2012 Stream 1.5 Activities

Louisa Nance, Barb Brown, Paul Kucera, Kathryn Newman, Christopher Williams, Mrinal Biswas, Laurie Carson, Tricia Slovacek

NCAR/RAL/JNT

NCAR

2012 RETROSPECTIVE EXERCISE

Stream 1.5 Retrospective Evaluation

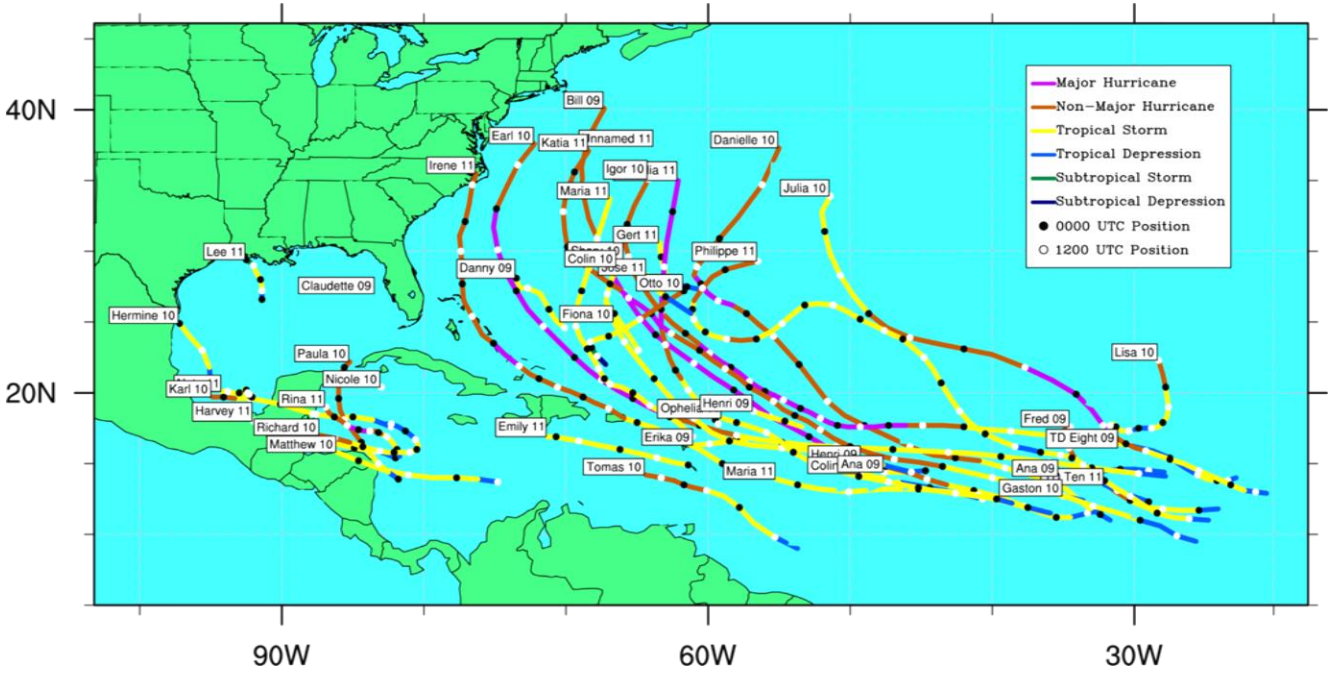
Goals

- Provide NHC with in-depth statistical evaluations of the candidate models/techniques directed at the criteria for Stream 1.5 selection
- Explore new approaches that provide more insight into the performance of the Stream 1.5 candidates

Selection criteria

- Track -
  - Explicit - 3-4% improvement over previous year's top-flight models
  - Consensus – 3-4% improvement over conventional model consensus track error
- Intensity –
  - improve upon existing guidance for TC intensity & RI

NCAR



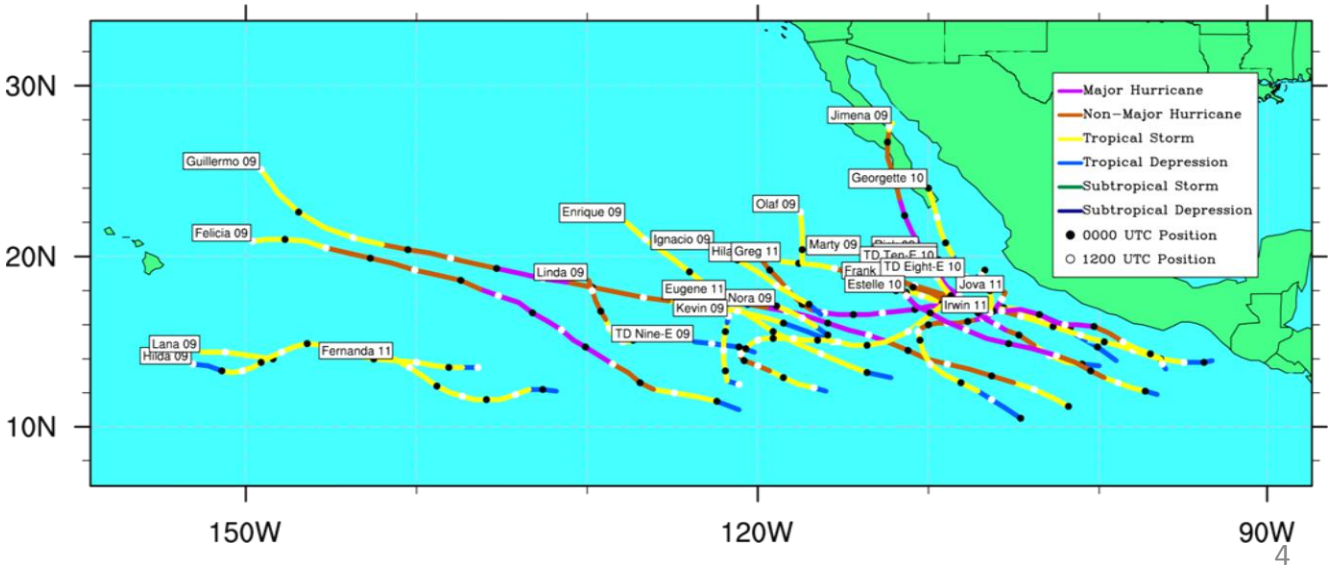
Atlantic Basin

2009: 8 storms

2010: 17 storms

2011: 15 storms

# of cases: 640



Eastern North Pacific Basin

2009: 13 storms

2010: 5 storms

2011: 6 storms

# of cases: 387

2012 Stream 1.5 Retrospective Participants

Organization	Model	Type	Basins	Config
MMM/SUNY-Albany	AHW	Regional-dynamic-deterministic	AL, EP	1
UW – Madison	UW-NMS	Regional-dynamic-deterministic	AL	1
NRL	COAMPS-TC	Regional-dynamic-deterministic	AL, EP	1
PSU	ARW	Regional-dynamic-deterministic	AL	2
GFDL	GFDL	Regional-dynamic-ensemble	AL, EP	2
GSD	FIM	Global-dynamic-deterministic	AL, EP	2
FSU	Correlation Based Consensus	Consensus (global/regional dynamic deterministic + statistical-dynamic)	AL	1
CIRA	SPICE	Statistical-dynamic-consensus	AL, EP	2

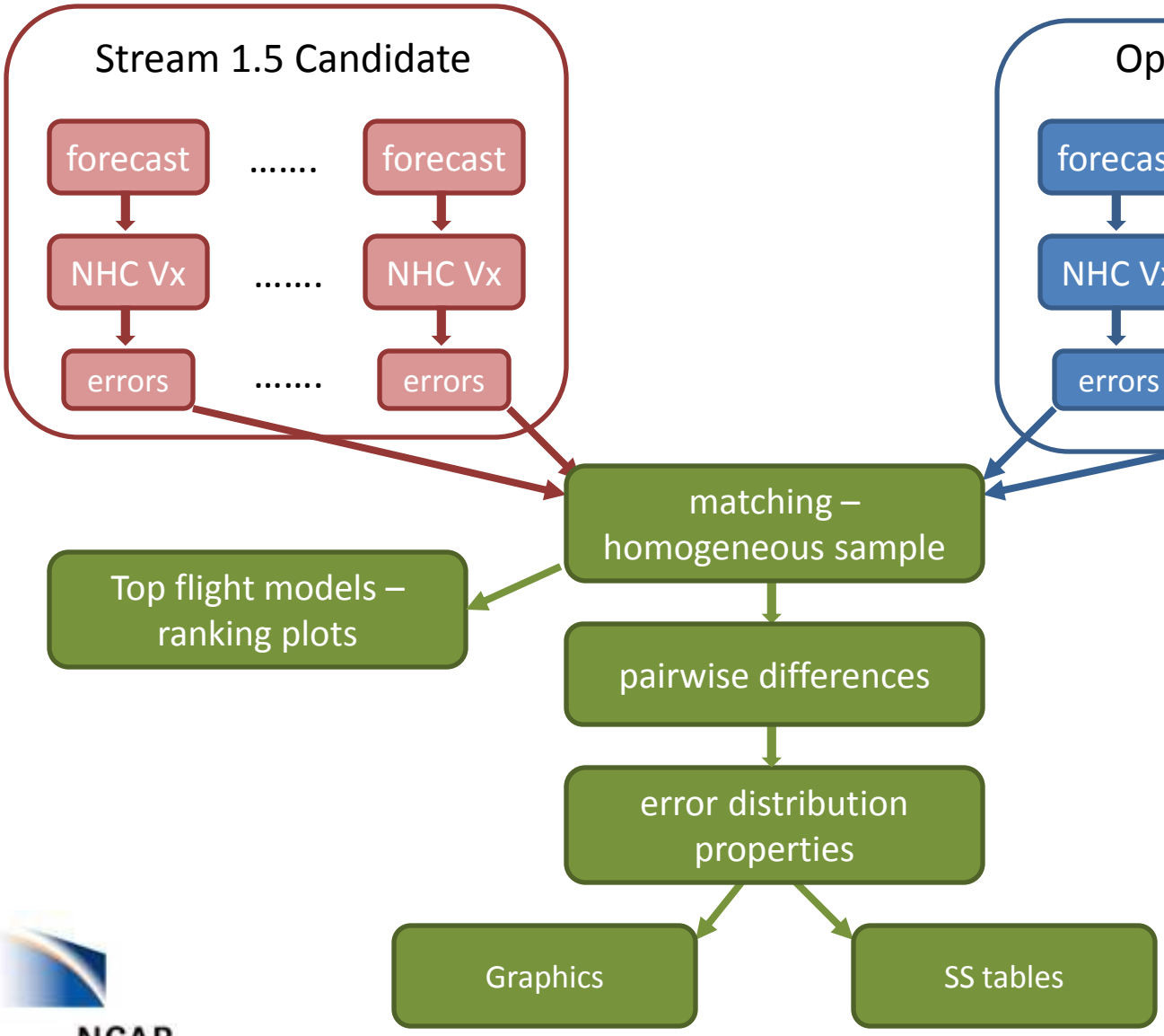
NCAR

Baseline Comparisons

Operational Baselines	Stream 1.5 configuration
Top flight models: Track – ECMWF, GFS, GFDL Intensity – DSHP, LGEM, GFDL	Stream 1.5
Consensus: Track (variable) AL: ECMWF, GFS, UKMET, GFDL, HWRF, GFDL-Navy EP: ECMWF, GFS, UKMET, GFDL, HWRF, GFDL-Navy, NOGAPS Intensity (fixed) AL & EP: Decay SHIPS, LGEM, GFDL, HWRF	AHW, ARW, UM-NMS, COAMPS-TC, FIM: Consensus + Stream 1.5 GFDL, SPICE: Consensus w/ Stream 1.5 equivalent replacement FSU-CBC: Direct comparison

NCAR

Methodology



Evaluation focused on early model guidance!

NCAR

All reports and graphics are available at:

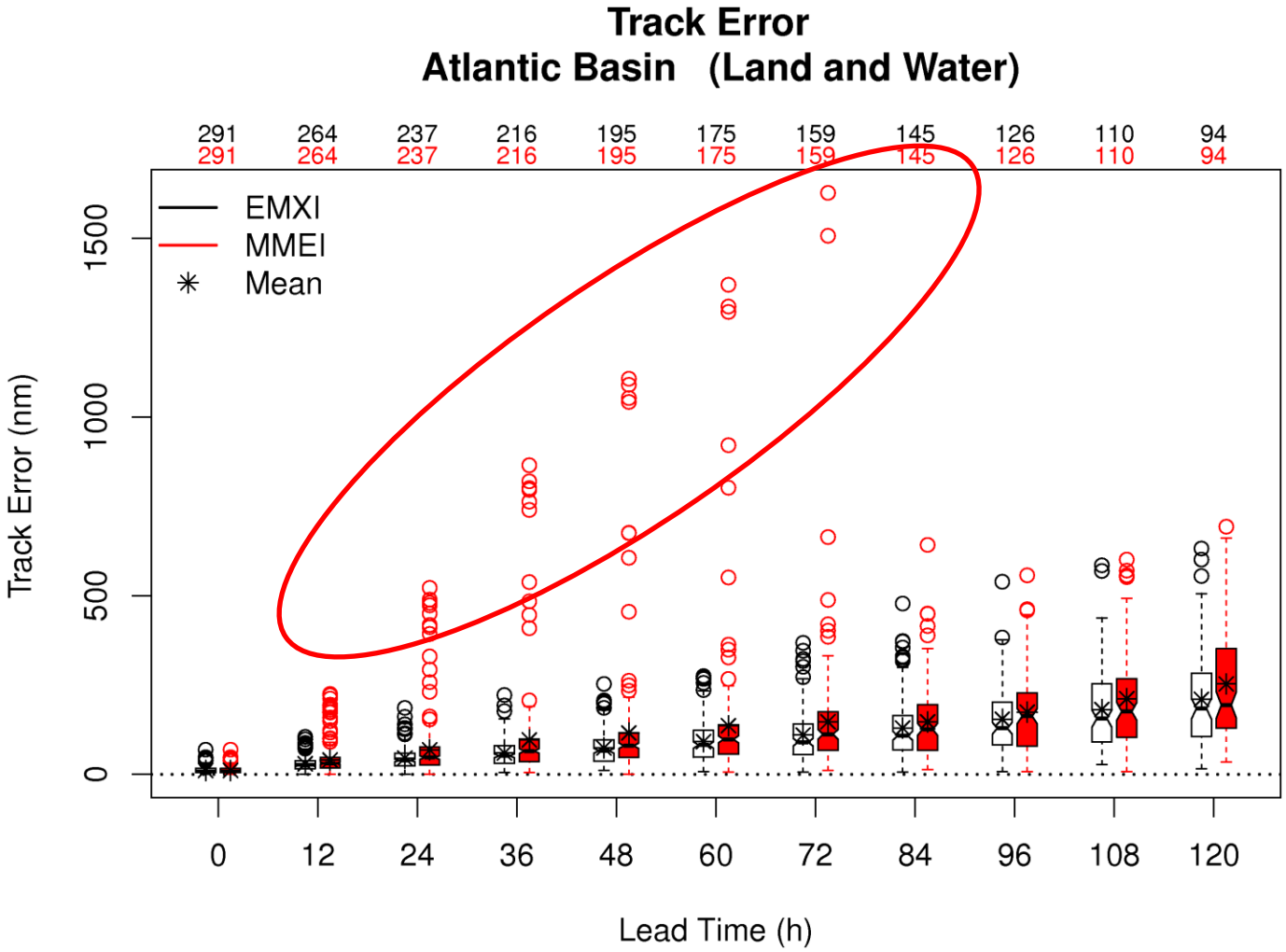
http://www.ral.ucar.edu/projects/hfip/h2012/verify/

SAMPLE RETRO RESULTS/DISPLAYS

NCAR

Error Distributions

Box Plots

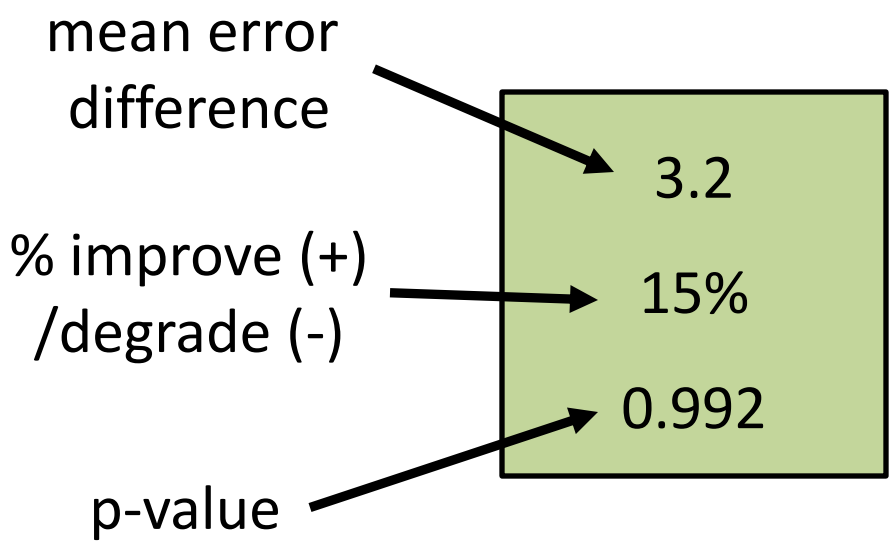


NCAR

Statistical Significance – Pairwise Differences

Summary Tables

Example COAMPS-TC



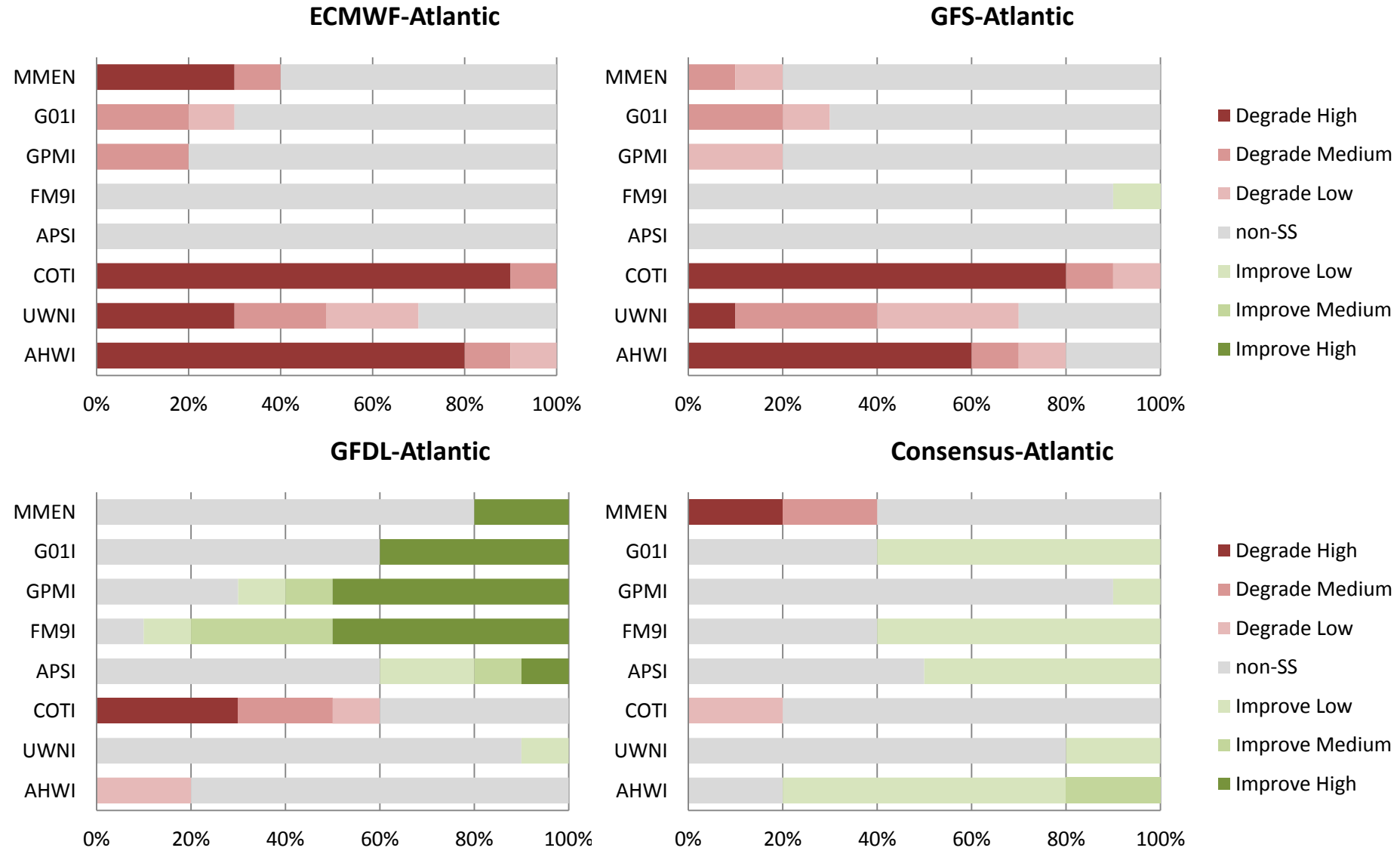
Practical Significance

		Track	Intensity
SS differences		Δ < -20	Δ < -2
		-20 < Δ < -10	-2 < Δ < -1
		-10 < Δ< 0	-1 < Δ< 0
		0 < Δ< 10	0 < Δ < 1
		10 < Δ< 20	1 < Δ < 2
		Δ > 20	Δ> 2
Not SS		Δ< 0	Δ < 0
		Δ> 0	Δ> 0

Forecast hour	0	12	24	36	48	60	72	84	96	108	120
GHMI	**0.0**	**-5.7**	**-12.4**	**-18.2**	**-21.5**	**-24.2**	**-23.6**	**-20.9**	**-23.4**	**-25.8**	**-28.6**
Track	**0%**	**-17%**	**-22%**	**-23%**	**-22%**	**-20%**	**-16%**	**-12%**	**-11%**	**-10%**	**-10%**
Land/Water	-	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.989**	**0.894**	**0.786**	**0.680**	**0.624**
GHMI	**0.0**	**-0.5**	**0.3**	**0.8**	**0.8**	**1.6**	**4.2**	**5.1**	**5.5**	**4.8**	**3.2**
Intensity	**0%**	**-6%**	**2%**	**5%**	**5%**	**9%**	**20%**	**24%**	**26%**	**23%**	**15%**
Land/Water	-	**0.987**	**0.546**	**0.625**	**0.576**	**0.954**	**0.999**	**0.999**	**0.999**	**0.999**	**0.992**

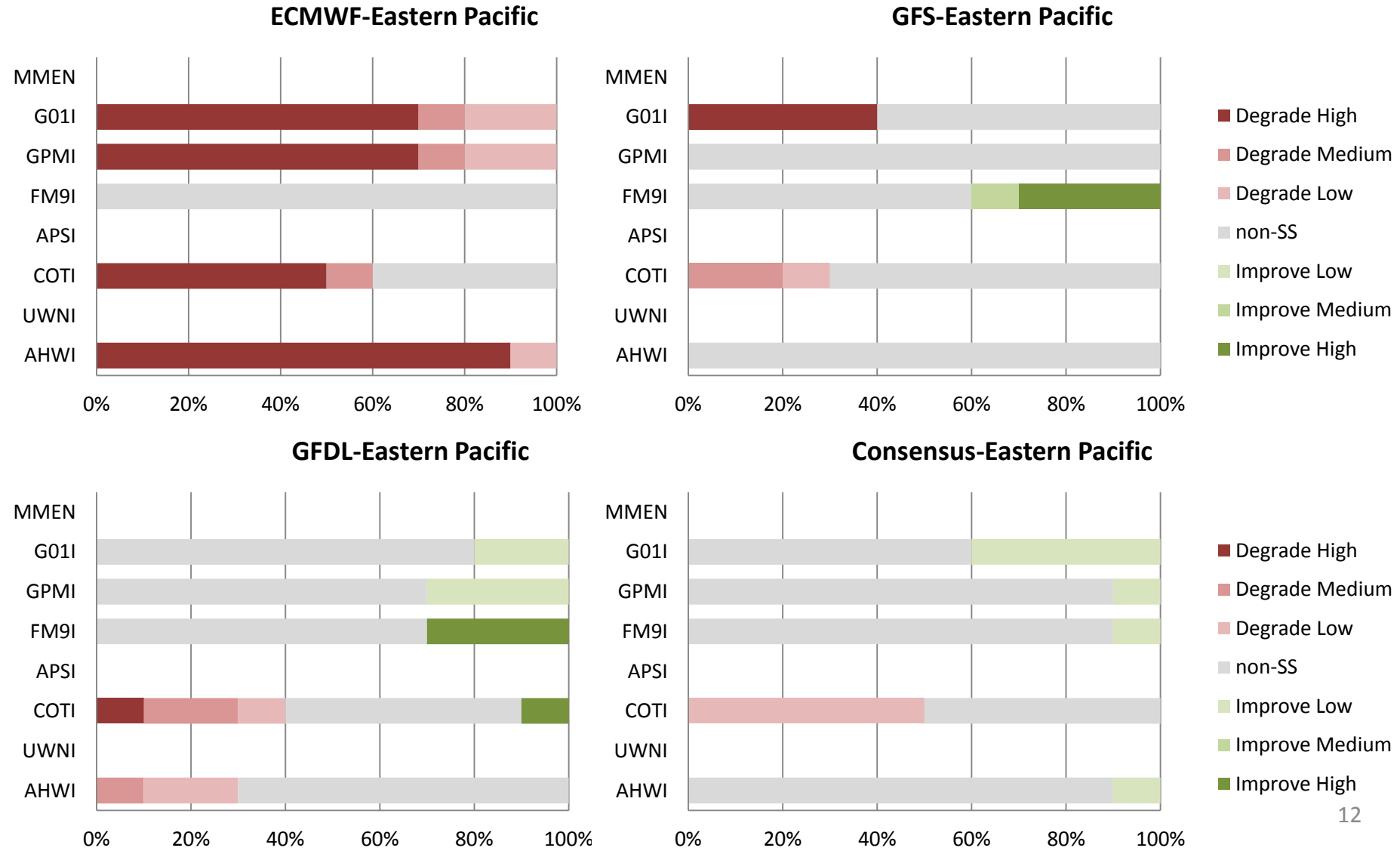
Statistically Significant Differences

Track Summary - Atlantic



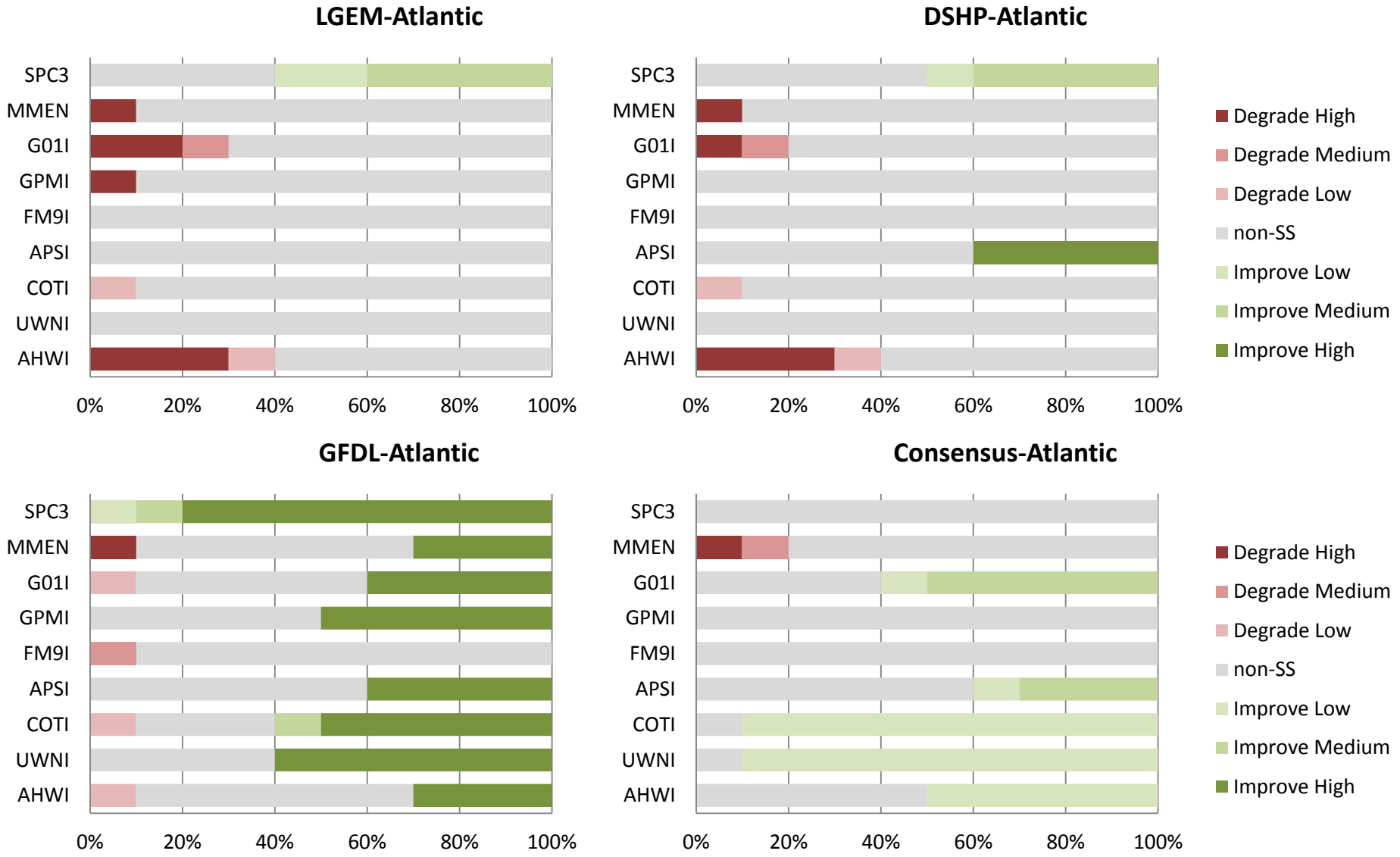
Statistically Significant Differences

Track Summary – Eastern Pacific



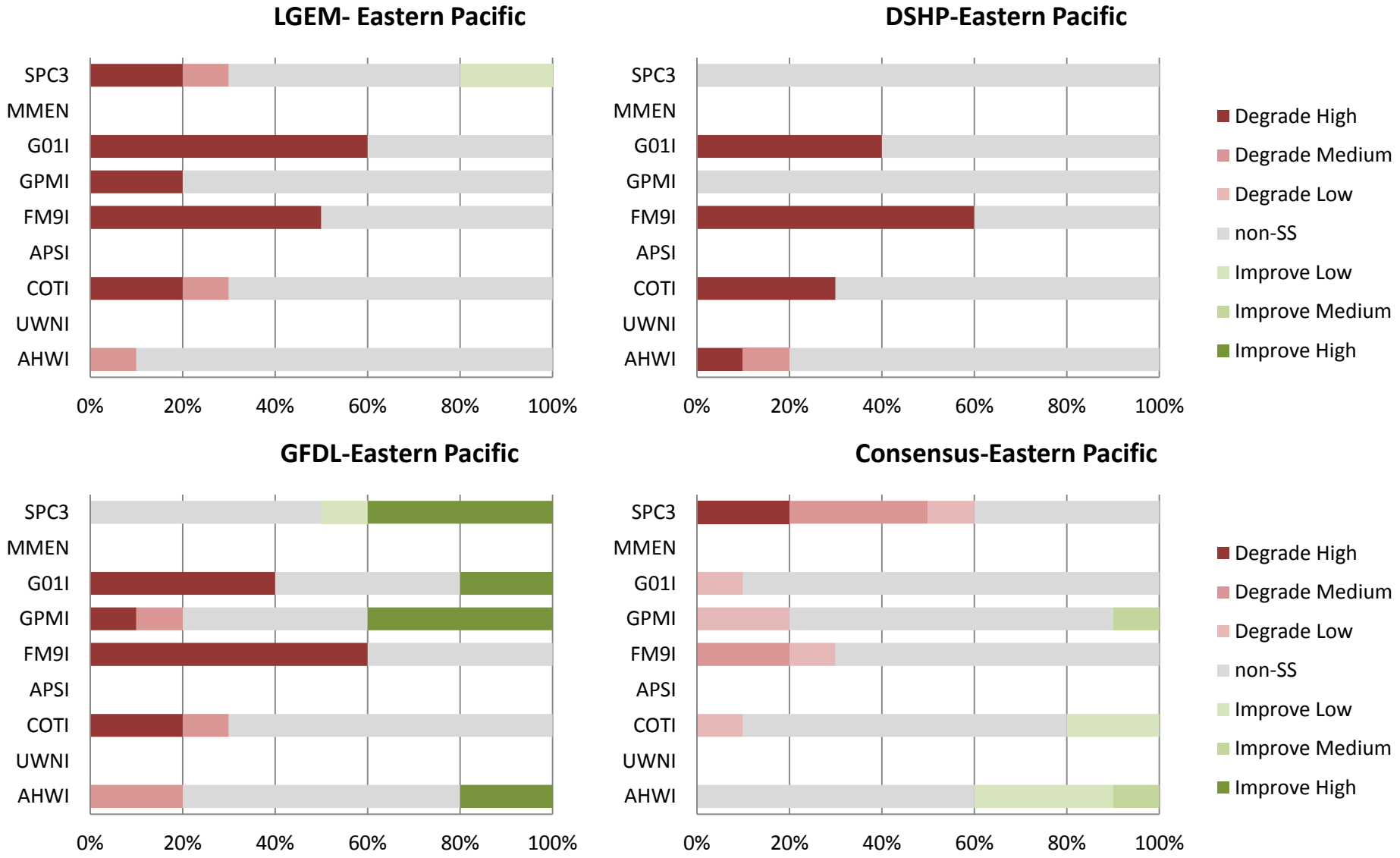
Statistically Significant Differences

Intensity Summary - Atlantic



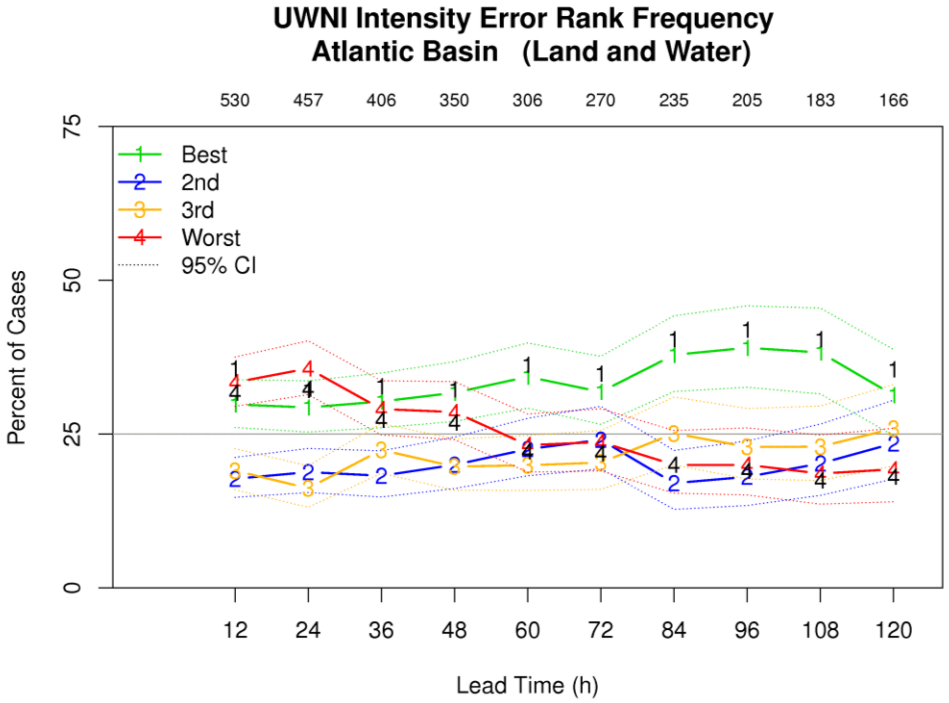
Statistically Significant Differences

Intensity Summary – Eastern Pacific



Comparison w/ Top-Flight Models

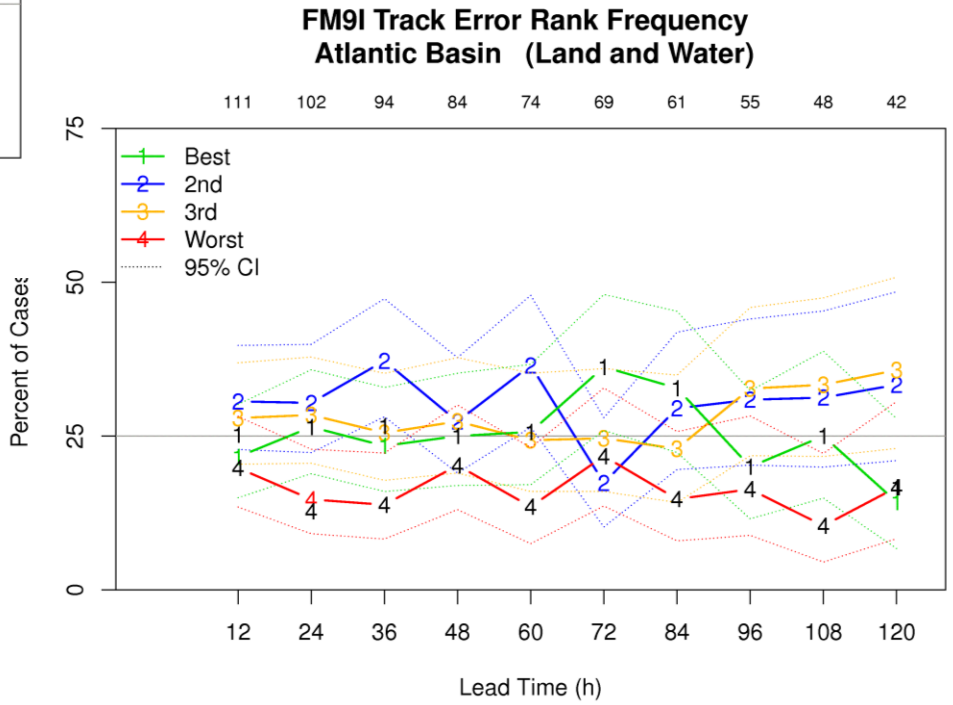
Rank Frequency



U of Wisconsin:

1st or last for shorter lead times

More likely to rank 1st for longer lead time



FIM:

CIs for all ranks tend to overlap

Method sensitive to sample size

NCAR

NHC's 2012 Stream 1.5 Decision

Organization	Model	Track	Track Consensus	Intensity	Intensity Consensus
MMM/SUNY-Albany	AHW		●		●
UW – Madison	UW-NMS				●
NRL	COAMPS-TC				●
PSU	ARW		●	●	●
GFDL	GFDL ensemble mean	●		●	
	No-bogus member	●		●	
GSD	FIM		●		
FSU	Correlation Based Consensus				
CIRA	SPICE			●	

NCAR

2013 Stream 1.5 Timeline

Activity	Deadline
Retrospective cases identified *(All TCs during Aug-Oct of 2010, 2011 & 2012)*	12 Nov '12
New Tier 3* data & forecast applications to run at NHC made known by researchers to NHC	1 Dec '12
NHC, Verification Team and TCMT determine evaluation metrics	Dec '12 – Mar '13
New Tier 3 data and forecaster applications to run at NHC delivered to NHC	15 Feb '13
List of potential Stream 1.5 participants & intended model characteristics (Notify TCMT of intent to participate by contacting Christopher Williams cwill@ucar.edu)	25 Feb '13
New Stream 1.5 participants: submit sample A-deck file to TCMT for review	15 Mar '13
Major annual ATCF upgrades completed	Early April '13
Completed retrospective runs submitted in A-deck format to TCMT	1 Apr '13
TCMT assessment of retrospective tests completed	15 May '13
Additional analysis requests by NHC, if any, undertaken by TCMT	15-31 May '13
NHC decision on prospective projects	31 May '13
Sample output from each approved model sent from TCMT to NHC	6 Jun '13
Technical preparations by model teams and NHC	1 Jun – 14 Jul '13
Test/shake-down for 2013 real-time activities	15-31 Jul '13
Stream 1.5 real-time activities	1 Aug – 31 Oct '13

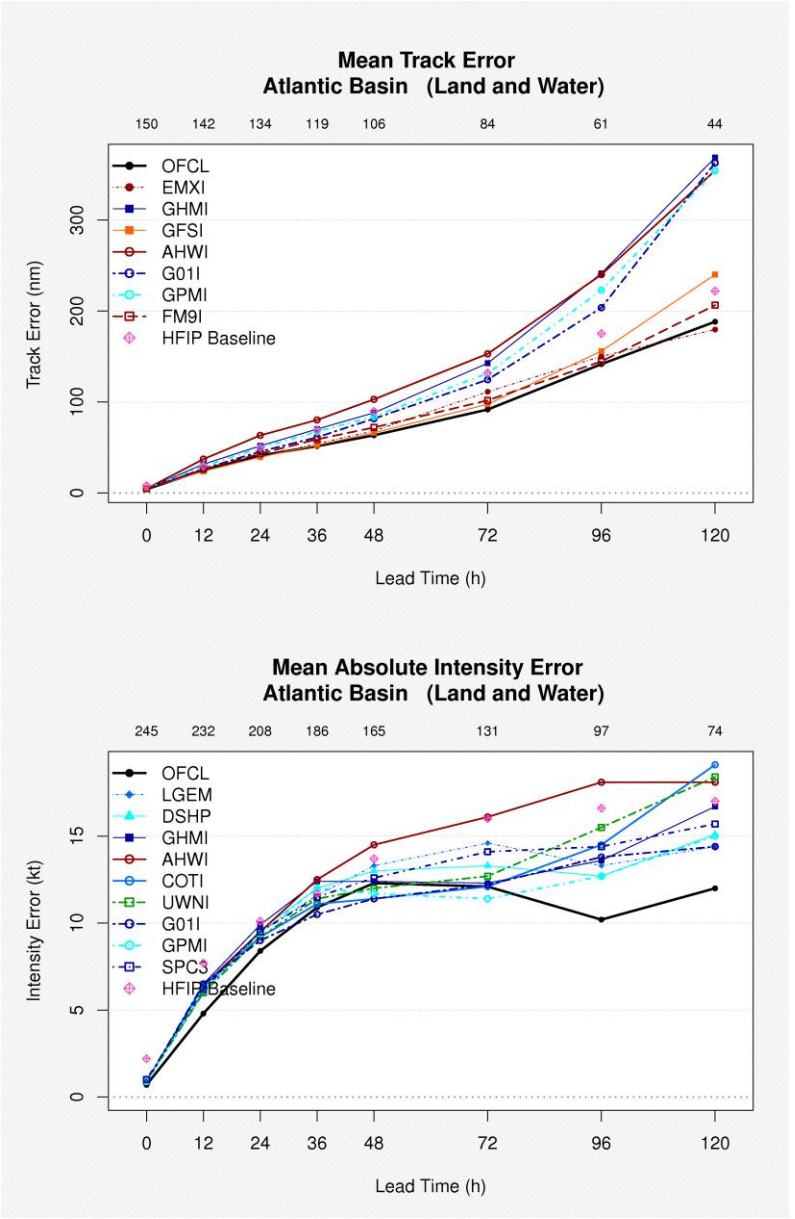
All graphics are available at:

http://www.ral.ucar.edu/projects/hfip/d2012/verify/

2012 DEMO

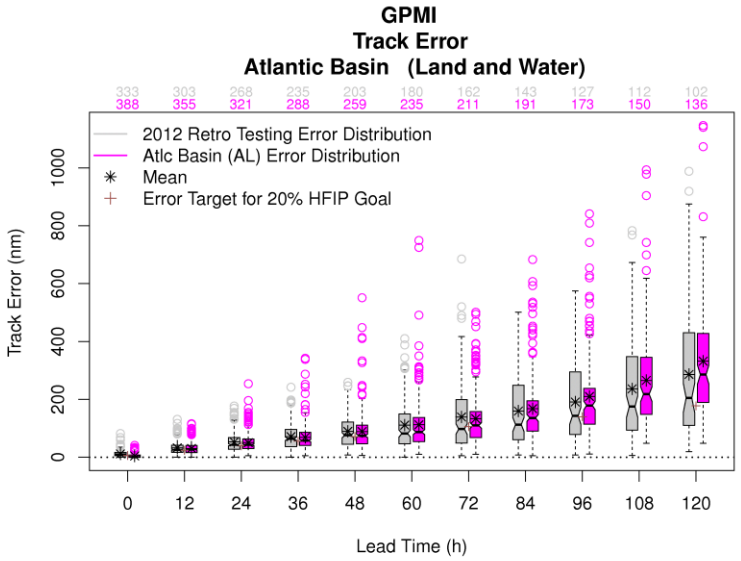
Demonstration Evaluation

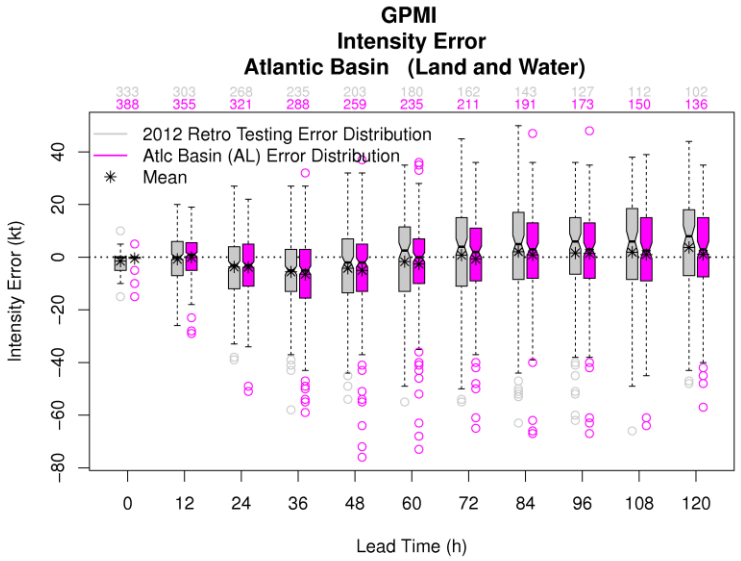
- Stream 1.5, 2.0 and operational models were evaluated for the 2012 HFIP Demonstration
- Models were evaluated with a homogeneous sample
- A variety of evaluations were conducted following the methodology of the Retrospective evaluation
- Mean track and intensity errors are presented on the right



Retrospective vs. Demonstration Evaluation

- Comparison of track and intensity error distributions
  - Retrospective (gray) vs. Demonstration (magenta) evaluations
  - All stream 1.5 candidates were evaluated
  - Example is for GPMI: error distributions have similar characteristics
- All results were summarized in statistically significant tables and compared between Retro and Demo performance





Retrospective vs. Demonstration Evaluation – SS Evaluation

Example – Consensus w/AHWI – 2012 Retrospective

	Forecast Hour	0	12	24	36	48	60	72	84	96	108	120
Atlantic Basin	TVCA	0.0	0.0	0.8	1.4	2.0	1.8	1.9	3.7	8.0	11.2	14.3
	Track	0%	0%	2%	2%	3%	2%	2%	3%	5%	6%	6%
	(Land and Water)	–	0.000	0.954	0.980	0.999	0.954	0.915	0.995	0.999	0.999	0.999
	ICON	0.0	0.1	0.4	0.7	0.8	0.8	0.7	0.2	0.1	0.2	0.4
	Intensity	0%	1%	4%	5%	6%	5%	4%	1%	1%	1%	2%
	(Land and Water)	–	0.682	0.999	0.999	0.999	0.999	0.999	0.495	0.261	0.382	0.575
	ICON	0.0	0.1	0.4	0.7	0.8	0.7	0.5	0.1	0.0	0.1	0.2
	Intensity	0%	1%	4%	5%	6%	5%	3%	1%	0%	1%	1%
	(Water Only)	–	0.682	0.999	0.999	0.999	0.999	0.987	0.261	0.000	0.197	0.310
Eastern North Pacific Basin	TVCE	0.0	0.3	0.9	1.4	1.5	1.1	0.0	0.8	0.3	0.5	2.0
	Track	0%	1%	2%	2%	2%	1%	0%	1%	0%	0%	1%
	(Land and Water)	–	0.866	0.975	0.919	0.788	0.508	0.000	0.233	0.070	0.085	0.382
	ICON	0.0	-0.1	0.0	0.3	0.6	0.7	0.9	1.1	1.2	1.7	1.8
	Intensity	0%	-2%	0%	2%	4%	4%	5%	6%	6%	9%	10%
	(Land and Water)	–	0.682	0.000	0.865	0.997	0.999	0.974	0.931	0.815	0.875	0.951
	ICON	0.0	-0.1	0.0	0.3	0.6	0.6	0.7	0.8	0.8	1.3	1.4
	Intensity	0%	-2%	0%	2%	4%	4%	4%	4%	4%	7%	8%
	(Water Only)	–	0.682	0.000	0.865	0.997	0.997	0.918	0.815	0.624	0.803	0.993

Retrospective vs. Demonstration Evaluation – SS Evaluation

Example – Consensus w/AHWI – 2012 Demonstration

	Forecast Hour	0	12	24	36	48	60	72	84	96	108	120
Atlantic Basin	TVCA	0.0	0.0	-0.8	-0.6	-1.0	-1.6	-3.9	-4.5	-7.0	-5.9	-5.5
	Track	0%	0%	-2%	-1%	-2%	-2%	-4%	-4%	-5%	-3%	-2%
	(Land and Water)	–	0.000	0.889	0.608	0.682	0.625	0.850	0.631	0.779	0.538	0.432
	ICON	0.0	0.1	0.2	0.2	0.0	0.1	0.4	0.2	0.2	0.5	0.5
	Intensity	0%	2%	2%	2%	0%	1%	4%	2%	2%	4%	4%
	(Land and Water)	–	0.682	0.953	0.953	0.000	0.197	0.816	0.494	0.310	0.680	0.786
	ICON	0.0	0.2	0.2	0.1	-0.1	-0.2	0.2	0.4	0.6	0.5	0.0
	Intensity	0%	4%	2%	1%	-1%	-2%	2%	3%	5%	4%	0%
	(Water Only)	–	0.954	0.953	0.382	0.261	0.310	0.310	0.494	0.680	0.785	0.000
Eastern North Pacific Basin	TVCE	0.0	-0.2	-2.0	-1.1	-1.6	0.6	2.7	5.9	16.2	46.6	-9.4
	Track	0%	-1%	-6%	-2%	-3%	1%	3%	6%	14%	32%	-10%
	(Land and Water)	–	0.382	0.902	0.481	0.476	0.140	0.320	0.306	0.600	0.995	–
	ICON	0.0	-0.1	-0.4	-0.2	0.0	-0.2	-0.1	-0.6	-1.1	-0.2	3.8
	Intensity	0%	-2%	-6%	-2%	0%	-3%	-1%	-8%	-11%	-2%	22%
	(Land and Water)	–	0.681	0.952	0.680	0.000	0.493	0.260	0.947	0.918	0.193	–
	ICON	0.0	-0.1	-0.4	-0.2	0.0	-0.2	-0.1	-0.6	-1.1	-0.2	3.8
	Intensity	0%	-2%	-6%	-2%	0%	-3%	-1%	-8%	-11%	-2%	22%
	(Water Only)	–	0.681	0.952	0.680	0.000	0.493	0.260	0.947	0.918	0.193	–

Retrospective vs. Demonstration Evaluation – SS Evaluation

Example – SPC3 – 2012 Retrospective

Forecast Hour		0	12	24	36	48	60	72	84	96	108	120
Atlantic Basin	GHMI	0.0	0.5	1.0	2.2	2.6	3.9	5.4	6.3	6.8	7.3	8.5
	Intensity	0%	6%	8%	15%	16%	21%	26%	29%	31%	33%	37%
	(Land and Water)	–	0.987	0.999	0.998	0.990	0.999	0.999	0.999	0.999	0.999	0.999
	GHMI	0.0	0.5	1.2	2.3	3.0	4.3	6.0	6.4	6.3	6.4	7.5
	Intensity	0%	6%	10%	15%	18%	23%	30%	30%	30%	31%	34%
	(Water Only)	–	0.987	0.983	0.996	0.993	0.998	0.999	0.999	0.998	0.997	0.997
	LGEM	0.0	0.2	0.4	1.0	1.1	1.8	1.5	1.3	0.6	1.0	1.7
	Intensity	0%	3%	4%	7%	7%	11%	9%	8%	4%	6%	11%
	(Land and Water)	–	0.954	0.999	0.999	0.994	0.999	0.997	0.936	0.546	0.732	0.910
	LGEM	0.0	0.1	0.5	1.0	1.0	1.5	1.1	0.8	0.4	0.6	1.0
	Intensity	0%	1%	4%	7%	7%	10%	7%	5%	3%	4%	6%
	(Water Only)	–	0.682	0.999	0.999	0.987	0.997	0.972	0.746	0.311	0.451	0.681
	DSHP	0.0	0.1	0.4	1.0	1.2	1.8	1.6	1.6	1.4	2.5	2.5
	Intensity	0%	1%	4%	7%	8%	11%	10%	9%	8%	15%	15%
	(Land and Water)	–	0.682	0.954	0.999	0.983	0.990	0.954	0.889	0.756	0.925	0.925
	DSHP	0.0	0.1	0.7	1.4	1.4	1.9	1.5	1.1	1.1	1.8	1.8
	Intensity	0%	1%	6%	10%	9%	12%	9%	7%	7%	11%	11%
	(Water Only)	–	0.682	0.999	0.999	0.995	0.993	0.938	0.728	0.602	0.769	0.768

Retrospective vs. Demonstration Evaluation – SS Evaluation

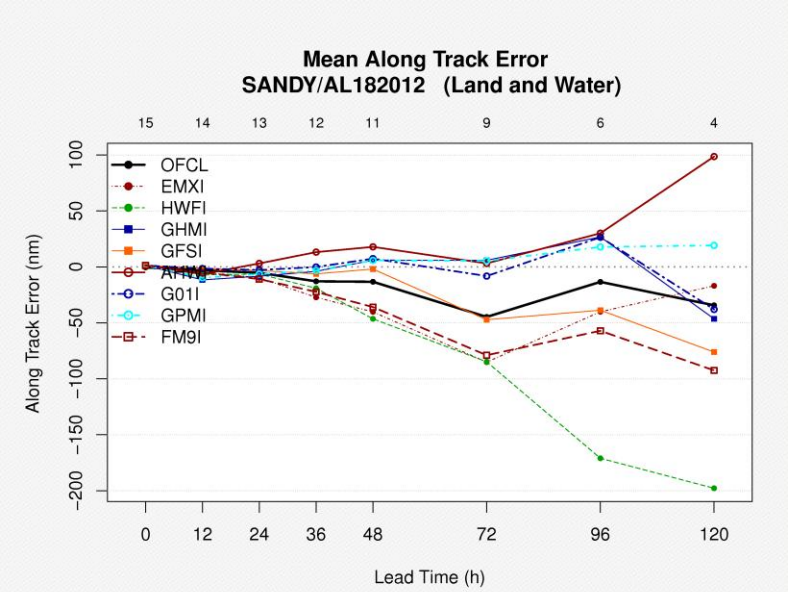
Example – SPC3 – 2012 Demonstration

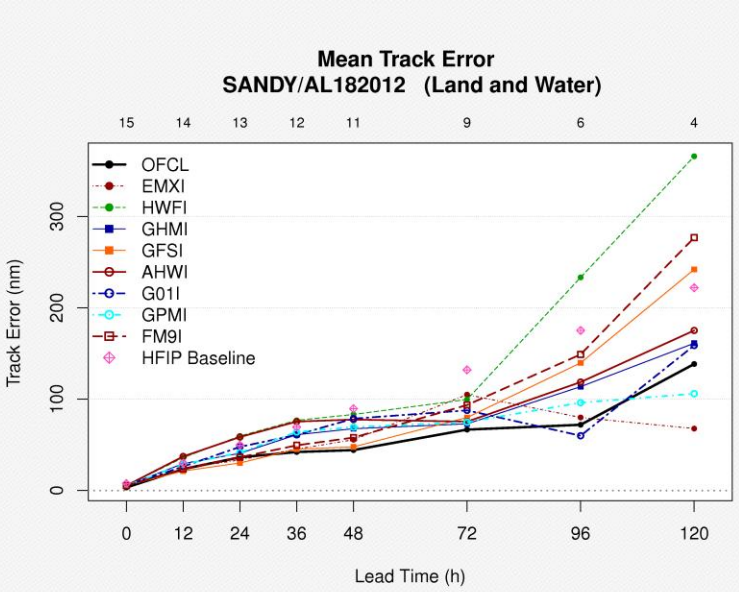
Forecast Hour		0	12	24	36	48	60	72	84	96	108	120
Atlantic Basin	GHMI	-0.1	0.3	0.7	1.1	0.2	-1.3	-1.6	-1.9	-1.5	-0.8	0.1
	Intensity	-10%	5%	7%	9%	2%	-10%	-12%	-14%	-11%	-5%	1%
	(Land and Water)	0.999	0.682	0.682	0.682	0.099	0.505	0.532	0.517	0.450	0.232	0.020
	GHMI	-0.1	0.4	1.1	1.7	0.2	-1.3	-1.1	-1.3	0.3	1.8	2.9
	Intensity	-10%	6%	11%	14%	2%	-10%	-8%	-9%	2%	11%	17%
	(Water Only)	0.999	0.816	0.883	0.842	0.084	0.445	0.327	0.315	0.068	0.437	0.696
	LGEM	-0.1	-0.1	0.1	0.3	0.9	1.1	0.7	-0.5	-0.4	-0.4	-0.6
	Intensity	-10%	-2%	1%	3%	7%	7%	5%	-3%	-3%	-3%	-4%
	(Land and Water)	0.999	0.682	0.383	0.865	0.975	0.994	0.918	0.524	0.575	0.431	0.768
	LGEM	-0.1	-0.1	0.0	0.4	0.9	1.0	0.6	0.1	-0.1	-0.5	-0.8
	Intensity	-10%	-2%	0%	4%	7%	7%	4%	1%	-1%	-4%	-6%
	(Water Only)	0.999	0.682	0.000	0.816	0.974	0.999	0.864	0.158	0.113	0.467	0.680
	DSHP	-0.1	0.0	0.2	0.3	0.5	0.1	-0.1	-1.1	-0.7	-0.1	-0.3
	Intensity	-10%	0%	2%	3%	4%	1%	-1%	-8%	-5%	-1%	-2%
	(Land and Water)	0.999	0.000	0.682	0.682	0.787	0.158	0.099	0.681	0.440	0.047	0.140
	DSHP	-0.1	-0.1	0.2	0.5	0.8	0.7	0.7	0.5	0.0	0.0	-0.9
	Intensity	-10%	-2%	2%	4%	6%	5%	5%	3%	0%	0%	-7%
	(Water Only)	0.999	0.682	0.682	0.903	0.953	0.837	0.755	0.350	0.000	0.000	0.382

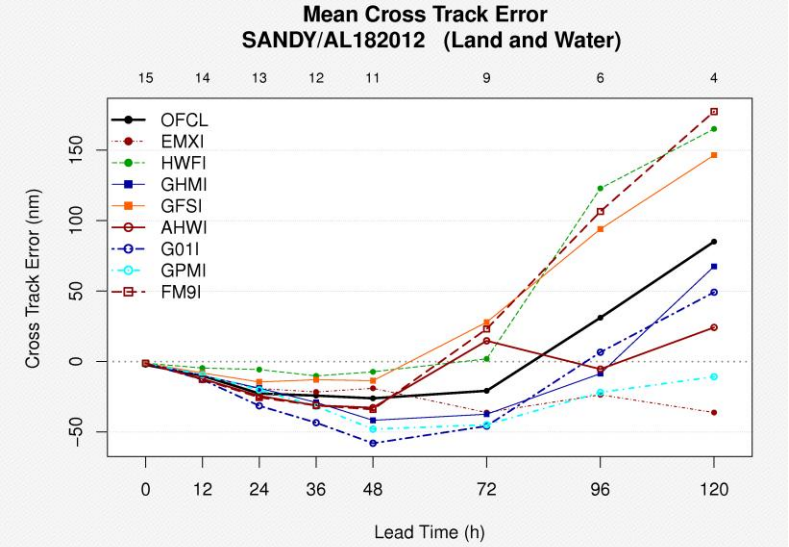
Summary

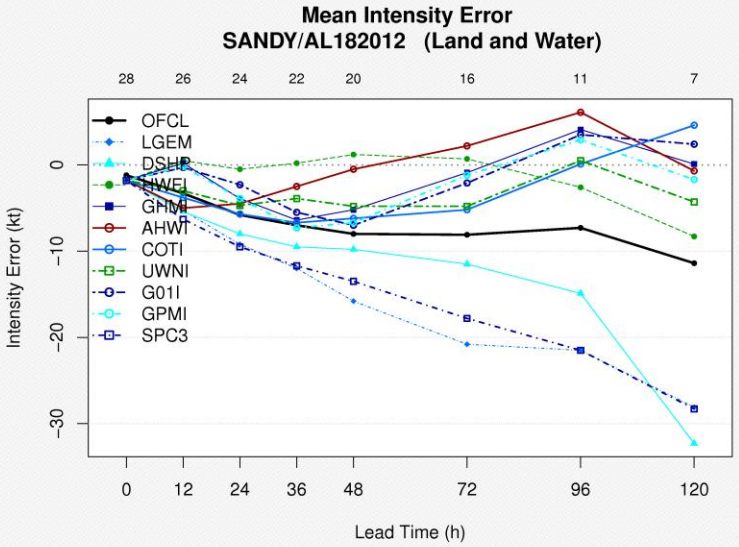
- All stream 1.5 models showed a neutral to a degradation in performance for the Demo compared to Retro evaluation
  - A few individual lead times did show SS improvements
- Tropical storms during the 2012 Demo were characteristically different than Retro cases?

Hurricane Sandy Evaluation



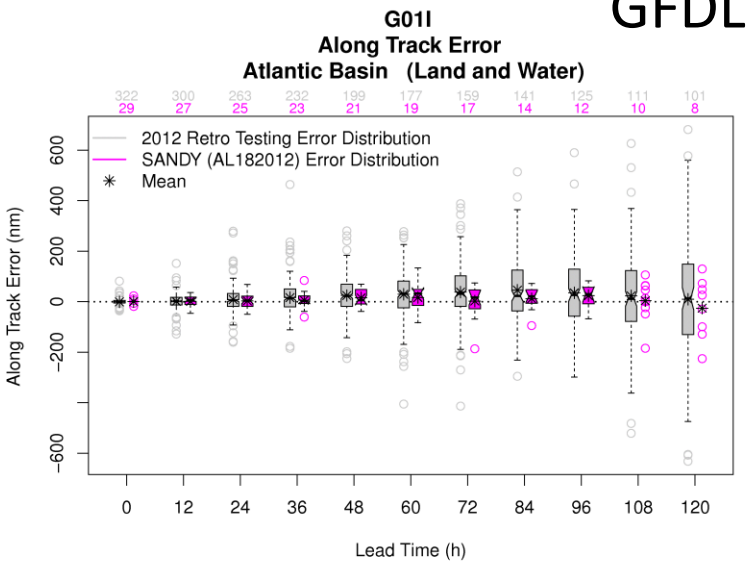


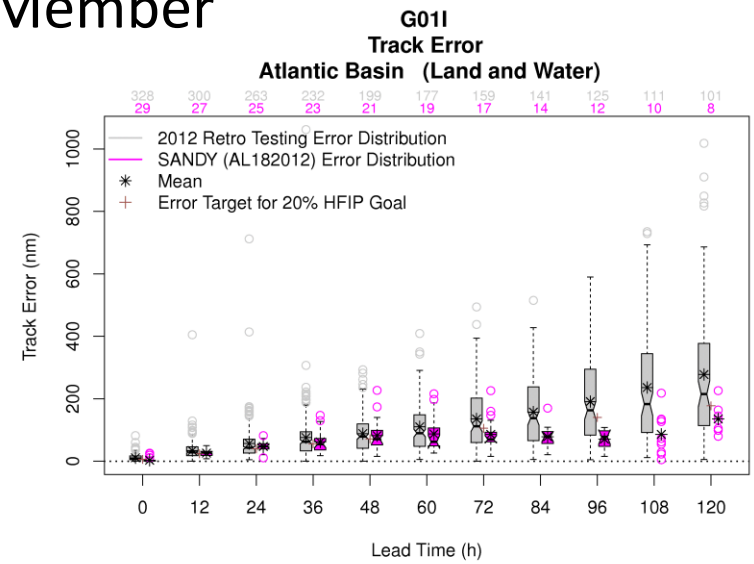


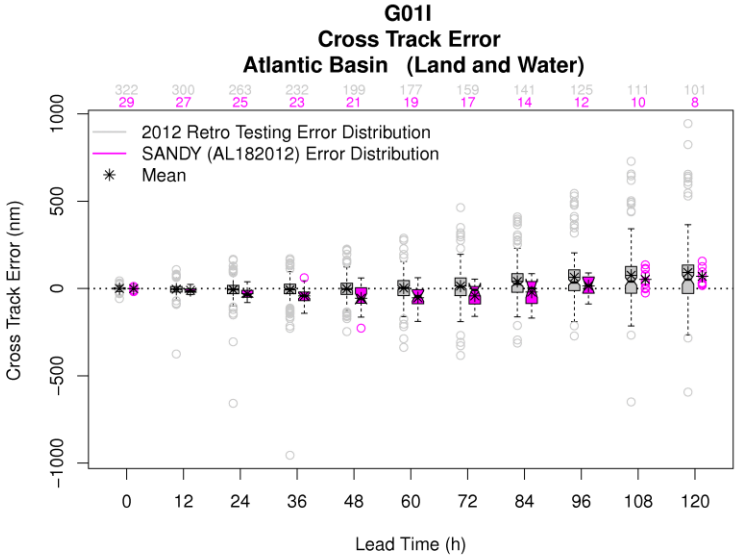


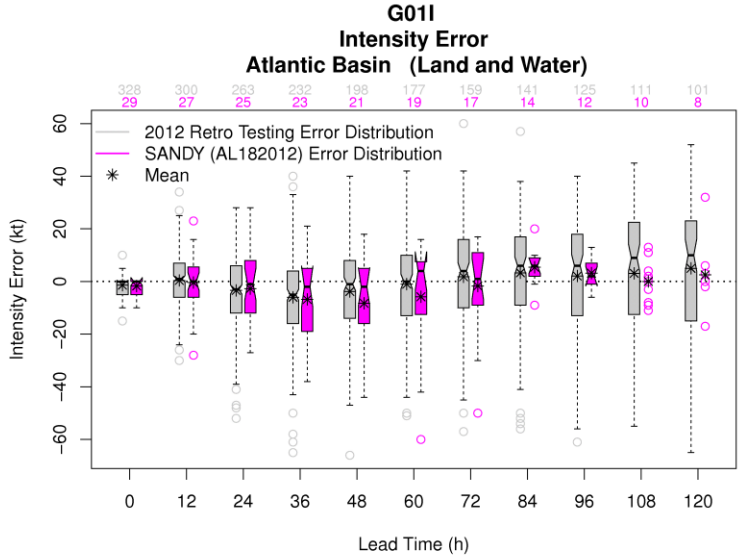
Hurricane Sandy Evaluation

GFDL 1st Ensemble Member



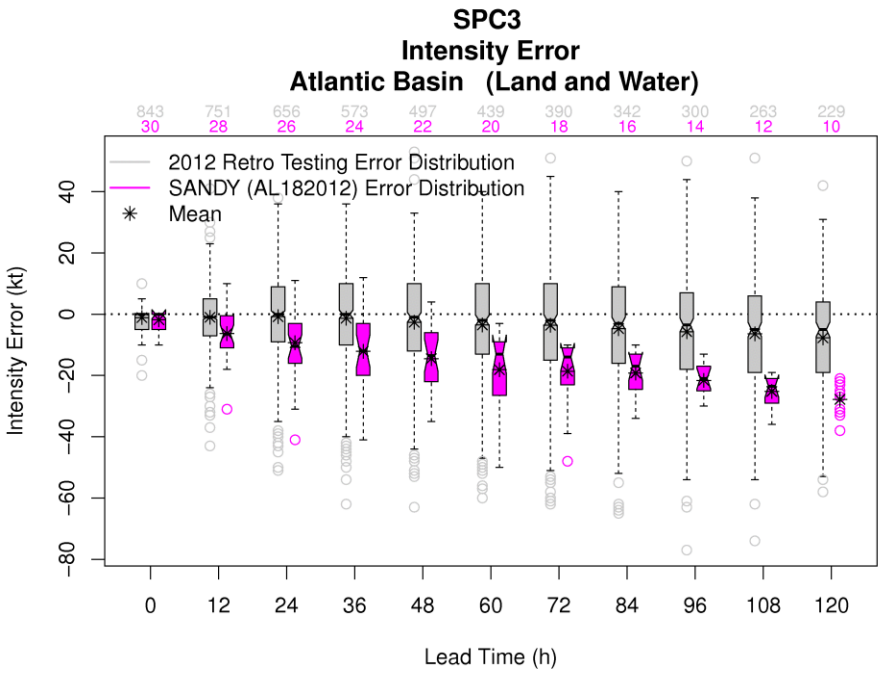


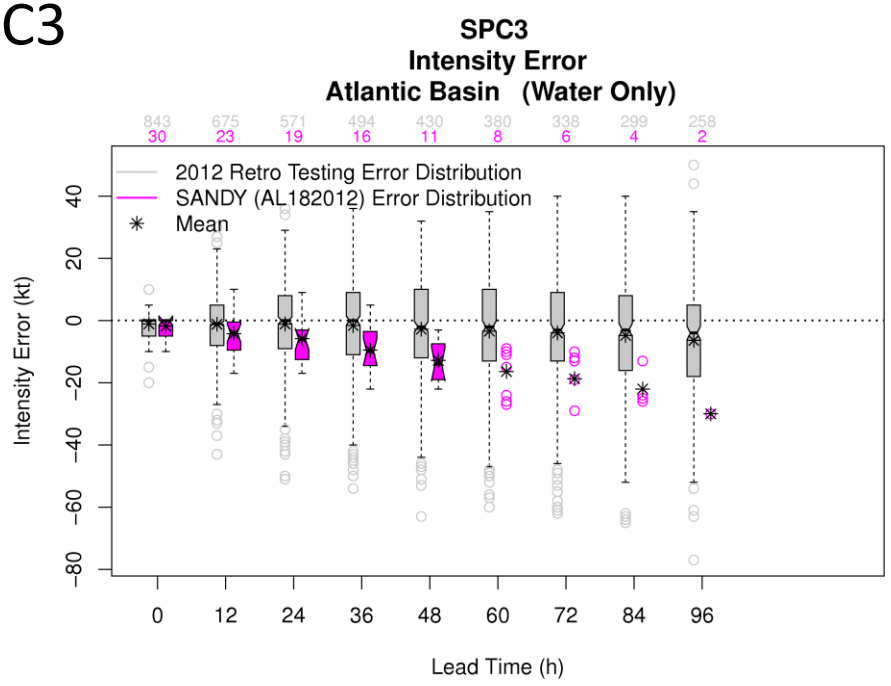




Hurricane Sandy Evaluation

SPC3





Online Access to HFIP Demonstration Evaluation Results

- Evaluation graphics are available on the TCMT website:
  - http://www.ral.ucar.edu/projects/hfip/d2012/verify/
- Wide variety of evaluation statistics are available:
  - Aggregated by basin or storm
  - Aggregated by land/water, or water only
  - Different plot types: error distributions, line plots, rank histogram, Demo vs. Retro
  - A variety of variables and baselines to evaluate