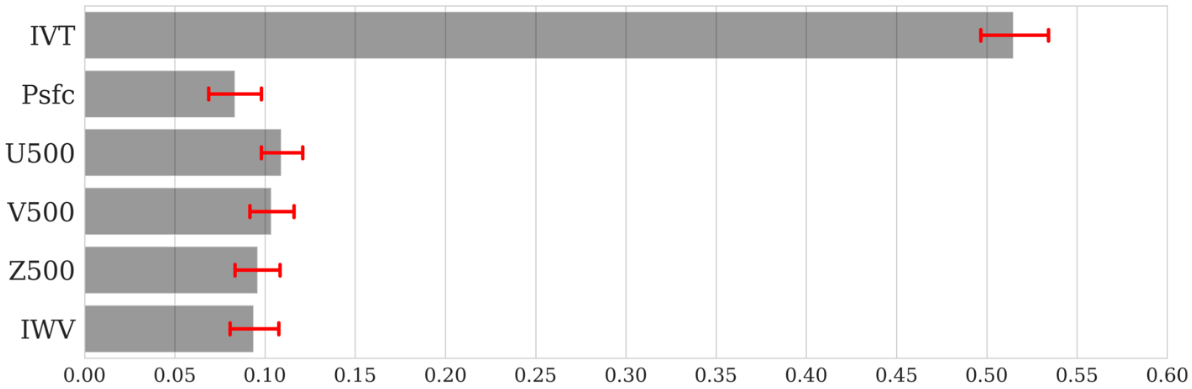

Figure. 3.3S. The mean values of the weights across every landfall locations received by each predictor as a result of the brute force optimization using the modified version with the bias correction (AnEn) as described in Section 2.4 (The Analog Ensemble for Rare Events). Red intervals indicate 1 standard deviation of the weight distribution (standard deviation across 144 station locations).

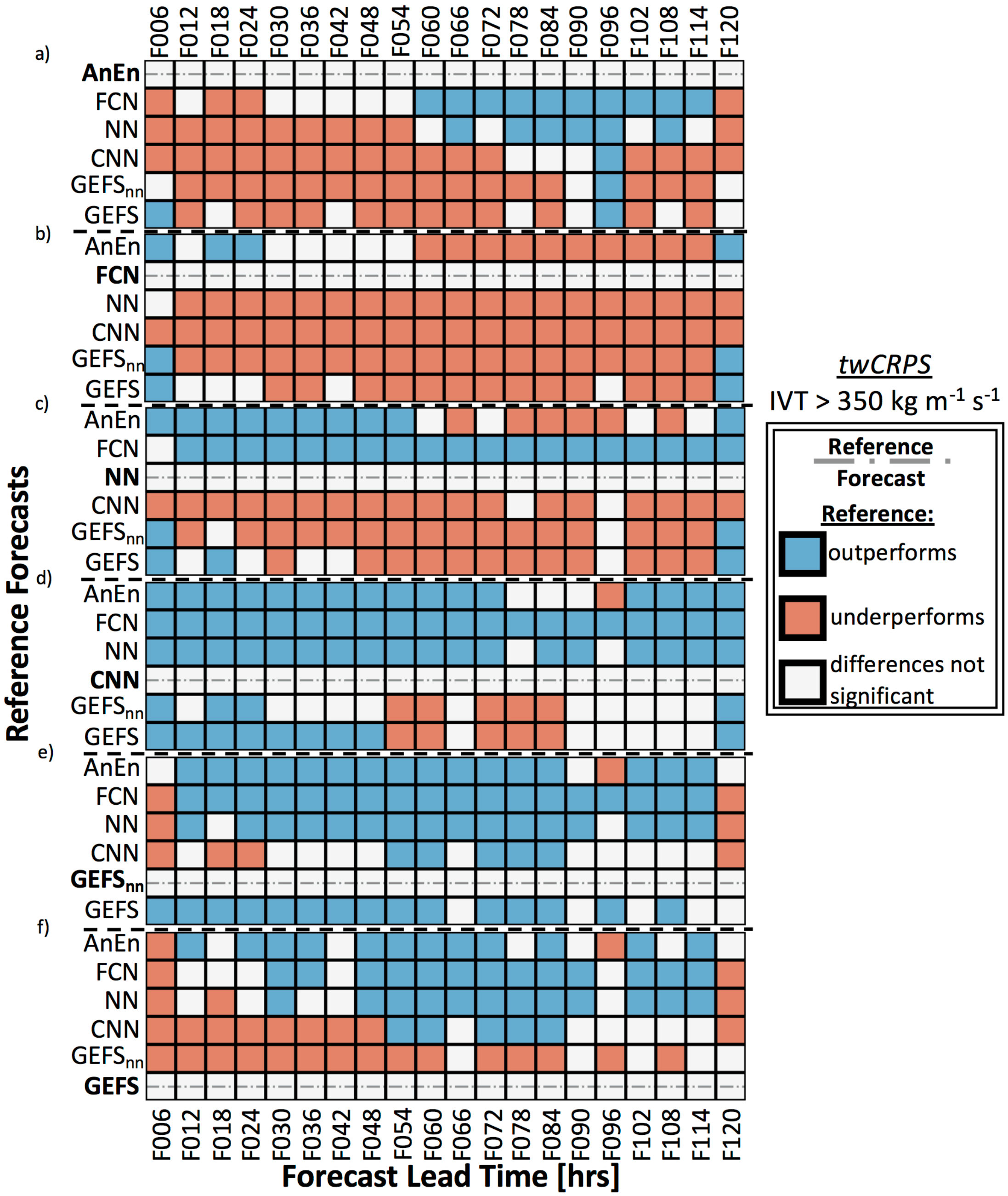

Figure. 3.4S.. Diebold-Mariano test for AnEn (a), FCN (b), NN (c), CNN (d), GEFSnn (e), and GEFS (f) as in Figure 3.3 but for twCRPS (threshold = [350 kg m^{-1} s^{-1}]).

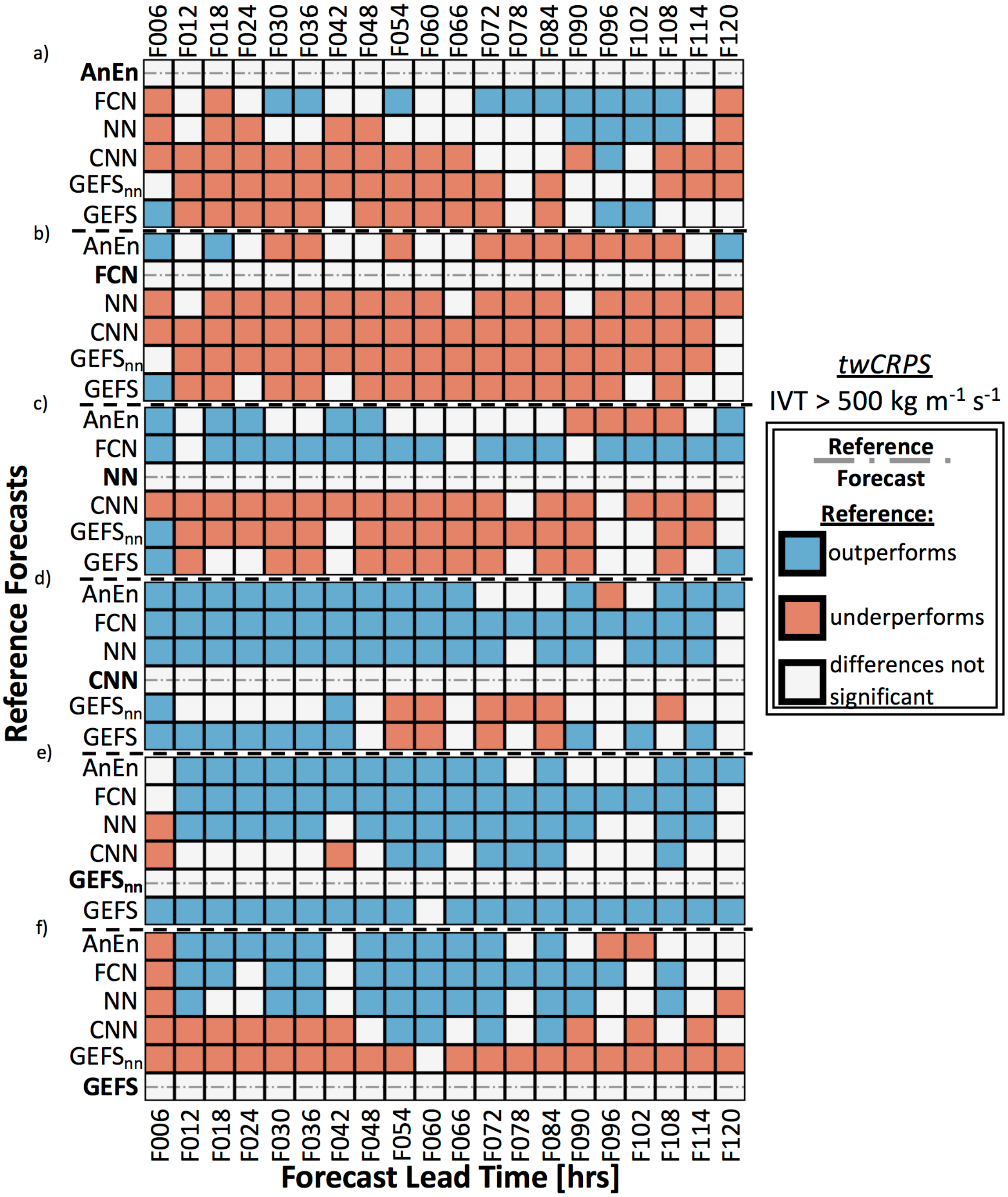

Figure. 3.5S. Diebold-Mariano test for AnEn (a), FCN(b),NN (c), CNN (d), GEFSnn (e), and GEFS (f) as in Figure 3.3 but for twCRPS (threshold = [500 kg m^{-1} s^{-1}]).

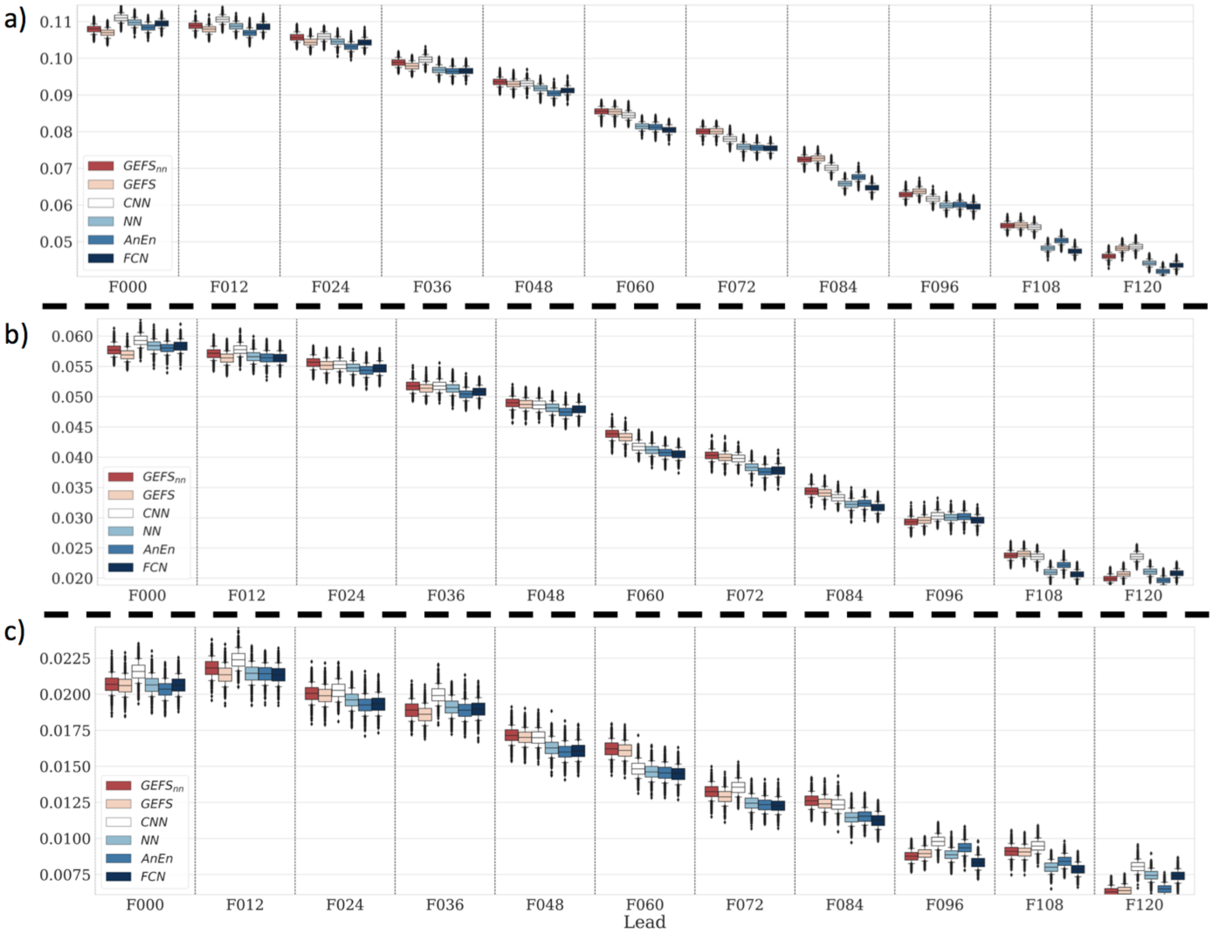

Figure. 3.6S. As in Figure 3.4. but for Brier decomposed Resolution at thresholds of 250 (a), 350 (b), and 500+ (c) [kg m^{-1} s^{-1}] forecasted units of IVT. This metric is positively oriented.

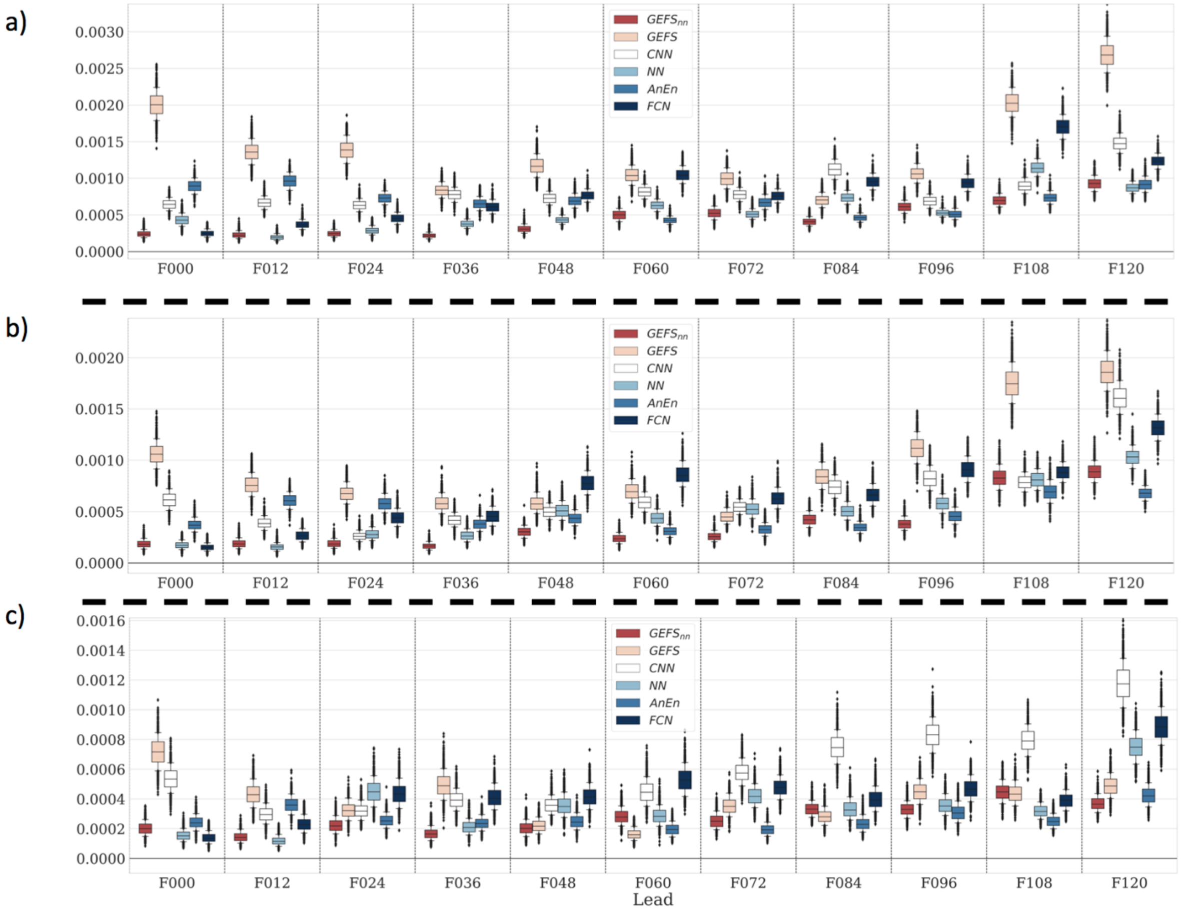

Figure 3.7S. As in Figure 3.4. but for Brier decomposed Reliability at thresholds of 250 (a), 350 (b), and 500+ (c) [kg m^{-1} s^{-1}] forecasted units of IVT. This metric is negatively oriented.

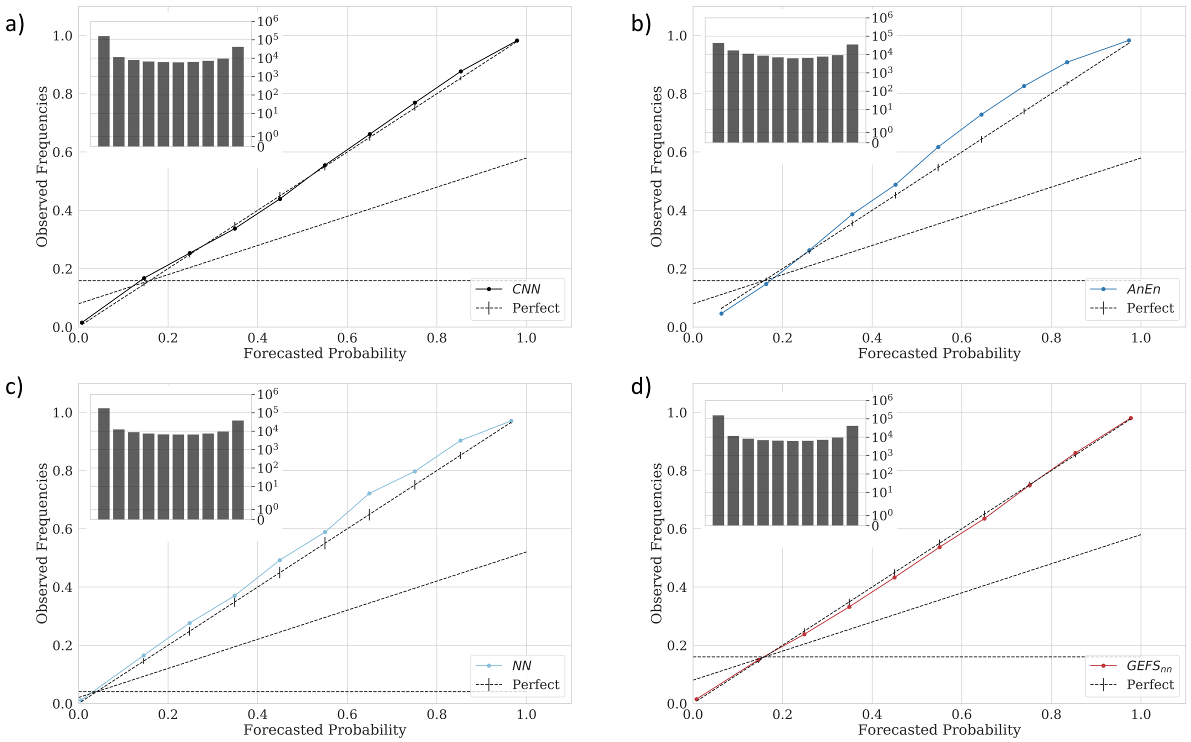

Figure 3.8S. reliability diagrams for the CNN (a), AnEn (b), NN (c), and GEFS$_{nn}$ (d) forecast systems for an exceedance threshold of 250 [kg m^{-1} s^{-1}] units of IVT. Results are based on observation- forecast pairs of all cross-validated testing years aggregated over all locations from 0-48 hour forecasts. The inset histograms show the frequencies for each of 10 forecast probability bins in log$_{10}$ scale. Bars on the diagonal indicate the bootstrapped 95% confidence interval of a perfect forecast.

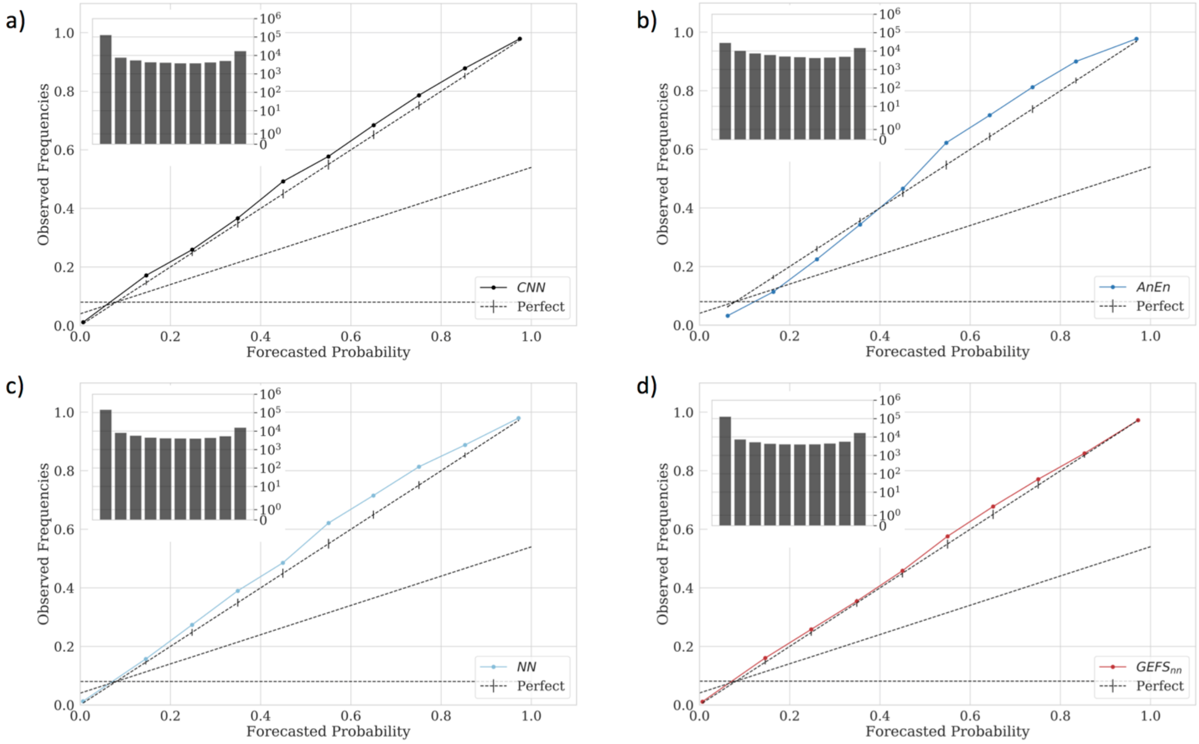

Figure 3.9S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.8s but for an exceedance threshold of 350 [kg m^{-1} s^{-1}] units of IVT.

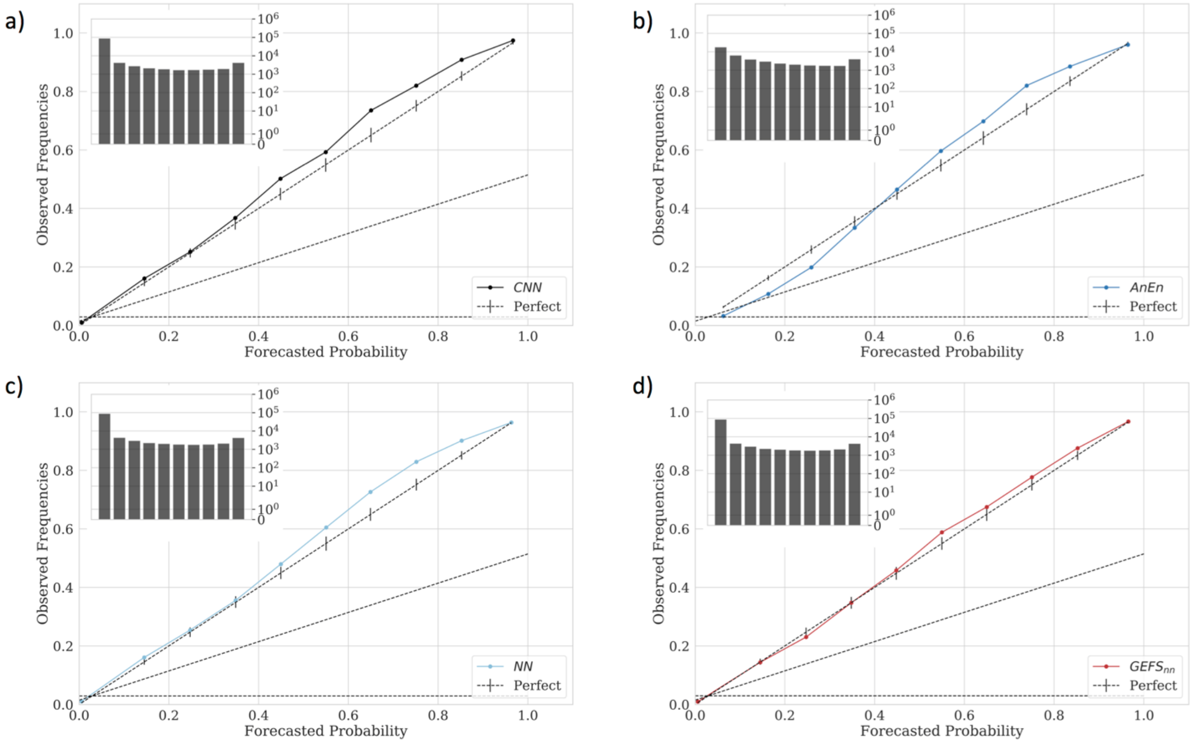

Figure 3.10S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.8S but for an exceedance threshold of 500 [kg m^{-1} s^{-1}] units of IVT.

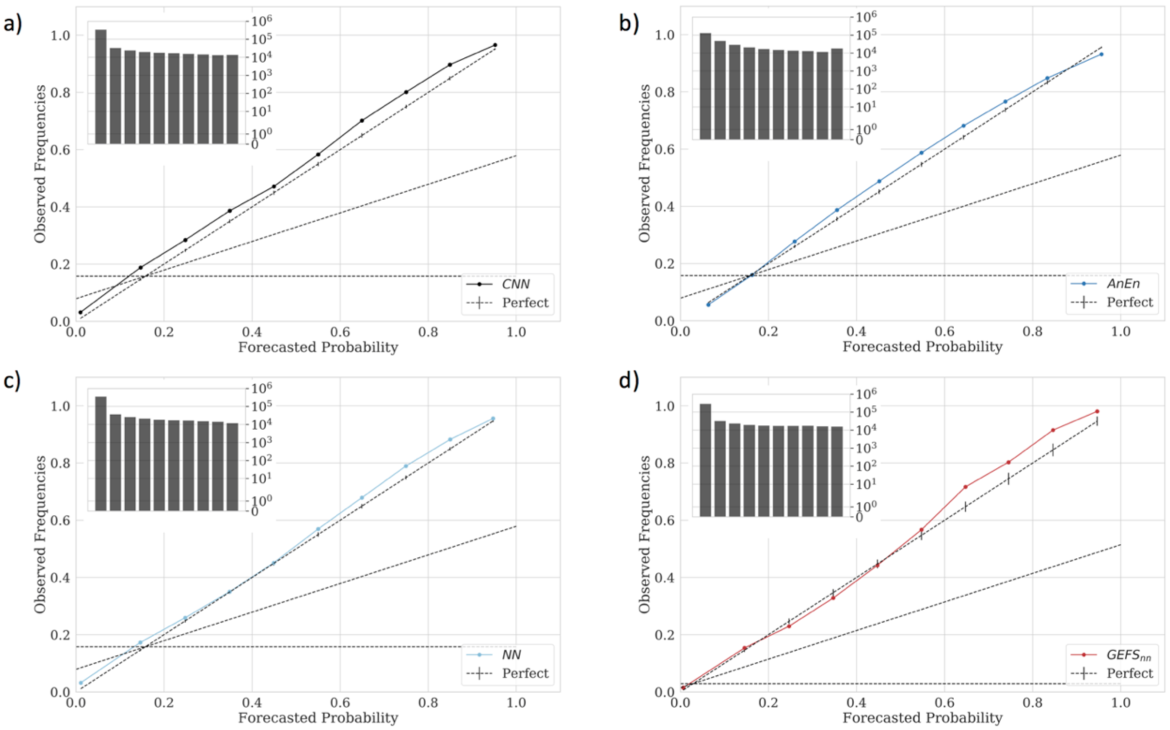

Figure 3.11S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.8S but for an exceedance threshold of 250 [kg m^{-1} s^{-1}] units of IVT for forecasts from 54-120 hours .

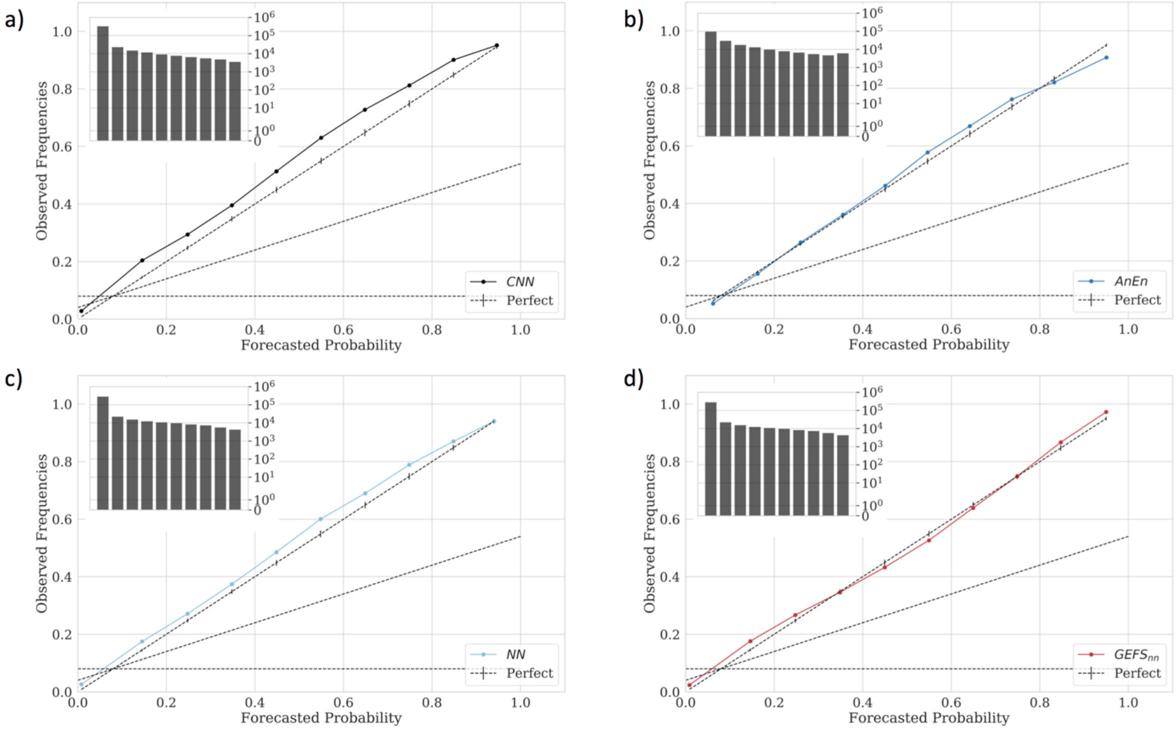

Figure 3.12S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.11s but for an exceedance threshold of 350 [kg m^{-1} s^{-1}] units of IVT.

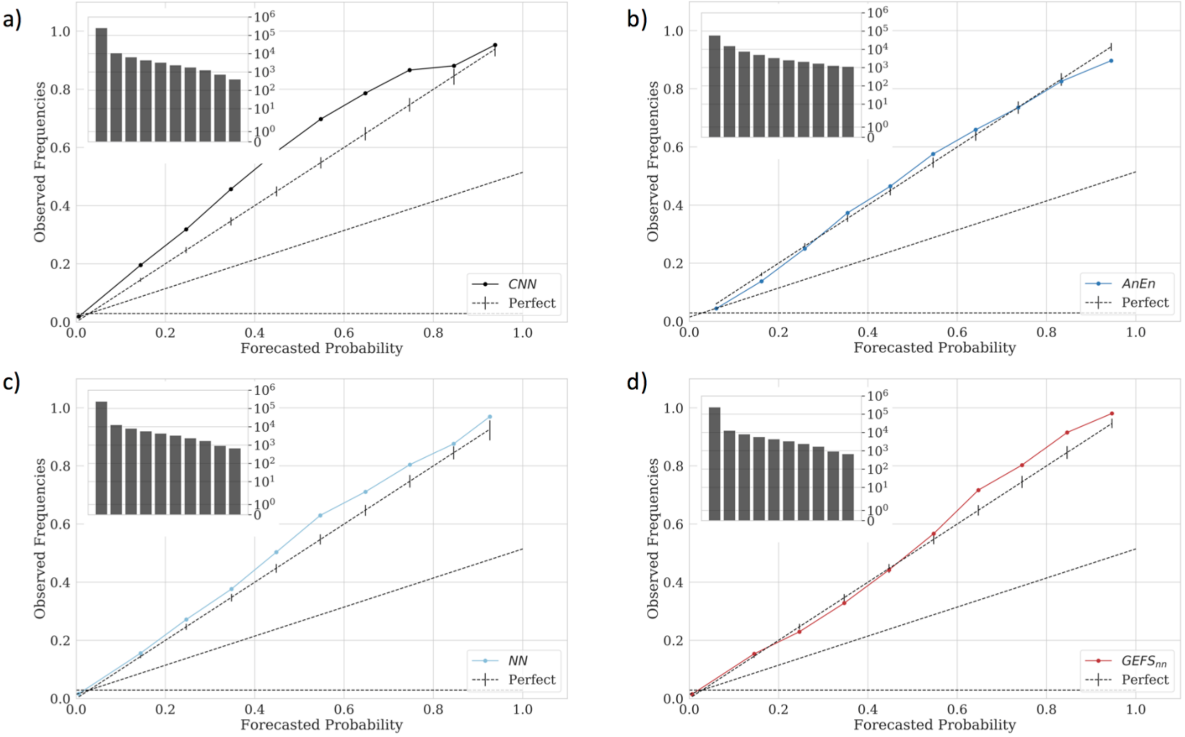

Figure 3.13S. Reliability diagrams for CNN (a), AnEn (b), NN (c), and GEFSnn (d) as in Figure 3.11S but for an exceedance threshold of 500 [kg m^{-1} s^{-1}] units of IVT.

Chapter 4

Monthly Modulations of ENSO Teleconnections: Implications for Potential Predictability in North America

Abstract

Using a high-resolution atmospheric general circulation model simulation of unprecedented ensemble size, we examine potential predictability of monthly anomalies under El Niño Southern Oscillation (ENSO) forcing and background internal variability. This study reveals the pronounced month-to-month evolution of both the ENSO forcing signal and internal variability. Internal variance in upper-level geopotential height decreases ~10% over the North Pacific during El Niño as the westerly jet extends eastward, allowing forced signals to account for a greater fraction of the total variability, and leading to increased potential predictability. We identify February and March of El Niño years as the most predictable months using a signal-to-noise analysis. In contrast, December, a month typically included in teleconnection studies, shows little-to-no potential predictability. We show that the seasonal evolution of SST forcing and variability leads to significant signal-to-noise relationships that can be directly linked to both upper-level and surface variable predictability for a given month. The stark changes in forced response, internal variability, and thus signal-to-noise across an ENSO season indicate that subseasonal fields should be used to diagnose potential predictability over North America associated with ENSO teleconnections. Using surface air temperature and precipitation as examples, this study provides motivation to pursue 'windows of forecast opportunity', in which

statistical skill can be developed, tested, and leveraged to determine times and regions in which this skill may be elevated.

4.1 Overview

El Niño-Southern Oscillation (ENSO) is the most influential mode of global climate variability. ENSO usually develops during early boreal summer, peaks in winter, and decays in spring. Eastern Pacific tropical SST anomalies associated with ENSO events result in anomalous convective tropical precipitation. The latent heating response in the Tropical Pacific drives divergent wind and vorticity anomalies in the upper troposphere, which communicate with the extratropics via Rossby waves. Due to the location of the extratropical divergence and the Asian-Pacific jet, quasi-stationary Rossby wave generation arises in preferred locations over the Pacific Basin (Bjerknes, 1969; Sardeshmukh & Hoskins, 1988; Wallace and Gutzler, 1981), anchoring geopotential height (GPH) anomalies, and influencing North American weather, largely through the well-studied Pacific North American (PNA) pattern (J. Bjerknes, 1969; Trenberth et al., 1998; Wallace & Gutzler, 1981a).

Atmospheric general circulation models (AGCM) are useful for examining the effect of ENSO on the predictability of the extratropical atmosphere (e.g., Branstator & Teng, 2017; Lau & Nath, 1996; Matsumura et al., 2010; Yang et al., 1998; Zheng et al., 2004). Ensemble members, influenced by similar lower boundary conditions but with perturbed initial conditions, result in a myriad of climate realizations which span the realistic range of atmospheric responses to boundary condition forcing. Lower-boundary forced signals manifest in the ensemble mean, working to make coherent anomalies despite the inter-ensemble member variability. However, the precise extratropical response to ENSO is difficult to determine as 1) there is year-to-year SST variability amongst ENSO events (e.g., Deser & Wallace, 1987; Johnson, 2013; Newman et al., 2011) resulting in an array of forced atmospheric responses (e.g., Barsugli & Sardeshmukh, 2002; Johnson & Kosaka, 2016) and 2) it exists within a background natural climate variability which acts to mask the SST forcing.

If the response to lower-boundary forcing is understood, then diagnosing and understanding the slow varying modes inherent to the land and sea surfaces (i.e., ENSO, seasonal snowpack, etc.) can aid in subseasonal-to-seasonal (S2S) predictions. Predictability is typically studied in a signal-to-noise (SN) framework, in which the influence of the forcing is set in ratio against natural variability. SN has been used in several previous studies to diagnose the predictability of ENSO driven cold-season extratropical circulation (e.g., Abid et al., 2015; Kumar & Hoerling, 1998; Peng & Kumar, 2005; Sardeshmukh et al., 2000). The SN can be increased via two pathways: 1) an increase in the influence of the forced component; for example, as prescribed by the influence of ENSO SST and atmospheric teleconnections 2) a significant decrease in atmospheric internal variability.

Many studies have demonstrated that the forced atmospheric response to interannual SST variations is important for the interannual variations in mid-latitude climates despite internal variability (e.g., Chen & Kumar, 2015; Kumar & Hoerling, 1995; Mizuta et al., 2017; Shukla & Wallace, 1983; Trenberth et al., 1998). Additionally, there is consensus that an increased atmospheric forced component associated with ENSO (dominantly in the warm phase) events leads to a higher seasonal predictability within the PNA region (e.g., Abid et al., 2015; Chen and Dool, 1999; Kumar & Hoerling, 1998; Peng & Kumar, 2005; Sardeshmukh et al., 2000) and over the North Atlantic (Ayarzagüena et al., 2018; Honda et al., 2005; Bernat Jiménez-Esteve and Domeisen, 2018). However, studies disagree on the magnitude of ENSO modulation on internal atmospheric variability. Sardeshmukh et al. (2000) show an increased (decreased) extratropical internal variability during El Niño (La Niña). Others observed negligible changes in the internal variability of GPH (Kumar and Hoerling, 1998) or associated surface variables (Chen and Kumar, 2015) conditioned on ENSO state. Kumar et al. (2000) documented a nonlinear ENSO modulation of internal atmospheric variability in the PNA region, with El Niño decreasing extratropical 500-hPa GPH internal variability over the North Pacific greater than La Niña increased internal variability. However, this did not significantly improve SN relative to the contribution of the ensemble mean shift. Abid et al., (2015) and Peng and Kumar, (2005) both report significant decreases (increases) in internal variability in El Niño (La Niña), leading to a significantly enhanced (diminished) SN relationship. However, there is evidence that these different conclusions may be due to the inclusion of different ENSO events and the number of examined ensembles, as SN does not vary wildly between models (Kang et al., 2011; Kang and Shukla, 2006).

Trenberth et al., (1998) review studies that have diagnosed tropical-extratropical interactions due to anomalous tropical SSTs, and reveal key factors in determining the extratropical response. These include the location and intensity of tropical circulation anomalies, the effects of the mean flow on planetary wave propagation and forcing, interactions with midlatitude storm tracks, and interference from the internal chaotic variability of the midlatitude circulation (Trenberth et al., 1998 and references therein). The extratropical atmosphere has been observed to respond nonlinearly to ENSO cold and warm events, with a dominant SST forced response occurring in the warm phase and a milder reaction during cold events (e.g., Hoerling et al., 1997; B Jiménez-Esteve & Domeisen, 2019). Additionally, the impact of the annual cycle on the global wind field, and thus the barotropic Rossby wave guide, leads to drastic dynamic changes in the background state upon on which low frequency forcing acts (R Seager et al., 2010; Souders et al., 2014). Therefore, studies which examine the departure from seasonal means rather than incorporating important month-to-month differences are less effective and potentially misleading, particularly in late winter early spring (Newman & Sardeshmukh, 1998). There has a been a recent reexamination of ENSO teleconnection and their extratropical manifestations (e.g., Chen & Kumar, 2015; Deser et al., 2017, 2018; Zhang et al., 2014). However, there has been much less work which resolves the significant intraseasonal differences sparking from a changing monthly background state.

Increasing computational resources enable AGCMs to now run at higher resolution, larger ensemble size, and utilize longer historical records. These added statistics permit a reexamination and further exploration of large-scale dynamics and their influence on extratropical predictability from a SN standpoint. In this study, we test the reliability of the PNA-like response, and the effects on temperature and precipitation anomalies associated with ENSO events. We employ a high-resolution, large ensemble AGCM to examine the dynamic effect of anomalous ENSO forcing, and the seasonal variations at monthly resolution. We then explore noticeable differences in month-to-month internal variability driven by changes in large scale dynamics within the PNA sector. The resulting monthly changes in SN relationships imply important changes in the level of predictability of given variables. Finally, to test the utility of the PNA driven changes, we diagnose whether the SN modulation manifests in monthly observed anomaly composites and also to improved predictive utility on monthly timescales. Utilizing in-situ observations, we construct a simple probabilistic framework and adopt an information theory based potential predictability (PP) perspective (Kleeman, 2002) to show the month-to-month impact of ENSO on temperature predictability.

4.2 Data and Methods

4.2.1 AGCM experiments

To diagnose the atmospheric response to prescribed SST conditions, we utilize monthly mean values from a 100-member ensemble AGCM. Ensemble data were produced by the Meteorological Research Institute AGCM, version 3.2 (Mizuta et al., 2017) at a horizontal spectral resolution with triangular truncation at wave-number 319 and linear Gaussian grid (TL319; equivalent to 60-km mesh) with 64 vertical layers (Murakami et al., 2012). The AGCM was driven by observation-based SST, sea-ice concentration, and radiative forcing (greenhouse gases, aerosols and ozone) from 1951-2010, derived from the Centennial In Situ Observation-Based Estimates (COBE/COBE-SST2) (Hirahara et al., 2014). Small SST perturbations based on slight adjustments to the empirical orthogonal functions of the interannual variation of SST analysis [see the appendix of Mizuta et al., (2017)] were added to the COBE SST to account for uncertainties in analysis (Hirahara et al., 2014). It has been shown that the spread in climate response due to the perturbed SST is comparable to that due to initial condition perturbations (Mizuta et al., 2017). Sea-ice concentration was derived from a quadratic equation on the sea-ice/SST relationship (Hirahara et al., 2014). This dataset, titled the Database for Probabilistic Description of Future Climate Change (d4PDF), has been used to evaluate historical variations of atmospheric responses to global SST variability (e.g., Kamae et al., 2017; Mei et al., 2019; Naoi et al., 2020). More details of the experimental setup can be found Mizuta et al. (2017) and Kamae et al. (2017).

The high-resolution of d4PDF, a state-of-the-art model with a physically consistent Northern Hemisphere atmospheric response to slowly varying mode forcing (e.g. ENSO, Pacific Decadal Oscillation, etc.) (Kamae et al., 2017) will likely represent transients (Hertwig et al., 2015), the atmospheric response to ENSO (Andrew Dawson et al., 2011), mid-latitude blocking (Davini et al., 2017), and major weather regimes (Andrew Dawson & Palmer, 2015) better than a low-resolution model as shown by previous studies referenced here.

4.2.2 Definition of ENSO and Compositing

We define the ENSO index as the 3-month running mean of COBE-SST2 anomalies in the Niño3.4 region ([5°S, 5°N];[170°W,120°W]). Anomalies are derived from centered 30-year base-periods updated every five years, in the exact manner as NOAA's Oceanic Niño Index (ONI). Years are classified as El Niño (La Niña) based on a DJF value greater (less) than (-)1K and a FMA value greater (less) than (-)0.5K. This criteria results in 10 El Niño and 9 La Niña years. Table 4.1 specifies the categorical state of each year. We note that 8 of the 10 examined El Niño events fall into the category defined in (Johnson & Kosaka, 2016a) which exceed the convective threshold in the eastern Pacific (~.7K DJFM average SST anomaly in region ([5°S, 5°N];[160°W,120°W]). Diagnosis of nonlinear responses between El Niño and La Niña states are performed by regressing variables on contemporaneous COBE-SST2 monthly anomalies in the Niño3.4 region for each state (Niño/Niña) independently, and examination of the slope of the fit. All values are demeaned (base period 1951-2010) and linearly detrended prior to the regression.

We examine monthly values for every model field. Temporal resolution is set at one-month intervals to focus on the intraseasonal dynamical atmospheric response to ENSO events. As El Niño's effects are largely pronounced in boreal winter (Philander, 1989) and SST anomaly peaks in early winter (Neelin et al., 2000), we focus on November-April. We examine only monthly anomaly fields. For demonstrative purposes, in a few instances, the figures show an anomaly + the background climatology, these exceptions will always be indicated in the figure caption. We compute monthly anomalies, for every field, by subtracting the climatology, derived from the monthly mean using base-period 1951-2010. We then linearly detrend each time series to reduce potential effects of secular climate change. El Niño/ La Niña composites are formed by averaging the monthly anomalies of the years defined in Table 4.1.

When testing significance on ensemble mean fields, we utilize bootstrap methods by resampling all of the ensemble mean monthly anomaly years in the record 1000 times and examine the 5th and 95th percentile from the synthetic distribution. When examining the observational record, we utilize the composite and sampling methods described in Deser et al. (2017), where ENSO events are treated as exchangeable and uncertainty in the composite mean is determined by random sampling with resampling, again we sample the events 1000 times to determine confidence intervals.

4.2.3 Variable Selection

We examine the Northern Hemisphere ENSO response on upper-level and surface variables. 200-hPa GPH and wind anomalies are examined. 200-hPa GPH is associated with strong teleconnection modes (Mo & Livezey, 1986) between the tropics and the extratropics via changes in the large-scale atmospheric circulation in the Pacific-American and Atlantic-European sectors. The 200-hPa wind field, particularly in the Pacific jet region, undergoes a seasonal extension and intensification through early winter (November–January) as the Northern Hemisphere midlatitude baroclinicity increases, reaching its greatest zonal extent in February, and then retracts and weakens through March and the early spring (see Fig.1 of (Newman & Sardeshmukh, 1998)] . Additionally, the 200-hPa zonal winds are modulated in El Niño (La Niña) winter with a southward (northward) shift, intensification (reduction) in magnitude, and thus an increased (decreased) zonal extent (see Fig. 4.1 vector anomalies and Fig. 7 of Jiménez-Esteve & Domeisen, (2018)).

We examine two-meter temperature (T2m) and precipitation which are directly linked to the above mentioned 200-hPa GPH and wind fields. During El Niño years anomalous southerly winds advect warm marine air over northwestern North America while anomalous northerlies bring cooler continental air masses to the southeastern United States. A strengthened storm track increases precipitation over much of the southwestern United States, while leaving the northwestern United States anomalously dry. We observe the opposite relationship for La Niña seasons (e.g., Dai & Wigley, 2000; Deser et al., 2018; Jong et al., 2016; Ropelewski & Halpert, 1986). We are motivated to study these surface variables, in tandem with upper-level dynamics, in order to improve ENSO based S2S forecasting accuracy, which benefits vast swaths of North America's populations.

4.2.4 Observations

Observed 200-hPa GPH anomalies, are derived from monthly data from the National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al., 1996) available on a 2.5° x 2.5°grid. Daily average T2m data is utilized from the NCEP surface gaussian product which is available on the native T-62 gaussian grid (approximately 1.875° x 1.875°) over North America (Kalnay et al., 1996). Finally, monthly observed precipitation is obtained from the National Oceanic and Atmospheric Administration’s precipitation reconstruction over land dataset interpolated onto a 1°x 1° grid (Chen et al., 2002). Every observed data set spans years 1951-2019. All the data sets were downloaded from www.esrl.noaa.gov/psd/.

4.2.5 Rossby Wave Source & Wave Activity Flux

We examine the 200-hPa Rossby wave source (RWS) (Sardeshmukh & Hoskins, 1988).

$$\mathrm{RWS} = -\zeta_{\mathrm{a}}\mathrm{D} - \overline{\mathrm{v}}_{\chi} \cdot \nabla\zeta_{\mathrm{a}} \;, \qquad (4.1)$$

The RWS is derived from the barotropic vorticity equation and locates vorticity forcing. RWS is computed using the magnitudes of the divergence (D), the absolute vorticity (ζ_a), and the irrotational component of the wind ($\overline{v}_{\chi}$). RWS can be decomposed to 1) $-\zeta_a D$, a vortex stretching term, representing the effects of divergence on vorticity change, and 2) $\overline{v}_{\chi} \cdot \nabla\zeta_a$, the absolute vorticity advection by divergent flow, provided by regions of strong vorticity gradient (i.e. subtropical jet). To compute RWS terms we use the windpharms Python package (Andrew Dawson, 2016).

Following Takaya & Nakamur (2001), we use horizontal 200-hPa wave activity flux (WAF) to explore the stationary Rossby wave sources and wave propagation in the extratropics. WAF is independent of the wave phase and parallel to the local group velocity of stationary Rossby waves. Monthly anomalies are regarded as perturbations. The horizontal flux is given as

$$W = \frac{Pcos\phi}{2|\overline{U}|}\left\{\begin{array}{c} \frac{U}{a^2cos^2\phi}(\psi_x'^2 + \psi'\psi_{xx}') + \frac{V}{a^2cos\phi}(\psi_x'\psi_y' - \psi'\psi_{xy}') \\ \frac{U}{a^2cos\phi}(\psi_x'\psi_y' - \psi'\psi_{xy}') + \frac{V}{a^2}(\psi_y'^2 + \psi'\psi_{yy}') \end{array}\right. , \qquad (4.2)$$

where P, U, V, ψ' and a are pressure (scaled by 1000 mb), zonal climatological wind velocity, meridional climatological wind velocity, perturbation geostrophic stream-function, and the radius of the earth. Subscript x denotes the longitudinal derivative $\frac{\partial}{\partial \lambda}$, y the latitudinal derivative $\frac{\partial}{\partial \phi}$, λ the longitude, and ϕ the latitude, respectively.

4.2.6 Variance Patterns

To extract the leading patterns of variability, we perform empirical orthogonal function (EOF) decomposition on monthly anomaly 200-hPa GPH fields of the ensemble mean and the internal variability fields. Decomposition is performed on each calendar month independently, and the full ensemble, and internal variability fields utilize all 100 members. All EOF patterns are area-weighted by the square-root of the cosine(latitude), prior to decomposition. We express the orthogonal spatial fields as the point-wise regression of each time series on the 1-standard deviation change of the temporal principal component (PC) modes.

4.2.7 Signal-to-Noise & Potential Predictability

With 100 ensemble members at 60km resolution, d4PDF is unmatched in SN literature, and provides a more constrained estimate of the forcing. The deviation from the forced response, or ensemble spread, gives an approximation of the atmospherically derived internal variability.

We define context dependent signal and noise as anomalies from the ensemble mean and spread, respectively, consistent with (Kumar & Hoerling, 1998). We note that the structure of atmospheric internal variability can, and in general does, depend on SST forcing. This dependence has been the subject of a number of papers (e.g., Abid et al., 2015; P. D. Sardeshmukh et al., 2000; Schubert et al., 2001). Strictly speaking, it is not valid to refer to internal variability simply as "noise," as this implies that it is independent of the forcing. However, for brevity we refer to SST independent internal variability as noise. We henceforth derive the climate signal, for any variable x, as the monthly mean anomaly of the ensemble mean state for individual months in a particular year (a) and ensemble members (i).

$$\overline{X_a} = \frac{1}{100}\sum_{i=1}^{100} X_{ia}\,, \qquad (4.3)$$

The internal variability of the system is what remains in each ensemble, after removing the forced signal. Deviation from the ensemble mean (equation 4.4) represents the variability of the atmosphere determined by any perturbation unassociated with the lower boundary condition and radiative forcing.

$$\overline{Y_a} = \frac{1}{100}\sum_{i=1}^{100}\left(X_{ia} - \overline{X_a}\right)^2, \qquad (4.4)$$

Spatial averaged (denoted by $\langle\quad\rangle$) signal and noise root-mean-square (RMS) terms are defined as $\left\langle \overline{X_a}^2 \right\rangle^{\frac{1}{2}}$ and $\left\langle \overline{Y_a} \right\rangle^{\frac{1}{2}}$ respectively with signal-to-noise (SN) being a representation of the ratio of the aforementioned terms ($SN = \frac{\left\langle \overline{X_a}^2 \right\rangle^{\frac{1}{2}}}{\left\langle \overline{Y_a} \right\rangle^{\frac{1}{2}}}$). This is analogous to the conventional assessment of potential predictability derived from standard ratio of variance analyses (Chervin, 1986; Kumar & Hoerling, 1995; Rowell, 1998). SN is positive, and values greater than 1 imply that signal is greater than noise. Grid point RMS is area-weighted by the square-root of the cosine(latitude) for spatially averaged fields.

4.2.8 Kullback-Leibler divergence

To help verify the AGCM findings on observations, we utilize the Kullback Leibler (KL) divergence to assess the potential predictability of a conditioned distribution against climatology.

$$KL = \sum_{i=1}^{I} (p_i)\, log_2 \left(\frac{p_i}{q_i}\right), \qquad (4.5)$$

The KL divergence is borrowed from information theory and measures (in units bits) the extent to which a distribution q can be discerned from p (Kullback, 1997). Here, since p and q are the conditioned and climatological distributions, respectively, the KL-divergence can be interpreted as the extent to which a particular condition (i.e. Niño3.4 > 1K) informs the model prediction beyond climatology alone. Formally, it measures the number of excess bits needed to represent the examined variable when the condition is ignored (Cover & Thomas, 2006; MacKay, 2003).

The use of the KL divergence for assessing the PP of a forecast was proposed by (Kleeman, 2002). It has also been used to evaluate the potential forecast skill for multiple atmospheric variables (DelSole, 2004; Roulston & Smith, 2002) and to evaluate the effect of ENSO on North American T2m (Schamberg et al., 2020). In our analysis i will represent categorical anomaly states of below normal, normal, and above normal ($i \in \{1,2,3\}$), using the 33rd and 66th percentiles to quantize these states. Confidence intervals are determined by bootstrap with resampling all years in the record 1000 times and examine the 5th and 95th percentile from the synthetic distribution.

4.3 Atmospheric Response to ENSO

It is important to note that the following results are reflective of the AGCM chosen for this analysis, and the ENSO event selection criteria. The sensitivity and response to SST forcing vary across individual models, resulting in varied ranges of internal variance and predictable ENSO forcing in the teleconnections. However, models with larger signals tend to have larger noise, making PP vary weakly across models (Kang & Shukla, 2006). Kang et al. (2011) showed that synoptic transients in the Pacific basin comprise a large fraction of the signal and noise associated with the PNA.

During El Niño, the Pacific warm pool (and thus anomalous precipitation) shifts eastward. Forced by strong divergence at the upper-levels in response to this precipitation, the Northern Hemisphere has a forced anomalous GPH response (Deser & Wallace, 1990). Figure 4.1, shows the ensemble mean monthly composite of anomalous GPH in February and March by ENSO state. The leading mode of forced tropical precipitation (not shown) has a correlation to the Niño3.4 index of 0.92 for the cold season (NDJFM, correlation on seasonal mean). Anomalous tropical SSTs peak during December and fade through the remainder of winter and early spring. However, precipitation in the tropics is not controlled solely by SSTs, but modulated by the convective threshold (Gadgil et al., 1984; Graham & Barnett, 1987; Johnson & Xie, 2010). Due to higher climatological SSTs in combination with a retained El Niño SST signature in late winter and early spring, the upper-level divergence, and thus teleconnection, remains active well beyond the peak of tropical SST anomalies (Y. Guo et al., 2019; Xie et al., 2018). Hoerling et al. (2001) accredits the convective threshold as the source of a longitudinal shift in the North Pacific teleconnection between strong and weak ENSO events. The forced tropical precipitation response for every examined field peaks in February and remains anomalously strong into March. The extratropical GPH ENSO response is well studied, and a pressure pattern similar to the PNA, emerges as a stationary Rossby wave (Wallace & Gutzler, 1981a). This PNA-like pattern is characterized by a deepened Aleutian Low (AL), an increased Canadian High, and a deepened Florida low pattern extending into the Atlantic. A clear longitudinal shift is evident in the magnitude of the GPH anomaly in the late El Niño season

(Fig 4.1a and 4.1b white dot, see supplemental Fig. 4.1S for the full wintertime season anomalies). Additionally, the ENSO forced North Atlantic response (Honda et al., 2001; Honda & Nakamura, 2001) is not apparent until February/March, and shows a weaker anomalous response to ENSO forcing than the PNA. DJFMA ensemble mean monthly composite of anomalous GPH by ENSO state are shown in the supplemental material (Fig. 4.1S).

ENSO anomalous upper-level winds are mostly geostrophic as evident by the 200-hPa anomalous wind vectors parallel to the 200-hPa GPH anomalies. The large-scale (synoptic) Pacific trough (ridge), is thus able to bring warm marine (cool polar) air into the North American West, altering the surface temperatures (Z. Q. Zhou et al., 2014) during an El Niño (La Niña) event. Additionally, there is a distinct latitudinal shift in the subtropical jet (Fig. 4.2) which migrates from north-to-south (~5° as measured by the maximum zonal winds) throughout the ENSO season (Fig. 4.2). It is observed that due to the deepened (shallowed) Aleutian pressure anomaly, the jet stream is magnified (diminished), moved southward, and extended (contracted) across the Pacific during El Niño (La Niña) (Norris, 2000) (Fig. 4.1 and Fig. 4.2). This alters the longitudinal location of the jet exit region, the region of highest variability (Athanasiadis et al., 2010).

4.3.1 Rossby Wave Source

The ENSO RWS is depicted in Figure 4.2. During El Niño (La Niña), anomalous divergence (convergence) is produced from deep tropical convection, flow peaks at the edges of the heating region, resulting in anomalous convergent (divergent) regions in the subtropics. In the North Pacific, the position of the jet anchors the source term and often determines the major Rossby wave response of the North Pacific (Hakim, 2003). Under the influence of the seasonal jet cycle, and the evolving ENSO precipitation signal (Fig. 4.2, inset), the peak response shifts east and west throughout the season, leading to a shifting center of action in the extratropical GPH response (Fig. 4.1, white dot). Interestingly, the teleconnection pattern shifts 7.5° (10°) from east to west alone between February and March of El Niño (La Niña) years, owing to the major contraction of the mean subtropical jet (Fig. 4.1). To first order, the monthly El Niño and La Niña response is symmetric, but the asymmetrical components, outside the strongest response regions, lead to important dynamic differences (Feng et al., 2017). The most notable asymmetry occurs in March where the El Niño composites show an extended positive RWS term that spans most of the Pacific while the La Niña counterparts RWS is relatively muted (Fig. 4.2g & 4.2h). Generally, asymmetry (in amplitude and position) is observed between the cold and warm composites, notably in the eastward shifted and amplified anomalous GPH and the Pacific extension of the RWS term (Figures 4.1 & 4.2). Zhang et al., (2014) and Feng et al., (2017) have recently reexamined the asymmetrical components the ENSO response and found it is driven by the background state of the atmosphere and plays an important role in how ENSO affects the North American climate.

Figure 4.3 shows the forced RWS and its components for each ENSO category in the characteristic RWS anchoring region for boreal winter (Andrew Dawson et al., 2010; Nie et al., 2019) (Lat: [25°N,40°N] Lon: [145°E, 155°W], inset map). We find a significant nonlinear RWS response to SST forcing between El Niño/La Niña seasons, with an increased sensitivity in El Niño periods (Fig. 4.3d). The nonlinearity is demonstrated by the slope of a linear fit calculated by regressing RWS on corresponding positive and negative Niño-3.4 anomalies (respectively) utilizing every monthly value in DJFM. The difference of these slopes is significant at the 10% level. We find no significant difference in the magnitude of the Niño-3.4 anomaly between El Niño/La Niña in any month over December–April (DJFMA; Fig. 4.3e).

This nonlinear response to categorical ENSO states is a well-noted phenomenon (e.g., Hoerling et al., 1997, 2001; Hoerling & Kumar, 2002; B Jiménez-Esteve & Domeisen, 2019; Johnson & Kosaka, 2016; Trascasa-Castro et al., 2019), although the exact source of the nonlinearity in the extratropics is still subject to debate (Frauen et al., 2014). Many studies point to the convective precipitation response to tropical SST as a contributing factor (e.g., Chung & Power, 2015; Hoerling et al., 2001). We find the observed RWS nonlinearity is alleviated somewhat (but remains significantly different), when regressing the RWS on tropical precipitation (not-shown). The nonlinear response is seen in the magnitude of difference in the vortex stretching term for either ENSO state. The anomaly difference of the RWS term for cold and warm states is greatest in March, where both ζ_a and D remain highly anomalous in the warm phase (Fig. 4.3c). For both phases of ENSO, RWS anomaly peaks in February, with near equal magnitudes in January and March of El Niño years (Fig. 4.3c). The absolute vorticity advection opposes the vortex stretching term, thus weakening the total RWS in DJF. However, the magnitude of $\overline{v_\chi} \cdot \nabla\zeta_a$ decreases back to climatology in March, diminishing the March RWS drop from the February RWS peak (Fig. 4.3b,c). April sees a near full decay of the RWS. We observe an asymmetric ENSO response in every examined monthly ensemble mean anomaly variable (GPH, RWS, divergent wind, etc.).

4.3.2 Wave Activity Flux

WAF is diagnosed using equation 4.2 to explore month-to-month Rossby wave propagation. Figure 4.4 shows the forced monthly composites of ENSO WAF (vector), the anomalous 200-hPa GPH (colorfill), and the anomalous RWS (contour). In both ENSO states, WAF emanates from the strong RWS at the exit of the Pacific jet through the Aleutian Low (AL) toward North America. The December El Niño Canadian limb of the teleconnection pattern shows a stronger anomalous signal than the corresponding Niña composite. By January in the El Niño season a canonical wave train has emerged, with the classic 4-pole PNA pattern. The January La Niña composite shows a strong AL signal but mostly insignificant WAF over land. WAF peaks in February, with a fully developed wave train pattern in both ENSO phases. This corresponds with the strongest PNA-like anomaly GPH pattern. The maximum Florida limb of the teleconnection pattern, for both WAF and GPH anomalies, is observed in March of Niño seasons. The Niña pattern has diminished greatly by March and both WAF and GPH appear relatively weak in April. Across the season the WAF shows an extreme asymmetry between El Niño and La Niña, varying with the asymmetric GPH anomalies. Interestingly, the El Niño/La Niña pathways appear different in Rossby wave propagation for the Florida low GPH anomalies with El Niño WAF showing a more southerly route (consistent with, Seager et al., 2010).

The Icelandic Low (IL) (~[64°N,30°W]) undergoes a seasonal shift in phase expression between early (ND) and late season (FM) ENSO states. This is a well-studied shift that is robust in the observational record (e.g., King et al., 2018) though climate models typically do not represent the early season mechanism well (Ayarzagüena et al., 2018). D4pdf captures the early-season IL anomalies in the ensemble mean (Fig. 4.4a,b), which stem from increased precipitation anomalies in the Gulf of Mexico leading to enhanced anomalous RWS at 250-hPa (Fig. 4.2a,b). A late-season emergence of an anomaly in the IL occurs in February/March of El Niño years. In February a wave train emanates as an extension of the PNA-like pattern, extending the Canadian High and Florida Low into the Atlantic. A large body of literature finds the late season ENSO influence on the IL is due to changes in stratospheric circulation (Trascasa-Castro et al., 2019 and references therein). We observe an additional tropospheric pathway with significant RWS terms stemming from increased precipitation (precipitation anomaly not shown) in the Gulf of Mexico and Florida region interacting with the Atlantic jet, which is energized and extended in El Niño years (Fig 4.2e,g). We see a particularly nonlinear RWS and WAF response between March of La Niña and El Niño years in this region. With El Niño leading to the shallowing of the surface IL anomalies, and the negative phase of the NAO. This late-season development of the IL, and peaking of the Florida low PNA-like pattern was also observed in multiple studies and referred to as the Aleutian-Icelandic low see-saw index (AII) (e.g., Honda et al., 2001, 2005; Honda & Nakamura, 2001).

4.3.3 Additional Sources of ENSO Forced Extratropical Waves

Though we focus on the dispersion of Rossby waves excited by tropical heating, extratropical waves are additionally generated and anchored due to barotropic energy conversion from the subtropical jet deceleration $\left(\frac{\delta\bar{U}}{\delta x} < 0\right)$ in the jet exit region and synoptic scale transient eddy vorticity fluxes. Both mechanisms are modulated by ENSO. Jet deceleration allows waves to effectively extract kinetic energy from the zonally asymmetric climatology, via an energy transfer from the climatological stationary eddies to the anomaly (Athanasiadis et al., 2010; Branstator, 1989; Feldstein, 2002; Simmons et al., 1983). The anomalous synoptic transient activity along the Pacific storm tracks -- which is extended eastward to the jet exit region during El Niño years (Harnik et al., 2010; R Seager et al., 2010) -- produces the seasonal-mean transient eddy vorticity flux convergence anomalies that reinforce the local signals of seasonal-mean circulation anomalies (Held et al., 1988; Straus & Shukla, 1997). Moreover, the downstream propagation of transient eddies from the Pacific to the Atlantic basin provide a tropospheric pathway for NAO related GPH anomalies (Jiménez-Esteve & Domeisen, 2018; Ying Li & Lau, 2012) during ENSO.

4.4 Signal vs. Noise

We examine the leading mode of variability in two categories: the internal variability, and the forced response (Fig. 4.5, column I and II, respectively). The leading mode of variability accounts for ~20-30% (month dependent) of the full variability (not shown), and both internal variability and the forced response have loadings in the PNA regions. However, distinct differences are observed. Note that the internal variability (Fig. 4.5, column I) patterns have a far southward extent of the Canadian high-pressure system that largely covers the western United States, and the forced response has a linked low pressure system between the AL and the Florida low (Fig. 4.5, column II). The forced pattern more closely resembles the El Niño composites (see loading locations of Fig. 4.4) and the anomaly strength in the principal component agrees with this finding (not shown).

Although the NAO loadings are present in the internal signal throughout boreal winter, the forced negative NAO signal does not emerge until February. The NAO, with the exception of very low frequency forcing signals, is not necessarily strongly forced by an oceanic mode (Stephenson et al., 2000). However, ENSO forced PNA/NAO patterns/signatures, can be spurred by the PNA's advection of air masses which lead to baroclinic waves forming the North Atlantic storm track (Pinto et al., 2011). By this mechanism, a negative interannual correlation between the intensities of the Aleutian and the Icelandic lows reaches a value of ~-0.7 between the indices averaged from February to Mid-March in observations. (Honda et al., 2001; Orsolini et al., 2008). During February and March, the leading forced modes (Fig. 4.5h,j) show loadings consistent with a negative NAO phase (Barnston & Livezey, 1987) that is correlated with the

Niño3.4 signal (Huang et al., 1998) and peaks in the late winter/early spring. There has been much work on the dispersive characteristics of climate models and seasonal-to-multiseasonal predictability of the NAO (e.g., Scaife & Smith, 2018; Shi et al., 2015; Weisheimer et al., 2019). The NAO fraction of variance is low compared to the forced ensemble counterpart in every month. However, it has been observed that the NAO is more predictable (in a signal-to-noise framework) than climate models typically represent it to be (Scaife & Smith, 2018; Siegert et al., 2016; Zhang & Kirtman, 2019) and a model post-processing variance adjustment (Smith et al., 2020) could show a more enhanced variance fraction of the full ensemble in the ensemble mean.

The leading mode pattern accounts for ~40-70% of the forced response variance and its principal component correlates with the Niño3.4 anomaly index at ~.65-.90, month dependent. The DJF average fraction of variance in the leading mode (~58%) agrees well with previous studies of ENSO forced variance (e.g., 53%: Kumar et al., 2005; 56.2%: Zhang et al., 2016). However, the forced PNA-like pattern is particularly dominant in FM (~66% of variance) and correlates strongly with the Niño3.4 index (~.9).

4.4.1 ENSO modulation of Internal Variance

Motivated by the important role the AL plays in modulating North American weather (e.g., Gibson et al., 2020), 200-hPa GPH signal and noise (Fig. 4.6a) over the North Pacific (Lat: [30°N,60°N], Lon: [165°E,130°W]) is diagnosed. Climatologically, GPH noise is greatest during boreal winter, peaks in February, and lowest in summer (Fig. 4.6a, solid black line). Internal variability is significantly different from climatology (though weakly) in February of La Niña years (Fig. 4.6a, solid blue line). We find a modulation of the GPH noise conditioned on the ENSO state (Fig. 4.6a). With adjustments of up to ~10% (by percentage difference) of modulated RMS across DJFM (Fig. 4.6a) in El Niño. This is a similar finding (though ~7%< (Abid et al., 2015)) in El Niño years. Abid et al. (2015) attributed the modulation of noise in ENSO years to extratropical transients, and not to increased tropical precipitation variability (as tropical variability, which is proportional to SST magnitude, increases (decreases) in El Niño (La Niña) years (Peng & Kumar, 2005)). The forced AL peaks in February of El Niño years diminishing slightly through March. Owing to decreased noise and increased signal in February and March, the regional SN approaches 1 in March.

Figure 4.7 displays the FM ENSO spatial modulation of the internal variability via monthly composites of GPH RMS noise with climatological noise depicted in solid contours. ENSO modulation is most apparent in JFM, with a peak in February. Noise modulation becomes effective for PP in March of El Niño years, as noise climatologically decays in concert with an El Niño mean shift (Fig. 4.6a & Fig. 4.7 , black contour). The internal variability is largely decreased (increased) during El Niño (La Niña), with the exception of the jet exit region, which is the highest source of variability in either ENSO state. The ENSO effect on internal variability is stronger in the warm phase than in the cool phase (See Fig 4.7c vs. Fig. 4.7d). La Niña noise in the northwest Pacific is significantly increased in DJF, peaks in February, and decays back to climatology by March (Fig. 4.6a). Abid et al. (2015) found similar diminished noise in the extratropical PNA region during El Niño events. Abid et al. (2015) point to the noise intensification associated with barotropic instability in the PNA region as a possible driver (Branstator, 1985; Simmons et al., 1983). Eastward (Westward) extensions (contractions) of the zonal jet are co-located with decreased (increased) noise in the western Pacific and over the southern United States. The areas of increased (decreased) El Niño (La Niña) noise ([~40°N,150°W]) are directly related to the shift in the Pacific jet exit region (Fig. 4.7). DJFMA

ENSO spatial modulation of the internal variability via monthly composites is shown in the supplemental material.

4.4.2 Signal-to-Noise Ratio

Using SN as a proxy for PP (Sardeshmukh et al., 2000), we examine 200-hPa GPH, T2m, and precipitation SN during ENSO events. Area averaged SN for GPH, T2m and precipitation is shown (Fig. 4.8a,b,c, respectively) in the PNA sector (defined here as [25°N, 70°N], [155°E, 60°W]). T2m SN is only accounted for over land. FMA GPH SN show a significant difference between ENSO categories (Fig. 4.8a). Temperature and precipitation show a significant difference in March and April (Fig. 4.8b). El Niño/La Niña Precipitation is significantly different in March (Fig. 4.8c). We observe a statistically significant (10%) nonlinearity (diagnosed as described above) of month-to-month SN across all variables conditioned on the Niño3.4 anomaly (Fig. 4.8d, T2m and precipitation not shown). Figure 4.9 shows the monthly composites of SN across North America for GPH and T2m. The teleconnection most dominantly affects T2m in El Niño in northwestern North America (NWNA), through the advection of warm marine air. The NWNA T2m SN increases in January, peaks in March, and remains elevated during April, shifting northward throughout the season. We theorize the April NWNA T2m SN to be a manifestation of a decreased snowpack from the previous month's warm temperature anomalies (Zhang et al., 2011). The American southeast T2m is also affected by the southernmost limb of the PNA pattern. Northern Mexico and Florida show the most consistent, and significant SN, which peaks in March of El Niño years. The temperature SN patterns match the Deser et al., (2018) observed and simulated ENSO anomaly seasonal composites well, but they occur in distinct months in boreal winter, rather than showing a full seasonal shift. This could indicate that averaging over a season acts to mute the forced ENSO signal. Additionally, La Niña SN is generally weaker in the T2m field, in agreement with diminished dynamic model forecast skill when compared to El Niño seasons (Chen et al., 2017).

GPH SN is greater over the PNA region, in EL Niño than La Niña, showing patterns which match the forcing signal (Fig. 4.4). Figure 4.10 shows the monthly composites of SN across North America for GPH and precipitation. La Niña SN only peaks in the southern half of the AL region (Fig. 4.10f, h), where internal variability is low (Fig. 4.7). GPH SN does not peak in the IL region, though in observations, Northern Canada and the Eastern U.S. show a significant shift in temperature anomalies (see, Deser et al., 2018). This is an indication that the northern limb of the PNA teleconnection response in d4PDF is potentially overdispersive.

We detected low precipitation SN across the ENSO seasons (Fig. 4.10). FM SN shows an emergent reflection of the well-studied meridional dipole of ENSO precipitation over western America (Dettinger et al., 1998). GPH patterns are often represented well (Flato et al., 2013), and an increased northern continental SN value in El Niño could be indicative of SST forced anomalous GPH patterns steering precipitation events away from the NWNA to impact more southerly locations. The largest source of SN in both ENSO states is in the Eastern Pacific (~[30°N,135°W]), highlighted by Zhou et al. (2014) , as enhanced (diminished) westerlies steer extratropical storms to a more southerly (northerly) position during El Niño (La Niña) causing increase (decreased) precipitation. Additionally, northern Mexico and Florida show a significant SN, magnified in La Niña years. Previous studies have shown significant influence of tropical SST anomalies on North American precipitation variability (e.g., Burgman & Jang, 2015; Dai, 2013; Meehl & Hu, 2006; Richard Seager et al., 2005). Accurately representing precipitation involves heavily parameterized processes, and linking to surface fields (topography, coastline, vegetation), making it difficult when compared to T2m representations.

4.5 ENSO Potential Predictability & Observations

Using SN as a proxy for PP, we have demonstrated month-to-month ENSO driven changes in GPH, T2m, and precipitation in the Database for Probabilistic Description of Future Climate Change (d4PDF) model ensemble. We now verify these findings on observations. The following analysis is performed on all years shown in Table 4.1 and extended to include years (2010-2019) which are beyond the d4PDF record. Apart from ’15/'16 (El Niño) and '10/'11 (La Niña), every added year is ENSO Neutral.

Figure 4.11 shows the d4PDF ensemble spread, d4PDF forced ensemble mean, observed composite mean, and every observed value of the PNA (Fig. 4.11a,b) and the Aleutian-Icelandic low see-saw index (AII) (Fig. 4.11c,d). The PNA is defined at 200-hPa by the four-point index described in (Wallace & Gutzler, 1981a) and is constructed using standardized anomaly time series at each point. The resulting index is normalized by the standard deviation of the combined DJF values. The AII is defined at 200-hPa in the characteristic regions described in (Honda et al., 2005), and is calculated as the normalized anomalous IL intensity subtracted from the normalized anomalous AL intensity. Each index uses values from 1951-2010 (the d4PDF period of record) to form the normalization climatology. The model mean and spread in the PNA/AII match the observed values well. PNA composite mean displacement for d4PDF and observations both peak in March of El Niño years with a near-zero anomaly shift in November, December and April. In agreement with the d4PDF, La Niña has a generally weaker mean shift and sits well within the d4PDF model spread. La Niña signal fades in March/April. The AII index observed mean sits well within the spread of the d4PDF model. Using the IL index alone gives a good fit between model spread and observations, but a dampened magnitude (not shown), compared to the AII index.

Figure 4.12 shows the observed monthly composite of anomalous precipitation and 200-hPa GPH from December to April by ENSO state. We now list noticeable similarities between the observed monthly anomalies (Fig. 4.12) and the SN relationships displayed in Fig. 4.10. 1) The significant AL and Florida Low GPH anomaly matches nearly exactly for each ENSO state across the full season. In La Niña the AL GPH composite is co-located with the low GPH noise anomaly shown in d4PDF (Fig. 4.7, column II)(~[40°N,150°W]). 2) December shows very little GPH or precipitation anomaly signal especially affecting the North American west coast. 3) La Niña composites show less significant anomaly than the El Niño counterpart, in precipitation and GPH. 4) The Gulf of Mexico and Florida are particularly attenuated in La Niña. 5) March El Niño precipitation extends farther into the continental United States. We note that specific months magnify specific anomaly loading locations throughout the ENSO season highlighted in the seasonal composite seen in Deser et al., (2017, 2018). For completeness, we show the same figure but for observed monthly temperature anomalies in the supplemental (Fig. 4.3S).

The largest precipitation pattern discrepancy occurs in the western United States shown in JFM, which is shifted into the eastern Pacific in the d4PDF SN (Fig. 4.10 vs. Fig. 4.12). Additionally, there is a clear model bias associated with the high-pressure limb of the PNA pattern in northeast Canada. This could be an indication of d4PDF overdispersiveness of the northern limb of the PNA pattern across the ensemble members, and a lack of forcing in the early season, which is consistent with the findings of Scaife and Smith (2018) and Smith et al. (2020) that the NAO is more predictable than climate models typically demonstrate.

To test the utility of the PNA-driven changes in SN on observations, we adopt an additional potential predictability metric developed from KL divergence (KLPP) to show the ENSO forcing on temperature predictability. We build distributions following the probabilistic framework in Johnson et al. (2014) and examine the T2m distributions for the weekly averaged temperature anomaly shifts conditioned on an ENSO state. These calculations are performed using observations, and not the d4PDF model. An observation is quantized into one of three divisions (below normal, normal, above normal), based on the highest probability tercile determined by the state of ENSO. The KL divergence is then computed [Eq. (4.5)]. We again show a monthly granularity to observe the evolution of potential forecast skill.

To illustrate, at every grid point we develop a climatological weekly temperature distribution across all years, using average weekly T2m observations. We use the 33rd and 66th percentiles to quantize the anomaly value into categorical states $T \in \{below\ normal, normal, above\ normal\}$. Therefore we have threshold values to divide anomalous temperature into equally probable categories $\left([P(below\,normal), P(normal), P(above\,normal)] = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]\right)$ for a climatological distribution (q, Eq. 4.5). Next, we examine the anomaly distribution conditioned on ENSO state against the climatological 33rd and 66th percentiles thresholds, and determine the categorical probability of each tercile of the conditioned distribution (e.g., $[P(below\,normal|Niño), P(normal|Niño), P(above\,normal|Niño)] = [0,0,1]$), where the probability is determined by the number of observed categorical states ([below, normal, above]) divided by the total number. The conditioned probability distribution (p, Eq. 4.5) is then compared to the climatological probability distribution (q) using equation 4.5. This is very similar to an evaluation of Climate Prediction Center's probabilities of tercile-based category product, and demonstrated to be an effective distance metric for ENSO effects on T2m (Schamberg et al., 2020). The KL divergence is a quantification of the information lost if a forecaster were to ignore that it was an ENSO year, and can be loosely thought of as a quantification of the forcing of the anomaly probability. Encouragingly, all of the KLPP patterns resemble the seasonal anomalies presented in (Deser et al., 2018).

Figure 4.13 shows the monthly T2m KLPP for DJFMA in respective El Niño (column I) and La Niña (column II) seasons, and the composite observed GPH anomaly (contour). KLPP is stippled for values significant at the 10% level. Largely, the observed KLPP matches the SN relationships displayed in Fig. 4.9. In agreement with the d4PDF, the results of the KLPP divergence indicate the following monthly patterns for T2m: 1) El Niño KLPP is larger than La Niña, 2) little to no KLPP exists in December for either El Niño or La Niña, 3) KLPP begins to develop over Mexico in January of El Niño and is strongest across the southern half of North America in FM, 4) January and February of La Niña years see a peak in the KLPP in the Gulf of Mexico and Florida region, 5) reliable NWNA KLPP emerges in February and peaks in March, and 6) KLPP in NWNA shifts northward in April of El Niño years and KLPP vanishes in April of La Niña years.

Differences between d4PDF SN and T2M KLPP exist. We note that these could be due to the internal variability of the atmosphere and the limited number of observations or attributable to d4PDF model biases. In observations there is a clear shift of the T2m associated with the high-pressure limb of the PNA pattern in northeast Canada in January (Fig. 4.13c). This could be an indication of d4PDF overdispersiveness of the northern limb of the PNA pattern across the ensemble members, and a lack of ENSO forcing in the early season. This is consistent with the findings of Ayarzagüena et al. (2018) and Smith et al. (2020) which show that the NAO is more predictable than climate models typically demonstrate. Additionally, February and March of La Niña years show a distinct KL divergence spike centered over Oregon/Washington. Figure 4.9a and 4.9b show very little SN in this region. This could be attributable to an over dispersion of Canadian limb of the PNA in La Niña seasons in the d4PDF, as a distinct trough is shown in March of observations (Figure 4.12h).

The presented KLPP has implications for the contemporaneous signal between tropical ENSO SSTs and North American T2m or precipitation. However, the conditional distributions developed are dependent only on the knowledge of the contemporaneous ENSO state and the present month. The correlation between February and March Niño-3.4 indices is 0.96 and the correlation between December and March is 0.87; thus, these findings have serious implications for monthly and seasonal forecast skill.

4.6 Summary and Discussion

Leveraging an atmosphere model ensemble, we examine the Northern Hemisphere's forced response to El Niño–Southern Oscillation (ENSO). We diagnose signal-to-noise (SN) relationships for 200-hPa geopotential height (GPH), 2-m temperature (T2m), and precipitation as a function of the amplitude and phase of tropical Pacific SST forcing, and amplitude of the natural variability at a monthly temporal resolution. Further, we verify the model findings by examining the potential predictability (PP) of those surface variables developed from observations with implications for subseasonal-to-seasonal (S2S) forecasting.

The forced teleconnection is examined with Rossby wave source (RWS) and wave activity flux analyses. The forced pattern is generally nonlinear and asymmetric with respect to categorical ENSO states, which has been noted in multiple studies (e.g., Abid et al., 2015; Feng et al., 2017; Johnson & Kosaka, 2016; Wenjun Zhang et al., 2019). The RWS cold season vortex stretching term is of weaker magnitude than it's warm phase counterpart; resulting in nonlinear Rossby wave forcing. The forced 200-hPa GPH is a consequence of this nonlinearity with warm events showing an increased amplitude as compared to their cold phase counterpart.

Appreciable dynamic evolution occurs on monthly timescales and is potentially an important component to increasing S2S forecast skill. The forced response evolves temporally across the ENSO season (November-April), due to differences in monthly strength and location of the tropically driven upper-level divergence and the Pacific jet. The combined effect of persistent forced signal and decreased atmospheric noise results in February and March showing the greatest PP in every examined variable, and December showing weak to no PP. The dominant signal for both the internal variability and the forced response is a Pacific North American (PNA) like pattern (Wallace and Gutzler, 1981). The pattern is particularly robust during February and March of warm phase events.

An open question remains around the forced El Niño PNA GPH anomaly in March and January. Though the RWS is nearly identical (Fig. 4.3), the March GPH anomaly is greater (Fig 4.4). This phenomenon is observed in other AGCM SN studies (e.g., see Fig. 3 in Honda et al., 2005). Jiménez-Esteve & Domeisen (2019) show a decrease in transient eddy forcing during March, therefore barotropic energy conversion from the jet exit region could be a potential pathway. The exact mechanism is not clear, and requires focused research.

Zhou et al., (2014) notes that in a warmer climate, the large-scale 200-hPa pattern associated with El Niño shifts eastward, associated with an eastward shift of the tropical precipitation pattern. Importantly, the Pacific maximum of precipitation, coincident with the jet exit region and the PNA teleconnection pattern (~40°N, 140°W), is projected to shift eastward in a warmer climate, impacting the western coast of North America. This coincides with the peak SN region in ENSO events (Fig. 4.10d-g) and could lead to an increase in skill for North American West Coast precipitation prediction. Additionally, the changes in circulation lead to an eastward and southward shifted temperature anomaly due to an increase in warm advection by the Aleutian low westerlies. These patterns imprint on late-season peak SN areas (Fig. 4.9g) and could increase forecast skill of temperature anomalies over large swaths of North America. This necessitates an intraseasonal exploration of the changes of ENSO SN in a warmer climate.

Month-to-month ENSO dynamics and the background seasonal cycle lead to distinct teleconnection patterns. These patterns result in a myriad of signal-to-noise relationships that can be exploited for forecasting. New interest has arisen for statistical models (i.e., deep learning) for S2S forecasting owing to recent computational advances, algorithmic toolbox development, and successes in the Earth sciences (e.g., Abid et al., 2015; Y. G. Ham et al., 2019). Proper training data periods must be utilized to capture these relationships and more skill may be gleaned from intraseasonal rather than seasonal algorithm development. This study joins Ayarzagüena et al., (2018) and King et al., (2018) in warning against seasonal mean analysis due to a shifting ENSO teleconnection and noise background state.

4.7 Acknowledgement

This work is supported by the National Science Foundation under Grant AGS 1637450 and by the U.S. Army Corps of Engineers (USACE)-Cooperative Ecosystem Studies Unit

(CESU), Grant W912HZ-15-2-0019 and the California Department of Water Resources Atmospheric River Program Phase 1 and 2 (Awards 4600010378 TO#15 Am 22 and 4600013361 respectively).

Chapter 4, in full, is a reprint of the material as it appears in Chapman, W. E., Subramanian, A. C., Xie, S. P., Sierks, M. D., Ralph, F. M., & Kamae, Y. (2021). Monthly Modulations of ENSO Teleconnections: Implications for Potential Predictability in North America. Journal of Climate. The dissertation author was the primary investigator and author of this paper

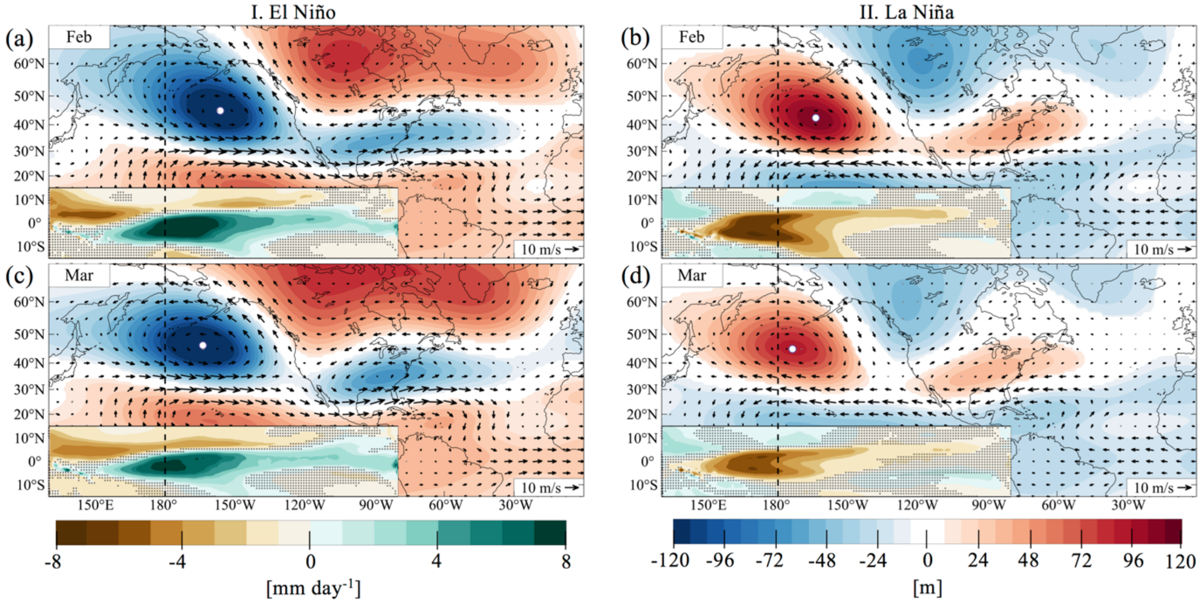

Figure 4.1 February and March ensemble mean monthly response to (column I) El Niño and (column II) La Niña: composites of anomalous mean 200-hPa geopotential height (colorfill), 200-hPa winds (vector), and tropical precipitation (inset 15°S–15°N, 130°E–80°W) for El Niño and La Niña. Anomalous geopotential height and black wind vectors are shown for significant locations. Insignificant wind vectors are shown in gray. Insignificant tropical precipitation is stippled. Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentile of the synthetic distribution. The white dot shows the Aleutian low center of action.

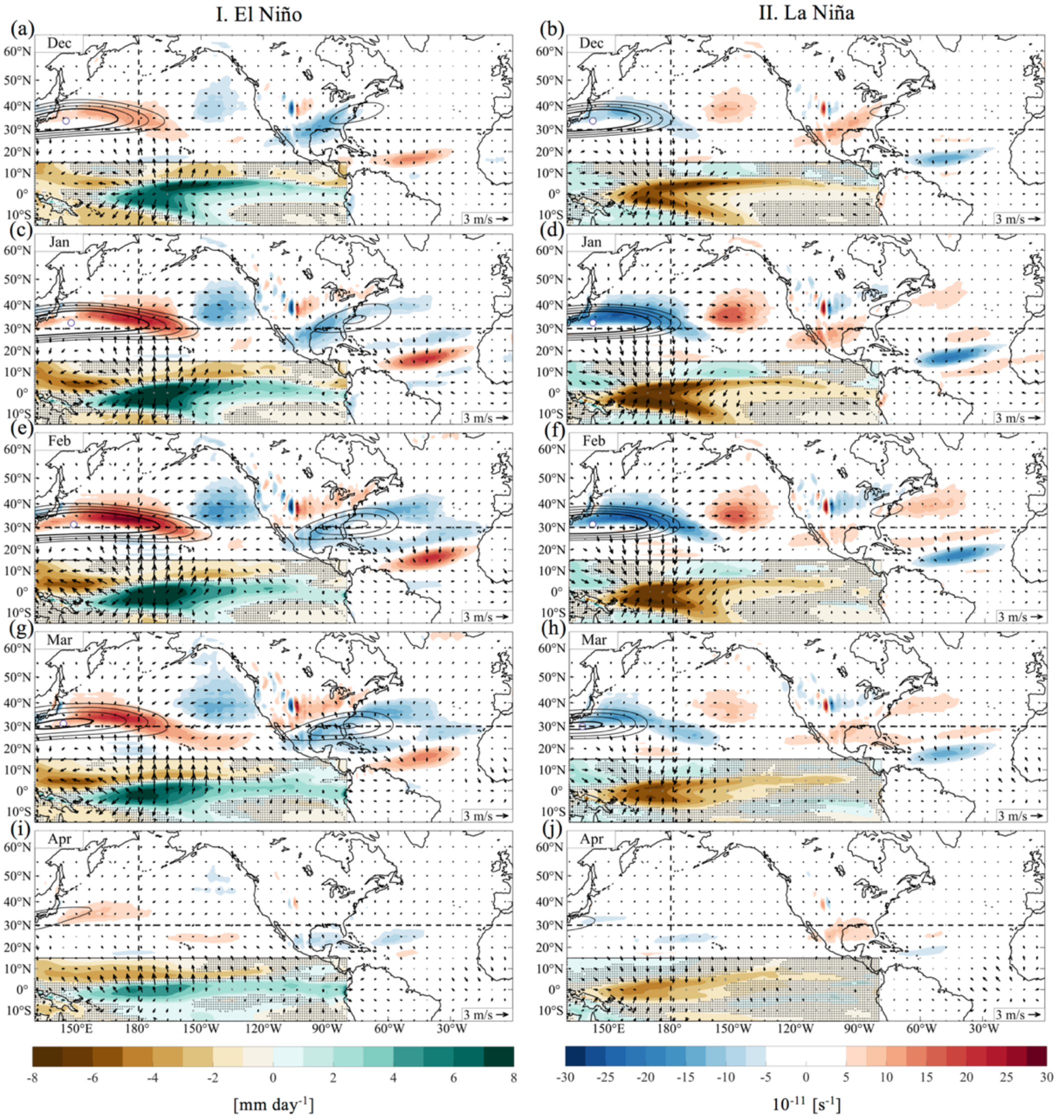

Figure 4.2 Ensemble mean monthly (December–April) response to (column I) El Niño and (column II) La Niña: composite 200-hPa anomalies of Rossby wave source (colorfill), and anomalous divergent winds (vector), along with anomalous tropical precipitation (inset 15°S–15°N, 130°E–80°W) and 200-hPa zonal wind (climatology + anomalies; 45, 50, 55, and 60 m s^{-1}shown with black contours; 60 m s^{-1} is shown in bold). Significant Rossby wave source is shown. Insignificant anomalous tropical precipitation is stippled. Significant vectors are shown in black and insignificant in gray. Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentiles of the synthetic distribution. The white dot shows the location of maximum 200-hPa zonal wind.

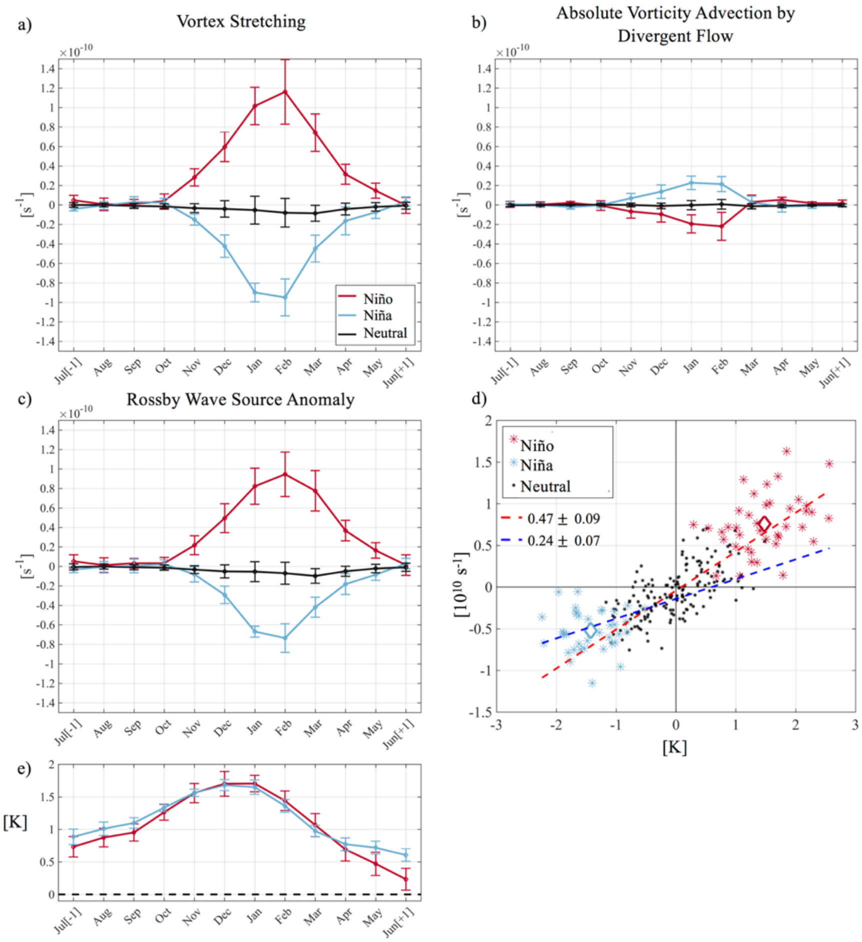

Figure 4.3 (a) Vortex stretching (VS). (b) Absolute vorticity advection by divergent flow (AVA). (c) Rossby wave source (RWS) anomaly for El Niño (red line), neutral (black line), and La Niña (blue line) year (categorized in Table 4.1) composites. (d) Average RWS anomaly index with respect to the Niño-3.4 anomaly for individual months DJFM. Red (blue) markers indicate El Niño (La Niña) years. Diamond marker indicates the class (Niño/Niña) composite mean. The dashed red (blue) line indicates the linear fit calculated using every positive (negative) Niño-3.4 anomaly. The slope of each line is shown with 2σ uncertainty determined by bootstrap, with resampling across all years, 1000 times. (e) Composite El Niño and La Niña (negative) anomaly SST in the Niño-3.4 region for years specified in Table 4.1; 5th and 95th confidence intervals are shown, determined by bootstrap with resampling 1000 times. VS, AVA, and RWS are area averaged in the region 25°–40°N, 145°–155°W.

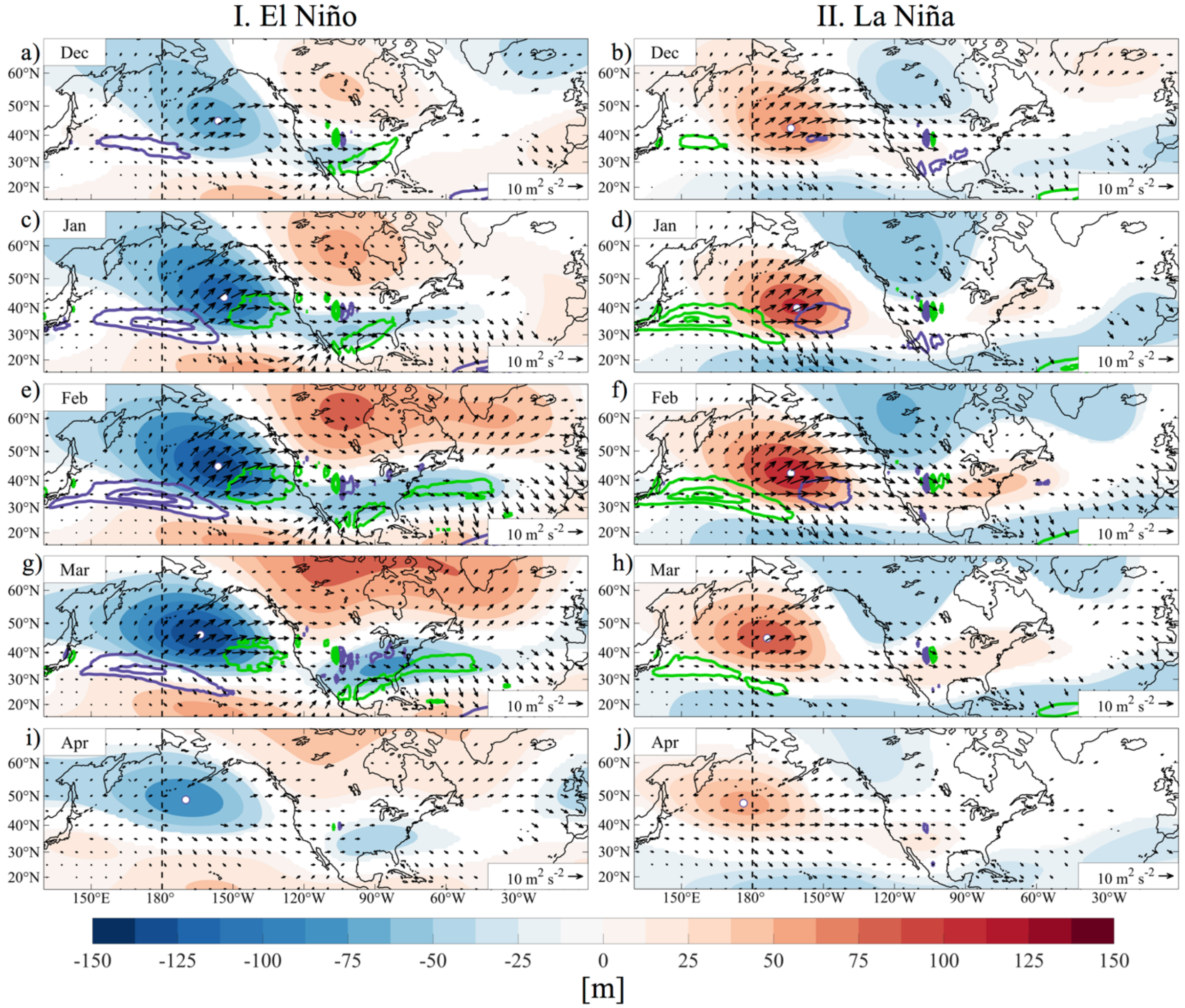

Figure 4.4 Ensemble mean monthly (DJFMA) response to (column I) El Niño and (column II) La Niña: 200-hPa TN wave activity flux (WAF) composite (vector), 200-hPa geopotential height anomaly (colorfill), and anomalous Rossby wave source [contour; purple (positive), green (negative); intervals ± at 10, 20, and 25 × 10^{-11} s^{-1}]. Only significant geopotential height is shown. Significant WAF vectors are shown in black. Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentiles of the synthetic distribution. White dot shows the Aleutian low center of action.

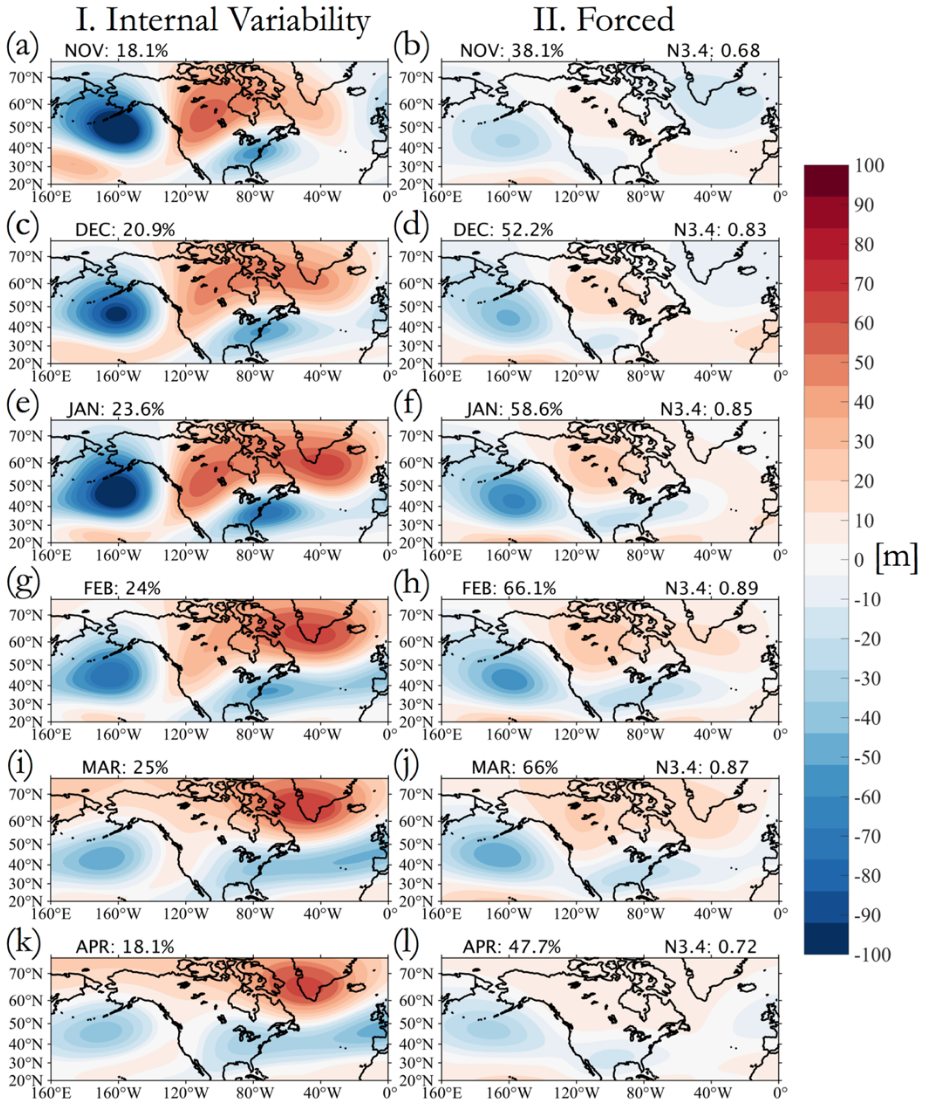

Figure 4.5 (column I) Internal variability and (column II) forced leading EOF mode of 200-hPa atmospheric geopotential height variability, calculated for each month individually, with percentage variability explained by this mode, for each month, and the correlation of the principal component to the concurrent Niño-3.4 anomaly index (at top right). PCs are normalized to unit variance.

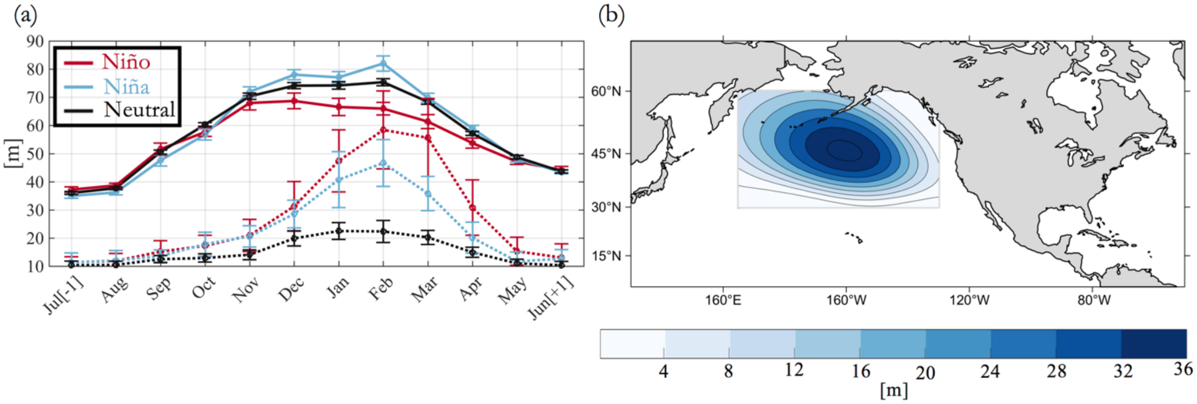

Figure 4.6 (a) Yearly development of composite RMS signal (dotted line) and noise (solid line) area averaged in (b) the region of interest (30°–60°N, 165°E–130°W) for El Niño (red), La Niña (blue), and neutral (black) years, as defined in Table 4.1, for monthly values from 1951–2010. Error bars show the 5th and 95th percentile bounds determined by bootstrap with resampling 1000 times across all El Niño years. The leading mode of variance (DJFMA seasonal mean) in the region of interest in (b), PC is normalized to unit variance.

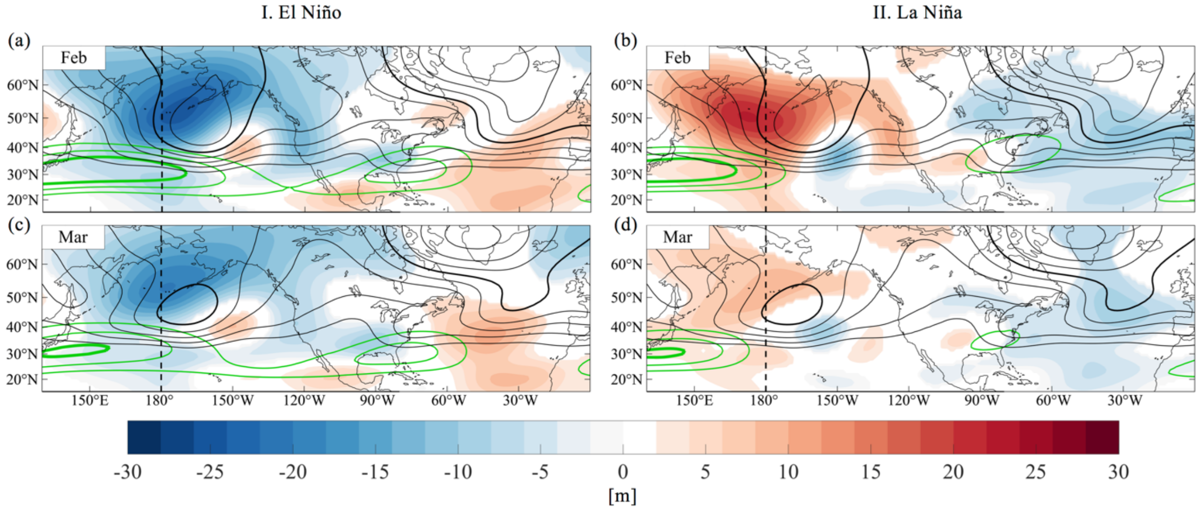

Figure 4.7 February and March RMS noise response to (column I) El Niño and (column II) La Niña, with a composite map of anomalous RMS noise 200-hPa geopotential height (colorfill); values that are not significant are not shown. The climatology of RMS noise is shown in black contours (60, 70, 80, 90, 100, 110, and 120 m; 100 m shown as the bold contour). Composite zonal wind is in green contours [40, 50, and 60 m s^{-1} (climatology + anomalies), the 60 m s^{-1} contour is shown in bold]. Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentile of the synthetic distribution.

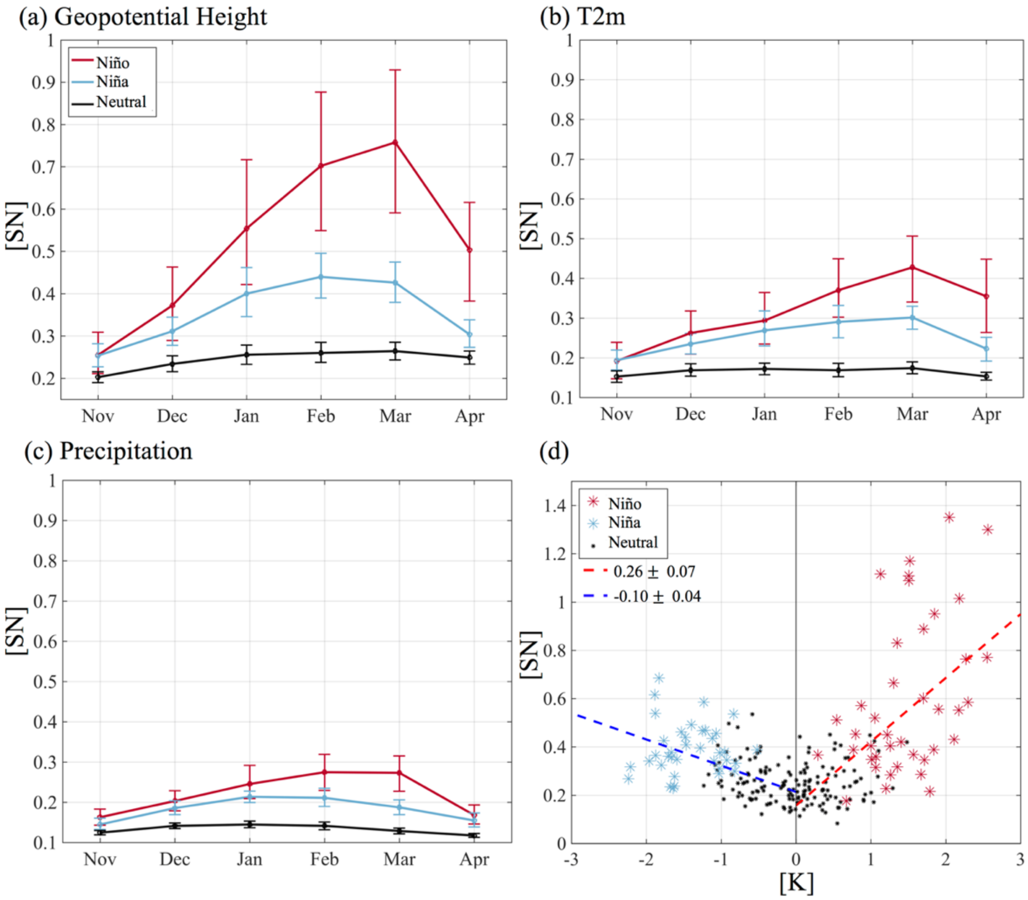

Figure 4.8 Area averaged (25°–70°N, 155°E–60°W) signal-to-noise ratio for (a) 200-hPa geopotential height (b) 2-m temperature (values over land only), (c) precipitation, and (d) geopotential height SN with respect to the Niño-3.4 anomaly, calculated for individual months (DJFM only). Red (blue) markers indicate El Niño (La Niña) years. Dashed red (blue) line indicates the linear fit calculated using every positive (negative) Niño-3.4 anomaly. The slope of each line is shown with 2σuncertainty. The 5th and 95th percentile confidence intervals are determined by bootstrap, with resampling across all years, 1000 times.

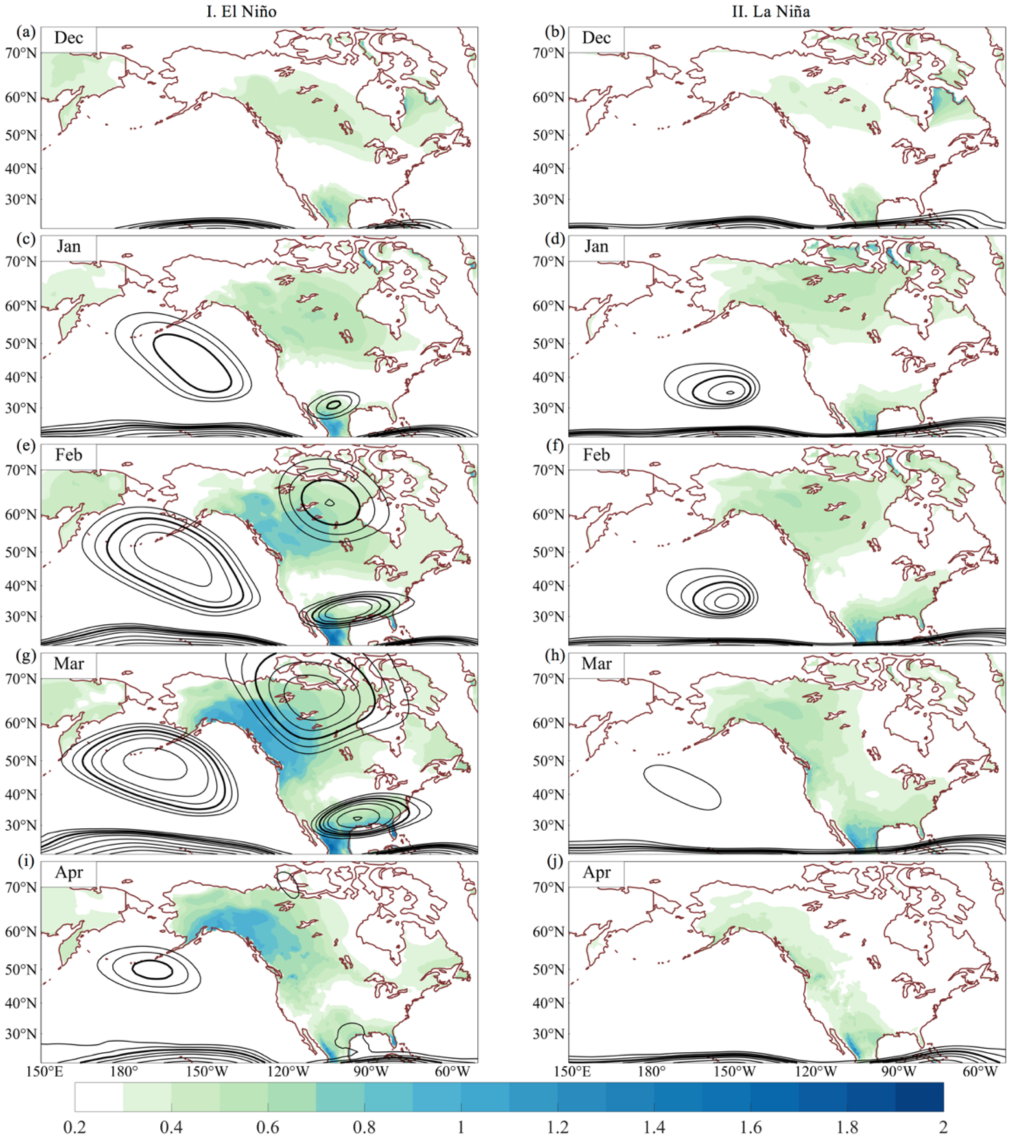

Figure 4.9. Monthly (DJFMA) signal-to-noise relationship for temperature (colorfill) and 200-hPa geopotential height (contour, 0.2 intervals beginning at 0.6, with 1 shown as the bold contour) for (column I) El Niño and (column II) La Niña years (as defined in Table 4.1).

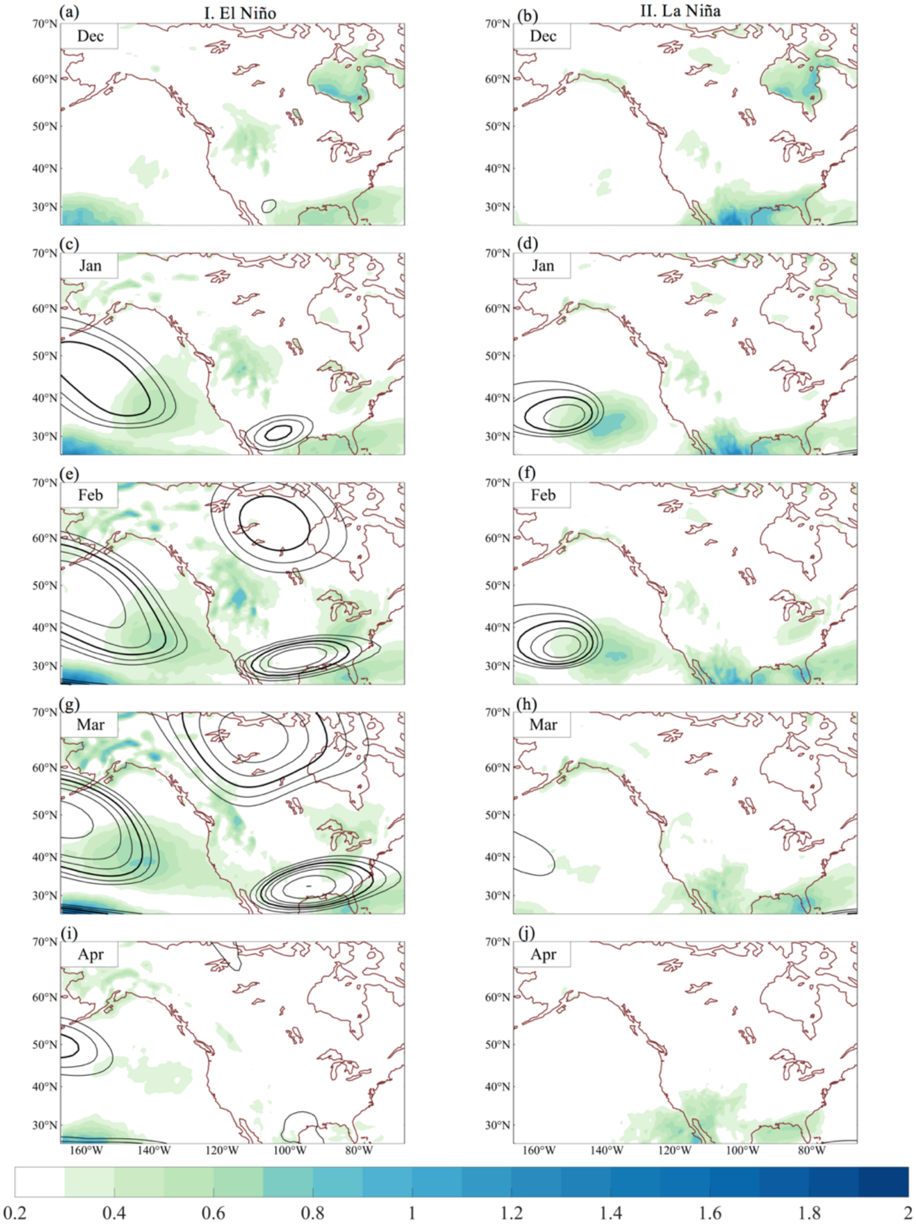

Figure 4.10. As in Fig. 4.9, but for North American precipitation (colorfill) and 200-hPa geopotential height (contour)

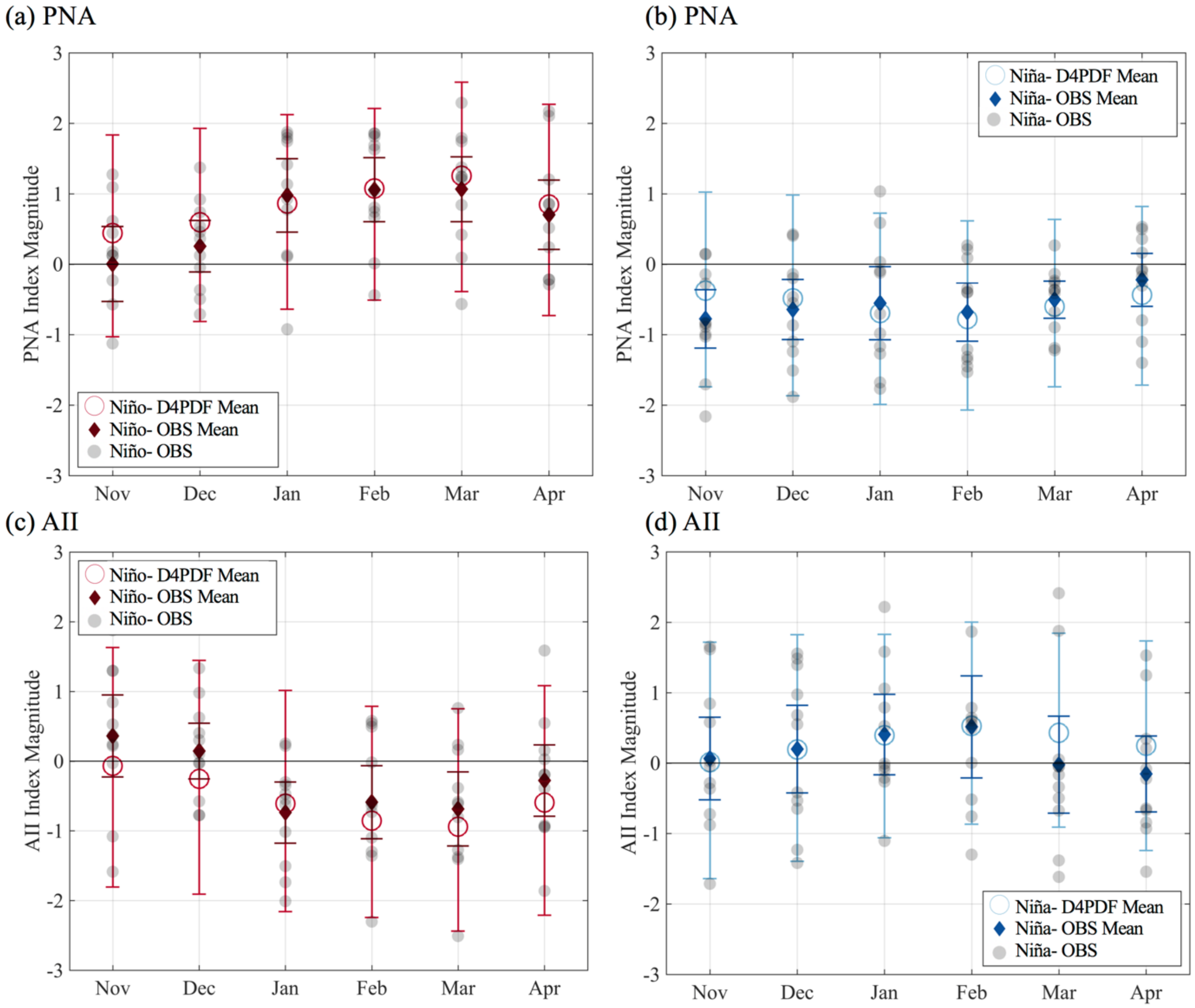

Figure 4.11. Monthly d4PDF (a),(b) PNA index and (c),(d) AII index mean and 5th and 95th percentiles across 100 ensemble members (open circle) and for observations (diamond, showing observation composite mean from ENSO years; observed values shown in gray dot) for (left) El Niño and (right) La Niña years. The observation mean spread is estimated from bootstrap with resampling 1000 times across years. The d4PDF ensemble intervals are estimated from bootstrap with resampling 1000 times across all members.

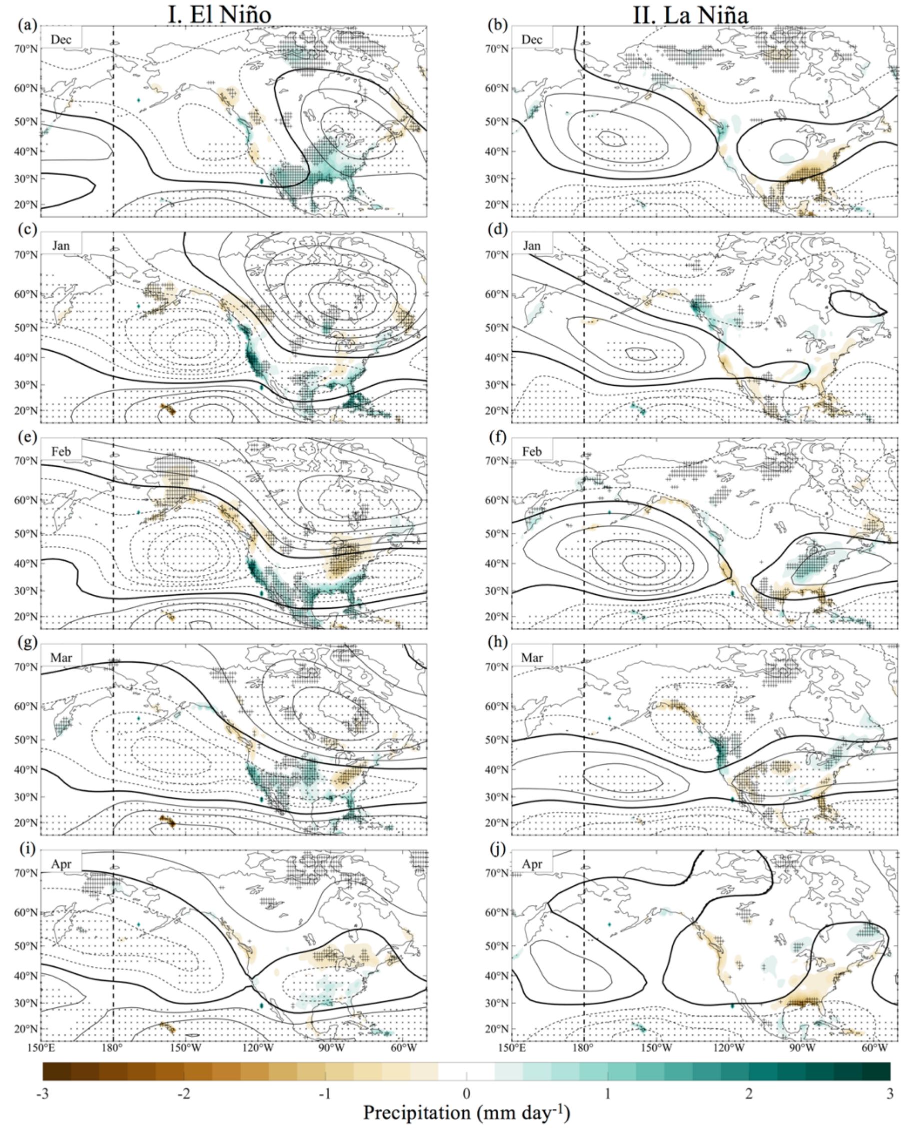

Figure 4.12. (column I) El Niño and (column II) La Niña monthly composite of observed precipitation (colorfill) and 200-hPa geopotential height (contour; negative dashed). Contour intervals are set at 20 m; the 0-m contour is shown in bold. Precipitation is stippled when significant (plus sign). Geopotential height is stippled when significant (star). Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentiles of the synthetic distribution.

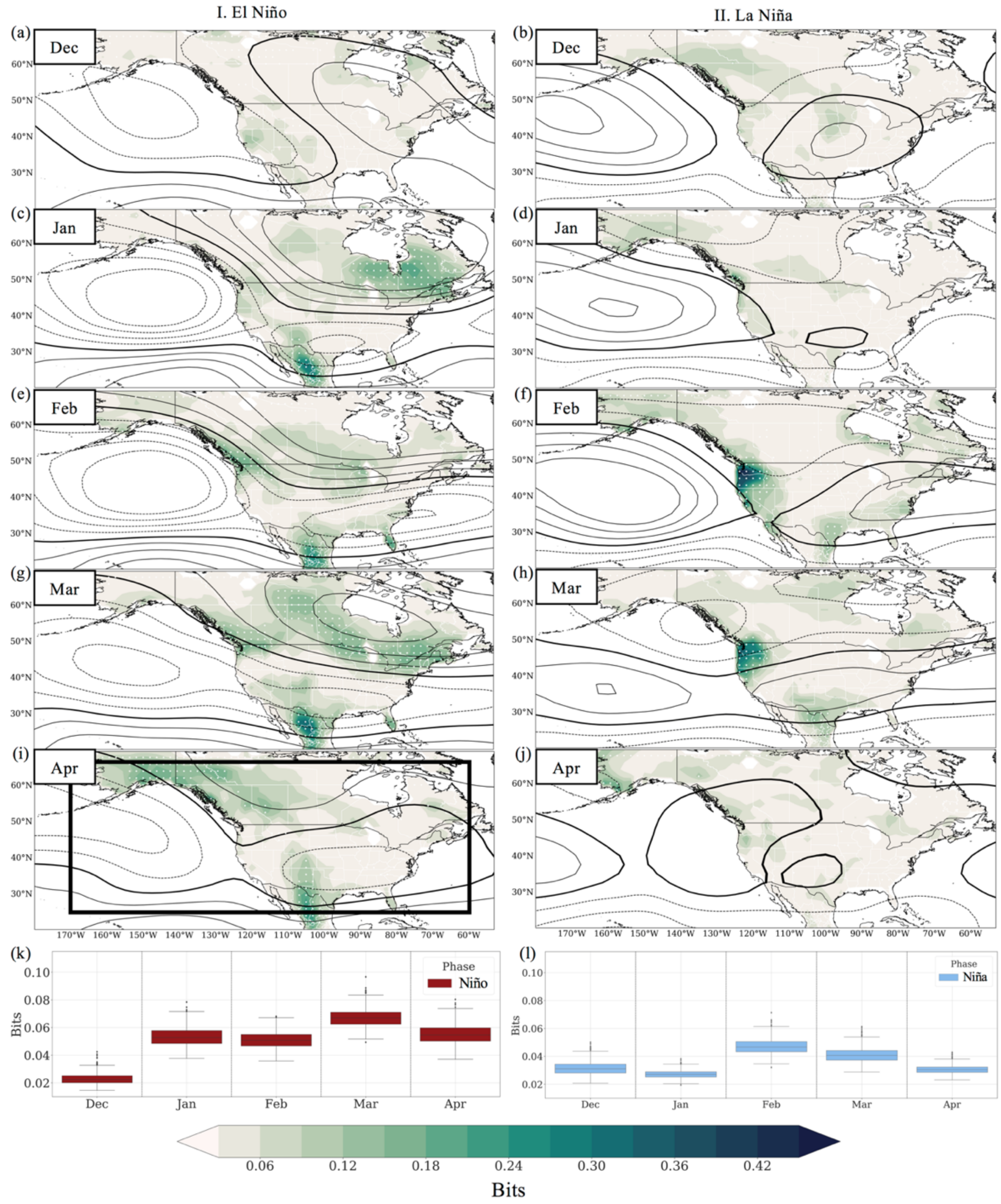

Figure 4.13. (a)–(j) Observed monthly (DJFMA) T2m Kullback–Leibler divergence (KL) (tercile discrete) for (column I) El Niño and (column II) La Niña years. Significant values of KL are stippled. Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentiles of the synthetic distribution. (k),(l) Land area averaged bits by month conditioned on ENSO phase (25°–65°N, 170°E–60°W, the region shown in (i)]. Contours show the observed 200-hPa GPH anomaly composite in 20-m intervals; 0-m contour shown in bold.

Table 4.1. The d4PDF defined ENSO states by year.

	El Niño	La Niña	Neutral
Condition	DJF ONI > 1K	DJF ONI < -1K	-1K < DJF ONI > 1K
Year	1957/58, 1965/66, 1968/69, 1972/73, 1982/83, 1986/87, 1991/92, 1994/95, 1997/98, 2009/10	1955/56, 1970/71, 1973/74, 1975/76, 1984/85, 1988/89 1998/99, 1999/2000, 2007/08	Remaining

4.8 Chapter 4 – Supporting Information

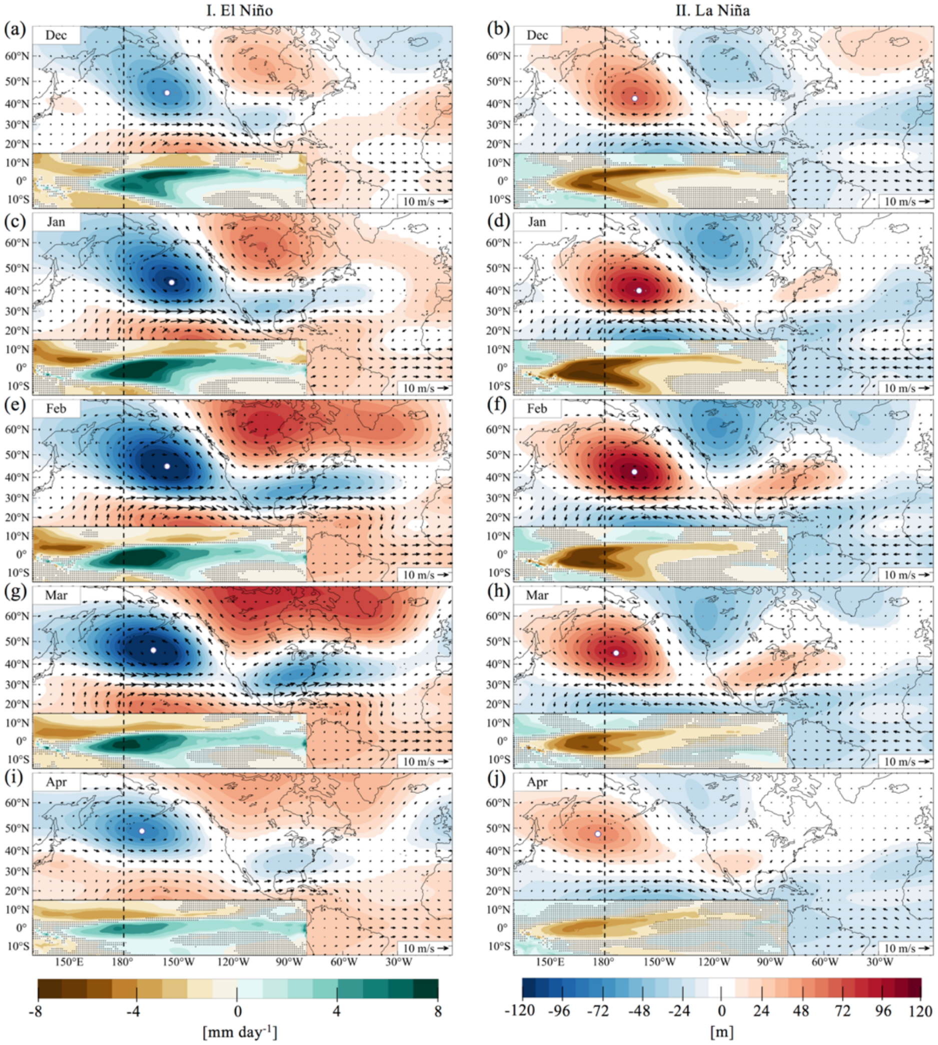

Figure 4.1S. Ensemble mean monthly (Dec-Apr) response to El Niño (column I) and La Niña (column II): composites of anomalous mean 200-hPa geopotential height (colorfill), 200-hPa winds (vector), and tropical precipitation (inset Lat: [15°S, 15°N], Lon: [130°E,80°W]) for El Niño (column I) and La Niña (column II). Anomalous geopotential height and wind vectors are shown for the 10% confidence interval. Insignificant wind vectors are shown in gray. Insignificant tropical precipitation is stippled. Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentile. White dot shows the Aleutian Low center of action.

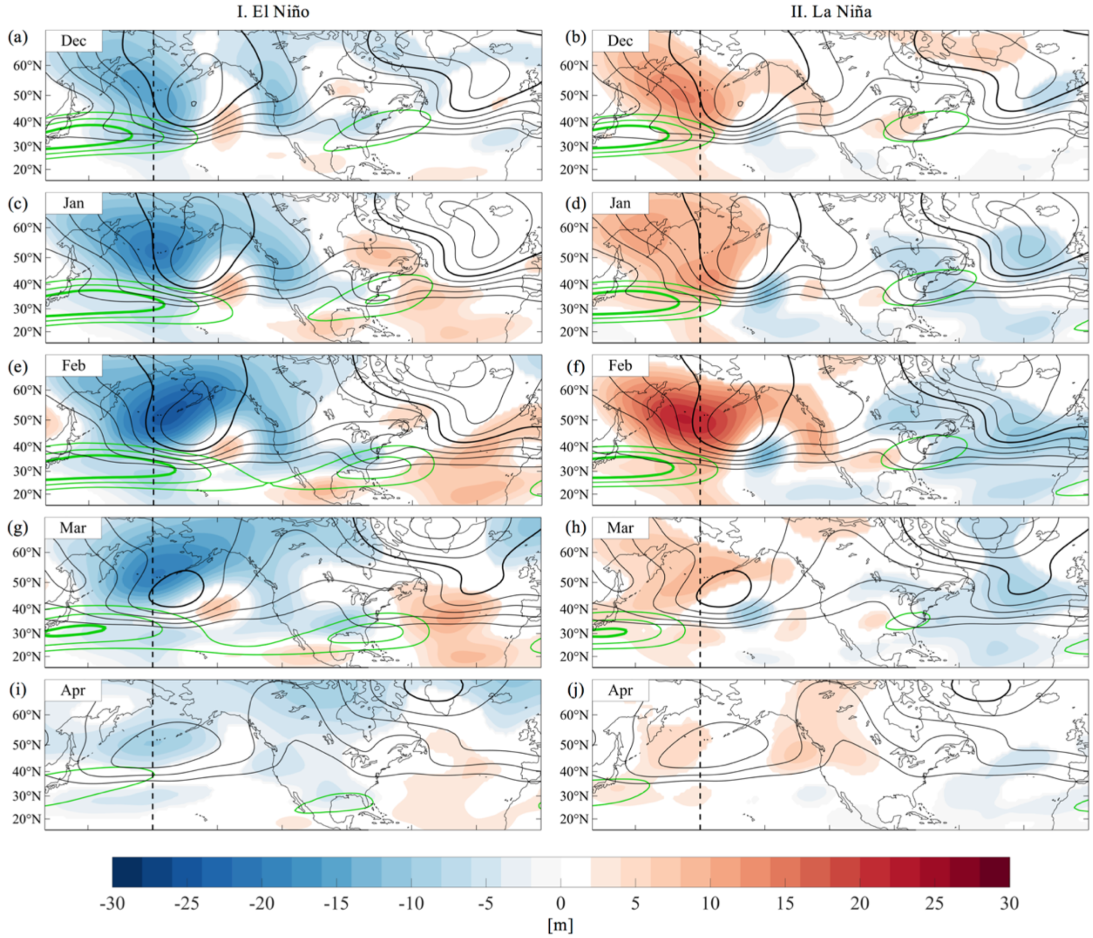

Figure 4.2S. RMS noise (Dec-Apr) response to El Niño (column I) and La Niña (column II): composite map of anomalous RMS noise 200-hPa geopotential height (colorfill), values that are not significant (10%) are not shown. The climatology of RMS noise is shown in black contours [m] [60, 70, 80, 90, 100, 110, 120]; 100m shown as the bold contour). Composite zonal wind is in green contour [ms^{-1} 40, 50, 55, 60] (climatology + anomalies). The confidence intervals are determined by bootstrap, with resampling across all years, 1000 times.

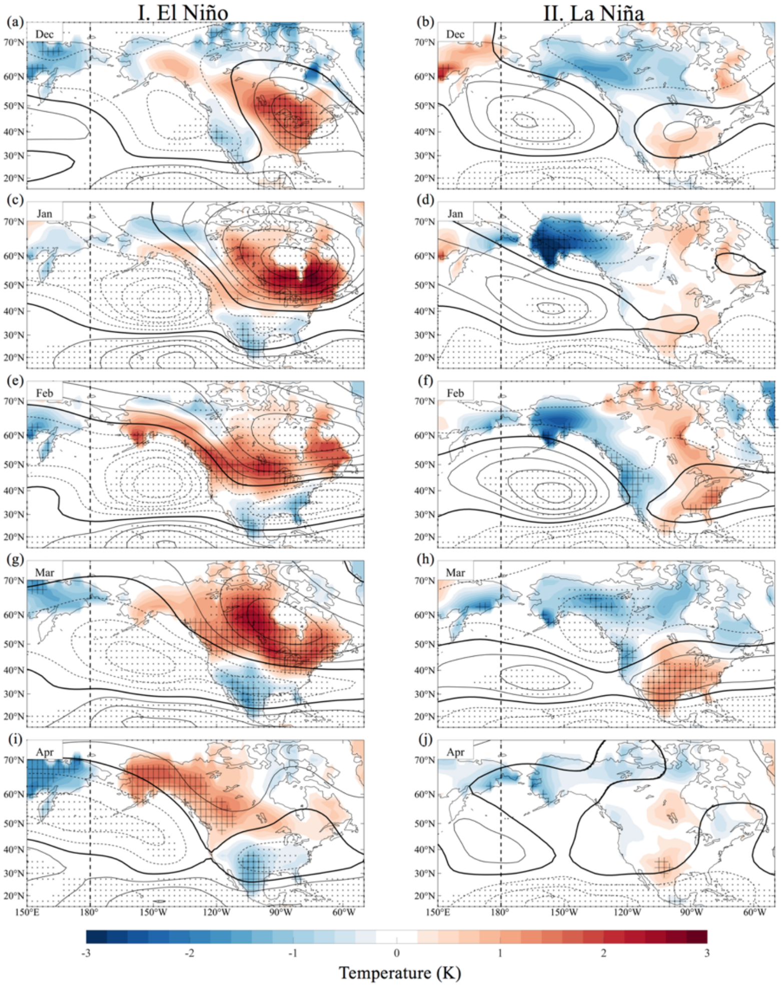

Figure 4.3S. El Niño (column I) and La Niña (column II) monthly composite of observed temperature (colorfill) and 200-hPa geopotential height (contour, negative dashed). Contour intervals are set at 20 meters; 0 m contour shown in bold. Temperature is stippled when significant (plus sign). Geopotential height is stippled when significant (star). Significant confidence intervals are determined by bootstrap, with resampling across all years 1000 times, and examination of the 5th and 95th percentile of the synthetic distribution.

Chapter 5

Subseasonal PNA Forecast Skill and Tropical PNA Drivers in the ECMWF 20^{th} Century Hindcast

Abstract

Using ensemble hindcasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) coupled model of the 20th century (period 1901–2010), we investigate the subseasonal forecast skill of the Pacific North American (PNA) pattern and the spatiotemporal evolution in the covariability of the PNA and 1) tropical sea surface temperatures (SST) and 2) the Madden Julian Oscillation (MJO) in both the November and February initializations. We find significant intraseasonal dependence of forecast skill and tropical forcing. The February initializations show a much more skillful subseasonal PNA forecast (compared to the November initializations). Additionally, the forecast skill derived from the low-frequency variability of the initial condition is much more valuable in February than in November. We investigate two known drivers of subseasonal PNA forcing, El Niño Southern Oscillation (ENSO) SSTs and the MJO. The covariability in the ensemble mean and ensemble spread is investigated with week-reliant singular value decomposition (SVD), which treats each variable in a given average weekly forecast sequence as a single time step. The leading mode of the ensemble mean represents the coevolution of the ENSO/PNA and MJO/PNA response. The leading mode of the ensemble spread in the SST/PNA SVD shows only response in the extratropical atmosphere forcing the tropical ocean indicating that on subseasonal time-scales the uncertainty in SST

ensemble spread shows little influence to PNA predictability. The MJO is revealed as the leading mode of ensemble spread in the tropical atmosphere. The February MJO/PNA SVD shows strong PNA modulation beginning in forecast week 3 with the growth of a northeast Pacific cyclonic/anticyclonic retrograding west and enforcing the PNA pattern. This pattern is notably lacking in the November initialization. Due to the large sample size provided by this simulation, we show that uncertainty in the MJO significantly influences uncertainty in the PNA forecast by forecast week 3.

5.1 Introduction

The Pacific-North American (PNA) teleconnection pattern is characterized by a Rossby wave train with four loading centers, which spans from the central tropical Pacific across the whole of North America (Wallace & Gutzler, 1981b). It is the leading mode of Northern Hemisphere midlatitude atmospheric variability (e.g., Chen & den Dool, 2003), is present in time-scales of ranging from weeks to years (Blackmon et al., 1984), and drives significant weather and climate anomalies over North America (e.g., Archambault et al., 2008; Gibson et al., 2020; Gutzler et al., 1988; Leathers et al., 1991, and many others). Therefore, predicting the PNA, particularly at long forecast horizons, is of utmost societal importance for North America.

Three main mechanisms have been noted for the development and persistence of the PNA pattern: 1) Poleward propagation of Rossby wave trains which are excited by tropical convection (e.g., Hoskins & Karoly, 1981). 2) Barotropic amplification of zonally asymmetric climatological flow in which the rapid growth of the PNA arises when the wave field has a spatial structure what projects spatially onto an unstable normal mode, which resembles the PNA. This particular modal structure is well suited to extract energy from the zonally varying northern hemisphere jet (e.g., Nakamura et al., 1987). 3) Amplification through a positive feedback onto the growing teleconnection pattern by high-frequency eddy vorticity fluxes (e.g., Egger & Schilling, 1983). This study looks to examine subseasonal model forecast skill, therefore we focus on tropical drivers with long system memory. These tropical fields are forecasted with much greater skill than the midlatitude drivers which are characteristic of the second and third mechanisms.

Theoretical and observational frameworks have been well-established which indicate that a coherent fluctuation is found between the PNA pattern and tropically derived convection. Particularly, the Madden-Julian Oscillation (MJO, Madden & Julian, 1971), and the El Niño Southern Oscillation (ENSO) have been cited in numerous studies for their role in sparking tropical convection which in turn drives vorticity sourcing in the midlatitudes (e.g., Henderson et al. 2020; Hoerling et al. 1997; Horel & Wallace 1981; Mori & Watanabe 2008). Additionally, particular focus has been paid to the background state of the extratropical atmosphere for driving PNA response to tropical forcing (Dawson et al. 2011; Henderson et al. 2017; Sardeshmukh & Hoskins 1988). The impact of the annual cycle on the global wind-field and thus the PNA's Rossby wave guide leads to significant dynamic monthly evolution of the midlatitude response to vorticity forcing (William E. Chapman et al., 2021). Therefore, studies that focus on a seasonal mean rather than accounting for the seasonal development of the background state will yield potentially misleading results by mixing the derived model skill across various degrees of forcing response (Newman & Sardeshmukh, 1998). However, likely do to the relatively short length of the observational record, much less focus has been paid to the intraseasonal development of PNA forecast skill and the tropical drivers of the PNA teleconnections when compared with seasonal forecasting.

The goal of this study is to examine the variability of the PNA forecast skill and PNA drivers on subseasonal timescales, as they evolve across a boreal winter, in an unprecedently large and long-running coupled seasonal forecast model: the European Center for Medium-Range Weather Forecast's (ECMWF) coupled hindcast of the 20th century. We systematically explore the PNA forecast skill from model initializations begun in November and February from using daily-to-monthly averaged predictability measures, and examine the tropically derived PNA model forcing and error growth. Additionally, this study leverages week-reliant multiple covariance analysis (MCA) to examine the co-evolution of forecasted fields and their associated teleconnection patterns. The ensemble mean is often used as the final forecast, while the ensemble spread (deviation from the mean) is used as a measure of prediction uncertainty. However, the spread can also be used for examining the intrinsic variability of the coupled tropical-midlatitude systems (e.g., Ma et al., 2017, 2021), while allowing for greatly increased degrees of freedom compared to the mean or observational space. Therefore, spread can be used to closely examine the growing modes of coupled variability in the model uncertainty space. Here, we examine the spatiotemporal coevolution of the MJO/PNA and ENSO/PNA teleconnection in the ensemble mean and spread in the ECMWF Hindcast for the November and February initializations.

This paper is organized as follows. In Section 2, data and analysis procedures are described. In Section 3, we leverage forecast skill measures to examine the forecasted versus observed PNA fields. In Section 4, we examine temporal variability of the tropically derived PNA forcing and uncertainty error growth in the November and February initializations. Section 5 gives summary and discussion.

5.2 Data

5.2.1 Observations

Daily averaged upper (200-hPa) and lower (850-hPa) zonal and meridional winds, geopotential height (500-hPa & 200-hPa) and sea surface temperature (SST) for the period 1901-2010 are obtained from the ECMWF Coupled Re-Analysis of the 20th Century (CERA-20C, Laloyaux et al., 2018). In CERA20-C, SST is derived from the Hadley Centere Sea Ice and Sea Surface Temperature dataset (HadISST) version 2.1.0.0 (Titchner & Rayner, 2014). These data are used for all forecast verification and all available observations are used to define the mean climatology. CERA-20C assimilates only surface pressure and marine wind observations. To reduce uncertainty associated with different spatial resolution, CERA-20C data are regrid to a 2.5° x 2.5° horizontal resolution using a 1st and 2nd order conservative remapping scheme (Schulzweida et al. 2006). Likewise, all model output, described below, are interpolated to this common grid resolution prior to any analysis calculation.

5.2.2 ECMWF 20th Century Hindcast Model

The coupled 20th century hindcast experiment (CSF-20C) is examined. CSF-20C was developed with ECMWF's Integrated Forecasting System (IFS) coupled model version cycle 41r1 which includes state-of-the-art atmospheric, land surface, oceanic, and sea-ice components (Weisheimer et al. 2020). The atmospheric resolution is run at T_L255 (~80 km) horizontally, with 91 vertical levels. The ocean resolution is 1° horizontally with 42 vertical levels. In our analysis, the atmosphere and ocean are regrid to a 2.5° x 2.5° horizontal resolution using a 1st and 2nd order conservative remapping scheme (Schulzweida et al. 2006). CSF-20C is initialized with ECMWF’s first coupled reanalysis of the twentieth century, CERA-20C (Laloyaux et al. 2018), which provides data from 1901 to 2010. Only surface observations in the atmospheric (i.e. surface pressure and marine winds, but no satellite data) and observed subsurface temperatures and salinity profiles in the ocean were assimilated in CERA-20C. We examine the hindcasts initialized on the first of November, and the first of February. These experiments are each run for 4 months. In this analysis, we examine the first 46 days of each model run. The November and February hindcasts consist of 51 and 25 ensemble members, respectively. The ensembles were created by a combination of stochastic perturbations to the model physics in the atmosphere and the 10 members of the CERA-20C. These experiments were designed to mimic ECMWF’s operational forecasts as much as possible to enable a clear comparison with a real-time forecasting system where only information before the initial date is available to use. Time-varying forcings from greenhouse gases, the solar cycle and volcanic aerosols were all prescribed.

5.2.3 ENSO Index

All Composite El Nino/Southern Oscillation (ENSO) variability was evaluated using the revised multivariate ENSO index (MEI, Wolter & Timlin 2011, https://www.psl.noaa.gov/enso/mei/). El Niño (La Niña) events are designated based on a MEI threshold greater (less) than (-)1°K at forecast initialization.

5.2.4 Model Drift

Numerical Weather Prediction (NWP) systems are designed to make accurate predictions over short forecast leads and are not subject to the same physical constraints that global climate models adhere to. Thus, NWP models are typically subject to systematic model drift in long-term forecasts. This includes subseasonal and seasonal prediction systems, which have been shown to develop pronounced spatially-dependent patterns of mean model drift on time-scales of 1-4 weeks (Vitart, 2004; Weigel et al. 2008). Prior to any analysis on a forecasted field, the lead-dependent bias is removed. Following Vitart (2004), we first calculate each ensemble mean lead time dependent bias by examining its bias on each specified date over the entire 110 year forecast period. The bias for a specific forecast date is taken as the average reforecast bias over all available calendar dates spanning the 28-day interval centered on that forecast date, and this bias is subtracted from the forecast produced by each of the corresponding ensemble members.

5.2.5 PNA Index

The PNA index is calculated as the principal component time series of the 1st mode of atmospheric variability, in the region [10°N,80°N], [140°E, 60°W] for a 30-day climatological window (110-year) centered on the forecast day of interest in the area-weighted 200-hPa CERA20-C geopotential height anomaly (Z200a) observations. These patterns are then projected onto the Z200a forecasted fields to obtain the principle component time series forecast. The PNA index is scaled by dividing by the square-root of the leading eigen-value of the decomposition (resulting in a unit variance time-series). Figure 5.1 shows the leading EOF from the November and February periods, respectively. We choose to define the PNA index in this manner, rather than a more standard grid-point definition (e.g., Wallace & Gutzler 1981) as significant intraseasonal development of the midlatitude jet leads to latitudinally/longitudinally shifted loading centers across a full boreal season (Chapman et al., 2021). However, it has been shown that the grid-point PNA indices and EOF developed indices are highly correlated when calculated over similar seasons (O'Reilly et al., 2017b). The PNA indices are calculated in this manner for the CERA-20C data set, hereafter called the reference PNA index, and for each member in the hindcast. The ensemble mean PNA forecast is calculated by averaging the individual hindcast member PNA indices.

5.2.6 Rossby Wave Source

We examine the 200-hPa Rossby wave source (RWS) (Sardeshmukh & Hoskins 1988).

$$RWS = -\xi D - \bar{v}_\chi \cdot \nabla\xi$$

The RWS is derived from the barotropic vorticity equation and locates vorticity forcing. RWS is computed using the magnitudes of divergence (D), the absolute vorticity (ξ), and the irrotational components of the wind ($\bar{v}_\chi$). RWS can be rewritten in the form of RWS anomaly (e.g., Hsu 1996; Seo & Lee 2017; Takahashi & Shirooka 2014; Wang et al. 2018):

$$(1)\ RWS' = -\bar{\xi}\nabla \cdot v'_\chi - \xi'\nabla \cdot \bar{v}_\chi - v'_\chi \cdot \nabla\bar{\xi} - \bar{v}_\chi \cdot \nabla\xi'$$

$$RWS' = S_1 + S_2 + S_3 + S_4$$

Here, the prime represents the intraseasonal anomalies and the overbar denotes the seasonal mean. The seasonal mean, and intraseasonal anomalies are calculated for each initialization independently. For the forecasted CSF-20C fields, we leverage the 46-day mean of the November and February initializations of each forecast, respectively. S_1 and S_2 are the vorticity forced by the interaction between anomalous divergence and mean vorticity, and by the interaction between mean divergence and anomalous vorticity, respectively. S_3 denotes the mean absolute vorticity advection by anomalous divergence flow and S_4 represents the anomalous absolute vorticity advection by mean divergence flow. RWS has been extensively used to examine the generation of Rossby waves excited by diabatic heating (e.g., Chapman et al. 2021; Johnson & Kosaka 2016; Kosaka & Nakamura 2006). To compute RWS terms we use the windpharms python package (Dawson 2016).

5.2.7 MJO Calculation

We follow the methods of Lin et al. (2008) for calculation of the real time MJO indices. A combined EOF analyses is performed based on Wheeler & Hendon (2004), except that instead of using outgoing longwave radiation (OLR) to represent tropical convection, we leverage velocity potential (VP) at 200 hPa (Ventrice et al. 2013). VP is the inverse Laplacian of divergence and acts as a smoother measure of convective activity than OLR and emphasizes the planetary-scale aspects of the divergent circulation -- spreading the MJO signal across the entire globe. Starting from the unfiltered observed daily averaged data of the CERA-20c reanalysis ensemble mean for VP and zonal wind at 850-hPa and 200-hPa from 1979-2010, the time mean, and the first three harmonics of the daily climatology are removed at every grid-point. Next, the time-series is filtered, by removing the grid-point time-mean of the previous 120 days. Removing the previous 120-day average eliminates most of the interannual variability, including the effects of ENSO. A meridional band average is then taken from 15°S to 15°N for the three fields. Each variable is then normalized by its own zonal average of temporal standard deviation, then the fields are combined and decomposed. The resulting structures of the EOF modes are very similar to Ventrice et al. (2013) and are shown in the supplementary material (Figure. 5.4S).

The same steps described above are then applied to the CSF-20C forecast data to calculate the first two modes of the real-time multivariate MJO index (RMM), RMM1 and

RMM2 in the forecast, respectively. In the forecasts, we do not have the previous 120 days of data (as the integration starts on forecast day 0), so we leverage the observations before the start date of the integration to replace the missing data. The observed and forecasted RMM1 and RMM2 are normalized by the standard deviations of the observed RMM1 and RMM2 index.

To evaluate MJO forecast skill, we compute the bivariate correlation (COR) or RMM1 and RMM2 compared to observations and define a skillful forecast as have a COR > 0.5. COR is defined as $COR(\tau) = \frac{\sum_{t=1}^{N}\{RMM_1^o(t)RMM_1^m(t,\tau)+RMM_2^o(t)RMM_2^m(t,\tau)\}}{\sqrt{\sum_{t=1}^{N}\{RMM_1^o(t)^2+RMM_2^o(t)^2\}}\sqrt{\sum_{t=1}^{N}\{RMM_1^m(t,\tau)^2+RMM_2^m(t,\tau)^2\}}}$, where $RMM_1^o(t)$ and $RMM_2^o(t)$ are the first two prinicipal components of the RMM analysis for the verification dataset at time t, and $RMM_1^m(t,\tau)^2$ and $RMM_2^m(t,\tau)^2$ are the first two prinicipal components of the RMM analysis for the forecast dataset at time t and forecast lead τ, and N is the total number of days considered.

5.2.8 Week Reliant MCA

To examine the spatiotemporal evolution in the covariability of examined forecast fields, we conduct week-reliant multiple covariance analysis (MCA), via singular value decomposition (SVD) (Deser & Timlin 1997; Wallace et al. 1992). This is structurally similar to the month reliant SVD analysis (Ma et al. 2017; Ma et al. 2021) (also called extended empirical orthogonal functions (Weare & Nasstrom, 1982)). The SVD is performed on concatenated ensemble member × 110 years x 7 weeks record of forecasted fields in the CSF-20C. For example, the forecast matrix is (N_x, N_y, N_w, N_{ens}, N_{yr}), where N_x and N_y are the grid point numbers in the zonal and meridional direction, respectively, N_w is the weekly mean forecast value (we examine out to 6 forecast weeks), N_{ens} is the ensemble size and N_{yr} is the number of years. First, we focus on the ensemble-mean variability by obtaining the ensemble-mean anomalies defined as the deviations from the climatological mean (N_x, N_y, N_w). Then the matrix ($N_x \times N_y \times N_w$, N_{yr}) formed. We form right and left heterogenous MCA fields by forming the covariance matrix of two examined forecast fields, SVD is then performed to decompose this covariance matrix.

Additionally, we examine the ensemble spread MCA modes by subtracting the ensemble mean (N_x, N_y, N_w, N_{yr}) from the debiased hindcast (N_x, N_y, N_w, N_{ens}, N_{yr}). The matrix ($N_x \times N_y \times N_w$, $N_{ens} \times N_{yr}$) is used to form the covariance matrix for the left and right heterogenous forecast fields, and we conduct the SVD analysis on this matrix. Hence, the conventional time dimension is enlarged by the ensemble size. A covariance matrix is constructed by treating the forecast anomalies in the weekly sequence as a single time-step. In this way we obtain the heterogeneous component (PC), and the heterogenous fields reflect the temporal evolution of the anomalies. To display the leading SVD modes, we regress the forecasted anomalies on to the normalized PCs (expansion coefficients) of the heterogenous fields. We calculate the correlation between the examined expansion coefficients and use a student's t test to determine the significance of these correlations.

MCA decomposition via SVD yields the leading modes of covariance between two datasets, shows the total dataset covariance, and gives the fraction of covariance explained by each covariate mode. It does not, however, give the fraction of the variance in each data set that is explained by these covariate modes. It can be easily shown that the fraction of variance in the examined datasets explained by each mode, can be obtained through manipulation of the MCA expansion coefficients. The fraction of variance explained by mode k is $VarFrac(k) =$

$100 \cdot \left[\frac{\sum_{t=1}^{T} a_{kt}^2}{\sum_{t=1}^{T}\sum_{m=1}^{M} a_{tm}^2}\right]$, where a is the MCA expansion coefficient, t is the discrete time index, and m carries the same dimension as the input spatial index ($N_x \times N_y \times N_w$). We refer the reader to Prohaska (1976) and Storch and Zwiers (1999) for and in-depth examination of the linear decomposition of coupled fields.

5.3 PNA Forecast Skill

The PNA ensemble spread (defined here as the average ensemble standard deviation) and rms error is shown for the daily averaged November (Fig. 5.2a) and February (Fig. 5.2b) model initializations. Ensemble spread and rms error are basic evaluations of the dispersion characteristics and prediction skill of an ensemble system, as it is valuable for an ensemble forecasting system to have the ability to forecast its own error (e.g., Hopson 2014; Molteni et al. 1996). Ensemble spread shows the sensitivity to initial condition and model uncertainty error growth and rms error measures the model prediction accuracy against the observational record. In a perfect ensemble system, a single ensemble member is indistinguishable from observations, if this is the case, the RMSE and ensemble spread are equivalent. Forecast error grows monotonically and eventually saturates, representing the upper limit of daily weather predictability (Lorenz 1969). The ensemble spread for both initialization times is notably under dispersive (smaller than RMSE) at all shorter lead times. The CSF-20C was originally designed for seasonal forecasting, and only weak initial perturbation was applied (Weisheimer et al. 2020), which is likely the cause of the initial under dispersion. The saturated November spread is larger than the February spread (~1.0 vs. ~0.82) and occurs earlier in the forecasting period (~20 days vs. ~25 days) suggesting that November forecasts are less certain that their February counterparts (Stensrud et al.1999).

For subseasonal to seasonal (S2S) forecasts, it has been shown that time and space averaging can help to improve forecast skill by isolating the low-frequency component of atmospheric variability (e.g., Deflorio 2019). However, care must be taken when averaging, as time-averaging reduces the phase decorrelation rates of the averaged fields and thus eliminates instabilities which propagate with periods less than the averaging time, and it also reduces the climatological variance of the averaged fields (Tribbia & Baumhefner 1988). We define a forecast as skillful when error variance is equal to climatological variance, or when anomaly correlation is great than 0.5. We now investigate the impact of time-averaging the PNA signal at weekly intervals.

Figure 5.3. shows the Pearson-correlation between forecasted PNA ensemble mean and observed PNA as a function of lead time in the November initialization (Fig. 5.3a) and the February initialization (Fig. 5.3b). The PNA is index is calculated on a rolling mean atmosphere state, at one-week intervals from 1 to 28 days, and the full record is used (1901-2010) in the correlation. The indicated forecast/observed correlation is shown at the central day of the rolling mean forecast. The February subseasonal correlation skill is significantly greater than the November initializations. The 7-day averaged November forecast correlation drops below 0.5 in week 2 (central day 8). Increasing the average window at 1-week intervals increases forecast skill by approximately two days per week averaged, with correlation greater than 0.5 out to day 20 when a 4-week average is leveraged. The 7-day February forecast correlation drops below 0.5 in week 4 (central day 27). Again, increasing the average window at 1-week intervals increases forecast skill by approximately two days per week averaged, with correlation greater than 0.5 past day 32 when a 4-week average is leveraged. This finding is consistent with the concept that predictability can be improved by taking temporal or spatial averages as large-scale variability does not change rapidly and the growth of initial errors is relatively slow for low-frequency components (e.g., den Dool & Saha, 1990; Lorenz, 1969; Younas & Tang, 2013). Additionally, in AGCM simulations, it has been shown that the late winter 200-hPa signal-to-noise (SN) peaks in February and remains significantly elevated in March over the PNA region this is especially attenuated in ENSO years (Chapman et al. 2020), while the early boreal winter season (ND) has relatively low ENSO driven SN (Kumar & Hoerling 1998). The intraseasonal disparity of forecast correlation skill supports that conclusion.

The pseudo-persistence forecast is shown in Figure 5.3c & 5.3d. Here, pseudo-persistence is defined ny retaining the first N-day averaged period of the forecast system as the forecasted anomaly and tests the impact of the low frequency initial condition certainty. As expected, the autocorrelation increases with time averaging from 7 to 28 days. When temporally averaged, every forecast field in the November initialization is less skillful than its respective persistence forecast, indicating that most of the forecast skill (when the forecast is averaged forward in time), is contained in the initial condition. This lower skill anomaly potentially indicates a significant shift between the large-scale forcing features which act in the early boreal winter. The February persistence forecasts vastly outperform the November forecast, indicating a more stable low-frequency atmosphere in the late boreal winter. For example, the February 7-day average persistence forecast drops below 0.5 at 14-days (Fig. 5.3d), whereas the actual forecast retains skill out to day 27. This indicates that the role of boundary forcing, outside of the initial condition is particularly attenuated during the February/March period, but not in November/December. Thus, the overall increase in forecast skill with time averaging can be attributed to 1) averaging weakening the noise, making the initial condition signal less dissipated, and 2) the role of boundary conditions and weather forcing patterns (i.e., ENSO and

MJO) become increasingly more influential with time averaging. We note that the one day forecast for February (November) vastly outperforms its respective persistence forecast with forecast correlation above 0.5 out to 21 (7) days and a persistence forecast correlation above 0.5 out to 8 (5) days.

We now investigate the phase dependence of subseasonal PNA forecast skill, by calculating the relative operating characteristic (ROC, Swets 1973) at various PNA threshold values (e.g., Kharin & Zwiers 2003; Mason & Graham 1999). The ROC curve provides a measure of the probabilistic forecast skill of a binary event, here, defined as the PNA index exceeding (or not) a prescribed threshold, and compares the forecast hit rate and false-alarm rate. The ROC skill score (ROCSS, defined as double the area under the ROC curve, minus one) is positive for skillful forecasts, equal to one for a perfect forecast and negative for forecasts less skillful than climatology. Figure 5.4. shows the phase dependent ROCSS curve for the November initialization (Fig. 5.4a) and the February initialization (Fig. 5.4b) centered on forecast day 28. The PNA is index is calculated on a rolling mean atmosphere state, at one-week intervals from 7 to 28 days, and the full record is used (1901-2010) in the ROCSS calculation. We choose to show the forecast centered on day 28 but note that the general findings derived from this figure are robust across every day examined. Confidence intervals are shown and are derived by performing bootstrap resampling 1000 times with replacement and give a statistical significance estimate to the ROCSS. We note that the (-)1.1 threshold represents the (10th) 90th percentiles in the PNA observational record when calculated over both forecast initializations.

November initializations show significantly lower ROCSS across every PNA threshold for the 7 and 28 day averaged forecast models when compared with the February ROCSS.

Positive PNA events above the 0.9 threshold, in the 7-day average, are not significantly forecasted better than climatology, whereas forecasting negative PNA events show a significantly better forecast. The 28-day average ROCSS is significantly more skillful than climatology at every threshold, though largely this is attributable to the low-frequency component of the initial condition (Fig 5.3c), for November, and additional boundary condition forcing for February. The February initialization performs remarkably well at predicting positive PNA events with average ROCSS of ~0.7-0.75 when the PNA observed at greater than a 0.1 threshold. We note that the November 28-day averaged, negative PNA anomaly shows signs of significant skill. Though decadal change is not the main focus of this study we show the same ROCSS for 1950-2010 in supplemental Figure 5.1S. The February ROCSS is relatively unchanged, whereas the November ROCSS drops significantly. Indicating that the dominant share of forecast skill in the November initialization is in the first 40 years of the record. This is a period when fewer observations are assimilated into the CERA-20C reanalysis verification product, and the deviation from the CERA initialization is greatest. Further focused study is required to determine whether the observed ROCSS in November is indicative of decadal change, or spurious sampling error. However, we find no evidence of boundary forcing effectively modulating the November PNA forecast period and therefore the later conclusion appears more probable.

5.4 Variability and Drivers of S2S PNA Predictability

In the previous section we discussed the models skill in forecasting the PNA across subseasonal time scales, we now seek to explain differences in the intraseasonal forecast skill by examining two major known drivers of PNA S2S forecast skill to examine the intraseasonal growth of PNA forcing and growth of PNA uncertainty: 1) ENSO (e.g., Straus & Shukla, 2002; Wallace & Gutzler, 1981; L. Wang & Robertson, 2019; Younas & Tang, 2013, and many others) and 2) the MJO (Mori & Watanabe, 2008; Riddle et al., 2013; Tseng et al. 2018; Tseng et al. 2020; Vitart & Molteni, 2010).

5.4.1 4-week Averaged Response

5.4.1.1 Tropical SSTs

ENSO is the dominant mode of global interannual variability, and it impacts weather in North America through its dominant teleconnection pattern, the PNA (Wallace & Gutzler, 1981b). However, recent work has highlighted that the intraseasonal evolution of the ENSO related PNA teleconnection's signal-to-noise ratio at monthly timescales varies drastically across boreal winter (Chapman et al. 2021). In the following section we will refer to ([5N,5S], [170W, 100W]) as the Nino3.4 region. Figure 5.4b shows a notable increase in skill associated with the 4-week averaged positive PNA events in February which is not present in the November initialization. It has been documented that El Niño events typically drive positive PNA patterns (Hoskins & Karoly, 1981), and this is a robust signal in both observations and climate models (e.g., Deser et al. 2017). Chapman et al. (2021) extensively explored the monthly modulations of SN in the PNA region associated with ENSO and found that SN should peak in February/March El Niño years and have associated weak SN in November/December. The CSF-20C model skill gives evidence to show that those findings hold in a coupled forecast setting. We first briefly examine the 4-week averaged forecast to explain this skill discrepancy. The MJO is the dominant mode of tropical intraseasonal variability, however, a full MJO cycle is realized in a period of approximately 30 to 60 days (Madden & Julian, 1971). The MJO/PNA teleconnection growth and decay is typically expressed in ~10 days period (Mori & Watanabe, 2008). Therefore, ENSO will be the sole focus of this 4-week averaged forcing section.

Figure 5.5a and 5.5b shows the 4-week averaged ensemble mean PNA response to SSTs averaged in the region Nino3.4 region, centered on forecast day 28 for the November and February model initializations, respectively. We choose to show the forecast centered on day 28 but note that the general findings derived from this figure are robust across every day examined (not shown). It is clear that tropical SSTs play a dominant role in determining the PNA response in February, accounting for ~50% of variance in the PNA forecast. However, the November forecast shows a very weak response to the concurrent Nino3.4 SSTs.

We diagnose this seasonal discrepancy using an RWS' vorticity sourcing framework (Sardeshmukh & Hoskins, 1988). During ENSO, convective anomalies, associated with the anomalous SSTs lead to anomalous tropical divergences which propagate into the midlatitudes, interact with a strong background absolute vorticity source (the midlatitude jet), and induce Rossby waves (Hoskins & Karoly 1981). Figure 5.5g and 5.5h show the 4-week averaged, ensemble mean RWS' response in the region: [40N,30S]; [140E, 180] to SSTs in the Nino3.4 region centered on forecast day 28 for the November and February model initializations, respectively. The February initialization has a much greater RWS' response to tropical SSTs than the November initialization, and is much more sensitive to ENSO events. The RWS' is comprised of two, major components anomalous divergence ($\nabla \cdot v_{\chi}$), and absolute vorticity (ξ). It has been shown that the anomalous SSTs associated with ENSO in boreal winter peak in December, and the discrepancy between the November and February ENSO SSTs is not significant (see scatter in x-axis Figure 5.5, or Chapman et al. (2021), Fig 5.3e). Therefore, we examine the divergence associated with ENSO SST in Figure 5.5e and Figure 5.5f, the February initialization provides does provide greater divergent component than the November initialization.

To further diagnose the dominant terms that contribute to the RWS' and thus the downstream teleconnection, RWS' anomalies are examined with the individual terms in Eq. (1). The resulting scatter of tropical SSTs and RWS'- term S1 $(-\bar{\xi}\nabla \cdot v'_{\chi})$ are nearly identical to Figure 5.5g, indicating that it is the dominant term driving the RWS' (not-shown). This finding is in agreement with multiple previous studies (e.g., Hsu, 1996; Wang et al., 2020). The midlatitude anomalous divergent flow $(\nabla \cdot v'_{\chi})$ is slightly weaker between the November and December initializations (Fig. 5.5e and Fig. 5.5f), but not at a magnitude to explain the entire discrepancy between the model RWS' response of the November and December initializations (Fig. 5.5g and Fig. 5.5h). Therefore, the seasonal-mean background absolute vorticity $(-\bar{\xi})$, must play an additional significant role. It has been shown that the 200-hPa jet undergoes a seasonal extension and intensification through early winter (November-January) as the northern hemisphere midlatitude baroclinicity increases, reaching its greatest zonal extent in February (Newman & Sardeshmukh, 1998). Additionally, the 200-hPa zonal winds are modulated by El Niño (La Niña) winter with a southward (northward) shift, intensification (reduction) in magnitude, and thus an increased (decreased) zonal extent across a boreal winter (e.g., Jiménez-Esteve & Domeisen 2018) modulating the absolute vorticity term with the ENSO season. For completeness, we show the 4-week averaged ensemble mean composite, centered on forecast day 28 of RWS, divergent wind, and climatological + anomalous jet in each ENSO season in the supplementary material Figure 5.2s. The composite figures are in agreement with the above findings.

PNA events have been shown to be less skillfully forecast when initialized in the negative phase (e.g., Palmer 1988) due to significantly more extraction of kinetic energy from the seasonal mean flow and large scale deflection of synoptic scale eddies during the PNA positive phase (Lin & Derome, 1996), leading to a greater internal variability when the PNA is in the positive phase. While the difference is not significant from a ROCSS standpoint, the negative PNA events forecast skill have a greater skill variance than their positive counterpoints. Negative PNA events are strongly associated with La Niña events (Fig. 5.5a). The findings we present here agree with that result. Chapman et al. (2021) noted a significant modulation atmospheric internal variability associated with ENSO events, with increased (decreased) atmospheric noise associated with La Niña (El Niño) events which could indicate more variable forecast skill. Figure 5.6 shows the percent modulation of the ensemble spread of Z200a per one standard deviation of change in the averaged tropical SSTs in the region [5N,5S], [170W, 120W]. In El Niño (La Niña) seasons, atmospheric internal variability is reduced (increased) by ~12% across the PNA region per standard deviation of Nino3.4 temperatures. Following Simmons et al. (1983), it is observed that the difference in barotropic conversion of kinetic energy from the time mean to the low-frequency eddies between ENSO phases, bares a very similar spatial structure to Figure 5.6 (not-shown), and likely drives the variability discrepancy. Additionally, Sardeshmukh et al. (2000) noted that with the increased precipitation signal associated with El Niño (La Niña), the variance in the tropical west pacific also increased, potentially contributing additional system noise associated with the midlatitude teleconnection and increasing forecast uncertainty. This same phenomenon is observed in the

CSF-20C with standard deviation of divergent flow being significantly greater in El Niño seasons than in La Niña seasons (2.6e-6 s^{-1} and 2.2e-6 s^{-1} in region [5N,5S] [170W, 100W], respectively). However, the midlatitude ENSO associated standard deviations of divergent flow show an opposite relationship to the tropical SSTS. The standard deviation of divergent flow being significantly greater in La Niña seasons than in El Niño seasons (2.5e-6 s^{-1} and 2.25e-6 s^{-1} in region [40N,30N] [140W, 180], respectively).

5.4.2 Weekly Averaged Response

Recent work in subseasonal prediction has highlighted the value of weekly time averaging to eliminate high variability noise in the climate system, yielding increased atmospheric predictability (e.g., Deflorio et al., 2019; Younas & Tang, 2013). We now turn to the weekly averaged development of tropically derived PNA forcing and prediction uncertainty growth and associated ENSO and MJO derived teleconnections.

5.4.2.1 Tropical SSTs

Figure 5.7. shows the 7-14 day averaged lagged response of Z200a (black contour), the 0-7 averaged lagged response of RWS' (colorfill) and divergent wind (vector) and the 200-hpa zonal wind (climatology, green contour) composite in El Niño (Fig. 5.7a) and La Niña (Fig. 5.7b) in the CSF-20C. The composite is calculated across every ensemble member individually and conditioned on the MEI defined ENSO index. Additionally, we characterize model bias in the Z200a field, and show where the model is biased high in stippled regions and biased low in hatched regions. Bias is calculated with the same method described in Deser et al. (2017). To determine bias, at every point we randomly sample the full ensemble simulation an equal number of times as the observed record conditioned on the ENSO phase of interest. This random compositing is performed 2000 times. We calculate the 5th and 95th percentiles of these pseudo composites distributions, the stippled and hatched regions show the areas where the observed composites lie outside of the bootstrap- sampled composite distributions.

In November, the divergent wind anomaly which reaches the West-Pacific extratropics is weak, additionally the climatological zonal wind is notably weaker when compared to later in the boreal winter. This results in a weak expression of $RWS'$and thus little vorticity forcing to generate the PNA pattern. Generally, there is no indication of a resolved PNA pattern. The La Niña composite indicates a weakly anomalous Aleutian Low (AL), however a bias is observed in the Bering sea. In El Niño seasons, no significant Z200a is observed in the extratropics. Notably, the anomalies in the North Atlantic Oscillation region are missing in November entirely.

In February, strong divergent/convergent wind is observed stemming from the tropics and interacting with the midlatitude jet. The RWS' is characteristic of the ENSO teleconnection, with high anomalies co-located with the northern flank of the midlatitude jet in the North-West Pacific (Qin & Robinson 1993; Sardeshmukh & Hoskins 1988). This generates the cyclonic/anticyclonic anomaly over the extra-tropical North Pacific and the PNA pattern. The February initialization shows a well-developed PNA response pattern in both ENSO phases, with particularly strong anomalies in the Aleutian Low (AL) region. The February El Niño composite shows model bias in a zonally stunted AL anomaly, which should show a larger extent to the west and east. Additionally, the low-pressure pattern observed over the Southeast United States shows a notable lack of depression. The February La Niña pattern shows a characteristic negative PNA, but the bias indicates that the Aleutian the low anomaly is too high, indicating that the model's teleconnection is too reactive to La Niña events in late boreal winter.

Ensemble Mean Anomaly Evolution

We investigate the spatiotemporal coevolution of ensemble-mean SSTs (region: [15S,15N]; [0E,359E]) and Z200a in the PNA region ([25N,70N]; [140E, 60W]) through week-reliant MCA. The weekly averaged SSTs and Z200a are used as right and left fields, respectively. Figure 5.8 shows the left and right heterogenous fields of the first week-reliant MCA mode in the November (Fig. 5.8a) and February (Fig. 5.8b) initializations for the first six weeks of the forecast period. Additionally, the RWS' (red contour) and 200 hPa divergent wind (vector) regressed on the normalized Z200a expansion coefficient is shown. The November fields explain 70.6% of covariance, 5% of variance in the GPH field and 68.7% of variance in the tropical SSTs. Tropical SSTs clearly show an ENSO like expression with anomalous SSTs peaking in the mid tropical pacific. The divergent flow to the midlatitudes is extremely weak and RWS' fades quickly after week 2. As was observed in Figure 5.7a, the midlatitude Z200a atmospheric response to tropical SST forcing is nearly negligible with a weakly discernable AL pattern emerging in weeks 0-2 but quickly dissipating. On weekly timescales it appears that there is little to no link in the ensemble mean of tropical SSTs and the PNA pattern.

The February patterns (Fig. 5.8b) explain 84.6% of covariance, 13.1% of variance in the GPH field and 55.8% of variance in the tropical SSTs from weeks 0-5. The heterogenous tropical SST pattern clearly show an ENSO expression, with tropical divergent winds consistent with the Figure 5.6 composites, and consistent RWS' remaining throughout the entire 6-week forecast. The Z200a anomaly slowly weakens across the forecast, as the ensemble members spread -- diminishing the effective anomaly in the ensemble mean and forecast reliance on the initial condition. By week 6, only the boundary condition forcing remains (see Fig. 5.2b & Fig. 5.3b), and the forced ensemble pattern resembles the leading mode of atmospheric variability (Fig. 5.1b). The evolution of the SST pattern is small across the forecast period, owing to the slow evolution ENSO related anomalies, but the maximum SST anomaly in the ensemble mean is diminished in week 5 by 0.18 [K] compared to the forecast initialization.

SST Ensemble Spread Evolution

We now turn our attention to the spatiotemporal evolution of the ensemble SST spread and its effect on the PNA predictability. Figure 5.9 shows the evolution of the leading mode of variance in the ensemble spread of tropical SSTs (region: [15S,15N]; [0E,359E]) for the November and February initializations (Figure 5.9a & 5.9b). This mode represents ~5% of total variance in both initializations. The anomalous westerly winds appear to slow the trade winds and induce a wind-evaporation-SST (WES) feedback (Xie & Philander, 1994), in which cross equatorial flow decelerates the trades in the warmer hemisphere. These trade wind anomalies can often be induced by cross-basin interaction of large-scale climate modes, particularly the North Atlantic Oscillation (Amaya & Foltz, 2014; Chiang et al. 2002). Additionally, they are early indications of the arrival of the North Atlantic meridional mode (Chang et al. 1997) , and the North Pacific meridional mode (Chiang & Vimont 2004). The above analysis indicates that anomalies in tropical SSTs directly influence the expressed PNA. However, though this the cross-basin interaction has been shown to affect future ENSO variability (e.g., Ma et al., 2021), the time-scales act on the order of months to seasons, and not weeks. To investigate the ensemble spread's affect influencing the extratropical atmosphere, we correlate the PCS of these patterns to Z200a, the North Atlantic Oscillation and not the PNA emerges (not shown).

Though it is not the focus of this work, we investigate this spatiotemporal coevolution further we conduct a week-reliant MCA on the ensemble spread of SSTs (region: [15S,15N]; [0E,359E]) and Z200a in the north Atlantic region ([20N,80N]; [90W, 40E]). In the patterns (shown in supplemental Figure 5.3s) for the February initialization, we observe a development of the NAO in forecast week 2 which spurs Pacific to Atlantic flow. Ma et al. (2021) showed that the weakened northeast trade winds in the North Tropical Atlantic drive an anomalous flux of latent energy into the ocean, increasing SSTs there, driving more westerly flow. The westerly flow is accompanied by anomalous northerlies in the Northeastern subtropical Pacific, which are induced by negative surface latent heat flux anomalies. These anomalies eventually develop into the North Pacific Meridional Mode.

Through MCA analysis, no relationship can be discovered between the ensemble spread of the PNA pattern and ensemble spread of the tropical SSTs with any significant correlation. Younas & Tang (2013) showed that, despite having improved potential predictability, weekly subseasonal PNA forecast prediction skill was not significantly improved when similar models were run with sea surface coupling when compared to their uncoupled counterparts. Here, uncoupled models mean the atmospheric circulation is driven by persistent ocean forcing. The results presented here agrees with that result and offers an explanation for this lack of forecast improvement, as the slowly evolving SSTs are forced by the atmospheric evolution, but do not gain the tropical temperature spread anomalies necessary to drive a feedback into the midlatitude vorticity sourcing on the examined timescales.

5.4.2.2 Madden Julian Oscillation Teleconnection

The MJO is characterized by the eastward propagation of planetary-scale convective anomalies with travel across the Indo-Pacific warm pool with a typical phase speed of ~5 m s^{-}

^{1}. A full planetary MJO cycle is typically realized on the order of 30-60 days [as summarized in reviews, (W. K.-M. Lau & Waliser, 2011; C. Zhang, 2005)]. Supplemental Figure 5.4S shows the velocity potential at 200 – hPa of the 8 standard phases of the MJO from in the CSF-20C product for the November and February model initializations. The MJO teleconnects to the midlatitudes via two mechanisms 1) diabatic heating anomalies associated with the convective wave activity forcing upper level divergent winds, which propagate north and interact with the strong absolute vorticity gradient of the subtropical jet and induce a Rossby wave response (Mori & Watanabe, 2008) 2) via excitation of Rossby waves flanking to the east and west of the convective region which extract kinetic energy from the climatological mean flow and force a poleward propagating wave train (Adames & Wallace 2014). Figure 5.10 shows the 7-14 day averaged lagged response of Z200a (black contour), the 0-7 averaged lagged response of RWS' (colorfill) and divergent wind (vector) and the 200-hpa zonal wind (climatology, green contour) composite of the 8 phases active phases of the MJO in the CSF-20C. The composite is calculated across every ensemble member individually and conditioned on an RMM magnitude of greater than 1 in each phase. Additionally, we characterize model bias in the Z200a field, and show where the model is biased high in stippled regions and biased low in hatched regions. Bias is calculated with the same method as Deser et al. (2017), which is described in the section above.

Multiple examples exist in the literature of the examined teleconnection response to active MJO events both in idealized simulations (e.g, Adames & Wallace, 2014; Mori & Watanabe, 2008) and in observations (e.g., Henderson et al. 2017; Tseng et al. 2019; Wang et al., 2020a, 2020b) have found direct forcing of the PNA pattern. The February initialization (Fig 5.10b) shows a good representation of the PNA like teleconnection, linked with the tropical

MJO signal. In phases 1-3, MJO convection is centered over the Indian ocean (Figure 5.4s), generating a negative RWS' midlatitude response and a significant negative PNA. In phases 6-7, MJO convection is centered over the maritime continent and western Pacific, generating a positive RWS' midlatitude response and a significant positive PNA pattern. The patterns shown here are direct parallels to those identified in Wang et al. (2020a). A notable bias exists in the limb of the PNA centered over the Hawaiian region, in both positive and negative phases of the PNA. This region is a characteristic response region which flanks the equatorial Gill response to tropical heating (Adames & Wallace, 2014), acting as a generation cite for the Rossby wave response, and requires an in depth examination.

The November initialization (Fig. 5.10a) shows a different response and does not contain the often observed PNA-like wave train characteristic of tropical heating. The RWS' is generally muted compared to its the February counterpart. However, significant Z200a anomalies exist, but appear to have manifest as higher wave numbers than the PNA. The seasonality of the MJO impacts on North America temperature have been briefly examined (Jenney et al. 2019), and have been shown to evolve across a boreal winter season. The modulation of seasonal manifestations of the MJO teleconnection found here supports this finding. However, no process based intraseasonal evolution of the MJO teleconnection exists in the literature that the authors could identify. Figure 5.4S, shows that the November initializations contain strong divergent/convergent winds over the Indian Ocean, but there is a striking lack of divergent wind over the western tropical Pacific. Convection anomalies are in general muted over the Pacific November initialization. Figure 5.10 suggests a significant modulation of the teleconnection given the divergent wind and background state across a boreal winter, and a focused study is needed. To examine the discrepancy between the November and

February teleconnection we show the four terms of the RWS' and the divergence in the characteristic PNA RWS' anchoring region ([20N,40N];[140E,170W]) in Figure 5.11. The first term in the RWS' (S1, $-\bar{\xi}\nabla \cdot v_{\chi}'$) dominates, which is consistent with multiple previous studies (e.g., Hsu, 1996; Wang et al., 2020). This shows that the background seasonal absolute vorticity and the MJO-induced divergence play important roles in the RWS', clearly a stronger divergence term is seen in the February initialization (Fig. 5.10d). Interestingly, S3 ($v_{\chi}' \cdot \nabla\bar{\xi}$) the advection of mean absolute vorticity by anomalous divergent wind shows an important signal in the November initialization. Mechanistically, the teleconnection response to S3 is examined in Seo & Lee (2017), who found that shorter waves first travel along the westerly jet and then emanate at the jet exit region, prior to formation of the full PNA response. This mechanism could offer one insight into the discrepancy between the November and February Z200a patterns.

Ensemble Mean Anomaly Evolution

We investigate the spatiotemporal coevolution of VP (region: [15S,15N]; [0E,359E]) and Z200a in the PNA region ([25N,70N]; [140E, 60W]) through week-reliant MCA of the ensemble mean. The weekly averaged VP and Z200a are used as right and left fields, respectively. Figure 5.12 shows the left and right heterogenous fields of the first week-reliant MCA mode in the November (Fig. 5.12a) and February (Fig. 5.12b) initializations for the first six weeks of the weekly averaged forecast period. Additionally, the RWS' (red contour) and 200 hPa tropical divergent wind (vector) regressed on the normalized Z200a expansion coefficient is shown. A very clear propagating MJO is shown in the VP field. The November fields explain 26.1% of covariance, and just 5.9% of variance in the GPH field and 15.35% of variance in the tropical VP. The VP shows a very clear propagating wave which travels at a rate similar to the MJO phase speed (J. Wang et al., 2020b). Similar to the composite structures shown in Figure 5.10, the MCA November does not show a PNA like pattern, rather it shows high wave number Rossby wave structures initially emanating over the Asian continent. This gives further evidence that the early season teleconnection of the MJO does not manifest as a PNA pattern.

The February fields explain 39.9% of covariance, and 12.5% of variance in the GPH field and 16.35% of variance in the tropical VP. The February fields display a very clear PNA pattern in Z200a which persists through week 3 before losing its general structure week 4 and 5 as MJO phase amplitude weakens in the VP field. The MJO manifests as a phase 6 event, with divergent wind over the maritime continent. By the end of the 46-day forecast the MJO has traveled into phase 0, and the anomaly signal has weakened significantly as the anomaly variance is spread out of the ensemble mean and the ensemble mean tends towards climatology.

Ensemble Spread Anomaly Evolution

We now turn our attention to the spatiotemporal evolution of the ensemble VP spread and thus the error growth's effect on the PNA predictability. Figure 5.13 shows the evolution of the leading mode of variance in the ensemble spread of tropical VP region: [15S,15N]; [0E,359E]) for the November and February initializations (Figure 5.13a & 5.13b). This mode represents ~18% of total variance in both initializations. The ensemble spread analysis reveals that the leading mode of ensemble uncertainty in the tropical convective systems on weekly timescales is the MJO. A coherent anomaly in the ensemble spread does not emerge until week 2, indicating that the ensemble spread is relatively low up until that point. The bivariate MJO forecast correlation is shown in supplemental Figure 5.5S. At week 2 the MJO still has a skillful forecast (~0.7 correlation), but it is clear that anomalous forecast deviations have begun to emerge. Further, they are accompanied by divergent wind (shown in vector) which can communicate with the extratropics and add to the uncertainty of the PNA forecast.

We examine the spatiotemporal coevolution VP (region: [15S,15N]; [0E,359E]) and Z200a in the PNA region ([25N,70N]; [140E, 60W]) through week-reliant MCA of the ensemble spread. Figure 5.14 shows the left and right heterogenous fields of the first week-reliant MCA mode in the November (Fig. 5.14a) and February (Fig. 5.14b) initializations for the first six weeks of the weekly averaged forecast period. Additionally, the RWS' (red contour) and 200 hPa tropical divergent wind (vector) regressed on the normalized Z200a expansion coefficient is shown. The VP field readily matches the leading mode of ensemble spread (shown in Fig. 5.13) and is identified as the emerging MJO signal. The November fields explain 25.6% of covariance, and 3.9% of variance in the GPH field and 6.7% of variance in the tropical VP. A weak teleconnection emerges in week 2; sparked by the growing divergent winds generated in by the emerging MJO signal. Week 3-5 show Rossby wave trains emerging from the Asian continent and spanning the entirety of the Pacific. The patterns are similar to the Phase 4 Rossby wave train shown in the November Z200a/MJO phase composites (Fig. 5.10a). Significant Z200a anomalies emerge throughout the 6-week forecast, but do not manifest as the PNA signal.

The February fields explain 43.0% of covariance, and 7.5% of variance in the GPH field and 11.6% of variance in the tropical VP. In week 2, a VP anomaly and Z200a first emerges. Weeks 3-5 show the clear development of an internal mode of the positive PNA pattern, with the corresponding MJO traveling from phases 5-8. It is important to note that the SVD analysis is sign invariant, thus this pattern also represents the negative forcing of the PNA pattern from MJO phases 1-3. The MJO first manifests as a dipole over the Indian Ocean and Maritime continent, and weak anomalous RWS is overserved in the midlatitudes. This initial MJO growth condition represents model spread in the convective region just east of the maritime continent (corresponding to MJO phase 6/7 in the positive expression and phase 1/2 in the negative expression). Interestingly, this corresponds with the location optimal PNA growth conditions via tropical convection, and the optimal growth conditions in the Z200a field in Henderson et al., (2020), where linear inverse modeling is used to isolate the MJO derived PNA pattern. In week two the first signs of a cyclonic anomaly emerges over the east Pacific (~15°N, 140°W), this pattern retrogrades west and strengthens in weeks 3 and 4, merging with the lower limb of the PNA pattern, centered over the Hawaiian Isles. The easterly winds driven by this pattern collide with the westerly winds associated with MJO phase 5/6 over the Indian Ocean. This forces further midlatitude divergence and strengthens the associate RWS' anchored in the midlatitude jet, reinforcing the PNA pattern. A relatively small percent of variance is explained by the displayed 6 week evolution of the Z200a (7.5%). However, when MCA is performed to examine covariance between individual weeks 3, 4, or 5 (rather than the full 7 week cycle), the variance explained increased to ~15-20% of the Z200a spread field.

The February ensemble spread shows a clear influence on the uncertainty of the PNA forecasts in the MCA analysis. The internal modes of the MJO begins to be influential on the Z200a in week three in the weekly averaged forecasts. We now test whether this accounts for a significant portion of the uncertainty in the spread of the forecasted PNA index. To test this, we set up a simple linear model to examine the relationship between the anomalous spread of the MJO RMM indices 1 and 2 and the anomalous spread of the PNA forecast seven days later.

The model is express as $\sigma(PNA(\tau))' = x_1\sigma(RMM1(\tau-7))' + x_2\sigma(RMM2(\tau-7))'$, where τ is the examined forecast day. We fit this function for values x_1 and x_2 by minimizing the mean squared error. We then examine the correlation between the model fit and the observed PNA spread. Figure 5.15 shows the correlation of this model in the November and February initializations (Fig. 5.15a & 5.15b, respectively) throughout the 46-day forecast. The forecast is averaged forward in time with a simple 7 day rolling average filter. Additionally, we show the evolution of the PNA ensemble spread (red line). The model is shown to be significantly skillful in week 3 (red dots on forecast day ~17), of the February initialization. Showing that the uncertainty in the RMM indices can explain ~15% of variance in the uncertainty of the PNA forecast, lagged 7 days later. The November initialization does not show this same relationship. This is expected, as we have shown that the MJO teleconnection in November does not manifest as the PNA signal. Upon examination, we find that uncertainty in $RMM1$ is the dominant source of the forecast skill, with greater uncertainty in $RMM1$ driving greater uncertainty in the forecasted PNA index. RMM1 represents convection over the maritime continent, thus uncertainty in that location seems to drive the MJO related uncertainty in the February PNA forecast.

5.5 Summary and Discussion

This study investigates the subseasonal forecast skill of the Pacific-North American (PNA) pattern in the European Center for Medium-Range Weather Forecast's (ECMWF) coupled hindcast of the 20th century, across multiple time scales. We find that the February initializations are much more skillful than their analogous November counterparts. The November forecasts do not appear to be subject to boundary layer forcing and the dominant portion of their forecast skill is attributable to accurate representation of the slowly evolving background state in the initial condition. The November daily averaged forecasts of PNA have a correlation to observations of 0.5 or greater only out to forecast day 7. Weekly averaging increases the forecast skill by roughly four days per week with a final 4-week average forecast retaining skill out to day 22. A pseudo-persistence forecast demonstrates that the forecast skill in November is entirely due to the initial condition. Alternatively, the February forecasts are much more skillful. Correlation remains above 0.5 for the 1-day averaged forecast out to 20 days. Weekly averaging again adds forecast skill of roughly four days per week averaged. The 4-week forecast skill drops below 0.5 at 42 days. The pseudo-persistence forecast demonstrates that boundary condition forcing (via slowly varying climate modes) are clearly in affect in February.

We examine the two tropically derived drivers of the PNA, 1) El Niño Southern Oscillation (ENSO) and 2) the Madden-Julian Oscillation (MJO, Madden & Julian, 1971). Through, composite analysis and optimal correlation analysis in the ensemble mean, we find no tropical teleconnection to the PNA in the November forecast. Despite a strong tropical SST anomaly, the ENSO/MJO teleconnections do not appear to have begun, and thus boundary conditions forcing the forecast skill is low. Despite having roughly equal magnitudes of tropical divergence in November and February, the ENSO forced divergence does not reach the midlatitudes. Additionally, the midlatitude divergence associated with the MJO is much weaker in the midlatitudes in November in every phase. Compounding the lack of divergence, in November a weaker subtropical jet leads to comparatively absolute vorticity sourcing, resulting in a weak Rossby wave source (Sardeshmukh & Hoskins, 1988) for both MJO and ENSO events. Alternatively, the February initializations have a strong PNA/ENSO teleconnection and

PNA/MJO teleconnection, observed in both composite analysis and optimal correlation analysis of the model ensemble mean.

We then examine the growing modes of forecast uncertainty via ensemble spread analysis. The ensemble spread analysis enables a much larger sample size and available observations and highlights the growing anomalies driving forecast divergence from the ensemble mean. On weekly time-scales we can identify no feedback between uncertainty in the ensemble SST forecast and uncertainty on the PNA. Due to the slowly evolving nature of the uncertainty growth in the sea surface anomalies do not grow large enough to drive PNA divergence. We identify the MJO as the largest growing mode of forecast uncertainty in the tropical atmosphere accounting for ~18% of variance in both November and December. Despite still having a skillful MJO forecast, in week 3, independent MJOs have developed in the model spread significantly affecting the PNA forecast certainty. Particularly error growth in RMM1 drives significant error growth in the PNA.

Finally, empirical model development and particularly modern machine learning methods for subseasonal forecasting, continue to grow in popularity as more data driven forecast systems show that they have model skill to rival numerical weather prediction (e.g., Rasp et al., 2020). It is import to train these methods with the representative data in order to optimize forecast skill. This study demonstrates that there is a significant evolution of tropical forcing, and persistence in the PNA across a boreal winter season. Care must be taken when designing empirical methods in order to account for this intraseasonal forecast evolution.

5.6 Acknowledgement

This chapter in part is currently being prepared for submission for publication of the material. Chapman W. E., Subramanian, A. C., Xie S. P., Ralph, F. M.. The dissertation author was the primary investigator and author of this paper.

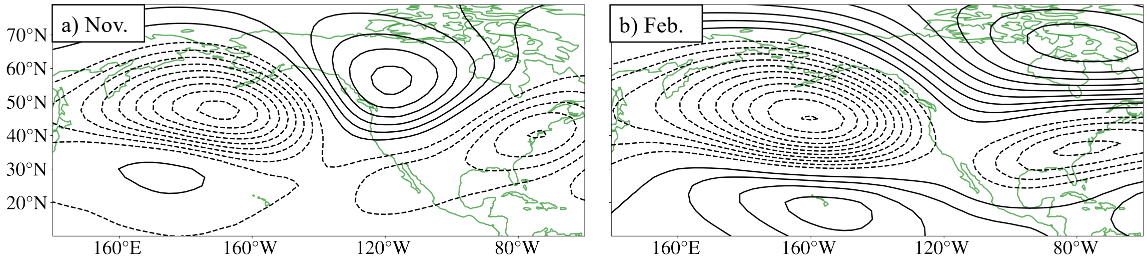

Figure 5.1 Leading modes of observed variability [10N,80N], [140E, 60W] in the area weighted 200-hPa CERA20-C observations over the period 1901-2010 for climatological first 30 days of a) November and b) February, representing 21.0% and 32.0% of total variance, respectively.

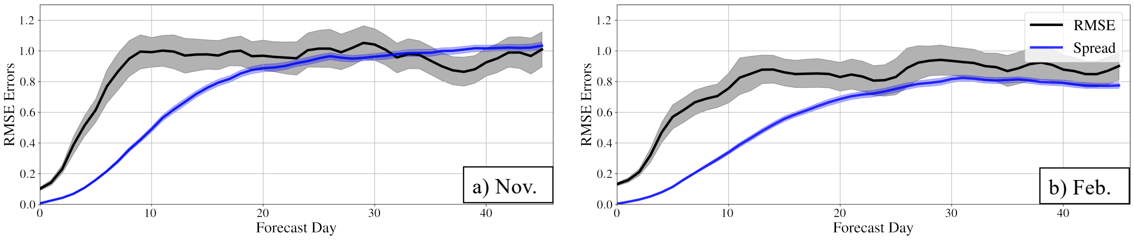

Figure 5.2 PNA RMSE (CERA-20C vs. CSF-20C) and ensemble spread as a function of forecast lead at daily time-scales a) November initialization and b) February initializations in the CSF-20C.

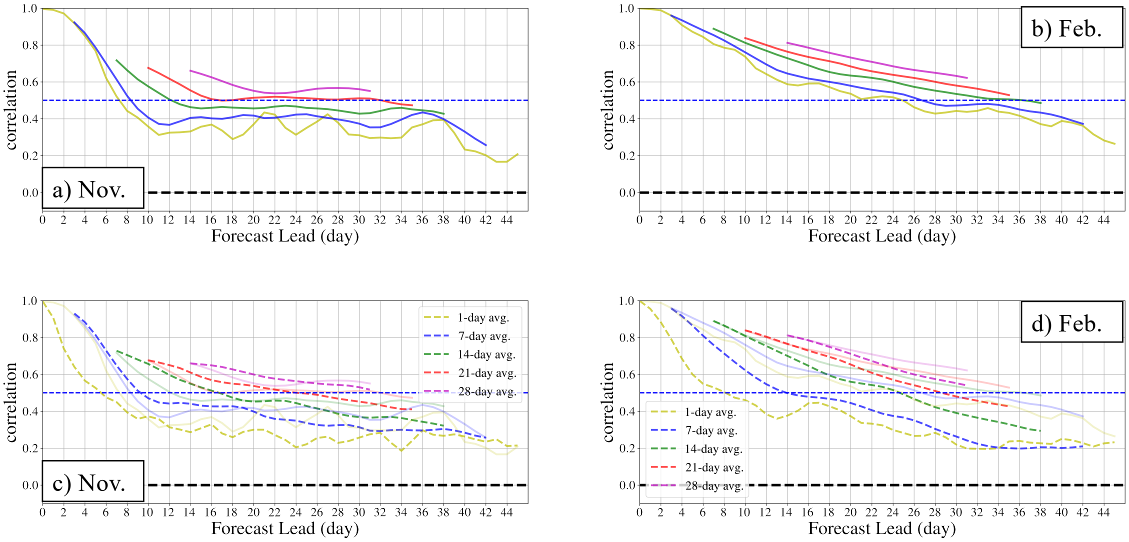

Figure 5.3 PNA correlation as a function of lead time for different time scales in the November (a) and February (b) forecast initializations and correlation of the initial averaged N-days (see legend) held constant (see text for details) in the November (c) and February (d) forecast initializations (dashed lines). Additionally, forecast correlation is shown as reference lines (solid lines) for panels (c) and (d), in their respective forecast months.

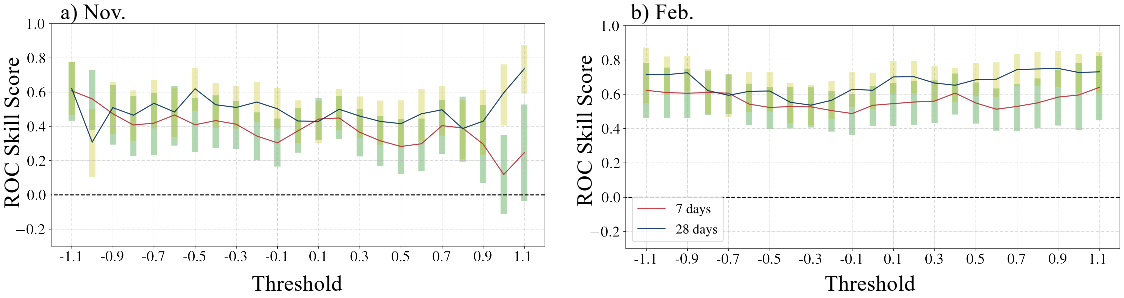

Figure 5.4 The ROC skill score for PNA events at different thresholds, for the 4-weeks averaged forecast, centered on forecast day 28. The 90-10% uncertainty range is also shown, as determined by bootstrap with resampling 1000 times.

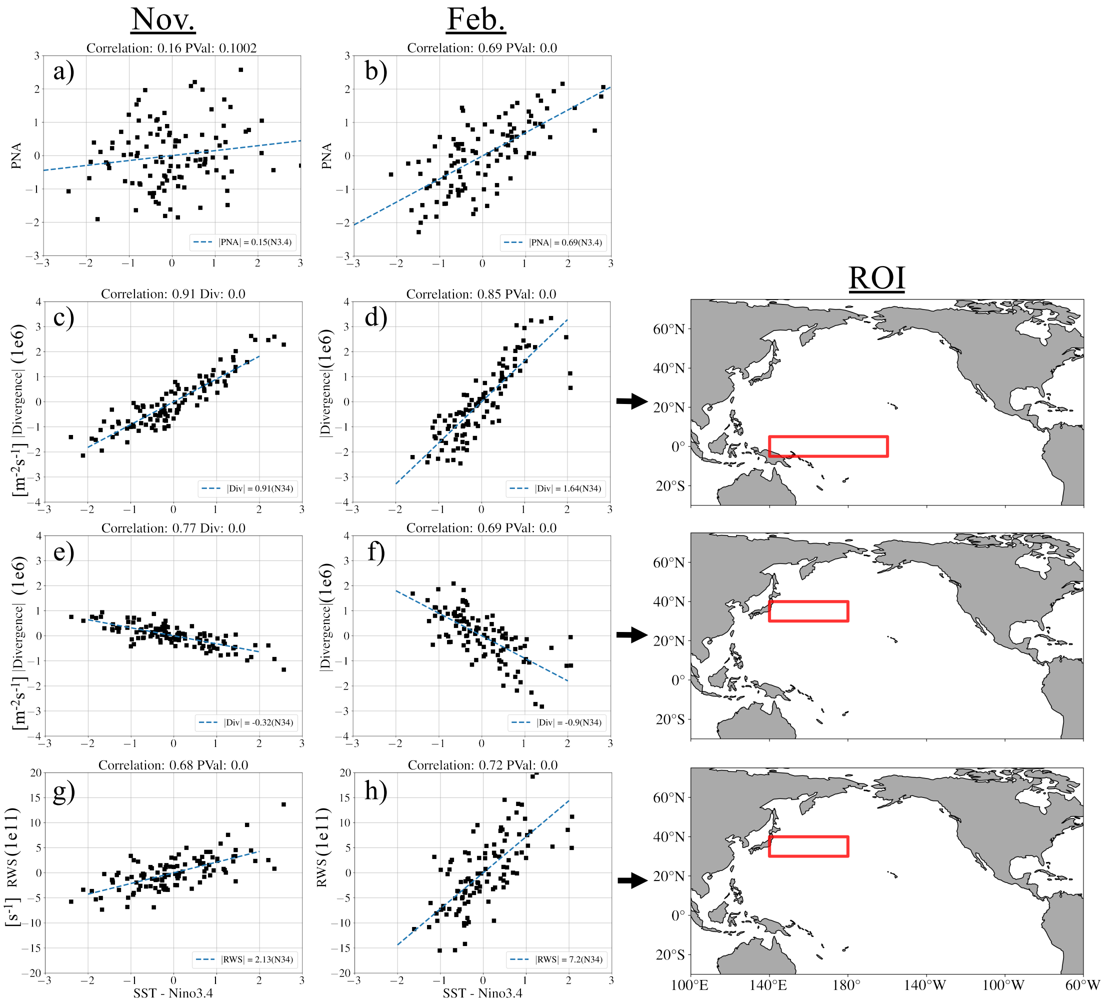

Figure 5.5 Scatter diagrams of PNA Magnitude (a,b), Tropical 200-hPa Divergence (c,d), Midlatitude 200-hPa Divergence (e,f), and 200-hPa Rossby Wave Source anomaly (g,h). In the ensemble mean of the 4-week average forecast of the ensemble mean in the November (a,c,e,g) and February (b,d,f,h) initializations against the concurrent SST anomaly averaged in the region: [5N,5S], [170W, 100W]. Regressions lines show the linear relationship between the variables, correlation and p-value are shown above each panel. The averaging region of interest is for each variable is shown in the panels on the right (see text).

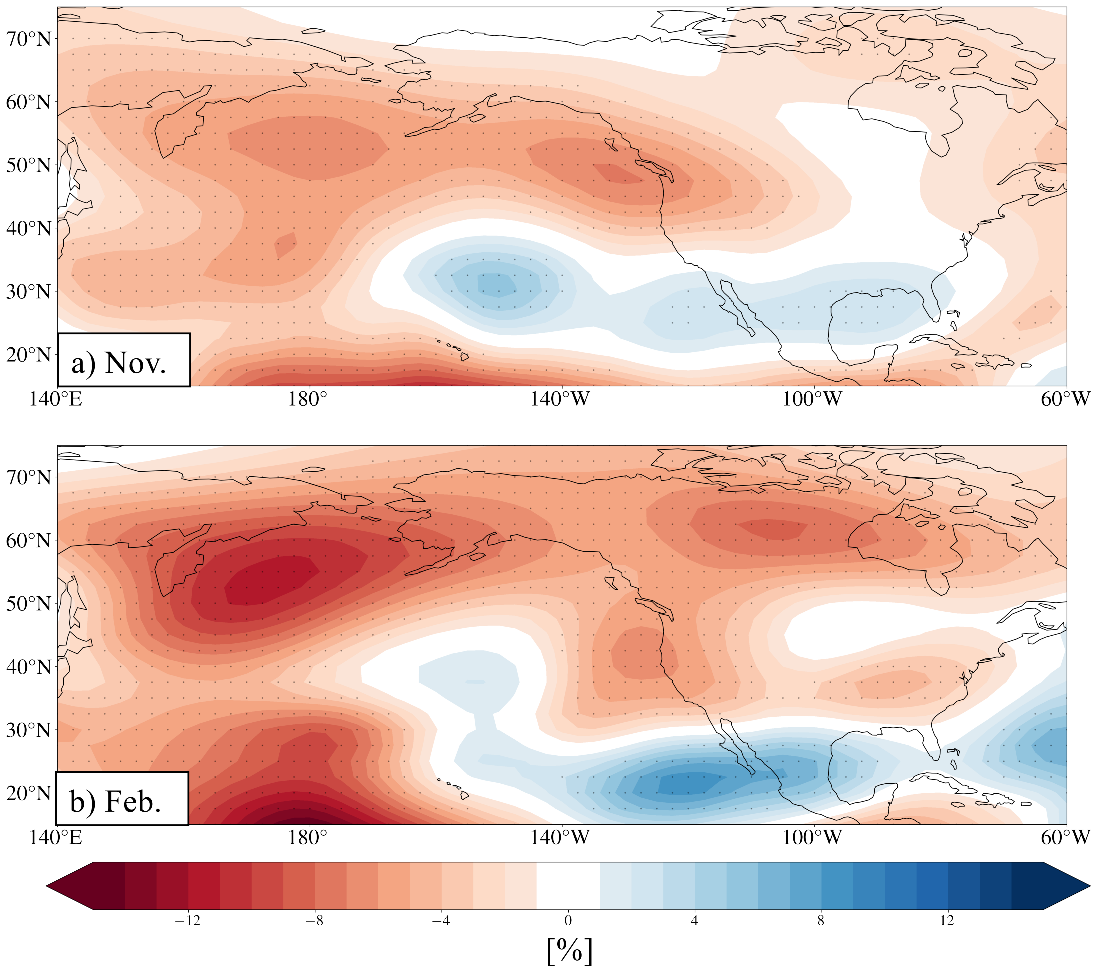

Figure 5.6 Percent of modulation to atmospheric variability as a function of 1 standard deviation change in averaged SST anomaly in the region: [5N,5S], [170W, 100W] via regression of the ensemble spread on concurrent SST anomalies in the November (a) and February (b) forecasts. Stippling shows variable significance determined by boostrap with replacement (1000 times) and examination of the 95th and 5th percentiles of the pseudo distribution.

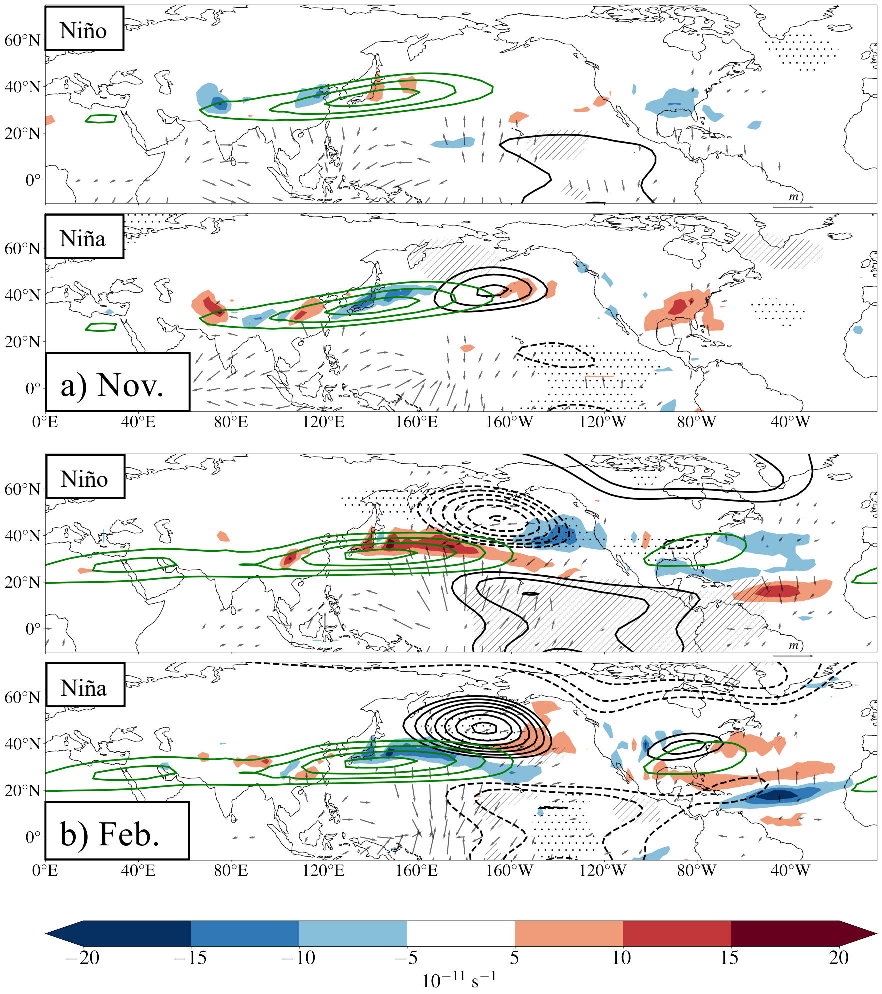

Figure 5.7 7-14 day averaged lagged response of Z200a (contour, black 10m interval, positive: solid; negative: dashed), 0-7 Rossby wave source anomaly averaged lagged response (colorfill), and climatological 200-hPa zonal wind (green; contour levels: [40,50,60,70,80] ms^{-1}) in each ENSO phase, for the November (a) and February (b) model initializations. Stippled regions indicate CSF-20C is systematically biased high, hatched regions indicate CSF-20C is systematically biased low.

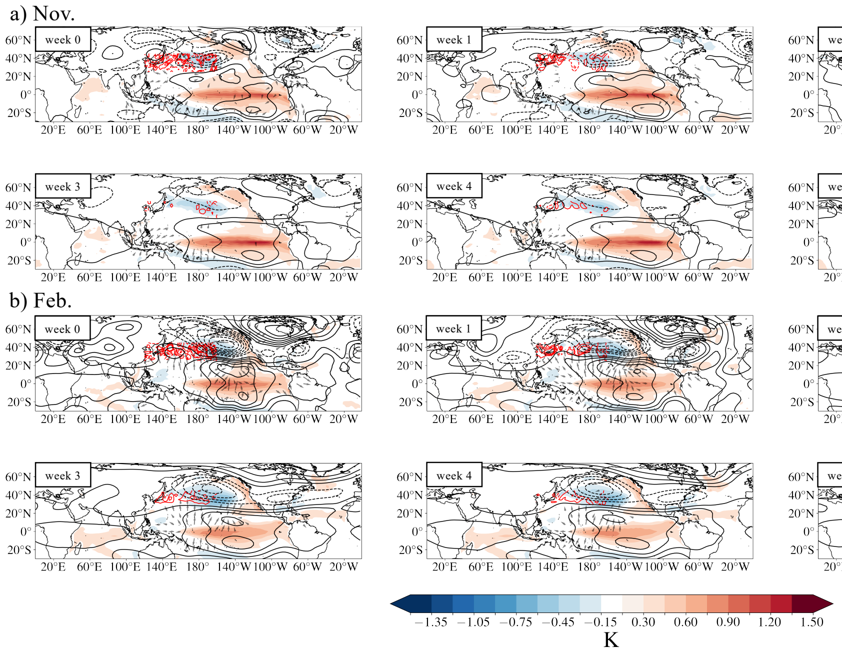

Figure 5.8 Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble mean weekly averaged SST anomalies (region: [-15S,15N],[0W, 360W]) and geopotential height (region: ([25N,70N]; [140E, 60W]). Additionally we show the non-rotational component of the wind (vector, reference arrow: 4 ms^{-1}, shown only when the windspeed is above 0.5 ms^{-1}) and Rossby Wave Source anomaly (contour, red [-15,-10,-5,5,10,15]*1e11 s^{-2}) regressed on the left expansion coefficient normalized to unit variance. Anomaly evolution is shown on averaged weekly intervals in the November (a) and February (b) initializations, the fields represent 70.6% and 68.7% of covariance, respectively.

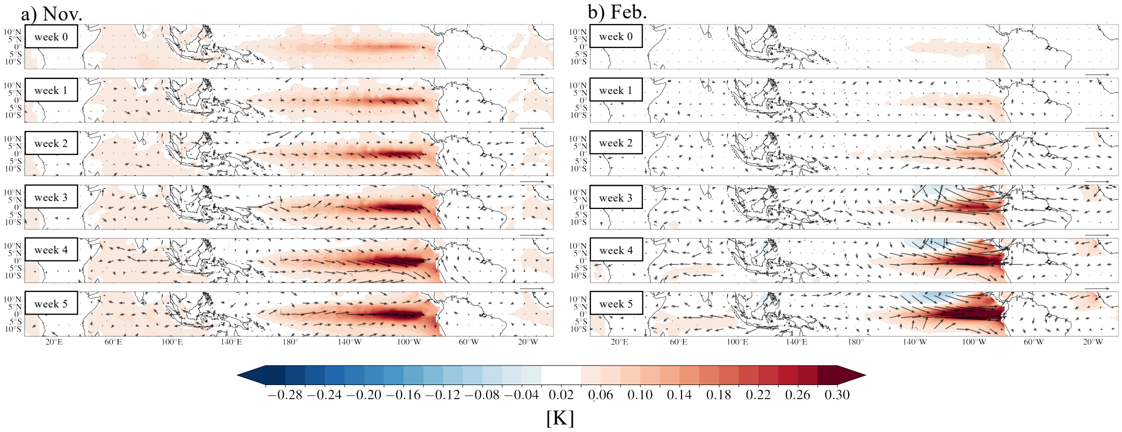

Figure 5.9 Anomalous sea surface temperature and zonal and meridional wind 850-hPa regressed on the extended EOF of SST (week 0-5, region: [-15S,15N],[0W, 360W]) intermember spread normalized to unit variance in, a) November initialization and b) February initializations, representing ~5% and ~5% of total variance, respectively. Reference wind vector represents 3 m/s.

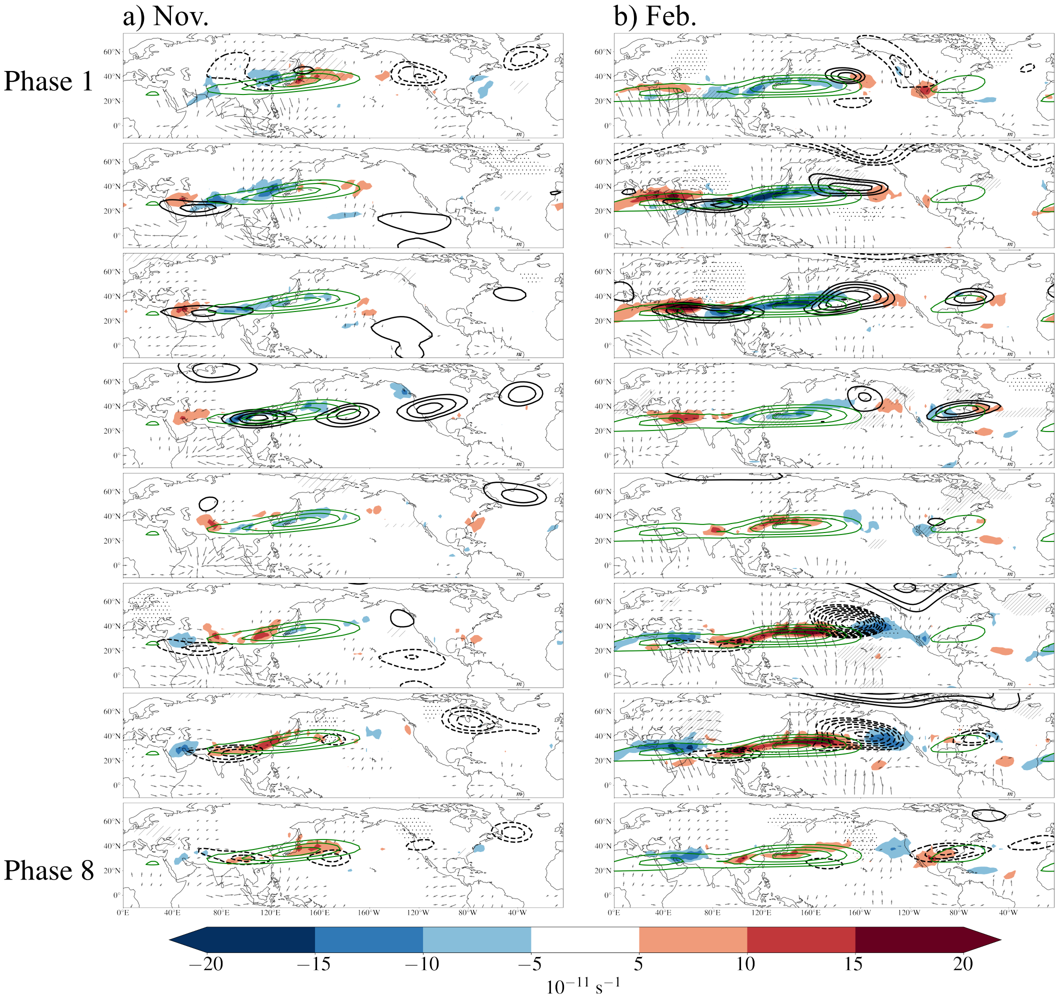

Figure 5.10 7-14 day averaged lagged response of Z200a (contour, black 10m interval, positive: solid; negative: dashed), 0-7 Rossby wave source anomaly averaged lagged response (colorfill), and climatological 200-hPa zonal wind (green; contour levels: [40,50,60,70,80] ms^{-1}) in each MJO phase, for the November (a) and February (b) model initializations. Stippled regions indicate CSF-20C is systematically biased high, hatched regions indicate CSF-20C is systematically biased low.

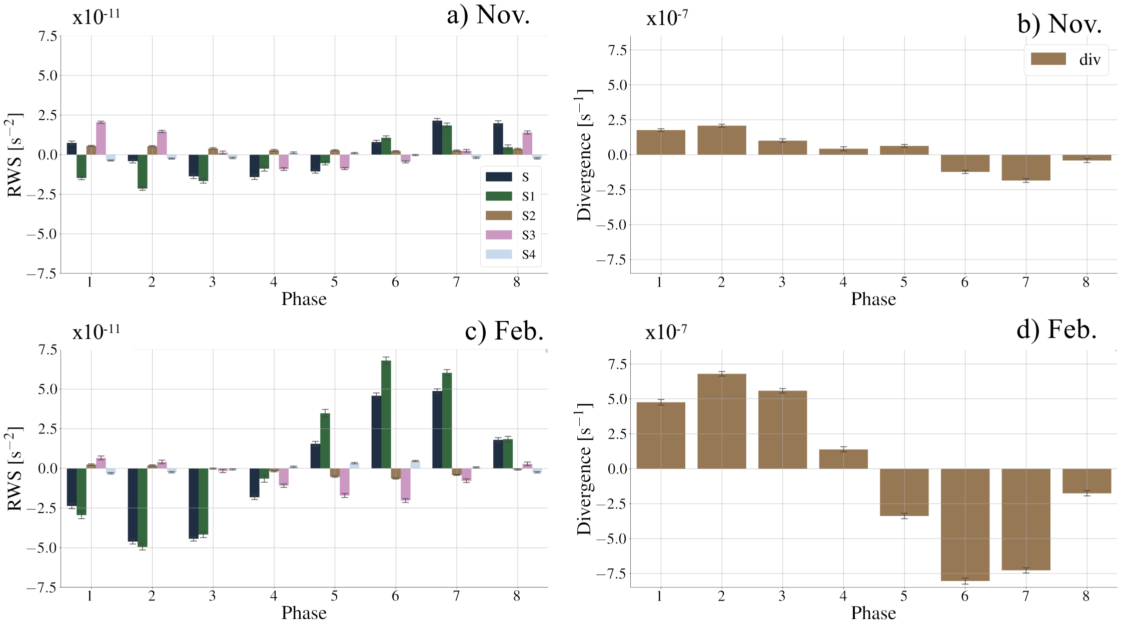

Figure 5.11 0-7 day averaged lagged four components of the Rossby Wave Source anomaly (a,c) and the divergence (b,d) in the eight MJO phases in region: [20N,40N];[140E,170W] for the November (a,b) and February (c,d) model initializations. Error bars represent the 10th and 90th confidence interval as determined by bootstrap with replacement 1000 times.

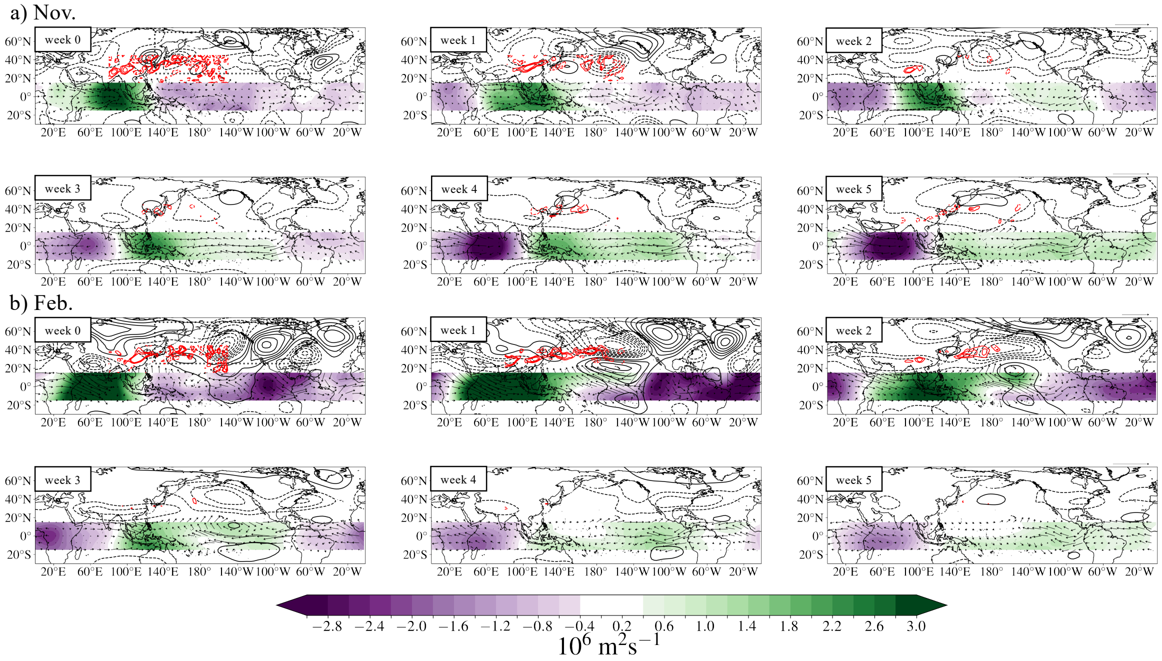

Figure 5.12 Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble mean weekly averaged Velocity Potential anomalies (region: [-15S,15N],[0W, 360W]) and geopotential height (region: ([25N,70N]; [140E, 60W]). Additionally we show the non-rotational component of the wind (vector, reference arrow: 4 ms^{-1}) and Rossby Wave Source anomaly (contour, red [-15,-10,-5,5,10,15]*1e11 s^{-2}) regressed on the left expansion coefficient normalized to unit variance. Anomaly evolution is shown on averaged weekly intervals in the November (a) and February (b) initializations, the fields represent 26.1% and 39.9% of covariance, respectively.

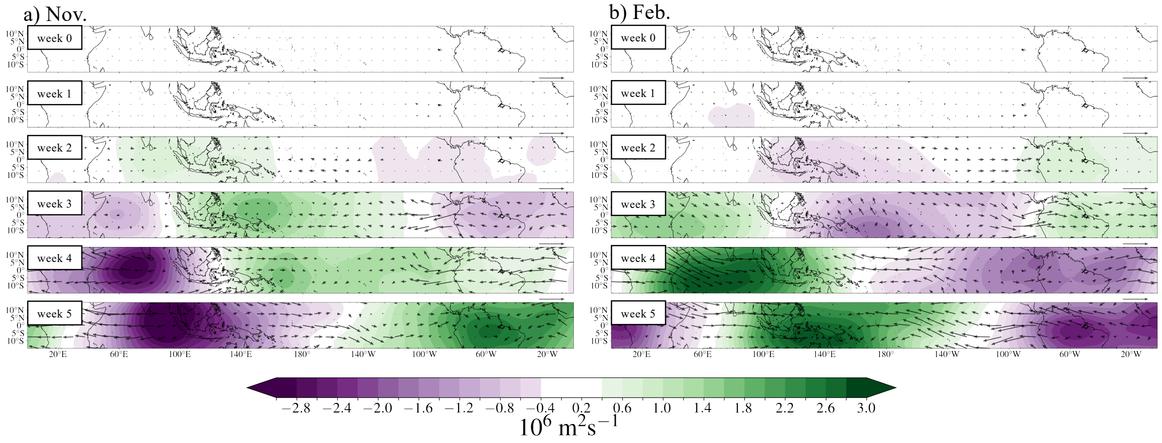

Figure 5.13 Anomalous velocity potential and zonal and meridional wind 200-hPa regressed on the extended EOF of SST (week 0-5, region: [-15S,15N],[0W, 360W]) intermember spread normalized to unit variance in, a) November initialization and b) February initializations, representing ~18% and ~18% of total variance, respectively. Reference wind vector represents 3 ms^{-1} in weeks 0-2 and 8 ms^{-1} in weeks 3-5.

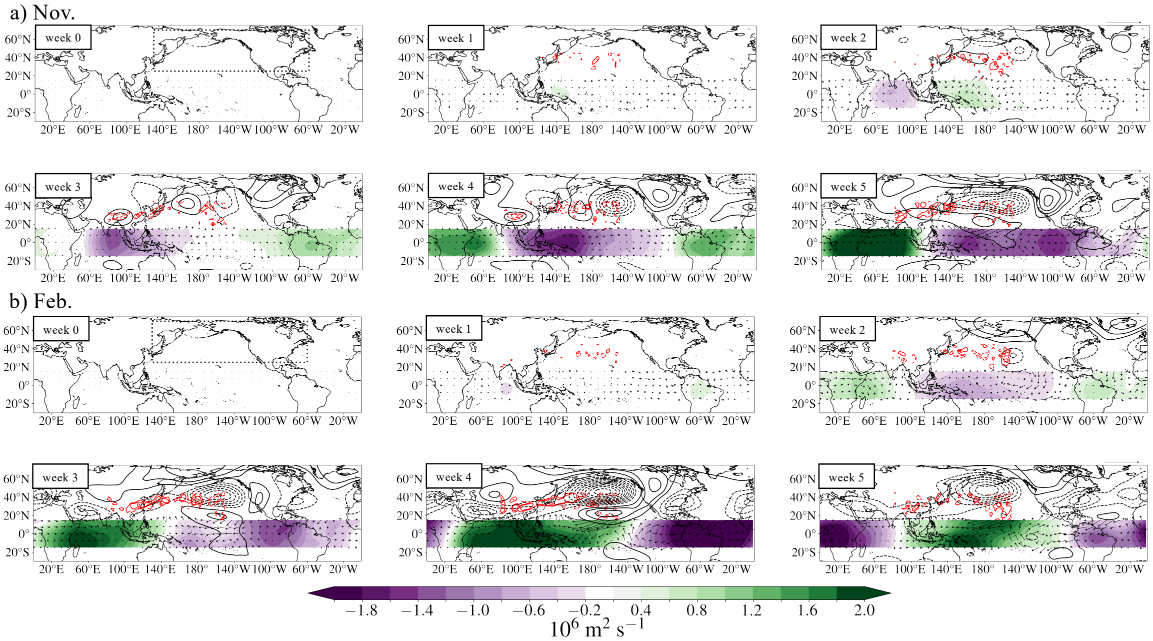

Figure 5.14 Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble spread weekly averaged Velocity Potential anomalies (region: [-15S,15N],[0W, 360W]) and geopotential height (region: ([25N,70N]; [140E, 60W]). Additionally we show the non-rotational component of the wind (vector, reference arrow: 3 ms^{-1} in weeks 0-2; 8 ms^{-1} in weeks 3-5) and Rossby Wave Source anomaly (contour, red [-15,-10,-5,5,10,15]*1e11 s^{-2}) regressed on the left expansion coefficient normalized to unit variance. Anomaly evolution is shown on averaged weekly intervals in the November (a) and February (b) initializations, the fields represent 25.6% and 43.0% of covariance, respectively.

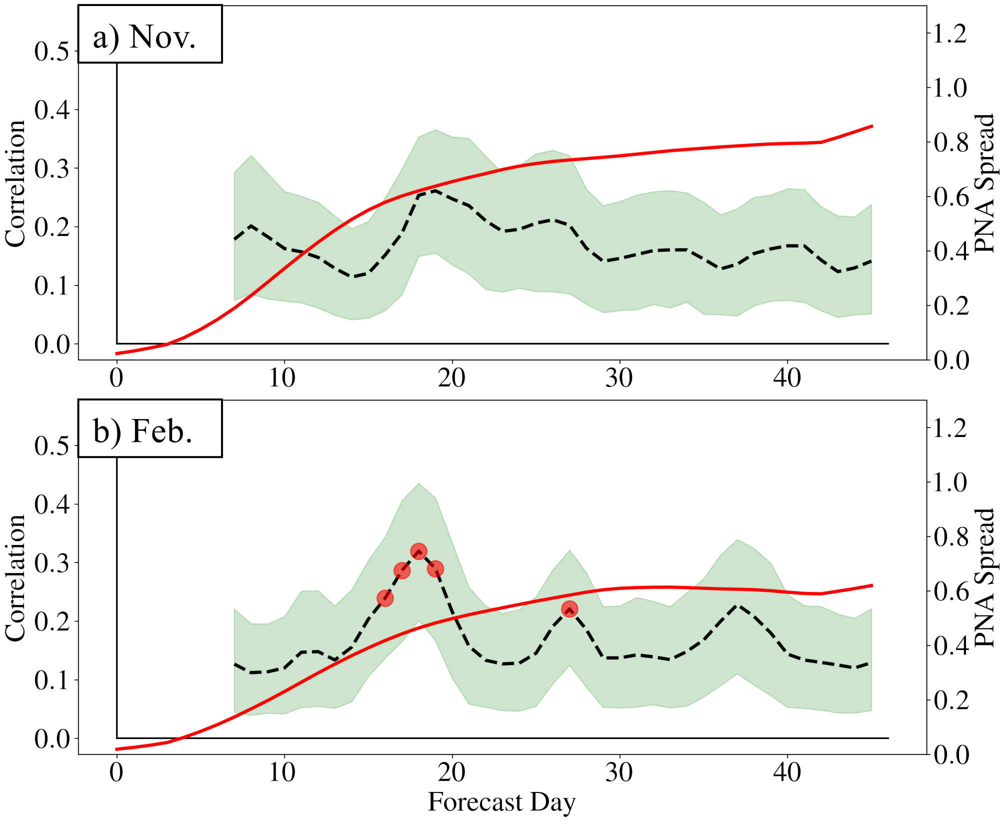

Figure 5.15 Correlation as a function of forecast lead time of the anomalous spread in the weekly averaged RMM indices and the anomalous spread in the 7-day lagged weekly averaged PNA forecast. Red dots indicate when the regression is significant at the 5% level as determined by a standard t-test. Additionally, PNA ensemble spread is shown (red line; right y-axis).

5.7 Supporting Information

Here, we present additional figures and information for Chapter 4 in support of the material presented above.

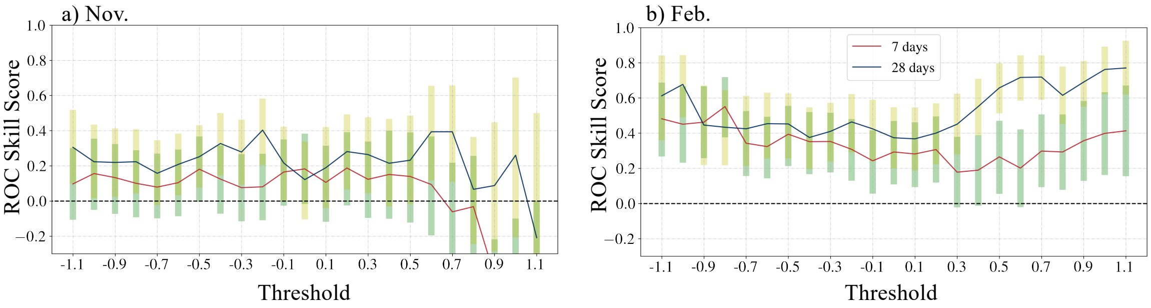

Figure 5.1S The ROC skill score for PNA events at different thresholds, for the 4-weeks averaged forecast, centered on forecast day 28. The 90-10% uncertainty range is also shown, as determined by bootstrap with resampling 1000 times.

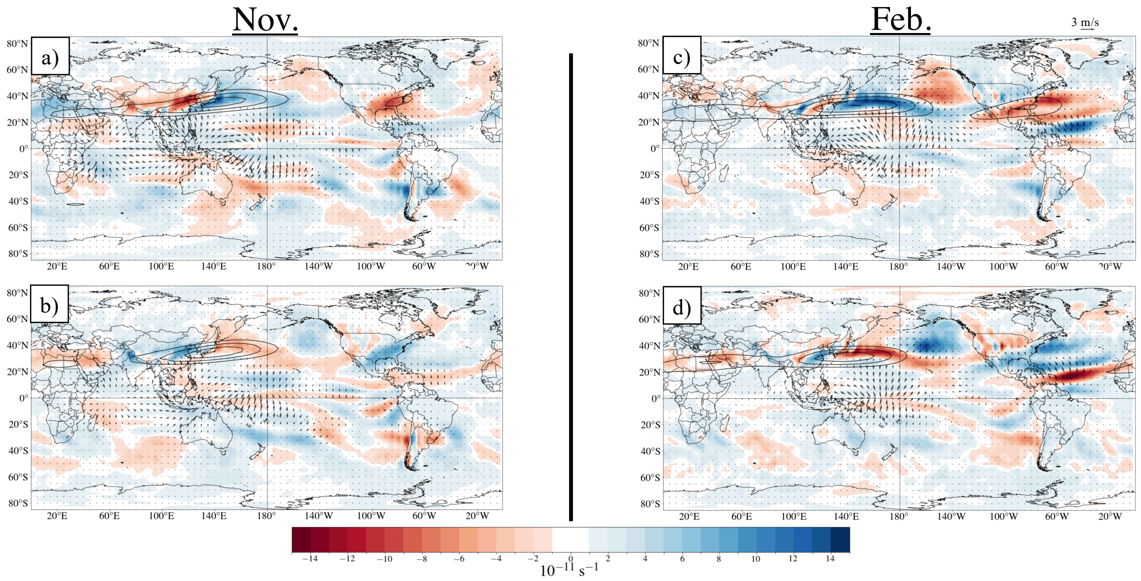

Figure 5.2S Composite Rossby wave source (colorfill), irrotational components of the 200-hPa wind (vector), and zonal wind at 200-hPa (climatology + anomaly, contour [40,50,60,70,80] ms^{-1}) for El Niño (a,c) and La Niña (b,d) seasons in the November (a,b) and February (c,d) CSF-20C model initializations.

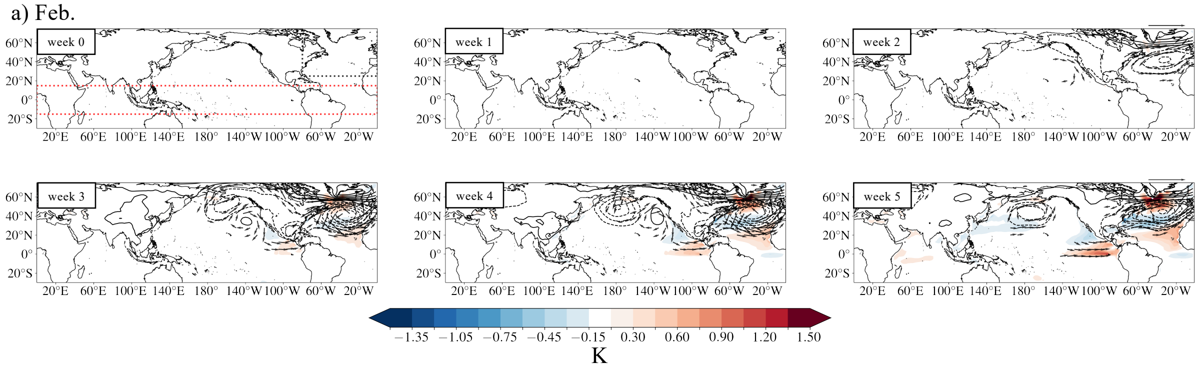

Figure 5.3S Left (contour; black at 10m intervals) and right (color shading) heterogenous fields of the first week-reliant MCA mode of the ensemble spread weekly averaged Velocity Potential anomalies (region: [-15S,15N],[0W, 360W]) and geopotential height (region: ([20N,80N]; [55W, 0]). Additionally we show the non-rotational component of the wind at 850-hPa(vector, reference arrow: 4 ms^{-1}) and Rossby Wave Source anomaly (contour, red [-15,-10,-5,5,10,15]*1e11 s^{-2}) regressed on the left expansion coefficient normalized to unit variance. Anomaly evolution is shown on averaged weekly intervals in the February (a) initializations, the fields represent 45.5% of covariance.

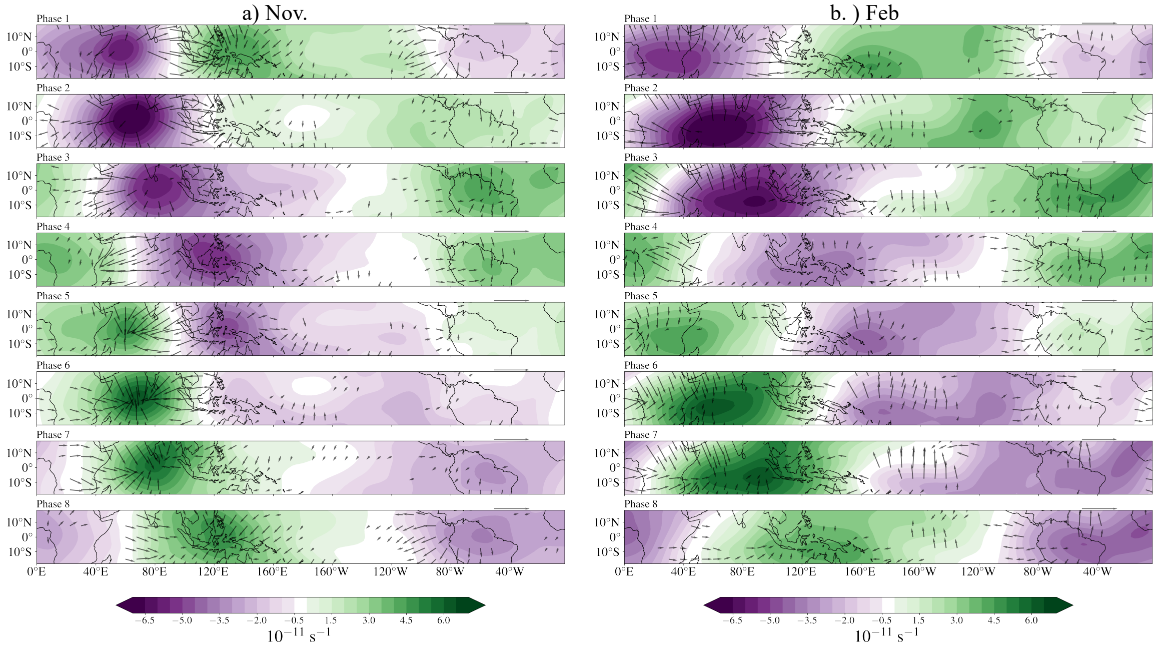

Figure 5.4S Composite of anomalous VP at 200-hPa and 200-hPa irrotational components of the wind anomalies (vector) for each phase of the MJO in the November (a) and February (b) initializations. Reference vector is 3 ms^{-1}. Negative VP represents upper level divergence.

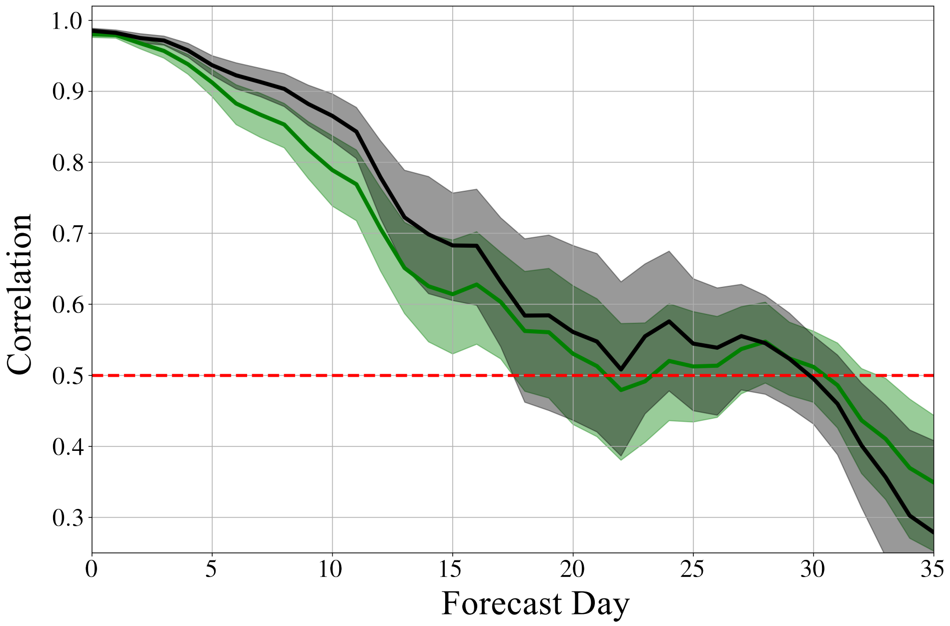

Figure 5.5S Bivariate correlation of the CSF-20C MJO and CERA-20C MJO observations by forecast lead compiled for the November and February initializations in when initialized in active (black) and inactive (green) phases of the MJO.

Chapter 6

Conclusion

6.1 Summary of Contributions

The purpose of this thesis is to examine and advance North American weather predictability from weather to subseasonal time-scales. Specifically, it focuses on 1) developing machine learning/deep learning methods and models to improve predictability through numerical weather prediction (NWP) post-processing on weather time-scales (0-7 days) and 2) examining the physical mechanisms which govern the evolution of the predictable components and noise components of teleconnection modes on subseasonal time-scales (7 days - 1 month).

NWP deficiencies (e.g., sub-grid parameterization approximations), nonlinear error growth associated with the chaotic nature of the atmosphere, and initial condition uncertainty lead initial small forecast errors to eventually result in weather predictions which are as skillful as random forecasts. A portion of these forecast errors are inherent to the NWP models alone, systematic biases. The first two chapters of this dissertation developed cutting-edge vision-based deep learning algorithms to advance the current state-of-the-art NWP post-processing and correct systematic NWP biases.

Chapter 2 tests the utility of convolutional neural networks (CNN) as a postprocessing framework for improving the National Center for Environmental Prediction's Global Forecast System's (GFS) integrated vapor transport (IVT) forecast field over the Eastern Pacific and Western United States. IVT is the characteristic field of atmospheric rivers, which provide over 65% of yearly precipitation at some western U.S. locations. The method reduces full field root mean squared error (RMSE) at forecast leads from 3 hours to 7 days (9-17% reduction), while increasing correlation between observations and predictions (0.5-12% increase). This represents a ~1-2-day lead time improvement in RMSE. Decomposing RMSE shows that random error and conditional biases are predominantly reduced. Systematic error is reduced up to 5-days forecast lead, but accounts for a smaller portion of RMSE. This work demonstrates CNNs potential to improve forecast skill out to 7 days for precipitation events affecting the western U.S.

Chapter 3 takes a step further and leverages Deep Learning (DL) post-processing methods to obtain reliable and accurate probabilistic forecasts from single-member numerical weather predictions of IVT. Using a 34-year reforecast, based on the Center for Western Weather and Water Extremes West-WRF mesoscale model of North American West Coast IVT, the dynamically/statistically derived 0–120-hour probabilistic forecasts for IVT under atmospheric river (AR) conditions are tested. These predictions are compared to the Global Ensemble Forecast System (GEFS) dynamic model and the GEFS calibrated with a neural network. Additionally, the DL methods are tested against an established, but more rigid, statistical-dynamical ensemble method (the Analog Ensemble). The findings show, using continuous ranked probability skill score and Brier skill score as verification metrics, that the DL methods compete with or outperform the calibrated GEFS system at lead times from 0-48 hours and again from 72-120 hours for AR vapor transport events. Additionally, the DL methods generate reliable and skillful probabilistic forecasts. The implications of varying the length of the training dataset are examined and the results show that the DL methods learn relatively quickly and ~10 years of hindcast data are required to compete with the GEFS ensemble. Additionally, this chapter lays out a transfer learning framework which can be readily applied to new model updates, or in the case of this work, to mitigate the potential effects of changing error statistics due to secular climate change, or slowly varying decadal variability.

The second half of this thesis shifts focus to subseasonal time scales and examines predictability in the Pacific North American (PNA) sector in boreal winter. Particularly, it investigates the physical mechanisms involved in the intraseasonal modulation of atmospheric Signal-to-Noise (SN), and how it is affected by slowly varying climate modes (ENSO and MJO). These mechanisms are explored using a fully coupled hindcast of the 20th century, showing that the increased SN leads to high model forecast skill at subseasonal timescales in particular forecast windows of opportunity. Additionally, we reveal the MJO as the largest growing mode of tropical forecast uncertainty which directly influence PNA forecast certainty.

Chapter 4 leverages a high-resolution atmospheric general circulation model simulation of unprecedented ensemble size, and examines the potential predictability of monthly anomalies under El Niño Southern Oscillation (ENSO) forcing and background internal variability. This chapter reveals the pronounced month-to-month evolution of both the ENSO forcing signal and internal variability. Internal variance in upper-level geopotential height decreases ~10% over the North Pacific during El Niño as the westerly jet extends eastward, allowing forced signals to account for a greater fraction of the total variability, and leading to increased potential predictability. We identify February and March of El Niño years as the most predictable months using a signal-to-noise analysis. In contrast, December, a month typically included in teleconnection studies, shows little-to-no potential predictability. This chapter shows that the seasonal evolution of SST forcing, and variability leads to significant signal-to-noise relationships that can be directly linked to both upper-level and surface variable predictability for a given month. The stark changes in forced response, internal variability, and thus signal-to-noise across an ENSO season indicate that subseasonal fields should be used to diagnose potential predictability over North America associated with ENSO teleconnections. Using surface air temperature and precipitation as examples, this study provides motivation to pursue ‘windows of forecast opportunity’, in which statistical skill can be developed, tested, and leveraged to determine times and regions in which this skill may be elevated.

In Chapter 5, using ensemble hindcasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) coupled model of the 20th century (period 1901–2010), the subseasonal forecast skill of the Pacific North American (PNA) pattern and the spatiotemporal evolution in the covariability of the PNA and 1) tropical sea surface temperatures (SST) and 2) the Madden Julian Oscillation (MJO) in both the November and February initializations is investigated. Significant intraseasonal dependence of forecast skill and tropical forcing is demonstrated. The February initializations show a much more skillful subseasonal PNA forecast (compared to the November initializations). Additionally, the forecast skill derived from the low-frequency variability of the initial condition is much more valuable in February than in November. Two known drivers of subseasonal PNA forcing are investigated, El Niño Southern Oscillation (ENSO) SSTs and the MJO. The covariability in the ensemble mean and ensemble spread is investigated with week-reliant singular value decomposition (SVD), which treats each variable in a given average weekly forecast sequence as a single time step. The leading mode of the ensemble spread in the SST/PNA SVD shows only response in the extratropical atmosphere forcing the tropical ocean indicating that on subseasonal time-scales the uncertainty in SST ensemble spread shows little influence to PNA predictability. The MJO is revealed as the leading mode of ensemble spread in the tropical atmosphere. The February MJO/PNA SVD shows strong PNA modulation beginning in week 3 with the growth of a northeast Pacific cyclonic/anticyclonic retrograding west and enforcing the PNA pattern. However, this pattern is notably lacking in the November initialization. Due to the large sample size provided by this simulation, we show that uncertainty in the MJO significantly influences uncertainty in the PNA forecast.

6.2 Future Directions

Machine learning / deep learning (ML / DL) is currently in an explosive growth phase for algorithmic development of NPW post-processing. This dissertation presents the first use case of convolutional neural networks, a type of computer vision-based DL, for model post-processing. Compared to previous post-processing algorithms, these methods offer extreme flexibility to ingest spatiotemporal ancillary predictors and have unique training regimes which systematically prevent overfitting to data. Chapter 3 introduced a method of uncertainty quantification for deterministic forecast fields, this method outperforms fully dynamically derived forecast models in every examined probabilistic skill metric. However, the forecasted fields are prescribed from variational distributions; non-parametric methods, could potentially yield a more flexible forecast system that could be used ubiquitously across desired predictand variables, and early work has shown that this is potentially an exciting pathway forward (e.g., Bremnes, 2020). Simple DL latent space analog matching has yielded initial exciting results in this domain, and is an echo back in time to the original forecasters who looked for forecast analogs (see introduction) to predict the weather.

Post-processing systems rarely are used to inform on the physics of model error growth, yet these systems can accurately characterize this growth based on the input parameters. Thus, there is scope for these systems to be leveraged to examine the largest growing modes of forecast uncertainty in order to inform on dynamic model ensemble creation. Finally, we have seen the success of these methods of identifying and correcting models post-forecast. Online learning and adjustment could be leveraged in order to actively correct known climate model biases by discovering the conditional error tendencies within nudging algorithms, early work has shown this can be successful in simplified forecast models (Brajard et al., 2021). For this application, the network would be required to obey known conservation laws to prevent model drift in long system integration. Beucler et al., (2021) has shown model 'constraint network layers' enforce physical law conservation into climate model simulations for use in subgrid parameterizations, this same concept could readily be applied to the model bias problem.

The demand for skillful forecasts at lead times of two weeks to two months continues to grow. Accurate forecasts, at these time scales would dramatically affect nearly every modern societal sector. As the desired forecast sit outside the theoretical limit of NWP predictability, subseasonal forecast skill will not be achieved by a simple extension of our current forecast systems (or a dramatic increase of model resolution). This dissertation makes it clear that even leveraging our most influential climate modes (ENSO / MJO) on intraseasonal timescales requires intentional study of the physics which determine the downstream atmospheric forcing, or modulation of internal variability. Thus, in statistical model development, neglecting the physics of the teleconnection response will result in suboptimal utilization of training data which is detrimental to potential forecast skill. Subseasonal forecasts are neither weather nor seasonal forecasts, and they should not be treated as such. Blending statistical/dynamical approaches could lead to our most successful forecast systems, and have shown early promise (Henderson et al., 2020; Weyn et al., 2021) .

Moving forward, there are several outstanding questions adjacent to this dissertation that can be the focus of future research. The intraseasonal dependence of MJO and ENSO PNA teleconnection forecast skill has been examined, yet O'Reilly et al., (2020) has shown the decadal strength of the seasonal teleconnections is significantly modulated by decadal climate mode variability this calls into question whether subseasonal forcing and internal variability is equally affected by decadal variability? Initial work with the ECMWF 20th century hindcast shows a significant modulation of teleconnection strength in given MJO phases across the last decade, yet more focused research could be illuminating. It has been shown that in a future climate the MJO teleconnection is actively affected by the mean state midlatitude jet amplification (W. Zhou et al., 2020), understanding of the decadal variability which modulates this bias would lead to a more complete picture of the future state of these teleconnection patterns.

This dissertation has shown the dependence of the internal variability and forcing of the atmosphere on the midlatitude background state. The MJO is touted as the leading mode of intraseasonal tropical variability and is likely the largest source of subseasonal forecast skill in ENSO neutral years. Yet, focused simplified modeling studies on the intraseasonal development of the MJO teleconnection are notably lacking in the literature. These studies are drastically important, as statistical methods are actively being developed to issue forecasts in the subseasonal time range and the trustworthy observational record is woefully short in order to develop robust teleconnection statistics. Open questions remain about the growth of the PNA teleconnection, the initialization mechanism of optimal teleconnection growth (tropical divergence or flanking Rossby waves) forced by the MJO, and the role of synoptic-scale transients in generating and maintaining the teleconnection (Adames & Wallace, 2014; Henderson et al., 2020).

Chapter 7

References

Abadi, Martín, Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. http://arxiv.org/abs/1603.04467

Abbe, C. (1895). The Needs of Meteorology. *Science*, *1*(7), 181–182. https://doi.org/10.1126/science.1.7.181

Abid, M. A., Kang, I.-S., Almazroui, M., & Kucharski, F. (2015). Contribution of Synoptic Transients to the Potential Predictability of PNA Circulation Anomalies: El Niño versus La Niña. *Journal of Climate*, *28*(21), 8347–8362. https://doi.org/10.1175/JCLI-D-14-00497.1

Adames, Á. F., & Wallace, J. M. (2014). Three-dimensional structure and evolution of the MJO and its relation to the mean flow. *Journal of the Atmospheric Sciences*, *71*(6), 2007–2026.

Alessandrini, S, Delle Monache, L., Sperati, S., & Cervone, G. (2015). An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, *157*, 95–110.

Alessandrini, Stefano, Sperati, S., & Monache, L. D. (2019). Improving the Analog Ensemble Wind Speed Forecasts for Rare Events. *Monthly Weather Review*, *147*(7), 2677–2692. https://doi.org/10.1175/MWR-D-19-0006.1

Amaya, D. J., & Foltz, G. R. (2014). Impacts of canonical and Modoki El Niño on tropical Atlantic SST. *Journal of Geophysical Research: Oceans*, *119*(2), 777–789.

Archambault, H. M., Bosart, L. F., Keyser, D., & Aiyyer, A. R. (2008). Influence of large-scale flow regimes on cool-season precipitation in the northeastern United States. *Monthly Weather Review*, *136*(8), 2945–2963.

Athanasiadis, P. J., Wallace, J. M., & Wettstein, J. J. (2010). Patterns of Wintertime Jet Stream Variability and Their Relation to the Storm Tracks. *Journal of the Atmospheric Sciences*, *67*(5), 1361–1381. https://doi.org/10.1175/2009JAS3270.1

Ayarzagüena, B., Ineson, S., Dunstone, N. J., Baldwin, M. P., & Scaife, A. A. (2018). Intraseasonal Effects of El Niño–Southern Oscillation on North Atlantic Climate. *Journal of Climate*, *31*(21), 8861–8873. https://doi.org/10.1175/JCLI-D-18-0097.1

Baran, S., & Baran, Á. (2021). Calibration of wind speed ensemble forecasts for power generation. *ArXiv Preprint ArXiv:2104.14910*.

Baran, S., & Lerch, S. (2015). Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, *141*(691), 2289–2299.

Barnston, A. G., & Livezey, R. E. (1987). Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns. *Monthly Weather Review*, *115*(6), 1083–1126. https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2

Barsugli, J. J., & Sardeshmukh, P. D. (2002). Global Atmospheric Sensitivity to Tropical SST Anomalies throughout the Indo-Pacific Basin. *Journal of Climate*, *15*(23), 3427–3442. https://doi.org/10.1175/1520-0442(2002)015<3427:GASTTS>2.0.CO;2

Bellier, J., Zin, I., & Bontron, G. (2017). Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Monthly Weather Review*, *145*(9), 3529–3544.

Bengio, Y. (2009). Learning Deep Architectures for AI. In *Foundations and Trends® in Machine Learning* (Vol. 2, Issue 1). https://doi.org/10.1561/2200000006

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, *126*(9), 98302.

Bjerknes, J. (1969). ATMOSPHERIC TELECONNECTIONS FROM THE EQUATORIAL PACIFIC. *Monthly Weather Review*, *97*(3), 163–172. https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2

Bjerknes, V. (1900). *Das dynamische Princip der Cirkulationsbewegungen in der Atmosphäre*.

Bjerknes, V. (1904). Das Problem der Wettervorhers-age, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteor. Z.*, *21*, 1–7.

Bjerknes, V. (1910). *Dynamic Meteorology and Hydrography: Part [1]-2,[and atlas of plates]* (Issue 88). Carnegie Institution of Washington.

Bjerknes, V., Rubenson, R., & Lindstedt, A. (1898). *Ueber einen Hydrodynamischen Fundamentalsatz und seine Anwendung: besonders auf die Mechanik der Atmosphäre und des Weltmeeres*. Kungl. Boktryckeriet. PA Norstedt \& Söner.

Blackmon, M. L., Lee, Y. H., Wallace, J. M., & Hsu, H.-H. (1984). Time variation of 500 mb height fluctuations with long, intermediate and short time scales as deduced from lag-correlation statistics. *Journal of Atmospheric Sciences*, *41*(6), 981–991.

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). *Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car*. 1–8. http://arxiv.org/abs/1704.07911

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200086.

Branstator, G. (1985). Analysis of General Circulation Model Sea-Surface Temperature Anomaly Simulations Using a Linear Model. Part II: Eigenanalysis. *Journal of the Atmospheric Sciences*, *42*(21), 2242–2254. https://doi.org/10.1175/1520-0469(1985)042<2242:AOGCMS>2.0.CO;2

Branstator, G. (1989). Low-Frequency Patterns Induced by Stationary Waves. *Journal of the Atmospheric Sciences*, *47*(5), 629–649. https://doi.org/10.1175/1520-0469(1990)047<0629:LFPIBS>2.0.CO;2

Branstator, G., & Teng, H. (2017). Tropospheric Waveguide Teleconnections and Their Seasonality. *Journal of the Atmospheric Sciences*, *74*(5), 1513–1532. https://doi.org/10.1175/JAS-D-16-0305.1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bremnes, J. B. (2020). Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Monthly Weather Review*, *148*(1), 403–414.

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, *22*(3), 651–661.

Burgman, R. J., & Jang, Y. (2015). Simulated U.S. Drought Response to Interannual and Decadal Pacific SST Variability. *Journal of Climate*, *28*(12), 4688–4705. https://doi.org/10.1175/JCLI-D-14-00247.1

Camporeale, E., & Carè, A. (2021). ACCRUE: Accurate and Reliable Uncertainty Estimate in Deterministic models. *International Journal for Uncertainty Quantification*, *11*(4).

Carter, G. M., Dallavalle, J. P., & Glahn, H. R. (1989). Statistical Forecasts Based on the National Meteorological Center's Numerical Weather Prediction System. *Weather and Forecasting*, *4*(3), 401–412. https://doi.org/10.1175/1520-0434(1989)004<0401:SFBOTN>2.0.CO;2

Chang, E. K. M., & Yu, D. B. (1999). Characteristics of wave packets in the upper troposphere. Part I: Northern Hemisphere winter. *Journal of the Atmospheric Sciences*, *56*(11), 1708–1728.

Chang, P., Ji, L., & Li, H. (1997). A decadal climate variation in the tropical Atlantic Ocean from thermodynamic air-sea interactions. *Nature*, *385*(6616), 516–518.

Chapman, W.E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving Atmospheric River Forecasts With Machine Learning. *Geophysical Research Letters*, *46*(17–18). https://doi.org/10.1029/2019GL083662

Chapman, William E., Subramanian, A. C., Xie, S.-P., Sierks, M. D., Ralph, F. M., & Kamae, Y. (2021). Monthly Modulations of ENSO Teleconnections: Implications for Potential Predictability in North America. *Journal of Climate*, *34*(14), 5899--5921. https://doi.org/10.1175/JCLI-D-20-0391.1

Chapman, William Eric, Subramanian, A., Delle Monache, L., Xie, S.-P., & Ralph, F. M. (2019). *Improving Atmospheric River Forecasts With Machine Learning Geophysical Research Letters*. 627–635. https://doi.org/10.1029/2019GL083662

Chen, L.-C., den Dool, H., Becker, E., & Zhang, Q. (2017). ENSO precipitation and temperature forecasts in the North American Multimodel Ensemble: Composite analysis and validation. *Journal of Climate*, *30*(3), 1103–1125.

Chen, M., & Kumar, A. (2015). Influence of ENSO SSTs on the spread of the probability density function for precipitation and land surface temperature. *Climate Dynamics*, *45*(3), 965–974. https://doi.org/10.1007/s00382-014-2336-9

Chen, M., Xie, P., Janowiak, J. E., & Arkin, P. A. (2002). Global Land Precipitation: A 50-yr Monthly Analysis Based on Gauge Observations. *Journal of Hydrometeorology*, *3*(3), 249–266. https://doi.org/10.1175/1525-7541(2002)003<0249:GLPAYM>2.0.CO;2

Chen, W. Y., & den Dool, H. (2003). Sensitivity of teleconnection patterns to the sign of their primary action center. *Monthly Weather Review*, *131*(11), 2885–2899.

Chen, W. Y., & Dool, H. M. Van Den. (1999). Significant change of extratropical natural variability and potential predictability associated with the El Niño/Southern Oscillation. *Tellus A: Dynamic Meteorology and Oceanography*, *51*(5), 790–802. https://doi.org/10.3402/tellusa.v51i5.14493

Chervin, R. M. (1986). Interannual Variability and Seasonal Climate Predictability. *Journal of the Atmospheric Sciences*, *43*(3), 233–251. https://doi.org/10.1175/1520-0469(1986)043<0233:IVASCP>2.0.CO;2

Chiang, J. C. H., Kushnir, Y., & Giannini, A. (2002). Deconstructing Atlantic Intertropical Convergence Zone variability: Influence of the local cross-equatorial sea surface temperature gradient and remote forcing from the eastern equatorial Pacific. *Journal of Geophysical Research: Atmospheres*, *107*(D1), ACL--3.

Chiang, J. C. H., & Vimont, D. J. (2004). Analogous Pacific and Atlantic meridional modes of tropical atmosphere--ocean variability. *Journal of Climate*, *17*(21), 4143–4158.

Chollet, F. (2015). Keras. *GitHub Repository*. https://github.com/fchollet/keras

Chollet, F., & others. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*, ascl--1806.

Chung, C. T. Y., & Power, S. B. (2015). Modelled Rainfall Response to Strong El Niño Sea Surface Temperature Anomalies in the Tropical Pacific. *Journal of Climate*, *28*(8), 3133–3151. https://doi.org/10.1175/JCLI-D-14-00610.1

Cobb, A., Michaelis, A., Iacobellis, S., Ralph, F. M., & Delle Monache, L. (2021). Atmospheric river sectors: Definition and characteristics observed using dropsondes from 2014--20 CalWater and AR Recon. *Monthly Weather Review*, *149*(3), 623–644.

Cordeira, J. M., & Ralph, F. M. (2021). A Summary of GFS Ensemble Integrated Water Vapor Transport Forecasts and Skill Along the US West Coast during Water Years 2017--2020. *Weather and Forecasting*.

Cordeira, J. M., Ralph, F. M., Martin, A., Gaggini, N., Spackman, J. R., Neiman, P. J., Rutz, J. J., & Pierce, R. (2017). Forecasting atmospheric rivers during CalWater 2015. *Bulletin of the American Meteorological Society*, *98*(3), 449–459.

Corringham, T. W., Ralph, F. M., Gershunov, A., Cayan, D. R., & Talbot, C. A. (2019). *Atmospheric rivers drive flood damages in the western United States*. 1–8.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory 2nd edition (wiley series in telecommunications and signal processing)*.

Dacre, H. F., Clark, P. A., Martinez-Alvarado, O., Stringer, M. A., & Lavers, D. A. (2015). How do atmospheric rivers form? *Bulletin of the American Meteorological Society*, *96*(8), 1243–1255. https://doi.org/10.1175/BAMS-D-14-00031.1

Dai, A. (2013). The influence of the inter-decadal Pacific oscillation on US precipitation during 1923--2010. *Climate Dynamics*, *41*(3), 633–646. https://doi.org/10.1007/s00382-012-

1446-5

Dai, A., & Wigley, T. M. L. (2000). Global patterns of ENSO-induced precipitation. *Geophysical Research Letters*, *27*(9), 1283–1286. https://doi.org/10.1029/1999GL011140

Davini, P., Von Hardenberg, J., Corti, S., Christensen, H. M., Juricke, S., Subramanian, A., Watson, P. A. G., Weisheimer, A., & Palmer, T. N. (2017). Climate SPHINX: Evaluating the impact of resolution and stochastic physics parameterisations in climate simulations. *Geoscientific Model Development*, *10*.

Dawson, Andrew. (2016). Windspharm: A High-Level Library for Global Wind Field Computations Using Spherical Harmonics. *Journal of Open Research Software*, *4*(1). https://doi.org/10.5334/jors.129

Dawson, Andrew, Matthews, A., & Stevens, D. (2010). Rossby wave dynamics of the North Pacific extra-tropical response to El Niño: Importance of the basic state in coupled GCMs. *Climate Dynamics*, *37*, 391–405. https://doi.org/10.1007/s00382-010-0854-7

Dawson, Andrew, & Palmer, T. N. (2015). Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Climate Dynamics*, *44*(7–8), 2177–2193.

Deflorio, M. J., Waliser, D. E., Guan, B., Lavers, D. A., Martin Ralph, F., & Vitart, F. (2018). Global assessment of atmospheric river prediction skill. *Journal of Hydrometeorology*, *19*(2), 409–426. https://doi.org/10.1175/JHM-D-17-0135.1

Deflorio, M. J., Waliser, D. E., Ralph, F. M., & Guan, B. (2019). *Experimental Subseasonal - to - Seasonal (S2S) Forecasting of Atmospheric Rivers Over the Western United States*. 242–265. https://doi.org/10.1029/2019JD031200

DeHaan, L. L., Martin, A. C., Weihs, R. R., Delle Monache, L., & Ralph, F. M. (2021). Object-based Verification of Atmospheric River Predictions in the Northeast Pacific. *Weather and Forecasting*.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., & Stull, R. B. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, *111*(D24).

Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., & Stull, R. (2006). Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *Journal of Geophysical Research Atmospheres*, *111*(5), 1–15. https://doi.org/10.1029/2005JD006311

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., & Stull, R. (2011). Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions. *Monthly Weather Review*, *139*(11), 3554–3570. https://doi.org/10.1175/2011mwr3653.1

DelSole, T. (2004). Predictability and Information Theory. Part I: Measures of Predictability. *Journal of the Atmospheric Sciences*, *61*(20), 2425–2440. https://doi.org/10.1175/1520-0469(2004)061<2425:PAITPI>2.0.CO;2

den Dool, H. M. (1989). A new look at weather forecasting through analogues. *Monthly Weather Review*, *117*(10), 2230–2247.

den Dool, H. M. (1994). Searching for analogues, how long must we wait? *Tellus A*, *46*(3), 314–324.

den Dool, H. M., & Saha, S. (1990). Frequency dependence in forecast skill. *Monthly Weather Review*, *118*(1), 128–137.

Deser, C., Simpson, I. R., McKinnon, K. A., & Phillips, A. S. (2017). The Northern Hemisphere extratropical atmospheric circulation response to ENSO: How well do we know it and how do we evaluate models accordingly? *Journal of Climate*, *30*(13), 5059–5082.

Deser, C., Simpson, I. R., Phillips, A. S., & McKinnon, K. A. (2018). How Well Do We Know ENSO's Climate Impacts over North America, and How Do We Evaluate Models Accordingly? *Journal of Climate*, *31*(13), 4991–5014. https://doi.org/10.1175/JCLI-D-17-0783.1

Deser, C., & Timlin, M. S. (1997). Atmosphere--ocean interaction on weekly timescales in the North Atlantic and Pacific. *Journal of Climate*, *10*(3), 393–408.

Deser, C., & Wallace, J. M. (1987). El Niño events and their relation to the Southern Oscillation: 1925–1986. *Journal of Geophysical Research: Oceans*, *92*(C13), 14189–14196. https://doi.org/10.1029/JC092iC13p14189

Deser, C., & Wallace, J. M. (1990). Large-Scale Atmospheric Circulation Features of Warm and Cold Episodes in the Tropical Pacific. *Journal of Climate*, *3*(11), 1254–1281. https://doi.org/10.1175/1520-0442(1990)003<1254:LSACFO>2.0.CO;2

Dettinger, M. D., Cayan, D. R., Diaz, H. F., & Meko, D. M. (1998). North–South Precipitation Patterns in Western North America on Interannual-to-Decadal Timescales. *Journal of Climate*, *11*(12), 3095–3111. https://doi.org/10.1175/1520-0442(1998)011<3095:NSPPIW>2.0.CO;2

Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., & Cayan, D. R. (2011). Atmospheric rivers, floods and the water resources of California. *Water*, *3*(2), 445–478.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business \& Economic Statistics*, *20*(1), 134–144.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). *Learning a Deep Convolutional Network for Image Super-Resolution* (pp. 184–199). https://doi.org/10.1007/978-3-319-10593-2_13

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018

Egger, J., & Schilling, H.-D. (1983). On the theory of the long-term variability of the atmosphere. *Journal of Atmospheric Sciences*, *40*(5), 1073–1085.

Epstein, E. S. (1969). Stochastic dynamic prediction. *Tellus*, *21*(6), 739–759.

Feldstein, S. B. (2002). Fundamental mechanisms of the growth and decay of the PNA teleconnection pattern. *Quarterly Journal of the Royal Meteorological Society*, *128*(581), 775–796. https://doi.org/10.1256/0035900021643683

Feng, J., Chen, W., & Li, Y. (2017). Asymmetry of the winter extra-tropical teleconnections in the Northern Hemisphere associated with two types of ENSO. *Climate Dynamics*, *48*(7), 2135–2151. https://doi.org/10.1007/s00382-016-3196-2

Fish, M. A., Wilson, A. M., & Ralph, F. M. (2019). Atmospheric river families: Definition and associated synoptic conditions. *Journal of Hydrometeorology*, *20*(10), 2091–2108.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., & Rummukainen, M. (2013). Evaluation of climate models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741–882). Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.020

Frauen, C., Dommenget, D., Tyrrell, N., Rezny, M., & Wales, S. (2014). Analysis of the Nonlinearity of El Niño–Southern Oscillation Teleconnections*. *Journal of Climate*, *27*(16), 6225–6244. https://doi.org/10.1175/JCLI-D-13-00757.1

Gadgil, S., Joseph, P. V, & Joshi, N. V. (1984). Ocean--atmosphere coupling over monsoon regions. *Nature*, *312*(5990), 141–143. https://doi.org/10.1038/312141a0

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, *32*(5), 1819–1840.

https://doi.org/10.1175/waf-d-17-0010.1

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., & others. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., … Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gershunov, A., Shulgina, T., Ralph, F. M., Lavers, D. A., & Rutz, J. J. (2017). Assessing the climate-scale variability of atmospheric rivers affecting western North America. *Geophysical Research Letters*, *44*(15), 7900–7908. https://doi.org/10.1002/2017GL074175

Ghazvinian, M., Zhang, Y., Seo, D.-J., He, M., & Fernando, N. (2021). A novel hybrid artificial neural network-Parametric scheme for postprocessing medium-range precipitation forecasts. *Advances in Water Resources*, *151*, 103907.

Gibson, P. B., Chapman, W. E., Altinok, A., Delle Monache, L., DeFlorio, M. J., & Waliser, D. E. (2021). Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*, *2*(1), 159. https://doi.org/10.1038/s43247-021-00225-4

Gibson, P. B., Waliser, D. E., Guan, B., DeFlorio, M. J., Ralph, F. M., & Swain, D. L. (2020). Ridging associated with drought across the western and southwestern United States: characteristics, trends, and predictability sources. *Journal of Climate*, *33*(7), 2485–2508.

Gill, A. E. (1980). Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society*, *106*(449), 447–462. https://doi.org/10.1002/qj.49710644905

Glahn, H. R., & Lowry, D. A. (1972). The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, *11*(8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118.

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business \& Economic Statistics*, *29*(3), 411–422.

Gowan, T. M., Steenburgh, W. J., & Schwartz, C. S. (2018). Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Weather and Forecasting*, *33*(3), 739–765.

Graham, N. E., & Barnett, T. P. (1987). Sea Surface Temperature, Surface Wind Divergence, and Convection over Tropical Oceans. *Science*, *238*(4827), 657–659. https://doi.org/10.1126/science.238.4827.657

Guan, B., & Waliser, D. E. (2015a). Detection of atmospheric rivers: Evaluation and applicationof an algorithm for gl obal studies. *Journal of Geophysical Research Atmospheres*, *120*. https://doi.org/10.1002/2015JD024257

Guan, B., & Waliser, D. E. (2015b). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *Journal of Geophysical Research: Atmospheres*, *120*(24), 12514–12535. https://doi.org/10.1002/2015JD024257

Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *ArXiv Preprint ArXiv:1604.06737*.

Guo, Y., Wen, Z., Chen, R., Li, X., & Yang, X.--Q. (2019). Effect of boreal spring precipitation anomaly pattern change in the late 1990s over tropical Pacific on the atmospheric teleconnection. *Climate Dynamics*, *52*(1), 401–416. https://doi.org/10.1007/s00382-018-4149-8

Gutzler, D. S., Rosen, R. D., Salstein, D. A., & Peixoto, J. (1988). Patterns of interannual variability in the Northern Hemisphere wintertime 850 mb temperature field. *Journal of Climate*, *1*(10), 949–964.

Hacker, Josh P, Ha, S.-Y., Snyder, C., Berner, J., Eckel, F. A., Kuchera, E., Pocernich, M., Rugg, S., Schramm, J., & Wang, X. (2011). The US Air ForceWeather Agency's mesoscale ensemble: scientific description and performance results. *Tellus A: Dynamic Meteorology and Oceanography*, *63*(3), 625–641.

Hacker, Joshua P., & Rife, D. L. (2008). A Practical Approach to Sequential Estimation of Systematic Error on Near-Surface Mesoscale Grids. *Weather and Forecasting*, *22*(6), 1257–1273. https://doi.org/10.1175/2007waf2006102.1

Hakim, G. J. (2003). Developing Wave Packets in the North Pacific Storm Track. *Monthly Weather Review*, *131*(11), 2824–2837. https://doi.org/10.1175/1520-0493(2003)131<2824:DWPITN>2.0.CO;2

Ham, Y.-G., Kim, J.-H., Kim, E.-S., & On, K.-W. (2021). Unified deep learning model for El Niño/Southern Oscillation forecasts by incorporating seasonality in climate data. *Science Bulletin*.

Ham, Y. G., Kim, J. H., & Luo, J. J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3), 550–560.

Hamill, T. M., & Colucci, S. J. (1997). Verification of Eta--RSM short-range ensemble forecasts. *Monthly Weather Review*, *125*(6), 1312–1327.

Hamill, T. M., & Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, *134*(11), 3209–3229.

Harnik, N., Seager, R., Naik, N., Cane, M., & Ting, M. (2010). The role of linear wave refraction in the transient eddy–mean flow response to tropical Pacific SST anomalies. *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 2132–2146. https://doi.org/10.1002/qj.688

Haupt, S. E., Chapman, W., Adams, S. V, Kirkwood, C., Hosking, J. S., Robinson, N. H., Lerch, S., & Subramanian, A. C. (2021). Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200091.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *In Proceedings of CVPR*, *19*, 770–778. https://doi.org/10.1016/0141-0229(95)00188-3

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hecht, C. W., & Cordeira, J. M. (2017). Characterizing the influence of atmospheric river orientation and intensity on precipitation distributions over North Coastal California. *Geophysical Research Letters*, *44*(17), 9048–9058. https://doi.org/https://doi.org/10.1002/2017GL074179

Held, I. M., Lyons, S. W., & Nigam, S. (1988). Transients and the Extratropical Response to El Niño. *Journal of the Atmospheric Sciences*, *46*(1), 163–174. https://doi.org/10.1175/1520-0469(1989)046<0163:TATERT>2.0.CO;2

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., & Haiden, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, *41*(24), 9197–9205.

Henderson, S. A., Maloney, E. D., & Son, S.-W. (2017). Madden--Julian oscillation Pacific teleconnections: The impact of the basic state and MJO representation in general circulation models. *Journal of Climate*, *30*(12), 4567–4587.

Henderson, S., Vimont, D. J., & Newman, M. (2020). The Critical Role of Non-Normality in Partitioning Tropical and Extratropical Contributions to PNA Growth. *Journal of Climate*, 6273–6295. https://doi.org/10.1175/JCLI-D-19-0555.1

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust Hyperparameter Optimization for Machine Learning. *SoftwareX*.

Hertwig, E., von Storch, J.-S., Handorf, D., Dethloff, K., Fast, I., & Krismer, T. (2015). Effect of horizontal resolution on ECHAM6-AMIP performance. *Climate Dynamics*, *45*(1), 185–211. https://doi.org/10.1007/s00382-014-2396-x

Hinton, G. E. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hirahara, S., Ishii, M., & Fukuda, Y. (2014). Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, *27*(1), 57–75. https://doi.org/10.1175/JCLI-D-12-00837.1

Hoerling, M. P., & Kumar, A. (2002). Atmospheric Response Patterns Associated with Tropical Forcing. *Journal of Climate*, *15*(16), 2184–2203. https://doi.org/10.1175/1520-0442(2002)015<2184:ARPAWT>2.0.CO;2

Hoerling, M. P., Kumar, A., & Xu, T. (2001). Robustness of the Nonlinear Climate Response to ENSO’s Extreme Phases. *Journal of Climate*, *14*(6), 1277–1293. https://doi.org/10.1175/1520-0442(2001)014<1277:ROTNCR>2.0.CO;2

Hoerling, M. P., Kumar, A., & Zhong, M. (1997). El Niño, La Niña, and the nonlinearity of their teleconnections. *Journal of Climate*, *10*(8), 1769–1786.

Homleid, M. (1995). Diurnal Corrections of Short-Term Surface Temperature Forecasts Using the Kalman Filter. *Weather and Forecasting*, *10*(4), 689–707. https://doi.org/10.1175/1520-0434(1995)010<0689:DCOSTS>2.0.CO;2

Honda, M., Kushnir, Y., Nakamura, H., Yamane, S., & Zebiak, S. E. (2005). Formation, Mechanisms, and Predictability of the Aleutian–Icelandic Low Seesaw in Ensemble

AGCM Simulations. *Journal of Climate*, *18*(9), 1423–1434. https://doi.org/10.1175/JCLI3353.1

Honda, M., & Nakamura, H. (2001). Interannual seesaw between the Aleutian and Icelandic lows. Part II: Its significance in the interannual variability over the wintertime Northern Hemisphere. *Journal of Climate*, *14*(24), 4512–4529.

Honda, M., Nakamura, H., Ukita, J., Kousaka, I., & Takeuchi, K. (2001). Interannual Seesaw between the Aleutian and Icelandic Lows. Part I: Seasonal Dependence and Life Cycle. *Journal of Climate*, *14*(6), 1029–1042. https://doi.org/10.1175/1520-0442(2001)014<1029:ISBTAA>2.0.CO;2

Hopson, T. M. (2014). Assessing the ensemble spread--error relationship. *Monthly Weather Review*, *142*(3), 1125–1142.

Horel, J. D., & Wallace, J. M. (1981). Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Monthly Weather Review*, *109*(4), 813–829.

Hoskins, B. J., & Karoly, D. J. (1981). The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of Atmospheric Sciences*, *38*(6), 1179–1196.

Hsu, H.-H. (1996). Global view of the intraseasonal oscillation during northern winter. *Journal of Climate*, *9*(10), 2386–2406.

Huang, J., Higuchi, K., & Shabbar, A. (1998). The relationship between the North Atlantic Oscillation and El Niño-Southern Oscillation. *Geophysical Research Letters*, *25*(14), 2707–2710. https://doi.org/10.1029/98GL01936

Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K., & Mackey, L. (2019). Improving Subseasonal Forecasting in the Western U . S . with Machine Learning. *25th KDD*, 7–12. https://dblp.org/db/conf/kdd/kdd2019.html#HwangOCPM19

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. http://arxiv.org/abs/1502.03167

Jenney, A. M., Nardi, K. M., Barnes, E. A., & Randall, D. A. (2019). The seasonality and regionality of MJO impacts on North American temperature. *Geophysical Research Letters*, *46*(15), 9193–9202.

Jiménez-Esteve, B, & Domeisen, D. I. V. (2019). Nonlinearity in the North Pacific Atmospheric Response to a Linear ENSO Forcing. *Geophysical Research Letters*, *46*(4), 2271–2281. https://doi.org/10.1029/2018GL081226

Jiménez-Esteve, Bernat, & Domeisen, D. I. V. (2018). The tropospheric pathway of the ENSO--North Atlantic teleconnection. *Journal of Climate*, *31*(11), 4563–4584.

Johnson, N. C. (2013). How Many ENSO Flavors Can We Distinguish?*. *Journal of Climate*, *26*(13), 4816–4827. https://doi.org/10.1175/JCLI-D-12-00649.1

Johnson, N. C., Collins, D. C., Feldstein, S. B., L'Heureux, M. L., & Riddle, E. E. (2014). Skillful Wintertime North American Temperature Forecasts out to 4 Weeks Based on the State of ENSO and the MJO*. *Weather and Forecasting*, *29*(1), 23–38. https://doi.org/10.1175/WAF-D-13-00102.1

Johnson, N. C., & Kosaka, Y. (2016a). The impact of eastern equatorial Pacific convection on the diversity of boreal winter El Niño teleconnection patterns. *Climate Dynamics*, *47*(12), 3737–3765.

Johnson, N. C., & Kosaka, Y. (2016b). The impact of eastern equatorial Pacific convection on the diversity of boreal winter El Niño teleconnection patterns. *Climate Dynamics*, *47*(12), 3737–3765. https://doi.org/10.1007/s00382-016-3039-1

Johnson, N. C., & Xie, S. (2010). Changes in the sea surface temperature threshold for tropical convection. *Nature Geoscience*, *3*(November), 3–6. https://doi.org/10.1038/ngeo1008

Jong, B.-T., Ting, M., & Seager, R. (2016). El Niño's impact on California precipitation: Seasonality, regionality, and El Niño intensity. *Environmental Research Letters*, *11*(5), 54021.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software, Articles*, *90*(12), 1–37. https://doi.org/10.18637/jss.v090.i12

Junk, C., Delle Monache, L., Alessandrini, S., Cervone, G., & Von Bremen, L. (2015). Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. *Meteor. Z*, *24*, 361–379.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., … Joseph, D. (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, *77*(3), 437–472. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Kamae, Y., Mei, W., Xie, S. P., Naoi, M., & Ueda, H. (2017). Atmospheric rivers over the Northwestern Pacific: Climatology and interannual variability. *Journal of Climate*, *30*(15), 5605–5619. https://doi.org/10.1175/JCLI-D-16-0875.1

Kang, I.-S., Kug, J.-S., Lim, M.-J., & Choi, D.-H. (2011). Impact of transient eddies on extratropical seasonal-mean predictability in DEMETER models. *Climate Dynamics*, *37*(3), 509–519. https://doi.org/10.1007/s00382-010-0873-4

Kang, I.-S., & Shukla, J. (2006). Dynamic seasonal prediction and predictability of the monsoon. In *The Asian Monsoon* (pp. 585–612). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-37722-0_15

Kharin, V. V, & Zwiers, F. W. (2003). On the ROC score of probability forecasts. *Journal of Climate*, *16*(24), 4145–4150.

King, M. P., Herceg-Bulić, I., Bladé, I., García-Serrano, J., Keenlyside, N., Kucharski, F., Li, C., & Sobolowski, S. (2018). Importance of Late Fall ENSO Teleconnection in the Euro-Atlantic Sector. *Bulletin of the American Meteorological Society*, *99*(7), 1337–1343. https://doi.org/10.1175/BAMS-D-17-0020.1

Kingma, P., Diederik, B., & Lei, J. (2014). Adam: A method for stochastic optimization. *ArXiv*.

Kirkwood, C., Economou, T., Odbert, H., & Pugeault, N. (2021). A framework for probabilistic weather forecast post-processing across models and lead times using machine learning. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200099.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., Dewitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., … Wood, E. F. (2014). The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Kleeman, R. (2002). Measuring Dynamical Prediction Utility Using Relative Entropy. *Journal of the Atmospheric Sciences*, *59*(13), 2057–2072. https://doi.org/10.1175/1520-0469(2002)059<2057:MDPUUR>2.0.CO;2

Kosaka, Y., & Nakamura, H. (2006). Structure and dynamics of the summertime Pacific-Japan teleconnection pattern. *Quarterly Journal of the Royal Meteorological Society*, *132*(619), 2009–2030. https://doi.org/10.1256/qj.05.204

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 950–957.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kumar, A., Barnston, A. G., Peng, P., Hoerling, M. P., & Goddard, L. (2000). Changes in the Spread of the Variability of the Seasonal Mean Atmospheric States Associated with ENSO. *Journal of Climate*, *13*(17), 3139–3151. https://doi.org/10.1175/1520-0442(2000)013<3139:CITSOT>2.0.CO;2

Kumar, A., & Hoerling, M. P. (1995). Prospects and Limitations of Seasonal Atmospheric GCM Predictions. *Bulletin of the American Meteorological Society*, *76*(3), 335–345. https://doi.org/10.1175/1520-0477(1995)076<0335:PALOSA>2.0.CO;2

Kumar, A., & Hoerling, M. P. (1998). Annual cycle of Pacific-North American seasonal predictability associated with different phases of ENSO. *Journal of Climate*, *11*(12), 3295–3308. https://doi.org/10.1175/1520-0442(1998)011<3295:ACOPNA>2.0.CO;2

Kumar, A., Zhang, Q., Peng, P., & Jha, B. (2005). SST-Forced Atmospheric Variability in an Atmospheric General Circulation Model. *Journal of Climate*, *18*(19), 3953–3967. https://doi.org/10.1175/JCLI3483.1

Kuo, C. C. J., Zhang, M., Li, S., Duan, J., & Chen, Y. (2019). Interpretable convolutional neural networks via feedforward design. *Journal of Visual Communication and Image Representation*, *60*, 346–359. https://doi.org/10.1016/j.jvcir.2019.03.010

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A., Matheson, M., Deslippe, J., Fatica, M., Prabhat, & Houston, M. (2018). *Exascale Deep Learning for Climate Analytics*. http://arxiv.org/abs/1810.01993

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L., Hersbach, H., & others. (2018). CERA-20C: A coupled reanalysis of the twentieth century. *Journal of Advances in Modeling Earth Systems*, *10*(5), 1172–1195.

Lamjiri, M. A., Dettinger, M. D., Ralph, F. M., & Guan, B. (2017). Hourly storm characteristics along the U.S. West Coast: Role of atmospheric rivers in extreme precipitation. *Geophysical Research Letters*, *44*(13), 7020–7028. https://doi.org/10.1002/2017GL074193

Lau, N.-C., & Nath, M. J. (1996). The Role of the "Atmospheric Bridge" in Linking Tropical Pacific ENSO Events to Extratropical SST Anomalies. *Journal of Climate*, *9*(9), 2036–2057. https://doi.org/10.1175/1520-0442(1996)009<2036:TROTBI>2.0.CO;2

Lau, W. K.-M., & Waliser, D. E. (2011). *Intraseasonal variability in the atmosphere-ocean climate system*. Springer Science \& Business Media.

Lavers, D. A., Waliser, D. E., Ralph, F. M., & Dettinger, M. D. (2016). Predictability of horizontal water vapor transport relative to precipitation: Enhancing situational awareness for forecasting western U.S. extreme precipitation and flooding. *Geophysical Research Letters*, *43*(5), 2275–2282. https://doi.org/10.1002/2016GL067765

Leathers, D. J., Yarnal, B., & Palecki, M. A. (1991). The Pacific/North American teleconnection pattern and United States climate. Part I: Regional temperature and precipitation associations. *Journal of Climate*, *4*(5), 517–528.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lerch, S., & Thorarinsdottir, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, *65*(1), 21206.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., & Gneiting, T. (2017). *Forecaster ' s Dilemma : Extreme Events and Forecast Evaluation*. *32*(1), 106–127. https://doi.org/10.1214/16-STS588

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539.

Li, Ying, & Lau, N.-C. (2012). Impact of ENSO on the Atmospheric Variability over the North Atlantic in Late Winter—Role of Transient Eddies. *Journal of Climate*, *25*(1), 320–342. https://doi.org/10.1175/JCLI-D-11-00037.1

Li, Yuanpeng, Lang, J., Ji, L., Zhong, J., Wang, Z., Guo, Y., & He, S. (2020). Weather Forecasting Using Ensemble of Spatial-Temporal Attention Network and Multi-Layer Perceptron. *Asia-Pacific Journal of Atmospheric Sciences*, 1–14.

Lin, H., Brunet, G., & Derome, J. (2008). Forecast skill of the Madden--Julian oscillation in two Canadian atmospheric models. *Monthly Weather Review*, *136*(11), 4130–4149.

Lin, H., & Derome, J. (1996). Changes in predictability associated with the PNA pattern. *Tellus A*, *48*(4), 553–571.

Long, J., Shelhamer, E., & Darrell, T. (2015a). Fully Convolutional Networks for Semantic Segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7298965

Long, J., Shelhamer, E., & Darrell, T. (2015b). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Lorenz, E. N. (1969a). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences*, *26*(4), 636–646.

Lorenz, E. N. (1969b). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289–307.

Ma, J., Xie, S.-P., & Xu, H. (2017). Contributions of the North Pacific meridional mode to ensemble spread of ENSO prediction. *Journal of Climate*, *30*(22), 9167–9181.

Ma, J., Xie, S.-P., Xu, H., Zhao, J., & Zhang, L. (2021). Cross-basin Interactions between the Tropical Atlantic and Pacific in the ECMWF Hindcasts. *Journal of Climate*, *34*(7), 2459–2472.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Madden, R. A., & Julian, P. R. (1971). Detection of a 40--50 day oscillation in the zonal wind in the tropical Pacific. *Journal of Atmospheric Sciences*, *28*(5), 702–708.

Magnusson, L., & Källén, E. (2013). Factors Influencing Skill Improvements in the ECMWF Forecasting System. *Monthly Weather Review*, *141*(9), 3142–3153. https://doi.org/10.1175/mwr-d-12-00318.1

Martin, A., Ralph, F. M., Demirdjian, R., DeHaan, L., Weihs, R., Helly, J., Reynolds, D., & Iacobellis, S. (2018). Evaluation of atmospheric river predictions by the WRF Model using aircraft and regional mesonet observations of orographic precipitation and its forcing. *Journal of Hydrometeorology*, *19*(7), 1097–1113.

Mason, S. J., & Graham, N. E. (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, *14*(5), 713–725.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096.

Matsumura, S., Huang, G., Xie, S.-P., & Yamazaki, K. (2010). SST-Forced and Internal Variability of the Atmosphere in an Ensemble GCM Simulation. *Journal of the Meteorological Society of Japan*, *88*(1), 43–62. https://doi.org/10.2151/jmsj.2010-104

McCarty, W, Coy, L., Gelaro, R., Huang, A., Merkova, D., Smith, E. B., Sienkiewicz, M., & Wargan, K. (2016). MERRA-2 input observations: Summary and assessment. *NASA Tech. Rep. NASA/TM-2016-104606*, *46*, 51.

McCarty, Will, Coy, L., Gelaro, R., Huang, A., Merkova, D., Smith, E. B., Sienkiewicz, M., & Wargan, K. (2016). MERRA-2 Input Observations: Summary and Assessment. *Technical Report Series on Global Modeling and Data Assimilation*, *46*(October). https://gmao.gsfc.nasa.gov/pubs/docs/McCarty885.pdf

McCollor, D., & Stull, R. (2008). Hydrometeorological Accuracy Enhancement via Postprocessing of Numerical Weather Forecasts in Complex Terrain. *Weather and Forecasting*, *23*(1), 131–144. https://doi.org/10.1175/2007WAF2006107.1

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., & Williams, J. K. (2017). Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/bams-d-16-0123.1

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Meech, S., Alessandrini, S., Chapman, W., & Delle Monache, L. (2021). Post-processing rainfall in a high-resolution simulation of the 1994 Piedmont flood. *Bulletin of Atmospheric Science and Technology*. https://doi.org/10.1007/s42865-020-00028-z

Meehl, G. A., & Hu, A. (2006). Megadroughts in the Indian Monsoon Region and Southwest North America and a Mechanism for Associated Multidecadal Pacific Sea Surface Temperature Anomalies. *Journal of Climate*, *19*(9), 1605–1623. https://doi.org/10.1175/JCLI3675.1

Mei, W., Kamae, Y., Xie, S.-P., & Yoshida, K. (2019). Variability and Predictability of North Atlantic Hurricane Frequency in a Large Ensemble of High-Resolution Atmospheric Simulations. *Journal of Climate*, *32*(11), 3153–3167. https://doi.org/10.1175/JCLI-D-18-0554.1

Mizuta, R., Murata, A., Ishii, M., Shiogama, H., Hibino, K., Mori, N., Arakawa, O., Imada, Y., Yoshida, K., Aoyagi, T., Kawase, H., Mori, M., Okada, Y., Shimura, T., Nagatomo, T., Ikeda, M., Endo, H., Nosaka, M., Arai, M., … Kimoto, M. (2017). Over 5,000 years of ensemble future climate simulations by 60-km global and 20-km regional atmospheric models. *Bulletin of the American Meteorological Society*, *98*(7), 1383–1398. https://doi.org/10.1175/BAMS-D-16-0099.1

Mo, K. C., & Livezey, R. E. (1986). Tropical-Extratropical Geopotential Height Teleconnections during the Northern Hemisphere Winter. *Monthly Weather Review*, *114*(12), 2488–2515. https://doi.org/10.1175/1520-0493(1986)114<2488:TEGHTD>2.0.CO;2

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, *122*(529), 73–119.

Moorthi, Ss., Pan, H.-L., & Caplan, P. (2001). Changes to the 2001 NCEP Operational MRF/AVN Global Analysis/ Forecast System. *NWS Technical Procedures Bulletin*, *484*, 1–14.

Mori, M., & Watanabe, M. (2008). The growth and triggering mechanisms of the PNA: A MJO-PNA coherence. *Journal of the Meteorological Society of Japan. Ser. II*, *86*(1), 213–236.

Murakami, H., Mizuta, R., & Shindo, E. (2012). Future changes in tropical cyclone activity projected by multi-physics and multi-SST ensemble experiments using the 60-km-mesh MRI-AGCM. *Climate Dynamics*, *39*(9–10), 2569–2584. https://doi.org/10.1007/s00382-011-1223-x

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, *12*(4), 595–600.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Icml*.

Nakamura, H., Tanaka, M., & Wallace, J. M. (1987). Horizontal structure and energetics of Northern Hemisphere wintertime teleconnection patterns. *Journal of Atmospheric Sciences*, *44*(22), 3377–3391.

Naoi, M., Kamae, Y., Ueda, H., & Mei, W. (2020). Impacts of Seasonal Transitions of ENSO on Atmospheric River Activity over East Asia. *Journal of the Meteorological Society of Japan. Ser. II*, *advpub*. https://doi.org/10.2151/jmsj.2020-027

Nardi, K. M., Barnes, E. A., & Ralph, F. M. (2018). Assessment of numerical weather prediction model reforecasts of the occurrence, intensity, and location of atmospheric rivers along the West Coast of North America. *Monthly Weather Review*, *146*(10), 3343–3362.

Nayak, M. A., Villarini, G., & Lavers, D. A. (2014). On the skill of numerical weather prediction models to forecast atmospheric rivers over the central United States. *Geophysical Research Letters*, *41*(12), 4354–4362.

Neelin, J. D., Jin, F.-F., & Syu, H.-H. (2000). Variations in ENSO Phase Locking. *Journal of Climate*, *13*(14), 2570–2590. https://doi.org/10.1175/1520-0442(2000)013<2570:VIEPL>2.0.CO;2

Newman, M., & Sardeshmukh, P. D. (1998). The impact of the annual cycle on the North Pacific/North American response to remote low-frequency forcing. *Journal of the Atmospheric Sciences*, *55*(8), 1336–1353.

Newman, M., Shin, S.-I., & Alexander, M. A. (2011). Natural variation in ENSO flavors. *Geophysical Research Letters*, *38*(14). https://doi.org/10.1029/2011GL047658

Nie, Y., Zhang, Y., Yang, X.-Q., & Ren, H.-L. (2019). Winter and Summer Rossby Wave Sources in the CMIP5 Models. *Earth and Space Science*, *6*(10), 1831–1846. https://doi.org/10.1029/2019EA000674

Nielsen, M. (2015). Neural networks and deep learning. *San Francisco, CA: Determination Press*, *2018*. http://neuralnetworksanddeeplearning.com/

Norris, J. R. (2000). Interannual and Interdecadal Variability in the Storm Track, Cloudiness, and Sea Surface Temperature over the Summertime North Pacific. *Journal of Climate*, *13*(2), 422–430. https://doi.org/10.1175/1520-0442(2000)013<0422:IAIVIT>2.0.CO;2

O'Reilly, C. H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T. N., Schaller, N., & Woollings, T. (2017a). Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophysical Research Letters*, *44*(11), 5729–5738.

O'Reilly, C. H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T. N., Schaller, N., & Woollings, T. (2017b). Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophysical Research Letters*, *44*(11), 5729–5738. https://doi.org/10.1002/2017GL073736

O'Reilly, C. H., Weisheimer, A., MacLeod, D., Befort, D. J., & Palmer, T. (2020). Assessing the robustness of multidecadal variability in Northern Hemisphere wintertime seasonal forecast skill. *Quarterly Journal of the Royal Meteorological Society*, *146*(733), 4055–4066.

Orsolini, Y. J., Kvamstø;, N. G., Kindem, I. T., Honda, M., & Nakamura, H. (2008). Influence of the Aleutian-Icelandic Low Seesaw and ENSO onto the Stratosphere in Ensemble Winter Hindcasts. *Journal of the Meteorological Society of Japan. Ser. II*, *86*(5), 817–825. https://doi.org/10.2151/jmsj.86.817

Palmer, T. N. (1988). Medium and extended range predictability and stability of the Pacific/North American mode. *Quarterly Journal of the Royal Meteorological Society*, *114*(481), 691–713.

Peng, P., & Kumar, A. (2005). A Large Ensemble Analysis of the Influence of Tropical SSTs on Seasonal Atmospheric Variability. *Journal of Climate*, *18*(7), 1068–1085. https://doi.org/10.1175/JCLI-3314.1

Philander, S. G. (1989). El Niño, La Niña, and the southern oscillation. *International Geophysics Series*, *46*, X--289.

Pinto, J. G., Reyers, M., & Ulbrich, U. (2011). The variable link between PNA and NAO in observations and in multi-century CGCM simulations. *Climate Dynamics*, *36*(1), 337–354. https://doi.org/10.1007/s00382-010-0770-x

Prohaska, J. T. (1976). A technique for analyzing the linear relationships between two meteorological fields. *Monthly Weather Review*, *104*(11), 1345–1353.

Qin, J., & Robinson, W. A. (1993). On the Rossby wave source and the steady linear response to tropical forcing. *Journal of Atmospheric Sciences*, *50*(12), 1819–1823.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174.

Ralph, F. M., Cannon, F., Tallapragada, V., Davis, C. A., Doyle, J. D., Pappenberger, F., Subramanian, A., Wilson, A. M., Lavers, D. A., Reynolds, C. A., & others. (2020). West Coast forecast challenges and development of atmospheric river reconnaissance. *Bulletin of the American Meteorological Society*, *101*(8), E1357--E1377.

Ralph, F. M., Dettinger, M. D., Schick, L. J., & Anderson, M. L. (2020). Introduction to atmospheric rivers. In *Atmospheric rivers* (pp. 1–13). Springer.

Ralph, F. M., Neiman, P. J., & Wick, G. A. (2004). Satellite and CALJET Aircraft Observations of Atmospheric Rivers over the Eastern North Pacific Ocean during the Winter of 1997/98. *Monthly Weather Review*, *132*(7), 1721–1745. https://doi.org/10.1175/1520-0493(2004)132<1721:SACAOO>2.0.CO;2

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., Schick, L. J., & Smallcomb, C. (2018). A Scale to Characterize the Strength and Impacts of Atmospheric Rivers. *Bulletin of the American Meteorological Society*, *100*(2), 269–289. https://doi.org/10.1175/bams-d-18-0023.1

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., Schick, L. J., & Smallcomb, C. (2019). A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, *100*(2), 269–289.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020MS002203

Rasp, S., & Lerch, S. (2018). Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649–667.

Richardson, L. F. (1922). *Weather Prediction by Numerical Process*. Cambrige Univ. Press.

Riddle, E. E., Stoner, M. B., Johnson, N. C., L'Heureux, M. L., Collins, D. C., & Feldstein, S. B. (2013). The impact of the MJO on clusters of wintertime circulation anomalies over the North American region. *Climate Dynamics*, *40*(7–8), 1749–1766

Robertson, D. E., Shrestha, D. L., & Wang, Q. J. (2013). Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrology and Earth System Sciences*, *17*(9), 3587–3603.

Roeger, C., Stull, R., McClung, D., Hacker, J., Deng, X., & Modzelewski, H. (2003). Verification of Mesoscale Numerical Weather Forecasts in Mountainous Terrain for Application to Avalanche Prediction. *Weather and Forecasting*, *18*(6), 1140–1160. https://doi.org/10.1175/1520-0434(2003)018<1140:vomnwf>2.0.co;2

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Ropelewski, C. F., & Halpert, M. S. (1986). North American Precipitation and Temperature Patterns Associated with the El Niño/Southern Oscillation (ENSO). *Monthly Weather Review*, *114*(12), 2352–2362. https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2

Roulston, M. S., & Smith, L. A. (2002). Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, *130*(6), 1653–1660. https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2

Rowell, D. P. (1998). Assessing Potential Seasonal Predictability with an Ensemble of Multidecadal GCM Simulations. *Journal of Climate*, *11*(2), 109–120. https://doi.org/10.1175/1520-0442(1998)011<0109:APSPWA>2.0.CO;2

Sardeshmukh, P. D., Compo, G. P., & Penland, C. (2000). Changes of probability associated with El Niño. *Journal of Climate*, *13*(24), 4268–4286. https://doi.org/10.1175/1520-0442(2000)013<4268:COPAWE>2.0.CO;2

Sardeshmukh, Prashant D, Compo, G. P., & Penland, C. (2000). Changes of probability associated with El Niño. *Journal of Climate*, *13*(24), 4268–4286.

Sardeshmukh, Prashant D, & Hoskins, B. J. (1988). The generation of global rotational flow by steady idealized tropical divergence. *Journal of the Atmospheric Sciences*, *45*(7), 1228–1251.

Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *Npj Climate and Atmospheric Science*, *1*(1), 28. https://doi.org/10.1038/s41612-018-0038-4

Schamberg, G., Chapman, W., Xie, S.-P., & Coleman, T. P. (2020). Direct and Indirect Effects—An Information Theoretic Perspective. *Entropy*, *22*(8), 854.

Scher, S. (2018). Toward Data-Driven Weather and Climate Forecasting: Approximating a

Simple General Circulation Model With Deep Learning. *Geophysical Research Letters*, *45*(22), 12,616-12,622. https://doi.org/10.1029/2018GL080704

Scher, Sebastian, & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830–2841. https://doi.org/10.1002/qj.3410

Scher, Sebastian, & Messori, G. (2019). Weather and climate forecasting with neural networks: using GCMs with different complexity as study-ground. *Geoscientific Model Development Discussions*, *March*, 1–15. https://doi.org/10.5194/gmd-2019-53

Scheuerer, M., & Hamill, T. M. (2015). Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored , Shifted Gamma Distributions *. *Monthly Weather Review*, 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1

Scheuerer, M., Switanek, M. B., Worsnop, R. P., & Hamill, T. M. (2020). Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. *Monthly Weather Review*, *148*(8), 3489–3506. https://doi.org/10.1175/MWR-D-20-0096.1

Schubert, S. D., Suarez, M. J., Chang, Y., & Branstator, G. (2001). The Impact of ENSO on Extratropical Low-Frequency Noise in Seasonal Forecasts. *Journal of Climate*, *14*(10), 2351–2365. https://doi.org/10.1175/1520-0442(2001)014<2351:TIOEOE>2.0.CO;2

Schulz, B., & Lerch, S. (2021). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *ArXiv Preprint ArXiv:2106.09512*.

Schulzweida, U. (2019). *CDO User Guide (Version 1.9.6)*. https://doi.org/10.5281/zenodo.2558193

Schulzweida, U., Kornblueh, L., & Quast, R. (2006). CDO user's guide. *Climate Data Operators, Version*, *1*(6).

Seager, R, Naik, N., Ting, M., Cane, M. A., Harnik, N., & Kushnir, Y. (2010). Adjustment of the atmospheric circulation to tropical Pacific SST anomalies: Variability of transient eddy propagation in the Pacific–North America sector. *Quarterly Journal of the Royal Meteorological Society*, *136*(647), 277–296. https://doi.org/10.1002/qj.588

Seager, Richard, Kushnir, Y., Herweijer, C., Naik, N., & Velez, J. (2005). Modeling of Tropical Forcing of Persistent Droughts and Pluvials over Western North America: 1856–2000. *Journal of Climate*, *18*(19), 4065–4088. https://doi.org/10.1175/JCLI3522.1

Seo, K.-H., & Lee, H.-J. (2017). Mechanisms for a PNA-like teleconnection pattern in response to the MJO. *Journal of the Atmospheric Sciences*, *74*(6), 1767–1781.

Shi, W., Schaller, N., Macleod, D., Palmer, T. N., & Weisheimer, A. (2015). Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research Letters*, *42*(5), 1554–1559. https://doi.org/10.1002/2014GL062829

Shukla, J., & Wallace, J. M. (1983). Numerical Simulation of the Atmospheric Response to Equatorial Pacific Sea Surface Temperature Anomalies. *Journal of the Atmospheric Sciences*, *40*(7), 1613–1630. https://doi.org/10.1175/1520-0469(1983)040<1613:NSOTAR>2.0.CO;2

Siegert, S., Bröcker, J., & Kantz, H. (2012). Rank histograms of stratified Monte Carlo ensembles. *Monthly Weather Review*, *140*(5), 1558–1571.

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R., & Arribas, A. (2016). A Bayesian Framework for Verification and Recalibration of Ensemble Forecasts: How Uncertain is NAO Predictability? *Journal of Climate*, *29*(3), 995–1012. https://doi.org/10.1175/JCLI-D-15-0196.1

Simmons, A. J., Wallace, Jm., & Branstator, G. W. (1983). Barotropic wave propagation and instability, and atmospheric teleconnection patterns. *Journal of Atmospheric Sciences*, *40*(6), 1363–1392.

Smith, D. M., Scaife, A. A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., Bilbao, R., Borchert, L. F., Caron, L.-P., Counillon, F., Danabasoglu, G., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Estella-Perez, V., Flavoni, S., Hermanson, L., Keenlyside, N., Kharin, V., … Zhang, L. (2020). North Atlantic climate far more predictable than models imply. *Nature*, *583*(7818), 796–800. https://doi.org/10.1038/s41586-020-2525-0

Sodemann, H., & Stohl, A. (2013). Moisture Origin and Meridional Transport in Atmospheric Rivers and Their Association with Multiple Cyclones*. *Monthly Weather Review*, *141*(8), 2850–2868. https://doi.org/10.1175/mwr-d-12-00256.1

Souders, M. B., Colle, B. A., & Chang, E. K. M. (2014). The Climatology and Characteristics of Rossby Wave Packets Using a Feature-Based Tracking Technique. *Monthly Weather Review*, *142*(10), 3528–3548. https://doi.org/10.1175/MWR-D-13-00371.1

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., & Rogers, E. (1999). Using ensembles for short-range forecasting. *Monthly Weather Review*, *127*(4), 433–446.

Stensrud, D. J., & Skindlov, J. A. (2002). Gridpoint Predictions of High Temperature from a Mesoscale Model. *Weather and Forecasting*, *11*(1), 103–110. https://doi.org/10.1175/1520-0434(1996)011<0103:gpohtf>2.0.co;2

Stensrud, D. J., & Yussouf, N. (2003). Short-Range Ensemble Predictions of 2-m Temperature and Dewpoint Temperature over New England. *Monthly Weather Review*, *131*(10), 2510–2524. https://doi.org/10.1175/1520-0493(2003)131<2510:sepomt>2.0.co;2

Stephenson, D. B., Pavan, V., & Bojariu, R. (2000). Is the North Atlantic Oscillation a random walk? *International Journal of Climatology*, *20*(1), 1–18. https://doi.org/10.1002/(SICI)1097-0088(200001)20:1<1::AID-JOC456>3.0.CO;2-P

Stone, R. E., Reynolds, C. A., Doyle, J. D., Langland, R. H., Baker, N. L., Lavers, D. A., & Ralph, F. M. (2020). Atmospheric river reconnaissance observation impact in the Navy global forecast system. *Monthly Weather Review*, *148*(2), 763–782.

Storch, H. Von, & Zwiers, F. W. (1999). Statistical Analysis in Climate Research. *Journal of the American Statistical Association*, *95*, 1375. https://doi.org/10.1017/CBO9780511612336

Straus, D. M., & Shukla, J. (1997). Variations of Midlatitude Transient Dynamics Associated with ENSO. *Journal of the Atmospheric Sciences*, *54*(7), 777–790. https://doi.org/10.1175/1520-0469(1997)054<0777:VOMTDA>2.0.CO;2

Straus, D. M., & Shukla, J. (2002). Does ENSO force the PNA? *Journal of Climate*, *15*(17), 2340–2358. https://doi.org/10.1175/1520-0442(2002)015<2340:DEFTP>2.0.CO;2

Swets, J. A. (1973). The relative operating characteristic in psychology: a technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, *182*(4116), 990–1000.

Takahashi, C., & Shirooka, R. (2014). Storm track activity over the North Pacific associated with the Madden-Julian Oscillation under ENSO conditions during boreal winter. *Journal of Geophysical Research: Atmospheres*, *119*(18), 10–663.

Takaya, K., & Nakamura, H. (2001). A Formulation of a Phase-Independent Wave-Activity Flux for Stationary and Migratory Quasigeostrophic Eddies on a Zonally Varying Basic Flow. *Journal of the Atmospheric Sciences*, *58*(6), 608–627. https://doi.org/10.1175/1520-0469(2001)058<0608:AFOAPI>2.0.CO;2

Tao, Y., Gao, X., Hsu, K., Sorooshian, S., & Ihler, A. (2016). A Deep Neural Network Modeling Framework to Reduce Bias in Satellite Precipitation Products. *Journal of Hydrometeorology*, *17*(3), 931–945. https://doi.org/10.1175/JHM-D-15-0075.1

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Theocharides, S., Makrides, G., Livera, A., Theristis, M., Kaimakis, P., & Georghiou, G. E. (2020). Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Applied Energy*, *268*, 115023.

Thorarinsdottir, T. L., & Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(2), 371–388.

Thorarinsdottir, T. L., & Johnson, M. S. (2012). Probabilistic wind gust forecasting using nonhomogeneous Gaussian regression. *Monthly Weather Review*, *140*(3), 889–897.

Titchner, H. A., & Rayner, N. A. (2014). The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, *119*(6), 2864–2889.

Toms, B. A., Kashinath, K., Prabhat, & Yang, D. (2019). *Deep Learning for Scientific Inference from Geophysical Data: The Madden-Julian Oscillation as a Test Case*. http://arxiv.org/abs/1902.04621

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242–264). IGI global.

Toth, Z., & Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, *74*(12), 2317–2330.

Trascasa-Castro, P., Maycock, A. C., Scott Yiu, Y. Y., & Fletcher, J. K. (2019). On the linearity of the stratospheric and Euro-Atlantic sector response to ENSO. *Journal of Climate*, *32*(19), 6607–6626.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., & Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *Journal of Geophysical Research: Oceans*, *103*(C7), 14291–14324. https://doi.org/10.1029/97JC01444

Tribbia, J. J., & Baumhefner, D. P. (1988). Estimates of the predictability of low-frequency variability with a spectral general circulation model. *Journal of Atmospheric Sciences*, *45*(16), 2306–2318.

Tseng, K-C, Barnes, E. A., & Maloney, E. D. (2018). Prediction of the midlatitude response to strong Madden-Julian Oscillation events on S2S time scales. *Geophysical Research Letters*, *45*(1), 463–470.

Tseng, Kai-Chih, Maloney, E., & Barnes, E. (2019). The consistency of MJO teleconnection patterns: An explanation using linear Rossby wave theory. *Journal of Climate*, *32*(2), 531–

548.

Tseng, Kai-Chih, Maloney, E., & Barnes, E. A. (2020). The consistency of MJO teleconnection patterns on interannual time scales. *Journal of Climate*, *33*(9), 3471–3486.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z. Ben, Bhend, J., Dabernig, M., Cruz, L. De, Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., … Ylhaisi, J. (2021). Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, *102*(3), E681–E699. https://doi.org/10.1175/BAMS-D-19-0308.1

Ventrice, M. J., Wheeler, M. C., Hendon, H. H., Schreck, C. J., Thorncroft, C. D., & Kiladis, G. N. (2013). A modified multivariate Madden--Julian oscillation index using velocity potential. *Monthly Weather Review*, *141*(12), 4197–4210.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 1096–1103. https://doi.org/10.1145/1390156.1390294

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H. S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., … Zhang, L. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, Frédéric. (2004). Monthly forecasting at ECMWF. *Monthly Weather Review*, *132*(12), 2761–2779.

Vitart, Frédéric, & Molteni, F. (2010). Simulation of the Madden--Julian oscillation and its teleconnections in the ECMWF forecast system. *Quarterly Journal of the Royal Meteorological Society*, *136*(649), 842–855.

Wallace, J. M., & Gutzler, D. S. (1981a). Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, *109*(4), 784–812.

Wallace, J. M., & Gutzler, D. S. (1981b). Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. In *Monthly Weather Review* (Vol. 109, Issue 4, pp. 784–812). https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2

Wallace, J. M., Smith, C., & Bretherton, C. S. (1992). Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *Journal of Climate*, *5*(6), 561–576.

Wang, J., Kim, H.-M., Chang, E. K. M., & Son, S.-W. (2018). Modulation of the MJO and North Pacific storm track relationship by the QBO. *Journal of Geophysical Research: Atmospheres*, *123*(8), 3976–3992.

Wang, J., Kim, H., Kim, D., Henderson, S. A., Stan, C., & Maloney, E. D. (2020a). MJO teleconnections over the PNA region in climate models. Part I: Performance-and process-based skill metrics. *Journal of Climate*, *33*(3), 1051–1067.

Wang, J., Kim, H., Kim, D., Henderson, S. A., Stan, C., & Maloney, E. D. (2020b). MJO teleconnections over the PNA region in climate models. Part II: Impacts of the MJO and basic state. *Journal of Climate*, *33*(12), 5081–5101.

Wang, L., & Robertson, A. W. (2019). Week 3--4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dynamics*, *52*(9), 5861–5875.

Wang, X., & Bishop, C. H. (2003). A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *Journal of Atmospheric Sciences*, *60*(9), 1140–1158.

Warner, M. D., Mass, C. F., & Salathé, E. P. (2012). Wintertime Extreme Precipitation Events along the Pacific Northwest Coast: Climatology and Synoptic Evolution. *Monthly Weather Review*, *140*(7), 2021–2043. https://doi.org/10.1175/mwr-d-11-00197.1

Weare, B. C., & Nasstrom, J. S. (1982). Examples of Extended Empirical Orthogonal Function Analyses. *Monthly Weather Review*, *110*(6), 481–485. https://doi.org/10.1175/1520-0493(1982)110<0481:EOEEOF>2.0.CO;2

Webb, R. S., Nowak, K., Cifelli, R., & Brekke, L. D. (2017). Sub-Seasonal Climate Forecast Rodeo. *AGU Fall Meeting Abstracts*, *2017*, PA14A--02.

Weigel, A. P., Baggenstos, D., Liniger, M. A., Vitart, F., & Appenzeller, C. (2008). Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review*, *136*(12), 5162–5182.

Weisheimer, A., Befort, D. J., MacLeod, D., Palmer, T., O?Reilly, C., & Str?mmen, K. (2020). Seasonal Forecasts of the Twentieth Century. *Bulletin of the American Meteorological Society*, *101*(8), E1413–E1426. https://doi.org/10.1175/BAMS-D-19-0019.1

Weisheimer, A., Decremer, D., MacLeod, D., O’Reilly, C., Stockdale, T. N., Johnson, S., & Palmer, T. N. (2019). How confident are predictability estimates of the winter North Atlantic Oscillation? *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 140–159. https://doi.org/10.1002/qj.3446

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *ArXiv Preprint ArXiv:2102.05107*.

Wheeler, M. C., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932.

Wick, G. A., Neiman, P. J., Ralph, F. M., & Hamill, T. M. (2013). Evaluation of Forecasts of the Water Vapor Signature of Atmospheric Rivers in Operational Numerical Weather Prediction Models. *Weather and Forecasting*, *28*(6), 1337–1352. https://doi.org/10.1175/WAF-D-13-00025.1

Wilks, D. (2016). "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2273.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications: A Journal of Forecasting, Practical Applications, Training Techniques and Modelling*, *16*(3), 361–368.

Wilks, D. S. (2010). Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 2109–2118.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Wilks, D. S., & Hamill, T. M. (2007). Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, *135*(6), 2379–2390. https://doi.org/10.1175/mwr3402.1

Wilson, A. M., Chapman, W., Payne, A., Ramos, A. M., Boehm, C., Campos, D., Cordeira, J., Garreaud, R., Gorodetskaya, I. V, Rutz, J. J., & others. (2020). Training the next generation of researchers in the science and application of atmospheric rivers. *Bulletin of the American Meteorological Society*, *101*(6), E738--E743.

Wirth, V., Riemer, M., Chang, E. K. M., & Martius, O. (2018). Rossby wave packets on the midlatitude waveguide—A review. *Monthly Weather Review*, *146*(7), 1965–2001.

Wolter, K., & Timlin, M. S. (2011). El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI. ext). *International Journal of Climatology*, *31*(7), 1074–1087.

Wu, L., Seo, D.-J., Demargne, J., Brown, J. D., Cong, S., & Schaake, J. (2011). Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *Journal of Hydrology*, *399*(3–4), 281–298.

Xie, S.-P., Peng, Q., Kamae, Y., Zheng, X.-T., Tokinaga, H., & Wang, D. (2018). Eastern Pacific ITCZ Dipole and ENSO Diversity. *Journal of Climate*, *31*(11), 4449–4462.

https://doi.org/10.1175/JCLI-D-17-0905.1

Xie, S.-P., & Philander, S. G. H. (1994). *A coupled ocean-atmosphere model of relevance to the ITCZ in the eastern Pacific*. Tellus A.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 802–810.

Yang, X.-Q., Anderson, J. L., & Stern, W. F. (1998). Reproducible Forced Modes in AGCM Ensemble Integrations and Potential Predictability of Atmospheric Seasonal Variations in the Extratropics. *Journal of Climate*, *11*(11), 2942–2959. https://doi.org/10.1175/1520-0442(1998)011<2942:RFMIAE>2.0.CO;2

Younas, W., & Tang, Y. (2013). PNA predictability at various time scales. *Journal of Climate*, *26*(22), 9090–9114. https://doi.org/10.1175/JCLI-D-12-00609.1

Yu, F., & Koltun, V. (2015). *Multi-Scale Context Aggregation by Dilated Convolutions*. http://arxiv.org/abs/1511.07122

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into Deep Learning. *ArXiv Preprint ArXiv:2106.11342*.

Zhang, C. (2005). Madden-julian oscillation. *Reviews of Geophysics*, *43*(2).

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155. https://doi.org/10.1109/TIP.2017.2662206

Zhang, T., Hoerling, M. P., Perlwitz, J., Sun, D.-Z., & Murray, D. (2011). Physics of U.S. Surface Temperature Response to ENSO. *Journal of Climate*, *24*(18), 4874–4887. https://doi.org/10.1175/2011JCLI3944.1

Zhang, T., Hoerling, M. P., Perlwitz, J., & Xu, T. (2016). Forced Atmospheric Teleconnections during 1979–2014. *Journal of Climate*, *29*(7), 2333–2357. https://doi.org/10.1175/JCLI-D-15-0226.1

Zhang, T., Perlwitz, J., & Hoerling, M. P. (2014). What is responsible for the strong observed asymmetry in teleconnections between El Niño and La Niña? *Geophysical Research Letters*, *41*(3), 1019–1025. https://doi.org/10.1002/2013GL058964

Zhang, Wei, & Kirtman, B. (2019). Estimates of Decadal Climate Predictability From an Interactive Ensemble Model. *Geophysical Research Letters*, *46*(6), 3387–3397. https://doi.org/10.1029/2018GL081307

Zhang, Wenjun, Wang, Z., Stuecker, M. F., Turner, A. G., Jin, F.-F., & Geng, X. (2019). Impact of ENSO longitudinal position on teleconnections to the NAO. *Climate Dynamics*, *52*(1), 257–274. https://doi.org/10.1007/s00382-018-4135-1

Zheng, X., Sugi, M., & Frederiksen, C. S. (2004). Interannual Variability and Predictability in an Ensemble of Climate Simulations with the MRI-JMA AGCM. *Journal of the Meteorological Society of Japan. Ser. II*, *82*(1), 1–18. https://doi.org/10.2151/jmsj.82.1

Zhou, W., Yang, D., Xie, S. P., & Ma, J. (2020). Amplified Madden–Julian oscillation impacts in the Pacific–North America region. *Nature Climate Change*, *10*(7), 654–660. https://doi.org/10.1038/s41558-020-0814-0

Zhou, Z. Q., Xie, S. P., Zheng, X. T., Liu, Q., & Wang, H. (2014). Global warming-induced changes in El Niño teleconnections over the North Pacific and North America. *Journal of Climate*, *27*(24), 9050–9064. https://doi.org/10.1175/JCLI-D-14-00254.1

Zhu, Y., & Newell, R. E. (1998). A proposed algorithm for moisture fluxes from atmospheric rivers. *Monthly Weather Review*, *126*(3), 725–735.