P580

Experimental Model Evaluations for Tropical Cyclone Forecasting

L. Nance, M. Biswas, B. Brown, T. Fowler, P. Kucera, K. Newman, J. Vigh, and C. Williams

National Center for Atmospheric Research, Boulder, CO

Joint Numerical Testbed Program Tropical Cyclone Modeling Team

Background

What is HFIP?

The Hurricane Forecast Improvement Project is a 10-year effort to accelerate improvements in 1–5-day forecasts for hurricane track, intensity, storm surge & to reduce forecast uncertainty, w/ an emphasis on rapid intensity change.

HFIP approach…

Invest in data assimilation & model development by the research community, as well as observations. Overall program includes an annual retrospective exercise that tests the most promising innovations.

Our role…

Provide objective verification statistics for contributed forecasts w/ a focus on performance relative to top-performing operational models. Evaluation is used to select experimental models that will be incorporated into NHC's real-time numerical guidance.

Results from all past exercises available at: http://www.rap.ucar.edu/jnt/tcmt/

2013 Modeling Group Participants

Organization	Model
MMM/SUNY-Albany	AHW
UW – Madison	UW-NMS 4km
	UW-NMS ensemble
NRL	COAMPS-TC
PSU	ARW
GFDL	GFDL ensemble mean
GSD	FIM
FSU	Multi-Model Super Ensemble
CIRA	SPICE
EMC	HWRF ensemble

Operational Baselines

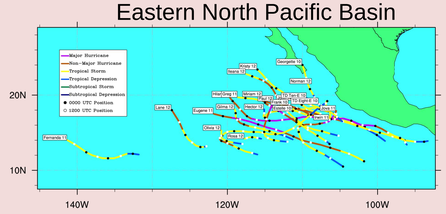
Track:	Intensity:
ECMWF	DSHP
GFS	LGEM
GFDL	GFDL
Variable consensus (ECMWF, GFS, UKMET, GFDL, HWRF)	Fixed consensus (DSHP, LGEM, GFDL, HWRF)

Acknowledgments: The authors would like to thank the participating HFIP modeling groups. Without all their hard work towards improving their models and generating retrospective forecasts, this work would not be possible. We would also like to thank James Franklin of NHC for all his fruitful discussions that contributed to refining and expanding our evaluation approach. NCAR is sponsored by the National Science Foundation. This work was funded by and performed under the auspices of the NOAA's Hurricane Forecast Improvement Project (HFIP).

2013 Retrospective Cases

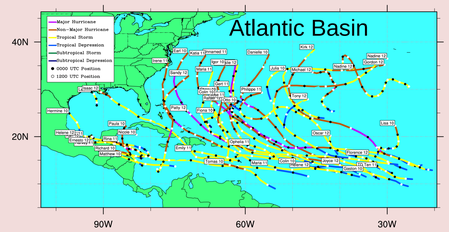
All tropical & subtropical cyclones in Atlantic & eastern North Pacific basins for Aug-Oct 2010-12

Eastern North Pacific Basin



2010: 5 storms
2011: 6 storms # of cases: 315
2012: 11 storms

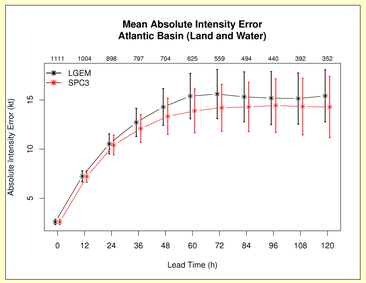
Atlantic Basin

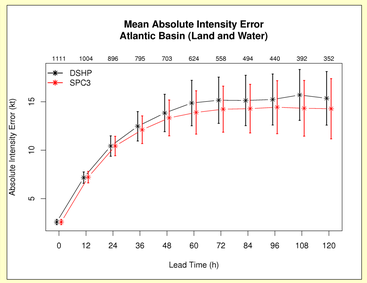


2010: 17 storms
2011: 14 storms # of cases: 897
2012: 15 storms

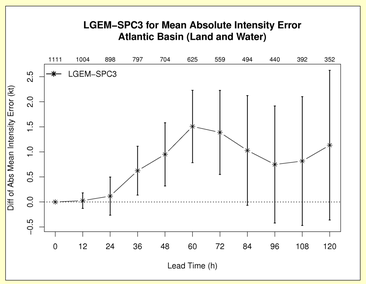
Pairwise Difference Tests

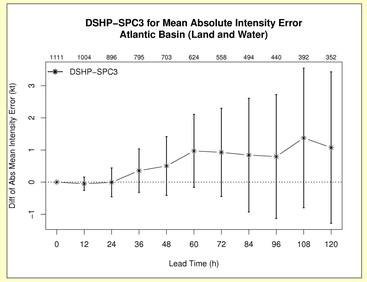
2-sample test often leads to statistically indistinguishable means due to inherent variability of the sample.





Pairwise tests remove variability due to common forecast challenge (i.e. more powerful test).





Summary Statistical Significance Tables

Pairwise tests for multiple baselines leads to large amount of information that needs to be condensed into a form NHC can easily determine key aspects of performance. Shading indicates difference is SS & darkness of shading is used to distinguish size of differences (magnitude or percent change), providing useful visual cues while retaining detailed info.

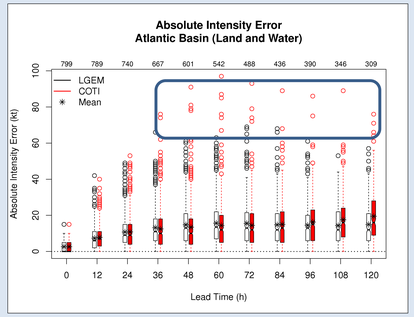
	Track	Intensity
SS differences	% ≤ -6	% ≤ -10
	-6 < % ≤ -4	-10 < % ≤ -5
	-4 < % < 0	-5 < % < 0
	0 < % < 4	0 < % < 5
	4 ≤ % < 6	5 ≤ Δ < 10
	% ≥ 6	% ≥ 10
Not SS	% < 0	% < 0
	% > 0	% > 0

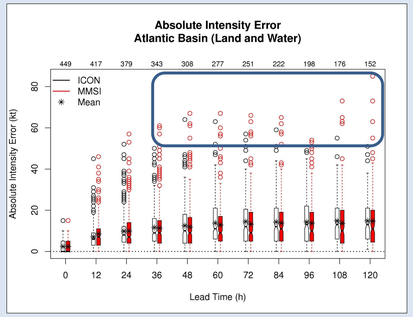
mean error difference → 1.4
% improve (+) /degrade (-) → 11%
p-value → 0.973

Forecast Hour		12	24	36	48	60	72	84	96	108
Atlantic Basin	GHMI Land/Water	0.3 5% 0.970	0.5 4% 0.735	1.4 11% 0.973	1.7 11% 0.954	2.5 15% 0.980	3.7 21% 0.996	4.5 24% 0.997	5.0 26% 0.998	5.1 26% 0.998
	GHMI Water Only	0.4 6% 0.992	1.0 9% 0.973	1.8 13% 0.986	2.1 14% 0.980	3.0 18% 0.986	4.3 23% 0.995	4.5 24% 0.987	5.1 26% 0.991	5.2 27% 0.985
	LGEM Land/Water	0.0 0% 0.257	0.1 1% 0.453	0.6 5% 0.988	1.0 7% 0.997	1.5 10% 0.999	1.4 9% 0.999	1.0 7% 0.933	0.7 5% 0.789	0.8 5% 0.786
	LGEM Water Only	0.0 0% 0.353	0.4 3% 0.984	0.8 6% 0.999	1.1 8% 0.999	1.5 10% 0.999	1.4 9% 0.997	1.0 7% 0.901	0.8 6% 0.740	0.9 6% 0.764
	DSHP Land/Water	-0.1 -1% 0.373	0.0 0% 0.023	0.4 3% 0.698	0.5 4% 0.719	1.0 7% 0.905	0.9 6% 0.814	0.8 6% 0.648	0.8 6% 0.581	1.4 9% 0.785
	DSHP Water Only	0.0 0% 0.305	0.4 4% 0.983	0.7 5% 0.963	0.9 6% 0.898	1.3 9% 0.953	1.3 9% 0.898	1.0 7% 0.642	0.9 6% 0.566	1.4 9% 0.735
	ICON Land/Water	-0.4 -6% 0.999	-0.9 -9% 0.999	-0.8 -7% 0.974	-0.9 -7% 0.923	-0.3 -3% 0.386	0.0 0% 0.030	0.1 1% 0.113	0.3 2% 0.222	0.8 5% 0.481
	ICON Water Only	-0.3 -4% 0.999	-0.4 -4% 0.951	-0.6 -5% 0.890	-0.6 -5% 0.713	0.0 0% 0.012	0.6 4% 0.431	0.6 4% 0.338	0.8 5% 0.416	1.3 8% 0.619

Error Distributions

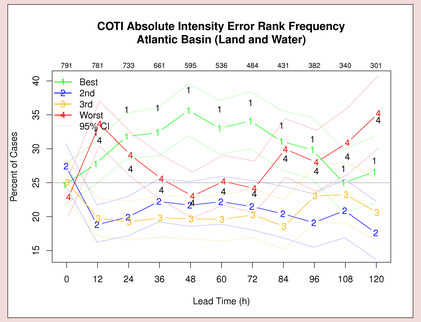
The error distributions can also provide valuable information about the performance of experimental models. A model whose performance is as good as or perhaps slightly better than the operational models for a majority of the sample, but produces forecasts with errors significantly larger than the largest operational errors is not viewed favorably by the forecaster.

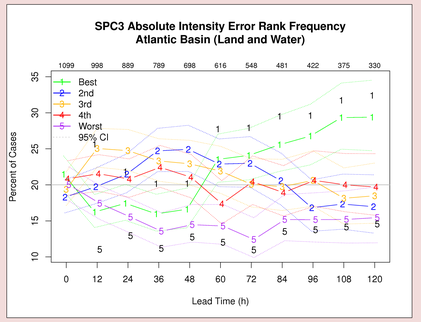




Rank Frequency

The rank frequency approach provides information about how the experimental model performed with respect to the operational baselines as a group. Ranks 1 (smallest error) – 4 or 5 (largest error) are assigned to all model errors, w/ ties assigned randomly. The frequency of each rank for the experimental model is displayed in a line plot by lead time.





Summary

By applying a variety of statistical tests to retrospective forecasts provided by the HFIP model development groups, the TCMT provides NHC w/ an in-depth picture of the experimental models' strengths & weaknesses w/ respect to track & intensity forecasts. Pairwise tests provide a powerful method for determining whether the mean errors for the experimental models are different those for the operational baselines. Working in consultation w/ NHC's Hurricane Specialists, the TCMT was able to develop summary SS tables for the pairwise difference tests that concisely provide detailed information of interest to NHC. By comparing the properties of the error distributions, the TCMT is also able to highlight areas of concern (large outliers) or favorable improvements (significant reduction in the largest errors) related to the experimental model performance that are not immediately apparent when only considering the mean statistics. And finally, rank frequency tests provide a look at how the experimental model performs w/ respect to the operational models as a group. Moving forward, the TCMT is adding new approaches to its evaluation directed at providing more diagnostic information about the experimental models' performance to provide insight to the on-going development process.