2015-12-02

Tropical Cyclone Intensity Forecast Error Predictions and Their Applications

Kieran T. Bhatia
University of Miami, kieran.bhatia@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/1537

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
UNIVERSITY OF MIAMI

TROPICAL CYCLONE INTENSITY FORECAST ERROR PREDICTIONS AND THEIR APPLICATIONS

By

Kieran T. Bhatia

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

December 2015
UNIVERSITY OF MIAMI

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

TROPICAL CYCLONE INTENSITY FORECAST ERROR
PREDICTIONS AND THEIR APPLICATIONS

Kieran T. Bhatia

Approved:

David S. Nolan, Ph.D.
Professor of Meteorology
and Physical Oceanography

Benjamin P. Kirtman, Ph.D.
Professor of Meteorology
and Physical Oceanography

Sharanya J. Majumdar, Ph.D.
Associate Professor of Meteorology
and Physical Oceanography

Dean of the Graduate School

Mark DeMaria, Ph.D.
Branch Chief
Technology and Science Branch
National Hurricane Center
Miami, Florida
This dissertation aims to improve tropical cyclone (TC) intensity forecasts by exploring the connection between intensity forecast error and parameters representing initial condition uncertainty, atmospheric flow stability, TC strength, and the large-scale environment surrounding a TC. After assessing which of these parameters have robust relationships with error, a set of predictors are selected to develop a priori estimates of intensity forecast accuracy for Atlantic basin TCs. The applications of these forecasts are then discussed, including a multimodel ensemble that unequally weights different intensity models according to the situation. The ultimate goal is to produce skillful forecasts of TC intensity error and use their output to enhance intensity forecasts.

To test if intensity forecast accuracy is linked to storm-specific characteristics and the synoptic situation experienced by the TC, a situation-dependent binning technique is used to verify past model performance. Three TC models, the Logistic Growth Equation Model (LGEM), Decay Statistical Hurricane Intensity Prediction Scheme (DSHP), and Geophysical Fluid Dynamics Laboratory (GFDL) Hurricane Model, as well as the National Hurricane Center (NHC) official forecast (OFCL), are evaluated for 24-, 48-,
and 72-hour intensity forecasts in the Atlantic basin. The bias, absolute error (AE), and skill relative to a benchmark model are binned in physically meaningful ways, and $t$ tests are used to measure the robustness of the results. The statistical significance established between different bins indicates that forecast error is often related to the nature of the particular storm and surrounding atmospheric environment.

Proxies for atmospheric flow stability and initial condition error are analyzed to supplement the dynamical parameters, and the expanded predictor pool is used to develop the Prediction of Intensity Model Error (PRIME) model. PRIME is trained to forecast both the AE and bias of Atlantic basin TCs from 2007 to 2014. These predictions of forecast error are formulated using a multiple linear regression scheme and applied to 12 to 120 hour intensity forecasts for DSHP, LGEM, Hurricane Weather Research and Forecasting Model Interpolated (HWFI), and GFDL Hurricane Model Interpolated (GHMI). The performance of PRIME is assessed by comparing its predictions of AE and bias to the climatology of these quantities for each of the models. Using paired $t$ tests, PRIME error forecasts are found to be more accurate than climatological error forecasts at the 99% significance level for bias and at the 95% significance level for AE.

PRIME forecasts are then tested for their ability to lower the error in the original intensity forecasts of each model. PRIME bias forecasts serve as real-time corrections to intensity forecasts and reduce their AE for all forecast hours and models. Also, PRIME AE and bias predictions are used to create unequally-weighted and modified equally-weighted ensembles of HWFI, GHMI, LGEM, and DSHP. Both of these types of ensembles are compared to ICON (equally-weighted multimodel consensus of the four models) and are found to frequently outperform it.
Additionally, following a similar framework to PRIME, multiple logistic regression is used to produce probabilistic PRIME (P-PRIME) forecasts because end users often care about the percentage likelihood of error exceeding a threshold rather than just a deterministic quantity. Brier Skill Scores, Ranked Probability Skill Scores, sharpness diagrams, and reliability diagrams demonstrate that PRIME probabilistic forecasts are skillful and reliable. The positive results and numerous applications of PRIME forecasts suggest they could be valuable to the hurricane forecasting community.
Acknowledgments

I would like to thank my advisor, David Nolan, for his guidance and patience for five years. His intelligent questioning and scientific integrity are traits that I will always admire. I would also like to express my gratitude to the rest of my doctoral committee, Mark DeMaria, Sharan Majumdar, and Ben Kirtman. A special thanks goes to Mark DeMaria. He answered more emails than I care to share while remaining unfailingly thoughtful and polite in his responses. Mark’s expertise is noticeable every time you talk to him, and I aspire to become a scientist like him one day. I also appreciate the help along the way from other scientists, Thomas Hamill, Jeff Whitaker, Shali Mohleji, Amy Clement, and Keith Brill.

To my friends at RSMAS and those who have kept in touch with me during my Ph.D., I am forever in debt. There are unfortunately too many individual names to list here but this group of people made my Miami years very special. I am especially grateful to my friends who have visited during my Ph.D.; I appreciate you all so much for the fun and memories we created. The larger RSMAS family also deserves a big thank you because the unique and interesting people I met at RSMAS made this Ph.D. less of a job and more of a constant exciting adventure.

Most importantly, my family has been and always will be everything. Jeremy, Rachel, Deepak, and Rebecca - thank you for your laughter, love, and acceptance. Your visits, phone calls, and messages provided the support I needed to accomplish this goal.
# Table of Contents

LIST OF FIGURES ..................................................................................................... vi

LIST OF TABLES ....................................................................................................... xiii

Chapter 1. Introduction .............................................................................................. 1
  1.1. Background and Motivation ................................................................. 1
  1.2. Atlantic Basin Intensity Error Trends ........................................... 8

Chapter 2. Situation-Dependent Verification of Intensity Forecasts ...................... 17
  2.1. Background and Motivation ................................................................. 17
  2.2. Data ...................................................................................................... 19
  2.3. Methodology ....................................................................................... 22
  2.4. Results .................................................................................................. 25
    2.4.a. One-variable Histograms ........................................................... 27
    2.4.b. Joint Histograms .......................................................................... 31
    2.4.c. Other Notable Results ................................................................. 33
  2.5. Summary and Conclusions ................................................................. 37

Chapter 3. Prediction of Intensity Model Error (PRIME) ........................................ 64
  3.1. Background and Motivation ................................................................. 64
  3.2. Data ...................................................................................................... 67
  3.3. Regression Analysis .............................................................................. 72
  3.4. Independent Testing of PRIME ............................................................ 79
    3.4.a. Separate Regression Models for Bias and AE ................................. 80
    3.4.b. Average Statistics of PRIME with and without Proxies .............. 82
    3.4.c. Average Statistics of PRIME for Land and No Land Cases ............ 82
    3.4.d. Comparison of Average Statistics for PRIME and R-PRIME ...... 83
      3.4.d.1. PRIME AE Forecasts ............................................................ 84
      3.4.d.2. PRIME Bias Forecasts ......................................................... 87
    3.4.e. Annual Variability of PRIME Performance .................................. 88
    3.4.f. Case Studies ................................................................................. 90
  3.5. Summary of PRIME and Potential Improvements ............................... 94

Chapter 4. Improving Intensity Forecasts with PRIME ......................................... 126
  4.1. Background and Motivation ................................................................. 126
  4.2. Upgraded PRIME and R-PRIME ....................................................... 129
  4.3. PRIME and R-PRIME Bias-corrected Models .................................... 134
  4.4. PRIME and R-PRIME Modified Ensembles ...................................... 135
    4.4.a. PRIME Results ............................................................................ 137
    4.4.b. R-PRIME Results ....................................................................... 140
  4.5. Summary and Conclusions ................................................................. 143
List of Figures

Fig 1.1: Time series of the MAE for NHC OFCL forecasts of track (top) and intensity (bottom) in the Atlantic basin. MAE is shown for 24, 48, 72, 96, and 120 hour forecasts with linear fits to each forecast hour indicated by the dashed lines............................. 12

Fig. 1.2: Time series of the MAE for NHC OFCL intensity forecasts in the Atlantic basin. Forecasts are included in the calculation of MAE if DSHP, LGEM, HWFI, and GHMI intensity forecasts exist as well as best track verification. Besides these criteria, the NHC OFCL verification rules are followed to determine which storms are verified. .......... 13

Fig. 1.3: Similar to figure 1.2 except skill score is plotted instead of MAE. ............ 14

Fig. 1.4: The dashed lines represent the time series of MAE for NHC OFCL forecasts of intensity and have the same values as the solid lines in figure 1.2. The solid lines show the time series of MAE of the best model (selected from DSHP, LGEM, GHMI, and HWFI) available for guidance. Section 1.2 provides an explanation of how MAE of the best model is determined. ................................................................. 15

Fig. 2.1: MAE of the DSHP, GFDL, LGEM, and OFCL 48-hour intensity forecasts as a function of 48-hour average potential intensity. The numbers at the bottom of each bar represent the amount of cases that are used to calculate the MAE of each bin. ........40

Fig. 2.2: MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average potential intensity. ................................................................. 41

Fig. 2.3: MAE of the DSHP, GFDL, LGEM, and OFCL 48-hour intensity forecasts as a function of 48-hour average potential intensity. Similar to figure 2.1 except equal-sized bins are used instead of manually-selected bins. ........................................................ 42

Fig. 2.4: MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 0-hour shear. ................................................................. 43

Fig. 2.5: MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average shear. ................................................................. 44

Fig 2.6: MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity. ................................................................. 45

Fig 2.7: Bias of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity. ................................................................. 46

Fig. 2.8: SS of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity. ................................................................. 47
Fig. 2.9: MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity and 0-hour shear magnitude. The numbers at the bottom of each box represent the amount of cases that are used to calculate the MAE of each bin. ............................................................... 48

Fig. 2.10: Bias of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity and 0-hour shear magnitude. .............................................. 49

Fig. 2.11: SS of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity and 0-hour shear magnitude. ............................................. 50

Fig. 2.12: MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction. ............... 51

Fig. 2.13: Bias of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction. ............... 52

Fig. 2.14: SS of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction. ............... 53

Fig. 2.15: MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction. Similar to figure 2.12 except with only two shear direction bins. ............................................... 54

Fig. 2.16: MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed. ................................................................. 55

Fig. 2.17: MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed and 0-hour potential intensity. ................................. 56

Fig. 2.18: Bias of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed and 0-hour potential intensity. ................................. 57

Fig. 2.19: SS of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed and 0-hour potential intensity. ................................. 58

Fig. 3.1: Time series of annual 24-120 hour absolute intensity forecast error of the best model available for guidance. The linear fit for each forecast hour is included as the dashed line. Section 1.2 provides an explanation of how the absolute error of the best model is determined. ................................................................. 98

Fig. 3.2: 120-hour HWFI intensity forecast bias plotted against deviation of HWFI intensity forecast from the ensemble mean for the 2007–14 Atlantic seasons. The dashed line represents the linear fit to the data. ................................................................. 99
Fig. 3.3: 72-hour GHMI absolute error plotted against GHMI absolute intensity forecast deviation from the ensemble mean (top left), ensemble intensity spread (top right), average latitude bottom left), and forecast intensity (bottom right) for the 2007–14 Atlantic seasons. The dashed lines represent the linear regression fits to the data. 

Fig. 3.4: Histograms showing the 72-hour real-time GHMI absolute error for the 2007-2013 Atlantic hurricane seasons before (top) and after (bottom) the Box-Cox transformation. A $\lambda$ of 0.22 along with the original absolute error values is inputted into equation (3) to obtain the bottom histogram. 

Fig. 3.5: 108-hour retrospective DSHP bias plotted against 0-hour distance to land (LDIS) for the 2008-2013 Atlantic hurricane seasons before (top) and after (bottom) a second-order polynomial transformation of the predictor. The top figure shows the raw LDIS values for the DSHP model plotted against 108-hour bias. The dashed red curve is fit using data from all forecast intervals. The bottom figure displays bias versus fitted LDIS. The solid blue line represents the linear fit to the data. The $R$ values in the top right corner of each plot capture the linear correlation of the independent and dependent variable. 

Fig. 3.6: Idealized scenario explaining the different behavior certain predictors displayed when plotted against bias (top image) and absolute error (bottom image). The solid blue lines indicate the linear fits to the data. R values describing linear correlation of the predictor and predictand are included in bottom right corner of the images. This predictor shows higher linear correlation with absolute error than bias. 

Fig. 3.7: Similar to figure 3.6 except this predictor shows higher linear correlation with bias than absolute error. 

Fig. 3.8: Predicted AE versus predicted $|bias|$ for 72-hour LGEM forecasts. The 2007-2013 hurricane seasons are used to develop the regression formula for these PRIME predictions and 2014 serves as the verification period. The solid blue line indicates the linear fit to the data, and the $R$ value of the line is 0.10. 

Fig. 3.9: The average absolute error of PRIME $|bias|$ forecasts, PRIME AE forecasts, and climatological absolute error forecasts for HWFI, GHMI, LGEM, and DSHP. Error statistics are calculated using cases from 2007-2014 that are described in table 3.1. 

Fig. 3.10: The average absolute error of absolute error forecasts for PRIME with proxies, PRIME without proxies, and climatological forecasts for HWFI, GHMI, LGEM, and DSHP. Error statistics are calculated using cases from 2007-2014 that are described in table 3.1. 

Fig. 3.11: Similar to figure 3.10 except the plots apply to bias predictions. 

Fig. 3.12: The average absolute error of PRIME, retrospective PRIME (R-PRIME), and climatological absolute error forecasts for HWFI, GHMI, LGEM, and DSHP. Error
statistics are calculated using cases from 2008-2014 that are available for both PRIME and R-PRIME.

Fig. 3.13: Similar to figure 3.12 except the plots apply to bias predictions.

Fig. 3.14: The yearly skill score of retrospective PRIME (R-PRIME) absolute error forecasts for each model and forecast hour. The skill score of the total sample is demarcated with a dashed black line and cases from 2008, 2010, 2011, 2012, and 2014 are represented by different colored lines. A positive skill score indicates PRIME has lower absolute error than climatology. The 2009 and 2013 seasons are excluded from the plots because of the non-representative behavior described in section 3.4.

Fig. 3.15: Similar to figure 3.14 except it shows the skill score of retrospective PRIME (R-PRIME) bias forecasts normalized by climatological bias forecasts.

Fig. 3.16: Predicted bias versus true bias for 96-hour DSHP forecasts. The 2008 and 2010-2014 hurricane seasons are used to develop the regression formula for these retrospective PRIME (R-PRIME) predictions and 2009 serves as the verification period. The dashed line indicates the linear fit to the data, and the $R$ value of the line is included in the top right corner of the figure.

Fig. 3.17: Predicted absolute error (AE) versus true AE for 48-hour LGEM forecasts. The 2008-2009 and 2011-2014 hurricane seasons are used to develop the regression formula for these retrospective PRIME (R-PRIME) predictions, and 2010 serves as the verification period. The dashed line indicates the linear fit to the data, and the $R$ value of the line is included in the top right corner of the figure. The red box highlights cases where SPRD and ADEM are especially poor predictors.

Fig. 3.18: The observed best-track intensity (solid black line) and the 72-hour, 96-hour, and 120-hour bias-corrected HWFI intensity forecasts for Tropical Storm Dorian in 2013. Retrospective PRIME (R-PRIME) bias forecasts are used to adjust the original intensity forecast of HWFI. 72-hour bias-corrected forecasts are indicated by blue x’s, 96-hour bias-corrected forecasts are indicated by green triangles, and 120-hour bias-corrected forecasts are indicated with red plusses.

Fig. 4.1: 96-hour retrospective DSHP transformed AE vs 96-hour DSHP forecast intensity (FINT) for the 2009-2014 Atlantic hurricane seasons before (top) and after (bottom) a third-order polynomial transformation of the predictor. The top figure shows the raw DSHP FINT values plotted against 96-hour retrospective DSHP transformed AE. The dashed red curve is fit using data from all forecast intervals during 2009-2014. The bottom figure shows transformed AE vs fitted FINT. The solid blue line represents the linear fit to the data. The $R$ values in the top right corner of each plot capture the linear correlation of the independent and dependent variable.

Fig. 4.2: 96-hour retrospective DSHP bias vs 0-hour divergence (0DIV) for the 2008-2013 Atlantic hurricane seasons before (top) and after (bottom) a third-order polynomial
transformation of the predictor. The top figure shows the raw 0DIV values plotted against 96-hour retrospective DSHP bias. The dashed red curve is fit using data from all forecast intervals during 2008-2013. The bottom figure shows bias vs fitted 0DIV and the linear fit to the data. ............................................................................................................. 146

Fig. 4.3: 108-hour retrospective HWFI bias vs. forecast average latitude (ALAT) for the 2008-2012 and 2014 Atlantic hurricane seasons before (top) and after (bottom) a two-peak Gaussian transformation of the predictor. The top figure shows the raw ALAT values plotted against 108-hour retrospective HWFI bias. The dashed red curve is fit using data from all forecast intervals during 2008-2012 and 2014. The bottom figure shows bias vs fitted ALAT and the linear fit to the data. ............................................ 147

Fig. 4.4: The average absolute error of HWFI, GHMI, LGEM, and DSHP before and after PRIME bias corrections. The dashed lines represent the bias-corrected models and the solid lines represent the real-time models. Error statistics are calculated using cases from 2007-2014 that are listed in table 3.1. ................................................................. 148

Fig. 4.5: The average absolute error of retrospective HWFI, GHMI, LGEM, and DSHP before and after R-PRIME bias corrections. The dashed lines represent the bias-corrected models and the solid lines represent the retrospective models. Error statistics are calculated using cases from 2008-2014 that are listed in table 3.2. ......................................... 149

Fig. 4.6: The average absolute error of the Bias-Corrected ensemble, Exclude Worst (AE) ensemble, Correct Worst ensemble, and ICON forecasts. These ensembles were formulated using real-time forecasts of DSHP, LGEM, GHMI, and HWFI from 2007-2014 and verified with best track information. The first three ensembles were modified using PRIME forecasts of AE and bias. ............................................................................ 150

Fig. 4.7: The average absolute error of ICON and the Unequal (AE), Unequal (AE SQR), and Unequal (BIAS) ensembles. These ensembles were formulated using real-time forecasts of DSHP, LGEM, GHMI, and HWFI from 2007-2014 and verified with best track information. The first three ensembles were unequally-weighted using PRIME forecasts of AE and bias. ................................................................. 151

Fig. 4.8: The average absolute error of ICON and the Unequal (AE), Unequal (BIAS), and Bias-Corrected ensembles. These PRIME modified ensembles recorded the lowest AE. Only AE for 48-120 hour forecasts is plotted to capture how the AE of the ensembles separate at longer forecast hours. ............................................................................. 152

Fig. 4.9: The AE of the ICON and PRIME Bias-Corrected ensemble are plotted for cases when their forecasts differ by more than 2.5 and 5 knots. The 2.5 knots cases are indicated with black lines, and “(2.5)” is added to names of the ensembles. The 5 knots cases are indicated with red lines and “(5)” is added to the names of the ensembles. The dashed lines represent ICON forecasts and the solid lines represent Bias-Corrected ensemble forecasts. Table 4.1 lists the number of forecasts for the different cases at each forecast hour. ............................................................................. 153
Fig. 4.10: The AE of the ICON and PRIME Bias-Corrected ensemble are plotted for cases when their forecasts differ by more than 10 knots. The dashed lines represent ICON forecasts and the solid lines represent Bias-Corrected ensemble forecasts. The third column of Table 4.1 lists the number of forecasts at each forecast hour. .................. 154

Fig. 4.11: The average absolute error of the Bias-Corrected ensemble, Exclude Worst (AE) ensemble, Correct Worst ensemble, and ICON forecasts. These ensembles were formulated using retrospective forecasts of DSHP, LGEM, GHMI, and HWFI from 2008-2014 and verified with best track information. The first three ensembles were modified using R-PRIME forecasts of AE and bias. .......................................................... 155

Fig. 4.12: The average absolute error of ICON and the Unequal (AE), Unequal (AE SQR), and Unequal (BIAS) ensembles. These ensembles were formulated using retrospective forecasts of DSHP, LGEM, GHMI, and HWFI from 2008-2014 and verified with best track information. The first three ensembles were unequally-weighted using R-PRIME forecasts of AE and bias. .......................................................... 156

Fig. 4.13: The average absolute error of ICON and the Unequal (AE), Unequal (AE SQR), and Bias-Corrected ensembles. These R-PRIME modified ensembles recorded the lowest average AE during 2008-2014. Only AE for 48-120 hour forecasts is plotted to capture how the AE of the ensembles separate at longer forecast hours. ................... 157

Fig. 4.14: The AE of the ICON and R-PRIME Unequal (AE SQR) ensemble are plotted for cases when their forecasts differ by more than 1.5 and 3 knots. The 1.5 knots cases are indicated with black lines and “(1.5)” is added to names of the ensembles. The 3 knots cases are indicated with red lines and “(3)” is added to the names of the ensembles. The dashed lines represent ICON forecasts and the solid lines represent Unequal (AE SQR) ensemble forecasts. Table 4.2 lists the number of forecasts for the different cases at each forecast hour. .......................................................... 158

Fig 5.1: (a) Reliability diagram for P-PRIME forecasts of 108-hour DSHP AE greater than 20 knots during the period 2008-2014. Dots show the observed frequency of AE exceeding 20 knots, conditional on each of the 10 possible probability forecast bins. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins. ...................... 183

Fig 5.2: (a) Reliability diagram for P-PRIME forecasts of 120-hour DSHP AE greater than 20 knots during the period 2008-2014. Dots show the observed frequency of AE exceeding 20 knots, conditional on each of the 10 possible probability forecast bins. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins. ...................... 184

Fig 5.3: (a) Reliability diagram for P-PRIME forecasts of 96-hour HWFI bias greater than 0 knots during the period 2008-2014. Bins with less than 5 entries are not included
in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins. ................................................................. 185

Fig 5.4: (a) Reliability diagram for P-PRIME forecasts of 84-hour LGEM AE greater than 20 knots (blue line: U) and AE lower than 10 knots (red line: L) during the period 2008-2014. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the lower tercile. (c) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the upper tercile. .......................... 186

Fig 5.5: (a) Reliability diagram for P-PRIME forecasts of 60-hour LGEM bias greater than 15 knots (blue line: U) and bias lower than 15 knots (red line: L) during the period 2008-2014. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the lower tercile. (c) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the upper tercile. ................................. 187
List of Tables

Table 1.1: Dynamical and proxy predictors for PRIME. Abbreviations are listed for each predictor. Bold predictor abbreviations indicate predictors whose 0-hour and forecast average value are both used. If the predictor value varies depending on the model, a Y is listed in the third column. An N is used if the same data is used to produce the predictor for every model. .......................................................... 16

Table 2.1: Number of verified forecasts (best track verification), for DSHP, GFDL, LGEM, and OFCL at each of the forecast periods. These cases’ totals are for Atlantic basin storms between 2006 and 2010. ................................................................. 59

Table 2.2: MAE in knots, using best track verification, for DSHP, GFDL, LGEM, and OFCL at each forecast period. ................................................................. 60

Table 2.3: Bias in knots, using best track verification, for DSHP, GFDL, LGEM, and OFCL at each forecast period. ................................................................. 61

Table 2.4: MAE in knots, using 0-hour operational analyses, for DSHP, GFDL, LGEM, and OFCL at each forecast period. ...................................................... 62

Table 2.5: Bias in knots, using 0-hour operational analyses, for DSHP, GFDL, LGEM, and OFCL at each forecast period. ...................................................... 63

Table 3.1: The average bias and absolute error of HWFI, GHMI, LGEM, and DSHP is calculated using the real-time forecasts of Atlantic basin storms between 2007 and 2014. The first column of the table represents the number of verified real-time forecasts. For each model and forecast interval, the first entry in a table cell is the bias and the second entry is the absolute error. ................................................................. 116

Table 3.2: Similar to table 3.1 except the entries correspond to the statistics of verified retrospective forecasts for each model at each forecast interval. These values are calculated using Atlantic basin storms between 2008 and 2014. ......................................... 117

Table 3.3: The percent of the total variance explained ($R^2$) by PRIME for dependent verification of absolute error forecasts for each model and forecast hour. These values are calculated using the same cases as table 3.1. ................................................................. 118

Table 3.4: Similar to table 3.3 except the $R^2$ values apply to bias forecasts for each model and forecast hour. ................................................................. 119

Table 3.5: Normalized regression coefficients for the three best predictors of absolute error for each model. These weighting coefficients were derived from the retrospective version of PRIME (R-PRIME), using 2008-2014 as the training period. ...................... 120
Table 3.6: Similar to table 3.5 except the normalized regression coefficients apply to the best predictors of bias for each model. .............................................................. 121

Table 3.7: The percent of the total absolute error variance explained ($R^2$) by PRIME for land (left column) and no land (right column) cases for each model and forecast interval. Bold numbers are used to indicate which column has a higher $R^2$ value for the model-forecast hour pair. .......................................................... 122

Table 3.8: Similar to table 3.7 except the $R^2$ values apply to bias forecasts. ........ 123

Table 3.9: The percent of the total absolute error variance explained ($R^2$) by retrospective (left column) and real-time (right column) PRIME for each model and forecast interval. Bold numbers are used to indicate which version of PRIME has a higher $R^2$ value for the model-forecast hour pair. ..................................................................................... 124

Table 3.10: Similar to table 3.9 except the $R^2$ values apply to bias forecasts. ........ 125

Table 4.1: The percent of the total real-time cases listed in table 3.1 where ICON and the PRIME Bias-Corrected ensemble differ by 2.5, 5, and 10 knots. Each column represents the magnitude of the difference between the intensity forecasts of ICON and the Bias-Corrected ensemble. ................................................................. 159

Table 4.2: The percent of the total retrospective cases listed in table 3.2 where ICON and the R-PRIME Unequal (AE SQR) ensemble differ by 1.5 and 3 knots. Each column represents the magnitude of the difference between the intensity forecasts of ICON and the Unequal (AE SQR) ensemble. ............................................................................... 160

Table 5.1: The Brier Skill Scores of P-PRIME binary forecasts of AE greater than 20 knots for each model and forecast interval. These results apply to retrospective forecasts, which were generated by applying the 2014 version of each model to the 2008-2014 hurricane seasons. ............................................................ 188

Table 5.2: Similar to table 5.1 except the Brier Skill Scores apply to P-PRIME forecasts of bias greater than 0 knots. ................................................................. 189

Table 5.3: The Ranked Probability Skill Scores of P-PRIME ternary forecasts of AE for each model and forecast interval. The thresholds set for the predictand were 10 and 20 knots of AE. ................................................................. 190

Table 5.4: The Ranked Probability Skill Scores of P-PRIME ternary forecasts of bias for each model and forecast interval. The thresholds set for the predictand were -15 knots and 15 knots of bias. ................................................................. 191

Table 5.5: The BSS of P-PRIME ternary forecasts of LGEM AE for each forecast interval. The BSS of each event outcome is represented with a column in the table. 192
Table 5.6: The BSS of P-PRIME ternary forecasts of LGEM bias for each forecast interval. The BSS of each event outcome is represented with a column in the table.
Chapter 1. Introduction

1.1 Background and Motivation

There are several methods for quantifying the value of a meteorological forecast: social science studies monitor how forecasts influence the well-being and behavior of individuals, economic studies measure how forecasts affect financial systems and policy decisions, and scientific studies evaluate forecasts using a variety of metrics that compare predicted quantities to reality (Katz and Murphy 1993). In all cases, a weather forecast holds no innate value and only becomes meaningful to society when it influences the decisions made by its end users (Murphy 1993). Using this logic, tropical cyclone (TC) forecasts appear incredibly valuable. At least a few times a year, emergency managers, public officials, businesses, and citizens make costly and life-changing decisions based on TC forecasts. As a result, it is imperative that these forecasts are both skillful and reliable.

For the duration of this dissertation, the discussion will specifically revolve around TC intensity forecasts. The word “intensity” refers to the wind speed of a TC at the appropriate forecast or verification time. More precisely, TC intensity is defined as the maximum one-minute average sustained surface (10 m) wind (National Weather Service Directive System 10-604). For consistency, any measure of intensity for the rest of this dissertation will have the units of knots (1 kt = 0.52 m s⁻¹). Over the last 25 years, intensity forecasts have barely improved in the Eastern Pacific and Atlantic basins, while the errors in operational track forecasts have considerably decreased (Cangialosi and Franklin 2015). The error trends in Atlantic basin intensity forecasts are particularly
relevant to this dissertation and are discussed further in section 1.2. TC intensity forecast performance can also differ considerably between models, years, and storms. The mediocre progress and erratic skill of TC intensity forecasts has diminished their value and demands attention from the scientific community.

Creating forecasts for a meteorological variable, such as TC intensity, is a complex problem due to initial condition errors, imperfect model formulations, and the inherent uncertainty associated with the particular atmospheric flow pattern (Kalnay and Dalcher 1987). To mitigate the effects of these forecast obstacles and improve TC intensity forecast performance, the meteorological research community recommends finer spatial and temporal resolution of dynamical models, more advanced data assimilation techniques to better incorporate real-time measurements into forecast equations, and the acquisition of additional observations by deploying more instrumentation (Zhang 2011). Finer horizontal, vertical, and temporal spacing produces dynamical features that span meters and minutes which are important for predicting intensity (Nolan et al. 2013; Gopalakrishnan et al. 2011; Gentry and Lackmann 2010). However, for the foreseeable future, computational constraints will only allow operational models to calculate wind speeds every few hours and kilometers. More sophisticated data assimilation schemes (4D-Variational [4D-VAR]) require tedious calculations and are too inefficient to be included in today’s real-time TC forecasts. Finally, a large improvement in the quality and quantity of observations is unlikely due to financial limitations (National Research Council 2012). Therefore, a near-term improvement in TC intensity forecasts must involve an alternative solution.
A potential way to increase the value of TC intensity forecasts is to skillfully quantify the error expected in individual forecasts of the premier hurricane models. Current operational TC forecasts do not provide a situation-dependent estimation of intensity forecast error, which suggests that each forecast will have approximately the same error as the climatological average\(^1\). The National Hurricane Center (NHC) includes a wind speed probability graphic in their suite of products that provides users with an array of probabilities for different intensity outcomes. However, the percentage likelihood of the different intensity outcomes only considers case-dependent track error information and does not use intensity uncertainty information unique to the particular storm (or its environment) being evaluated (DeMaria et al. 2013). End users who rely on the wind speed probabilities for important financial decisions and public safety cannot easily translate the presented data into a measure of confidence in the deterministic forecast. Additionally, the general public rarely sees this forecast product and typically only receives uncertainty information in the form of the NHC forecast track “cone of uncertainty”. As a result, the general public and end users are unaware of the uncertainty in TC intensity forecasts and under these circumstances, they often incorrectly estimate uncertainty themselves (Joslyn and Savelli 2010).

TC forecasting agencies can provide much-needed support to decision makers by adding case-dependent forecasts of intensity forecast error to their suite of products. The economic value of having prior knowledge whether a particular forecast will be more or less reliable than average is well documented (Pielke and Carbone 2002; Katz and Murphy 1997; Wilks and Hamill 1995). Although applying such predictions to TC

\(^1\) Throughout the dissertation, the term climatology is used to represent the average weather statistics accumulated over historical data.
intensity forecasts is only recently receiving attention, it is not a new concept that an a priori expectation of forecast error is a necessary part of every forecast (Kalnay and Dalcher 1987; Molteni and Palmer 1991; Tennekes et al. 1987; Palmer and Tibaldi 1988).

In fact, Epstein (1969) provided the first framework for estimating forecast error in real-time over 45 years ago when he created a stochastic-dynamic (variation of the ensemble technique) forecast technique to model the future state of the atmosphere as a probability distribution. In his seminal publication, Epstein postulated that the spread of the future probability distribution was correlated with the level of uncertainty in the initial and future state of the atmosphere as well as the flaws in the model formulations. Soon after, Leith (1974) suggested the more computationally efficient ensemble forecast technique of “Monte Carlo forecasting” and hypothesized that the dispersion between the forecasted states of different ensemble members was related to the skill of the ensemble mean forecast. Both of these studies concluded that a more uncertain atmosphere results in more disagreement between ensemble members and less forecast skill.

Since then, several publications have acknowledged the relationship between forecast error and ensemble spread and applied it to predicting the forecast error of large-scale fields of atmospheric variables (see the reviews by Ehrendorfer 1997; Wobus and Kalnay 1995). These error predictions achieved varying levels of success and there appears to be an upper limit for the predictability of forecast error obtainable from ensemble forecasting (Wobus and Kalnay 1995). However, ensemble techniques represent only one approach to anticipating the error of a particular forecast. Statistical methods are an alternative that are simpler and less computationally expensive.
Previous studies using a statistical approach to predicting forecast error have focused on proxies\(^2\) for the intrinsic stability of the atmospheric state and the quality of the initial conditions of a model as independent variables. The uncertainty in the atmospheric state is usually represented by parameters that capture the persistence of the atmosphere, large-scale flow pattern, and skill of recent forecasts (Molteni and Palmer 1991; Palmer and Tibaldi 1988). Variables that represent the uncertainty in the initial state are less straightforward but examples consistent with physical reasoning include discrepancies between different models’ initial analyses (Kalnay and Dalcher 1986) and the recent change in the forecasted variable. After choosing a pool of predictors, a predictand related to the performance of the forecast (typically root-mean-square error [RMSE], anomaly correlation coefficient [ACC], or mean absolute error [MAE]) is selected. Then, a statistical regression analysis is applied to a training dataset and the most statistically significant predictors are used for real-time predictions of forecast error in the verification dataset. Statistical techniques have exhibited some drawbacks in past studies (Ehrendorfer 1997) but their potential has yet to be realized for TC intensity error forecasts. TC models exhibit different error patterns than the forecasts of global and regional atmospheric fields. The large fluctuations between adjacent forecasts and the well-documented relationships between storm evolution and environment (Frank and Ritchie 2001; Rogers et al. 2003; Zhang and Sippel 2009) are promising for skillful predictions of TC intensity forecast error.

This dissertation discusses how statistical techniques can be applied to easily-accessible data to create more reliable TC intensity guidance. As a first step towards

---

\(^2\) Proxy is a term used throughout this dissertation to label an empirically-derived variable that attempts to encapsulate the uncertainty caused by atmospheric properties or model deficiencies.
producing error predictions to accompany each TC intensity forecast, Bhatia and Nolan (2013; hereafter, BN13) examined the relationship between synoptic parameters, TC attributes, and forecast errors. Chapter 2 of this dissertation presents the results of BN13. In that study, storm speed, latitude, initial intensity, potential intensity, wind shear magnitude, and direction of the wind shear vector were plotted against mean absolute error (MAE), bias, and skill score (SS). The performance of the operational Atlantic basin intensity models significantly varied based on these parameters. Although the results of BN13 addressed conventional wisdom about which dynamical variables lead to better forecasts of TC intensity and highlighted the different strengths of TC models, it also prompted additional questions. What is the relative importance of situation-defining parameters for TC model performance? For example, how reliable is a forecast when the initial latitude of a TC is in a low-confidence regime but its intensity is in a high-confidence regime? If certain features and environments make TCs inherently more difficult to forecast, can parameters representing them predict TC intensity forecast error in real time?

Motivated by the desire to address the questions raised by BN13, one of the primary goals of this dissertation is to develop a regression model that uses various meteorological parameters to forecast the performance of TC intensity forecasts. Forecasting a TC is an incredibly complex problem that potentially involves countless predictors. Therefore, it is necessary to follow a methodology to select the minimum number of variables that sufficiently capture the different regimes conducive to forecasts with higher or lower error and weight the variables according to importance. In a manner similar to the development of the Statistical Hurricane Intensity Prediction Scheme
(SHIPS; DeMaria and Kaplan 1994; DeMaria et al. 2005) model, a multiple linear regression (Draper and Smith 1998) technique called the Prediction of Intensity Model Error (PRIME) model (Bhatia and Nolan 2015) is created to forecast the absolute error (AE) and bias of four “early” Atlantic basin intensity models. The independent variables used in PRIME include the parameters discussed in BN13, as well as proxies for atmospheric flow stability and the quality of each model’s initial conditions. Table 1.1 lists the entire set of predictors used in this dissertation along with their abbreviated names and whether information unique to each model is involved in the calculation of the predictor. The development of the PRIME model is described in more detail in chapter 3 of this dissertation.

PRIME error forecasts significantly outperformed climatological error forecasts at all forecast intervals, which implies that PRIME output could provide more than just error guidance. Chapter 4 tests the ability of PRIME to reduce intensity forecast error. PRIME bias forecasts are utilized to correct TC model intensity forecasts and significantly reduced the error for all models and forecast hours. Additionally, two types of ensemble post-processing techniques involving PRIME are proposed to improve upon traditional multimodel ensembles. For the first technique, PRIME error forecasts are used to eliminate or bias-correct models that are forecasted to have higher errors, and then the ensemble mean of the modified models is calculated. Secondly, an inverse-weighting (higher errors correspond to lower weights) approach is applied to PRIME AE and bias predictions where each model is weighted proportionally to their expected error.

In chapter 5, Probabilistic PRIME (P-PRIME) forecasts are produced using a similar framework to PRIME but multiple logistic regression replaces multiple linear regression.
Verification of P-PRIME forecasts confirms they are also skillful. Chapter 6 presents an overall summary of this dissertation as well as potential future work.

1.2 Atlantic Basin Intensity Error Trends

Over the last 25 years, the MAE of Atlantic basin OFCL intensity forecasts issued by NHC has remained largely constant. During the same period, it is well accepted that the associated operational track forecasts have improved dramatically. Figure 1.1 documents these trends with a time series of the errors in operational TC track forecasts from NHC for the Atlantic Basin at 24, 48, 72, 96, and 120 hours (top), and a similar plot for the intensity errors (bottom). The linear fit for each forecast hour is demarcated with a dashed line. Clearly, besides the last eight years, the MAE of intensity forecasts has barely changed or even increased at some forecast hours. A recent study by DeMaria et al. (2014) pointed out that this lack of improvement in intensity forecasts is largely attributable to operational models delivering inconsistent results to NHC forecasters. The authors demonstrated that more trust in guidance is now warranted, and following the Atlantic basin intensity models available from 1989-2012 more closely would have led to a more statistically-significant downward trend. To arrive at this surprising result, DeMaria et al. considered the scenario where NHC forecasters could theoretically select the model or consensus forecasting technique with the lowest average error over a multiple-season period (picking the best model in each year was considered unrealistic for NHC forecasters) and use its forecast to produce the OFCL forecast. Although this pre-season selection of a forecasting technique does not perfectly simulate the subjective work of forecasters, it is a reasonable representation of humans attempting to determine the best-performing individual model or consensus forecast.
At first glance, the recent decrease in error seems justified. 2007 was the last year that a major model was added to the suite of intensity guidance for Atlantic basin TCs, and additional forecasting resources have surfaced in recent years. The two best performing statistical models currently available, the Decay Statistical Hurricane Intensity Prediction Scheme (DSHP: Decay-SHIPS; DeMaria and Kaplan 1994; DeMaria et al. 2005; DeMaria et al. 2006) and Logistic Growth Equation Model (LGEM; DeMaria 2009), were implemented for operational use in 2000 and 2006, respectively. The Atlantic basin early\(^3\) dynamical models with the lowest error, the Geophysical Fluid Dynamics Laboratory Hurricane Model Interpolated (GFDL interpolated: GHMI; Bender et al. 2007) and Hurricane Weather Research and Forecasting Model Interpolated (HWRF interpolated: HWFI; Gopalakrishnan et al. 2010) became operational in 2006 and 2007, respectively\(^4\). The recent introduction of HWFI, GHMI, LGEM, and DSHP to the operational community has resulted in an unprecedented period of skillful guidance for forecasters and end users of TC intensity forecasts. In figure 1.2, the MAE of 24, 48, 72, 96, and 120 hour Atlantic basin OFCL forecasts is plotted for the 2007-2014 seasons with dashed linear trends overlaid. The storms included in this analysis are selected based on the verification rules used by NHC when it determines OFCL intensity errors. When applying the Student’s \(t\) test, the slopes of the regression lines are found to be significantly different from 0 at the 95\% level for each forecast hour.

---

\(^3\) Intensity models are labeled as either early or late, depending on whether or not they are available to guide the NHC forecaster during the forecast cycle. For example, consider the 1200 UTC (12Z) forecast cycle, which begins with the 12Z synoptic time and ends with the release of the OFCL forecast at 15Z. The 12Z GFDL run is not available to the forecaster until about 16Z so it is called a late model. The GHMI would interpolate data from the 6Z run to provide its forecast by the 12Z deadline and therefore is an early model.

\(^4\) A general discussion of the post-processing technique required to interpolate the HWRF and GFDL forecasts from the previous cycle to the current forecast time, thereby creating HWFI and GHMI, is described in Sampson et al. (2006).
However, the noticeable drop in MAE since 2007 in figures 1.1 and 1.2 and statistical significance observed by DeMaria et al. (2014) is largely due to less challenging seasons. SS is an appropriate metric for distinguishing whether models have improved because it compares their forecasts to a benchmark forecasting tool. OCD5 is used here as the benchmark model because it is formulated by combining a simple track and intensity model, the 5-Day Decay Statistical Hurricane Intensity Forecast (Decay-SHF5: DSF5; Knaff et al. 2003) and the CLIPER5 (CLP5; Neumann 1972; Aberson 1998). SS is calculated as follows:

\[
SS = 100 \times \left( 1 - \left( \frac{E_{\text{MODEL}}}{E_{\text{BENCHMARK\_MODEL}}} \right)^2 \right)
\]

Here, \(E_{\text{MODEL}}\) represents the error (i.e. AE, MSE, etc.) for the forecast model being tested and \(E_{\text{BENCHMARK\_MODEL}}\) represents the error of the benchmark model. In this case, a positive SS indicates a model obtains greater accuracy than the OCD5 model, which translates to a model outperforming a forecast based solely on climatology and persistence. If forecasts are easier during a hurricane season, then in theory, OCD5 will perform at a higher level, and each model will need lower errors to achieve a positive SS.

Figure 1.3 is similar to figure 1.2 but the dependent variable is SS instead of MAE. Although the trend line of each forecast hour shows an increase in SS from 2007-2014, the Student’s \(t\) test reveals none of the slopes of the regression lines are statistically different from 0 at the 90% level. In other words, the SS of TC intensity forecasts issued by the NHC has not significantly increased over the last eight years even with the heralded upgrades to the suite of operational models. This surprising trend prompts the question: was objective intensity guidance sufficient to improve the NHC OFCL forecasts? This question is addressed here by considering an idealized scenario where a
priori information is available to select the model (DSHP, LGEM, HWFI, or GHMI) with the lowest AE for every individual forecast. Figure 1.4 shows the MAE of this “best model” scenario at every forecast hour with the corresponding OFCL MAE. Using a paired $t$ test, the MAE of the best model scenario is found to be significantly less than the MAE of the OFCL forecast at each forecast hour for the 2007-2014 period. Therefore, there is major incentive to develop an objective technique to reliably anticipate the best performing model at the time of the forecast. One of the main goals of this dissertation is to close the gap between the best model forecast and the OFCL forecast to improve skill.
Fig. 1.1

Time series of the MAE for NHC OFCL forecasts of track (top) and intensity (bottom) in the Atlantic basin. MAE is shown for 24, 48, 72, 96, and 120 hour forecasts with linear fits to each forecast hour indicated by the dashed lines.
Fig. 1.2

Time series of the MAE for NHC OFCL intensity forecasts in the Atlantic basin. Forecasts are included in the calculation of MAE if DSHP, LGEM, HWFI, and GHMI intensity forecasts exist as well as best track verification. Besides these criteria, the NHC OFCL verification rules are followed to determine which storms are verified.
Fig. 1.3

Similar to figure 1.2 except skill score is plotted instead of MAE.
Fig. 1.4

The dashed lines represent the time series of MAE for NHC OFCL forecasts of intensity and have the same values as the solid lines in figure 1.2. The solid lines show the time series of MAE of the best model (selected from DSHP, LGEM, GHMI, and HWFI) available for guidance. Section 1.2 provides an explanation of how MAE of the best model is determined.
<table>
<thead>
<tr>
<th><strong>Dynamical Predictors</strong></th>
<th><strong>Abbreviation</strong></th>
<th><strong>Changes with Model? (Y/N)</strong></th>
<th></th>
<th><strong>Proxies</strong></th>
<th><strong>Abbreviation</strong></th>
<th><strong>Changes with Model? (Y/N)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>LAT</td>
<td>N</td>
<td>Standard deviation of the intensity forecast ensemble</td>
<td>SPRD</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Longitude</td>
<td>LON</td>
<td>N</td>
<td>Absolute DFEM</td>
<td>ADEM</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Vertical shear magnitude</td>
<td>SHR</td>
<td>N</td>
<td>Deviation of intensity forecast from ensemble mean</td>
<td>DFEM</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Storm speed</td>
<td>SSPD</td>
<td>N</td>
<td>Deviation of track forecast from ensemble mean</td>
<td>DTRK</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Sin(shear direction)</td>
<td>SHRDIR</td>
<td>N</td>
<td>Forecasted intensity change</td>
<td>FCIC</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Ocean heat content</td>
<td>OHC</td>
<td>N</td>
<td>Absolute FCIC</td>
<td>AFIC</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Potential intensity</td>
<td>POT</td>
<td>N</td>
<td>Previous 12-hour intensity change</td>
<td>P12C</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Vorticity</td>
<td>VOR</td>
<td>N</td>
<td>Previous 12-hour error</td>
<td>P12E</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Divergence</td>
<td>DIV</td>
<td>N</td>
<td>Distance to land</td>
<td>LDIS</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Relative humidity</td>
<td>RH</td>
<td>N</td>
<td>Forecast distance to land</td>
<td>FLND</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Initial intensity</td>
<td>0INT</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forecast intensity</td>
<td>FINT</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% area of GOES cold pixels</td>
<td>GCLD</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation of GOES brightness temperature</td>
<td>GBRT</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1

Dynamical and proxy predictors for PRIME. Abbreviations are listed for each predictor. Bold predictor abbreviations indicate predictors whose 0-hour and forecast average value are both used. If the predictor value varies depending on the model, a Y is listed in the third column. An N is used if the same data is used to produce the predictor for every model.
Chapter 2. Situation-Dependent Verification of Intensity Forecasts

2.1 Background and Motivation

All weather forecasts should not be treated as equal. The quality of forecasts varies considerably due to two main sources of error. First, scientists cannot perfectly observe and simulate the atmosphere based on computational and instrumentation limitations. Second, forecasts will always be negatively affected by the intrinsic instabilities and nonlinearity of the atmospheric state during the forecast (Palmer and Tibaldi 1988). The fluctuating performance of weather forecasts is particularly noticeable for TC intensity forecasts where two consecutive forecasts can occasionally differ in error by over 100 miles per hour. As a result, an a priori expectation of the accuracy of every weather forecast, including TC intensity forecasts, is necessary. Such an expectation is easily provided by climatological error forecasts which are derived by verifying a past set of model predictions, and averaging the resulting forecast errors. However, the goal of this chapter is to determine if there are situations where model performance differs significantly from average error. Understanding the relationship between a particular atmospheric regime and forecast performance in past hurricane seasons would be useful for quantifying the level of uncertainty in individual forecasts. These results provide the foundation for real-time confidence guidance to accompany each deterministic intensity forecast, which would increase the value of forecasts without necessarily reducing forecast error.
As a technique to identify the situations associated with good and bad TC intensity forecasts, this study links forecast error to parameters that represent the strength of a TC and the synoptic environment surrounding a TC. The situation-dependent performance of LGEM, DSHP, SHF5, GFDL, and OFCL are evaluated with different performance metrics. SHF5 uses a statistical algorithm based on climatology and persistence and is one of the simplest statistical models; for this reason, it is considered a satisfactory benchmark for mean forecast errors (DeMaria et al. 2007). However, unlike OCD5, SHF5 does not account for intensity decay over land. As a result, SHF5 errors are slightly higher than OCD5, and it is considered a weaker benchmark model. The Navy version of GFDL (GFDN; Bender et al. 2007) is not included in our analysis because both GFDL and GFDN solve the dynamical equations nearly identically and their verification statistics for intensity forecasts are very similar (GFDL produces slightly better results at shorter forecast lengths). DSHP is the better-performing inland decay version of the SHIPS model and therefore replaces it in the analysis. A brief summary of the models assessed and their methodologies are available at the NHC model summary page.

The performance of each model is evaluated by computing the MAE, bias, and SS relative to the SHF5 model for 24-, 48-, and 72-hour forecasts in the Atlantic basin from 2006 to 2010. This study focuses on the shorter forecast periods because these forecasts have not improved at the same rate as the longer ones (Cangialosi and Franklin 2011). Additionally, dynamical parameters are forecasted more accurately at shorter forecast lengths (McNoldy et al. 2012). These performance metrics are binned according to the magnitude of six parameters (“predictors”) and computed for each of the different

---

5Throughout this chapter, the OFCL forecast generated by NHC is occasionally referred to as a “model" forecast for convenience. The author is aware that the OFCL forecast is created by a NHC employee who uses a synthesis of model guidance and their own expertise to produce the forecast.
models. Conventional one-variable histograms (hereafter referred to as histograms) and
two-variable (joint) histograms are created to display the forecast performance metrics
based on the bins.

The data analyzed to calculate model performance is described in section 2.2. Section
2.3 discusses the methods used to process the data and compute statistical significance.
Section 2.4 displays a sample of the histograms and joint histograms created with
emphasis placed on the inclusion of statistically significant results. Section 2.5 provides
the summary and conclusions of chapter 2.

2.2 Data

The 24-, 48-, and 72-hour intensity forecasts for all models are located in the National
Oceanographic and Atmospheric Administration’s (NOAA) Automated Tropical Cyclone
Forecast (ATCF; Sampson and Schrader 2000) “a-deck” files. The forecasts are verified
with two different datasets: the NHC “best track” digital database (Landsea and Franklin
2013) and the 0-hour operational intensity estimates from the aforementioned a-deck. In
the a-deck files, the intensity forecasts archived for the GFDL, DSHP, LGEM, and SHF5
models are recorded to the nearest knot, while the OFCL forecast is recorded to the
nearest five knots. The best track intensities are also provided to the nearest five knots.

Forecasts are verified with the best track and operational intensities of each model,
but this chapter focuses on the results using best track intensities. Comparing forecasts
against the best track data offers a more consistent and accurate verification technique
because the disagreement between operational analyses from different models can reach
45 knots. Also, best track values are a combination of TC data from many diverse sources
(surface observations, ship and buoy reports, aircraft measurements, dropsonde measurements, and satellite observations), yielding a more well-informed intensity estimate (Landsea and Franklin 2013). Summary statistics in section 2.4 will highlight the differences between the verification techniques, but the best track results should be interpreted as more reliable.

The predictor values are available in the stext (SHIPS) files. These files contain intensity forecasts for DSHP, SHIPS, and LGEM as well as all of the data that are used to train these statistical regression models. The predictor data in the SHIPS files are computed using output from the National Centers for Environmental Prediction’s (NCEP) Global Forecast System (GFS). The tested predictors include initial intensity (operational analyses), potential intensity (POT), storm speed, 850-200 hPa wind shear magnitude (hereafter referred to as shear), 850-200 hPa wind shear direction (hereafter referred to as shear direction), and latitude. For each forecast, we use the average of each of these predictors during the forecast period (e.g., for a 24-hour forecast, the average of initial shear and each 6-hourly forecasted shear until 24 hours). These particular predictors are selected for analysis from numerous parameters in the SHIPS files because they are both well observed and important for forecasting TC intensity change (table 3 in DeMaria and Kaplan 1998).

The POT predictor in SHIPS is the difference between the maximum POT (MPI) and the current storm intensity (DeMaria and Kaplan 1994b). The MPI is determined empirically from the sea surface temperature (SST) (DeMaria and Kaplan 1994), while the theoretical MPI of Bister and Emanuel (1998) is a function of SST, the atmospheric

---

6 Files are available at ftp://rammftp.cira.colostate.edu/demaria/SHIPS/stext_oper/.
temperature profile, and the atmospheric moisture profile. As a result, the SHIPS MPI occasionally differs from the theoretical MPI. The shear predictor is the difference between the 850- and 200-hPa wind vectors. From 2007 to 2010, the horizontal wind components used to determine shear were computed from a spatial average (vortex removed) of all the GFS model grid points within 500 km of the 850 hPa storm center at the appropriate height levels. In 2006, the spatial average consisted of grid points between radii of 200 km and 800 km because the vortex removal technique had not been implemented (Knaff et al. 2007). In both scenarios, the shear predictor is particularly adept at capturing the large-scale environment of a TC. The direction of the 850-200 hPa shear vector has units of degrees and follows the convention that the shear is coming from the given heading. For example, a value of 90 degrees means the shear vector is pointing west, 180 degrees is a shear vector pointing north, etc. All of the assessed predictors are available in real time and can serve as useful tools for predicting the error of TC intensity forecasts.

The five-year period between 2006 and 2010 for the Atlantic basin is an exceptional sample for statistical analysis because the evaluated models received no major upgrades during this time period. Only the underlying model, GFS, which provides initial conditions into the GFDL, DSHP, and LGEM, evolved considerably during the five hurricane seasons7. These GFS upgrades were experienced by each model concurrently, thereby keeping the dataset consistent between the models. Also, the models selected for this study were operational for the duration of the 2006 to 2010 Atlantic hurricane seasons (for this reason, HWRF is not included in our analysis). As a result, the sample

7 Documentation for GFS upgrades is available at http://www.emc.ncep.noaa.gov/gmb/STATS/html/model_changes.html
sizes are comparable for the different models and a large number of verified forecasts are available. It is important to note that data for “invests” (low pressure areas monitored by forecasters for possible development) are not included in our analysis. As a result, the performance of the different models is calculated using storms that at least became tropical or subtropical depressions during their life cycles.

However, the model verification presented here contains more cases (forecasts with corresponding 0-hour best track verification) than what is listed from 2006 to 2010 in the NHC verification report (Cangialosi and Franklin 2011). Similar to this study, the NHC excludes invests from their verification statistics. The discrepancy in the number of cases between the two verification results originates from the way the NHC treats weakening storms. If a TC is forecasted to dissipate then the NHC will not include that storm in their OFCL forecast performance metrics. For example, a storm with a current intensity of 60 knots and a 48-hour forecast for a 25-knot low pressure system (“LO” in the best track file) is not included in the NHC verification results even if the verification exists. Additionally, when a hurricane is forecasted to transition into an extratropical storm (“EX” in best track file), NHC excludes these cases (J. Cangialosi 2012, personal communication). Results from both scenarios are included in our analysis, thereby expanding the sample size for each forecast time. Table 2.1 shows the number of forecasts verified for the models at each of the forecast times.

2.3 Methodology

There are twelve synoptic variables used as predictors at each forecast time: six are taken from the initial forecast time and six are derived from averages over the forecast
time period. To evaluate intensity forecasts for each model, histograms are made for individual predictors by selectively binning a predictor and plotting the performance metrics based on those bins. Joint histograms are created by graphing the performance metrics against two predictors. Each square on the joint histograms represents a range of values for each predictor, and the square is shaded a symbolic color to indicate the magnitude of MAE, SS, or bias. Binning is accomplished through two different methods: either dividing the data into three approximately “equal-sized” bins or selecting arbitrary bin ranges to gain insight on whether certain synoptic regimes yield anomalous results. The equal-sized bins are determined by collecting all the predictor values for the different models and splitting them up into thirds based on the values of the predictor. The convention for plotting both single and joint histograms is that each bin includes data that falls on the exact value of its upper limit, while exact values on the lower limit belong to the preceding bin. The only exception is the lowest bin in each figure, which includes values from both bin boundaries.

The 0-hour and time-averaged predictors for each forecast time lead to many possible predictor combinations for data analysis. The twelve different predictors for each of the three forecast times and performance metrics lead to the creation of 108 histograms for each of the four models. The joint histograms lead to fifteen possible outcomes of two predictor combinations (shear paired with intensity, shear paired with storm speed, shear paired with POT, etc.). These combinations can be plotted for both 0-hour and time-averaged predictors for three forecast times and three forecast hours (time-averaged predictors are only paired with time-averaged predictors and 0-hour predictors are only paired with 0-hour predictors). As a result, the total figures required to display the
different permutations are 270 joint histograms for each of the four models. Due to space limitations, we will be showing only a small sample of the figures, focusing on the statistically significant results.

After creating the standard and joint histograms, two-variable $t$ tests (e.g., Wilks 2006) are carried out to establish statistical significance between the different models and synoptic regimes within each model. When determining the significance between different bins in a histogram for a particular model (e.g., the 40 to 70 knot initial intensity compared to the 70 to 100 knot initial intensity for the GFDL model), an unpaired $t$ test is conducted to test if the differences in the means of individual bins are significant. Equation (5.8) from Wilks (2006), adjusted to account for serial correlation between the forecasts (Wilks equation 5.12), is used to determine the Gaussian test statistic $z$, which is converted to a $p$ value. When the $p$ value is less than the significance threshold of 0.05 for the two-sided test, the difference in the means of individual bins is considered statistically significant. In joint histograms, there are smaller bin sizes and additional bins to compare against so it is harder for a bin to achieve statistical significance against all other bins.

For establishing significance between the same bin in different models (e.g., comparing the 40 to 70 knot initial intensity bin for GFDL with the 40 to 70 knot initial intensity bin for LGEM), paired $t$ tests are used. In this case, equation (5.11) from Wilks (2006), again adjusted to account for serial correlation, is used to determine the Gaussian test statistic $z$, which is converted to a $p$ value. The $t$ test is paired in this scenario, because the data values making up the corresponding bins in different models are
observed simultaneously. The two-sided $t$ test $p$ value of 0.05 is again used as the statistical significance threshold.

SS is used to describe the relationship of the forecast accuracy of DSHP, LGEM, GFDL, and OFCL to a benchmark model. Equation 1.1 is used to calculate SS except SHF5 replaces OCD5 as the as the benchmark model in the SS calculation. The SS performance metric is particularly important, because it identifies forecasts that tend to be the most difficult climatologically and where a model substantially improves upon the baseline model. In equation (1), when the MAE of SHF5 is 0 for an individual forecast, the SS calculation results in an undefined quantity. Paired $t$ tests require individual bin entries to be compared, so $t$ tests are not carried out for SS histograms.

### 2.4 Results

In this section, the performance of the intensity forecasts for four operational models during five Atlantic basin hurricane seasons are discussed in detail. Tables 2.2 and 2.3 respectively show the average MAE and bias for the different models at each forecast time. Tables 2.4 and 2.5 display similar information to tables 2.2 and 2.3 except the forecasts are verified with operational intensity estimates for the individual models instead of the best track data. In general, the best track verification results in lower intensity errors for every model and identifies OFCL as the best performing model. DSHP is the worst performing model for 48-hour and 72-hour forecasts while GFDL is the worst at 24 hours. The operational intensity verification results are considerably different; using these data, the GFDL consistently provides the poorest intensity forecasts.
while the LGEM is the best performing model for 48-hour and 72-hour intensity forecasts (OFCL is the best at 24 hours).

Tables 2.3 and 2.5 are generally consistent irrespective of the verification technique, although the best track verification usually results in lower bias values. The tables also indicate that longer forecast lead times are associated with larger positive biases. It appears that the statistical-dynamical models account for model bias with a forecast correction whereas the dynamical model, GFDL, does not and consequently displays the largest mean bias. The discrepancies between the best track and operational intensity tables demonstrate that the verification method employed, and specifically different models’ analysis of 0-hour intensity, can greatly influence the conclusions one makes about the performance of different models. Although the topic of seeking the “best estimate” of intensity is outside the scope of this paper, the effects of using two different verification datasets demonstrate this issue deserves further attention. Subsequent analysis will only focus on the more reliable best track verification results.

Several important observations should be highlighted from the multitude of figures created. First, the verification trends at different forecast lead times occasionally show discrepancies. Certain bin ranges are statistically significant for all models or a particular model at one forecast hour but exhibit dissimilar behavior at another forecast hour. Secondly, more statistically significant bins are present for the time-averaged predictors than 0-hour predictors when verifying long-range forecasts. The more statistically robust correlations with time-averaged parameters are largely because they account for the time variation of the large-scale flow over the longer forecast period. Also, when evaluating the accuracy of a model at a particular lead time, it is important to
review all the different performance metrics as a way to develop understanding for the observed forecast error. When a model records high MAE values for a bin, an extreme positive or negative bias could be responsible. If the difference in the mean bias between this bin and the other bins is statistically significant, then it might be possible to apply a conditional bias correction or upgrade a flaw in the model. If the same bin shows comparatively high SS even with anomalously large MAE, then it is clear that the model is still improving on the benchmark model. Moreover, a high MAE bin might incorrectly imply a model deficiency whereas the high SS value signals the large error is due to an unpredictable synoptic pattern. Finally, adding a second predictor is found to provide useful information about the relationship between the surrounding meteorological conditions and forecast error. A histogram will frequently highlight a bin range as anomalously different from the mean of the forecast lead time, but by incorporating another synoptic predictor, more information is available about the particular combination of parameters that is leading to the most anomalous forecast errors.

2.4.a One-Variable Histograms

Figures 2.1 and 2.2 convey how predictor values can be differently associated with forecast performance depending on the forecast lead time. Figure 2.1 shows the MAE of the 48-hour intensity forecasts for the DSHP, GFDL, LGEM, and OFCL models plotted against 48-hour average POT. The black numbers at the bottom of each histogram entry represent the number of cases in each bin. The histogram indicates that TCs with lower forecasted POT produce lower errors for all models, while the 140 to 160 knot POT bin is the worst performing bin for every model. Additionally, an unpaired $t$ test reveals that both GFDL and OFCL record a statistically significant difference between the 100 to 120
knot bin and every other bin. A paired $t$ test demonstrates the MAE of this bin in GFDL and OFCL is significantly different from the corresponding bin in LGEM and DSHP (but not against each other). On the other hand, the 140 to 160 knot bin is not statistically significant against more than one bin in each of the models.

Figure 2.2 is similar to figure 2.1, but it shows the MAE of 72-hour forecasts plotted against 72-hour average POT. In figure 2.2, the 100 to 120 knot bin also contains the lowest MAE for every model but unlike figure 2.1, this bin is not statistically significant against every other bin in any model. Also, the 140 to 160 knot bin does not record the highest MAE in GFDL and OFCL. Figures 2.1 and 2.2 show that the different lead time error statistics communicate a different message about the effect of POT on model performance. The 48-hour MAE histogram highlights a definitive average POT range where GFDL and OFCL excel, while a similar deduction is not possible for the 72-hour case. Although it is not a statistically robust conclusion, figure 2.1 also highlights a bin where all models perform the worst. The physical justification for these results is not straightforward but the method used to create a DSHP forecast suggests that they should be expected. DSHP uses a different set of regression coefficients for each forecast length (DeMaria and Kaplan 1994). In other words, the same dynamical parameters (storm speed, shear, etc.) are weighted differently based on the forecast period (24-hour, 48-hour, etc.). The results presented here support the practice of time-dependent weighting, and further exploration of the dynamical properties behind these trends could be useful.

Figure 2.3 contains the same predictor and performance metric as figure 2.1 but with equal-sized bins. The figure captures the same general trend as figure 2.1, with MAE rising as POT increases. However, there are fewer bins in figure 2.3, which prevents it
from highlighting that the highest forecast average POT does not actually have the largest MAE. Figure 2.1 shows that the 140 to 160 knot bin has higher MAE than the 160 to 180 knot bin. Since the equal-sized bin graphics frequently reveal trends similar to the manually-selected bins but often miss important details in the performance metrics, the following results will focus on manually-selected bins.

Figures 2.4 and 2.5 provide an example of how using average parameters as predictors often results in more statistically significant differences between bins for long-range forecasts. Figure 2.4 displays the 72-hour MAE histograms for the four models with 0-hour shear as the predictor. There are no bins in the GFDL and DSHP histograms that are statistically significant compared to any of the other bins for each model. In fact, the 30 to 40 knot initial shear bin compared to the 10 to 20 knot initial shear bin in LGEM and OFCL are the only statistically significant bin comparisons. However, in figure 2.5, there are much larger discrepancies between the MAE of different bins. In this figure, the predictor is the 72-hour average forecast shear. For the GFDL model, an unpaired $t$ test shows the anomalously low MAE of the 30 to 40 knot shear bin is statistically significant at the 99% level compared to every other GFDL bin. OFCL also achieves low MAE for 72-hour average shear between 20 and 30 knots; the MAE of this bin along with the highest shear bin is significant against both of the low shear bins for OFCL at the 99% level. Additionally, the 20 to 30 knot shear bin for LGEM and DSHP are statistically significant against the two lower shear bins in each model. Therefore, it appears the mean forecast shear for a TC over the longer forecast period is better linked to MAE than the initial shear. This observation is a common theme throughout all of the predictors.
Figures 2.6, 2.7, and 2.8 illustrate the various analysis methods utilized in this study and emphasize how using multiple performance metrics can provide more conclusive results. These figures show the MAE, bias, and SS histograms for 24-hour forecasts with 0-hour intensity as the independent variable. Although table 2.2 lists OFCL forecasts as the most skillful at 24 hours, figure 2.6 indicates OFCL contains the bin (100 to 130 knots) with the highest MAE out of all the models. LGEM is the best performing model in this bin range, with almost 5 knots less MAE than OFCL. A paired $t$ test reveals the difference in the means of the 100 to 130 knot intensity bin for the two models is statistically significant. Therefore, for 24-hour forecasts of major hurricanes, it appears LGEM guidance was not heavily weighted.

Figure 2.7 provides the bias values for the same predictor and bin ranges, supplementing the MAE results with a possible explanation for the surprising poor performance of OFCL for high intensity TCs. All models have a high positive bias when the 0-hour intensity falls between 100 and 130 knots. The OFCL bin has the largest bias with a mean of 12.4 knots for 95 verified 24-hour forecasts. A $t$ test calculation indicates the difference between the mean bias of this high intensity bin and every other OFCL bin is statistically significant. Additionally, a paired $t$ test is used to compare OFCL’s 100 to 130 knot bin with the same bin in other models; OFCL’s bias for this bin is statistically significant compared to all the other models. In summary, it appears that OFCL forecasts have higher MAE than other models for high intensity storms because forecasts in this bin have an anomalously high positive bias. A possible explanation for this surprising result is that the NHC is employing a “better safe than sorry” protocol for strong hurricanes and consequently, maintaining or intensifying strong TCs in their forecasts.
This technique is popular for hurricanes approaching land because an underestimate in intensity for strong hurricanes could cause civilians and emergency managers to inadequately prepare, resulting in additional fatalities and monetary loss.

Figure 2.8 contains the same bins as figures 2.6 and 2.7 but they are plotted against SS. Clearly, OFCL forecasts achieve the least SS for the high intensity bin, which is also observed to have the highest bias and MAE. However, OFCL still obtains a SS of 16.6% and GFDL, LGEM, and DSHP all have at least one bin that achieves lower SS. The fact that the 100 to 130 knot intensity bin achieves the highest MAE but does not record the lowest SS emphasizes OFCL is struggling in a regime that is inherently uncertain. Nevertheless, all the other analyzed models attain considerably higher SS values for short-range forecasts of strong hurricanes.

2.4.b Joint Histograms

To better define the environment that is lowering the SS for all models in high intensity storms, 0-hour shear is paired with 0-hour intensity to create a joint histogram. Figures 2.9, 2.10, and 2.11 show respectively the MAE, bias, and SS in joint histograms for 24-hour forecasts. The white numbers at the bottom right corner of each box represent the number of cases per bin. It is more difficult to establish statistical significance between the less populated bins in joint histograms so only three bins are prescribed for each independent variable. The reduction in bins requires the 0-hour intensity predictor to use different bin ranges than those seen in figures 2.6 through 2.8; the intensity bin ranges are instead chosen to approximately represent tropical depressions, tropical storms, and hurricanes. Still, figure 2.9 shows a similar trend to previous histograms with
MAE increasing as intensity increases. When initial intensity is greater than 70 knots and the 850-200 hPa shear is between 10 and 20 knots, OFCL MAE is 16.1 knots, almost six knots greater than the average 24-hour OFCL MAE. This high-intensity, medium-shear bin represents the synoptic regime with the largest error in each model, so it is clear TC intensity is difficult to forecast in these situations regardless of the model. A t test demonstrates that the difference in the MAE between the OFCL high-intensity, medium-shear bin and almost every other OFCL bin is statistically significant. The only exception is the intensity bin of greater than 70 knots and a shear between 0 and 10 knots; the difference in the means of these two bins achieves a p value of only 0.09.

It is important to mention that OFCL also has the bin with the lowest MAE in figure 2.9. When the initial intensity of a TC is between 0 and 35 knots and the 0-hour shear is greater than 20 knots, OFCL MAE is only 5.1 knots; the mean of this bin is significantly different from every other OFCL bin. All models appear to have small errors for this bin but only OFCL has a bin that is significant against all of the other bins within the model. Figure 2.10 shows the bias of the different models with the same predictors as figure 2.9. In all of the models, the forecasts for hurricane-strength TCs have a strong positive bias, which agrees well with figure 2.7. The large positive biases are collocated with the largest MAE values in figure 2.9, which suggests the biases could be responsible for the higher MAEs. As expected, the high-intensity, medium-shear bin for the OFCL joint histogram records the highest positive bias.

Figure 2.11 shows SS as the performance metric. All three bins for the 10 to 20 knot shear range for the OFCL forecasts show low SS. This observation is consistent with the previous two figures. The superior SS of the GFDL, LGEM, and DSHP models for
hurricane strength storms is largely attributable to the lower biases seen in figure 2.10. OFCL attains a positive SS (even with the highest MAE) for these stronger TCs because SHF5 is even worse at forecasting in this inherently chaotic regime. From the analysis of figures 2.9 to 2.11, it is apparent that medium shear environments are contributing the most error for forecasts involving high intensity storms. Some hypotheses are presented on why strong storms lead to poor intensity forecasts but it is less clear why medium shear is detrimental to forecast performance. It is possible the shear range between 10 and 20 knots is associated with less reliable 24-hour forecasts because models are struggling with understanding when moderate shear either fosters or stifles TC development. Unlike with very weak or strong shear, models and forecasters do not always agree how a medium-shear environment will affect TC intensification. It is likely that this uncertainty for medium shear is largely due to the current deficiencies of dynamical models, which limit their ability to capture the small-scale interactions that determine how shear exactly interacts with storm structure.

2.4.c Other Notable Results

Determining the background environments that affect the performance of the statistical models is another focal point of this investigation. DSHP and especially LGEM have emerged as two of the best performing intensity models over the last decade (Cangialosi and Franklin 2011). Regardless, LGEM and DSHP typically struggle with 850-200 hPa easterly shear, especially at lower latitudes (M. DeMaria 2012, personal communication). The data analysis conducted in this study is particularly adept at testing such a hypothesis, and we created figures for all three forecast times and performance metrics. Three joint histograms, figures 2.12, 2.13, and 2.14, are presented with latitude
and shear direction as the independent variables. Even though these figures show only 72-hour intensity forecast verification results, shorter forecast lead times confirm LGEM and DSHP have high errors in the low-latitude, easterly-shear regime (not shown).

20°N is selected as the cutoff between low and high latitudes, and only two latitude bins are used (below 20°N and above 20°N) to keep the sample size large for all bins. Shear direction is divided into four bins to capture the effects of the shear vector pointing in all four cardinal directions. Figure 2.12 shows the MAE for the four models and confirms LGEM and DSHP are less skillful for low latitudes and an 850-200 hPa shear vector that is directed west. In fact, all models have their largest MAE when the 72-hour average latitude falls between 0-20° N and the 72-hour average 850-200 hPa shear direction is between 0 and 90° (pointing from the northeast quadrant). The bin with the largest MAE occurs in the LGEM joint histogram; the orange color in the low latitude, northeasterly shear bin represents a MAE of 32.2 knots. This MAE is over 15 knots larger than the average 72-hour LGEM MAE. DSHP also performs poorly for this bin, recording a MAE that exceeds 30 knots. For both of these models, a paired $t$ test shows that the mean of this bin is not larger at a statistically significant level than the corresponding bin in GFDL and OFCL due to the small sample size of the involved bins. However, an unpaired $t$ test reveals that the mean of the low latitude, northeasterly shear bin in LGEM and DSHP is significantly different than all the other bins in the respective models (except for the 90° to 180° shear direction, 0 to 20°N bin in each model and the 0° to 90° shear direction, >20°N bin in DSHP).

Figure 2.13 depicts joint histograms with the same independent variables but bias is used as the performance metric. For the 0 to 90° shear direction with 0 to 20°N latitude
bin, DSHP and LGEM have a positive bias but the mean bias of this bin is not statistically significant against any bin. Therefore, the extremely high MAE in these two bins is not necessarily attributable to bias. Figure 2.14 illustrates that the SS of this bin for LGEM and DSHP is considerably lower than the same bin in other models. LGEM and DSHP obtain a SS of 12.9% and 17.9% while GFDL and OFCL have a SS of 30.5% and 33.2%. Although the positive SS for DSHP and LGEM is surprising, this result is possible because SHF5 performs very poorly in this bin as well.

Figure 2.15 is very similar to figure 2.12 except the number of shear direction bins is reduced so there are only westerly and easterly shear bins. These larger bin results agree with the previous figures that all models are performing poorly at low latitudes with easterly shear. The low-latitude, easterly-shear bin for LGEM obtains the highest MAE out of all the bins. This bin’s MAE is different than all other bins in LGEM at a \( p \) value of 0.08 or lower. Therefore, the results shown are consistent irrespective of the binning of the predictors and therefore present an accurate diagnosis of forecast error dependence on TC latitude and shear direction. Furthermore, the analysis of the figures suggests that forecasters should avoid relying on DSHP and especially LGEM if a TC is in the low-latitude, easterly-shear synoptic regime. These statistical-dynamical models appear to excel overall because their performance in westerly shear and high latitudes compensates for their forecasts in the troublesome environmental conditions.

There are some interesting atmospheric patterns where GFDL performs significantly worse than the other models. Figure 2.16 shows the MAE of 24-hour intensity forecasts with 0-hour storm speed as the independent variable. The >15 knot storm speed bin in the GFDL histogram contains a noticeably higher MAE than any other bin in all models. A
paired t test reveals that this GFDL bin is statistically significant compared to the corresponding bin in every other model. With the exception of the five to ten knot 0-hour storm speed bin (p value of 0.12), the MAE of the other GFDL bins are also different from the highest storm speed bin at a statistically significant level. Surprisingly, the other models show no real trend in MAE compared to the translation speed of the TC. Therefore, we can conclude that 24-hour forecasts by GFDL for storms travelling at greater than 15 knots are not only worse than other GFDL forecasts but are also worse than other models at forecasting these fast-moving storms.

By pairing 0-hour storm speed with 0-hour POT, we gain further insight on the environmental conditions that are leading to the large GFDL errors for fast-moving storms. 0-hour POT is selected as the second predictor instead of initial intensity because it is associated with more statistically significant results. In figure 2.17, both 0-hour storm speed and 0-hour POT are used as independent variables for a MAE joint histogram. The GFDL joint histogram captures two high storm speed bins where the MAE is substantially larger than the average MAE for 24-hour forecasts. The GFDL bin that represents a 0-hour storm speed between 15 to 22.5 knots and a 0-hour POT between 150 to 180 knots has a MAE of 17.7 knots, which is over six knots larger than the GFDL 24-hour forecast average. The 120 to 150 knot POT bin for GFDL also has a very high MAE value, 15.2 knots. Due to the small sample size of the higher POT bin, only the 120 to 150 knot POT bin is different from the corresponding bin in all the other models at a statistically significant level. The other three models do not distinguish a regime in storm speed-POT space that causes anomalously poor forecasts. Figure 2.18 displays bias as a function of the same two independent variables. Neither bin with anomalously high MAE
records large bias values. In figure 2.19, SS is plotted against 0-hour storm speed and 0-hour POT. The GFDL bin corresponding to a 0-hour storm speed between 15 to 22.5 knots and a 0-hour POT between 120 to 150 knots is highlighted as the bin with the lowest SS. In fact, the SS of this bin is -16.3% for 92 verified forecasts. A forecaster could use this information when producing short-range intensity predictions of fast-moving TCs by avoiding GFDL guidance.

2.5 Summary and Conclusions

This study represents one of the first attempts to distinguish if TC intensity model performance is dependent on the synoptic environment. A small sample of figures is presented with accompanying t tests to provide an example of the possible deductions from this innovative binning technique. The statistical significance established between bins conveys that there is robust evidence that forecast error is often related to the surrounding atmospheric environment. Important conclusions are discussed for the overall behavior of all models as well as the environment-based performance of each model. For all models, we frequently observed that the same predictor will lead to varying evaluations of model performance depending on the forecast period. In other words, bins that captured a significant synoptic regime for particular models at a short forecast lead time would not necessarily be as statistically robust at a longer forecast lead time. Also, for long-range forecasts, time-averaged predictors are found to produce more statistically significant results than 0-hour predictors.

Several observations about individual models are also possible using this regime-dependent analysis. Although strong hurricanes lead to poor 24-hour forecasts for every
model, OFCL records the highest MAE. The bias histograms suggest this anomalously poor performance is attributable to a strong positive bias. Pairing initial intensity with initial shear reveals that hurricanes in medium shear are responsible for the high error observed for hurricane-strength storms in the MAE histogram. For 24-hour forecasts with a 0-hour storm speed above 15 knots, GFDL performs significantly worse than other models. The joint histogram with initial storm speed and initial POT as predictors shows that a storm speed between 15 to 22.5 knots and a POT between 120 to 150 knots results in negative SS. A synoptic regime that is not conducive to good forecasts for DSHP and LGEM is also presented. When a TC is at low latitudes in easterly shear, both DSHP and LGEM struggle for all forecast times (only results for 72-hour forecasts are shown). The general trends and the individual inferences about each model that emerge from this situation-based evaluation technique emphasizes the utility of this more detailed validation of TC intensity forecasts.

Although the results provide support for developing real-time confidence guidance for each model’s intensity forecasts, these binned objective statistics have a variety of uses. Firstly, by knowing when models are consistently underperforming or succeeding, forecasters can gain intuition as to what situations produce forecasts that deserve higher or lower confidence. If a TC is approaching land and models are in a high-confidence regime, then emergency managers can focus their evacuations and storm preparations accordingly as a result of the larger reliability of a landfalling prediction. Secondly, a handful of statistical, dynamical, and “hybrid” (a mixture of the two) models have recently been developed but no individual model consistently excels (DeMaria and Gross 2003). If forecast solutions diverge, knowledge of which model is reliable in a given
situation can help NHC forecasters decide which model to favor and consequently produce better verifying forecasts. Finally, if forecasts of forecast errors reveal that certain environmental conditions or dynamical instabilities in the current flow pattern lead to more or less accurate forecasts then further investigation into these regimes is worthwhile. Modelers can focus their efforts into improving a model in less reliable situations and exploring the dynamical mechanisms that cause low-confidence regimes.
Fig. 2.1

MAE of the DSHP, GFDL, LGEM, and OFCL 48-hour intensity forecasts as a function of 48-hour average potential intensity. The numbers at the bottom of each bar represent the amount of cases that are used to calculate the MAE of each bin.
MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average potential intensity.

Fig. 2.2
Fig. 2.3

MAE of the DSHP, GFDL, LGEM, and OFCL 48-hour intensity forecasts as a function of 48-hour average potential intensity. Similar to figure 2.1 except equal-sized bins are used instead of manually-selected bins.
Fig. 2.4

MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 0-hour shear.
Fig. 2.5

MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average shear.
Fig 2.6

MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity.
Bias of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity.
Fig. 2.8

SS of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity.
MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity and 0-hour shear magnitude. The numbers at the bottom of each box represent the amount of cases that are used to calculate the MAE of each bin.
Fig. 2.10

Bias of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity and 0-hour shear magnitude.
Fig. 2.11

SS of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour intensity and 0-hour shear magnitude.
Fig 2.12

MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction.
Fig. 2.13

Bias of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction.
Fig. 2.14

SS of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction.
Fig. 2.15

MAE of the DSHP, GFDL, LGEM, and OFCL 72-hour intensity forecasts as a function of 72-hour average latitude and 72-hour average shear direction. Similar to figure 2.12 except with only two shear direction bins.
Fig. 2.16

MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed.
Fig. 2.17

MAE of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed and 0-hour potential intensity.
Fig. 2.18

Bias of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed and 0-hour potential intensity.
Fig. 2.19

SS of the DSHP, GFDL, LGEM, and OFCL 24-hour intensity forecasts as a function of 0-hour storm speed and 0-hour potential intensity.
<table>
<thead>
<tr>
<th>Forecast Period (hr)</th>
<th>24</th>
<th>48</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSHP</td>
<td>1322</td>
<td>1162</td>
<td>1012</td>
</tr>
<tr>
<td>GFDL</td>
<td>1243</td>
<td>1070</td>
<td>910</td>
</tr>
<tr>
<td>LGEM</td>
<td>1320</td>
<td>1158</td>
<td>1008</td>
</tr>
<tr>
<td>OFCL</td>
<td>1259</td>
<td>1090</td>
<td>937</td>
</tr>
</tbody>
</table>

Table 2.1

Number of verified forecasts (best track verification), for DSHP, GFDL, LGEM, and OFCL at each of the forecast periods. These cases’ totals are for Atlantic basin storms between 2006 and 2010.
<table>
<thead>
<tr>
<th>Forecast Period (hr)</th>
<th>24</th>
<th>48</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSHP</td>
<td>11.06</td>
<td>15.48</td>
<td>18.32</td>
</tr>
<tr>
<td>GFDL</td>
<td>11.85</td>
<td>15.16</td>
<td>17.62</td>
</tr>
<tr>
<td>LGEM</td>
<td>10.90</td>
<td>14.60</td>
<td>16.38</td>
</tr>
<tr>
<td>OFCL</td>
<td>10.20</td>
<td>13.90</td>
<td>15.72</td>
</tr>
</tbody>
</table>

Table 2.2

MAE in knots, using best track verification, for DSHP, GFDL, LGEM, and OFCL at each forecast period.
Table 2.3

Bias in knots, using best track verification, for DSHP, GFDL, LGEM, and OFCL at each forecast period.

<table>
<thead>
<tr>
<th>Forecast Period (hr)</th>
<th>24</th>
<th>48</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSHP</td>
<td>0.83</td>
<td>1.74</td>
<td>1.49</td>
</tr>
<tr>
<td>GFDL</td>
<td>1.86</td>
<td>4.85</td>
<td>7.68</td>
</tr>
<tr>
<td>LGEM</td>
<td>-0.87</td>
<td>-0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>OFCL</td>
<td>1.51</td>
<td>2.50</td>
<td>3.76</td>
</tr>
<tr>
<td>Forecast Period (hr)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>48</td>
<td>72</td>
</tr>
<tr>
<td>DSHP</td>
<td>10.85</td>
<td>16.20</td>
<td>19.08</td>
</tr>
<tr>
<td>GFDL</td>
<td>11.62</td>
<td>16.61</td>
<td>19.16</td>
</tr>
<tr>
<td>LGEM</td>
<td>10.70</td>
<td>15.19</td>
<td>17.25</td>
</tr>
<tr>
<td>OFCL</td>
<td>10.53</td>
<td>15.40</td>
<td>17.43</td>
</tr>
</tbody>
</table>

Table 2.4

MAE in knots, using 0-hour operational analyses, for DSHP, GFDL, LGEM, and OFCL at each forecast period.
<table>
<thead>
<tr>
<th>Forecast Period (hr)</th>
<th>24</th>
<th>48</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSHP</td>
<td>1.51</td>
<td>3.09</td>
<td>3.24</td>
</tr>
<tr>
<td>GFDL</td>
<td>4.22</td>
<td>7.33</td>
<td>9.85</td>
</tr>
<tr>
<td>LGEM</td>
<td>-0.58</td>
<td>-0.20</td>
<td>1.01</td>
</tr>
<tr>
<td>OFCL</td>
<td>1.99</td>
<td>3.14</td>
<td>3.69</td>
</tr>
</tbody>
</table>

Table 2.5

Bias in knots, using 0-hour operational analyses, for DSHP, GFDL, LGEM, and OFCL at each forecast period.
Chapter 3. Prediction of Intensity Model Error (PRIME)

3.1 Background and Motivation

Forecasters and end users of TC intensity forecasts would greatly benefit from a reliable expectation of model error to counteract the lack of consistency in forecast performance. The observed day-to-day, model-to-model, and storm-to-storm fluctuations in forecast quality prevent individuals from gauging how much certainty to place in a single forecast. As a result, climatological forecasts and forecaster intuition are used to estimate the skill of TC intensity forecasts. Although these techniques are a reasonable benchmark, they cannot reliably indicate whether a particular model forecast will perform better than average or if the atmospheric regime is inherently more difficult for all models to forecast. This chapter will show that using resources currently available, statistical techniques can be utilized to create real-time forecasts of forecast error and provide a new objective tool to add value to TC intensity forecasts.

As discussed in chapter 2, the binning of model performance based on TC strength and environmental conditions highlights certain situations that are conducive to better or worse TC intensity forecasts. Multiple idealized numerical modelling studies (Tao and Zhang 2014; Zhang and Tao 2013; Zhang et al. 2014) have also linked synoptic-scale conditions and storm properties to forecast error. Although BN13 first recommended that these verification trends supplied a foundation for skillful error predictions of TC

---

8 Note that here the term "climatological forecasts" refers to the expectations of model intensity errors based on intensity forecast errors averaged over several years.
intensity forecasts, the statement that “no forecast is complete without a forecast of forecast skill” is almost 30 years old (Tennekes et al. 1987). In the hurricane research community, Aberson (1997) conducted the first investigation into using statistical analysis to forecast TC track forecast error. In that study, parameters related to the synoptic situation, atmospheric flow stability of the environment surrounding a TC, and recent model performance were selected as predictors for error forecasts of the barotropic hurricane track forecast model, VICBAR (Vic Ooyama barotropic model). The multiple linear regression analyses only showed modest results, but the linear discriminant analyses were effective at predicting VICBAR model performance. For the small independent data sample evaluated in that study, a statistical scheme was able to distinguish between good and bad forecasts.

More recently, Goerss (2007) and Goerss and Sampson (2014) used stepwise multiple linear regression to predict the error of a consensus track model (CONU) and the AE of two consensus intensity models (IVCN and S5YY). The predictors used for the forecasts of consensus intensity and track forecast error were obtained solely from the real-time ATCF files. For the prediction of CONU track error, Goerss (2007) showed the regression analysis was able to explain a large portion of the variance of independent data, ranging from 23% to 46% for the 2003 Atlantic hurricane season. However, when predicting IVCN AE for the 2012 Atlantic hurricane season, only 2% to 5% of the variance of the independent data was explained (Goerss and Sampson 2014). The modest skill in the intensity AE predictions suggests either the ATCF predictors were not identifying environments with significant error trends, or that by averaging the different
model forecasts to create a consensus, useful information on the error relationships unique to the individual models was lost.

In this chapter, we build on previous results of BN13 by testing the ability of the PRIME model to forecast the AE and bias of LGEM, DSHP, GHMI, and HWFI. PRIME forecasts are evaluated at each 12-hour interval from 12 to 120 hours during the 2007-2014 Atlantic hurricane seasons. A second, retrospective version of PRIME is validated for the 2008-2014 Atlantic hurricane seasons. This version of PRIME is developed from reforecasts of select storms during the 2008-2013 seasons using the 2014 version of each model (obtained from V. Tallapragada and M. DeMaria 2014, personal communication). The AE and bias of both versions of PRIME are then compared to their respective climatologies to determine their skill. In order to simulate PRIME performance in an operational setting, the results focus on independent verification calculated using cross validation (Wilks 2011).

The procedure used to create the SHIPS model (DeMaria and Kaplan 1994) is generally followed for the development of PRIME. Dynamical quantities as well as parameters capturing the quality of the initial analysis and intrinsic instability of the synoptic situation are inputted as independent variables in a stepwise multiple linear regression formula. However, the standard regression techniques were slightly modified to account for the properties of the predictands. A power transformation is applied to AE before it is forecasted with PRIME because AE has a lower bound of 0 and positive skewness. The details of the AE transformation are described in further detail in section 3.3. Additionally, to account for nonlinear relationships, additional predictors are created by empirically fitting individual independent variables to each dependent variable with
four different functions: Gaussian, two-peak Gaussian, second order polynomial, and third order polynomial. Each of these new predictors is tested for improvement to PRIME.

The data used to develop PRIME are described in section 3.2. In section 3.3, the regression model is discussed, including the selection of predictors, required modifications to standard multiple linear regression, and weighting of predictors. Section 3.4 provides the results of the independent data testing of PRIME. Potential applications of PRIME and conclusions are presented in section 3.5.

3.2 Data

This chapter and the next two chapters focus on the models that are essential to the guidance of Atlantic basin OFCL intensity forecasts. In contrast to BN13, this chapter only includes early models that are available by the time the OFCL forecast is released. Two other early versions of the GFDL hurricane model, GFDI and GFNI (Rennick 1999) were omitted from the analysis because they perform slightly worse than the GHMI, display nearly identical error trends, and have smaller sample sizes. As mentioned in section 1.2, DSHP, LGEM, GHMI, and HWFI were all introduced as TC intensity guidance in the last fifteen years and the latter three models surfaced in either 2006 or 2007. As a result, NHC forecasters have received considerably more resources during the years considered in this study. Revisiting the idealized best model scenario of section 1.2, it is important to emphasize the best model available as guidance for the Atlantic basin OFCL intensity forecasts significantly improved from 2007-2014. Figure 3.1 shows the AE of this scenario (same data used to create the solid lines in figure 1.4) at every forecast hour with the corresponding best-fit lines. Using the Student’s $t$ test, the slope of
the regression line for each forecast hour was found to be significantly different from 0 at
the 95% level (except 120 hours which was significant at the 90% level). The improved
guidance suggests that anticipating the best-performing model at the time of the forecast
could substantially reduce OFCL errors.

The intensity forecasts, verification, and predictors needed to develop PRIME are
from the same sources as chapter 2. Storms were selected using the same verification
rules followed by NHC when it determines OFCL intensity errors. A PRIME forecast of
bias and AE was only computed for a particular forecast time if the TC was at least a
tropical or subtropical depression initially and at verification, all four models had an
intensity forecast available in the real-time ATCF files, and a verifying TC intensity was
available in the best-track dataset. These requirements resulted in a dataset that was
completely homogeneous among the models. There was only a small discrepancy in the
sample sizes for PRIME compared to OFCL verification because PRIME satisfied the
additional requirement that all the predictors and intensity forecasts of the four models
existed. Storms over the ocean and land were tested independently and together, and the
results are presented in section 3.4.

Table 1.1 contains all of the predictors, their abbreviated names, and whether the
predictor changed with each model. In chapter 2, the tested predictors included initial
intensity (0INT), potential intensity (POT), storm speed (SSPD), 850-200 hPa shear
(SHR), 850-200 hPa shear direction, and latitude (LAT). All of these predictors were also
included in the development of PRIME, but shear direction was replaced by sin(shear
direction) (SHRDIR). This update mathematically resolves the discontinuity between a
shear direction of 359 degrees and 1 degree. Several dynamical quantities available in the
SHIPS and ATCF files were added to the predictor pool: forecast intensity (FINT), longitude (LON), 850 hPa environmental (0-1000 km radial average) vorticity (VOR), 200 hPa environmental (0-1000 km radial average) divergence (DIV), 850-700 hPa environmental (200-800 km radial average) relative humidity (RH), ocean heat content (OHC), standard deviation of GOES channel 4 (infrared) brightness temperature (50-200 km radial average) (GBRT), and percent area of GOES channel 4 brightness temperature below -20°C (cold pixels, 50-200 km radial average) (GCLD). A thorough description of the data used to calculate both the GOES predictors and the OHC predictor is available in DeMaria et al. (2005). As in chapter 2, both the initial value and the average of each of these predictors during the forecast period were used as predictors. Forecast OHC was occasionally missing in the SHIPS files and in these situations, the forecast average OHC was calculated using the available forecast times. Additionally, there were a few instances when forecasted latitude was not included in the SHIPS files but the rest of the predictors were available. The forecasted latitude of the individual model was utilized in this scenario.

To supplement the synoptic predictors, proxies for atmospheric flow stability and initial condition error were also computed. The forecasted intensity change during the forecast period (FCIC), deviation of each model’s intensity forecast from the mean of the models (DFEM), standard deviation of the intensity forecasts (SPRD), deviation of each model’s track forecast from the mean of the models (DTRK), and initial (LDIS) and forecasted distance to land (FLND) were identified as potential indicators of the uncertainty in the synoptic regime. The previous 12-hour intensity change (P12C) and the error of the most recently verified 12-hour forecast (P12E) were the only proxies used
to represent initial condition error. If one of the models was not run at the forecast time 12 hours before, then the average of P12E and P12C from the available models filled in the missing predictors. The absolute and signed (positive and negative) values of FCIC, DFEM, P12C, and P12E were all assessed as predictors for both bias and AE. As expected, the absolute values of FCIC and DFEM had higher correlations with AE than the signed versions of the variables, and the signed versions of FCIC and DFEM had higher correlations with bias. Conversely, P12C and P12E had higher correlations with both AE and bias as signed predictors. With the additions of these proxies, a total of 33 predictors were used in the stepwise multiple linear regression formula for both AE and bias.

These proxies were integral to the success of PRIME, especially for the version of PRIME using real-time models, where DFEM was the leading predictor of bias forecasts for all models and forecast intervals. As an example, figure 3.2 shows 120-hour HWFI intensity forecast bias for the 2007-2014 seasons plotted against DFEM. The high positive correlation of HWFI bias and DFEM ($R$ value of 0.71) highlights how closely bias is coupled with this predictor. Absolute DFEM (ADEM) was one of the leading predictors of AE but its performance was rivaled by FINT, average latitude (ALAT), and SPRD. Examples of the correlations for the four leading predictors of AE are shown in figure 3.3. 72-hour GHMI AE is plotted against ADEM (top left), SPRD (top right), ALAT (bottom left), and FINT (bottom right). Even though these predictors have the largest correlations with AE for this forecast interval and model, the $R$ values are much lower than the previous bias case. Section 3.4 further demonstrates the importance of
proxies by comparing the accuracy of PRIME error predictions that include and exclude them.

As previously mentioned, the eight-year period between 2007 and 2014 for the Atlantic basin is an important sample for statistical analysis because it represents the longest period that the evaluated models are simultaneously operational. However, it is important to mention all of the models as well as the underlying model, GFS, evolved considerably during the eight hurricane seasons. The upgrades to the physics and grid spacing of the dynamical models as well as the added predictors to the statistical models have recently lowered intensity forecast errors. Such model inconsistencies within the data sample are not ideal for statistical analysis. To produce a more homogeneous dataset, the length of the training period of PRIME was varied so the models were more consistent over the sample. In general, these adjustments did not benefit PRIME but the climatological forecasts of AE and bias improved with the shorter training period.

A second, more helpful approach to account for the changes to the models over the time series is using retrospective runs of each model to train PRIME. The 2007 hurricane season retrospective forecasts were not available for any of the models, and the 2009 and 2013 seasons were missing for GHMI. Additionally, there were about 15%-40% less retrospective cases than real-time cases for 2008-2012, because the retrospective HWFI forecasts were only generated when tail-wind Doppler radar data were available.

To partially augment the smaller sample size of the retrospective runs, PRIME was also developed with a combination of real-time and retrospective runs. For this version of PRIME, the real-time GHMI forecasts from 2009 and 2013 were added to the sample. The other three models produced retrospective forecasts for these years so the addition of
these GHMI cases adds several cases to the retrospective sample for all models\(^9\) without noticeably contaminating the integrity of the statistical analysis. PRIME was then trained using the same methodology followed for the real-time model output and compared to the climatology of the retrospective models. The version of PRIME that included the 2009 and 2013 real-time GHMI data supplementing the retrospective data outperformed the version with solely retrospective runs. Moving forward, this augmented version of PRIME will be referred to as R-PRIME\(^{10}\). Tables 3.1 and 3.2 demonstrate the discrepancy in performance of the models used to develop R-PRIME and PRIME. Table 3.1 lists the AE, bias, and sample size for the real-time intensity forecasts while table 3.2 lists similar quantities for the retrospective intensity forecasts.

### 3.3 Regression Analysis

The goal of this study is to skillfully predict the AE and bias of the intensity forecasts of four operational TC models using information specific to the TC and surrounding environment. PRIME was developed with a stepwise multiple linear regression scheme similar to the SHIPS model (DeMaria and Kaplan 1994). The prediction equation associated with multiple linear regression is:

\[
y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k
\]  

(2)

where \(y\) is either bias or AE, \(k\) is the number of predictors, \(x_k\) is the set of predictors, and \(b_k\) is the set of weighting coefficients. The regression procedure chose the weighting coefficients that minimized the sum of squares of the residuals for forecasts of bias or AE given the observed predictor values. Separate regressions were performed for each

\(^9\) As mentioned previously, PRIME only produced an error forecast when all models had intensity forecasts available.

\(^{10}\) The version of PRIME using the real-time HWFI, GHMI, LGEM, and DSHP intensity forecasts is simply referred to as PRIME.
forecast interval, predictand, and model because BN13 demonstrated that the relationships between error (bias and AE) and dynamical variables can change depending on these conditions. The regression equation started with all of the predictors and then the least significant predictor was removed. This backward-stepping procedure continued until the weighting coefficients associated with the remaining predictors were all significant. A predictor was deemed significant if a standard $F$ statistic test found there was a 95% likelihood or higher that its regression coefficient was different from 0.

To simplify the operational implementation of PRIME and allow users to compare the importance of predictors at different forecast hours, the same set of predictors was used at all forecast intervals for each model. A predictor was included in this final pool for a model as long as it was significant for at least one forecast interval. However, preliminary results suggest that PRIME errors can be lowered using different sets of predictors at each forecast interval. The benefits of changing the predictors based on forecast length will be further explored in a future study.

Multiple linear regression is an appropriate statistical model for PRIME because the data satisfies a majority of the key assumptions of the technique and the resulting regression coefficients can be easily interpreted. However, the distribution of the AE predictand for a fixed time interval is not normal and causes errors that are heteroscedastic. As a result, a power transformation is necessary to transform the positively skewed AE distribution to an approximately Gaussian distribution, creating more homoscedastic data for the linear regressions (Wilks 2011). Specifically, the approach outlined in Box and Cox (1964) can be followed to transform the original AE values with the function:
\[ T(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases} \]  

(3)

Here, \( \lambda \) is the transformation parameter and \( x \) represents the original AE values. The best \( \lambda \) for (3) will maximize the log-likelihood function for the Gaussian distribution:

\[ L(\lambda) = \frac{n}{2} \ln[s^2(\lambda)] + (\lambda - 1) \sum_{i=1}^{n} \ln(x_i) \]  

(4)

In this equation, \( n \) is the sample size, \( s^2(\lambda) \) is the sample variance of the data after the transformation, and \( \lambda \) and \( x_i \) are the same variables included in (3). Different values of \( \lambda \) are inserted into (4) and the one yielding the largest \( L(\lambda) \) is considered the most appropriate. When \( x_i \) equals zero, the natural log term causes (4) to become undefined. To deal with this discontinuity, a positive constant (2 knots) was added to each \( x_i \) so that all the data was shifted to the positive half of the number line. This shift was then offset after producing the AE predictions.

Each power transformation only uses the training dataset and does not involve the sample that validates PRIME. Figure 3.4 is included as an example of an AE distribution before and after a Box-Cox transformation is implemented. The top plot in figure 3.4 shows the AE of the 72-hour GHMI real-time forecasts for the PRIME training data of the 2014 season (2007-2013), and the bottom plot shows the transformed AE for that forecast interval and model using a \( \lambda \) of 0.22. Both qualitative and quantitative comparisons of the two distributions illustrate that the Box-Cox transformation helps create a more normally distributed predictand.

In an ideal multiple linear regression scheme, there should also be approximately linear relationships between the dependent variables and independent variable. To account for some predictors exhibiting nonlinear relationships with error, Gaussian, two-
peak Gaussian, second-order polynomial, and third-order polynomial functions were applied to predictors. For the predictors with the observed nonlinear patterns with bias and AE, each of these functions was empirically fit using the training data. Due to the requirement that the fitting parameters remain constant for the different forecast hours, the functions were fit using data from all forecast hours. The transformed predictor then replaced the original predictor and was tested using independent verification for improvement to PRIME.

Although nonlinear functions were applied to several independent variables for both bias and AE, only LDIS significantly improved independent verification results over all forecast intervals and models for a predictand (bias) and version of PRIME (both). Negative values of LDIS indicate the storm is currently over land and high positive values indicate the storm is far from land; these scenarios are typically associated with high forecast certainty along with small biases for all models. However, small positive values of LDIS characterize storms that are expected to dissipate over land but often continue over the ocean and maintain their strength. Models have high positive intensity bias in these uncertain situations. Figure 3.5 demonstrates how the LDIS predictor is empirically fit using a second-order polynomial to produce a more linear relationship with bias. The top image of figure 3.5 shows the curvilinear relationship between LDIS and 108-hour retrospective DSHP bias for the 2008-2013 hurricane seasons with the corresponding linear correlation ($R$) value of the linear fit. The red dashed line is the second-order polynomial that best fits all forecast intervals and whose coefficients were used to transform the LDIS values for the 2014 R-PRIME forecasts of DSHP bias. The bottom image of figure 3.5 shows fitted LDIS plotted against 108-hour DSHP bias with
the increased $R$ value. With the inclusion of the fitted LDIS predictor, the errors in the bias predictions of R-PRIME were decreased by 0%-6% (with similar percentages for PRIME). Thus, the final version of the R-PRIME and PRIME bias forecasts contained the original 32 predictors and one fitted predictor.

The predictor with the clearest nonlinear relationship with AE was RH (both average and initial). Intermediate values of RH (45%-75%) were associated with higher AE than the large or small values but the restriction on maintaining the same predictors irrespective of forecast hour was particularly detrimental to fitting a Gaussian or second-order polynomial. The exact medium RH values with the peak AE varied depending on the forecast hour and as a result, the empirical fits detracted from certain forecast hours enough where the average improvement over all forecast intervals was not significant for a majority of the models. Nevertheless, there was evidence that AE peaks for a range of mid-level RH values.

When R-PRIME and PRIME were verified with the developmental sample (dependent verification), it unsurprisingly performed best with the entire set of 33 predictors. Tables 3.3 and 3.4 show the percent of the total variance explained ($R^2$) by PRIME for dependent verification of AE and bias forecasts for each model and forecast hour. These $R^2$ values are similar to those achieved with R-PRIME (not shown). However, to minimize the error for the independent verification, it is important not to overfit the regression on the developmental sample with too many independent variables (Neumann et al. 1977). The optimal number of independent variables for the training of each version of PRIME was determined by first calculating the correlations of every variable with each predictand over all the forecast hours in the training dataset. The
predictors were then ranked based on the magnitude of the correlation and a smaller group of the best predictors were used in the regression formulas. Depending on the model and predictand, the optimal number of predictors varied between 2 and 8.

Once the optimal predictors were specified to PRIME for a model, it was possible to gain some physical insight on how predictors are related to forecast error. The dependent and independent variables were normalized before each regression to allow comparisons between the regression coefficients of different variables and forecast intervals. Tables 3.5 and 3.6 list the three most significant predictors and their weighting coefficients for AE and bias respectively. For each model, the predictors are ordered (bottom is most significant) based on the average magnitude of their weighting coefficients over all forecast intervals. Using the 2008-2014 hurricane seasons as the training period, these weighting coefficients are derived from the optimal predictors of R-PRIME. PRIME highlights similar predictors and yields comparable weighting coefficients but to avoid the possibility of real-time model upgrades influencing the analysis of the weighting coefficients, these tables focus on the retrospective results. Only three predictors are included for each model because the signs of the coefficients of these predictors are consistent with physical reasoning and significant at more forecast intervals than all the other predictors.

Table 3.5 shows similar AE predictors are found to be significant amongst the different models; ADEM, FINT, 0INT, ALAT, and SPRD are the only five predictors represented. ADEM is listed for all the models, and the sign of its weighting coefficients are intuitive. When a model’s intensity forecast deviates from the ensemble mean, PRIME unsurprisingly expects a larger AE for its intensity forecasts at all forecast
intervals. SPRD is positively correlated with AE because of the well-documented relationship between spread and skill (Whitaker and Loughe 1998). ALAT is listed as one of the three most important predictors for LGEM, GHMI, and DSHP. These models show negative regression coefficients for ALAT at all forecast intervals, largely because a majority of the rapid intensification events, which are associated with high AE, occur at low latitudes. Additionally, BN13 showed there are more easterly shear cases at lower latitudes, which are associated with higher AE. For HWFI and GHMI, FINT is listed as an important predictor. The high positive weighting coefficients for FINT can be explained by the models incorrectly predicting rapid intensification and eyewall replacement cycles; both of these processes are poorly resolved by the operational models. For similar reasons, weaker 0INT is associated with higher AE for HWFI. Storms further away from their maximum possible intensity have more potential for intensification and thus, more room for error.

Table 3.6 contains the weighting coefficients of the three most important predictors of bias. GHMI and LGEM have the same three predictors listed: FINT, 0INT, and DFEM. FINT has large negative weighting coefficients for both models, especially at shorter forecast lead times. In other words, the models underestimate the magnitude of intensification for storms expected to be strong and decay for storms expected to be weak. Meanwhile, the positive weighting coefficients for 0INT likely signal that models are maintaining strong storms at high intensities for too long and not anticipating the intensification of weak storms. A similar pattern was observed in chapter 2. The strongest predictor for all the models is the DFEM predictor. DFEM is positively correlated with bias because a model’s deviations from the ensemble mean typically have the same sign.
as its forecast bias. The fitted LDIS predictor coefficients are not inherently intuitive after the predictor transformation but an inspection of figure 3.5 can explain their positive sign. The second-order polynomial function helps to shift positive bias events to larger x values and negative bias events to smaller x values. HWFI and DSHP have two unique predictors that are significant: ALAT and forecast average divergence (ADIV). The high positive correlations of HWFI bias with ALAT could be related to higher latitudes having more land and as a result, more storms that interact with land unexpectedly. Also, lower latitudes experience more RI cases where models considerably underforecast intensity change, hence the negative biases. The sign of the ADIV predictor does not have a straightforward physical explanation but PRIME signals that DSHP either overstates the intensification of storms that have divergence aloft or underestimates the weakening of storms with convergence aloft.

3.4 Independent Testing of PRIME

In this section, we present the independent verification statistics of PRIME and R-PRIME bias and AE forecasts. To infer the future operational performance of PRIME, cross validation was applied to the bias and AE predictions for both versions of PRIME. For the cross validation, all but one of the years were used as the training data, and then the excluded year was used for validation; this procedure was repeated for all years. Unless specified, the same 33 predictors were included in the preliminary predictor pool for AE, and fitted LDIS replaced the original LDIS predictor for bias. The methodology discussed in section 3.3 was followed to determine the ideal number of predictors for each model and predictand; consequently, the number of predictors involved in each iteration of PRIME often varied. Before PRIME was evaluated, a final adjustment was
applied to AE and bias predictions. When the PRIME regression formula for a particular model-forecast interval combination yielded a negative value for an AE prediction, the output was rounded up to 0 knots. The reasoning behind the modification is nonnegative AE values are not allowed mathematically. Secondly, PRIME bias predictions that implied forecast intensities below 20 knots were changed to prevent forecast intensities dropping below this threshold. The bias adjustment was necessary because storms below 20 knots were not included in the verification.

3.4.a Separate Regression Models for Bias and AE

Separate regression models were developed for forecasts of AE and bias because of their inherently different behavior. Section 3.3 discussed the variables that were more adept at forecasting either bias or AE. As a result, the post-processing step of computing the absolute value of the PRIME bias (|bias|) forecasts does not result in accurate forecasts of AE. Two idealized sets of data are presented in figures 3.6 and 3.7 to justify the development of distinct versions of PRIME for each predictand. In figure 3.6, ten data points are plotted for predictor values against bias (top) and AE (bottom). In both panels, the $R$ of the data is included in the bottom right corner. The predictor values show almost no linear trend with bias but when the same predictor values are plotted against the absolute value of bias (AE), there is a strong linear relationship. This predictor would be useful for forecasting AE but not bias. The uncorrelated nature of AE and bias is further emphasized by calculating their $R$ value; it is only 0.06.

The opposite case is considered in figure 3.7. Ten different data points are plotted for a fictitious predictor against bias and AE, and the $R$ values are located in the bottom left corner of each image. For this situation, bias exhibits a strong linear trend with the
predictor while AE does not. Although both figures represent fabricated data, they highlight the different possible relationships between the predictors and predictands. There are also storms and hurricane seasons where bias generally maintains the same sign (i.e. all the data points in figure 3.6 would have positive bias). For these scenarios, the predictor values would have \( R \) values of similar magnitudes with both AE and bias.

In practice though, bias and AE are weakly correlated, and thus, PRIME forecasts of AE vary considerably from PRIME forecasts of \(|bias|\). Figure 3.8 provides an example of a forecast hour, model, and year where the PRIME forecasts of AE and \(|bias|\) disagree. In figure 3.8, the x-axis contains PRIME forecast values of AE for 72-hour LGEM forecasts from 2014, and the y-axis contains the corresponding PRIME \(|bias|\) forecast values. The \( R \) value for this sample is only 0.1, indicating that the PRIME equation for predicting bias and AE is very different.

To determine if this result is representative of a larger sample of forecasts, PRIME AE forecasts were compared to PRIME \(|bias|\) forecasts for the sample listed in table 3.1. Rather than verify PRIME forecasted bias against the true bias, \(|bias|\) was compared to true AE. Figure 3.9 shows the average AE of PRIME AE predictions, PRIME \(|bias|\) predictions, and climatological AE predictions. At all forecast hours, PRIME AE forecasts perform better than PRIME \(|bias|\) forecasts. Additionally, the climatological AE forecasts have lower average errors than PRIME \(|bias|\) forecasts at almost all forecast hours. Therefore, figures 3.6-3.9 demonstrate that taking the absolute value of PRIME bias forecasts is not appropriate for forecasting AE.
3.4.b Average Statistics of PRIME with and without Proxies

As mentioned, regression models perform best when there are a limited yet effective set of independent variables. Even though the correlations between the proxies and error are high in past seasons, it is worth exploring if the dynamical predictors listed in Table 1.1 are sufficient for producing skillful PRIME error predictions. Using the dataset listed in Table 3.1, PRIME was developed with and without proxies in the predictor pool to evaluate their importance. Figure 3.10 illustrates the AE of PRIME AE predictions for the two predictor pools. The solid and dashed black lines respectively represent PRIME with and without proxies. The solid red line corresponds to error predictions using climatology. For all forecast hours and models, PRIME outperforms the version of PRIME with no proxies. Figure 3.11 is similar to Figure 3.10 but applies to PRIME bias predictions. Clearly, the proxies improve PRIME more for the bias forecasts. In fact, PRIME bias predictions with the proxies are significantly better than the no proxy PRIME for all forecast hours and models. Even though the improvement in the AE forecasts is less impressive, all of the PRIME forecasts for the rest of the chapter will include proxies.

3.4.c Average Statistics of PRIME for Land and No Land Cases

As a way to test if PRIME error statistics are significantly different when TCs interact with land, PRIME was independently developed for “land” and “no land” cases. Intensity forecasts are classified as land cases if the TC is forecasted to traverse land in the period leading up to forecast verification (including verification time), over land when the forecast is made, or reported by best track verification as over land in the period leading up to forecast verification. Table 3.7 lists the percent of the total variance of the true AE
values explained by PRIME AE forecasts for land and no land cases. The first column for each model represents the land $R^2$ values and the second column represents the no land $R^2$ values. The bold numbers are used to indicate which column has the higher $R^2$ value for the particular forecast interval. These values were calculated using independent verification for the sample in table 3.1. Approximately 90% of the total forecasts are composed of the no land cases. For all models excluding GHMI, PRIME performs considerably better for the land cases, largely because the average AE and AE variance is larger when storms interact with land. Table 3.8 is similar to table 3.7 but the $R^2$ values apply to bias forecasts. The PRIME bias forecasts of all models generally perform better for land cases but the PRIME bias forecasts of the statistical models show the largest improvement. PRIME AE and bias forecasts also record higher SSs compared to climatology for storms interacting with land (not shown).

Unfortunately, best track information is only available after a forecast is verified so this technique to differentiate land and no land cases would not be possible in real time. Using information available to forecasters in real-time, land cases can also be defined as situations where the models forecast the TC to cross land or are currently over land. In this scenario, PRIME showed a minor improvement compared to no land cases but the differences between the two samples were rarely significant. Therefore, the land and no land cases are combined for the rest of this chapter to increase sample sizes, maintain PRIME as a real-time tool, and improve the performance of PRIME.

3.4.d Comparison of Average Statistics for PRIME and R-PRIME

To compare the two versions of PRIME to each other as well as climatology, the sample size of PRIME was limited to only include the cases available for R-PRIME.
Therefore, average statistics were computed using substantially less cases than those listed in table 3.1. Before presenting the results, it is important to identify the strengths of each version of PRIME. R-PRIME has the advantage that the formulation of the models is unchanged between the training and verification period, which translates into more consistent statistical relationships between the predictors and predictand. However, the smaller sample size of retrospective data might not be as representative of the broader spectrum of TCs that occur across many Atlantic basin hurricane seasons. In particular, the retrospective forecast sample contained less intense hurricanes and fewer storms that interacted with land, which were two situations where PRIME excelled in comparison to climatological forecasts. A more sizeable retrospective dataset would undoubtedly improve the capabilities of R-PRIME. Another potential consequence of underrepresenting certain storms is the skill of the forecasts based solely on climatology is probably overstated. The climatology of the retrospective models would not be a reliable tool for a future comprised of Atlantic hurricane seasons like 2004 or 2005.

3.4.d.1 PRIME AE Forecasts

Figure 3.12 shows the average AE of R-PRIME, PRIME, and climatological AE predictions. The number of cases for each forecast interval is approximately equal to those listed in table 3.2. In each model subplot, the black and blue lines respectively indicate data originating from R-PRIME and PRIME. The solid lines illustrate PRIME results while the dashed lines represent climatological results. For all forecast intervals and models, R-PRIME outperformed PRIME, and both versions of PRIME had smaller errors than their respective climatological forecasts. A paired t test (e.g. equation 5.11 from Wilks 2011), adjusted for serial correlation, determined that the differences between
PRIME and climatology errors for all forecast intervals, models, and versions of PRIME were significant at the 95% level. Additionally, both versions of PRIME were able to forecast the AE of the models’ intensity forecasts significantly better than the models forecasted intensity.

Overall, climatological forecasts were better at predicting AE when retrospective data were used instead of the real-time data. This result is not surprising because when using retrospective data, the models remain consistent throughout the time series, and the errors vary less from season to season. The climatological forecasts of the two dynamical models, HWFI and GHMI, showed significantly more improvement (up to 50% at some forecast hours) between the real-time and retrospective datasets than the statistical models. This substantial reduction in the error of the retrospective climatological forecasts of HWFI and GHMI is likely attributable to their intensity forecasts progressing more than DSHP and LGEM. The lower AEs of HWFI and GHMI are associated with fewer high-error forecast events and thus, less variance in their AE forecasts. In fact, the intensity forecasts of the retrospective versions of these models had 75% lower AE variance than the real-time versions for select forecast intervals. When models improve and the variability of forecast errors is reduced, it is harder for PRIME to explain the variance of the errors, and climatology becomes a more viable estimate for future forecast AE.

For similar reasons, PRIME generally improved upon climatology more for extended-range instead of short-range AE forecasts. Recent studies by Torn and Snyder (2012) and Landsea and Franklin (2013) suggested that average best track error can reach ~10 knots, which is approximately equal to the AE of intensity forecasts at short forecast intervals.
The systematic error in the shorter forecast intervals translates into random noise that PRIME cannot anticipate. At longer forecast intervals, the variance in the AE is much larger, which provides PRIME with an opportunity to detect statistical trends and improve upon climatology.

SS represents another method to summarize how PRIME AE compared with the AE of climatological forecasts. Using equation (1) as a guide, SS essentially normalizes PRIME errors with climatology errors. A positive SS represents an improvement upon climatology, with the highest SS being 100%. Averaged over all the years of R-PRIME AE forecasts, SS ranged from 6-9% for HWFI, 5-10% for GHMI, 5-12% for LGEM, and 6-15% for DSHP. Although HWFI was the model with the lowest AE in its R-PRIME AE forecasts, it did not have the highest SSs because of how well its climatological forecasts performed. For PRIME, SSs ranged from 7-11% for HWFI, 6-16% for GHMI, 4-10% for LGEM, and 6-10% for DSHP. When the sample size was expanded to the original totals shown in table 3.1, PRIME performed slightly better and climatological forecasts performed slightly worse. The SSs increased to 7-13% for HWFI, 8-17% for GHMI, 6-11% for LGEM, and 6-10% for DSHP.

The final statistical measure presented here to summarize the average performance of PRIME AE forecasts is the percent of the total variance of the true AE values explained by PRIME AE forecasts. Table 3.9 lists the $R^2$ for each model, forecast interval, and version of PRIME. The first column for each model represents R-PRIME $R^2$ values and the second column represents PRIME $R^2$ values. The bold numbers are used to indicate whether PRIME or R-PRIME has the higher $R^2$ value for the particular forecast interval. These values were calculated using independent verification for the homogeneous sample
of R-PRIME and PRIME. Many of the trends observed in figure 3.12 were further emphasized by table 3.9. In general, the $R^2$ values of both versions of PRIME were larger at longer forecast intervals. PRIME performed better for the dynamical models than the statistical models, and the opposite was true for R-PRIME.

3.4.d.2 PRIME Bias Forecasts

Figure 3.13 is almost identical to figure 3.12 except it shows the average AE of PRIME and R-PRIME for bias predictions compared to climatological forecasts. The colors and lines have the same interpretation as in figure 3.12. Again, R-PRIME had smaller errors than PRIME, and both versions of PRIME had smaller errors than their respective climatological forecasts. Paired $t$ tests revealed that, except for the 96- and 108-hour HWFI R-PRIME bias forecasts, the differences between PRIME and climatological forecast AE for all forecast intervals, models, and versions of PRIME were significant at the 99% level. PRIME generally improved on climatological forecasts for the longer forecast intervals, which was also observed for PRIME AE forecasts (omitting the aforementioned HWFI forecast intervals). This trend is attributable to the additional leeway for PRIME to capture the variability in the signal of these longer forecasts.

Similar to the AE forecasts, the performance of the climatological bias forecasts for the statistical models barely changed from the real-time to retrospective versions of the model, while the behavior of the dynamical models was quite different. Using retrospective forecasts instead of real-time forecasts, the AE of the climatological predictions of GHMI and HWFI bias were approximately 20%-40% less for a majority of the forecasts times. Regardless, PRIME generally showed larger SSs for bias predictions compared to AE predictions. Averaged over all the years of R-PRIME bias forecasts, SSs
ranged from 1-7% for HWFI, 9-15% for GHMI, 9-17% for LGEM, and 8-21% for DSHP. For PRIME, SSs were 8-29% for HWFI, 11-22% for GHMI, 5-15% for LGEM, and 7-15% for DSHP. When the sample size of PRIME was expanded to the original totals shown in table 3.1, the SSs were almost identical.

Table 3.10 is similar to table 3.9 but lists the percent of the total variance explained for bias instead of AE. Although many of the trends discussed for AE forecasts were also visible for bias forecasts, the $R^2$ values were higher for bias forecasts. The larger variability in bias from forecast-to-forecast and stronger correlations between the predictors and forecast bias are likely responsible for this improvement. The $R^2$ values for the longer forecast intervals of HWFI showed the largest decrease of all the models between PRIME and R-PRIME. The bias of the HWFI retrospective forecasts had significantly less variance in than the real-time forecasts, which decreased the opportunities for PRIME to detect a signal in the bias. LGEM and DSHP attained substantially higher $R^2$ values for the retrospective forecasts compared to the real-time forecasts and appeared to benefit the most from the consistent models between the training and verification periods. Overall, the different statistical measures summarizing the average performance of PRIME bias and AE forecasts indicate that PRIME would represent a significant upgrade to climatological forecasts of intensity error.

3.4.e Annual Variability of PRIME Performance

To better diagnose the interseasonal variability of PRIME for the sample discussed above, the annual performance of R-PRIME and PRIME was analyzed. SS is again used to compare PRIME to climatological forecasts. We omit PRIME results from this section because R-PRIME had lower errors and similar yearly variations. However, it is
important to reiterate that the retrospective climatological forecasts were more accurate, which translated into lower SSs for R-PRIME compared to PRIME for most forecast intervals.

Figure 3.14 shows the yearly (colored lines) and entire 2008-2014 sample (dashed black line) SS of R-PRIME AE forecasts for all the forecast intervals of each model. The SS of the 2009 and 2013 hurricane seasons is not plotted for any model because they correspond to the best and worst performing seasons for a majority of the models, forecast intervals, and predictands. More importantly, these seasons are supplemented with real-time forecasts for GHMI and have fewer cases; therefore, their behavior is not representative of the entire sample. R-PRIME bias forecasts of HWFI and GHMI showed the largest changes in performance in 2013 and 2009, and possible reasons will be discussed in the next section.

In figure 3.14, R-PRIME AE forecasts achieve positive skill at a majority of the forecast intervals for DSHP and LGEM. PRIME LGEM forecasts only have negative skill for 12-96 hour forecasts in 2010 and 24-48 hour forecasts in 2009 (not shown). PRIME DSHP forecasts have negative skill for 12-84 hour forecasts in 2010, 108-hour forecasts in 2014, 60-72 hour forecasts in 2013 (not shown), and 24-48 hour forecasts in 2009 (not shown). Figure 3.14 also shows PRIME AE forecasts for HWFI and GHMI have forecast hours with negative SSs scattered throughout the different seasons. However, all models avoid negative SSs that exceed 10%, which suggests that PRIME was never a notable downgrade to climatological forecasts. At the same time, the SSs for the 2008 and 2012 seasons stay above 10% for every model and forecast interval.
Figure 3.15 demonstrates that R-PRIME also has positive results for bias forecasts. In these plots, the forecasts for statistical models have higher SSs than for dynamical models. The variability in yearly performance is also higher for the statistical models. Over the entire time series, DSHP and LGEM each has three forecast hours with negative SS, but the magnitude of these SSs is small. For the 2009 and 2013 seasons, the R-PRIME bias forecasts for both of these models never record a negative SS for a forecast hour, and several forecast hours have SSs above 30%. HWFI and GHMI also have only a handful of forecast hours with negative SS and appear to consistently improve upon climatology. Both models achieve their highest average SS in 2009 and their lowest average SS in 2013 (not shown). By comparing the four plots in figure 3.15, 2012 was the only year where PRIME has positive SS for all forecast intervals and models. Further investigation is needed to understand the behavior of PRIME during this season because it is also arguably the strongest for PRIME AE forecasts.

3.4.f Case Studies

The 2009 season was relatively inactive and had substantially fewer verified retrospective forecasts than other seasons. However, Hurricanes Ida, Bill, and Fred all reached category 2 strength or higher, and both Bill and Ida interacted with land. As a result, the fluctuating bias in these forecasts provided PRIME with opportunities to improve on climatology. In 2009, R-PRIME excelled for the bias forecasts of all models and forecast periods but the 96-hour DSHP bias forecasts were particularly skillful. Figure 3.16 displays the relationship between DSHP true bias and predicted bias for 96-hour R-PRIME forecasts. The PRIME regression equation that produced these bias predictions was:
Bias = 0.66 × DFEM + 0.12 × 0 Hr LON + 0.12 × LDIS - 0.23 × AVERAGE SHRDIR + 0.02 × 0 Hr DIV \hspace{1cm} (4)

Each ‘x’ in the scatter plot represents one verified forecast where the dependent variable is true bias and the independent variable is the bias predicted by (4). The dashed line shows the linear fit to the data. For the 24 verified forecasts, R-PRIME was able to explain 85.5% of the total variance in the bias, and the AE of its forecasts was only 7.6 knots compared to 19.0 knots for climatology. These impressive statistics can be attributed to the training period for the season selecting adept predictors for this forecast interval and model. Using the 2008 and 2010-2014 hurricane seasons, DFEM was determined to be the most significant predictor of error for 96-hour DSHP bias forecasts. The large positive weighting coefficient of DFEM derived from the training data was especially appropriate for the 2009 season. The ensemble mean of the four models served as a reliable intensity forecast and when DSHP deviated from the ensemble mean, a similar bias was observed. The second most important predictor was average SHRDIR. The negative coefficient in (4) indicates that shear from the west quadrant (SHRDIR = -0.707 to -1) have positive biases and shear from the east quadrant have negative biases (SHRDIR = 0.707 to 1). Both of these predictors captured observed trends for the 2009 season which enabled accurate R-PRIME bias forecasts.

In figure 3.14, the 2010 season was the worst season for R-PRIME AE forecasts for all models. During 2010, there were more long-lasting hurricanes than any other season but most of them avoided land. Additionally, these storms experienced rapid intensity changes at low latitudes, which led to high intensity forecast errors for all models. Averaging the predicted AE for all the storms in 2010, R-PRIME routinely expected between 20%-40% less AE than the validated values. Figure 3.17 shows the R-PRIME
predicted AE for 48-hour LGEM forecasts plotted against the true AE. This figure highlights the underestimation of the AE of LGEM for a particular forecast period but similar trends could be shown for other models and forecast periods during 2010.

The negative bias of R-PRIME AE forecasts for this model and forecast interval was caused by the two leading predictors not functioning well for this particular hurricane season. For the high-error storms boxed in figure 3.17, all of the models were either in modest agreement for the intensity forecasts but wrong or individual models deviated from the ensemble mean correctly. As a result, SPRD and ADEM did not have their expected strong positive correlations with AE. In these situations, PRIME might be improved by adding more models to the ensemble to help enhance its accuracy and ability to capture uncertainty. Also, a more expansive retrospective forecast sample size would allow PRIME to be trained with additional forecasts of strong storms and benefit future versions of PRIME.

There were 13 named storms in 2013 but only two reached hurricane intensity and no storms became category 2 strength. A hurricane count of 2 was the lowest seasonal total since 1982, and a season with no hurricanes exceeding category 1 strength had not happened since 1968; this anomalous behavior negatively affected R-PRIME bias forecasts for HWFI and GHMI. To explain the poor performance of R-PRIME for this season, Figure 3.18 shows all the 72-hour, 96-hour, and 120-hour HWFI intensity forecasts during Tropical Storm Dorian adjusted by the corresponding R-PRIME bias forecasts. The blue x’s, green triangles, and red pluses represent the bias-corrected 72-hour, 96-hour, and 120-hour intensity forecasts, respectively. The black line indicates the
best-track intensities of Dorian during the specified days in July. Clearly, R-PRIME was expecting sustained or increased intensity while Dorian weakened.

Evaluating the sign and magnitude of the weighting coefficients for the bias predictors of HWFI can provide a foundation for understanding these errors. ALAT, DFEM, POT, and fitted LDIS were the four main predictors for the 72-120 hour HWFI bias forecasts. For these forecast intervals, ALAT had negative coefficients and fitted LDIS had positive coefficients, and because Dorian was located at low latitudes away from land, these predictors correctly forecasted a positive bias. The POT and DFEM predictors had the largest weighting coefficients and were expected to be positively correlated with bias. These predictors were primarily responsible for the negative biases forecasted by R-PRIME for the longer forecast intervals.

For Dorian and the 2013 season in general, LGEM and DSHP positively deviated from the ensemble mean while GHMI and HWFI intensity forecasts were consistently below the ensemble mean. Hence, DFEM was erroneously implying HWFI forecasts with negative biases for Dorian. Also, Dorian was a weaker than average storm with low POT which translated to high negative biases in other seasons. Negatives biases were not observed for Dorian as environmental forcing during the 2013 season weakened storms that would normally intensify. In theory, a more expansive retrospective training sample for PRIME would contain additional unique cases like Dorian and allow PRIME to anticipate storms with uncommon behavior. Additionally, a larger ensemble to improve the DFEM and SPRD predictors could develop PRIME into an even better forecasting tool. These case studies represent only a small portion of the investigation required to
explain unique behavior during particular seasons and storms for all the models. More comprehensive analysis is needed to improve future versions of PRIME.

3.5. Summary of PRIME and Potential Improvements

The Prediction of Intensity Model Error (PRIME) model is a real-time statistical scheme that predicts the absolute error (AE) and bias of the four primary Atlantic basin intensity forecast models. Error guidance was created by inputting dynamical parameters and proxies for atmospheric flow stability and initial condition uncertainty into stepwise multiple linear regression formulas. Two versions of PRIME were tested, one involving the real-time versions of the intensity models (PRIME) and the other involving retrospective versions (R-PRIME). The sample for R-PRIME spans from 2008-2014 and includes about two-thirds as many cases as PRIME, which spans from 2007-2014. For the bias forecasts of both versions of PRIME, deviation of a model’s intensity forecast from the ensemble mean (DFEM) was the leading predictor, and initial (0INT) and forecast intensity (FINT) were also consistently significant. It was harder to isolate a particular predictor as the most important for AE forecasts but the most significant predictors among the four models were intensity forecast spread (SPRD), average forecast latitude (ALAT), forecast intensity (FINT), and absolute intensity deviation from the ensemble mean (ADEM).

To simulate the performance of the PRIME model in an operational setting, a standard cross-validation method was utilized to independently verify the bias and AE forecasts. Comparing the homogeneous sample for the two versions of PRIME, R-PRIME was found to have lower AE in its forecasts than PRIME for every forecast interval, model, and predictand. Each version of PRIME was also tested separately
against forecasts based on the error climatology of their training samples. Due to much poorer climatological forecasts, PRIME improved more than R-PRIME on climatology. Using the sample sizes in tables 3.1 and 3.2, PRIME skill scores were 6-17% for AE forecasts and 6-29% for bias forecasts while R-PRIME skill scores were only 5-15% for AE forecasts and 1-21% for bias forecasts. For all forecast intervals, models, and versions of PRIME, the error differences between the climatological forecasts and PRIME AE forecasts were significant at the 95% level. PRIME bias forecasts were significantly better than the climatological forecasts at the 99% level for all forecast hours and models. R-PRIME bias forecasts were significantly better than the climatology forecasts at the 99% level for all but two forecast hours in the HWFI model.

Although these results indicate that PRIME has overall skill in the prediction of TC forecast error, there are certain situations where PRIME appears particularly valuable. Forecasts that involve bias as a predictand, longer forecast intervals, and storms that either interact or are forecasted to interact with land have considerably more variations in model performance. As a result, systematic errors are a smaller percentage of the total error, and PRIME can detect meaningful statistical trends. For similar reasons, PRIME achieves higher SSs when models struggle with their intensity forecasts. With larger error variances in intensity forecasts, climatology becomes less representative of future forecast error and PRIME becomes more useful. Admittedly though, PRIME represents a first step in providing accurate guidance of intensity model performance, and additional improvement is needed for the difficult error forecast situations discussed in this chapter.

The first update to PRIME that should provide immediate improvement is a larger sample of retrospective forecasts. The current sample has a shortage of high-intensity and
landfalling storms which could negatively affect PRIME forecasts when these situations occur operationally. Even though the presented results use four state-of-the-art models as ensemble members, more members are necessary to improve the predictors that involve the dispersion of the track and intensity forecasts of the ensemble. Logical additions to the ensemble would include the GFS model and the European Center for Medium-Range Weather Forecast (ECMWF) model. These global models have shown marked improvement in their TC intensity forecasts, and their forecasts are also available in the a-deck files. If the sample sizes of these models increase in future seasons or additional cases are added to the a-deck files of prior seasons, their forecasts could easily be added to the PRIME ensemble.

Better results are also possible if predictors varied with the forecast period. Currently, the predictors used with PRIME were based on the strongest predictors observed in the training dataset over all forecast periods. Additionally, the low-order polynomials and Gaussian functions that were applied to predictors could improve PRIME further if the empirical functions were trained for specific forecast hours. The skill of PRIME might also increase if additional proxies were included in the predictor pool that better captured the quality of the initial analysis.

In summary, this chapter confirms that the nature of the TC and its surrounding environment affect the performance of intensity models, and this information can be used to provide skillful a priori estimates of intensity forecast error. The success of this preliminary version of PRIME suggests these forecasts could have a number of applications for the operational community. Currently, PRIME produces only deterministic forecasts of error but studies have shown communicating the uncertainty
associated with TC intensity forecasts as a probability might be more useful for end users (National Research Council 2006). As a result, probabilistic forecasts using PRIME will be featured in chapter 5. PRIME error forecasts could also be used to bias-correct the individual models in real time and improve the original intensity forecasts of the models. Chapter 4 will show that each model has significantly lower errors when PRIME bias corrections are applied.

Finally, PRIME forecasts will be used to unequally weight a multimodel ensemble comprising of the HWFI, GHMI, LGEM, and DSHP models. The NHC has used the intensity guidance of the equal-weighted multimodel consensus of these four models for years under the acronym ICON, and it routinely records the lowest forecast errors. The Florida State Super Ensemble (FSSE) is an example of a multimodel ensemble that does not assign equal weights to all models in the ensemble. Krishnamurti et al. (2000) developed a post-processing procedure for ensembles involving multiple linear regression and applied this method to many forecasting problems, including TC track and intensity forecasts. For the FSSE, the forecast errors of the individual intensity models during the training period are used to weight the models and predict future forecast error. In chapter 4, the “performance-controlling” parameters highlighted by PRIME will be used to post-process the multimodel ensemble of the four models discussed here with the hope of producing more skillful forecasts than FSSE and equal-weighting schemes.
Fig. 3.1

Time series of annual 24-120 hour absolute intensity forecast error of the best model available for guidance. The linear fit for each forecast hour is included as the dashed line. Section 1.2 provides an explanation of how the absolute error of the best model is determined.
Fig. 3.2

120-hour HWFI intensity forecast bias plotted against deviation of HWFI intensity forecast from the ensemble mean for the 2007–14 Atlantic seasons. The dashed line represents the linear fit to the data.
Fig. 3.3

72-hour GHMI absolute error plotted against GHMI absolute intensity forecast deviation from the ensemble mean (top left), ensemble intensity spread (top right), average latitude (bottom left), and forecast intensity (bottom right) for the 2007–14 Atlantic seasons. The dashed lines represent the linear regression fits to the data.
Fig. 3.4

Histograms showing the 72-hour real-time GHMI absolute error for the 2007-2013 Atlantic hurricane seasons before (top) and after (bottom) the Box-Cox transformation. A $\lambda$ of 0.22 along with the original absolute error values is inputted into equation (3) to obtain the bottom histogram.
Fig. 3.5

108-hour retrospective DSHP bias plotted against 0-hour distance to land (LDIS) for the 2008-2013 Atlantic hurricane seasons before (top) and after (bottom) a second-order polynomial transformation of the predictor. The top figure shows the raw LDIS values for the DSHP model plotted against 108-hour bias. The dashed red curve is fit using data from all forecast intervals. The bottom figure displays bias versus fitted LDIS. The solid blue line represents the linear fit to the data. The $R$ values in the top right corner of each plot capture the linear correlation of the independent and dependent variable.
Fig. 3.6

Idealized scenario explaining the different behavior certain predictors displayed when plotted against bias (top image) and absolute error (bottom image). The solid blue lines indicate the linear fits to the data. R values describing linear correlation of the predictor and predictand are included in bottom right corner of the images. This predictor shows higher linear correlation with absolute error than bias.
Fig. 3.7

Similar to figure 3.6 except this predictor shows higher linear correlation with bias than absolute error.
Fig. 3.8

Predicted AE versus predicted |bias| for 72-hour LGEM forecasts. The 2007-2013 hurricane seasons are used to develop the regression formula for these PRIME predictions and 2014 serves as the verification period. The solid blue line indicates the linear fit to the data, and the R value of the line is 0.10.
Fig. 3.9

The average absolute error of PRIME |bias| forecasts, PRIME AE forecasts, and climatological absolute error forecasts for HWFI, GHMI, LGEM, and DSHP. Error statistics are calculated using cases from 2007-2014 that are described in table 3.1.
The average absolute error of absolute error forecasts for PRIME with proxies, PRIME without proxies, and climatological forecasts for HWFI, GHMI, LGEM, and DSHP. Error statistics are calculated using cases from 2007-2014 that are described in table 3.1.
Fig. 3.11

Similar to figure 3.10 except the plots apply to bias predictions.
Fig. 3.12

The average absolute error of PRIME, retrospective PRIME (R-PRIME), and climatological absolute error forecasts for HWFI, GHMI, LGEM, and DSHP. Error statistics are calculated using cases from 2008-2014 that are available for both PRIME and R-PRIME.
Fig. 3.13

Similar to figure 3.12 except the plots apply to bias predictions.
The yearly skill score of retrospective PRIME (R-PRIME) absolute error forecasts for each model and forecast hour. The skill score of the total sample is demarcated with a dashed black line and cases from 2008, 2010, 2011, 2012, and 2014 are represented by different colored lines. A positive skill score indicates PRIME has lower absolute error than climatology. The 2009 and 2013 seasons are excluded from the plots because of the non-representative behavior described in section 3.4.
Fig. 3.15

Similar to figure 3.14 except it shows the skill score of retrospective PRIME (R-PRIME) bias forecasts normalized by climatological bias forecasts.
Fig. 3.16

Predicted bias versus true bias for 96-hour DSHP forecasts. The 2008 and 2010-2014 hurricane seasons are used to develop the regression formula for these retrospective PRIME (R-PRIME) predictions and 2009 serves as the verification period. The dashed line indicates the linear fit to the data, and the $R$ value of the line is included in the top right corner of the figure.
Predicted absolute error (AE) versus true AE for 48-hour LGEM forecasts. The 2008-2009 and 2011-2014 hurricane seasons are used to develop the regression formula for these retrospective PRIME (R-PRIME) predictions, and 2010 serves as the verification period. The dashed line indicates the linear fit to the data, and the $R$ value of the line is included in the top right corner of the figure. The red box highlights cases where SPRD and ADEM are especially poor predictors.
The observed best-track intensity (solid black line) and the 72-hour, 96-hour, and 120-hour bias-corrected HWFI intensity forecasts for Tropical Storm Dorian in 2013. Retrospective PRIME (R-PRIME) bias forecasts are used to adjust the original intensity forecast of HWFI. 72-hour bias-corrected forecasts are indicated by blue x’s, 96-hour bias-corrected forecasts are indicated by green triangles, and 120-hour bias-corrected forecasts are indicated with red plusses.
<table>
<thead>
<tr>
<th># of Cases</th>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1917</td>
<td>12</td>
<td>-1.7, 7.7</td>
<td>-0.9, 7.8</td>
<td>-0.9, 7.5</td>
<td>-0.4, 7.4</td>
</tr>
<tr>
<td>1715</td>
<td>24</td>
<td>-2.1, 10.3</td>
<td>-1.2, 10.5</td>
<td>-0.9, 10.2</td>
<td>0.6, 10.1</td>
</tr>
<tr>
<td>1512</td>
<td>36</td>
<td>-1.2, 11.9</td>
<td>-0.4, 12.5</td>
<td>-0.6, 11.7</td>
<td>1.5, 11.6</td>
</tr>
<tr>
<td>1323</td>
<td>48</td>
<td>0.3, 12.9</td>
<td>1.3, 13.3</td>
<td>-0.1, 12.5</td>
<td>2.1, 12.5</td>
</tr>
<tr>
<td>1160</td>
<td>60</td>
<td>1.7, 13.7</td>
<td>2.6, 13.9</td>
<td>0.4, 13.0</td>
<td>2.4, 13.2</td>
</tr>
<tr>
<td>1023</td>
<td>72</td>
<td>2.8, 14.3</td>
<td>3.8, 14.3</td>
<td>1.3, 13.2</td>
<td>2.7, 13.4</td>
</tr>
<tr>
<td>886</td>
<td>84</td>
<td>3.3, 14.5</td>
<td>4.5, 14.2</td>
<td>1.8, 13.2</td>
<td>2.6, 13.5</td>
</tr>
<tr>
<td>762</td>
<td>96</td>
<td>3.9, 14.5</td>
<td>4.8, 14.0</td>
<td>2.3, 13.2</td>
<td>2.4, 13.5</td>
</tr>
<tr>
<td>664</td>
<td>108</td>
<td>4.2, 14.7</td>
<td>4.8, 13.8</td>
<td>2.5, 13.0</td>
<td>2.2, 13.5</td>
</tr>
<tr>
<td>582</td>
<td>120</td>
<td>4.6, 14.7</td>
<td>5.0, 13.9</td>
<td>2.7, 13.0</td>
<td>2.0, 13.4</td>
</tr>
</tbody>
</table>

Table 3.1

The average bias and absolute error of HWFI, GHMI, LGEM, and DSHP is calculated using the real-time forecasts of Atlantic basin storms between 2007 and 2014. The first column of the table represents the number of verified real-time forecasts. For each model and forecast interval, the first entry in a table cell is the bias and the second entry is the absolute error.
<table>
<thead>
<tr>
<th># of Cases</th>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1182</td>
<td>12</td>
<td>-1.9, 7.1</td>
<td>-1.1, 7.3</td>
<td>-1.0, 7.2</td>
<td>-0.4, 7.2</td>
</tr>
<tr>
<td>1095</td>
<td>24</td>
<td>-2.2, 9.3</td>
<td>-2.3, 9.7</td>
<td>-1.2, 10.0</td>
<td>0.4, 10.1</td>
</tr>
<tr>
<td>1012</td>
<td>36</td>
<td>-2.2, 10.6</td>
<td>-2.7, 11.5</td>
<td>-1.3, 11.7</td>
<td>0.8, 11.7</td>
</tr>
<tr>
<td>924</td>
<td>48</td>
<td>-2.4, 11.3</td>
<td>-2.6, 12.3</td>
<td>-1.6, 12.6</td>
<td>0.5, 12.6</td>
</tr>
<tr>
<td>838</td>
<td>60</td>
<td>-2.2, 11.8</td>
<td>-2.4, 12.4</td>
<td>-1.7, 13.2</td>
<td>0.3, 13.4</td>
</tr>
<tr>
<td>758</td>
<td>72</td>
<td>-1.7, 12.0</td>
<td>-1.5, 12.3</td>
<td>-1.4, 13.2</td>
<td>0.1, 13.4</td>
</tr>
<tr>
<td>679</td>
<td>84</td>
<td>-1.1, 11.8</td>
<td>-0.5, 12.2</td>
<td>-0.7, 12.8</td>
<td>0.0, 13.2</td>
</tr>
<tr>
<td>614</td>
<td>96</td>
<td>-0.7, 11.7</td>
<td>0.1, 12.4</td>
<td>-0.1, 12.5</td>
<td>-0.1, 13.0</td>
</tr>
<tr>
<td>547</td>
<td>108</td>
<td>0.0, 11.9</td>
<td>1.0, 12.7</td>
<td>0.5, 12.3</td>
<td>0.0, 13.0</td>
</tr>
<tr>
<td>492</td>
<td>120</td>
<td>1.0, 12.2</td>
<td>1.6, 12.6</td>
<td>0.8, 12.4</td>
<td>0.1, 12.8</td>
</tr>
</tbody>
</table>

Table 3.2

Similar to table 3.1 except the entries correspond to the statistics of verified retrospective forecasts for each model at each forecast interval. These values are calculated using Atlantic basin storms between 2008 and 2014.
Table 3.3

The percent of the total variance explained ($R^2$) by PRIME for dependent verification of absolute error forecasts for each model and forecast hour. These values are calculated using the same cases as table 3.1.

<table>
<thead>
<tr>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>13.7</td>
<td>14.4</td>
<td>13.7</td>
<td>13.0</td>
</tr>
<tr>
<td>24</td>
<td>12.3</td>
<td>13.7</td>
<td>13.7</td>
<td>11.6</td>
</tr>
<tr>
<td>36</td>
<td>14.4</td>
<td>16.8</td>
<td>16.0</td>
<td>11.6</td>
</tr>
<tr>
<td>48</td>
<td>15.2</td>
<td>16.0</td>
<td>20.3</td>
<td>21.2</td>
</tr>
<tr>
<td>60</td>
<td>19.4</td>
<td>17.6</td>
<td>21.2</td>
<td>22.1</td>
</tr>
<tr>
<td>72</td>
<td>22.1</td>
<td>20.3</td>
<td>20.3</td>
<td>20.3</td>
</tr>
<tr>
<td>84</td>
<td>25.0</td>
<td>24.0</td>
<td>25.0</td>
<td>23.0</td>
</tr>
<tr>
<td>96</td>
<td>28.1</td>
<td>29.2</td>
<td>26.0</td>
<td>28.1</td>
</tr>
<tr>
<td>108</td>
<td>36.0</td>
<td>34.8</td>
<td>25.0</td>
<td>33.6</td>
</tr>
<tr>
<td>120</td>
<td>38.4</td>
<td>38.4</td>
<td>30.3</td>
<td>38.4</td>
</tr>
<tr>
<td>Hours</td>
<td>HWFI</td>
<td>GHMI</td>
<td>LGEM</td>
<td>DSHP</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>12</td>
<td>23.0</td>
<td>23.0</td>
<td>21.2</td>
<td>16.0</td>
</tr>
<tr>
<td>24</td>
<td>29.2</td>
<td>30.3</td>
<td>26.0</td>
<td>21.2</td>
</tr>
<tr>
<td>36</td>
<td>34.8</td>
<td>42.3</td>
<td>27.0</td>
<td>25.0</td>
</tr>
<tr>
<td>48</td>
<td>41.0</td>
<td>42.3</td>
<td>30.3</td>
<td>28.1</td>
</tr>
<tr>
<td>60</td>
<td>42.3</td>
<td>43.6</td>
<td>30.3</td>
<td>28.1</td>
</tr>
<tr>
<td>72</td>
<td>46.2</td>
<td>43.6</td>
<td>30.3</td>
<td>29.2</td>
</tr>
<tr>
<td>84</td>
<td>49.0</td>
<td>43.6</td>
<td>32.5</td>
<td>32.5</td>
</tr>
<tr>
<td>96</td>
<td>54.8</td>
<td>47.6</td>
<td>38.4</td>
<td>41.0</td>
</tr>
<tr>
<td>108</td>
<td>62.4</td>
<td>51.8</td>
<td>41.0</td>
<td>47.6</td>
</tr>
<tr>
<td>120</td>
<td>67.2</td>
<td>57.8</td>
<td>43.6</td>
<td>50.4</td>
</tr>
</tbody>
</table>

Table 3.4

Similar to table 3.3 except the $R^2$ values apply to bias forecasts for each model and forecast hour.
Table 3.5

<table>
<thead>
<tr>
<th>Pred- ictors</th>
<th>HWFI</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>24</td>
<td>36</td>
<td>48</td>
<td>60</td>
<td>72</td>
<td>84</td>
<td>96</td>
<td>108</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>ADEM</td>
<td>0.15</td>
<td>0.13</td>
<td>0.16</td>
<td>0.13</td>
<td>0.17</td>
<td>0.18</td>
<td>0.14</td>
<td>0.04</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0INT</td>
<td>-0.13</td>
<td>-0.21</td>
<td>-0.14</td>
<td>-0.20</td>
<td>-0.21</td>
<td>-0.15</td>
<td>-0.17</td>
<td>-0.21</td>
<td>-0.22</td>
<td>-0.21</td>
<td></td>
</tr>
<tr>
<td>FINT</td>
<td>0.21</td>
<td>0.34</td>
<td>0.27</td>
<td>0.23</td>
<td>0.19</td>
<td>0.15</td>
<td>0.09</td>
<td>0.08</td>
<td>0.12</td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pred- ictors</th>
<th>GHMI</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>24</td>
<td>36</td>
<td>48</td>
<td>60</td>
<td>72</td>
<td>84</td>
<td>96</td>
<td>108</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>FINT</td>
<td>0.12</td>
<td>0.15</td>
<td>0.19</td>
<td>0.17</td>
<td>0.13</td>
<td>0.05</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>ALAT</td>
<td>-0.20</td>
<td>-0.10</td>
<td>-0.32</td>
<td>-0.48</td>
<td>-0.43</td>
<td>-0.27</td>
<td>-0.39</td>
<td>-0.24</td>
<td>-0.09</td>
<td>-0.18</td>
<td></td>
</tr>
<tr>
<td>ADEM</td>
<td>0.12</td>
<td>0.17</td>
<td>0.29</td>
<td>0.22</td>
<td>0.09</td>
<td>0.17</td>
<td>0.20</td>
<td>0.20</td>
<td>0.27</td>
<td>0.32</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pred- ictors</th>
<th>LGEM</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>24</td>
<td>36</td>
<td>48</td>
<td>60</td>
<td>72</td>
<td>84</td>
<td>96</td>
<td>108</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>SPRD</td>
<td>0.09</td>
<td>0.04</td>
<td>0.04</td>
<td>0.11</td>
<td>0.17</td>
<td>0.13</td>
<td>0.08</td>
<td>0.09</td>
<td>0.06</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>ALAT</td>
<td>-0.05</td>
<td>-0.08</td>
<td>-0.11</td>
<td>-0.15</td>
<td>-0.16</td>
<td>-0.15</td>
<td>-0.14</td>
<td>-0.17</td>
<td>-0.17</td>
<td>-0.17</td>
<td></td>
</tr>
<tr>
<td>ADEM</td>
<td>0.15</td>
<td>0.21</td>
<td>0.19</td>
<td>0.13</td>
<td>0.11</td>
<td>0.20</td>
<td>0.25</td>
<td>0.20</td>
<td>0.20</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pred- ictors</th>
<th>DSHP</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>24</td>
<td>36</td>
<td>48</td>
<td>60</td>
<td>72</td>
<td>84</td>
<td>96</td>
<td>108</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>ALAT</td>
<td>-0.03</td>
<td>-0.04</td>
<td>0.07</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.08</td>
<td>-0.06</td>
<td>-0.10</td>
<td>-0.13</td>
<td>-0.19</td>
<td></td>
</tr>
<tr>
<td>ADEM</td>
<td>0.15</td>
<td>0.15</td>
<td>0.11</td>
<td>0.16</td>
<td>0.20</td>
<td>0.21</td>
<td>0.24</td>
<td>0.20</td>
<td>0.14</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>SPRD</td>
<td>0.11</td>
<td>0.13</td>
<td>0.17</td>
<td>0.11</td>
<td>0.14</td>
<td>0.18</td>
<td>0.19</td>
<td>0.22</td>
<td>0.25</td>
<td>0.27</td>
<td></td>
</tr>
</tbody>
</table>

Normalized regression coefficients for the three best predictors of absolute error for each model. These weighting coefficients were derived from the retrospective version of PRIME (R-PRIME), using 2008-2014 as the training period.
Table 3.6

Similar to table 3.5 except the normalized regression coefficients apply to the best predictors of bias for each model.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Forecast Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12 24 36 48 60 72 84 96 108 120</td>
</tr>
<tr>
<td></td>
<td>HWFI</td>
</tr>
<tr>
<td>ALAT</td>
<td>0.01 0.06 0.11 0.17 0.19 0.18 0.15 0.13 0.14 0.11</td>
</tr>
<tr>
<td>LDIS</td>
<td>0.02 0.00 0.01 0.05 0.08 0.10 0.11 0.19 0.29 0.26</td>
</tr>
<tr>
<td>DFEM</td>
<td>0.35 0.35 0.37 0.32 0.31 0.33 0.36 0.36 0.31 0.41</td>
</tr>
<tr>
<td></td>
<td>GHMI</td>
</tr>
<tr>
<td>0INT</td>
<td>0.49 0.38 0.24 0.22 0.19 0.18 0.17 0.16 0.14 0.12</td>
</tr>
<tr>
<td>FINT</td>
<td>-0.52 -0.36 -0.21 -0.16 -0.10 -0.06 -0.06 -0.05 -0.02 0.01</td>
</tr>
<tr>
<td>DFEM</td>
<td>0.50 0.57 0.58 0.58 0.58 0.57 0.58 0.58 0.53 0.56</td>
</tr>
<tr>
<td></td>
<td>LGEM</td>
</tr>
<tr>
<td>FINT</td>
<td>-0.52 -0.36 -0.21 -0.16 -0.10 -0.06 -0.06 -0.05 -0.02 0.01</td>
</tr>
<tr>
<td>0INT</td>
<td>0.49 0.38 0.24 0.22 0.19 0.18 0.17 0.16 0.14 0.12</td>
</tr>
<tr>
<td>DFEM</td>
<td>0.50 0.57 0.58 0.58 0.58 0.57 0.58 0.58 0.53 0.56</td>
</tr>
<tr>
<td></td>
<td>DSHP</td>
</tr>
<tr>
<td>ADIV</td>
<td>0.02 0.02 0.02 0.05 0.07 0.09 0.10 0.11 0.09 0.06</td>
</tr>
<tr>
<td>LDIS</td>
<td>-0.04 -0.04 -0.01 0.03 0.04 0.07 0.09 0.15 0.22 0.25</td>
</tr>
<tr>
<td>DFEM</td>
<td>0.40 0.47 0.50 0.50 0.52 0.53 0.56 0.59 0.61 0.64</td>
</tr>
<tr>
<td>Hours</td>
<td>HWFI</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>12</td>
<td>10.1</td>
</tr>
<tr>
<td></td>
<td>8.5</td>
</tr>
<tr>
<td>24</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>8.1</td>
</tr>
<tr>
<td>36</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>10.5</td>
</tr>
<tr>
<td>48</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>8.5</td>
</tr>
<tr>
<td>60</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>8.3</td>
</tr>
<tr>
<td>72</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td>9.7</td>
</tr>
<tr>
<td>84</td>
<td>21.1</td>
</tr>
<tr>
<td></td>
<td>12.7</td>
</tr>
<tr>
<td>96</td>
<td>20.7</td>
</tr>
<tr>
<td></td>
<td>11.3</td>
</tr>
<tr>
<td>108</td>
<td>17.6</td>
</tr>
<tr>
<td></td>
<td>14.7</td>
</tr>
<tr>
<td>120</td>
<td>34.8</td>
</tr>
<tr>
<td></td>
<td>16.1</td>
</tr>
</tbody>
</table>

Table 3.7

The percent of the total absolute error variance explained ($R^2$) by PRIME for land (left column) and no land (right column) cases for each model and forecast interval. Bold numbers are used to indicate which column has a higher $R^2$ value for the model-forecast hour pair.
<table>
<thead>
<tr>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>5.4</td>
<td>13.9</td>
<td>25.1</td>
<td>21.2</td>
</tr>
<tr>
<td>24</td>
<td>19.0</td>
<td>20.6</td>
<td>35.6</td>
<td>22.7</td>
</tr>
<tr>
<td>36</td>
<td>32.0</td>
<td>39.5</td>
<td>38.1</td>
<td>34.8</td>
</tr>
<tr>
<td>48</td>
<td>46.4</td>
<td>35.2</td>
<td>50.7</td>
<td>38.9</td>
</tr>
<tr>
<td>60</td>
<td>54.5</td>
<td>25.3</td>
<td>44.1</td>
<td>35.4</td>
</tr>
<tr>
<td>72</td>
<td>49.5</td>
<td>28.5</td>
<td>55.4</td>
<td>44.7</td>
</tr>
<tr>
<td>84</td>
<td>55.0</td>
<td>37.9</td>
<td>60.3</td>
<td>52.2</td>
</tr>
<tr>
<td>96</td>
<td>59.6</td>
<td>37.1</td>
<td>38.7</td>
<td>34.0</td>
</tr>
<tr>
<td>108</td>
<td>57.6</td>
<td>31.6</td>
<td>35.6</td>
<td>46.2</td>
</tr>
<tr>
<td>120</td>
<td>53.2</td>
<td>28.5</td>
<td>29.3</td>
<td>23.4</td>
</tr>
</tbody>
</table>

**Table 3.8**

Similar to table 3.7 except the $R^2$ values apply to bias forecasts.
Table 3.9

The percent of the total absolute error variance explained ($R^2$) by retrospective (left column) and real-time (right column) PRIME for each model and forecast interval. Bold numbers are used to indicate which version of PRIME has a higher $R^2$ value for the model-forecast hour pair.

<table>
<thead>
<tr>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>7.3</td>
<td>5.7</td>
<td>2.7</td>
<td>7.8</td>
</tr>
<tr>
<td>24</td>
<td>5.7</td>
<td>7.3</td>
<td>3.2</td>
<td>6.9</td>
</tr>
<tr>
<td>36</td>
<td>5.8</td>
<td>9.1</td>
<td>8.0</td>
<td>10.1</td>
</tr>
<tr>
<td>48</td>
<td>6.8</td>
<td>6.6</td>
<td>8.6</td>
<td>11.5</td>
</tr>
<tr>
<td>60</td>
<td>5.6</td>
<td>5.9</td>
<td>4.8</td>
<td>9.0</td>
</tr>
<tr>
<td>72</td>
<td>5.1</td>
<td>8.7</td>
<td>4.7</td>
<td>6.3</td>
</tr>
<tr>
<td>84</td>
<td>4.8</td>
<td>9.6</td>
<td>7.9</td>
<td>6.5</td>
</tr>
<tr>
<td>96</td>
<td>5.0</td>
<td>11.0</td>
<td>7.3</td>
<td>9.1</td>
</tr>
<tr>
<td>108</td>
<td>6.0</td>
<td>15.1</td>
<td>13.9</td>
<td>16.4</td>
</tr>
<tr>
<td>120</td>
<td>9.7</td>
<td>19.2</td>
<td>15.6</td>
<td>16.7</td>
</tr>
<tr>
<td>Hours</td>
<td>HWFI</td>
<td>GHMI</td>
<td>LGEM</td>
<td>DSHP</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>12</td>
<td>12.5</td>
<td>14.2</td>
<td>18.1</td>
<td>16.8</td>
</tr>
<tr>
<td>24</td>
<td>11.6</td>
<td>18.2</td>
<td>24.7</td>
<td>25.6</td>
</tr>
<tr>
<td>36</td>
<td>11.1</td>
<td>21.8</td>
<td>26.5</td>
<td>25.9</td>
</tr>
<tr>
<td>48</td>
<td>7.3</td>
<td>24.4</td>
<td>26.9</td>
<td>25.0</td>
</tr>
<tr>
<td>60</td>
<td>6.6</td>
<td>26.1</td>
<td>28.8</td>
<td>27.6</td>
</tr>
<tr>
<td>72</td>
<td>9.3</td>
<td>31.5</td>
<td>30.1</td>
<td>30.2</td>
</tr>
<tr>
<td>84</td>
<td>11.7</td>
<td>36.0</td>
<td>29.9</td>
<td>32.3</td>
</tr>
<tr>
<td>96</td>
<td>11.0</td>
<td>40.1</td>
<td>27.5</td>
<td>34.3</td>
</tr>
<tr>
<td>108</td>
<td>7.3</td>
<td>41.5</td>
<td>23.1</td>
<td>34.6</td>
</tr>
<tr>
<td>120</td>
<td>14.6</td>
<td>44.8</td>
<td>25.2</td>
<td>37.5</td>
</tr>
</tbody>
</table>

**Table 3.10**

Similar to table 3.9 except the $R^2$ values apply to bias forecasts.
Chapter 4. Improving Intensity Forecasts with PRIME

4.1 Background and Motivation

The encouraging performance of PRIME\textsuperscript{11} broaches the question of whether it can serve as not only a forecasting supplement that quantifies how much confidence should be placed in an intensity forecast, but also as a tool for increasing the accuracy of an intensity forecast. To answer this question, PRIME error predictions are used as a post-processing tool for the TC models: DSHP, LGEM, GHMI, and HWFI. PRIME bias forecasts are first tested as corrections to intensity forecasts. The errors of the bias-corrected models are then compared to the errors of the original models to determine whether TC intensity forecasts were significantly\textsuperscript{12} improved. The output of PRIME is also utilized to assemble unique ensembles of the four models to better predict TC intensity.

Ensembles are produced from many parallel forecast realizations as a method to capture the uncertainty in the initial analysis as well as the deficiencies of a numerical model (Hamill et al. 2012). The ensemble mean forecast, which is the average of the forecasts of the different members of an ensemble, is more likely to achieve lower forecast errors than a single, higher-resolution model forecast (Leith 1974). For these reasons, ensembles are increasingly used for track and intensity forecasts of TCs and have recently produced better results than the best operational models (Torn and Hakim

\textsuperscript{11}Used throughout the chapter as an umbrella term for all forecasts created by the PRIME forecasting scheme. Unless contextually specified, “PRIME” refers to both the PRIME and R-PRIME models.

\textsuperscript{12}Unless otherwise specified, statistical significance is defined following the same methodology outlined in chapter 2 for paired $t$ tests.
2009; Zhang et al. 2011; Cangialosi and Franklin 2015). The discussion here will focus on ensembles that have members originating from different models or “multimodel ensembles”. Multimodel ensembles combine forecasts, typically by calculating the ensemble mean, from different models initialized at the same time to create a consensus forecast.

Using operational models as ensemble members often produces excellent results because each model is constructed with state-of-the-art parameterizations and initialization techniques. Each ensemble member provides forecasts that represent a realistic future state of the atmosphere and the spread of the forecasts often captures the range of potential solutions of the real atmosphere. The equally-weighted multimodel consensus forecasts, ICON and IVCN, have provided TC intensity guidance in the Atlantic basin since 2007. ICON is created by averaging the intensity forecasts of DSHP, LGEM, GHMI, and HWFI, while IVCN is formulated by averaging at least two of the intensity forecasts of DSHP, LGEM, GHMI, HWFI, and GFNI. According to NHC verification (Cangialosi and Franklin 2015; Cangialosi and Franklin 2013; Cangialosi and Franklin 2011) of Atlantic basin TCs, these consensus models typically register the lowest errors.

With the success of consensus TC intensity forecasts, it is surprising that alternatives to the equal-weighting technique have not surfaced. One of the potential explanations for the lack of research in this area is that many climate studies have demonstrated that combining different forecast systems with equal weights is typically the best approach to multimodel forecasting (Doblas-Reyes et al. 2005). Recently, DelSole et al. (2012) investigated the merits of unequal weighting and equal weighting for multimodel
forecasting of global precipitation and 2-meter temperature. They evaluated five models’ hindcasts for these fields from the ENSEMBLES dataset over the period 1960-2005 and recorded when the unequally-weighted ensemble forecasts had significantly smaller errors than the equally-weighted ensemble forecasts. Over the entire globe, unequal-weighting strategies rarely provided the more accurate forecasts for the two fields.

However, it is not clear whether these results are pertinent to the challenge of TC intensity forecasting. TC forecasts focus on a smaller portion of the globe and in limited regions, unequal weighting can provide significantly better results than equal weighting (DelSole et al. 2012). Also, with multiple TCs forecasted every year and numerous forecast verifications per storm, TC forecasts typically deal with larger sample sizes than climate forecasts. Development of a statistical model based on a training period of a few years for TC forecasting can reach 1000s of cases (BN13) while studies on unequal weighting in climate models routinely use less than 100 cases (Doblas-Reyes et al. 2005; Rodrigues et al. 2013). Larger sample sizes are critical for statistical post-processing techniques and often produce more conclusive results (Wilks 2006). Finally, there is considerable error variability between models based on the situation a TC is experiencing (BN13), whereas climate models in multimodel ensembles rarely show such large variations in skill. DelSole et al. (2012) explained that larger variations between ensemble members increase the chance that unequal weighting will outperform equal weighting.

For all these reasons, the unequally-weighted FSSE (Krishnamurti et al. 2000) model discussed in chapter 3 often outperforms the equally-weighted multimodel ensembles (Cangialosi and Franklin 2015). FSSE TC intensity forecasts are created by linearly
combining different models using a multiple linear regression framework. Krishnamurti et al. (2000) claimed that equally weighting models degrades the overall results of multimodel ensembles because inaccurate models are treated in the same way as accurate models in the ensemble mean. However, evaluating models irrespective of the particular atmospheric state defined by the TC appears to be a bigger flaw than weighting a weak model the same as a strong model (BN13). Based on chapters 2 and 3, a more effective scheme for multimodel ensembles would weight models depending on the synoptic conditions experienced by a TC as well as storm-specific characteristics.

As a result, the main goal of this chapter is to investigate if an unequally-weighted ensemble, based on PRIME output, can outperform ICON. PRIME AE and bias forecasts are used to create three unequally-weighted ensembles of DSHP, LGEM, GHMI, and HWFI. The performance of these ensembles is compared to a second set of ensembles that use PRIME bias and AE forecasts to eliminate or correct models before they are equally weighted. Section 4.2 discusses the upgraded predictor pools that provide the foundation for the PRIME forecasts used throughout this chapter. Section 4.3 presents the results of the PRIME bias-corrected models. Section 4.4 is devoted to the results of the ensemble intensity forecasts that are adjusted based on PRIME error forecasts. Conclusions are found in section 4.5.

4.2 Upgraded PRIME and R-PRIME

The original predictors tested in the development of PRIME are listed in table 1.1. In chapter 3, the fitted LDIS predictor was added to the bias predictor pool because it was the only fitted predictor, for either predictand, that improved the forecasts of all models, forecast intervals, and versions of PRIME. In this chapter, the condition for using a fitted
predictor is relaxed because forecast performance is prioritized over operational convenience. A fitted predictor is included in a regression formula for a model as long as, averaged over all forecast intervals, it improved independent verification results for a version of PRIME.

For PRIME and R-PRIME, the FINT predictor, empirically fit using a third-order polynomial, was found to improve transformed\textsuperscript{13} AE forecasts of DSHP, GHMI, and HWFI. Figure 4.1 provides an example of how a third-order polynomial was fitted to FINT to produce a more linear relationship with AE. The top image of figure 4.1 portrays the original curvilinear relationship between FINT and 96-hour retrospective DSHP AE for the 2009-2014 hurricane seasons with the corresponding $R$ value of the linear fit. As mentioned in chapter 3, PRIME and R-PRIME results are intended to simulate real-time performance so the training period (2009-2014) for determining the $R$-weighting coefficients of the third-order polynomial does not include the year (2008) that R-PRIME will forecast. The red dashed line is the third-order polynomial that best fits all forecast intervals and whose coefficients were used to transform the FINT values for the 2008 R-PRIME forecasts of DSHP AE. The bottom image of figure 4.1 shows fitted FINT plotted against 96-hour DSHP AE and the $R$ value of the modified data. The corresponding plots (not shown) created for PRIME, other forecast intervals, and other models captured similar behavior between FINT and AE before and after the fitting of FINT.

The positive linear correlation between FINT and AE discussed in chapter 3 is also visible in figure 4.1, but the nonlinear function highlights some unique behavior at low and high FINT. At high FINT, DSHP AE values deviate below the positive linear fit.

\textsuperscript{13} Note that the same Box-Cox transformation discussed in chapter 3 is applied to AE values before using multiple linear regression. For the rest of this chapter, “transformed AE” is replaced with “AE”.
These cases represent the strongest storms where eyewall replacement and RI have already occurred, and a TC is experiencing a quasi-steady intensity. For these low uncertainty situations, DSHP records anomalously low AE values. However, for low FINT, DSHP is often forecasting a TC that is expected to interact with land. There is a lot of uncertainty associated with these long-range forecasts and thus, low FINT values have AE values that deviate above a linear fit.

Initial divergence (0DIV), fitted with a third-order polynomial, was added to the predictor pool for PRIME and R-PRIME bias forecasts. After replacing the original 0DIV predictor with the fitted one, R-PRIME forecasts of DSHP and HWFI bias and PRIME forecasts of DSHP bias improved. Figure 4.2 shows 0DIV and fitted 0DIV plotted against 96-hour retrospective DSHP bias for the 2008-2013 hurricane seasons. The top panel of figure 4.2 contains the third-order polynomial that best describes the relationship between 0DIV and DSHP bias, over all forecast intervals, from 2008-2013. The polynomial coefficients of the red dashed line are used to transform the 0DIV values for the 2014 R-PRIME forecasts of DSHP bias. The bottom image of figure 4.2 shows fitted 0DIV plotted against 96-hour retrospective DSHP bias for the same forecast cases as the top image. After comparing the correlation coefficients of the top and bottom plots, it is obvious that the fitted 0DIV predictor provides an upgrade.

The physical reasoning for the shape of the red dashed curve in figure 4.2 is not clear. When 0DIV is between approximately -50 s\(^{-1}\) and 60 s\(^{-1}\), there is no trend observed. Between ~60 s\(^{-1}\) and ~120 s\(^{-1}\), DSHP has no large negative biases, but there is also no visible linear relationship. However, 0DIV above ~120 s\(^{-1}\) has a noticeable negative correlation with 96-hour DSHP bias. These high 0DIV cases are an example of DSHP
forecasting the weakening of TCs when they actually maintain their strength. It is possible that the formulation of DSHP is misjudging the connection between favorable upper level conditions and TC intensity but for a definitive explanation, further investigation is needed. The equivalent PRIME and HWFI figures are omitted here but show comparable behavior to figure 4.2. Other forecast intervals show similar behavior but the 0DIV values separating the different error regimes vary. However, as PRIME is currently formulated, the unique relationships between independent and dependent variables based on forecast length are not captured with the fitted predictors. Echoing the sentiment in chapter 3, another logical update to PRIME would involve nonlinear fitting functions that changed with the forecast intervals.

The final parameter added to the predictor pool was ALAT fitted with a two-peak Gaussian function. This predictor only significantly improved R-PRIME forecasts of HWFI bias. In figure 4.3, 108-hour retrospective HWFI bias for the 2008-2012 and 2014 Atlantic hurricanes seasons is plotted against ALAT before (top) and after (bottom) it was fitted with a two-peak Gaussian function. In the top panel, the red dashed line represents the two-peak Gaussian that best fits the data. In the bottom panel, the solid blue line is the linear fit of the transformed data. Comparing the $R$ values of the top and bottom panels demonstrates the two-peak Gaussian function improves the linear fit.

The unique shape of the nonlinear fit captures a distinct high positive bias group of forecasts for the HWFI model. For the storms forecasted to have an ALAT between approximately 13°N and 22°N, ALAT has a negative correlation with bias. However, for ALAT between roughly 23°N and 25°N, there is a noticeable positive jump (second peak is outside axes’ limits) in the dashed red line. This latitude range is frequented by TCs
and many of them interact with landmasses such as Cuba, the Yucatan Peninsula, and the Bahamas. A majority of the high positive bias cases in this ALAT range represent storms that either weakened over land more than HWFI expected or unexpectedly interacted with land. Above 25°N, ALAT has a small positive correlation with bias but there is basically no trend. The verification rules followed in this study as well as at NHC provide justification for the lack of high positive or negative bias events for these long-range, high-latitude HWFI forecasts. When a TC makes landfall, it typically decays and loses its tropical attributes within one or two days. At this point, a TC will be listed as either a “LO” or not listed in the best track file. In both these situations, verification of the TC intensity forecast is not possible and the high bias event is avoided.

For the remainder of this chapter, PRIME and R-PRIME forecasts are produced using the predictors in chapter 3 as well as the additional fitted predictors discussed here. As a result, the LGEM predictor pool for R-PRIME AE forecasts was comprised of the original 32 predictors in table 1.1, while the other models also have fitted FINT in their predictor pool. For all models, R-PRIME bias forecasts used the original 32 predictors and the fitted LDIS predictor discussed in chapter 3. R-PRIME bias forecasts for DSHP included 0DIV fitted with a third-order polynomial as an additional predictor. R-PRIME bias forecasts for HWFI also used this predictor as well as ALAT fitted with a two-peak Gaussian function. The same parameters involved in R-PRIME forecasts of AE are utilized in PRIME forecasts of AE. PRIME bias forecasts use the same predictor pool discussed in chapter 3 with the addition of fitted 0DIV for forecasts of DSHP bias.
4.3 PRIME and R-PRIME Bias-corrected Models

Following the methodology defined in chapter 3, optimal predictors were selected for each model, version of PRIME, and predictand. Using the updated PRIME and R-PRIME predictor pools, AE and bias forecasts were respectively created for the cases listed in table 3.1 and 3.2. As in chapter 3, cross validation was used to independently verify PRIME and R-PRIME forecasts. Although the new fitted predictors improved certain error forecasts, the main conclusions presented in chapter 3 remain largely unchanged.

Figure 4.4 shows the 2007-2014 average AE of DSHP, LGEM, GHMI, and HWFI real-time intensity forecasts with and without PRIME bias corrections. The plotted dashed lines represent the AE of the original model forecasts located in the a-deck files while the solid lines represent the AE of the model forecasts that are bias-corrected with PRIME. The data used to construct the dashed lines originate from table 3.1. For all forecast intervals, the AE of the bias-corrected models was significantly lower than the AE of the original models. The solid lines are grouped together because the weighting coefficient of the DFEM predictor is decisively the largest in the PRIME bias regression formula for all models and forecast hours. Although table 3.7 shows DFEM is also the most influential predictor for R-PRIME, the magnitude of its weighting coefficients (not shown) in the formulation of PRIME is considerably larger. As a result, when a model forecasts an intensity that deviates from the ensemble mean, the PRIME bias-corrected forecast will typically move closer to the ensemble mean.

Figure 4.5 is similar (note the y-axis limits are different) to figure 4.4 but it shows 2008-2014 average AE of DSHP, LGEM, GHMI, and HWFI retrospective intensity forecasts with and without R-PRIME bias corrections. Besides the 96- and 108-hour
HWFI bias-corrected forecasts, the AE of the bias-corrected forecasts are significantly lower than the original forecasts. At longer forecast hours, the performance of the different bias-corrected retrospective models varies much more than what was shown in figure 4.4. This divergence in model performance for longer forecast hours is likely due to other predictors rivalling DFEM’s importance in the R-PRIME regression formulas. The performance of bias-corrected DSHP is particular noticeable for 84-120 hour forecasts, where it had significantly lower AE than the other bias-corrected models.

In figure 4.5, the AE of LGEM and DSHP was clearly reduced the most because of the R-PRIME bias corrections. Chapter 3 discussed some of the reasons that PRIME forecasts for the statistical models improved more than those for the dynamical models when retrospective forecasts were used. It is also important to highlight that both DSHP and LGEM are developed with GFS dynamical fields, which are used to calculate the synoptic parameters that are predictors for all models in PRIME forecasts. Replacing the GFS predictors with the analogous forecasted dynamical fields from the HWFI and GHMI models could theoretically increase the skill of PRIME forecasts of HWFI and GHMI AE and bias.

4.4 PRIME and R-PRIME Modified Ensembles

Using PRIME and R-PRIME output, several modifications were applied to ICON, the standard equally-weighted ensemble of DSHP, LGEM, GHMI, and HWFI. For the first attempt at creating a better ensemble, a bias correction was applied to the model that was forecasted by PRIME to have the highest bias. Then, the ensemble mean was calculated to create the “Correct Worst” ensemble. The second unique ensemble formulated was called the “ Exclude Worst” ensemble. For this ensemble, PRIME forecasts were used to
isolate the model with the highest error and that model was excluded from the calculation of the ensemble mean. An Exclude Worst ensemble was created based on PRIME AE (Exclude Worst [AE]) and bias (Exclude Worst [bias]) predictions. The third experimental ensemble was constructed by taking the mean of all four model forecasts after they were corrected with PRIME bias forecasts. The resulting ensemble was called the “Bias-Corrected” ensemble.

The final three ensembles involved unequal weighting of the different models based on PRIME forecasts. Models were weighted inversely proportional to PRIME forecast error. For the “Unequal (AE)” ensemble, PRIME AE forecasts (PRIME_AE) quantified the confidence placed in each model with the equation:

$$ W_m = \frac{1}{\sum_{m=1}^{M} \text{PRIME}_A E_m} \cdot \frac{1}{\text{PRIME}_A E_m} . $$  \hspace{1cm} (5) 

Here, $M$ is equal to 4, the number of models. After calculating the weight, $W$, for each model $m$, the Unequal (AE) ensemble forecast intensity is calculated with:

$$ \text{FINT}_{\text{UNEQUAL (AE)}} = \sum_{m=1}^{M} W_m \times \text{FINT}_m . $$  \hspace{1cm} (6) 

Equation (6) shows each model’s forecast intensity, $FINT$, is multiplied by its computed $W$ to obtain Unequal (AE) ensemble forecast intensity, $FINT_{\text{Unequal (AE)}}$. A similar ensemble, “Unequal (AE SQR)”, was computed using (5) and (6), except both the numerator and denominator of (5) were squared. This alteration adds additional weight to the models that PRIME expects to have the lowest AE. The third unequally-weighted ensemble, “Unequal (BIAS)”, used PRIME bias forecasts. To create this ensemble, (5) and (6) were followed but the absolute value of the PRIME bias forecast replaced
PRIME_AE in (5). Intuitively, the absolute value of the PRIME bias forecast should be equivalent to the PRIME AE forecast but it is important to remember these two forecasts are calculated with separate regression formulas.

4.4.a PRIME Results

Figure 4.6 shows the AE of the Bias-Corrected ensemble, Exclude Worst (AE) ensemble, Correct Worst ensemble, and ICON forecasts at all forecast hours. The first three ensembles were modified using PRIME forecasts of AE and bias. All of these ensembles were formulated using real-time forecasts of DSHP, LGEM, GHMI, and HWFI from 2007-2014 and verified with best track information. The Exclude Worst (bias) ensemble was omitted from the figure to avoid additional clutter and because it was outperformed by the Exclude Worst (AE) ensemble at all forecast hours. The Bias-Corrected ensemble provided the forecasts with the lowest AE at all forecast intervals but was especially accurate for 96-120 hour forecasts. The Exclude Worst (AE) and Correct Worst ensembles only performed better than ICON at 108 and 120 hours.

Figure 4.7 displays the AE of ICON and the Unequal (AE), Unequal (AE SQR), and Unequal (BIAS) ensembles. At all forecast hours, the unequally-weighted ensembles have lower AE in their forecasts than the Exclude Worst (AE) and Correct Worst ensembles in figure 4.6. The Unequal (BIAS) ensemble provides the most accurate forecasts at every forecast hour except 96 and 108 hours. However, for all forecast hours, it has lower errors than ICON. Excluding the 96-hour forecasts, the 72-120 hour unequally-weighted ensemble forecasts show the most improvement over ICON forecasts.
The AE of ICON and the three best PRIME modified ensembles, the Bias-Corrected, Unequal (AE), and Unequal (BIAS) ensemble are plotted in figure 4.8. To better highlight the differences between the AE of these ensembles and ICON at the later forecast hours, only the AE of 48-120 hour forecasts are included. The Bias-Corrected ensemble has the lowest AE for all forecasts except 84-hour forecasts where the Unequal (BIAS) ensemble has the lowest. The AE of the Bias-Corrected ensemble is significantly (at the 92% level or higher) lower than the AE of ICON at 12, 24, 72, 108, and 120 hours. Combined, the other modified ensembles are significant at this level for only one forecast hour. The superior performance of the Bias-Corrected ensemble is expected because chapter 3 results demonstrated that PRIME bias forecasts are considerably better than its AE forecasts.

Figures 4.5-4.8 illustrate that when PRIME forecasts were used to correct or weight ensemble members, the AE of TC intensity forecasts were almost always lower than ICON. However, among all the PRIME modified ensembles, the largest observed improvement for a forecast hour was only a 5% decrease in AE (Bias-Corrected ensemble at 120 hours). This percentage decrease is statistically significant but end users would benefit if they could know situations beforehand where PRIME modified ensemble forecasts showed additional improvement over ICON. Confirming the results of DelSole et al. (2012), PRIME modified ensembles were the most skillful when there was higher variance between the forecasts of the ensemble members. Under these conditions, PRIME often isolated a model or two models that had more likely forecast solutions and the resulting forecast differed more from ICON.
Figures 4.9 and 4.10 provide examples of how the Bias-Corrected ensemble was especially valuable when its forecasts differ more from ICON. In figure 4.9, the AE of the ICON and Bias-Corrected ensemble are plotted for cases when their forecasts differ by more than 2.5 and 5 knots. For the 2.5 knots cases, ICON is referred to as ICON (2.5) and the Bias-Corrected ensemble is referred to as Bias-Corrected (2.5). Analogous naming procedures are used for the ensemble forecasts that differ by 5 knots. These values were subjectively selected to maintain large sample sizes while still portraying how larger discrepancies between the forecasts lead to larger differences in the performance of these ensembles. The dashed lines plotted in figure 4.9 represent the AE of ICON averaged over 2007-2014 while the solid lines provide the same information for the Bias-Corrected ensemble. Table 4.1 provides the percentage of the real-time forecasts where the Bias-Corrected ensemble and ICON differ by 2.5 and 5 knots. All forecast intervals containing at least 5% of the verified forecast cases between 2007 and 2014 are plotted in figure 4.9.

Compared to the original Bias-Corrected ensemble, the Bias-Corrected (2.5) and Bias-Corrected (5) ensembles showed greater improvement upon ICON (2.5) and ICON (5) at every forecast interval. The gap in performance between ICON and the Bias-Corrected ensemble was further increased when the difference in their forecasts was extended from 2.5 to 5 knots. From 48-120 hours, Bias-Corrected (5) improved upon ICON (5) between 3.5-11.3%. These forecast hours contained between 22% and 38% of the total forecasts, which allowed statistical significance to be established at each forecast hour.
To further emphasize the trend showed in figure 4.9, figure 4.10 graphs the AE of ICON and the PRIME Bias-Corrected ensemble when their forecasts differ by more than 10 knots. Table 4.1 indicates there are fewer cases where the forecasts of ICON and the Bias-Corrected ensemble differ by more than 10 knots. Only AE values for forecast hours between 48 and 120 are plotted because they contain at least 5% of the total forecasts. However, for these forecast hours, the Bias-Corrected (10) ensemble achieves significantly lower AE values that differ from ICON (10) by 13.0%-30.9%. Figures 4.9 and 4.10 demonstrate that the PRIME modified ensemble could be particularly useful in situations where it is forecasting an intensity that differs considerably from ICON.

4.4.b R-PRIME Results

Much like figure 4.6, figure 4.11 shows the AE of the Bias-Corrected ensemble, Exclude Worst (AE) ensemble, Correct Worst ensemble, and ICON forecasts at all forecast hours. For this figure, R-PRIME is used instead of PRIME to modify the ensembles and ensembles were formulated using retrospective forecasts from 2008-2014 instead of real-time forecasts. The Exclude Worst (bias) ensemble was omitted from the figure for the same reasons stated for figure 4.6. Depending on the forecast hour, either the Bias-Corrected ensemble or the Exclude Worst (AE) ensemble provided the forecasts with the lowest AE. The Correct Worst ensemble forecasts always verified with higher AE than ICON. In fact, there were no forecast intervals in figure 4.6 where an R-PRIME modified ensemble recorded an AE that was significantly lower than ICON AE.

Figure 4.12 displays the AE of ICON and the unequally-weighted ensembles: Unequal (AE), Unequal (AE SQR), and Unequal (BIAS). The Unequal (BIAS) ensemble is the least accurate unequally-weighted ensemble and has higher AE than ICON between
96 and 120 hours. In chapter 3, R-PRIME forecasts of bias were not as skillful as PRIME forecasts of bias so the observed deterioration in performance for the Unequal (BIAS) ensemble is not surprising. Excluding 12-hour forecasts, the Unequal (AE) and Unequal (AE SQR) ensembles have lower AE than ICON. However, these ensembles never achieve AE values that are significantly different from ICON. The SSs for R-PRIME forecasts were often lower than the SSs for PRIME forecasts, which explain the discrepancy in performance of the modified ensembles in figure 4.12 and figure 4.7. Additionally, as mentioned in chapter 3, the variance in the performance of the retrospective forecasts was considerably lower than what was observed for the real-time forecasts. As a result, the retrospective intensity forecasts produced by the 2014 version of the different models often showed more agreement and were more accurate, leaving less room for R-PRIME to detect error-prone situations.

Using the R-PRIME forecasts, the Bias-Corrected, Unequal (AE), and Unequal (AE SQR) ensembles performed the best. The AE of these ensembles as well as ICON are plotted for 48-120 hour forecasts in figure 4.13. Averaged over all forecast hours, the Unequal (AE SQR) ensemble had the lowest AE but the performance of the Unequal (AE) ensemble was comparable. The unequally-weighted ensembles showed the most improvement over PRIME for 84-120 hour forecasts.

Following the methodology used for PRIME modified ensembles, the best R-PRIME modified ensemble, Unequal (AE SQR) ensemble, was compared to ICON for forecast situations when the two ensembles diverged. In figure 4.14, the AE of the ICON and Unequal (AE SQR) ensembles are plotted for cases when their forecasts differ by more than 1.5 and 3 knots. The selected thresholds for analysis were smaller than those used in
figure 4.9 because the ICON and Unequal (AE SQR) ensembles’ forecasts rarely showed large deviations from each other\footnote{For a potential explanation, see discussion in chapter 3 and at the end of the previous paragraph.}. Figure 4.14 followed the same naming procedures used in figure 4.9. In figure 4.14, the dashed lines represent the AE of retrospective ICON averaged over 2008-2014 while the solid lines represent the AE of the Unequal (AE SQR) ensemble. Table 4.2 provides the percentage of the total retrospective forecasts where the Unequal (AE SQR) ensemble and ICON differ by 1.5 and 3 knots. All forecast intervals containing at least 5% of the total verified retrospective forecasts from 2008-2014 are plotted in figure 4.14.

At every forecast hour, the difference in the AE of the Unequal (AE SQR) (1.5) ensemble and ICON (1.5) was larger than the corresponding difference in the original ensembles. With the exception of the 60-hour forecasts, the Unequal (AE SQR) (3) ensemble accomplished the same feat. In general, the Unequal (AE SQR) (3) ensemble was more skillful (relative to ICON [3]) than the Unequal (AE SQR) (1.5) ensemble. However, the sample sizes for the 24-120 hour forecasts were not large enough to establish statistical significance between Unequal (AE SQR) (3) and ICON (3) or Unequal (AE SQR) (1.5) and ICON (1.5). When the two ensembles differed by more than 4, 5, etc. knots (not shown), Unequal (AE SQR) showed even more improvement over ICON but the sample sizes became very small.

4.5 Summary and Conclusions

PRIME forecasts were originally formulated to provide situation-dependent guidance for the expected bias and AE of the Atlantic basin TC intensity models: DSHP,
LGEM, GHMI, and HWFI. PRIME only uses information that is available before the official forecast time which means it is available to post-process the intensity forecasts of these models. The skill displayed by PRIME in chapter 3 coupled with it being available in real time implies it could serve as a tool to lower TC intensity forecast error. PRIME and R-PRIME bias forecasts were first tested for their ability to bias correct real-time and retrospective intensity forecasts. The PRIME bias-corrected models achieved AE values that were significantly lower than the original models’ AE values at all forecast hours, while R-PRIME accomplished the same feat, except for two forecast hours, with the retrospective models.

With the hope of improving upon the equally-weighted ensemble (ICON) of DSHP, LGEM, GHMI, and HWFI, PRIME error forecasts were also used to create unique ensembles of the four models. PRIME was tested for two different kinds of ensemble modifications. The first group of ensembles involved alterations to the individual models before taking the ensemble mean. The second group of ensembles used unequally-weighted ensemble members. For the real-time models, the best ensemble adjusted with PRIME error predictions was the Bias-Corrected ensemble. This ensemble was created by using PRIME bias forecasts to correct each model before the ensemble mean was calculated. The Bias-Corrected ensemble verified with lower AE than ICON at every forecast interval and for five forecast hours, it was significantly lower than ICON at the 92% level or higher. Using R-PRIME modifications for the retrospective models, the Unequal (AE SQR) ensemble performed the best. This ensemble weights models inverse-proportionally to their forecasted AE but the added exponent results in larger weights for the predicted stronger models. Excluding 12-hour forecasts, this ensemble outperformed
retrospective ICON at every forecast interval. However, the magnitude of improvement was not significant at any forecast interval.

The PRIME and R-PRIME modified ensembles provided the biggest upgrade over ICON when their forecasts deviated more from ICON forecasts. The PRIME Bias-Corrected ensemble forecasts had AE values up to 30% less than ICON when their forecasts differed by more than 10 knots. The Unequal (AE SQR) ensemble created with R-PRIME also showed additional improvement over retrospective ICON when its forecasts differed more from it. However, the Unequal (AE SQR) ensemble was observed to agree with ICON for more forecast cases because of the more accurate retrospective models. The performance of the PRIME and R-PRIME modified ensembles varied based on how well the PRIME and R-PRIME error forecasts performed. As a result, future work improving PRIME forecasts would likely lead to more accurate modified ensembles.
Fig. 4.1

96-hour retrospective DSHP transformed AE plotted against 96-hour DSHP forecast intensity (FINT) for the 2009-2014 Atlantic hurricane seasons before (top) and after (bottom) a third-order polynomial transformation of the predictor. The top figure shows the raw DSHP FINT values plotted against 96-hour retrospective DSHP transformed AE. The dashed red curve is fit using data from all forecast intervals during 2009-2014. The bottom image illustrates the relationship between transformed AE and fitted FINT. The solid blue line represents the linear fit to the data. The $R$ values in the top right corner of each plot capture the linear correlation of the independent and dependent variable.
Fig. 4.2

96-hour retrospective DSHP bias plotted against 0-hour divergence (0DIV) for the 2008-2013 Atlantic hurricane seasons before (top) and after (bottom) a third-order polynomial transformation of the predictor. The top figure shows the raw 0DIV values plotted against 96-hour retrospective DSHP bias. The dashed red curve is fit using data from all forecast intervals during 2008-2013. The bottom image displays the relationship between bias and fitted 0DIV. The solid blue line represents the linear fit to the data. The R values in the top right corner of each plot capture the linear correlation of the independent and dependent variable.
Fig. 4.3

108-hour retrospective HWFI bias plotted against forecast average latitude (ALAT) for the 2008-2012 and 2014 Atlantic hurricane seasons before (top) and after (bottom) a two-peak Gaussian transformation of the predictor. The top figure shows the raw ALAT values plotted against 108-hour retrospective HWFI bias. The dashed red curve is fit using data from all forecast intervals during 2008-2012 and 2014. The bottom figure shows the relationship between bias and fitted ALAT. The solid blue line represents the linear fit to the data. The $R$ values in the top right corner of each plot capture the linear correlation of the independent and dependent variable.
Fig. 4.4

The average absolute error of HWFI, GHMI, LGEM, and DSHP before and after PRIME bias corrections. The dashed lines represent the bias-corrected models and the solid lines represent the real-time models. Error statistics are calculated using cases from 2007-2014 that are listed in table 3.1.
Fig. 4.5

The average absolute error of retrospective HWFI, GHMI, LGEM, and DSHP before and after R-PRIME bias corrections. The dashed lines represent the bias-corrected models and the solid lines represent the original models. Error statistics are calculated using cases from 2008-2014 that are listed in table 3.2.
The average absolute error of the Bias-Corrected ensemble, Exclude Worst (AE) ensemble, Correct Worst ensemble, and ICON forecasts. These ensembles were formulated using real-time forecasts of DSHP, LGEM, GHMI, and HWFI from 2007-2014 and verified with best track information. The first three ensembles were modified using PRIME forecasts of AE and bias.
The average absolute error of ICON and the Unequal (AE), Unequal (AE SQR), and Unequal (BIAS) ensembles. These ensembles were formulated using real-time forecasts of DSHP, LGEM, GHMI, and HWFI from 2007-2014 and verified with best track information. The first three ensembles were unequally-weighted using PRIME forecasts of AE and bias.

Fig. 4.7
Fig. 4.8

The average absolute error of ICON and the Unequal (AE), Unequal (BIAS), and Bias-Corrected ensembles. These PRIME modified ensembles recorded the lowest AE. Only AE for 48-120 hour forecasts is plotted to capture how the AE of the ensembles separate at longer forecast hours.
The AE of ICON and the PRIME Bias-Corrected ensemble are plotted for cases when their forecasts differ by more than 2.5 and 5 knots. The 2.5 knots cases are indicated with black lines, and “(2.5)” is added to names of the ensembles. The 5 knots cases are indicated with red lines and “(5)” is added to the names of the ensembles. The dashed lines represent ICON forecasts and the solid lines represent Bias-Corrected ensemble forecasts. Table 4.1 lists the number of forecasts for the different cases at each forecast hour.
The AE of ICON and the PRIME Bias-Corrected ensemble are plotted for cases when their forecasts differ by more than 10 knots. The dashed lines represent ICON forecasts and the solid lines represent Bias-Corrected ensemble forecasts. The third column of Table 4.1 lists the number of forecasts at each forecast hour.
The average absolute error of the Bias-Corrected ensemble, Exclude Worst (AE) ensemble, Correct Worst ensemble, and ICON forecasts. These ensembles were formulated using retrospective forecasts of DSHP, LGEM, GHMI, and HWFI from 2008-2014 and verified with best track information. The first three ensembles were modified using R-PRIME forecasts of AE and bias.
Fig. 4.12

The average absolute error of ICON and the Unequal (AE), Unequal (AE SQR), and Unequal (BIAS) ensembles. These ensembles were formulated using retrospective forecasts of DSHP, LGEM, GHMI, and HWFI from 2008-2014 and verified with best track information. The first three ensembles were unequally weighted using R-PRIME forecasts of AE and bias.
The average absolute error of ICON and the Unequal (AE), Unequal (AE SQR), and Bias-Corrected ensembles. These R-PRIME modified ensembles recorded the lowest average AE during 2008-2014. Only AE for 48-120 hour forecasts is plotted to capture how the AE of the ensembles separate at longer forecast hours.
The AE of the ICON and R-PRIME Unequal (AE SQR) ensemble are plotted for cases when their forecasts differ by more than 1.5 and 3 knots. The 1.5 knots cases are indicated with black lines and “(1.5)” is added to names of the ensembles. The 3 knots cases are indicated with red lines and “(3)” is added to the names of the ensembles. The dashed lines represent ICON forecasts and the solid lines represent Unequal (AE SQR) ensemble forecasts. Table 4.2 lists the number of forecasts for the different cases at each forecast hour.
<table>
<thead>
<tr>
<th>Hours</th>
<th>(2.5)</th>
<th>(5)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>24</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>51</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>50</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>52</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>60</td>
<td>49</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>72</td>
<td>53</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>84</td>
<td>56</td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>96</td>
<td>55</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>108</td>
<td>61</td>
<td>34</td>
<td>15</td>
</tr>
<tr>
<td>120</td>
<td>66</td>
<td>38</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 4.1

The percent of the total real-time cases listed in table 3.1 where ICON and the PRIME Bias-Corrected ensemble differ by 2.5, 5, and 10 knots. Each column represents the magnitude of the difference between the intensity forecasts of ICON and the Bias-Corrected ensemble.
### Table 4.2

<table>
<thead>
<tr>
<th>Hours</th>
<th>(1.5)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>36</td>
<td>28</td>
<td>9</td>
</tr>
<tr>
<td>48</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>60</td>
<td>36</td>
<td>14</td>
</tr>
<tr>
<td>72</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>84</td>
<td>37</td>
<td>15</td>
</tr>
<tr>
<td>96</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td>108</td>
<td>37</td>
<td>17</td>
</tr>
<tr>
<td>120</td>
<td>36</td>
<td>16</td>
</tr>
</tbody>
</table>

The percent of the total retrospective cases listed in table 3.2 where ICON and the R-PRIME Unequal (AE SQR) ensemble differ by 1.5 and 3 knots. Each column represents the magnitude of the difference between the intensity forecasts of ICON and the Unequal (AE SQR) ensemble.
Chapter 5. Probabilistic PRIME (P-PRIME)

5.1 Background and Motivation

Forecasts of forecast skill are a useful tool for end users of weather forecasts but they are accompanied with their own uncertainty. Forecast skill predictions are commonly presented as deterministic with expected error given by a single value (see the reviews by Ehrendorfer 1997; Wobus and Kalnay 1995). This forecasting convention is a direct result of decision making based on natural hazards. Emergency managers and government officials are often forced into binary choices between evacuating and not evacuating, preparing and not preparing, or mitigating and not mitigating. In these situations, a deterministic forecast could lead to a more straightforward protocol, even though scientific information is hidden from the end users. However, several studies have emphasized this form of error quantification can be misleading (see review by NRC 2006). The atmosphere is inherently chaotic (Lorenz 1996) and as a result, a more suitable representation of its future state and the error in its future state is a probability distribution rather than a single value. Therefore, the production of a suite of error forecasts including both deterministic and probabilistic data appears to be necessary.

For operational TC forecast products, communication of forecast uncertainty varies depending on the predictand. NHC provides deterministic and probabilistic information in the cone of uncertainty graphic, which illustrates the forecasted track of a TC along with the uncertainty in the forecast. The size of the cone is based on the 5-year climatology of OFCL track forecast errors so its size does not change throughout the season. A similar graphic for TC intensity forecasts would be helpful but visualizing the
range of potential intensities is considerably more challenging than picturing the different directions a TC could travel. NHC attempts to capture the expected error in intensity forecasts with its wind speed probability graphic. However, this graphic combines track and intensity uncertainty information, and chapter 1 delineates the potential drawbacks in its current formulation. In this chapter, situation-dependent probabilistic forecasts of intensity are proposed. In contrast to the NHC cone of uncertainty, these forecasts will vary considerably throughout a hurricane season.

In previous chapters, the performance of PRIME, a deterministic error prediction technique for Atlantic basin TC intensity forecasts, was discussed. PRIME forecasts were produced using stepwise multiple linear regression and the forecasted quantities were bias and AE. Although PRIME was shown to skillfully capture the uncertainty associated with different forecast situations, the deterministic output might undermine PRIME performance and could be less helpful for end users. For example, there are situations where PRIME predicted an AE of 30 knots for a 48 hour intensity forecast and the actual AE was 50 knots. This PRIME forecast verifies with 20 knots of AE but its value is hidden by this seemingly large error. The large predicted AE of this particular PRIME forecast would effectively alert end users of a high uncertainty situation and consequently, helps forecasters and stakeholders understand there is potential for a variety of future TC intensities. Therefore, forecasting the probability of high or low AE and bias events rather than individual error values would likely be more useful for end users of intensity forecasts.

Similar to chapter 3, the current chapter reports the performance of forecasts of AE and bias for LGEM, DSHP, GHMI, and HWFI. However, multiple logistic regression
(Wilks 2005) is used here instead of multiple linear regression to produce probabilistic PRIME (P-PRIME) forecasts. P-PRIME forecasts are also evaluated at each 12-hour interval from 12 to 120 hours during the 2007-2014 Atlantic hurricane seasons. Only retrospective forecasts are used to develop P-PRIME because chapter 3 shows that R-PRIME and PRIME have similar behavior while R-PRIME produces forecasts with lower error. The predictor pool discussed in chapter 4 that was used for the development of the version of R-PRIME with the lowest errors is also used to develop P-PRIME for bias and AE forecasts.

Both binary and multinomial logistic regressions are conducted to test whether the binning of the predictand affects the skill of P-PRIME. For the binary logistic regression, a threshold is set for the predictand to transform AE or bias values into a dichotomous variable. For the multinomial logistic regression, two thresholds are used to convert the predictand into a ternary variable. Logistic regression is then applied to the dichotomous or trichotomous predictand. The exact details of the binning procedure as well as the logistic regression are discussed in section 5.2. After training P-PRIME for AE and bias forecasts, its errors are compared to the errors in the climatological forecasts of each predictand to determine the skill of P-PRIME. In order to simulate real-time performance, the results focus on independent verification calculated using cross validation (Wilks 2011).

The Brier Skill Score (BSS), Ranked Probability Skill Score (RPSS), reliability diagrams, and sharpness diagrams are used to assess the performance of P-PRIME. The P-PRIME forecasts created with binary logistic regression display positive skill for all models, forecast hours, and predictands. For the probabilistic forecasts created with
multinomial logistic regression, P-PRIME also has positive skill for all models, forecast hours, and predictands. Furthermore, by examining the lower and upper tercile of these probabilistic forecasts, considerably higher positive skill values are found. This result suggests P-PRIME could be particularly useful for forecast guidance when anomalous error behavior occurs. Coupling these results with the encouraging reliability and sharpness diagrams, it appears that P-PRIME would serve as a valuable operational tool to the TC forecasting community.

The data used to develop P-PRIME are described in section 3.2 and will not be discussed in this chapter. In section 5.2, the methodology for developing the logistic regression models is outlined. Section 5.3 shows the performance of P-PRIME using independent verification. Conclusions and future work to improve P-PRIME are presented in section 5.4.

5.2 Methodology

One of the benefits of using statistical methods instead of ensemble methods to forecast intensity forecast error is they allow easy quantification of the uncertainty associated with different forecast situations. The multiple linear regression scheme followed for PRIME forecasts in chapter 3 can be extended to produce the probability distribution associated with forecasted bias or AE. Using equation 9.41 in Wilks (2005), it is possible to convert the PRIME deterministic forecasts into probabilistic forecasts. However, this method assumes linear relationships between the independent and dependent variables, normally-distributed independent and dependent variables, and homoscedastic residuals. Although chapter 3 demonstrates that the data can be modified
to meet all of these assumptions, a forecasting system that does not require adjustments, which are inherently imperfect, is preferable.

Another approach to generating probabilistic forecasts using a linear regression framework is called Regression Estimation of Event Probabilities (REEP) (Glahn 1985). For this technique, the predictand is converted into a binary event and its only possible values become zero and one (indicating the predictand is below or above a particular threshold). Then, a standard multiple linear regression scheme can be trained and the resulting predicted values represent the probability the chosen threshold is exceeded. REEP has its own drawbacks. The regression formula with REEP can generate predicted probabilities outside of the 0 to 1 range which are not physically justified. Additionally, by applying a linear regression onto a binary predictand, the resulting residuals are not homoscedastic (Wilks 2005). Finally, there is no clear extension of REEP for forecast problems with more than two possible outcomes.

To avoid the deficiencies in the previously mentioned techniques, a multiple logistic regression model was implemented to develop P-PRIME. Logistic regression does not share the same assumptions regarding the normality, linearity, and homoscedasticity of the predictors and predictands. As a result, the Box-Cox transformation implemented in chapter 3 for AE data is not necessary prior to logistic regression. Logistic regressions that are fit to binary predictands follow the nonlinear equation:

$$y = \frac{1}{1 + \exp(-b_0 - b_1x_1 - b_2x_2 \cdots - b_kx_k)} \quad .$$  \hspace{1cm} (7)

The variables are the same as (2) but the predictand \(y\) is the probability of exceeding a bias or AE threshold rather than a bias or AE value. By looking at equation (7), it is clear that the predictand values are bounded zero and one, and the equation allows for each
predictor to be fit in a nonlinear way to the predictand. A statistical textbook such as Wilks (2005) can be consulted for a description of the log likelihood maximization procedure that is used to find the regression parameters. P-PRIME uses analogous equations for a predictand composed of three ordinal responses, and equation 2 in Ananth and Kleinbaum (1997) can serve as a guide for these calculations.

Before producing probability forecasts with multiple logistic regression, the predictand was transformed into either a binary or ternary variable. For binary P-PRIME AE forecasts, a threshold of 20 knots was set to transform AE below 20 knots to a 0 and AE above 20 knots to a 1. Although the selection of the threshold was subjective, multiple values were tested to separate high and low error events. For consistency, the same threshold was used for all forecast hours and models, which necessitated an AE value that was large enough to be considered a poor forecast for longer forecast intervals and the lesser models. Additionally, the threshold could not be too large where P-PRIME would rarely forecast AE values exceeding it. Two thresholds, 10 and 20 knots, were used for ternary P-PRIME AE forecasts. The probability of AE falling between 0 and 10 knots, 10 and 20 knots, and greater than 20 knots was computed. These bin ranges respectively correspond to good, average, and bad forecasts.

For binary P-PRIME bias forecasts, a threshold of 0 knots of bias was used to transform negative bias to a 0 and positive bias to a 1. This threshold was an obvious choice, because forecasters want to know the sign of the bias of an intensity forecast, and based on climatology, negative and positive biases are almost equally likely. The ternary P-PRIME bias forecasts’ thresholds were set at -15 knots of bias and 15 knots of bias. Therefore, the probability of bias less than -15 knots, between -15 and 15 knots, and
greater than 15 knots was calculated. The bias ranges separate high negative bias events, low bias events, and high positive bias events. For consistency with the AE forecasts, thresholds of -20 and 20 knots were also tested but resulted in a disproportionate amount of cases in the middle bin.

Although this chapter’s selection of thresholds was mainly a qualitative process, other studies have demonstrated that thresholds can be selected to optimize forecast performance (Wilks 2005; Splitt et al. 2014). However, the optimal threshold for a particular skill score does not always align with what is most beneficial for users of the forecast. Using a fictitious example, the skill scores of P-PRIME binary forecasts of bias could be maximized when the bias threshold is set to 30 knots. Probabilistic forecasts using this threshold would have little value because 30 knots of bias is rarely exceeded for all models and forecast hours. Therefore, the current thresholds could potentially be more useful for decision-making purposes because they maintain satisfactory sample sizes and separate meaningful error events. Regardless, future work will include thresholds that vary depending on the model and forecast hour because the definition of good and bad forecasts changes based on these circumstances.

P-PRIME was developed with the same a-deck intensity forecasts, best track verification, and predictor data as discussed in chapters 3 and 4. P-PRIME utilized solely retrospective forecasts and because the same criteria as chapter 3 was used to select storms for analysis, the sample sizes for P-PRIME were identical to those shown in table 3.2. The initial predictor pool included the fitted predictors from chapter 4 as well as the parameters shown in table 1.1. A slightly modified methodology compared to chapter 3 was followed to determine which of these predictors were optimal for each model and
predictand. Multiple logistic regression (instead of multiple linear regression) was applied to different combinations of the best predictors identified in the previous chapters, and the group that yielded the lowest average errors in the independent verification of the probabilistic predictions of forecast error was considered the optimal number. Similar to chapter 3, the optimal number of predictors varied, based on the model and predictand, between 2 and 8.

The dependent and independent variables were normalized before each regression to allow comparisons between the regression coefficients of different variables and forecast intervals. The best predictors for each model, predictand, and type of probabilistic forecast were almost identical to what is described in chapters 3 and 4. The predictor weighting coefficients (not shown) are not as intuitive as those involved with the linear regression because they represent the log of the odds ratio. In other words, if predictor $x_1$ in (7) corresponded to shear and $b_1$ equaled 1, then a shear value that was one standard deviation above the mean would lead to the odds of exceeding the threshold as $(e^1)$ approximately 2.716 times as likely. The weighting coefficients of the predictors and the corresponding odds ratios revealed the predictors displayed similar behavior to what was discussed in chapters 3 and 4. Therefore, a quantitative discussion of the predictors and weighting coefficients is not included in this chapter.

5.3 Independent Testing of P-PRIME

5.3.a P-PRIME Verification Tools

The performance of the P-PRIME probabilistic forecasts was evaluated using the BSS, RPSS, reliability diagrams, and sharpness diagrams. The cross-validation procedure
discussed in chapter 3 was followed for verification of the bias and AE predictions for both binary and ternary P-PRIME forecasts. As a result, the weighting coefficients of the predictors only incorporate information that would be available in real time and an operational version of P-PRIME should exhibit similar error statistics to the presented results. The Brier Score (BS; Brier 1950) is one of the most commonly used performance metrics for probability forecasts with dichotomous predictands and is calculated using the equation:

$$BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2 . \quad (8)$$

Here, $t$ represents an individual forecast, $f_t$ is the probability of an event occurring as forecasted by P-PRIME, $o_t$ is a variable that indicates whether the event occurred ($o_t = 1$) or not ($o_t = 0$), and $N$ is the number of forecasts that are evaluated. Smaller BS values represent more accurate forecasts with a perfect BS equaling 0.

BS is a single scalar quantity that captures the mean squared error for all forecast-observation pairs, which makes it hard to interpret. As a result, the BSS metric is used to compare the BS of P-PRIME to the BS of climatological forecasts. BSS is calculated as:

$$BSS = 100\% \times \frac{BS_{clim} - BS}{BS_{clim}} . \quad (9)$$

$BS_{clim}$ indicates the BS of forecasts that are based on the past probability of the event occurring. For a particular forecast hour and model, climatological forecasts are generated by calculating the mean and standard deviation of the training data. The distribution of the training data is assumed to be roughly Gaussian so the mean and
standard deviation are sufficient to calculate the probability of intensity error exceeding a threshold.

BSS has a more straightforward interpretation. If BSS is positive then the forecasting technique is improving upon climatology, and its magnitude indicates the degree of improvement. If the predictand has more than two possible outcomes, BS can be calculated for each bin individually by treating all other results not in the forecasted category as an \( o_i = 0 \) in (8). However, the ranked probability score (RPS) is a verification tool that is better suited to predictands that have more than two possible outcomes (Epstein 1969; Murphy 1970). RPS is preferred over BS for these situations because it is sensitive to the distance the forecasted quantity is from the observed quantity. For one forecast-observation pair, RPS is the sum of squared differences between the cumulative forecasted and observed event probabilities:

\[
RPS = \sum_{m=1}^{J} [Y_m - O_m]^2, \tag{10a}
\]

where

\[
Y_m = \sum_{i=1}^{m} f_i, \quad m=1,2, \ldots, J \tag{10b}
\]

and

\[
O_m = \sum_{i=1}^{m} o_i, \quad m=1,2, \ldots, J. \tag{10c}
\]

The cumulative forecast \( Y_m \) and observation \( O_m \) quantities are calculated in (10b) and (10c) with the same variables as (8) except with the addition of \( J \), which represents the number of predictand outcomes. Consider a hypothetical three-outcome example where \( f_1=0.5 \), \( f_2=0.3 \), and \( f_3=0.2 \). Also, \( o_2=1 \) which means \( o_1 \) and \( o_3=0 \). Then, \( Y_1=0.5 \), \( Y_2=0.8 \), and \( Y_3=1.0 \).
Y₂=0.8, Y₃=1, O₁=0, O₂=1, O₃=1, and RPS = 0.29 (0.25 + 0.04 + 0). A perfect RPS would have each Yₘ = Oₘ and would equal 0. In this chapter, RPS was averaged over all the forecast-observation pairs from 2008-2014 to produce one number that represented the performance of all P-PRIME forecasts with ternary predictands for a particular model, forecast interval, and predictand. Following the BSS example, RPSS is calculated like (9) by comparing P-PRIME RPS to climatology.

The final verification tool applied to P-PRIME is the reliability diagram (Wilks 2005). Unlike the RPSS or BSS, the reliability diagram provides a more holistic picture of the quality of a forecasting technique. The reliability diagram is a graphical device that shows how well the issued forecast probabilities align with observed event frequencies for different forecast probability bins. The first step in creating a reliability diagram is partitioning the forecast values into K non-overlapping bins. Ten equal-sized bins (B₁, B₂, …, B₁₀) were used for all of this chapter’s results, and the resulting bin ranges were 0-0.1, 0.1-0.2, …, 0.9-1.

Once a forecast is made, the forecasted probability identifies a bin, Bₖ, it is placed in. After grouping all forecast cases into bins, the average of all the forecast values in each Bₖ (for convenience, referred to as F_AVₖ) can be calculated. For all the forecast cases in Bₖ, it is also possible to calculate the frequency that the event was observed to occur (for convenience, referred to as O_FQₖ). Thus, O_FQₖ is simply the number of times the event happened divided by the number of total entries in bin Bₖ. Finally, the reliability diagram is produced by plotting F_AVₖ versus O_FQₖ for all bins. For perfect reliability, the forecast probability and the frequency of occurrence should be equal, and the plotted points should lie on the diagonal. For example, when P-PRIME predicts an 80%
probability that AE will exceed 20 knots then for perfect reliability, AE should exceed 20 knots on 80% of occasions on which the prediction is made. An accompanying graphic, called a sharpness diagram, was included with the reliability diagram to show the relative frequency of each forecast probability bin. A more detailed explanation of the reliability and sharpness diagrams is available in Wilks (2005).

To examine the average performance of P-PRIME for the 2008-2014 hurricane seasons, reliability and sharpness diagrams were created for all four models, binary and ternary predictands, bias and AE forecasts, and each of the ten forecast intervals. As a result, there were 40 reliability and sharpness diagrams produced for each model. Although reliability diagrams are designed for displaying the performance of binary predictands, the ternary scenario was addressed by repeating the calculations for each outcome. In these situations, a separate line showing the performance of P-PRIME for the highest and lowest tercile was incorporated into the reliability diagram. For all reliability diagrams, data were only plotted for bins with at least five entries to avoid misleading points on the graph. In order to stay concise with the probabilistic verification of P-PRIME, the RPSS and BSS metrics are used to summarize the overall P-PRIME performance and only a few reliability and sharpness diagrams are displayed.

5.3.b P-PRIME Binary Forecasts

Table 5.1 shows the 2008-2014 average BSSs of P-PRIME forecasts of AE greater than 20 knots. The BSS is positive for all models and forecast intervals. The number of cases for each forecast interval is equal to those listed in table 3.2. The BSSs show similar trends to the tables and figures in chapter 3 that presented the results of R-PRIME
deterministic AE forecasts. The BSSs of statistical models are higher than the dynamical models for a majority of the forecast intervals, and P-PRIME generally performs better for longer forecast intervals. In particular, P-PRIME performs best for forecasts of DSHP AE and worst for forecasts of HWFI AE. However, the 120-hour forecasts show a low BSS for all of the models, and the statistical models have a much lower BSS at this forecast interval compared to other long intervals. This trend is not visible in the SS of R-PRIME AE forecasts in chapter 3.

To gain insight on the unique performance of longer forecasts, figures 5.1a-5.1b and 5.2a-5.2b examine the 108- and 120-hour P-PRIME forecasts of DSHP AE. Figure 5.1a is the reliability diagram for P-PRIME forecasts of 108-hour DSHP AE. As expected, the P-PRIME forecasts show high resolution and calibration for all forecast probability bins. Figure 5.1b shows the sharpness diagram for the same forecasts as figure 5.1a. Approximately 50% of the forecast probabilities produced by P-PRIME fall between 0.1 and 0.2 or 0.2 and 0.3, which are close to the climatologically-most-likely forecast\textsuperscript{15} of 0.28. Still, 5.1b illustrates that almost 20% of 108-hour P-PRIME forecasts of DSHP AE exceed a probability of 0.5. Additionally, almost 15% of 108-hour P-PRIME forecasts of DSHP AE are in the 0-0.1 probability bin. The quality of these high-confidence\textsuperscript{16} forecasts enables forecasters and end users to trust P-PRIME when it signals a high likelihood or low likelihood of DSHP AE exceeding 20 knots. Therefore, figures 5.1a-5.1b justify the high BSS for this model-forecast hour combination.

\textsuperscript{15} Climatologically most-likely forecast is equivalent to the mean AE over the training dataset for a forecast hour-model combination.

\textsuperscript{16} Probabilistic forecasts in the low and high probability bins are called high confidence because P-PRIME is expressing high certainty that an event will or will not occur. High confidence in the error forecasts should not be confused with an expectation for a low-error intensity forecast.
Figures 5.2a-5.2b show the reliability and sharpness diagram for 120-hour P-PRIME forecasts of DSHP AE. For most probability bins in figure 5.2a, P-PRIME is well-calibrated. However, for the 28 forecasts with a probability between 0.6 and 0.9, P-PRIME has a high bias. In the 0.6-0.7, 0.7-0.8 bin, and 0.8-0.9 bin, P-PRIME predicted that AE will exceed 20 knots respectively 64%, 76%, and 84% of the time while the observed frequency of exceeding 20 knots for these bins was 50%, 50%, and 71%. Even though roughly 5% of P-PRIME forecasts fall in these bins, they are largely responsible for P-PRIME’s poor forecasting of 120-hour DSHP AE. Additionally, only 14% of P-PRIME forecasts predict a higher than 0.5 probability of 120-hour DSHP AE exceeding 20 knots. Therefore, P-PRIME has fewer and more inaccurate high-confidence forecasts for this forecast time, which negatively affects the reliability. Even though the analysis of figures 5.1a-5.1b and 5.2a-5.2b highlight which forecasts are causing the lower BSS of 120-hour P-PRIME DSHP forecasts, additional investigation is needed to understand the physical mechanisms leading to this P-PRIME behavior.

Table 5.2 shows the 2008-2014 average BSS of P-PRIME forecasts of bias greater than 0 knots. The BSS values are all positive and much larger than the BSS values of AE forecasts. Again, longer forecasts have higher BSS values, P-PRIME forecasts for statistical models outperform its forecasts for dynamical models, and the BSS values of DSHP and HWFI are the highest and lowest respectively. All of the forecast intervals for DSHP, LGEM, and GHMI have BSS values greater than 10% and with many forecast intervals over 15%, P-PRIME is unquestionably a valuable forecast tool. In chapter 3, potential explanations were proposed for the lower SSs reported for R-PRIME bias.
forecasts of HWFI. A reliability and sharpness diagram is included here to gain further insight.

Figures 5.3a-5.3b show the reliability and sharpness diagram for 96-hour P-PRIME forecasts of HWFI bias. This model-forecast hour pair was selected for analysis because it registered the lowest BSS. In figure 5.3a, the reliability diagram indicates that P-PRIME forecasts are fairly accurate for bins above 0.2. The observed low BSS is most likely caused by P-PRIME’s low bias for forecasts that range from 0.0-0.2. For the 35 forecasts in this probability range, P-PRIME forecasted a low likelihood of a positive bias, when in reality, positive bias events occurred 40-50% more often than expected. A future study will carefully analyze what predictors are suggesting a more negative bias than what is verified. Figure 5.3b highlights that most P-PRIME forecasts center around the climatologically-most-likely probability of 0.45. There is moderate spread in the forecast bins but an ideal probabilistic forecasting scheme would have more forecasts in the extreme bins.

5.3.c P-PRIME Ternary Forecasts

Table 5.3 shows the 2008-2014 average RPSS of P-PRIME multinomial forecasts of AE. For the ternary predictand, P-PRIME was designed to forecast the probability of AE appearing in three predicted ranges: less than 10 knots, between 10 and 20 knots, and greater than 20 knots. The RPSSs of all models and forecast intervals are positive, which suggests that the success of P-PRIME is not an artifact of fortuitous threshold selection. Except for 120-hour forecasts, P-PRIME is more skillful at the longer forecast intervals for the statistical models. The dynamical models have their higher RPSS entries at the
medium forecast intervals. The magnitude of the RPSSs for the dynamical models was slightly larger than their BSSs for the binary predictand, while the RPSSs of statistical models were slightly smaller.

Table 5.4 provides the 2008-2014 average RPSS of P-PRIME multinomial forecasts of bias. For the ternary forecasts, P-PRIME is trained to forecast the probability of bias that is less than -15 knots, between -15 and 15 knots, and greater than 15 knots. The RPSSs of all models and forecast intervals are positive and larger than the RPSSs of P-PRIME ternary forecasts of AE. The RPSSs of the P-PRIME bias forecasts for the ternary predictand are generally smaller than the BSSs shown for the binary predictand. Again, P-PRIME is more skillful for the statistical models compared to the dynamical models. For HWFI, LGEM, and DSHP, P-PRIME 84-hour forecasts have the highest RPSSs and for GHMI, 120-hour forecasts have the highest RPSS. There is no real trend in P-PRIME performance above 48 hours for any of the models.

In (10a), $Y_m$ and $O_m$ are both cumulative functions of probability components that must add to one, so the last term in the RPS sum is always 0. Only the first $J - 1 = 2$ terms in (10a) are needed to compute the RPS and RPSS and there is no framework available for checking the relative performance of the individual bins. As a result, the BSS metric was computed separately for each of the three AE events to test their relative accuracy. This analysis was repeated for all models and forecast intervals. As previously stated, the BSS can be misleading for the multiple predictand case because it does not penalize forecasts based on the distance the forecast probability is from the observed frequency. However, BSS is used here solely as a tool for comparing the relative performance of the different predictand events.
Table 5.5 shows the BSS of P-PRIME ternary forecasts of LGEM AE at all forecast intervals and for all predictand events. The BSS of each event outcome is represented with a column in the table. P-PRIME forecasts of LGEM AE exceeding 20 knots accomplished the highest BSS, and the forecasts of LGEM AE between 10 and 20 knots had the lowest BSS. P-PRIME forecasts of AE were consistently found to have less error in the extreme bins throughout the other models and forecast hours. Forecasters and end users would likely be concerned with the performance of P-PRIME in the outermost bins because these high and low error situations have a more straightforward interpretation for decision making. When a model is forecasted to likely have low errors (below 10 knots), the intensity forecast is viewed as reliable and more definitive preparations for a TC are possible. On the other hand, a model expected to have higher errors (greater than 20 knots) cannot be trusted, and it would be wise to wait for another forecast, focus on another model, or prepare for a range of possible hazards.

Figure 5.4a is a reliability diagram that shows 84-hour P-PRIME forecasts of LGEM AE greater than 20 knots (upper tercile: “U” on the plot) and lower than 10 knots (lower tercile: “L” on the plot). This forecast hour-model combination was selected for analysis because table 5.4 shows that it varies considerably among different terciles. For both of the extreme terciles, P-PRIME forecasts appear unbiased and well-calibrated. Figures 5.4b and 5.4c are the associated sharpness diagrams for the lower and upper tercile forecasts.

Figure 5.4b shows the sharpness diagram for 84-hour P-PRIME forecasts of LGEM AE below 10 knots of error. The bins spanning forecast probabilities between 0.3 and 0.6 contain approximately 60% of the total forecasts. These low-confidence error forecasts
are close to the climatologically-most-likely probability of 0.43. Ideally, a forecasting method should have more entries in the higher and low probability bins but as a consolation, figure 5.4a illustrates that P-PRIME produces reliable forecasts in these medium-confidence bins. Regardless, P-PRIME forecasts in this lower tercile have a lower BSS because climatological forecasts perform the best in this tercile (not shown).

Figure 5.4c shows the sharpness diagram for 84-hour P-PRIME forecasts of LGEM AE exceeding 20 knots of error. Based on climatology, the most-likely probability is 0.26. As a result, it is not surprising the 0.1-0.3 forecast probability range is heavily populated. Still, P-PRIME performs well for the forecasts with probabilities exceeding 0.5, while climatological forecasts are unable to issue high probabilities to high AE events. The information presented in figure 5.4a-5.4c and table 5.5 suggests that end users could hypothetically only use P-PRIME forecasts in the extreme probability terciles, and especially the upper tercile, if they wanted higher quality forecasts.

Table 5.6 is very similar to table 5.5 but shows the BSS of P-PRIME ternary forecasts of LGEM bias. P-PRIME forecasts of LGEM bias exceeding fifteen knots accomplished the highest BSS, and the forecasts of LGEM bias between -15 and 15 knots had the lowest BSS. This error behavior among the bins was consistent throughout the evaluation of all the models and a majority of the forecast hours. Figure 5.5a is a reliability diagram for 60-hour P-PRIME forecasts of LGEM bias greater than 15 knots (“U”) and less than -15 knots (“L”). Figure 5.5a confirms the P-PRIME forecasts in the upper and lower tercile are both reliable and well-calibrated. According to table 5.4, this model-forecast hour pair has the highest RPSS out of all models and forecast hours between 12 and 60 hours.
Figures 5.5b and 5.5c are the sharpness diagrams for the P-PRIME forecasts of 60-hour LGEM bias in the lower and upper tercile. Based on climatology, 0.19 is average probability that bias will fall in the upper tercile, and 0.25 is the average probability that bias will be in the lower tercile. P-PRIME forecasts have their highest bias, for either tercile, in the 0.5-0.6 bin for upper tercile forecasts. Fortunately, figure 5.5b shows that this bin has very few entries. Besides the poor performance in this bin and the lack of high-confidence forecasts in figures 5.5b and 5.5c, P-PRIME appears to be a very reliable and useful forecasting tool for this model-forecast hour pair.

5.4 Conclusions and Future Work

This chapter focused on the methodology and performance of a probabilistic forecasting system, Probabilistic PRIME (P-PRIME), for the error of Atlantic basin TC intensity forecasts. P-PRIME was developed to forecast the bias and AE of the DSHP, LGEM, GHMI, and HWFI models for the 2008-2014 hurricane seasons. The error predictions were applied to retrospective forecasts produced by the 2014 version of each model as a technique to maintain consistent model formulations throughout the time series. P-PRIME forecasts were produced using a multiple logistic regression scheme for both binary and ternary predictands. P-PRIME was an extension of the PRIME model discussed in chapters 3 and 4, and therefore the independent variables in the regression formula were the same as those discussed in those chapters. Additionally, the selection of the optimal predictors and methodology for creating P-PRIME closely follows the procedure outlined in chapter 3.
P-PRIME forecasts were evaluated using a combination of BSS, RPSS, reliability diagrams, and sharpness diagrams. The AE and bias climatology of the different models served as a benchmark forecast for the calculation of the skill scores. The BSSs of P-PRIME forecasts of AE and bias for the binary predictand case were positive for all forecast intervals and models. The P-PRIME bias forecasts achieved BSSs that frequently exceeded 15% and were much higher than the P-PRIME AE forecasts. For both bias and AE, P-PRIME achieved its highest BSSs for DSHP and its lowest BSSs for HWFI. The longer forecast intervals typically had larger BSSs than the shorter intervals but this trend was inconsistent among some of the longer forecast intervals.

The RPSSs of P-PRIME forecasts of AE and bias for the ternary predictand were also positive for all forecast intervals and models but were lower than the BSSs for the binary predictand case. The trends among the models and forecast intervals were generally the same for both the ternary and binary predictands. The BSSs of P-PRIME ternary forecasts of LGEM AE and bias were also presented to emphasize the lower and upper tercile had higher skill than the middle tercile. This result was particularly encouraging because it suggests P-PRIME is adept at forecasting the extreme error events.

The collection of reliability diagrams in this chapter highlighted the strength of P-PRIME. In general, the P-PRIME forecasts were both well-calibrated and unbiased. The associated sharpness diagrams showed that P-PRIME could improve by issuing more high-confidence forecasts. Still, in its current form, P-PRIME generated forecasts that deviated considerably from the climatologically-most-likely probability.
Although the presented results for P-PRIME forecasts appear promising, this chapter represents a preliminary investigation into probabilistic forecasts of TC intensity forecast error. P-PRIME could be improved by catering some of the predictors more to the logistic regression scheme. In this chapter, the optimal predictors are selected using almost the same methodology as PRIME. However, the logistic regression does not benefit from the polynomial and Gaussian transformations of the predictors. Logistic relationships between the predictors and predictand are preferred over high linear correlations. In future work, different ways to transform predictors to better serve a logistic regression will be explored.

Secondly, the thresholds used for the conversion of the predictand to a binary or ternary variable influence the performance of P-PRIME (Wilks 2005). The different TC models and forecast intervals have different error characteristics so the definition of a high-error forecast varies. Thresholds can be defined more systematically by picking error values that partition the same amount of cases for the different models. For example, with a binary predictand, it might be better to choose an AE threshold that is the median (or top 33%, 20%, etc.) AE for each model and forecast interval based on climatological error values. That way, the threshold mathematically has the same meaning throughout the models and forecast hours. In future iterations of P-PRIME, a variety of different thresholds will be tested to monitor how P-PRIME skill is affected.

Finally, the weighting coefficients of P-PRIME for different years and models need to be carefully analyzed. Initial results showed that the weighting coefficients indicate the predictors have relationships with error that are similar to those seen with the PRIME deterministic forecasts. However, the performance of P-PRIME fluctuates from year-to-
year and storm-to-storm and better understanding the origin of this variability would undoubtedly help increase the skill of P-PRIME.

P-PRIME provides a probabilistic alternative to PRIME when forecasting TC intensity forecast error. Using a probabilistic forecast system is beneficial to end users because it conveys the uncertainty in forecasts of the future state of the atmosphere. This chapter focused mainly on quantifying, not visualizing, the performance of probabilistic forecasts of AE and bias. Once a satisfactory version of P-PRIME is created, the next step is presenting the forecast output in a convenient way for end users. Ultimately, P-PRIME forecasts can augment the information available to decision-makers before a potential TC landfall.
Fig 5.1

(a) Reliability diagram for P-PRIME forecasts of 108-hour DSHP AE greater than 20 knots during the period 2008-2014. Dots show the observed frequency of AE exceeding 20 knots, conditional on each of the 10 possible probability forecast bins. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins.
Fig 5.2

(a) Reliability diagram for P-PRIME forecasts of 120-hour DSHP AE greater than 20 knots during the period 2008-2014. Dots show the observed frequency of AE exceeding 20 knots, conditional on each of the 10 possible probability forecast bins. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins.
(a) Reliability diagram for P-PRIME forecasts of 96-hour HWFI bias greater than 0 knots during the period 2008-2014. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins.
Fig 5.4

(a) Reliability diagram for P-PRIME forecasts of 84-hour LGEM AE greater than 20 knots (blue line: U) and AE lower than 10 knots (red line: L) during the period 2008-2014. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the lower tercile. (c) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the upper tercile.
Fig 5.5

(a) Reliability diagram for P-PRIME forecasts of 60-hour LGEM bias greater than 15 knots (blue line: U) and bias lower than 15 knots (red line: L) during the period 2008-2014. Bins with less than 5 entries are not included in the diagram. (b) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the lower tercile. (c) Sharpness diagram indicating the relative frequency of use of each of the forecast probability bins for the upper tercile.
<table>
<thead>
<tr>
<th># of Cases</th>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1182</td>
<td>12</td>
<td>0.8</td>
<td>0.8</td>
<td>2.8</td>
<td>0.8</td>
</tr>
<tr>
<td>1095</td>
<td>24</td>
<td>0.8</td>
<td>1.8</td>
<td>3.1</td>
<td>5.0</td>
</tr>
<tr>
<td>1012</td>
<td>36</td>
<td>4.3</td>
<td>9.1</td>
<td>4.4</td>
<td>6.1</td>
</tr>
<tr>
<td>924</td>
<td>48</td>
<td>1.8</td>
<td>7.9</td>
<td>6.6</td>
<td>8.4</td>
</tr>
<tr>
<td>838</td>
<td>60</td>
<td>3.8</td>
<td>7.5</td>
<td>9.1</td>
<td>10.9</td>
</tr>
<tr>
<td>758</td>
<td>72</td>
<td>3.1</td>
<td>7.0</td>
<td>9.3</td>
<td>9.9</td>
</tr>
<tr>
<td>679</td>
<td>84</td>
<td>5.1</td>
<td>5.6</td>
<td>12.7</td>
<td>18.5</td>
</tr>
<tr>
<td>614</td>
<td>96</td>
<td>3.6</td>
<td>2.0</td>
<td>14.8</td>
<td>18.3</td>
</tr>
<tr>
<td>547</td>
<td>108</td>
<td>4.8</td>
<td>3.7</td>
<td>10.5</td>
<td>18.1</td>
</tr>
<tr>
<td>492</td>
<td>120</td>
<td>5.5</td>
<td>4.8</td>
<td>6.6</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Table 5.1

The Brier Skill Scores of P-PRIME binary forecasts of AE greater than 20 knots for each model and forecast interval. These results apply to retrospective forecasts, which were generated by applying the 2014 version of each model to the 2008-2014 hurricane seasons.
<table>
<thead>
<tr>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>7.7</td>
<td>12.6</td>
<td>11.8</td>
<td>10.2</td>
</tr>
<tr>
<td>24</td>
<td>12.0</td>
<td>13.1</td>
<td>15.1</td>
<td>16.3</td>
</tr>
<tr>
<td>36</td>
<td>8.9</td>
<td>14.1</td>
<td>20.8</td>
<td>19.1</td>
</tr>
<tr>
<td>48</td>
<td>6.1</td>
<td>17.6</td>
<td>19.3</td>
<td>19.1</td>
</tr>
<tr>
<td>60</td>
<td>6.2</td>
<td>14.1</td>
<td>20.8</td>
<td>19.6</td>
</tr>
<tr>
<td>72</td>
<td>7.3</td>
<td>16.9</td>
<td>16.3</td>
<td>19.5</td>
</tr>
<tr>
<td>84</td>
<td>4.8</td>
<td>18.0</td>
<td>13.1</td>
<td>16.8</td>
</tr>
<tr>
<td>96</td>
<td>4.3</td>
<td>20.1</td>
<td>14.4</td>
<td>22.7</td>
</tr>
<tr>
<td>108</td>
<td>4.9</td>
<td>19.8</td>
<td>15.2</td>
<td>29.9</td>
</tr>
<tr>
<td>120</td>
<td>9.0</td>
<td>19.6</td>
<td>18.9</td>
<td>30.9</td>
</tr>
</tbody>
</table>

Table 5.2

Similar to table 5.1 except the Brier Skill Scores apply to P-PRIME forecasts of bias greater than 0 knots.
<table>
<thead>
<tr>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>6.1</td>
<td>3.2</td>
<td>3.9</td>
<td>6.0</td>
</tr>
<tr>
<td>24</td>
<td>5.1</td>
<td>2.7</td>
<td>4.9</td>
<td>9.5</td>
</tr>
<tr>
<td>36</td>
<td>7.0</td>
<td>7.6</td>
<td>4.4</td>
<td>8.7</td>
</tr>
<tr>
<td>48</td>
<td>6.2</td>
<td>7.2</td>
<td>5.6</td>
<td>6.5</td>
</tr>
<tr>
<td>60</td>
<td>6.3</td>
<td>7.3</td>
<td>6.3</td>
<td>10.1</td>
</tr>
<tr>
<td>72</td>
<td>4.4</td>
<td>8.1</td>
<td>7.2</td>
<td>9.6</td>
</tr>
<tr>
<td>84</td>
<td>3.8</td>
<td>6.5</td>
<td>8.2</td>
<td>14.1</td>
</tr>
<tr>
<td>96</td>
<td>2.7</td>
<td>4.8</td>
<td>9.2</td>
<td>13.9</td>
</tr>
<tr>
<td>108</td>
<td>4.9</td>
<td>3.8</td>
<td>8.0</td>
<td>11.4</td>
</tr>
<tr>
<td>120</td>
<td>5.8</td>
<td>4.1</td>
<td>6.5</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Table 5.3

The Ranked Probability Skill Scores of P-PRIME ternary forecasts of AE for each model and forecast interval. The thresholds set for the predictand were 10 and 20 knots of AE.
<table>
<thead>
<tr>
<th>Hours</th>
<th>HWFI</th>
<th>GHMI</th>
<th>LGEM</th>
<th>DSHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>5.3</td>
<td>7.1</td>
<td>5.8</td>
<td>7.2</td>
</tr>
<tr>
<td>24</td>
<td>4.0</td>
<td>8.1</td>
<td>12.5</td>
<td>13.2</td>
</tr>
<tr>
<td>36</td>
<td>6.6</td>
<td>15.6</td>
<td>13.4</td>
<td>15.4</td>
</tr>
<tr>
<td>48</td>
<td>4.4</td>
<td>12.7</td>
<td>13.5</td>
<td>13.3</td>
</tr>
<tr>
<td>60</td>
<td>2.8</td>
<td>9.4</td>
<td>16.1</td>
<td>15.4</td>
</tr>
<tr>
<td>72</td>
<td>4.6</td>
<td>10.1</td>
<td>17.6</td>
<td>19.0</td>
</tr>
<tr>
<td>84</td>
<td>8.9</td>
<td>12.8</td>
<td>18.7</td>
<td>20.4</td>
</tr>
<tr>
<td>96</td>
<td>5.2</td>
<td>12.8</td>
<td>15.4</td>
<td>19.4</td>
</tr>
<tr>
<td>108</td>
<td>2.2</td>
<td>16.3</td>
<td>10.0</td>
<td>13.8</td>
</tr>
<tr>
<td>120</td>
<td>6.5</td>
<td>19.0</td>
<td>10.5</td>
<td>11.3</td>
</tr>
</tbody>
</table>

Table 5.4

The Ranked Probability Skill Scores of P-PRIME ternary forecasts of bias for each model and forecast interval. The thresholds set for the predictand were -15 knots and 15 knots of bias.
<table>
<thead>
<tr>
<th>Hours</th>
<th>0 to 10</th>
<th>10 to 20</th>
<th>&gt;20</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>4.2</td>
<td>1.7</td>
<td>2.8</td>
</tr>
<tr>
<td>24</td>
<td>5.3</td>
<td>2.2</td>
<td>3.9</td>
</tr>
<tr>
<td>36</td>
<td>4.0</td>
<td>1.4</td>
<td>5.1</td>
</tr>
<tr>
<td>48</td>
<td>4.4</td>
<td>0.9</td>
<td>7.2</td>
</tr>
<tr>
<td>60</td>
<td>3.4</td>
<td>1.0</td>
<td>9.7</td>
</tr>
<tr>
<td>72</td>
<td>5.0</td>
<td>0.4</td>
<td>9.9</td>
</tr>
<tr>
<td>84</td>
<td>4.3</td>
<td>-0.3</td>
<td>12.9</td>
</tr>
<tr>
<td>96</td>
<td>4.5</td>
<td>0.9</td>
<td>14.9</td>
</tr>
<tr>
<td>108</td>
<td>5.6</td>
<td>-0.2</td>
<td>10.9</td>
</tr>
<tr>
<td>120</td>
<td>5.3</td>
<td>0.3</td>
<td>7.9</td>
</tr>
</tbody>
</table>

Table 5.5

The BSS of P-PRIME ternary forecasts of LGEM AE for each forecast interval. The BSS of each event outcome is represented with a column in the table.
<table>
<thead>
<tr>
<th>Hours</th>
<th>-15&gt;</th>
<th>-15 to 15</th>
<th>&gt;15</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.2</td>
<td>3.6</td>
<td>13.8</td>
</tr>
<tr>
<td>24</td>
<td>5.7</td>
<td>5.9</td>
<td>21.6</td>
</tr>
<tr>
<td>36</td>
<td>9.4</td>
<td>4.4</td>
<td>18.1</td>
</tr>
<tr>
<td>48</td>
<td>9.6</td>
<td>2.4</td>
<td>17.6</td>
</tr>
<tr>
<td>60</td>
<td>11.8</td>
<td>5.2</td>
<td>21.2</td>
</tr>
<tr>
<td>72</td>
<td>14.9</td>
<td>6.2</td>
<td>20.8</td>
</tr>
<tr>
<td>84</td>
<td>15.0</td>
<td>8.6</td>
<td>22.9</td>
</tr>
<tr>
<td>96</td>
<td>13.9</td>
<td>5.5</td>
<td>17.1</td>
</tr>
<tr>
<td>108</td>
<td>4.5</td>
<td>4.5</td>
<td>16.1</td>
</tr>
<tr>
<td>120</td>
<td>5.0</td>
<td>4.2</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Table 5.6

The BSS of P-PRIME ternary forecasts of LGEM bias for each forecast interval. The BSS of each event outcome is represented with a column in the table.
Chapter 6. Summary and Future Work

An accurate portrayal of the limitations of a weather forecast is one of the simplest and most effective ways to enhance forecast value. Therefore, the current lack of error guidance accompanying operational TC forecasts is very surprising. Even though the NHC cone of uncertainty graphic for TC track forecasts is widely used by the media, emergency managers, and general public, there is no TC forecasting center in the world that provides an operational product that conveys information about the typical uncertainty associated with a model’s TC intensity forecast. As a result, this dissertation provides the foundation for the first operational forecast product, the PRIME model, that communicates the expected error of a TC intensity forecast.

One of the most important distinctions between PRIME forecasts and the cone of uncertainty is PRIME can predict error values that differ from climatology. Chapter 2 of this dissertation offered the inspiration for this unique feature of PRIME. In that chapter, the performance of the DSHP, LGEM, GFDL, and OFCL TC intensity models were binned based on TC attributes and the environmental conditions surrounding a TC. The appropriate $t$ tests were then used to measure whether the AE, bias, and skill of these models were significantly different depending on the situation. For a majority of the models and forecast hours, there were significant differences in the average performance of forecasts in different bins. The varying levels of success that a model showed between bins demonstrated that the performance of individual model forecasts often deviate from the average performance of the model. Additionally, the discrepancies in the performance of different models for the same bin highlighted situations where certain models should be trusted more than others.
Building on the results of chapter 2, chapter 3 described how PRIME used the situation-dependent performance of DSHP, LGEM, GHMI, and HWFI to predict the bias and AE of these intensity models in real time. PRIME was developed using a stepwise multiple linear regression technique. The synoptic variables from chapter 2 as well as proxies for initial condition error and atmospheric flow stability served as predictors. The developmental sample for PRIME included all Atlantic basin DSHP, LGEM, GHMI, and HWFI TC intensity forecasts from 2007 to 2014 that met the verification criteria of NHC and had all predictors available. PRIME forecasts were tested using a cross-validation procedure where all the regression coefficients were determined from the forecast sample with one year removed. The removed year represented the validation dataset for PRIME forecasts. PRIME predictions of AE and bias verified with significantly lower errors than climatological forecasts for all forecast hours and models. A second version of PRIME called Retrospective PRIME (R-PRIME) was developed using retrospective forecasts from 2008-2014. R-PRIME also performed very well, and the AE of its error forecasts were lower than PRIME.

For both versions of PRIME, the performance of the AE and bias forecasts showed higher skill for longer forecast hours and less accurate intensity models. Under these conditions, the variability in the error from forecast-to-forecast was much higher and therefore, easier to detect with PRIME (Palmer and Tibaldi 1988). Larger error fluctuations also cause random errors to become a smaller percentage of the total error. This result suggests that, even though longer intensity forecasts, bias forecasts, and bad model forecasts all have higher error variance, error predictions are potentially the most useful in these situations.
The skill of PRIME error forecasts implied that PRIME output could have additional applications. In chapter 4, PRIME error forecasts were used to bias-correct DSHP, LGEM, GHMI, and HWFI intensity forecasts and create unique ensembles of the four models. For all real-time models, the average AE of the PRIME bias-corrected intensity forecasts was significantly lower than the average AE of the original intensity forecasts at every forecast hour. The bias-corrected retrospective models all achieved lower average AE than the original retrospective models but there were two forecast hours for HWFI where the AE difference was not statistically significant.

PRIME forecasts were also used to produce two types of ensembles. The PRIME ensemble modifications either involved equal-weighting of a modified group of models or unequal-weighting of the original group of models. For the real-time models, the best-performing ensemble was created using PRIME bias forecasts to correct each model before the ensemble mean was calculated. This “Bias-Corrected ensemble” verified with lower AE than ICON at every forecast interval and for five of the ten forecast hours, it was significantly lower than ICON at the 92% level or higher. For the retrospective models, the best ensemble involved R-PRIME AE forecasts unequally-weighting the different models. This ensemble was called the “Unequal (AE SQR) ensemble” and did not significantly improve upon ICON at any forecast interval. The skill relative to ICON for both the best PRIME and R-PRIME modified ensemble was found to increase as the forecasts differed more from ICON.

In the final chapter, Probabilistic PRIME (P-PRIME) was developed to produce probabilistic forecasts of AE and bias for the same forecast sample as R-PRIME. P-PRIME used a similar framework to R-PRIME but multiple logistic regression replaced
multiple linear regression as the statistical forecasting scheme. The RPSSs calculated for ternary P-PRIME forecasts and BSSs calculated for binary P-PRIME forecast both showed positive skill relative to climatology for all forecast hours, models, and predictands. Additionally, the plotted reliability diagrams and sharpness diagrams confirmed forecasts were generally well-calibrated and moderately confident. P-PRIME forecasts were the most skillful when predicting extreme error values. This result is particularly encouraging because forecasters and end users could greatly benefit from knowing the likelihood of high and low error events.

At the time BN13 was published, it was the first study to recommend error guidance to accompany every TC intensity forecast. As a result, the PRIME model represents an initial attempt at using statistical techniques to provide skillful forecasts of TC intensity error. Several potential upgrades to the formulation of PRIME were included in this dissertation but there are additional problems that future error prediction techniques must address. As mentioned by Judt (2014), there is an ongoing discussion in the atmospheric science community whether small-scale processes or the large-scale environment are more important for controlling TC intensity. Many of the predictors used for PRIME were defined by synoptic-scale variables. The internal convective and dynamical processes also have a large influence on TC intensity predictability (Shapiro and Willoughby 1982; Rozoff et al. 2009) and parameters capturing these mechanisms would be natural additions as predictors.

Additionally, the methodology employed for the development of PRIME could be extended for the forecasting of other TC hazards. As it is currently defined, TC intensity cannot be precisely measured (see discussion in Nolan et al. 2014), and it is likely that
different predictands capturing TC strength will be operationally forecasted in the future. Examples of other possible predictands are the central pressure of the TC or the radii of certain wind speeds. A statistical scheme similar to the one used to develop PRIME can be utilized to forecast these better observed intensity metrics. In theory, these metrics will have smaller random errors and a forecast model like PRIME could provide more accurate forecasts of them.

Multiple linear regression is a practical model for forecasting TC intensity error but several more complex statistical schemes are available. Nonlinear methods and neural networking are two potential alternatives for producing error forecasts. As mentioned in chapter 1, well-behaved ensembles can also provide expectations of error based on the spread of the member forecasts. After evaluating other error forecasting techniques, guidance on TC intensity forecast performance should also be produced in other TC-prone regions across the world. TC landfalls are more prevalent outside the Atlantic basin, so reliable error forecasts would naturally be valuable in the Pacific Ocean and Indian Ocean. If successful, these forecasts could be produced globally and lead to more informed protocol for hurricane evacuations and storm preparations, which ultimately saves lives.
References


DeMaria, M. 2010: Tropical cyclone intensity change predictability estimates using a statistical-dynamical model, *29th AMS Conference on Hurricanes and Tropical Meteorology*, Tucson, AZ.


Musgrave, K.D., B.D. McNoldy, and M. DeMaria, 2012: Creation of a statistical ensemble for tropical cyclone intensity prediction. 30th Conf. on Hurr. And Trop. Met., Ponte Vedra Beach, FL.


