Transition of Machine-Learning Based Rapid Intensification Forecasts to Operations

Andrew Mercer

Kimberly Wood

Northern Gulf Institute

Department of Geosciences

Mississippi State University

2018 Tropical Cyclone Operations and Research Forum

72nd Interdepartmental Hurricane Conference

March 14, 2018

MISSISSIPPI STATE UNIVERSITY

NGI Northern Gulf Institute a NOAA cooperative institute

Motivation

- Operational Atlantic Basin statistically-based RI forecasts (SHIPS-RII) are methodologically limited by two key areas:

  1. Use of linear methods limiting the ability of methods to describe inherent nonlinearity among meteorological variables within TCs.

  2. Statistical models require individual values of predictors, such that field averages of important TC characteristics are used (losing important spatial information)

- Impacts RI forecasts by driving reduced skill
  - Typically roughly 15% better than climatology
- Possible that machine learning and updated feature selection will improve upon these issues

Research Objectives

- The initial NGI-funded project had two distinct objectives

1) Identify spatial regions that are most distinct between RI and non-RI events using robust feature selection
2) Develop a machine learning ensemble to predict RI/non-RI classes, from which an ensemble probability of RI will be derived

- The primary objective of the JHT project is transitioning this ensemble into an operational RI forecast tool to assist current RI forecast decisions

Datasets

- TC characteristics obtained from HURDAT2, including:
  - Storm center latitude/longitude
  - Storm maximum wind speed
  - Storm minimum pressure
  - Storm speed
- RI defined as **30-kt increase in peak wind speed in 24 hours**, current primary operational definition
  - 7.9% (52) of 658 tested TC timesteps are RI timesteps
  - Other definitions are currently being developed, not provided in upcoming testbed

Datasets

- Forecast mode requires operational NWP output with long period of record for training
- Global Ensemble Forecast System - Reforecast (GEFS-R) proxy for forecast data
- GEFS-reforecast database characteristics
  - 1° latitude-longitude global grid spacing
  - 8 vertical levels
  - Once daily (0000 UTC) data, 192 forecast hours
- TC centric grids - 11° longitude by 15° latitude grid

Datasets

Variable name	Vertical levels (mb)
Geopotential height (m)	1000, 925, 850, 700, 500, 300, 200, 100
Temperature (K)	1000, 925, 850, 700, 500, 300, 200, 100
Zonal (*u*) wind speed (m s^{-1})	1000, 925, 850, 700, 500, 300, 200, 100
Meridional (*v*) wind speed (m s^{-1})	1000, 925, 850, 700, 500, 300, 200, 100
Specific Humidity (kg kg^{-1})	1000, 925, 850, 700, 500, 300
Mean Sea Level Pressure (Pa)	Surface
Sea Surface Temperature (K)	Surface
Latent Heat Flux (K m s^{-1})	Surface
Sensible Heat Flux (K m s^{-1})	Surface
Convective Available Potential Energy (J kg^{-1})	Surface
Convective Inhibition (J kg^{-1})	Surface
Pressure Vertical Velocity (Pa s^{-1})	850
Static Stability (m^4 s^2 kg^{-2})	925, 850, 700, 500
Equivalent Potential Temperature (K)	1000, 850, 700, 500, 300
Divergence (s^{-1})	200
Vorticity (s^{-1})	700, 500, 200
Vertical Shear (m/s)	850-200 mb layer

Italicized parameters computed from GEFS-R observed fields, 59 total layers considered

Feature Selection Methodology

- Robust feature selection on the GEFS-R data was required to filter down the 59 layers x 165 gridpoints per layer (9735 features over 658 observations)
- Feature selection completed in a two-step process

1) Layer-based feature selection via permutation tests, layers retained if $p < 0.01$
2) Pointwise feature selection – bootstrapped permutation testing for 500 replicates was completed at all gridpoints for "best" layers, points kept if median replicate $p < 0.01$

- Six GFS-R points retained, all of which were *u*-wind components (5 at 200 mb, 1 at 300 mb)

Feature Selection Methodology

- In addition to 6 GFS-R gridpoints retained, additional features describing TC characteristics and SHIPS-RII predictors were retained (17 total predictors)
- From HURDAT2
  - Storm speed
  - Previous 12-hour intensity change
  - Storm latitude and longitude
- From SHIPS output
  - Low-level relative humidity
  - Divergence
  - Wind shear
  - Maximum potential intensity
  - Ocean heat content
  - Dry air predictor

Machine Learning Ensemble

- Current SHIPS-RII forecasts provide RI probabilistic output
- To obtain probabilistic output for machine learning, a large number of methods and configurations were tested
- Optimally performing members were retained as part of a machine learning ensemble
- Three machine learning methods considered
  - Support Vector Machines (SVMs) – 28 configurations tested
  - Random Forests (RFs) – 125 configurations tested
  - Multilayer Perceptrons (MPs) – 48 configurations tested

Machine Learning Ensemble

- Bootstrap-based cross-validation with 300 pairwise bootstrap iterations (80% training/20% testing)
- Optimal members retained based on performing optimally (based on Heidke Skill Score on testing dataset) in at least 10 of the 300 bootstrap iterations
- Resulted in 5 SVM members, 18 RF members, and 18 MP members (a 41 member ensemble)
- Heidke Skill Score (HSS) results for each ensemble member, based on cross-validated median bootstrap HSS, provided on next slide

SVM Member Results

Member	Kernel	Cost	γ-value	HSS
SVM1	Poly-2	1	0.05	0.183
SVM2	RBF	1	0.05	0.270
SVM3	RBF	10	0.05	0.306
SVM4	RBF	1	0.1	0.319
SVM5	RBF	10	0.1	0.277

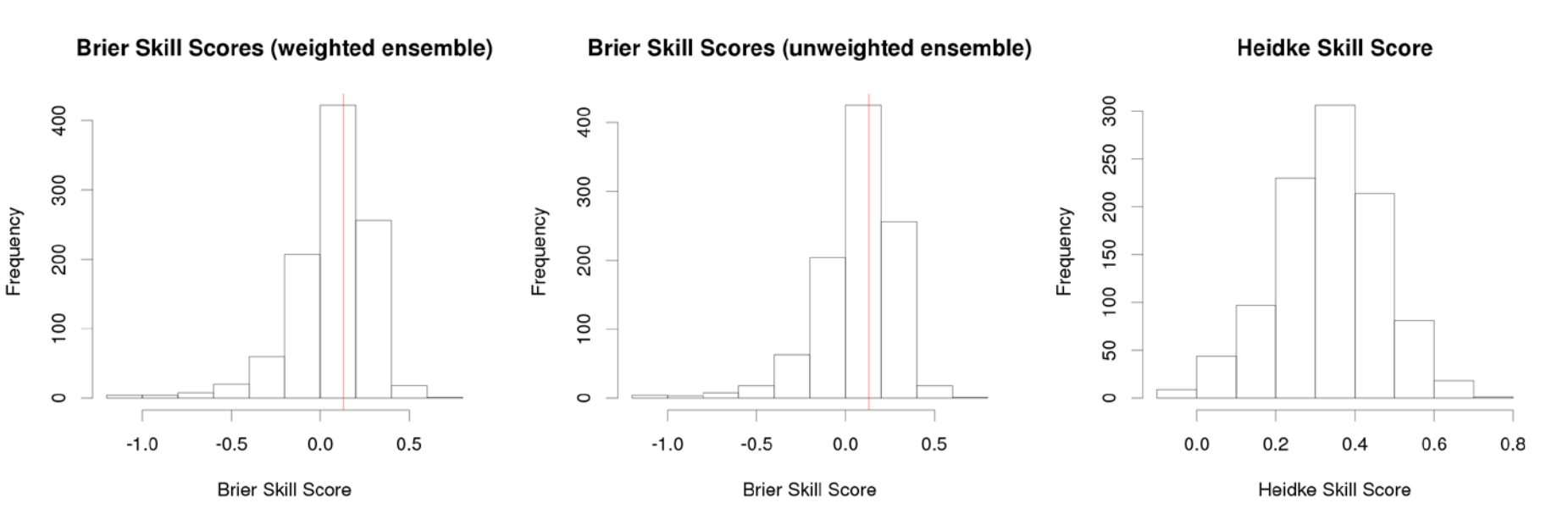
RF and MP Member Results

Member	Trees	Predictors	Cutoff	HSS
RF1	100	4	0.2	0.258
RF2	100	4	0.3	0.248
RF3	100	5	0.2	0.263
RF4	100	5	0.3	0.263
RF5	100	6	0.2	0.265
RF6	100	6	0.3	0.264
RF7	100	6	0.4	0.265
RF8	100	6	0.5	0.214
RF9	100	6	0.6	0.131
RF10	100	7	0.3	0.267
RF11	100	7	0.4	0.274
RF12	100	5	0.2	0.265
RF13	200	5	0.4	0.259
RF14	200	6	0.2	0.267
RF15	200	6	0.3	0.264
RF16	200	7	0.2	0.267
RF17	200	7	0.3	0.267
RF18	200	7	0.4	0.277

Member	Layers	Nodes	Epochs	HSS
MP1	4	10	100000	0.308
MP2	2	11	100000	0.309
MP3	2	12	100000	0.310
MP4	1	8	100000	0.308
MP5	4	8	100000	0.310
MP6	4	10	30000	0.309
MP7	1	11	30000	0.310
MP8	1	12	30000	0.309
MP9	3	8	30000	0.309
MP10	4	8	30000	0.306
MP11	2	9	30000	0.314
MP12	4	9	30000	0.309
MP13	1	10	50000	0.310
MP14	3	11	50000	0.307
MP15	4	12	50000	0.306
MP16	1	8	50000	0.313
MP17	1	9	50000	0.309
MP18	2	9	50000	0.313

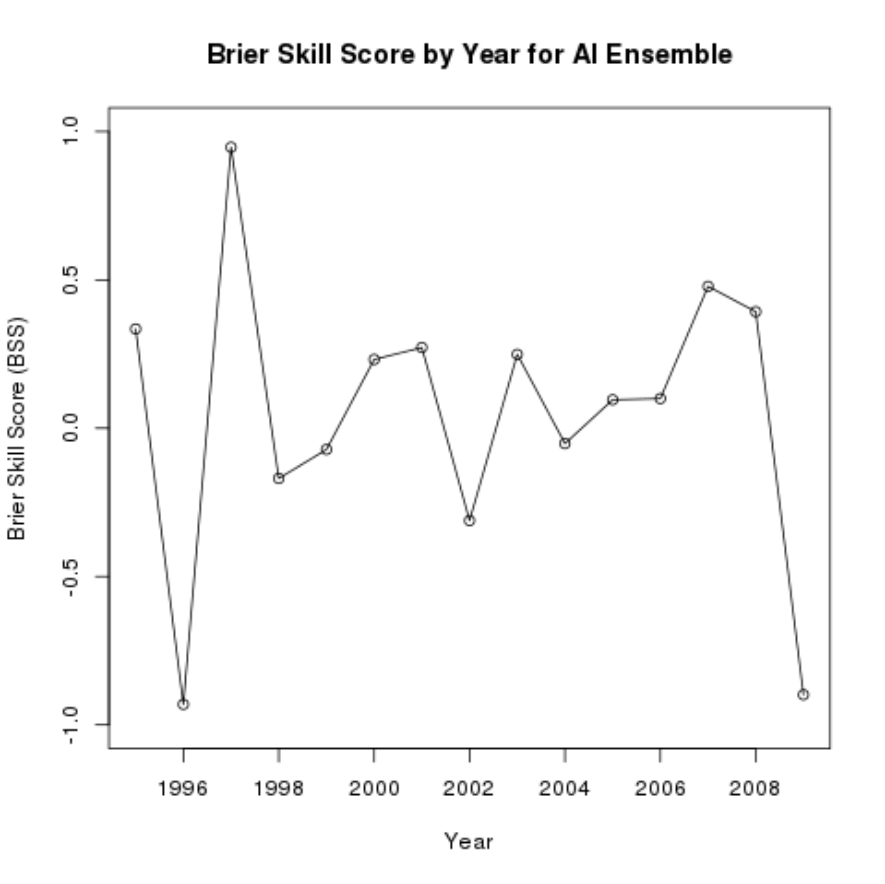
Full Ensemble Performance

- Brier Skill Score (BSS) for bootstrapped full ensemble performance (red line represents SHIPS-RII current performance, marked as 0.15)
- Weighted vs. unweighted, no real change



Full Ensemble Performance

- AI ensemble performance by year (tested year not included in training phase to simulate forecast mode)
  - Mean annual BSS = 0.04



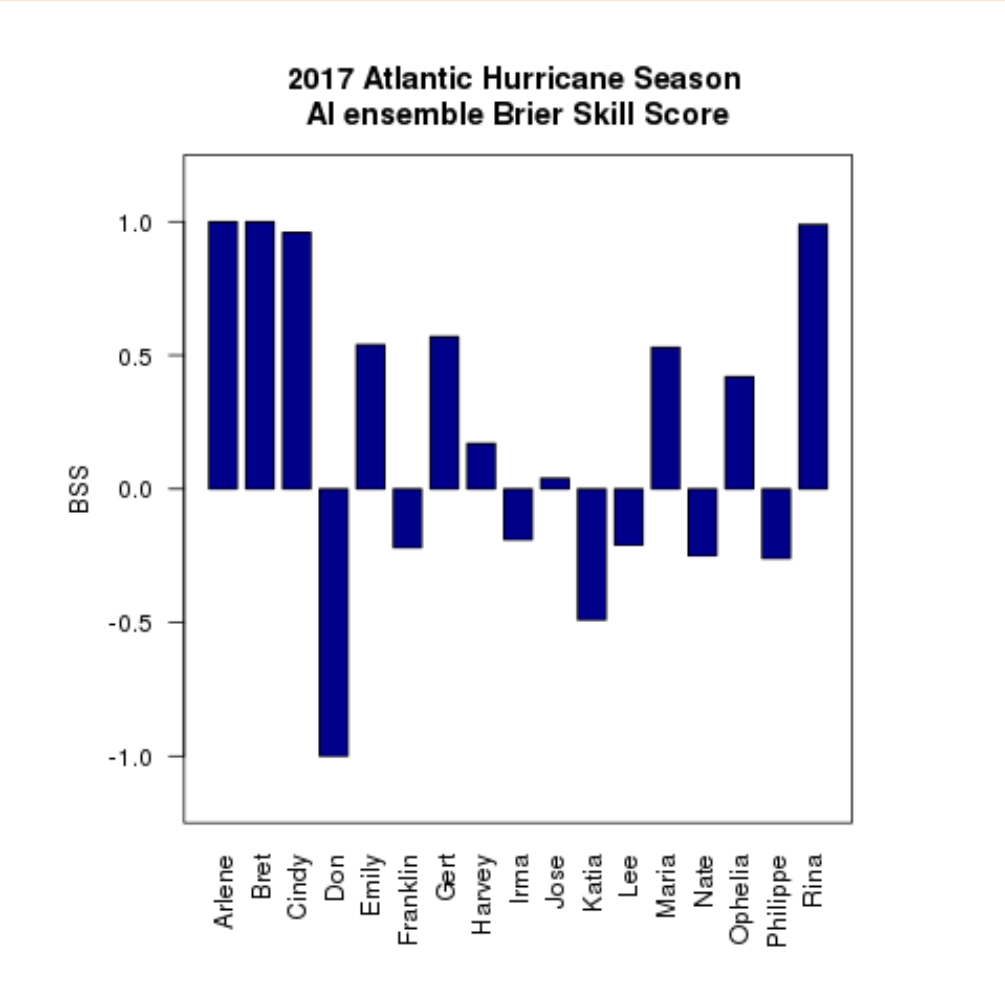
The 2017 Atlantic Hurricane Season

- AI ensemble tested locally in a quasi-operational environment during the 2017 Atlantic Hurricane Season
- Predictors from operational SHIPS, operational track information, and operational GFS (not GEFS-R)
- RI forecasts completed for all 24-hour forecast times for all Atlantic TCs. Skills for each storm were analyzed and compared against SHIPS-RII performance
- Performance measures
  - Global AI ensemble BSS: -0.01
  - SHIPS-RII BSS: 0.202
  - Bayesian BSS: 0.177
  - Logistic BSS: 0.202
  - Consensus BSS: 0.240

[a]*Based on climatology of 7.9% RI, frequency of RI in 2017 season was 10%*

The 2017 Atlantic Hurricane Season

- Brier Skill Score by Storm for Ensemble Probability



The 2017 Atlantic Hurricane Season

- Global Contingency Statistics (defining an RI forecast as 7.4% probability or higher per climatology)

Statistic	AI Ens	SHIPS-RII	Logistic	Bayesian	Consensus
PC	0.839	0.427	0.781	0.887	0.807
CSI	0.164	0.145	0.308	0.358	0.324
BIAS	1.237	6.658	3.132	1.395	2.763
FAR	0.744	0.854	0.689	0.547	0.667
POD	0.316	0.974	0.974	0.623	0.921
POFD	0.102	0.633	0.240	0.085	0.205
HSS	0.193	0.097	0.377	0.465	0.401
TSS	0.213	0.340	0.733	0.547	0.716

The 2017 Atlantic Hurricane Season

- Major findings from the 2017 season
  - AI ensemble seems to favor underprediction of RI timesteps/overprediction of non-RI timesteps
  - Ensemble performed very well in storms that did not undergo RI at any point in the life cycle
  - Ensemble performed well for storms originating in the western Atlantic (e.g. Hurricane Maria)
  - Performance degraded with storms originating in the eastern Atlantic (e.g. Hurricane Irma) or storms originating in the Caribbean (e.g. Hurricane Nate)
  - Poor performance with Don was due to one badly forecast timestep driving down the BSS significantly
  - BS_{climo} based on frequency of RI in the 658 timesteps used to create the ensemble. May not be the best estimate of climatology

- Example output
  - Sept. 19 2017 0000 UTC
  - AL15 (Maria)

2017 Atlantic AI Ensemble
09/19/2017
0000
AL152017

AI Ensemble Member	Prediction	PC (%)
SVM1	1	0
SVM2	1	0
SVM3	0	100
SVM4	0	100
SVM5	0	100
RF1	1	0
RF2	1	0
RF3	1	0
RF4	1	0
RF5	1	0
RF6	1	0
RF7	1	100
RF8	1	100
RF9	1	100
RF10	1	0
RF11	1	0
RF12	1	0
RF13	1	100
RF14	1	0
RF15	1	0
RF16	1	0
RF17	1	0
RF18	1	0
MP1	0	0
MP2	1	0
MP3	1	0
MP4	0	0
MP5	0	0
MP6	1	0
MP7	0	0
MP8	1	100
MP9	0	0
MP10	1	0
MP11	1	100
MP12	0	100
MP13	0	0
MP14	0	100
MP15	1	0
MP16	0	0
MP17	1	100
MP18	0	0

Ongoing/Future Work

- Preparations prior to Testbed
  - Developed a user guide and summary sheet explaining the different components of the output and their interpretation
  - Working on porting the code from R to Python for easier operational integration
  - Final preparations converting the output to e-deck format for easier integration into the ATCF
  - Setting up local website at Mississippi State to house the output for outside use as interested
- Future work
  - Continued feature selection work to identify improved GFS fields being used by the ensemble
    - Swarm optimization and genetic algorithms being considered
  - Development for other RI definitions and lead times
  - This work will be completed after the testbed

Questions?

If interested in training materials for AI ensemble, I will provide digital copies. Email me at aem35@misstate.edu.

Machine Learning Ensemble

- SVM configurations tested (28 total permutations)
  - Varied kernel function (7 tested) and cost function (4 tested)
- RF configurations tested (125 total permutations)
  - Varied grown trees (5 options tested), cutoff criterion (5 options tested, for uneven weighted samples), and predictors per tree (5 options tested)
- MP configurations tested (48 total permutations)
  - Varied stopping threshold (3 tested), hidden nodes (4 configurations tested) and hidden layers (4 configurations tested)
  - *Note with MPs, hidden node counts remained the same for all hidden layers, no deep learning was done*