U.S. NAVAL
RESEARCH
LABORATORY

Validation of rapid intensification forecasts from deterministic regional dynamical models

(… and some ensemble forecast products, time permitting)

Jonathan Moskaitis

Naval Research Laboratory, Monterey, CA

HFIP teleconference

03/08/2017

U.S. Naval Research Laboratory
Advancing research further than you can imagine ®

RI Validation: Introduction

- **Validation of rapid intensification for 2015 & 2016 real-time dynamical model forecasts of Atlantic, Eastern Pacific, Central Pacific, and Western Pacific TCs**

  CTCX : NRL demo COAMPS-TC with GFS ICs/BCs

  COTC : Operational COAMPS-TC with NAVGEM ICs/BCs

  HWRF : Operational, with GFS ICs/BCs

  GFDL : Operational, with GFS ICs/BCs

  GFDN : Operational, with NAVGEM ICs/BCs

- **Rapid Intensification (RI): 24 h intensity change ≥ 30 kt**

  - RI threshold is ~ 95th percentile of observed 24 h intensity change distribution in the Atlantic and Eastern Pacific (lower percentile in Western Pacific). It is by definition a rare event.

  - RI is a “yes/no” forecast with a “yes/no” observed predictand. Validation is based on the 2 x 2 contingency table and related metrics

RI Validation: Methodology

2 x 2 Contingency Table & Metrics

		RI observed	
		Yes	*No*
RI forecast	*Yes*	**HIT**	**FA**
	No	**MISS**	**CR**

Success rate (high is good)

SR = HIT / (HIT + FA) → Probability RI is observed, given that RI is forecast

Note: False alarm ratio = 1 – Success rate

Prob. of Detection (high is good)

POD = HIT / (HIT + MISS) → Probability RI is forecast, given that RI is observed

Threat Score (high is good)

TS = HIT / (HIT + MISS + FA) → Measure of accuracy with no “credit” for CRs

Note: Misses and false alarms considered equally bad

Bias Ratio (1 is ideal)

BR = (HIT + FA) / (HIT + MISS) → Rate RI is forecast / Rate RI is observed

RI Validation: Methodology

2 x 2 Contingency Table & Metrics

RI observed

RI forecast

	Yes	*No*
Yes	**HIT**	**FA**
No	**MISS**	**CR**

Success rate (high is good)

SR = HIT / (HIT + FA)

Prob. of Detection (high is good)

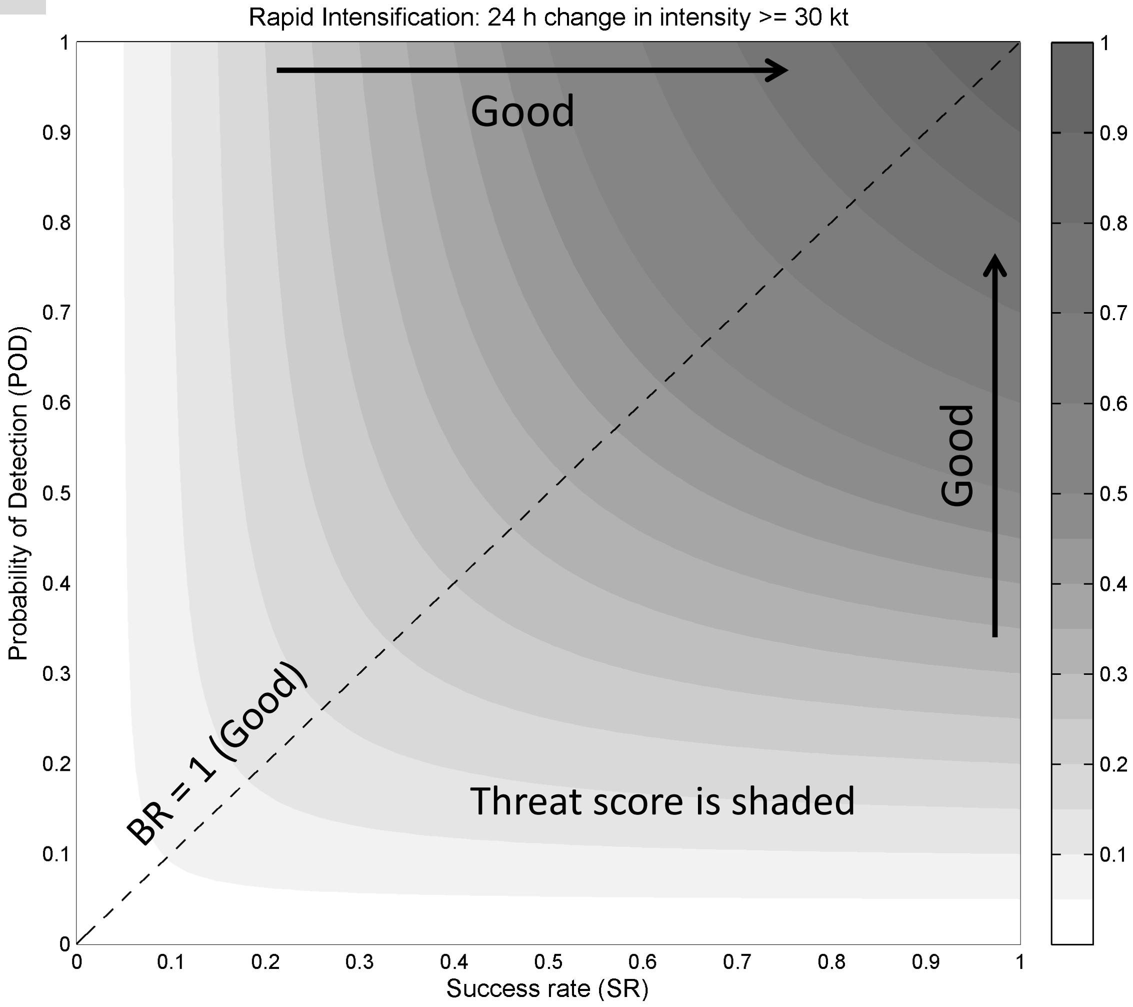
POD = HIT / (HIT + MISS)

Threat Score (high is good)

TS = HIT / (HIT + MISS + FA)

Bias Ratio (1 is ideal)

BR = (HIT + FA) / (HIT + MISS)



Plot adapted from Roebber 2009

RI Validation: Methodology

2 x 2 Contingency Table & Metrics

RI observed

RI forecast		Yes	No
	Yes	**HIT**	**FA**
	No	**MISS**	**CR**

Success rate (high is good)

SR = HIT / (HIT + FA)

Prob. of Detection (high is good)

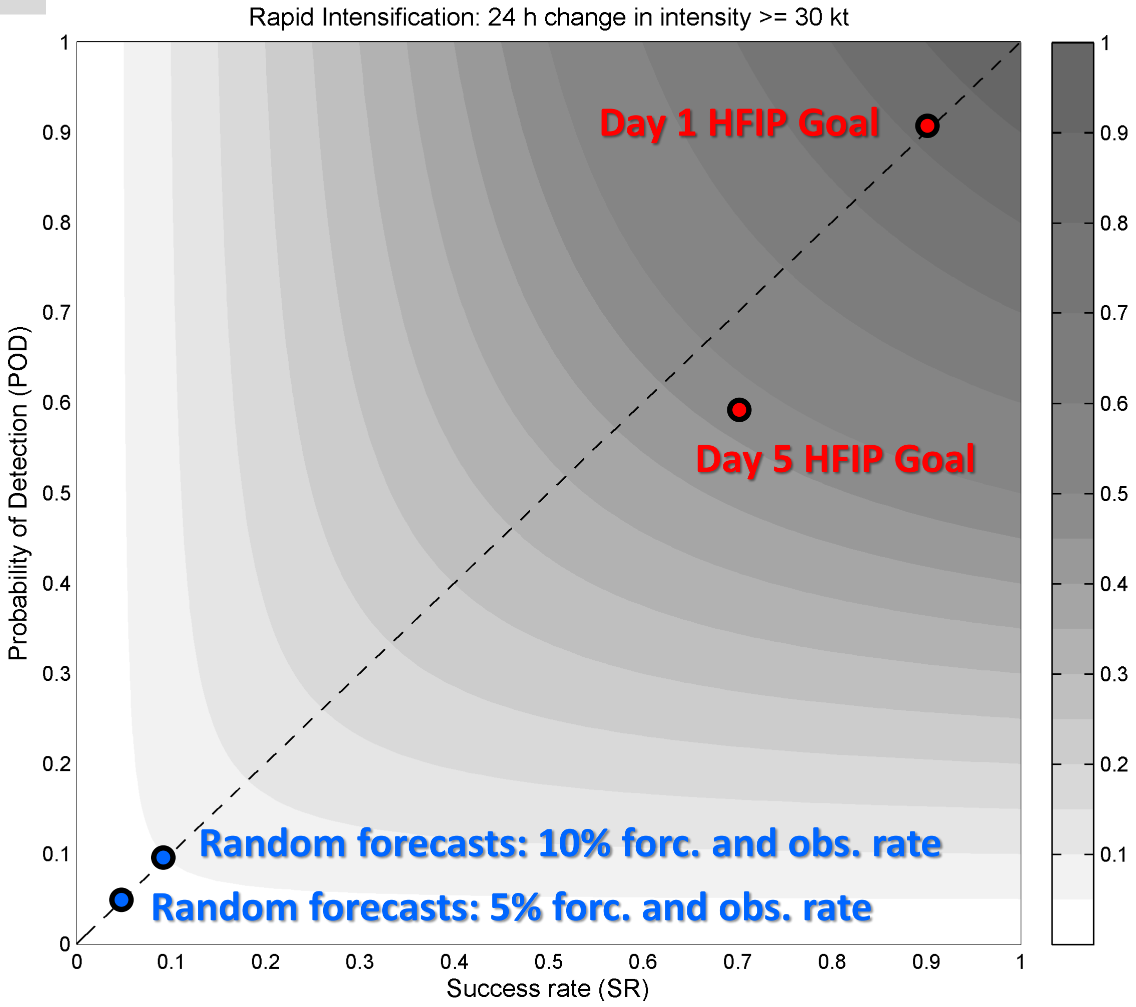
POD = HIT / (HIT + MISS)

Threat Score (high is good)

TS = HIT / (HIT + MISS + FA)

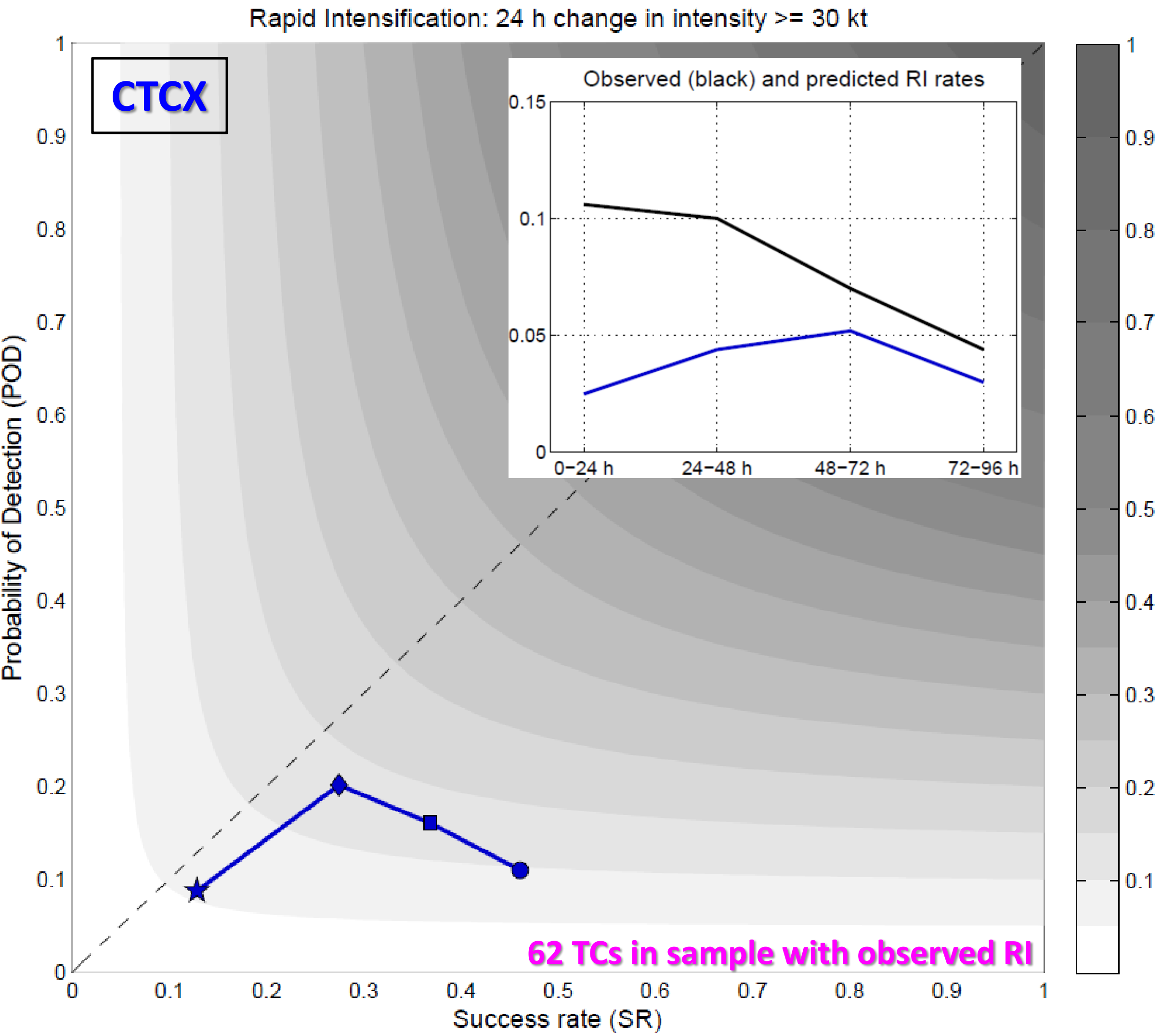
Bias Ratio (1 is ideal)

BR = (HIT + FA) / (HIT + MISS)



Plot adapted from Roebber 2009

RI Validation: Results



62 TCs in sample with observed RI

○	tau = 0−24 h through 18−42 h
□	tau = 24−48 h through 42−66 h
◇	tau = 48−72 h through 66−90 h
☆	tau = 72−96 h through 96−120 h

SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR
POD = prob(RI forecast | RI observed)
Above diag. prob(RI forecast) > prob(RI observed), vice versa below
Threat score (measure of forecast accuracy) grayscale shaded

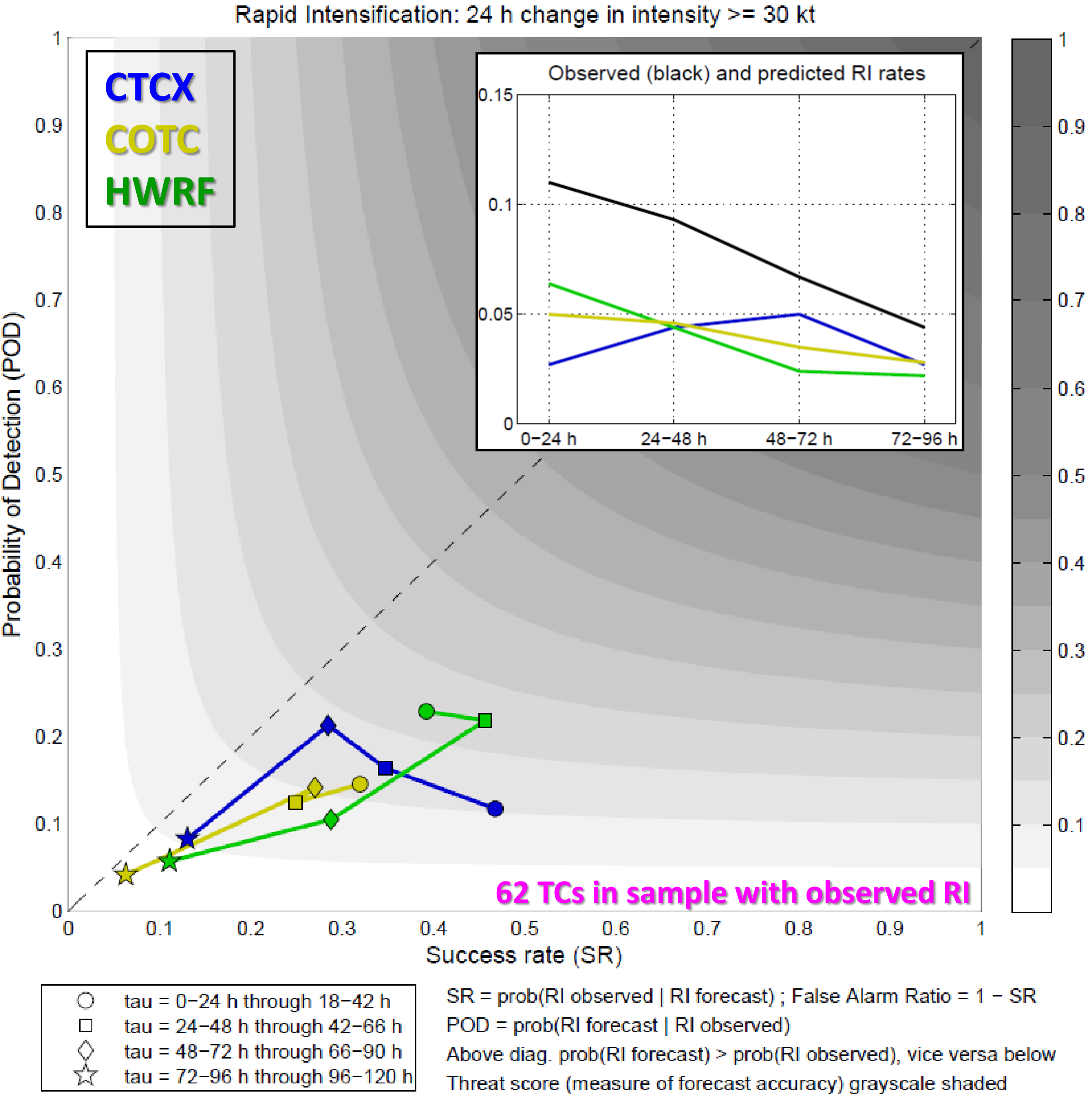
2015 & 2016: All basins

- Results are binned by lead time

  Tau = **0-24 h** through 18-42 h (circle)
  Tau = **24-48 h** through 42-66 h (square)
  Tau = **48-72 h** through 66-90 h (diamond)
  Tau = **72-96 h** through 96-120 h (star)

- Observed rate of RI decreases with forecast lead time
- Forecast rate of RI < Observed rate of RI, especially for early lead times
- Success rate > probability of detection (more misses than false alarms)
- Success rate decreases with lead time
- POD highest for 3rd lead time bin
- Threat score highest for 2nd and 3rd lead time bins

RI Validation: Results



Rapid Intensification: 24 h change in intensity >= 30 kt

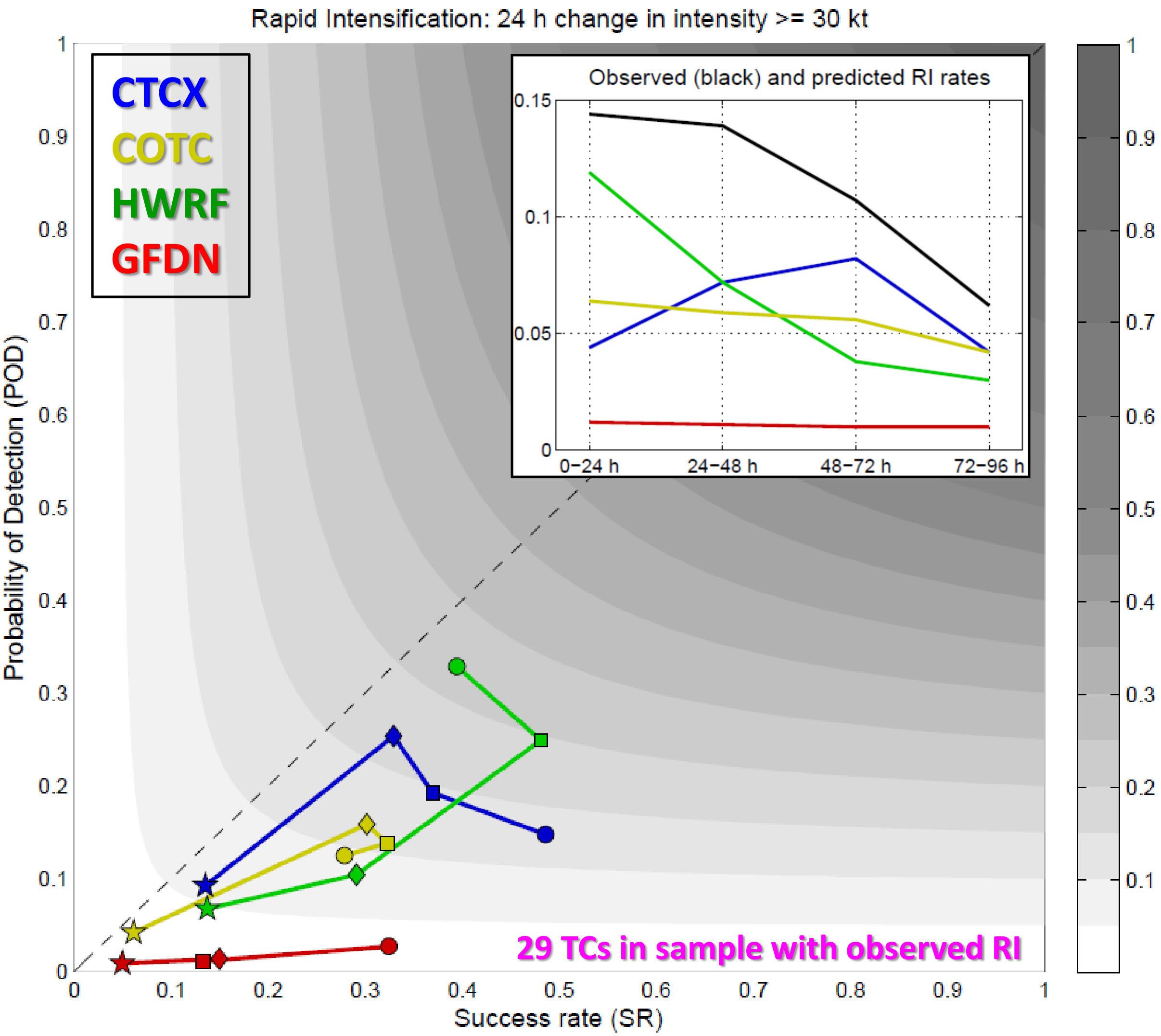
62 TCs in sample with observed RI

SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR
POD = prob(RI forecast | RI observed)
Above diag. prob(RI forecast) > prob(RI observed), vice versa below
Threat score (measure of forecast accuracy) grayscale shaded

2015 & 2016: All basins

- Homogeneous comparison
- All models underpredict the RI rate at all lead times (~0.5x obs. rate)
- Success rate > probability of detection
- Model performance declines with lead time; for last lead time bin metrics are similar to those of random forecasts
- HWRF performs best for first two lead time bins, CTCX for last two lead time bins (based on threat score)
- Dynamical model performance does not approach HFIP goal, but is skillful for the first three lead time bins

RI Validation: Results



SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR

POD = prob(RI forecast | RI observed)

Above diag. prob(RI forecast) > prob(RI observed), vice versa below

Threat score (measure of forecast accuracy) grayscale shaded

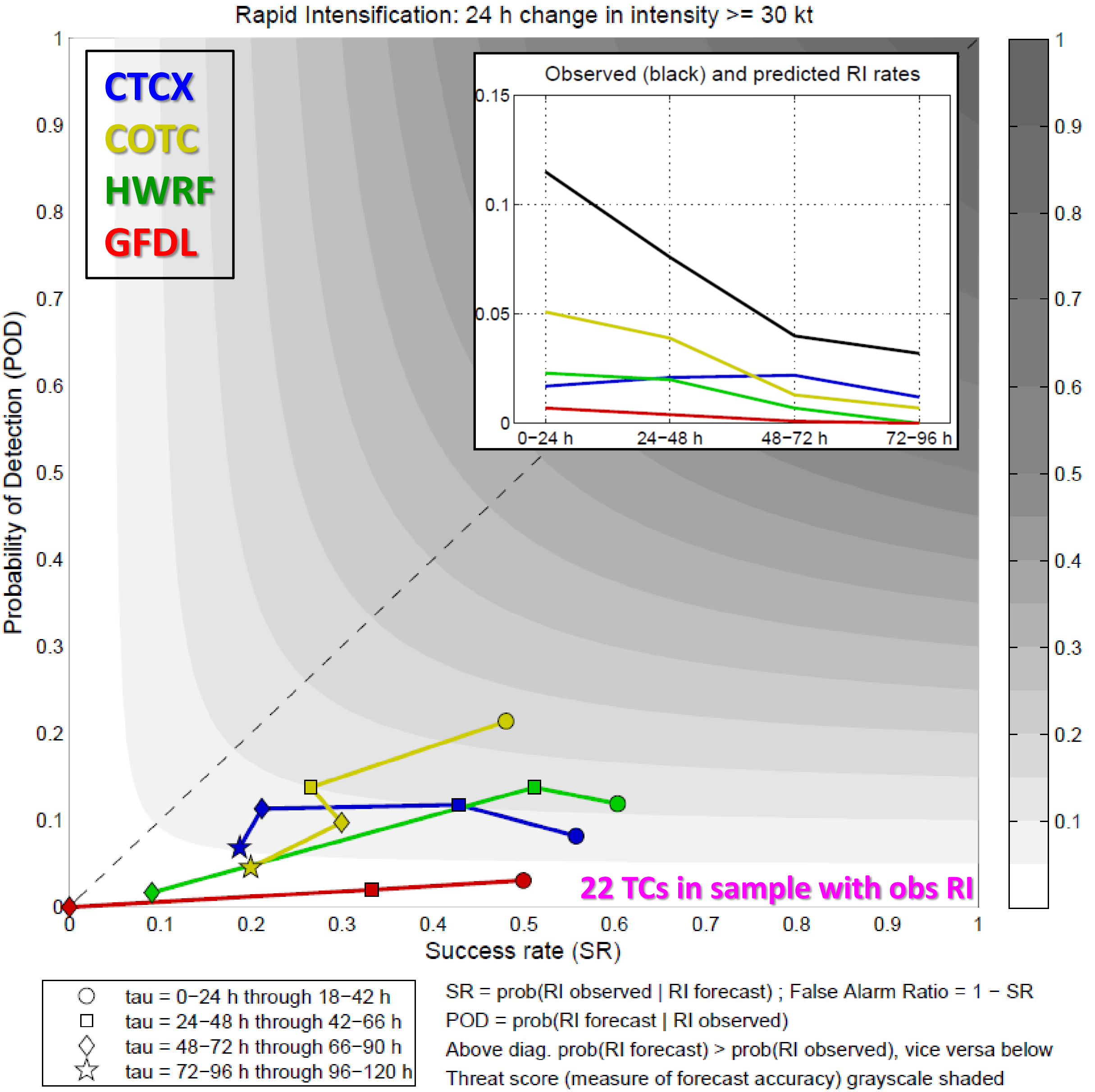
2015 & 2016: WestPac

- Relative to EastPac and Atlantic, observed rate of RI is higher, and model forecast performance is better
- All models underpredict the RI rate at all lead times. HWRF is best at earliest lead time bin and COAMPS-TC at later lead time bins
- Success rate > probability of detection
- HWRF performs best for first two lead time bins, CTCX for last two lead time bins (based on threat score)
- Except for GFDN, dynamical models are skillful for the first three lead time bins

Note: WestPac accounts for roughly half the 'All basins' sample

U.S. NAVAL RESEARCH LABORATORY

RI Validation: Results



SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR

POD = prob(RI forecast | RI observed)

Above diag. prob(RI forecast) > prob(RI observed), vice versa below

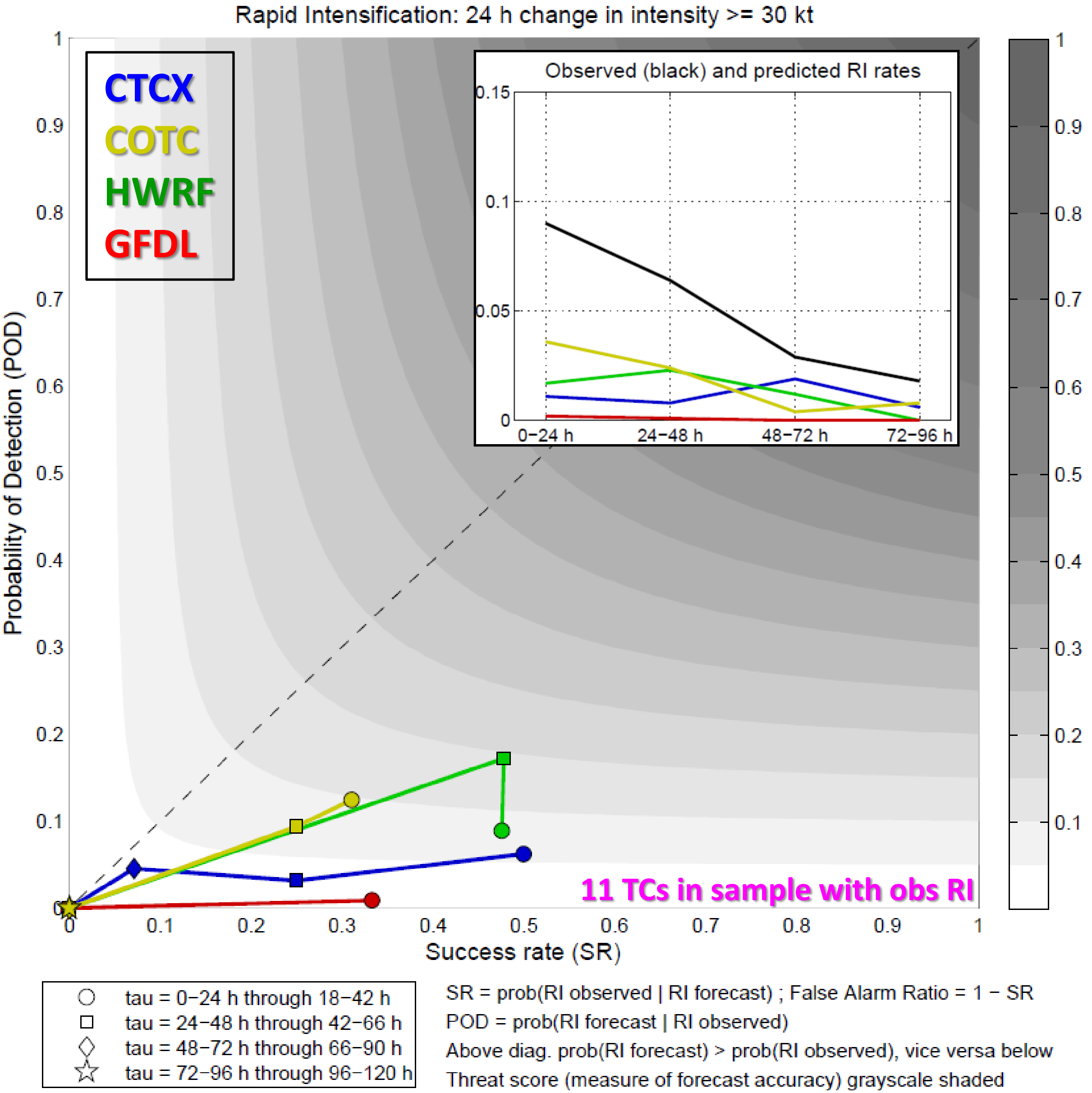
Threat score (measure of forecast accuracy) grayscale shaded

2015 & 2016: EastPac

- All models underpredict the RI rate at all lead times. Early lead times are particularly bad, especially for the GFS-based models
- Success rate >> probability of detection
- COTC best performing model for earliest lead time bin
- COTC and CTCX best performing models at the later lead time bins

RI Validation: Results

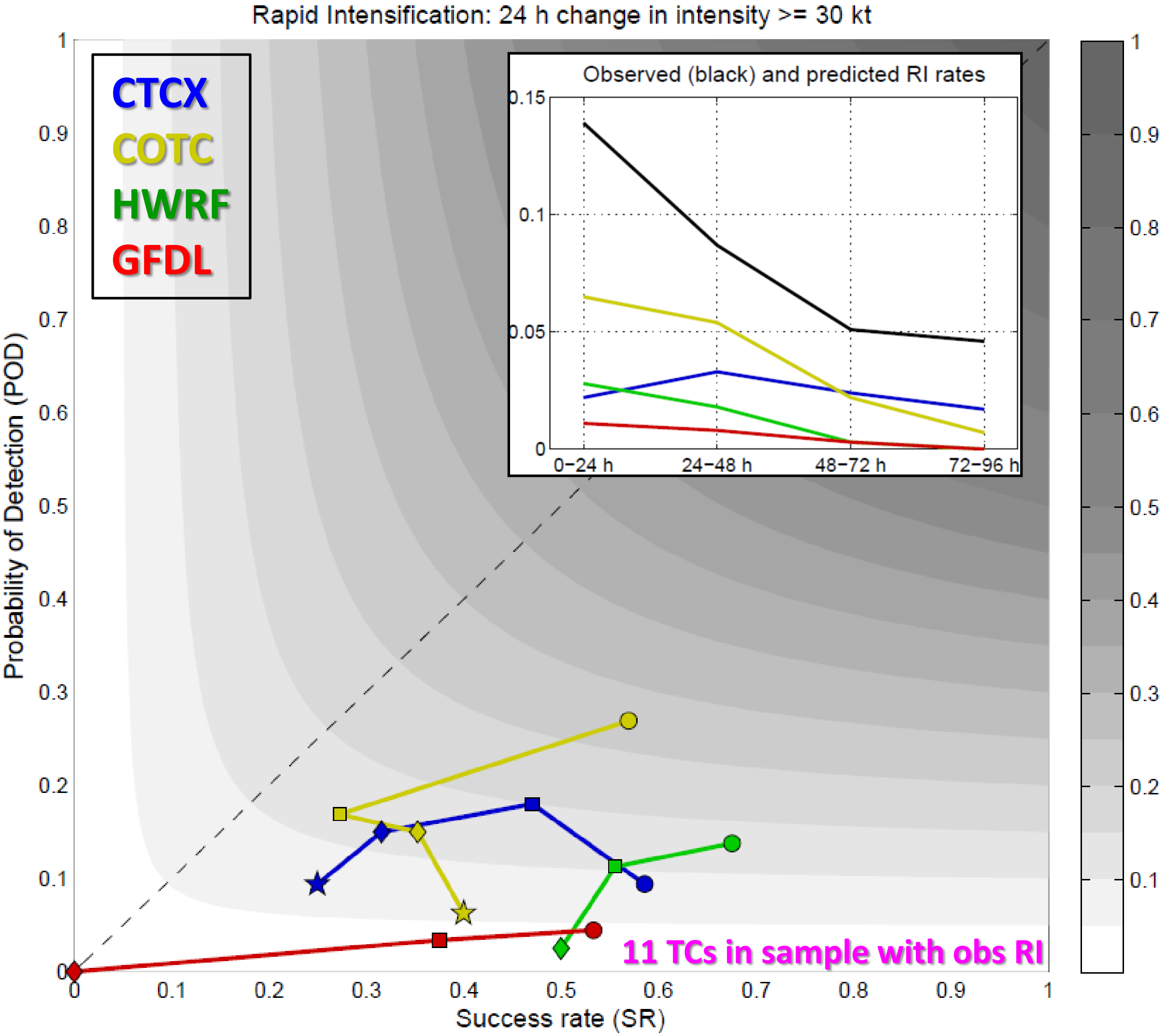
2016: EastPac



11 TCs in sample with obs RI

SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR
POD = prob(RI forecast | RI observed)
Above diag. prob(RI forecast) > prob(RI observed), vice versa below
Threat score (measure of forecast accuracy) grayscale shaded

RI Validation: Results



○ tau = 0−24 h through 18−42 h
□ tau = 24−48 h through 42−66 h
◇ tau = 48−72 h through 66−90 h
☆ tau = 72−96 h through 96−120 h

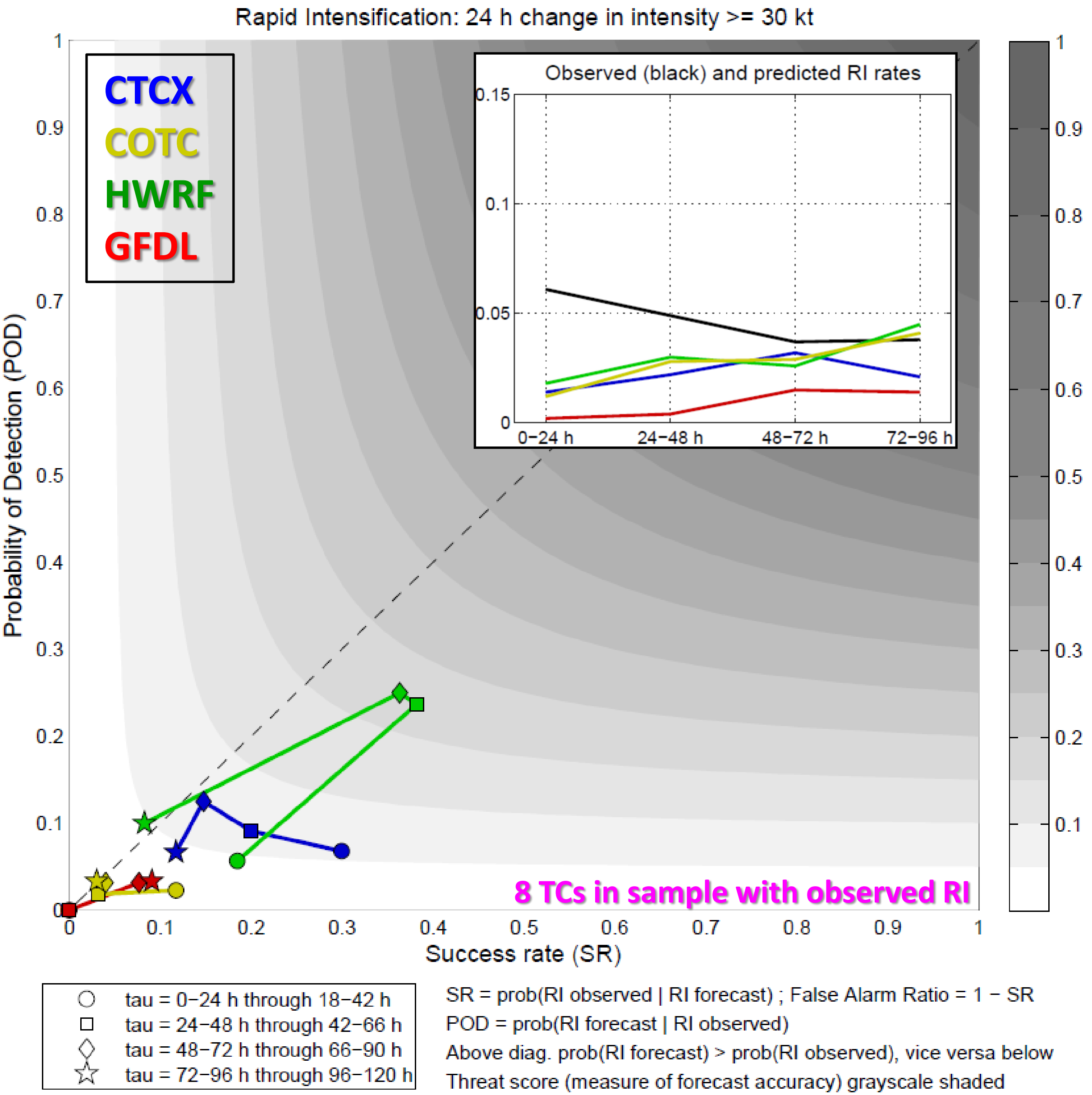
SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR
POD = prob(RI forecast | RI observed)
Above diag. prob(RI forecast) > prob(RI observed), vice versa below
Threat score (measure of forecast accuracy) grayscale shaded

2015: EastPac

- RI cases were apparently easier to predict in 2015 than in 2016. Maybe increased predictability from SST anomalies associated with El Niño?

- Beware of interpreting results for a single season/basin, or year-to-year changes in such results.

RI Validation: Results



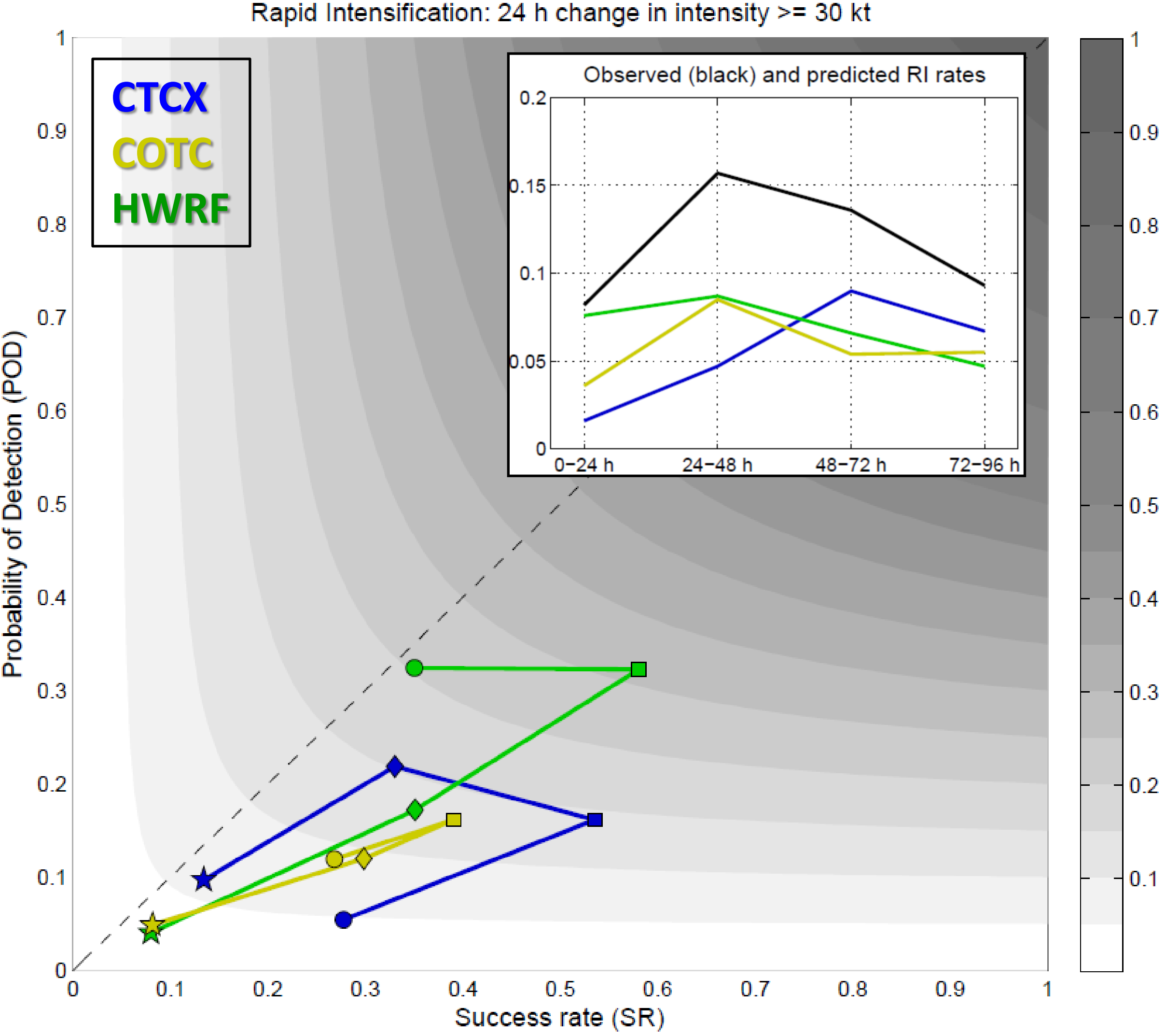
8 TCs in sample with observed RI

SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR
POD = prob(RI forecast | RI observed)
Above diag. prob(RI forecast) > prob(RI observed), vice versa below
Threat score (measure of forecast accuracy) grayscale shaded

2015 & 2016: Atlantic

- With fewer forecast cases and fewer observed RI events in 2015 and 2016 w.r.t. the other basins, undersampling is much bigger issue in Atlantic
- All models underpredict the RI rate at early lead times.
- HWRF and CTCX appear to have some skill, but reluctant to draw conclusions based on this sample

RI Validation: Results



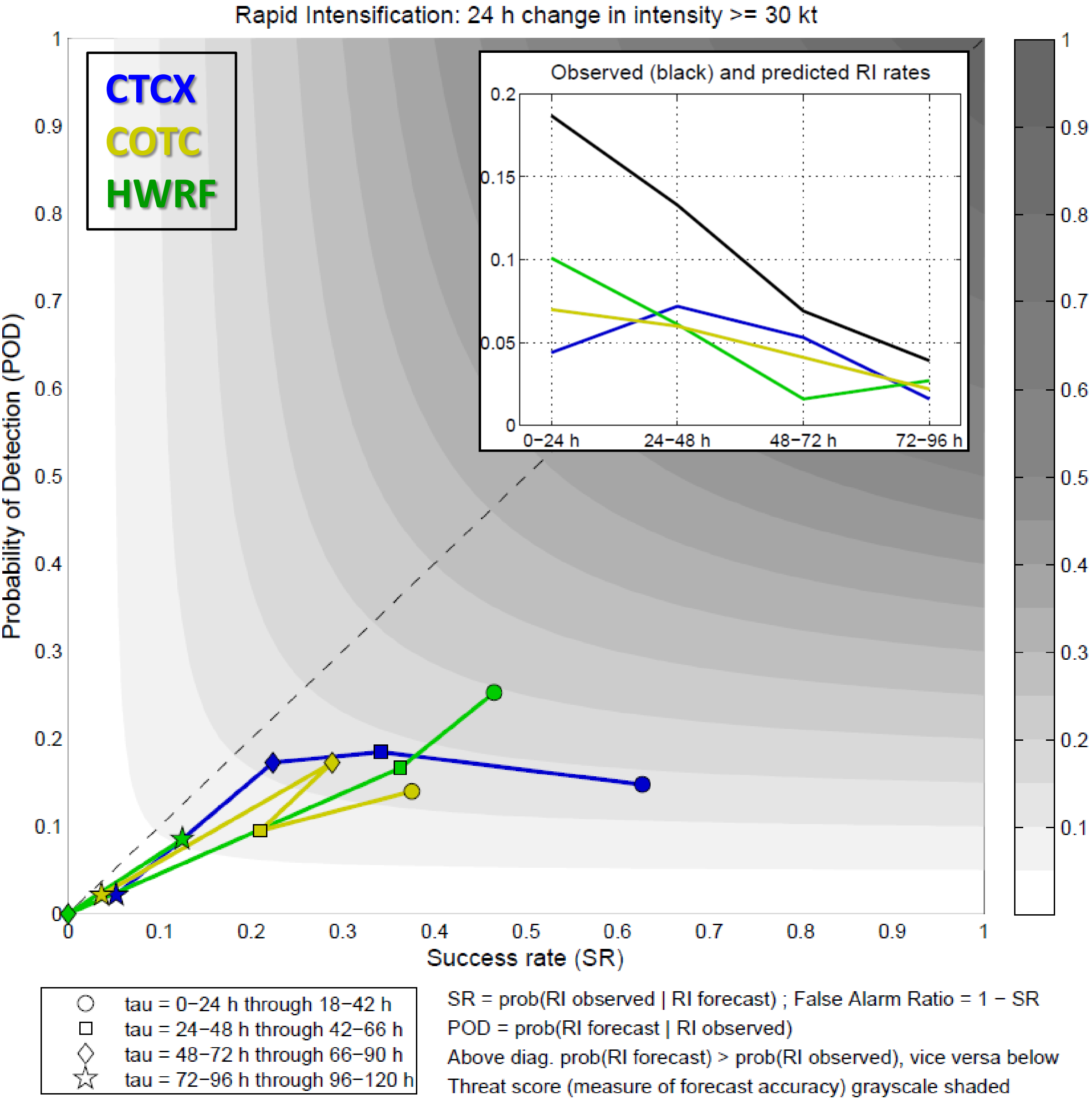
○ tau = 0−24 h through 18−42 h
□ tau = 24−48 h through 42−66 h
◇ tau = 48−72 h through 66−90 h
☆ tau = 72−96 h through 96−120 h

SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR
POD = prob(RI forecast | RI observed)
Above diag. prob(RI forecast) > prob(RI observed), vice versa below
Threat score (measure of forecast accuracy) grayscale shaded

Initial Vmax <= 40 kt

- Cases from 2015 & 2016, All basins
- Focus on results from first lead time bin (circles)
- HWRF has nearly the correct RI rate, COAMPS-TC forecast rate is far too low, especially CTCX
- HWRF has both POD and SR slightly above 0.3

RI Validation: Results



SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR

POD = prob(RI forecast | RI observed)

Above diag. prob(RI forecast) > prob(RI observed), vice versa below

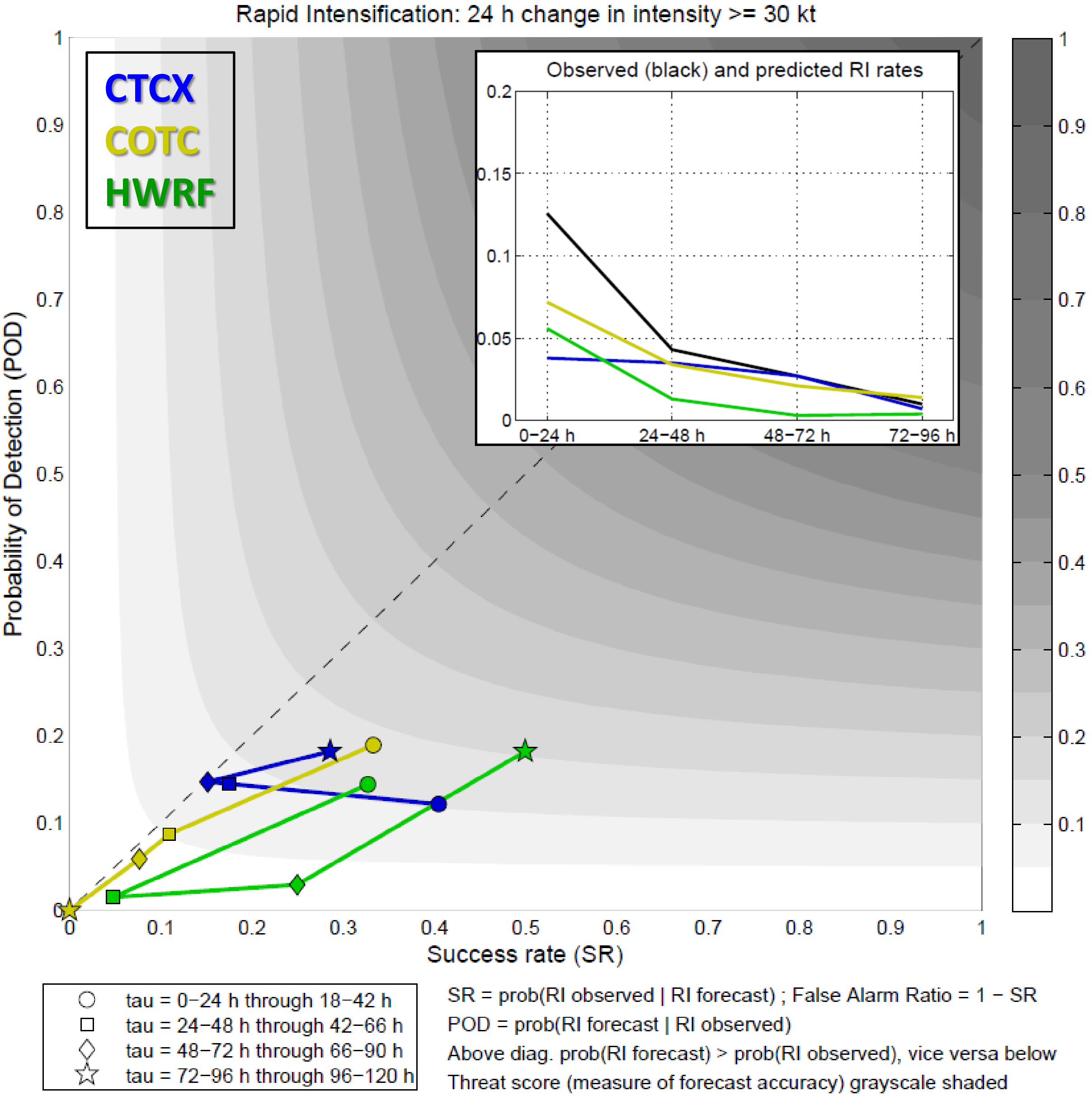
Threat score (measure of forecast accuracy) grayscale shaded

45 kt <= I. Vmax <= 60 kt

- Cases from 2015 & 2016, All basins
- Focus on results from first lead time bin (circles)
- Observed rate of RI is high relative to other categories of initial Vmax
- Models all underestimate obs RI rate
- CTCX has higher success rate than HWRF, but lower POD and threat score

U.S. NAVAL RESEARCH LABORATORY

RI Validation: Results



SR = prob(RI observed | RI forecast) ; False Alarm Ratio = 1 − SR

POD = prob(RI forecast | RI observed)

Above diag. prob(RI forecast) > prob(RI observed), vice versa below

Threat score (measure of forecast accuracy) grayscale shaded

65 kt <= I. Vmax <= 95 kt

- Cases from 2015 & 2016, All basins
- Focus on results from first lead time bin (circles)
- Models all underestimate obs RI rate
- Similar model performance; SR between 0.3 and 0.4, POD between 0.1 and 0.2
- HWRF performance worse than for TCs that are initial of TS & TD intensity

RI Validation: Conclusions

2015 & 2016: All basins

- Sample includes 62 TCs with observed RI, very active WestPac and EastPac
- Dynamical models underpredict (~0.5x) the observed rate of RI at all lead times
- Success rate > Probability of detection; miss more likely than false alarm
- Model performance varies according to TC initial intensity
- Dynamical models have skill for all but the latest lead times, relative to randomly predicting RI at the observed rate. However, performance is well short of HFIP goal.

2015 & 2016: Individual basins

- Performance is generally better in the Western Pacific than Eastern Pacific; Eastern Pacific has relatively low forecast rate of RI and low POD
- Atlantic has too few instances of RI to have a lot of confidence in results

Rapid intensification: Challenges

Validation challenges

- RI is rare by definition; difficult to accumulate sample with many observed RI instances
- Multi-basin, multi-year approach is most likely to give meaningful results, but makes a retrospective test of two model versions very computationally expensive
- Atlantic is particularly troublesome; to get ~60 TCs with observed RI (as in 2015-2016 multi-basin sample), would have to run 2004-2016 seasons.

Prediction challenges

- Models need to forecast RI more often to increase probability of detection … but this will be difficult without degrading success rate (i.e. more false alarms) and intensity mean absolute error
- All models struggle with 0-24 h RI rate for TCs with initial intensity > 40 kt. Why?
- Model performance is better in the Western Pacific than the Eastern Pacific (and Atlantic, perhaps). Why? Is it just that ΔVmax ≥ 30 kt in 24 h is more common in the Western Pacific?

TC ensemble forecast products

Jon Moskaitis, Will Komaromi, Alex Reinecke, Jim Doyle, Hao Jin

- **In 2014, 2015, and 2016 NRL ran a real-time COAMPS-TC ensemble**
- **Forecast products displayed on NRL web page for:**
  - COAMPS-TC ensemble
  - HWRF ensemble
  - GFDL ensemble
  - Multi-model combined ensemble

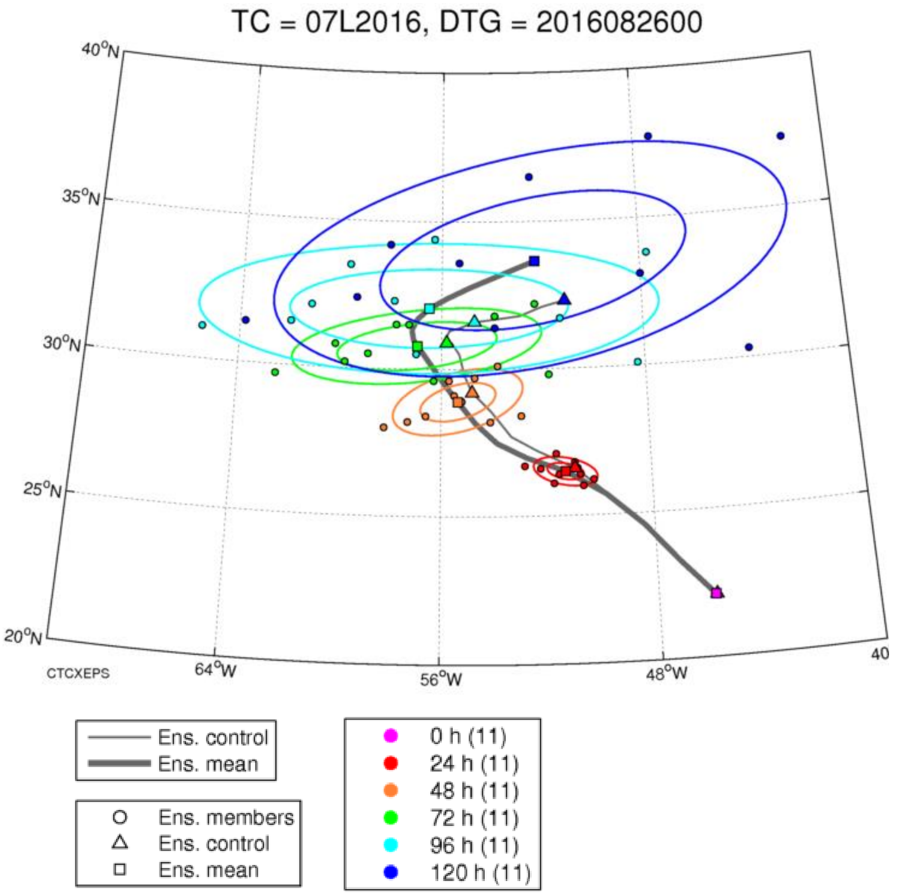
https://www.nrlmry.navy.mil/coamps-web/web/ens

- **Here, we review products available in 2016 and discuss future directions**

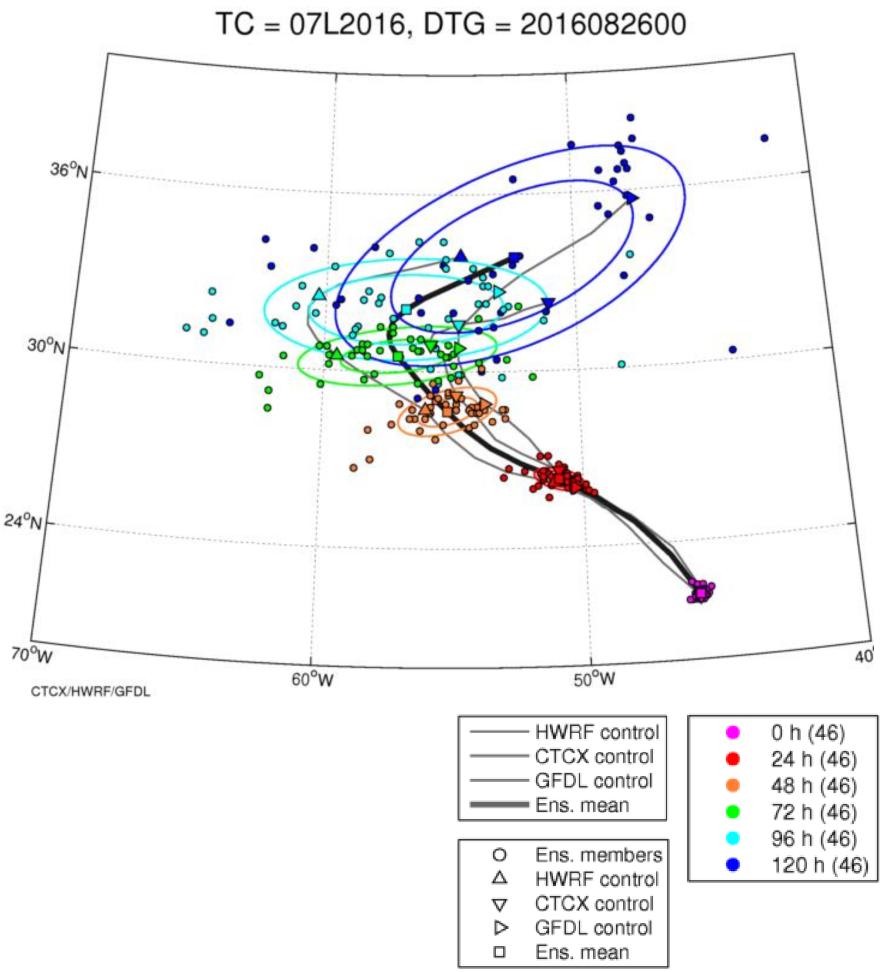
TC ensemble forecast products

Basic track forecast display

COAMPS-TC



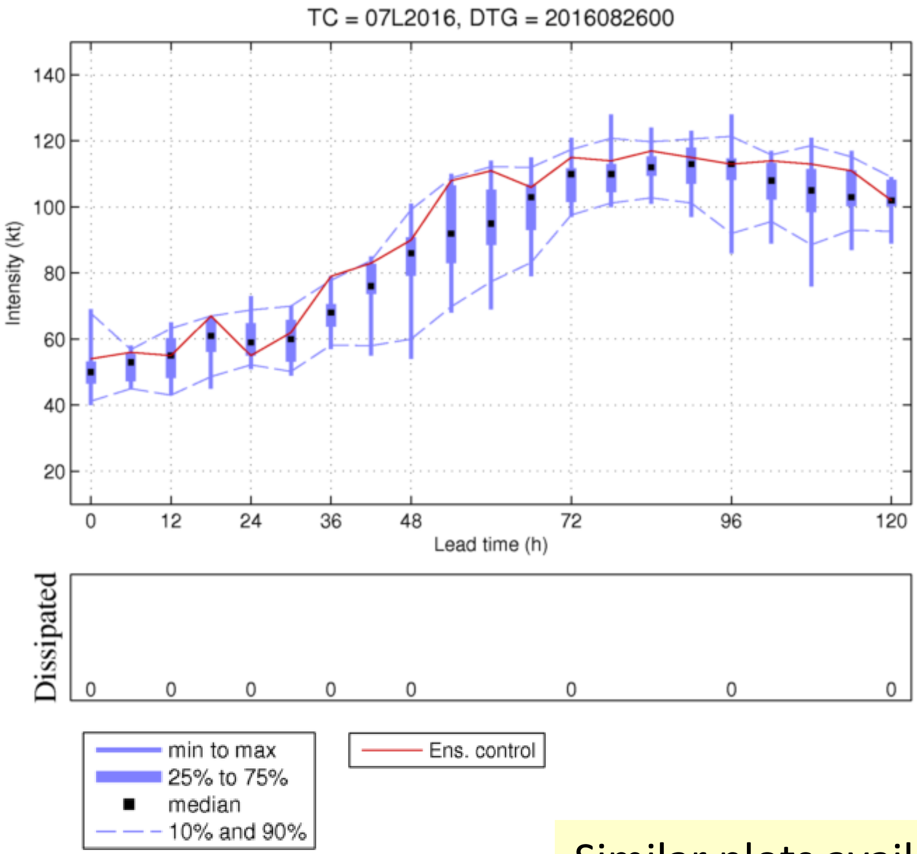
COAMPS-TC / HWRF / GFDL



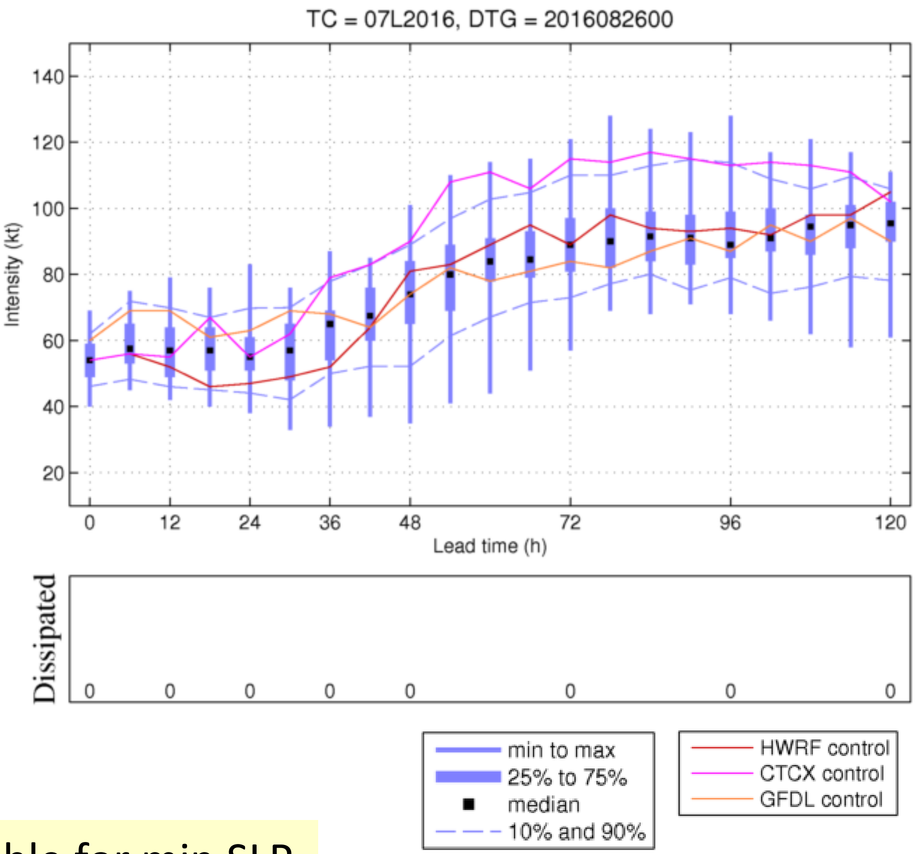
TC ensemble forecast products

Basic intensity forecast display

COAMPS-TC



COAMPS-TC / HWRF / GFDL



Similar plots available for min SLP

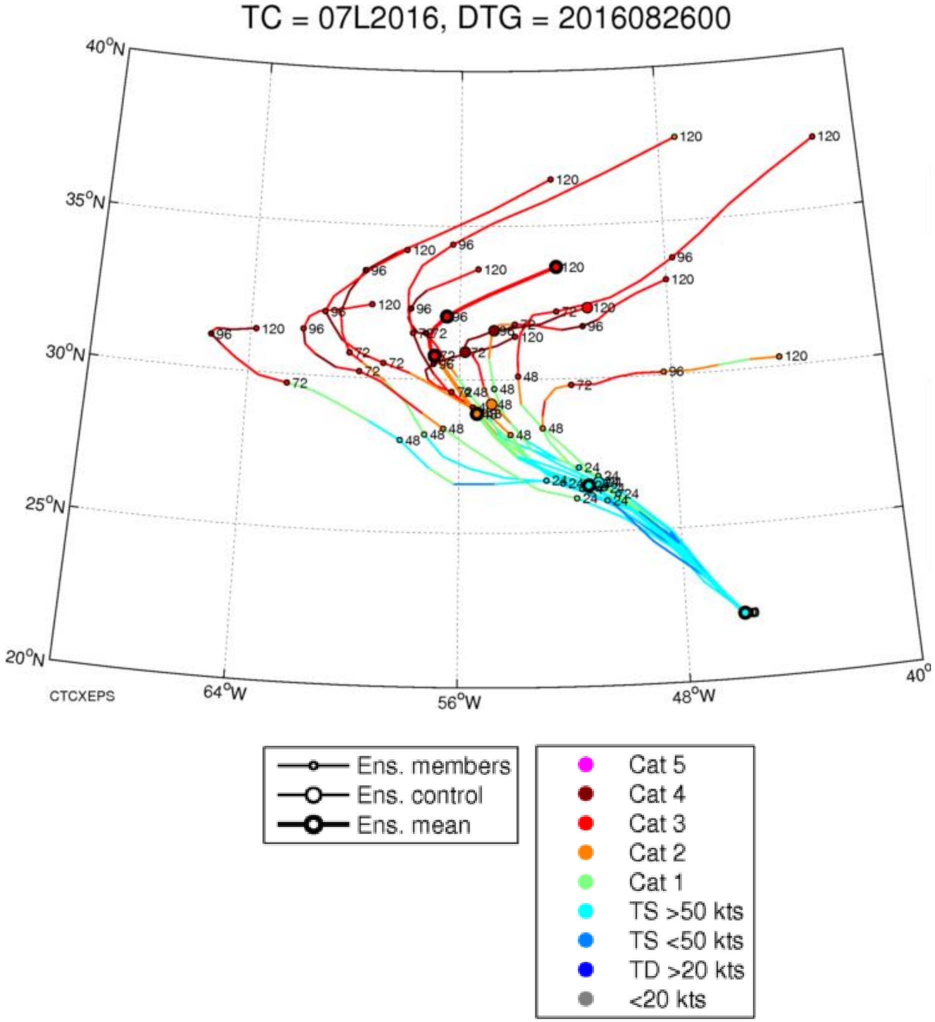
U.S. NAVAL RESEARCH LABORATORY

TC ensemble forecast products

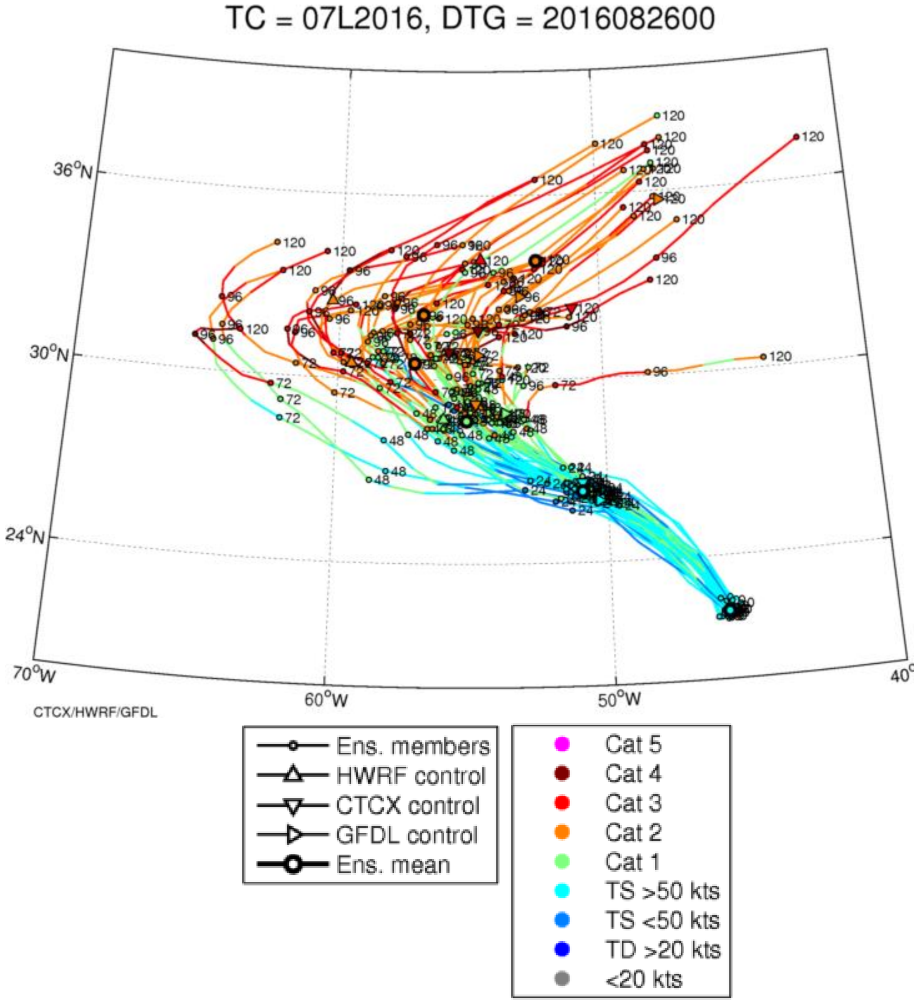
Track colored by forecast intensity

New for 2016

COAMPS-TC



COAMPS-TC / HWRF / GFDL



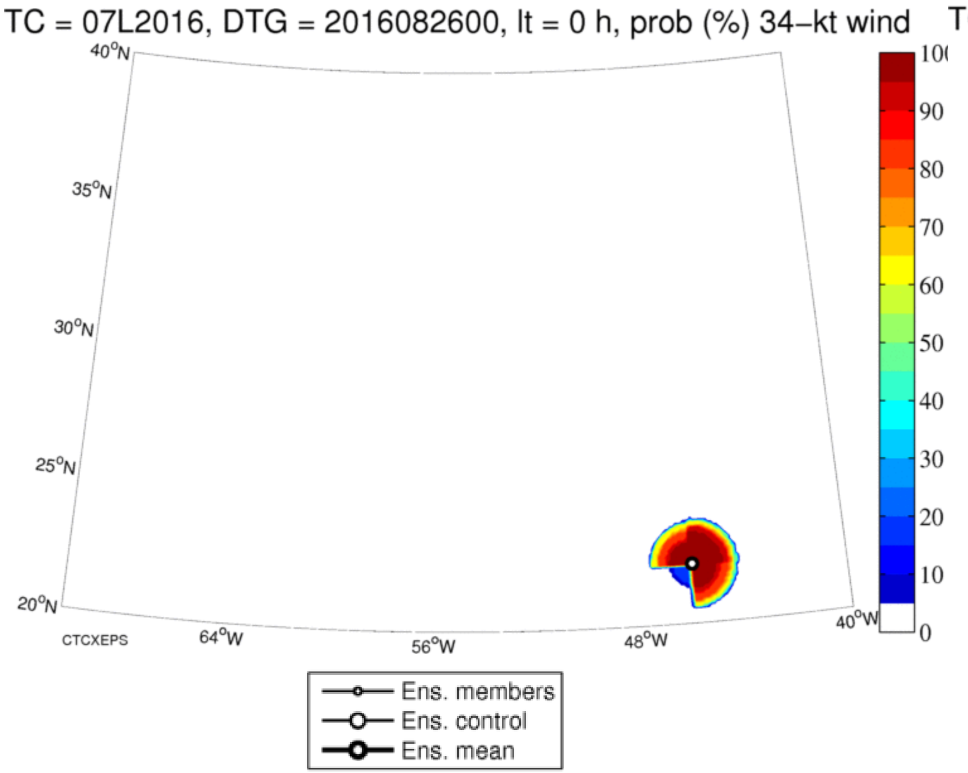
TC ensemble forecast products

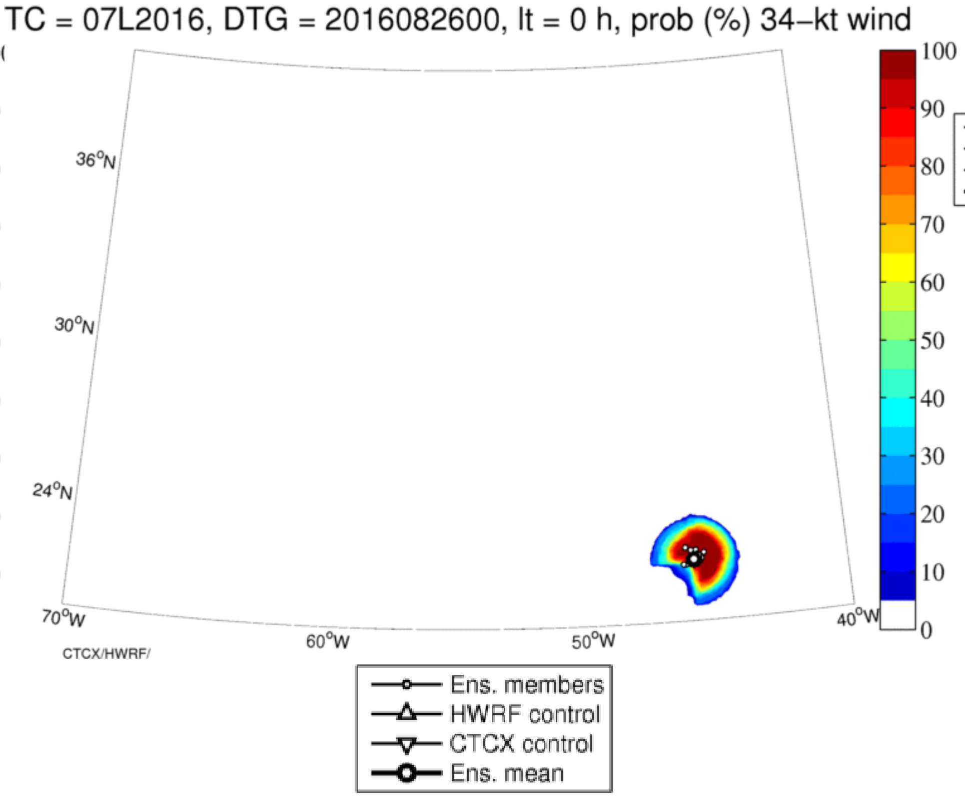
10-m wind threshold exceedance probability

New for 2016

COAMPS-TC

COAMPS-TC / HWRF





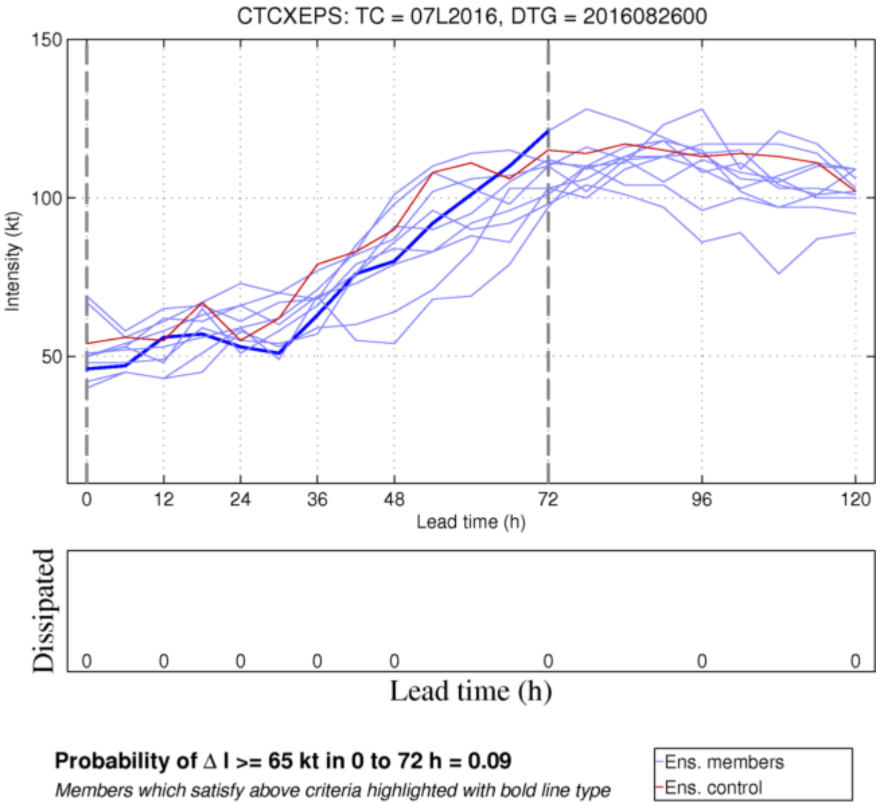
Available for 34 kt, 50 kt, and 64 kt thresholds, with both animations as shown above and static images for tau = 120 h

TC ensemble forecast products

Rapid intensification probability

New for 2016

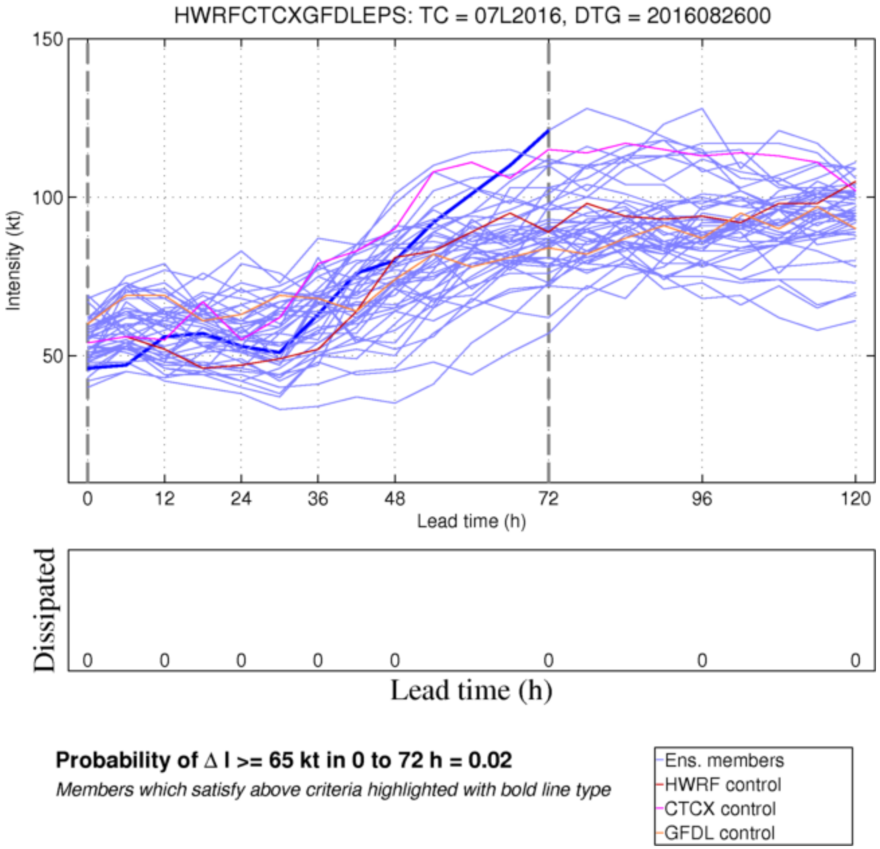
COAMPS-TC



Probability of Δ I >= 65 kt in 0 to 72 h = 0.09

Members which satisfy above criteria highlighted with bold line type

COAMPS-TC / HWRF / GFDL



Probability of Δ I >= 65 kt in 0 to 72 h = 0.02

Members which satisfy above criteria highlighted with bold line type

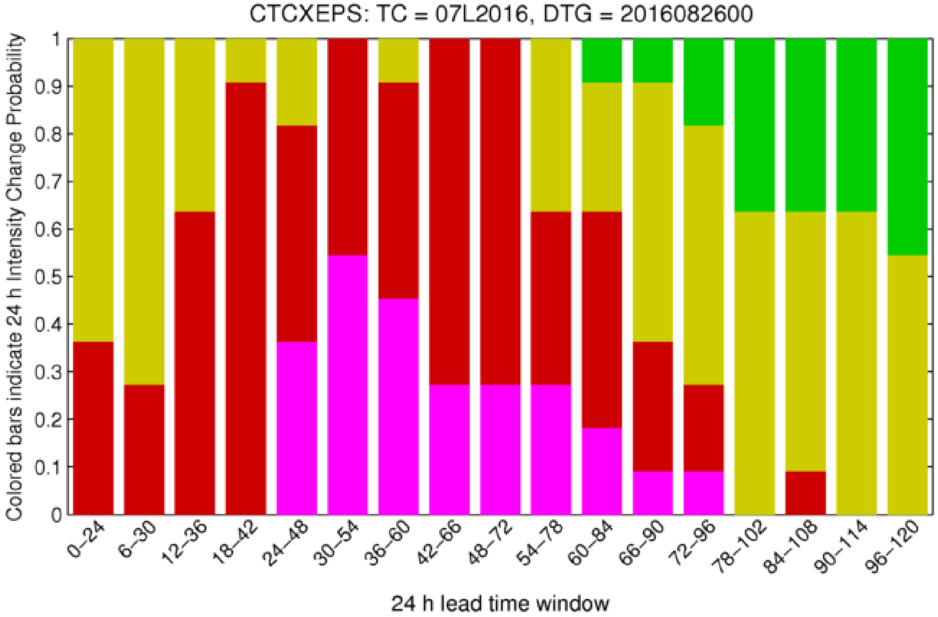
Available for ΔI ≥ 30 in 0 to 24 h, ΔI ≥ 55 in 0 to 48 h, and ΔI ≥ 65 in 0 to 72 h (as shown in example above)

TC ensemble forecast products

24 h intensity change probability

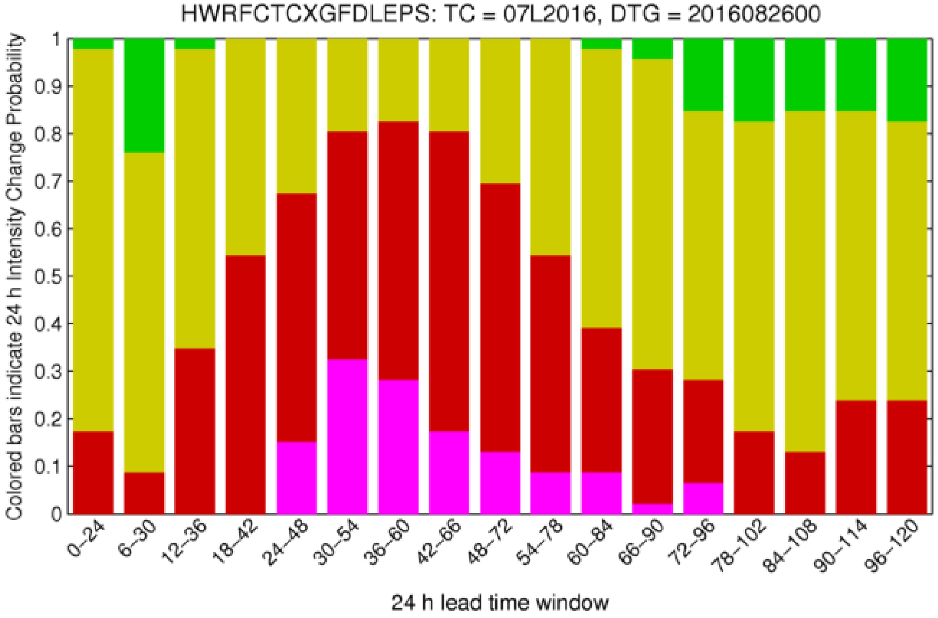
New for 2016

COAMPS-TC



Δ I >= 30 kt (Rapid Intensification)
10 kt <= Δ I < 30 kt (Moderate Intensification)
–10 kt < Δ I < 10 kt (Steady Intensity)
–30 kt < Δ I <= –10 kt (Moderate Weakening)
Δ I <= –30 kt (Rapid Weakening)
TC already dissipated or dissipates during window

COAMPS-TC / HWRF / GFDL



Δ I >= 30 kt (Rapid Intensification)
10 kt <= Δ I < 30 kt (Moderate Intensification)
–10 kt < Δ I < 10 kt (Steady Intensity)
–30 kt < Δ I <= –10 kt (Moderate Weakening)
Δ I <= –30 kt (Rapid Weakening)
TC already dissipated or dissipates during window

TC ensemble forecast products

Future directions

Deterministic prediction

- Under the assumption that the validating observation and ensemble forecast members are drawn from the same distribution, optimal deterministic forecast (for typical metrics like MAE, MSE) is central tendency of the ensemble

- However, if observational information becomes available between the forecast initial time and time the ensemble forecast is completed, it could potentially be used to re-weight the ensemble members to generate an improved deterministic prediction

Augmented deterministic prediction

- The COAMPS-TC ensemble can distinguish between low and high uncertainty cases, for both track and intensity

- The ensemble could be used to support a qualitative forecast uncertainty designation (e.g. high/medium/low) accompanying a deterministic forecast, or a quantitative measure of forecasts uncertainty (e.g. 90% confidence interval)

Probabilistic prediction

- We plan to continue producing and validating probabilistic, ensemble-based forecast products