2011 HFIP R & D Activities Summary: Accomplishments, Lessons Learned, and Challenges

R. Gall[11], F. Toepfer[11], S. Aberson[1], J.W. Bao[4], M. Bender[6], S. Benjamin[4], L. Bernardet[4], M. Biswas[5], B. Brown[7], C. Davis[6], M. DeMaria[8], J. Doyle[10], M. Fiorino[4], J. Franklin[9], I. Ginis[13], J. Goerss[10], S. Gopalakrishnan [1], T. Hamill[4], H.S. Kim[3], T. Krishnamurti [5], P. Kucera[7], Y. Kwon [3], W. Lapenta[3], S. Lord[3], T. Marchok[6], F. Marks[1], E. Rappaport[9], C. Reynolds[10], N. Surgi[3], V.Tallapragada[3], H. Tolman[3], G. Vandenberghe[3], Y. Weng[12] J. Whittaker[4], F. Zhang[12] ,Xuejin Zhang[1], M. Zupanski[2]

26 March 2012

[1]Atlantic Oceanographic and Meteorological Laboratory OAR/NOAA
[2]Colorado State University
[3]Environmental Modeling Center NCEP/NOAA
[4]Earth System Research Laboratory OAR/NOAA
[5]Florida State University
[6]Geophysical Fluid Dynamics Laboratory OAR/NOAA
[7]National Center for Atmospheric Research
[8]National Environmental Satellite Data Information Center NOAA
[9]National Hurricane Center NWS/NOAA
[10]Naval Research Laboratory, Monterey
[11]Office of Science and Technology, NWS/NOAA
[12]Pennsylvania State University
[13]University of Rhode Island

Table of Contents

Executive Summary

This report describes the activities and results of the Hurricane Forecast Improvement Program in 2011. It is organized around three themes, *Accomplishments* from 2011 HFIP testing, evaluation and development activities, *lessons learned* from negative results and the *challenges* faced by the program to achieve its goals. The main topics from each of these three categories are:

Recent Accomplishments

- The GFS model initialized with an EnKF Data assimilation system is showing significant ability to forecast tropical storm genesis out to 5 days or more
- Track forecasting skill with the global GFS/EnKF system at or better than the skill of the best global models in the world with 20% improvement for most forecast lead times
- The impact of high resolution data near the hurricane center using aircraft data (including the tail Doppler radar data) noted last year using the Penn State ARW model has been confirmed using HWRF – 20 to 40% improvement in intensity forecast
- The initialization system for the operational HWRF has been greatly improved
- There have been dramatic improvements in physical parameterizations in HWRF
- A 3 km version of HWRF, to be operational in 2012, is showing significant skill improvement over the current operational 9 km version at 84 hours and beyond
- New advancements in the COAMPS-TC data assimilation, initialization and physical parameterizations resulted in a greatly improved intensity forecast skill in 2011. COAMPS-TC was the only dynamical model to outperform the statistical guidance and exceeded the HFIP intensity goal for forecast times at 48 hours and beyond
- Statistical post processing of model output from two different statistical model approaches give the best intensity forecasts as compared to current dynamical and statistical models
- New model output products are providing more useful information from the models, especially ensembles
- New diagnostic verification tools are providing more useful information for model evaluation.

Lessons Learned

- Initialization of the regional models has improved but it still remains a major problem. It will need to be solved before we can make much progress on forecasting rapid intensification
- Physics packages suitable for models with resolutions of 3 km need to be improved

Challenges

- The biggest challenge to the results from HFIP will be for NOAA to identify the computer resources necessary to run the operational system that will be required to meet its goals
- The program needs a focus on developing appropriate physics packages for high resolution (3km) regional models
- Find ways to better use satellite data at resolutions appropriate for the hurricane

In the body of the text we discuss each of the above and provide an overview of the program and its goals.

1. Introduction

The year, 2005, saw record tropical cyclone and hurricane activity in the Atlantic, including two major storms that crossed the US coastline, Katrina and Rita. There were a total of 27 named storms that year of which 13 were hurricanes and 7 were major hurricanes. Throughout the decade 2000-2010, 13 hurricanes crossed the US coastline including the notable hurricanes of Charlie (2004), Wilma (2005) and Ike (2008), http://www.nhc.noaa.gov/archive/2011/index.shtml.

The heightened awareness of the danger to the US from hurricanes as a result of the record year in 2005 led to a number of studies on NOAA's ability to forecast hurricanes. The NOAA Science Advisory Board's (SAB) Hurricane Intensity Research Working Group (HIRWG) filed a report in October 2006 on recommendations to improve forecasts of hurricane intensity ; http://www.sab.noaa.gov/Reports/HIRWG_final73.pdf. Forecasting skill for hurricane intensity has not improved over the last two decades; http://www.nhc.noaa.gov/verification/verify5.shtml and the report makes recommendations on strategies to improve hurricane intensity forecasts.

On January 11, 2007 NSF National Science Board issued a report on the need for a National Hurricane Research Initiative http://www.nsf.gov/nsb/committees/archive/hurricane/initiative.pdf. Then in November 2007 NOAA released its response to the HIRWG report that included a proposed Hurricane Forecast Improvement Program:

"In response to the HIRWG report, NOAA convened a corporate hurricane summit developing unified strategy to address hurricane forecast improvements. On May 10, the NOAA Executive Council (NEC) established the NOAA Hurricane Forecast Improvement Project (HFIP), a 10-year effort to accelerate improvements in one to five day forecasts for hurricane track, intensity, storm surge and to reduce forecast uncertainty, with an emphasis on rapid intensity change." http://www.sab.noaa.gov/Reports/hirwg/2010/SAB_Nov07_HFIP_Response_to_HIRWG_FINAL.pdf.

The Hurricane Forecast Improvement Project (HFIP) was established within NOAA in June of 2007. In July 2008-July 2009 the President's Budget was amended to include +$13M for HFIP and this increment became part of NOAA's base budget.

The HFIP program, its goals and proposed methods for achieving those goals and recent results from the Program are described below and suggest that it is clearly on a path to the meet its goals on time.

2. The Hurricane Forecast Improvement Project

HFIP provides the basis for NOAA and other agencies to coordinate hurricane research needed to significantly improve guidance for hurricane track, intensity, and storm surge forecasts. It also engages and aligns the inter-agency and larger scientific community efforts towards addressing the challenges posed to improve hurricane forecasts. The goals of the HFIP are to improve the accuracy and reliability of hurricane forecasts; to extend lead time for hurricane forecasts with increased certainty; and to increase confidence in hurricane forecasts. These efforts will require major investments in enhanced observational strategies, improved data assimilation, numerical model systems, and expanded forecast applications based on the high resolution and ensemble-based numerical prediction systems.

The specific goals of the HFIP are:

- Reduce the average errors of hurricane track and intensity forecasts by 20% within five years and 50% in ten years with a forecast period out to 7 days
- Increase the probability of detecting rapid intensification at day 1 to 90% and 60% at day 5

The benefits from HFIP will significantly improve NOAA's forecast services through improved hurricane forecast science and technology. Forecasts of higher accuracy and greater reliability (i.e., user confidence) are expected to lead to improved public response, including savings of life and property.

NOAA recognizes that addressing the broad scope of the research and technology challenges associated with improving hurricane forecasts requires interaction with, and support of, the larger research and academic community. It is hypothesized that these very ambitious goals of the HFIP can only be met using high-resolution (~15 km) global atmospheric forecasting numerical models run as an ensemble in combination with regional models at even higher resolution (~1-5 km). Demonstrating this is very expensive, computationally, and hence HFIP has been building up a computational system in Boulder, Colorado where HFIP can demonstrate the techniques necessary to meet its goals. Only by demonstrating the value of high resolution and other techniques for improving the numerical guidance to hurricane forecasters is there any opportunity to obtain such a computational resource for operational hurricane forecasts.

For FY11, the HFIP program consisted of about $23M with $3M dedicated to enhancing computer capacity available to the Program. The funding for computing was used to enhance the HFIP system established in Boulder, Colorado in FY2009 and resulted in a machine called t-jet with 16,000 processors. About $7M of the $23M is part of the base funding for the National Hurricane Center (NHC) and Atlantic Ocean and Meteorology Laboratory (AOML) in Miami, the Environmental Modeling Center (EMC) at NCEP for hurricane model development and the Earth System Research Laboratory (ESRL). The remaining $13M was distributed to the following 1) various NOAA laboratories and centers: [Earth System Research Lab (ESRL), Geophysical Fluid Dynamics Laboratory (GFDL), National Environment Satellite Data and Information Service (NESDIS), and National Hurricane Center (NHC)] 2) the National Center for Atmospheric Research (NCAR) 3) the Naval Research Laboratory in Monterey (NRL), and 4) several universities: University of Wisconsin, The Pennsylvania State University, Colorado State University, Florida State University, University of Wisconsin and University of Rhode Island (awarded through a NOAA Announcement of Opportunity) and the National Oceanographic Partnership Program (NOPP). Specifically, $1M was contributed each year for three years to the NOPP, Announcement of Opportunity, for competed proposals related to improving understanding and prediction of hurricanes. The funding to NOPP from HFIP was matched by funding from the Office of Naval Research (ONR).

Distribution of the $10M was accomplished through recommendations from 9 teams focused on various components of the hurricane forecast problem. Current teams are listed in Table 1 including the co-leaders of each team and the various organizations represented on the teams. These teams are made up of over 50 members drawn from the hurricane research, development and operational community.

Table1. HFIP Development Teams

FY2010 Teams	Team Leads and Member's Organization
1. Global Model/Physics	*Stan Benjamin* (ESRL), *John Brown* (ESRL), AOML, NRL, GFDL, EMC, NRL
2. Regional Model/Physics	*Morris Bender* (GFDL), *Young Kwon* (EMC), AOML, NRL, ESRL URI, Old Dominion Univ, NCAR
3. Ensembles	*Zoltan Toth* (ESRL), *Carolyn Reynolds* (NRL), AOML, PSU, EMC, NHC, FSU
4. Data Assimilation/Vortex Initialization Team	*Jeff Whitaker* (ESRL), *Bill Lapenta* (EMC), AOML, NRL, CIRA, PSU
5. Verification Team	*Tim Marchok* (GFDL), *Barb Brown* (NCAR), NRL, NESDIS/STAR, AOML, NHC, EMC, ESRL, NWS/OST
6. Applications Development and Diagnostics	*Ed Rappaport* (NHC), *Mark DeMaria* (NESDIS/STAR), EMC, NRL, AOML, NCAR, ESRL, OU, AOML, FSU, NHC, AOML, NWS/OST
7. Hurricane Observations	*Sim Aberson* (AOML), *John Knaff* (NESDIS/STAR), NHC, EMC, NESDIS/STAR, ESRL, URI, NRL, AOC, NCAR, RSMAS, NCO, NCAR, NWS/OST
8. Ocean/Wave Models	*Hendrik Tolman* (EMC), *Halliwell* (AOML), URI, ESRL, NRL, RSMAS
9. Societal Impacts	*Bill Read* (NHC), *Jennifer Sprague*(NWS/OASST), NWS/SR, NWS/ER, FEMA,CT-EM, TX-EM, NC-EM, FL-EM, Weather Channel

HFIP is primarily focused on techniques to improve the numerical model guidance that is provided by NCEP operations to NHC as part of the hurricane forecast process. It is organized along two paths of development called Streams (Table 2). Stream1 assumes that the computing power available for operational hurricane forecast guidance will not exceed what is already planned by NOAA. The development for this stream has been in planning for several years by EMC and is augmented and expanded to involve the broad community by the HFIP including activities at the NOAA labs and centers.

HFIP Stream 2 does not put any restrictions on the availability of computer power available to NWS operations, and in fact, assumes that resources will be found to greatly increase available computer power in operations above that planned for the next 5 years. The purpose of Stream 2, therefore, is to demonstrate that the application of advanced science and technology developed under the auspices of HFIP along with increased computing will lead to the expected increase in accuracy and other aspects of forecast performance. Because the level of computing necessary to perform such a demonstration is large, the Program is developing its own computing system at NOAA/ESRL in Boulder Colorado.

A major part of Stream 2 is an experimental forecast system that is run each hurricane season. The purpose of this system is to evaluate strengths and weaknesses of promising new technology. As a result of the experimental testing, some components may be found to be of particular interest to the operational forecasters, and, if resources do not permit its implementation in the operational infrastructure, the Experimental Forecast System for the following season will emphasize those components and will provide specific output that is made available to NHC forecasters for evaluation. We refer to this component of the Experimental Forecast System as Stream 1.5. The stream 1.5 candidates undergo a level of testing similar to the testing at EMC for stream 1 operational implementations where the candidate models are run over many storms and many seasons. This process is overseen by the Tropical Cyclone Modeling Testbed (TCMT) at NCAR and uses a list of storms that have been selected by NHC: http://www.ral.ucar.edu/projects/hfip/includes/2011_stream_1.5_test_cases.pdf. The results are evaluated by TCMT and final selection for the upcoming hurricane season is made by NHC.

Table 2 outlines these various streams. Roughly half of the HFIP funding is going toward Stream 1 development activities. The remaining portion goes toward Stream 2 development and Stream 1.5 activities.

Table2. The Two Stream Strategy

Stream 1	Development to directly improve the current operational global and regional hurricane models. Assumes that the computing that will be available for operations is that currently being planned.
Stream 2	Assumes that operational computing can be substantially increased above current plans. Uses an HFIP developed computer system to test and evaluate new technology. Emphasis is on high resolution global and regional models run as ensembles. It will include a demonstration system run in real time each summer.
Stream 1.5	Stream 1.5 covers improved models and/or techniques that the NHC, based on prior assessments, wants to access in real-time during a particular hurricane season, but which cannot be made available to NHC by the operational modeling centers in conventional "production" mode. Stream 1.5 will be part of the summer demonstration system and will be forecaster defined.

To facilitate the transition of research to operations, HFIP has recognized the importance of having research and operations share the same code base, and HFIP has co-sponsored the Developmental Testbed Center (DTC) to make available and support HWRF (Hurricane WRF—the operational model at EMC) to the community. Code downloads, extensive documentation, and datasets can be found at http://www.dtcenter.org/HurrWRF/users. This support started in February 2010 with the DTC/EMC/MMM Joint Hurricane Workshop and the WRF for Hurricanes Tutorial. During FY10 and FY11, EMC and DTC established a common version of the operational codes for HWRF in the repository at DTC. As DTC ported the code to various platforms, conducted testing and evaluation, and established Reference Configurations (benchmarks of the code that can be used to assess future development or to point directions for model improvement), several model issues were uncovered and addressed, increasing the quality of the model. This unified version has been thoroughly tested and shown to give the same results as the operational codes. Thus with the repository in Boulder at the DTC, and rigorous code management protocols implemented by DTC and EMC, both the operations and research community are now drawing from the same central code repository facilitating the transitions of research results to operations.

3. The HFIP Model Systems

In HFIP we are assuming that the best approach to improving the forecast hurricane track beyond 4 days is through the use of high resolution global models run as an ensemble. We describe below the logic behind this assumption. For improvements in forecast of hurricane intensity, especially in the 1 to 5 day time range, the best approach is likely high resolution regional models, also run as an ensemble. The global model ensembles are likely to be limited in resolution to about 15-25 km for at least the next 5 years, because of computer limitations. Thus the only way to achieve the very high resolution of about 1-3 km necessary for resolving the inner core of the hurricane is with regional models. It is generally assumed that the inner core must be resolved before we can expect to see consistently accurate hurricane intensity forecasts (HIRWG Report).

3.1. High resolution ensemble approach

A single "deterministic" forecast by a particular numerical model has an inherent but unknown level of uncertainty; any two model forecasts starting from infinitesimally different initial states will grow differently with time, the amount of difference depending upon the weather situation. If the forecast is reproduced many times, each time introducing small initial differences, the result is called an ensemble, and the different model forecasts can potentially provide information on the confidence one should place in a particular forecast. In fact, a single deterministic model is inherently a single member of some ensemble and so would be part of the same uncertainty apparent in ensemble predictions. Note that ensembles can be created in several ways including changing initial conditions, using several models and altering some component of the model in some way such as the physics packages or output from the model's physics package. Frequently, but not always, the highest probability is that the correct forecast is near the mean, median or mode of the ensemble, though other ensemble realizations have a finite probability of being correct (Buizza 1997). Because the various forecasts diverge with time, emergency managers should be able to make more effective decisions when provided with ensemble guidance compared to being provided with a single forecast. Ensemble predictions can also be used by forecasters to analyze reasons for uncertainty and given more recent information available after ensemble initialization, make choices on how the forecast is going to go relative to the ensemble. For example if the uncertainty in hurricane track is being forced by movement (slow or fast) of a nearby mid latitude trough, more recent information on the speed of the trough than that available to the ensemble can help to discriminate between the various ensemble members as the more likely. High resolution is hypothesized to be necessary in these ensembles in order to adequately resolve the hurricane structure **(**HIRWG Report**)**, for the hurricane can alter the flow in which it is embedded and, in turn, this altered flow will impact the hurricane track and so also its intensity. To even begin to get structures in a forecast model which resemble actual hurricanes, resolutions of 15-25 km are likely necessary. Ideally, 20-30 members are computed to provide adequate estimates of the uncertainty with each ensemble member having a resolution of 15-25 km.

Beyond about three days, forecast guidance must come from global ensembles since the planetary-scale patterns interact with and influence the steering of the storm. After about three days, it has been shown (Reynolds et al. 2009, Hakim et al 2003, Langland et al. 2002, Palmer et al. 1998, Rabier et. al. 1996, Hoskins and Ambrizzi 1933, Chang, 1993) that the evolution of the atmospheric flow at a given location depends on atmospheric features distributed globally. Therefore, forecasts that extend out to 4-7 days require that the forecast models be global.

The potential value of high-resolution global ensembles has been demonstrated in part through forecasts from the international community such as the European Centre for Medium-Range Weather Forecasts (ECMWF**) (**Buizza et al. 2005). However, there is still much to be learned about high-resolution global modeling. The best way for the U.S. to make progress is to run the ensembles over enough cases such that statistical significance of the computed skill of the forecasts can be determined. Generally for hurricane forecasts this requires at least that the high-resolution ensemble be run over the most active few months of the hurricane season and every forecast period from genesis to decay (with 2 to 3 years of cases being even better at capturing the full range of tropical cyclone characteristics associated with inter-annual changes in environment, e.g., associated with El Nino events). This is an enormous computing challenge but it needs to be performed to demonstrate the value of the high-resolution forecast guidance over the guidance that is operationally available today.

Much the same can be said for regional ensembles, but here the emphasis shifts from track forecasts at longer forecast lead times, to intensity forecasts at medium forecast lead times. Much of the control of the intensity of the storm is thought to reside in the dynamics of the inner core region of the hurricane. If this is true, then the inner core must be resolved to account for these dynamics requiring a resolution of at least 1-5 km. The regional high-resolution ensembles are nested within high-resolution global ensembles, which provide the lateral boundary conditions and at least a first guess of initial conditions to the regional model.

3.2. The Global Models

The descriptions below are for the various individual global model components of the HFIP experimental system. Some, as noted later, are run as an ensemble.

The *Flow-following finite-volume Icosahedral Model* (FIM) is an experimental global model that can be run at various resolutions and uses initial conditions from a number of sources. It is currently using a fixed ocean underneath. It has been built by the NOAA Earth System Research Laboratory (ESRL) http://fim.noaa.gov/fimdocu_rb.pdf (Benjamin et. al. 2004).

There are two versions of the *Global Forecast System* (GFS) model currently running in the demonstration system. This includes a version of the current operational model run at the NOAA National Centers for Environmental Prediction (NCEP) and an experimental version of that model at ESRL. The main difference between the two versions is the initialization system: GSI for the operational model and EnKF for the ESRL version http://www.emc.ncep.noaa.gov/GFS/doc.php.

Currently a semi-Lagrangian version of the *Navy Operational Global Atmospheric Prediction System* (NOGAPS) is being developed, which will allow for efficient high-resolution forecasts http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA247216&Location=U2&doc=GetTRDoc.pdf.

Some specifics of the global models are shown in Table 3. Note that GFS/GSI and NOGAPS are operational global models.

Table 3. Specifications of the HFIP Global Models

Models	Horizontal resolution	Vertical levels	Cumulus Parameterization	Microphysics	PBL	Land Surface	Radiation	Initialization
FIM	27 km	64	From 2010GFS - Simplified Arakawa Schubert	Ferrier	GFS Non-Local PBL	Noah LSM	GFDL/RRTM	ESRL EnKF
GFS/EnKF	27 km	64	Simplified Arakawa Schubert	Ferrier	GFS Non-Local PBL	Noah LSM	GFDL scheme	ESRL EnKF
GFS/GSI	27 km	64	Simplified Arakawa Schubert	Ferrier	GFS Non-Local PBL	Noah LSM	GFDL	GSI
NOGAPS	41 km	42	Emanuel	N/A	NOGAPS	NOGAPS	Harshvardhan/ Fu-Liou	NAVDAS-AR

3.3. The Regional Models

The *Weather Research and Forecasting* (WRF) model is actually a modeling system that is comprised of many configurable components. The options for the dynamic core include the Advanced Research WRF (ARW) built by NCAR and the Non-hydrostatic Mesoscale Model (NMM) built by EMC. The ARW system includes an option for a simplified one dimensional model of the ocean and two interactive nests within the outer regional model. There are also several options for physics as well as initialization, post processing and verification systems http://www.dtcenter.org/HurrWRF/users/docs/index.php.

The *NCEP Hurricane WRF* (HWRF) is based on the Non-hydrostatic Mesoscale Model (NMM) dynamic core and has a movable, two-way nested grid capability for the 9 km inner nest. The coarse domain is 27 km resolution and covers a 75° x 75° region with 42 vertical layers. Advanced physics include atmosphere/ocean fluxes, coupling with the Princeton Ocean Model and the NCEP GFS boundary layer and deep convection. An experimental version of HWRF being developed by EMC and AOML has a third inner nest with a resolution of 3 km.

The *Penn State Regional Ensemble* is another configuration of the ARW system. It uses a static interactive inner nest but no interactive ocean http://hfip.psu.edu/realtime/AL2011/forecast_track.html (Zhang et. al. 2011).

The *Coupled Ocean/Atmosphere Mesoscale Prediction System* (COAMPS-TC) is the Navy next-generation tropical cyclone model developed and run by NRL Monterey (Doyle et al. 2012). It is a version of their COAMPS regional prediction system that is being run operationally and has an interactive ocean http://www.nrlmry.navy.mil/coamps-web/web/tc.

The *Wisconsin Model* was developed by Greg Tripoli of the University of Wisconsin and is being run as part of the HFIP Experimental Forecast System http://cup.aos.wisc.edu/will/HFIP/config.html.

The *GFDL Model* development is being led by Tim Marchok. It was run as part of the HFIP Experimental Forecast System as a regional ensemble prediction system and the basic model is the same as the GFDL operational system. See below for description of the GFDL Ensemble.

Specifics of the regional models are shown in Table 4. Note that GFDL (OPS) and HWRF (OPS) are the operational regional models.

Table 4: Specifications for the HFIP Regional models

Models	Nesting / Horizontal Resolution (km)	Vertical Levels core	Cumulus Parameterization	Microphysics	PBL	Land Surface	Radiation	Initial and Boundary Conditions	Initialization
HWRF (OPS)	**2 27/9**	**42 NMM**	**Simplified Arakawa Schubert**	**Ferrier**	**GFS Non-Local PBL**	**GFDL Slab Model**	**GFDL Scheme**	**GFS**	**Bogus vortex plus GSI 3DVAR**
GFDL (OPS)	**3 30/15/7.5**	**42 GFDL**	**Arakawa Schubert**	**Ferrier**	**GFS Non-Local PBL**	**Slab Model**	**Schwarz-kopf-Fels Scheme**	**GFS**	**GFDL Synthetic Bogus Vortex**
HWRF-HRD/EMC 27-9-3	**3 27/9/3**	**42 NMM**	**Simplified Arakawa Schubert**	**Ferrier**	**GFS Non-Local PBL**	**GFDL Slab Model**	**GFDL Scheme**	**GFS**	**Bogus vortex plus GSI 3DVAR**
HWRF-HRD 27-9	**9 27/9**	**42 NMM**	**Simplified Arakawa Schubert**	**Ferrier**	**GFS Non-Local PBL**	**GFDL Slab Model**	**GFDL Scheme**	**GFS**	**EnKF with aircraft data**
NCAR ARW	**3 12/4/1.3**	**36 ARW**	**New Kain Fritsch (12 km only)**	**WSM5**	**YSU**	**5-Layer Thermal Diffusion soil Model**	**RRTM (longwave) / Dudhia (shortwave)**	**GFS**	**EnKF method in a 6-hour cycling mode**
COAMPS-TC	**3 45/15/5 (15/5 km following the storm)**	**40 COAMPS**	**Kain Fritsch on 45 and 15 km meshes**	**Explicit microphysics (5 class bulk scheme)**	**Navy 1.5 Order Closure**	**Slab with the NOAH LSM as an option**	**Fu-Liou**	**GFS**	**3D-Var cycling data assimilation with synthetic observations**
Wisconsin Model	**UW NMS (3D enstrophy/ entropy/KE conserving dynamics core)**	**3-4 90/45/9 km 90/45/9/3 km**	**52**	**Modified Emanuel**	**Explicit bulk microphysics (cloud/rain/pristine/ aggregate/graupel)**	**1.5 Order Closure**	**WRF vegetation/land surface/Andreas emulsion layer**	**RRTM**	**GFS/GFDL**
Penn State ARW	**3 40.5/13.5/4.5 for ensemble forecast 1.5-km nest for control**	**35 ARW**	**Grell-Devenyi ensemble scheme (40.5 km only)**	**WSM 6-class graupel scheme**	**YSU**	**5-layer thermal diffusion scheme**	**RRTM (longwave) / Dudhia (shortwave)**	**GFS**	**EnKF with NOAA airborne radar**

3.4. Evaluation of the Stream 1.5 Candidates

There are a variety of global, regional, and statistical models being developed as part of the HFIP effort. For the 2011 retrospective evaluation, there were eight modeling teams that submitted a variety of models and configurations. The evaluation was conducted using storms from the Atlantic and Eastern Pacific basins sampled from the 2008-2010 hurricane seasons. TCMT evaluated the model configurations and provided results and summary reports to NHC. A consistent methodology was applied for each evaluation. A flowchart for the evaluation is shown in Figure 1. In summary, if the candidate was a late model, an interpolator was applied to the forecasts. All the forecasts and operational baselines (Table 5) were run through the NHC Verification software to compute the errors. Afterwards, the forecasts errors were matched to provide a homogeneous sample. Pairwise differences were generated and error distribution properties were computed. Results were displayed in a variety of graphical summaries and statistically significant tables. The NHC decisions based on these analyses are provided in Table A.1. in the Appendix.

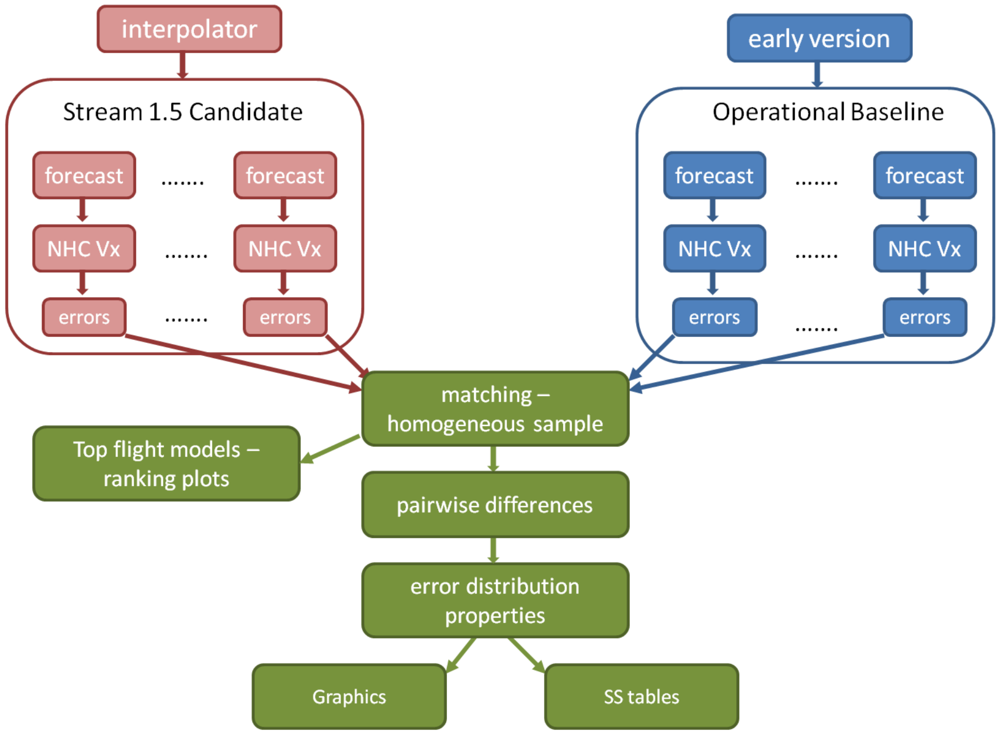


Figure 1: Flowchart showing the steps used in the retrospective evaluation of Stream 1.5 candidates.

Table 5: Baseline comparisons for the Stream 1.5 evaluation

Operational Baselines	Stream 1.5 configuration
Individual: Track – GFSI, GHMI, HWFI (HWRF only) Intensity – LGEM, GHMI, HWFI (HWRF only)	Stream 1.5
Variable Consensus - Track AL: GFSI, EGRI, GHMI, HWFI, GFNI EP: GFSI, EGRI, GHMI, HWFI, GFNI,NGPI Intensity AL & EP: DSHP, LGEM, GHMI, HWFI	ARW, UM-NMS, COAMPS-TC, FIM: Consensus + Stream 1.5 HWRF, GFDL, SPICE: Consensus w/ Stream 1.5 equivalent replacement
Average error of previous year's top flight models Track: GFSI, EGRI, GHMI Intensity: GHMI, DSHP, LGEM	Stream 1.5

3.5. Initialization systems

The initial state for the regional models was generally produced by removing the vortex from the initial state and then inserting a new vortex often known as a bogus vortex. In addition, downscaling directly from a global model is also used for some models. Other models such as the Penn State model and the NCAR ARW model use an EnKF initialization system.

The operational HWRF utilizes an advanced vortex initialization and assimilation cycle consisting of four major steps: 1) interpolate the global analysis fields from the Global Forecast System (GFS) onto the operational HWRF model grid 2) remove the GFS vortex from the global analysis 3) add the HWRF vortex modified from the previous cycle's 6-hour forecast (or use a synthetic bogus vortex for cold start) and 4) add satellite radiance and other observation data in the hurricane area (9 km inner domain). The major differences from the operational GFDL model initialization are steps 3) and 4).

A number of approaches were used to create the initial state for the global and regional models described above. The choices included:

1) The current operational model, Global Forecast System (GFS), interpolated to the higher resolution grid. The GFS uses the Grid point Statistical Interpolation (GSI) data assimilation system that has run operationally for many years. It is a three-dimensional variational approach (3D-VAR) http://www.dtcenter.org/com-GSI/users/docs/index.php , (Parrish et. al. 2003, Parrish et.al. 2003, Wu et.al. 2002, Parrish, et al. 1992, Cohn and Parrish 1991).
2) The NRL Atmospheric Variational Data Assimilation System (NAVDAS) is used to provide the initial conditions to the Navy global model. It has been a 3D-VAR system but starting late September 2009, it was upgraded to NAVDAS-AR (for accelerated representor), a four-dimensional variational approach (4D-VAR).
3) The Navy regional model, COAMPS-TC, makes use of synthetic observations based on the forecaster's best estimate of the tropical cyclone intensity and structure. The vortex is relocated in the background fields and then the synthetic observations are blended with the other conventional observations (dropsonde, aircraft, radiosonde, satellite) using the NAVDAS 3D-VAR system. One of the unique aspects of COAMPS-TC is that cycling data assimilation is used on all the meshes. http://www.nrlmry.navy.mil/docs/_NAVDAS01.pdf **(**Daley and Barker, 2001).
4) The Ensemble Kalman Filter (EnKF) is also an advanced assimilation approach (somewhat like 4D-VAR) that uses an ensemble to create background error statistics for a Kalman Filter (Tippett, et al , 2003, , Keppenne, 2000, Evensen et al., 1998, Houtekamer et. al. 1998). While this approach is still in the experimental stage in the U.S. it has shown considerable promise (Hamill et. al., 2011).
5) The Hybrid Variational-Ensemble Data Assimilation System (HVEDAS) combines aspects of the EnKF and 3D- or 4D-VAR using the ensemble of forecasts to estimate the covariances at the start of a 4D-VAR assimilation window. This technology is under development at NOAA/NCEP/EMC, NOAA/OAR/ESRL and NOAA/AOML/HRD. It is expected to be ready for testing in the 2012 season. This hybrid approach is likely to define the operational global data assimilation system for NOAA in the 5-year time-frame.

3.6. The HWRF community code repository

During 2009-2011 both EMC and DTC worked to update the operational version of HWRF from version 2.0 to the current version of WRF (version 3.3). This makes every component of the HWRF operational model connected with the community codes, allowing researchers access to the operational codes and making improvements in HWRF developed by the research community easily transferable into operations. This was one of the initial goals of the WRF program and has now been fully implemented for the first time with the hurricane models. The process of transforming HWRF onto a community code, which involves the porting to various platforms, creation of several levels of testing including regression tests and consistency checks, and the use of the model by a much larger group, has resulted in a higher quality and more robust code.

Figure 2 shows schematically how the process of moving the operational model into the WRF repository was achieved. The version 3.3 codes were carefully compared to the original operations code until the two versions gave identical answers. Subsequent changes to the operational code will now be entered directly into the unified version.

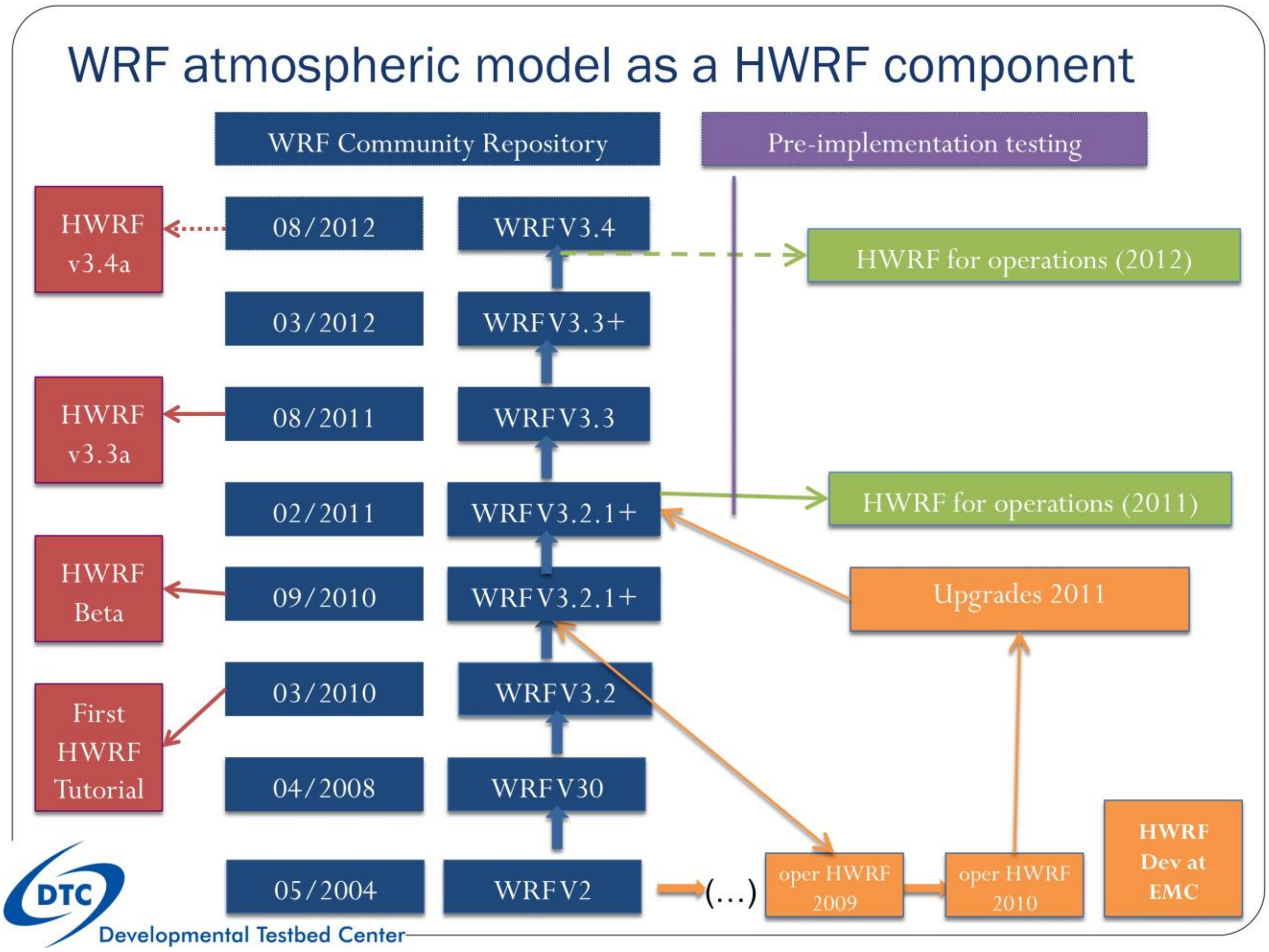


Figure 2 Schematic of the process for incorporating operational capabilities of the HWRF atmospheric component onto the community code (Developmental Testbed Center, personal communication 2011). Note that new capabilities for the 2012 operational model are being added directly onto the community code, so that merges between code developed in separate repositories is no longer need. This process reduces labor costs and human errors.

4. Meeting the HFIP goals

4.1. The HFIP Baseline

HFIP established a baseline, to measure progress toward meeting these goals, against which results from experimental and operational HFIP model guidance will be measured. These HFIP Performance Goals Baselines were developed by James Franklin and others at the National Hurricane Center (NHC) and are provided in Table 6 (track performance) and Table 7 (intensity performance). The CONS is a consensus of operational models and is the primary baseline for measuring improvements. OCD5 is a simple statistical model baseline for measuring improvements in forecast skill, and OFCL is the average NHC official forecast errors from the baseline period.

Table 6. HFIP Track Performance Baseline (nautical miles)

VT (h)	N	OFCL	OCD5	CONS
0	818	7.4	7.7	7.8
12	741	29.4	44.5	30.0
24	663	49.6	93.3	49.8
36	586	69.9	150.9	69.5
48	518	91.2	212.2	89.6
72	411	135.0	317.2	132.0
96	313	173.0	396.5	175.2
120	247	218.6	473.0	221.9

Table 7. HFIP Intensity Performance Baseline (knots)

VT (h)	N	OFCL	OCD5	CONS
0	820	1.9	2.2	2.2
12	745	7.2	8.3	7.7
24	667	10.4	11.5	10.1
36	590	12.6	14.2	11.7
48	522	14.6	16.1	13.7
72	415	17.0	17.8	16.0
96	316	17.5	19.3	16.6
120	250	19.0	19.3	17.0

4.2. Meeting the Track goals

Earlier we noted that forecasts beyond a few days become a global problem meaning the influence on a forecast for a particular location is coming from systems far removed from the local forecast region. This is of course an important consideration for hurricane track prediction out to 7 days, a goal of HFIP. We will note below that global models generally outperform the regional models on track at all lead-times. Thus the HFIP strategy for meeting the track goals is to improve the global models and run them as an ensemble at as high a resolution as possible. The regional models (which can resolve inner core processes) will focus on intensity forecasts. Generally, especially when run as an ensemble, the global models will not have sufficient resolution to resolve inner core structures and consequently will have a low bias in intensity anyway. Though this low bias can be removed from the forecasts, the global models still won't resolve the inner core and so can only indicate trends in intensity due to interactions with the larger scales in the atmosphere and ocean.

Perhaps it is not surprising that global models are more skillful with track than the regional models since track is largely a result of large scale processes which the global models can resolve and which the regional models must take from the global models. Improving the global models is expected to lead to improved intensity forecasts in the regional models since intensity often depends on the track of the storm (Emanuel et al. 2004).

The influence track has on intensity is especially obvious when comparing the storm tracks interaction with or without land**.** However, even when the tracks remain well away from land this can be true. An example from one of the HFIP regional models (Penn State ARW in Table 4) (Zhang et al. 2011, Weng and Zhang 2011, Snyder and Zhang 2003**)** is shown in Figure 3. At the midpoint of the forecast period the various tracks from a 60 member ensemble are divided into three categories (the third on the left of the envelope denoted orange), the third in the middle (green) and the third on the right (blue) based on the forecast storm positions at 18Z 7 September. This same color coding is used to identify the various intensity plots from the ensemble members displayed in a spaghetti plot. Note that the simulated storms on the left (south) side of the track are consistently stronger than those on the right. The observed track was toward the south side of the envelope and indeed the observed winds were consistent with the higher intensity for tracks in the southern third.

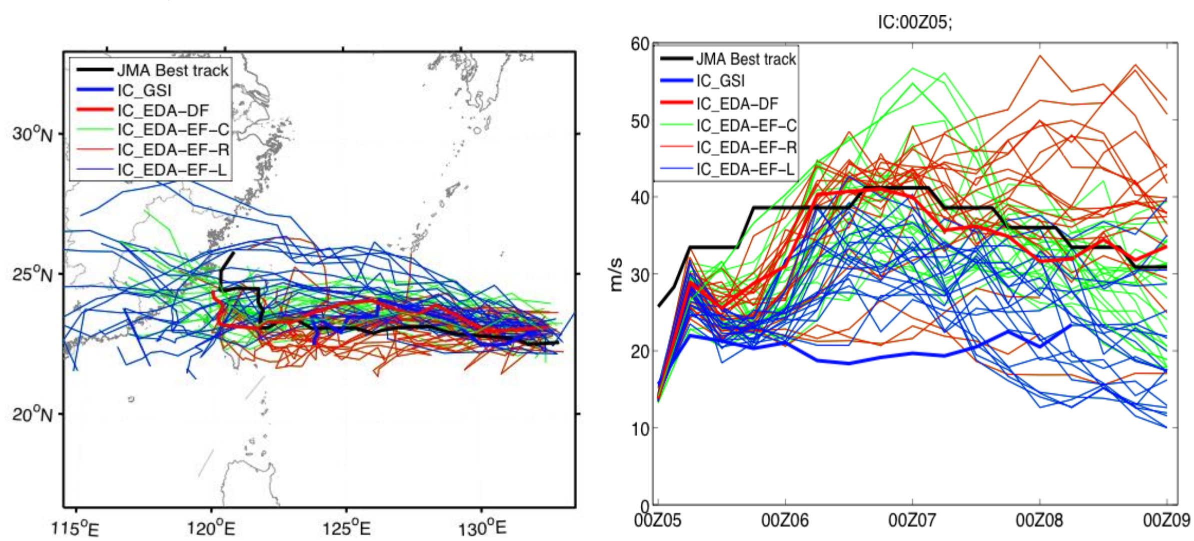


FIG. 3. Track (left panel) and intensity (right panel) forecasts from a 60 member Penn State ARW regional ensemble for Typhoon Morokot from 00Z 5 through 00Z 9 September 2010 in the Western Pacific. The black curve is the Best Track (analysis of Japan Meteorology Administration based on observations). The heavy orange line is a single deterministic run initialized with GFS-EnKF analysis at 00Z 05 September 2010. The orange curves signify the 1/3 ensemble members of the storm (20 members) that were located on the left side of the Best Track at 18Z 07 September 2010, green curves show other 1/3 members near the Best Track and blue curves denote the other 1/3 members located at the right of the Best Track at 18Z 07 September 2010 (F. Zhang 2010, personal communication).

One of the first efforts in HFIP was to improve the global models over what is currently being run operationally by NCEP since the global models generally provide better track forecasts. It was shown early in the program that using a more advanced data assimilation scheme than the one currently in use at NCEP (the operational Grid Statistical Interpolation (GSI) method) in just about any global model, including the current Global Forecast System (GFS), improved forecasts particularly in the tropics. Figure 4 compares forecasts of tropical winds at 250 mb in the GFS and FIM (see Table 3) using GSI with GFS and FIM using an Ensemble Kalman Filter (EnKF) approach.

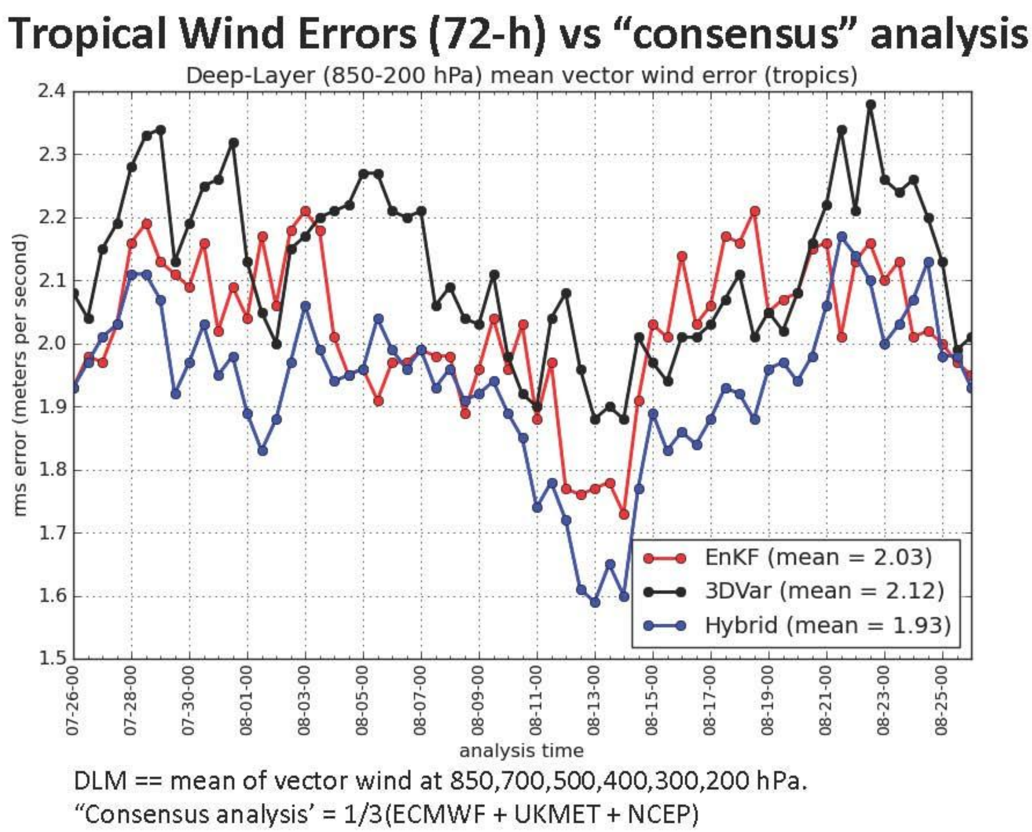


Figure 4. Verification statistics for forecasts run from initial conditions specified by the GSI operational data assimilation system, the EnKF experimental data assimilation system and the experimental hybrid data assimilation system. The statistics are for a 72 h forecast for the deep layer mean wind (850, 700, 500,400, 300, 200 hPa) in the tropics for the Global Forecast System (GFS) operational model run at 30 km using the various assimilation systems for initialization. The legend on the lower right of the figure indicates the initialization system used for each curve. The number in parentheses in the legend shows the average the RMS vector wind error for each configuration. Date is indicated along the horizontal axis using yymmdd at 0000 UTC. (Jeff Whitaker and Daryl Kleist, personal communication).

Note the GFS performs much better using the EnKF data assimilation approach as compared to GSI. All HFIP global models now use the EnKF system and most of the regional models will eventually switch over. Based on comparisons such as that in Figure 4, NCEP has embarked on a planned replacement of the current GSI data assimilation system with a hybrid system**.** This system combines the EnKF approach with the GSI approach and results for that system are shown in blue. The combined (hybrid system) clearly performs even better than the EnKF system alone.

We noted above that ensembles on average produce more skillful forecasts than any individual component of the ensemble run at the same resolution. Since one can consider a single deterministic run as just one component of some virtual ensemble, that deterministic run could be any member of that ensemble and so likely to have a higher average error than if the virtual ensemble had actually been run. Of course, single deterministic runs can outperform any ensemble if they use a superior model.

Figure 5 shows the track errors for various regional and global models including the operational model (GFS GSI—AVNO) and the ECMWF high resolution deterministic model for the 2010 hurricane season. Those with shades of purple are the regional models and the others are the global models. The official forecast errors are shown in orange. The green bar is the HFIP global ensemble using the GFS global model (the same as the operational NCEP operational model) but using the EnKF system to initialize. Note that the ensemble was run at low resolution, T254, as compared to the operational, GFS T574, and the ECMWF T1299. The error statistics are shown relative to the HFIP baseline (Table 6) which is why the official forecasts are close to the zero line (the baseline was defined from the official model guidance in available in 2009). Anything below is worse than the baseline and anything above is better.

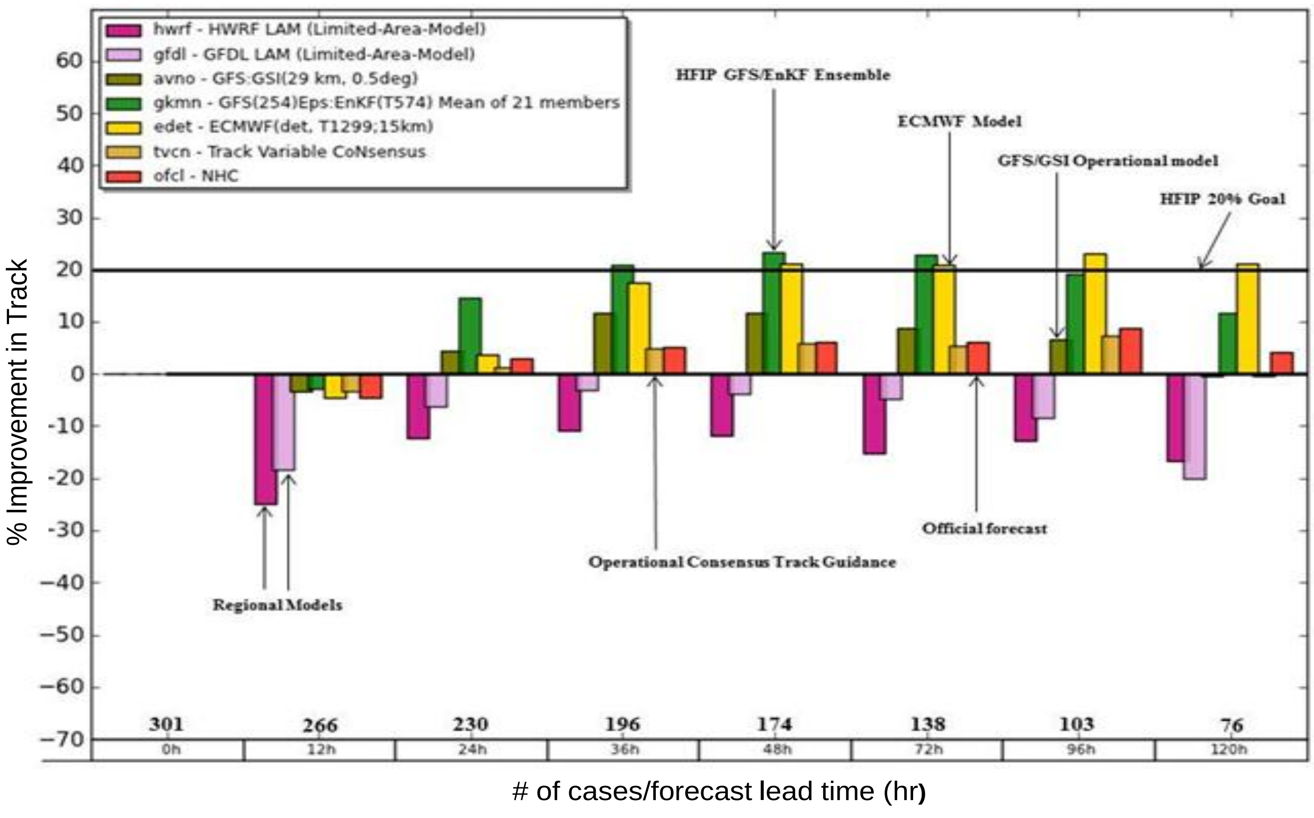


Figure 5 Average track errors for various models in the Atlantic. All storms in 2010-2011 are included; a total of 36 storms. The number of cases at each forecast lead time is shown across the bottom of the figure. HFIP 20% goal refers to the 5 year goal of reducing track errors by 20% relative to the HFIP baseline defined earlier(Mike Fiorino personal communication).

Note that even though the GFS/EnKF ensemble is the lowest resolution of any of the models shown, it still out performed all others in the figure including the very high resolution ECMWF for hurricane track. In fact the GFS ensemble reached the HFIP 5 year goal of a 20% improvement over the baseline for most forecast lead times. Also the track forecasts from the regional models shown here are below the baseline which is typical of all the regional models.

The superior performance of the GFS with EnKF ensemble has to be due to the data assimilation system since the model is exactly the same as the operational GFS/GSI. The ensemble is run at lower resolution than the GFS/GSI (T574) so resolution isn't the reason.

4.3. Reaching the Intensity Goals

HFIP expects that its intensity goals will be achieved through the use of high resolution regional models that can have resolution near the core of at least 3 km. This will allow those inner core processes that control intensity to be resolved. In addition, early results suggest that results from individual HFIP models can be used in statistical models to further increase the skill of the intensity forecasts.

In 2011, the suite of regional models that went into the HFIP regional ensemble are listed in Table 4 and all were part of Stream 1.5. In addition GFDL, NCAR ARW, COAMPS-TC and the Penn State ARW were run as individual ensembles. The GFDL ensemble was constructed from the operational GFDL model using NHC recommendations for varying the initial conditions and convective parameterization. The GFDL ensemble also served as a way to test sensitivities of a regional model to various storm parameters provided by NHC to describe the storm. Figure 6 lists the various members of the GFDL ensemble. The other two models vary initial conditions to construct the ensemble.

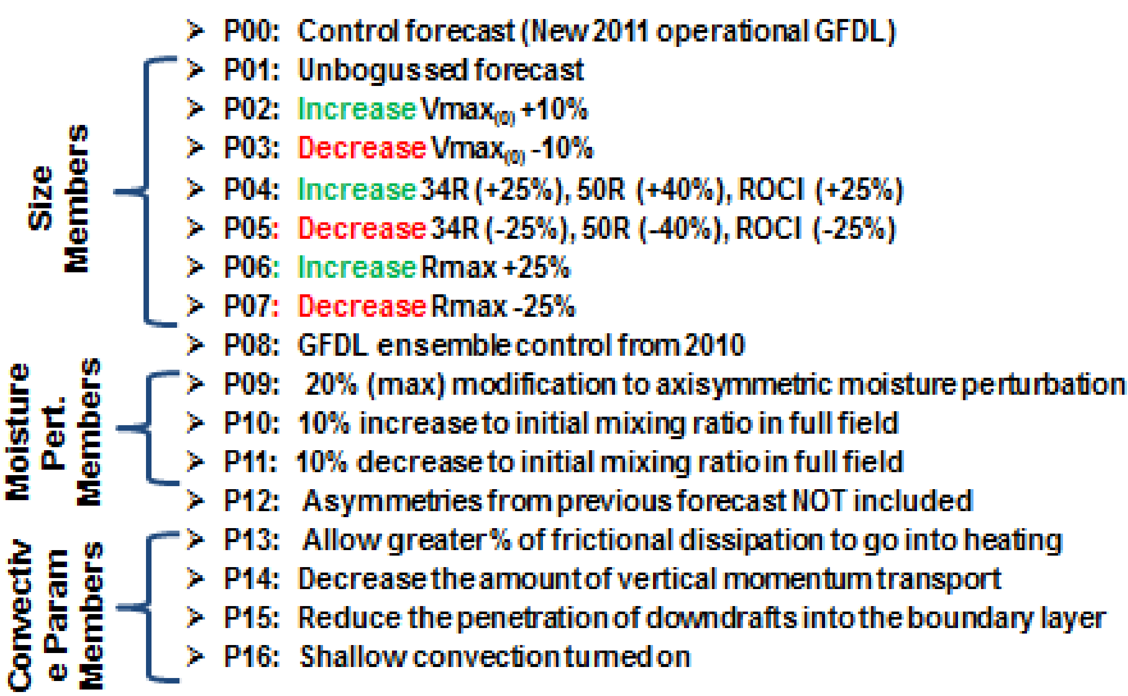


Figure 6. Seventeen member GFDL ensemble: 16 perturbed members and a control forecast. ROCI stands for radius of outermost closed isobar. An unbogussed forecast means that the initial condition used the raw vortex from the GFS global operational model and not the GFDL model initialization process.

In Figure 7, we show results from the 2011 hurricane season for the various models that are operational or are part of the HFIP stream 1.5. It includes all cases through the middle of September 2011 and the comparison is homogeneous. The two black lines in the diagram are for the official forecast (upper line) and for the HFIP baseline (lower line) that was described above. Note that all the lines above the official forecast are for the current operational models running at 7-9 km resolution (GFDL and HWRF including the GFDL ensemble which is a variation of the operational GFDL). The lines below the official forecast include the operational statistical models (DSHP = Decay Ships--Statistical Hurricane Intensity Prediction System, LGEM = Logistic Growth Equation Model) and the experimental stream 1.5 consensus statistical model, SPC3, that uses both DSHIP and LGEM using parameters defined from GFDL, HWRF and the operational GFS (see section 4.4). Solid lines are various HFIP stream 1.5 dynamical models. Note that for the cases in Figure 6, COAMPS-TC and the SPC3 statistical model showed a 10-20% improvement over the HFIP baseline after about 36 hours into the forecast. Of the Stream 1.5 models, COAMPS-TC and SPC3 performed better than the current operational models. COAMPS-TC (COTI) was the only dynamical model to outperform the statistical guidance at 24 h and beyond. COAMPS-TC has intensity forecast skill that exceeds the HFIP baseline at forecast times of 48 h and beyond (by more than 20%). All the models still have start up issues and it is taking about 12-24 h hours for the rapid error growth to decrease.

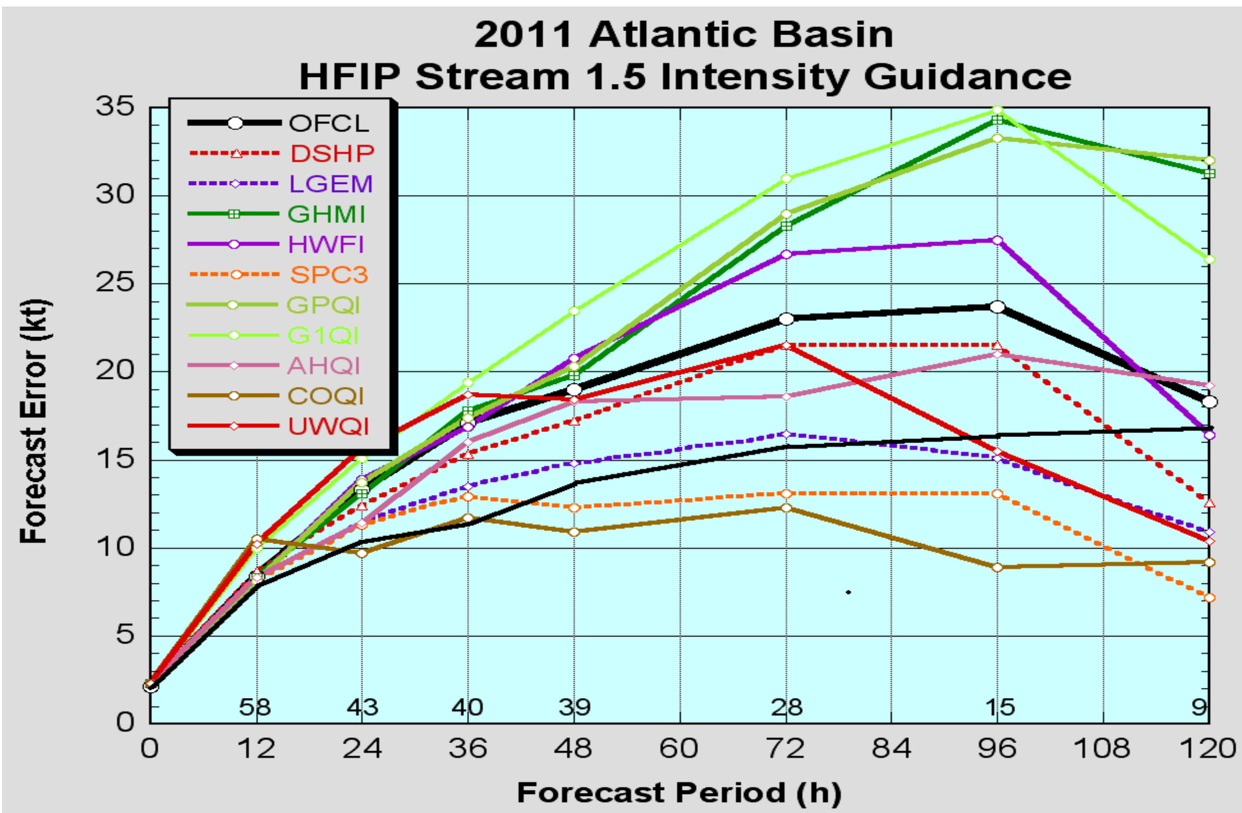


Figure 7. Intensity statistics from the Stream 1.5 models through mid-September. It is a homogeneous sample and shows absolute error. The dashed lines indicate statistical models and solid the dynamical model (from James Franklin, NHC). The various models are: OFCL = official forecast, DSHP = Decay Ships--Statistical Hurricane Intensity Prediction System, LGEM = Logistic Growth Equation Model, GHMI = operational GFDL model, HWFI = operational HWRF, SPC3 = Intensity consensus: 6 DSHP and LGEM with predictors from GFS/GFDL/HWRF, GPQI = GFDL Ensemble mean, G1Q1 = GFDL Ensemble member 1 (no bogus vortex), AHQI = NCAR Model, COQI = COAMPS TC and GFS initial conditions, UWQI = Univ of Wisc. Model

It was demonstrated by Fuqing Zhang and his group using the Penn State configuration of the ARW model that data collected by the NOAA P3 tail Doppler radar data in hurricanes can significantly improve forecasts of intensity over forecasts when the tail Doppler radar data are not included (Figure 8).   The EnKF data assimilation systems was used with the PSU model for this study.

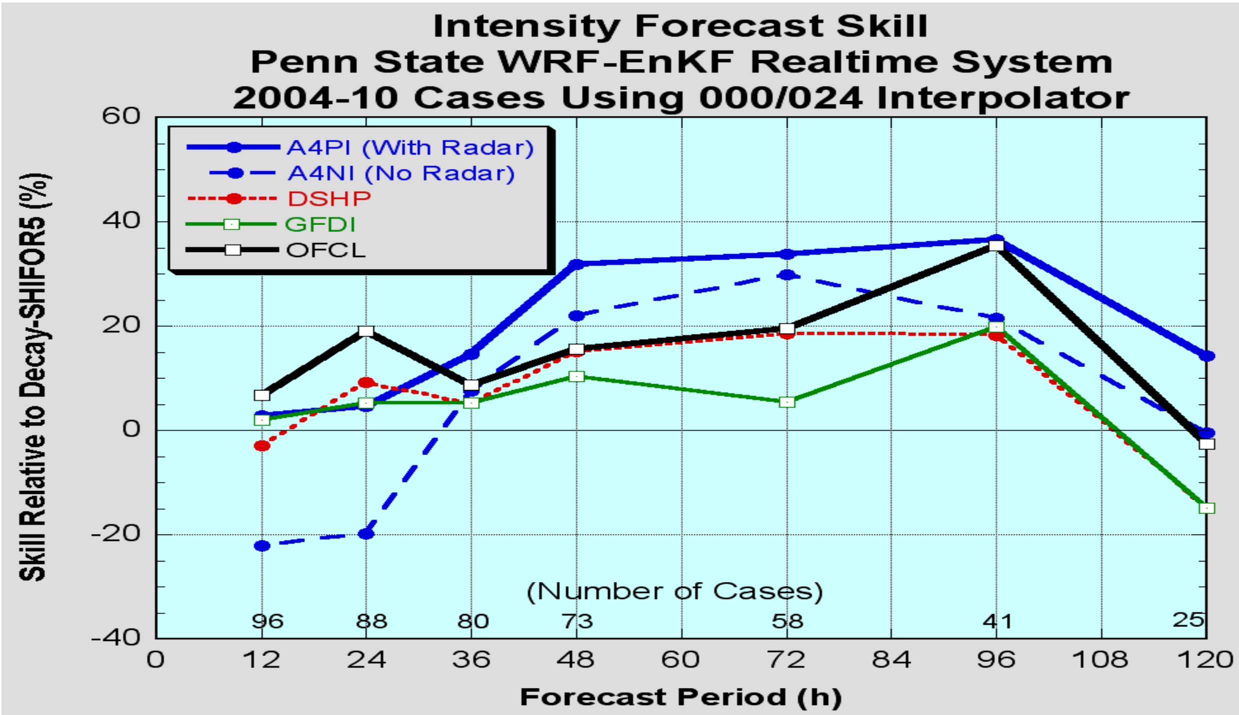


Figure 8. Impact of P3 tail Doppler radar data on intensity statistics from the PSU ARW using an EnKF data assimilation system. All cases with radar data from 2004-2010 are included. Data are personal communication from Fuqing Zhang (PSU) 2011.

The radar data from the PSU study (above) was verified by the group at HRD (Aksoy et. al. 2012) using the HWRF model and HEDAS data assimilation. The results compared to the HFIP baseline are shown in Figure 9. In the figure all cases for which radar was taken in storms from 2008-2010 are included.

Note that for all forecast lead times beyond 36 hours inclusion of radar data improves the forecasts by 20% and as much as 40% for these cases.   It is likely that this result is because the high resolution data near the core better defines the environmental flow within the hurricane and hence the impact out to 5 days. It avoids the use of filtering and bogusing to create initial conditions that are used in most regional hurricane models. Note also there is a problem in the early part of the forecast where the model is still adjusting to the initial conditions. Fixing this initial problem is a major current focus of the HFIP Program.

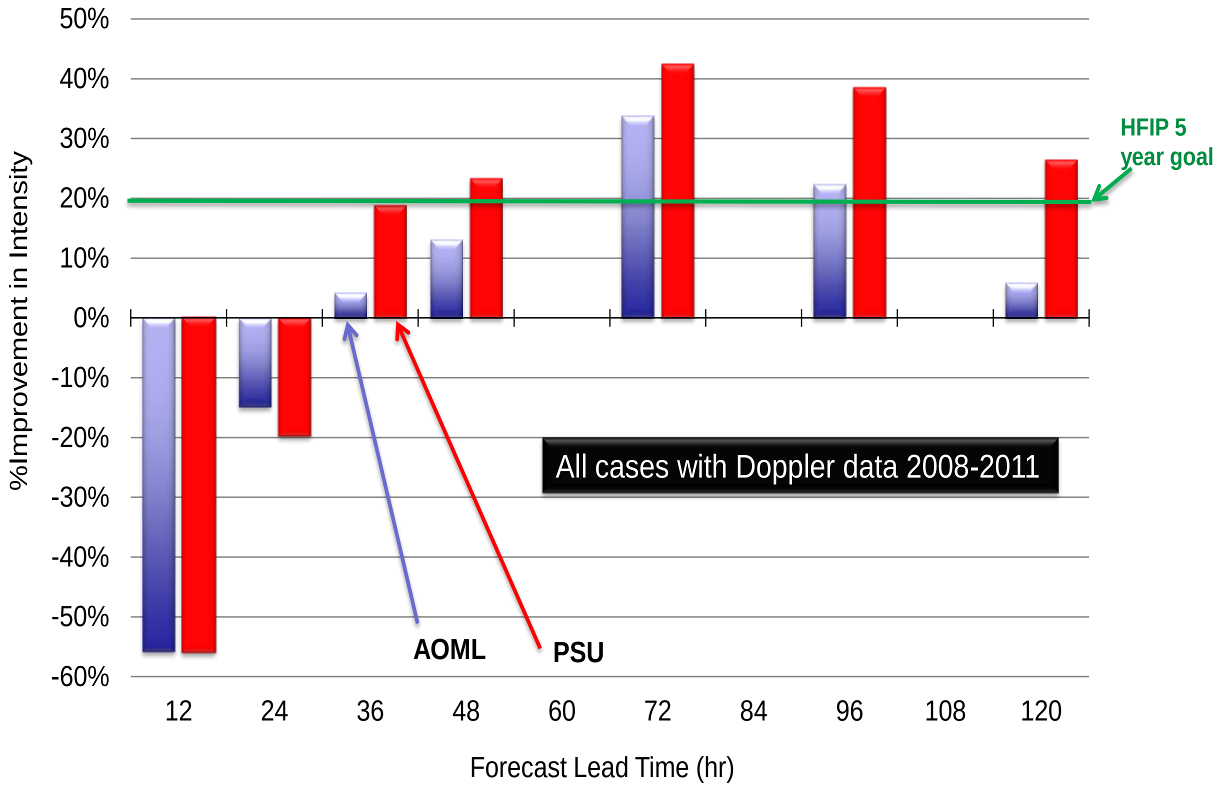


Figure 9. Impact of P3 tail Doppler radar data on intensity statistics from the HRD/EMC 27/9/3 HWRF using HEDAS and the PSU ARW using an EnKF data assimilation system. All cases with radar data from 2008-2010 are included. It is a homogeneous sample and shows error relative to the HFIP baseline (posivive means above the HFIP baseline). Blue bars are for the PSU Model and red for HWRF. Number of cases available at each forecast lead time shown across the top. Data are personal communication from Sim Aberson (HRD) and Fuqing Zhang (PSU) 2011.

The above results suggest that the newer regional models (COAMPS-TC, HWRF and the Penn State model) may provide perhaps a 10% improvement in intensity as compared to the HFIP baseline and adding the tail Doppler radar and other aircraft data, when available, may add an additional 20%-40%. This would meet the HFIP goal of 20% for those storms for which tail Doppler radar data is collected. For other storms HFIP is pursuing the use of high resolution satellite data near the hurricane. However, it should be noted for the small homogeneous sample, the models that assimilated radar observations did not have errors as low as COAMPS-TC, which did not assimilate radar data.

4.4. Statistical Post Processing of Model Output

Much of the discussion above focused on using model improvement to achieve the HFIP goals. We have already stated in section 4.3 that the statistical models (DSHIP, LGEM and SPC3) performed among the best as predictors of hurricane intensity. In fact, SPC3 provides comparable improvement as the radar data described above when compared to the operational statistical models. SPC3 (Figure 7) is an "ensemble" of six members created by using HWRF, GFDL and GFS operational models each with DSHIPS and LGEM. Figure 10 is further indication of the impact of statistical models and shows a comparison between the current operational HWRF and

GFDL models by themselves and a combination of DSHIP/LGEM using parameters determined from each of those two operational models.

Note that for both the GFDL and HWRF statistical models, using parameters from the respective operational model, gave an improvement of up to 20% over the parent operational model.

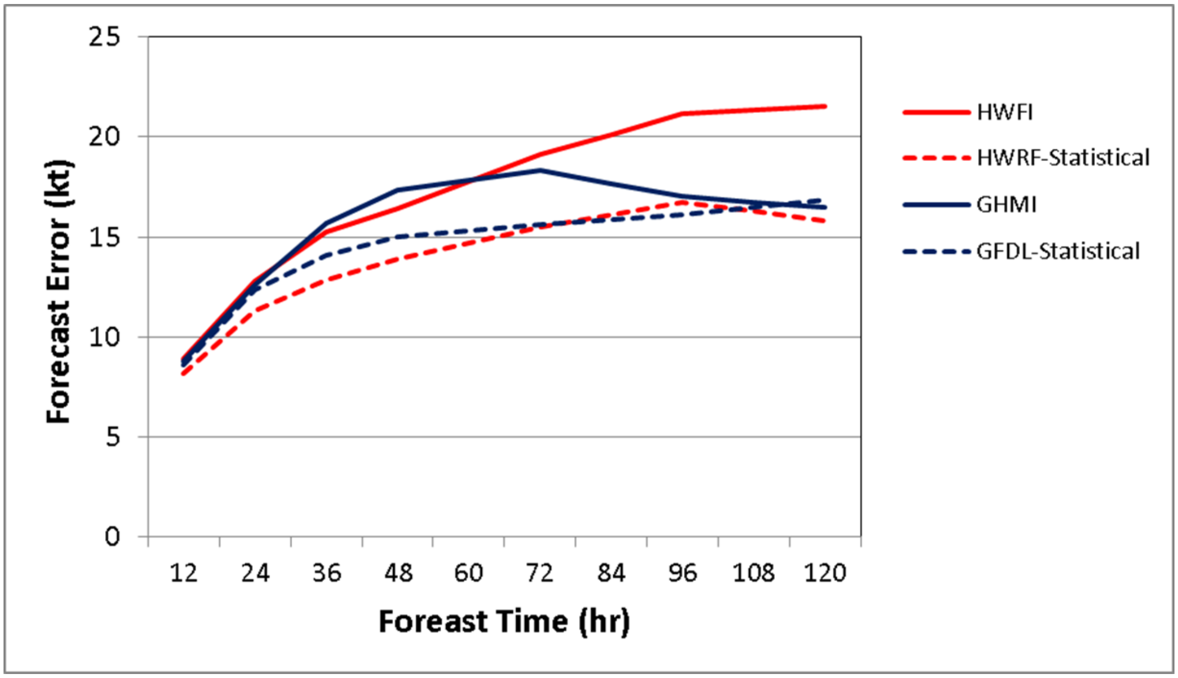


Figure 10. The impact of including the operational regional models (HWRF, GFDL) in the SHIPS statistical intensity model. All cases for 2008-2010 are included (Mark DeMaria, personal communication)

It is likely statistical post processing along with model improvements, the use of ensembles, improved data assimilation and inclusion of more data such as aircraft and satellite data promises to be the key to meeting the HFIP goals.

4.5. Reaching Rapid Intensification Goals

A major goal of HFIP is to improve the detection of rapid intensification (RI) and weakening decay (RW) with a high probability of detection and a low false alarm rate. At this point none of the HFIP dynamical models are capable of providing reliable forecasts of RI reliably in the first 36 hours. The global models are not able to resolve the inner core processes that are likely to be very important in the RI process and all the regional models have serious spin up (and spin down) problems (Figures 7 and 9). In some cases during the spin up (or spin down) period, intensity changes can meet or exceed the RI limits (30 mb in 12 hours). Occasionally, one or more of the HFIP regional models can do very well capturing RI (Figure 11 for example) but that figure is the exception rather than the rule. The spin up (spin down) problem is related to model initialization that has to be resolved before progress can be achieved toward the RI goal, especially in the most important first 36 hours.

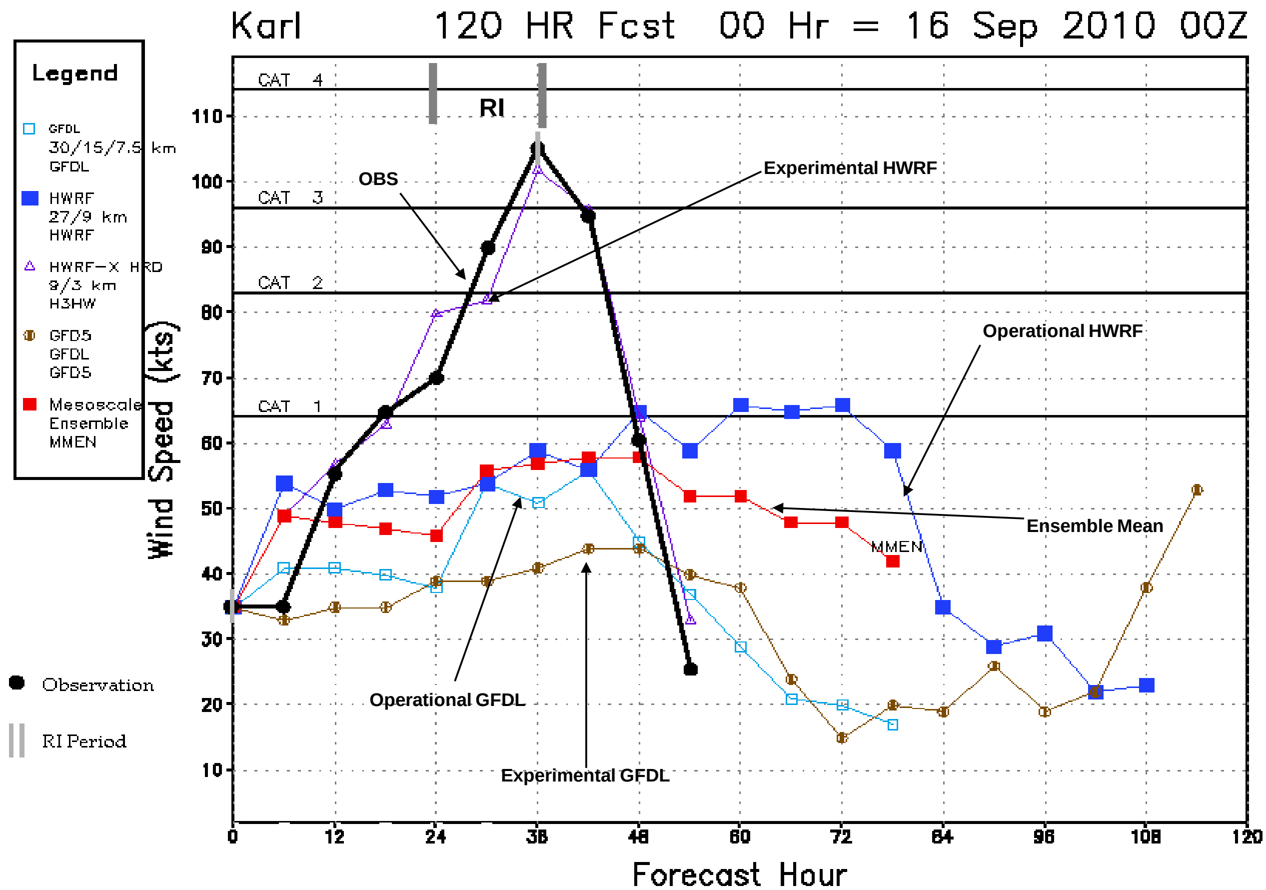


Figure 11. Wind speed forecasts from various models for hurricane Karl, 2010. The observed wind is shown as the black line. MME is the mean of HWRF, GFDL, and Experimental HWRF(Gopalakrishnan-personal communication)

The solution is likely to come from using data assimilation systems rather than some form of a bogus vortex initialization that is employed by most regional hurricane models. Additionally, advancements in physical parameterizations, particularly the microphysics and hurricane boundary layer representations, are needed. These issues are a current focus of the HFIP program.

5. Recent Accomplishments

In this and the following two sections we will outline and describe results from the summer 2011 demonstration program that indicate promising results beyond what is outlined in the previous sections, some lessons learned and remaining challenges.

Recent results include:

- The GFS model initialized with an EnKF Data assimilation system is showing significant ability to forecast tropical storm genesis out to 5 days or more

- Track forecasting skill with the global GFS/EnKF system at or better than the skill of the best global models in the world with 20% improvement for most forecast lead times
- The impact of high resolution data near the hurricane center using aircraft data (including the tail Doppler radar data) noted last year using the Penn State ARW model has been confirmed using HWRF – 20 to 40% improvement in intensity forecast
- The initialization system for the operational HWRF has been greatly improved
- There have been dramatic improvements in physical parameterizations in HWRF
- A 3 km version of HWRF, to be operational in 2012, is showing significant skill improvement over the current operational 9 km version at 84 hours and beyond
- New advancements in the COAMPS-TC data assimilation, initialization and physical parameterizations resulted in a greatly improved intensity forecast skill in 2011. COAMPS-TC was the only dynamical model to outperform the statistical guidance and exceeded the HFIP intensity goal for forecast times at 48 hours and beyond
- Statistical post processing of model output from two different statistical model approaches give the best intensity forecasts as compared to current dynamical and statistical models
- New model output products are providing more useful information from the models, especially ensembles
- New diagnostic verification tools are providing more useful information for model evaluation.

5.1. Forecasting Tropical Storm Genesis Out to 5 Days or More.

The GFS/EnKF global system that HFIP ran during the 2011 demo season showed a remarkable ability to forecast hurricane at least out to 5 days and perhaps to 7. We have just completed a tracker system (by Tim Marchok at GFDL) so statistical confirmation of this result will be forthcoming but we present below some subjective evidence of genesis skill. We do note here that at least some of the global models besides the GFS/EnKF system show similar skill, particularly the ECMWF and the operational GFS (an example is shown later). The subjective evaluation of genesis skill for both the West Pacific and the Atlantic in 2011 suggests a high Probability of Detection (POD) and a low False Alarm Rate (FAR).

The EnKF global system is outlined in Table 3 and some skill results are shown in Figures 4 and 5. It was run as a 20 member ensemble at a resolution of T254 or about 60km. Figure 12 shows a forecast starting at 1200Z 18 August, 2011 for hurricane Irene using the GFS/EnKF ensemble. The hurricane was named 0000Z on August 24 or 3 and a half days after the start of this forecast. Shown are the various ellipses of position determined by the ensemble that contain 80% of the members. The ellipses that elongate along the track indicate a predominant speed uncertainty and ellipses that elongate perpendicular to the track indicate a predominant position uncertainty left or right. The green track shown in the figure is the best track from NHC after the storm was named. The forecast was excellent though slightly to the left of the actual track after 7 days. The forecast was begun when NHC initially indicated that the wave that was to become Irene was listed as an investigation area for possible development.

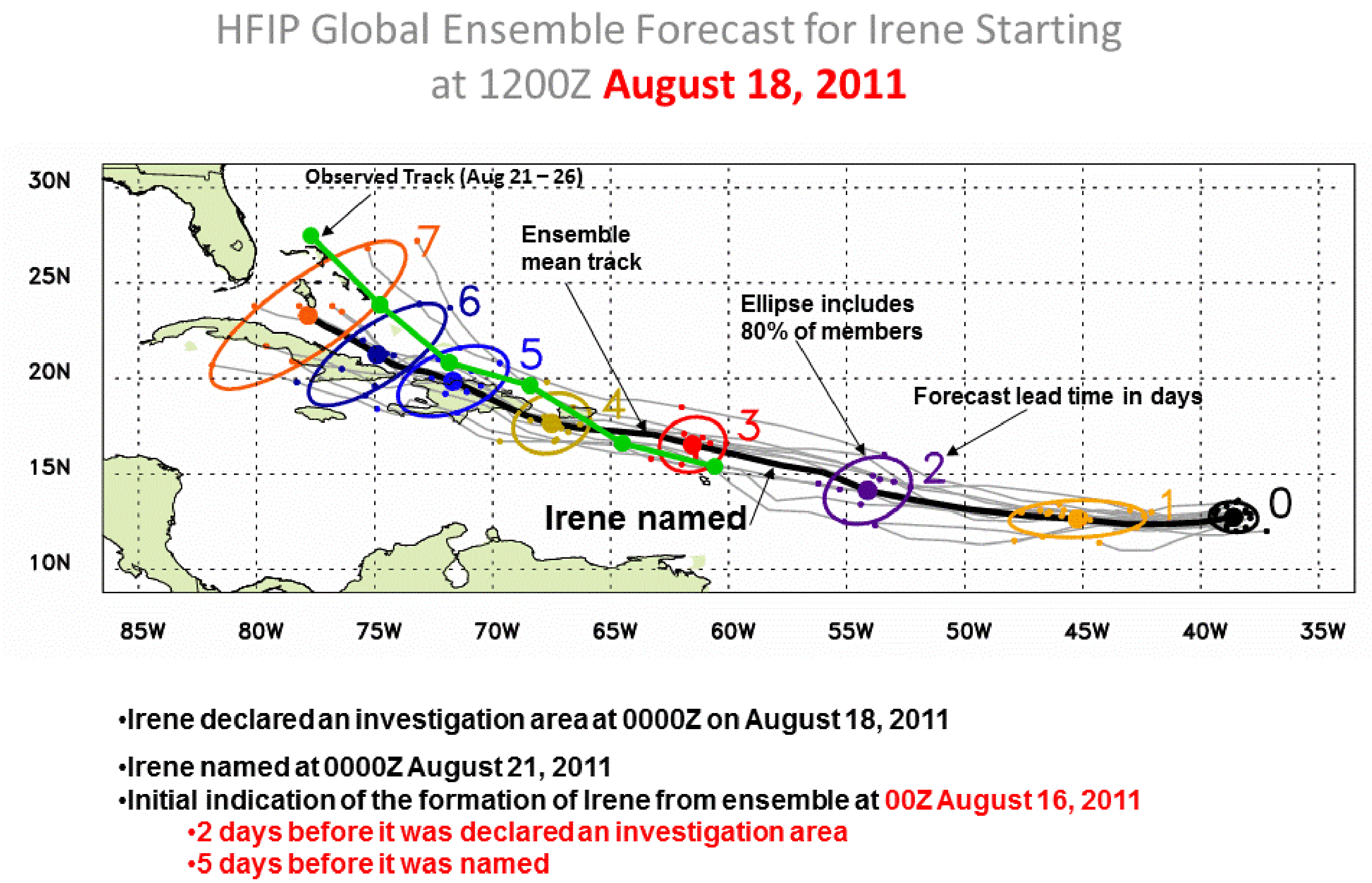


Figure 12. Seven day ensemble forecast made by the GFS/EnKF global system starting 1200Z August 18, 2011. This was a 22 member ensemble (including the ops GFS and a control version of the GFS/EnKF) and various details of the plot are noted on the figure.

Figure 13 shows two "postage stamp" presentations of forecasts from the various members of the GSF/EnKF ensemble. The top panel was initialized at 0600Z on August 16, two days before the wave was listed as an investigation area. By visually counting those members that were indicating a tropical storm after 7 days (17 out of 22) one could say that genesis of a tropical storm after a week was 77% based on the ensemble forecast and very near its observed position (between days 4 and 5 on Figure 12). This forecast was two days before the wave that became Irene was identified as an investigation area and 5 days before it was named.

The second panel in Figure 13 shows the "postage stamps" from the forecast initiated at 1200Z 18 August, 2011 (Figure 12). Again, by counting members with storms, all but 1 (95%) forecast a storm and half indicated that the storm would be strong (50%). At this time Irene was a hurricane

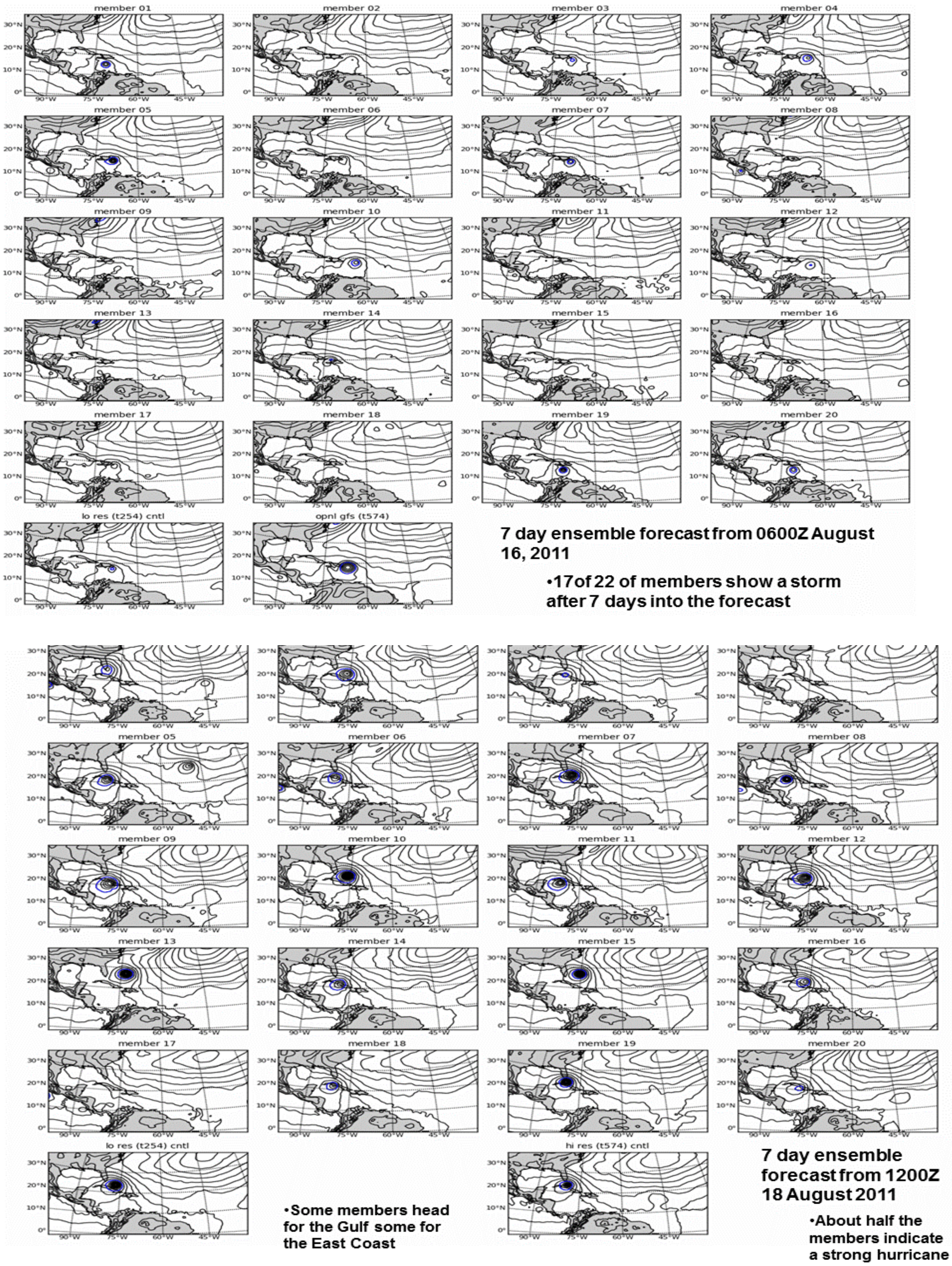


Figure 13. "Postage stamps" depicting surface pressure (mb) from each GFS/EnKF ensemble member at day 7 of the forecast. See text for further discussion. Upper panel for forecast initialized at 0600Z August 16, 2011, and lower panel initialized at 1200Z August 18, 2011.

This is just one example, for Irene, but we observed similar skill for most of the other storms in the Atlantic and West Pacific this season with the exception of a couple of storms that were very small when they formed (such as Jose). Remember that the resolution of the GFS/EnKF ensemble was about 60 km which cannot resolve storms that are essentially growing from a single thunderstorm.

Figure 14 is a hypothetical product that could be generated from information from a global ensemble forecast. It is for August 16 near the time that the forecast in the upper panel of Figure 13 was initialized. It is the official Graphical Tropical Weather Outlook generated by NHC several times a day. Figure 14 shows tropical storm Gert and a yellow ellipse that indicated an investigation area (an area where NHC was watching for tropical cyclone development). That area eventually became Harvey. The white ellipse off Africa is the wave that became Irene five days before Irene was named. The white circle is the area where Irene was expected to form from the forecast initiated near the time of Figure 12. This kind of ensemble information could easily be shown to forecasters and we show a real example in section 5.8.

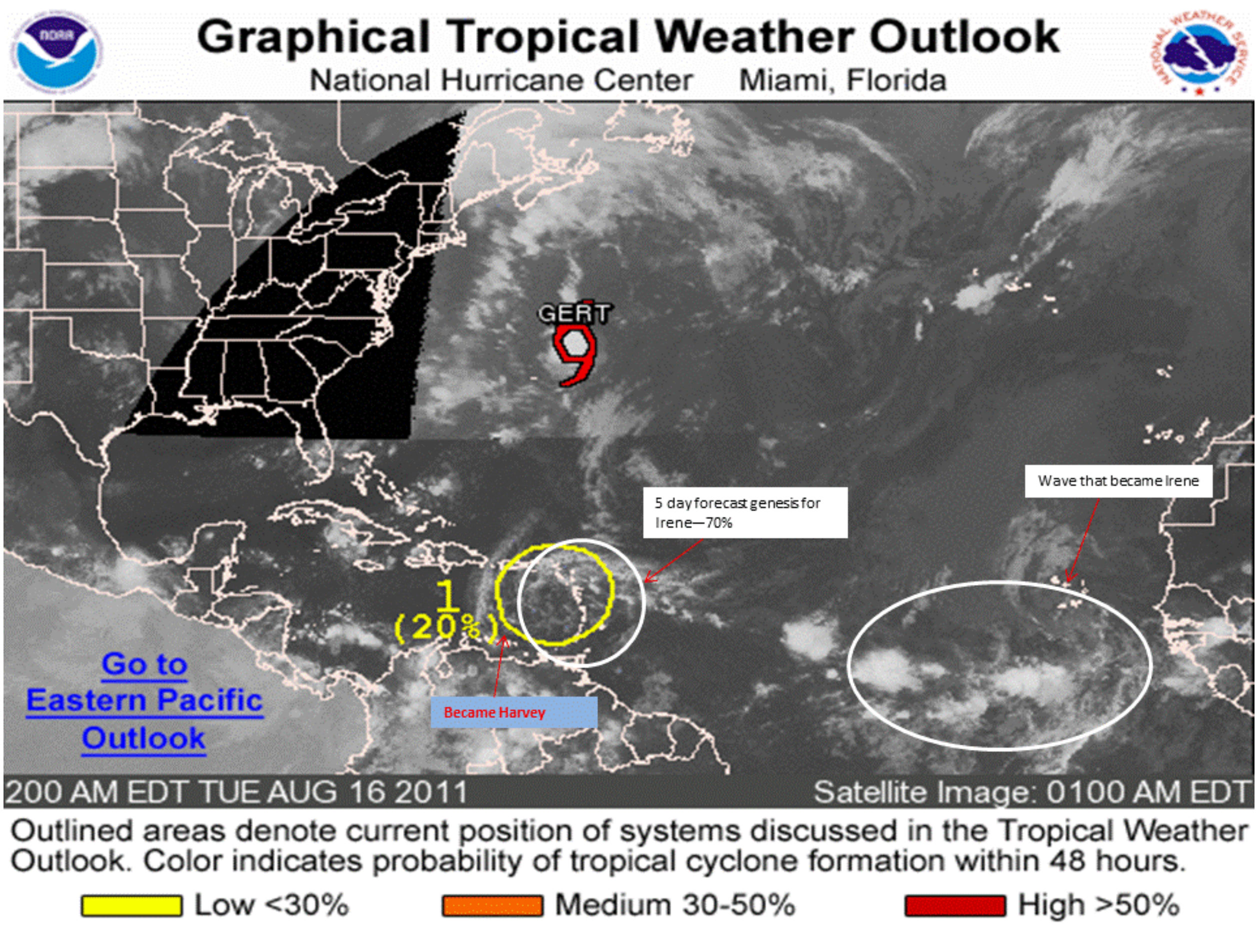


Figure 14. Graphical Tropical Weather Outlook form NHC, 16 August, 2011. Hypothetical product for genesis shown by white ellipses. See figure for other notation.

5.2. *Track forecasting skill*

In section 4.2 we discussed the skill in the HFIP GFS global model in track forecasting. The skill seems to be coming from the EnKF DA assimilation system since the skill is considerably higher than the corresponding model, the operational GFS that uses the GSI data assimilation system. It has approximately the same skill as the ECMWF model which is widely viewed as the best model for hurricane track in the world. Note the track skill is near or exceeding the 20% improvement relative to the HFIP baseline.

Figure 15 shows additional evidence of this skill by comparing the GFS/EnKF system to four operational global ensembles including those for NCEP, ECMWF, CMC, and UKMO. The HFIP GFS/EnKF system provides more accurate track predictions as compared to all centers except ECMWF (which uses a 4DVAR DA system thought to be comparable to EnKF) where the two systems have similar skill.

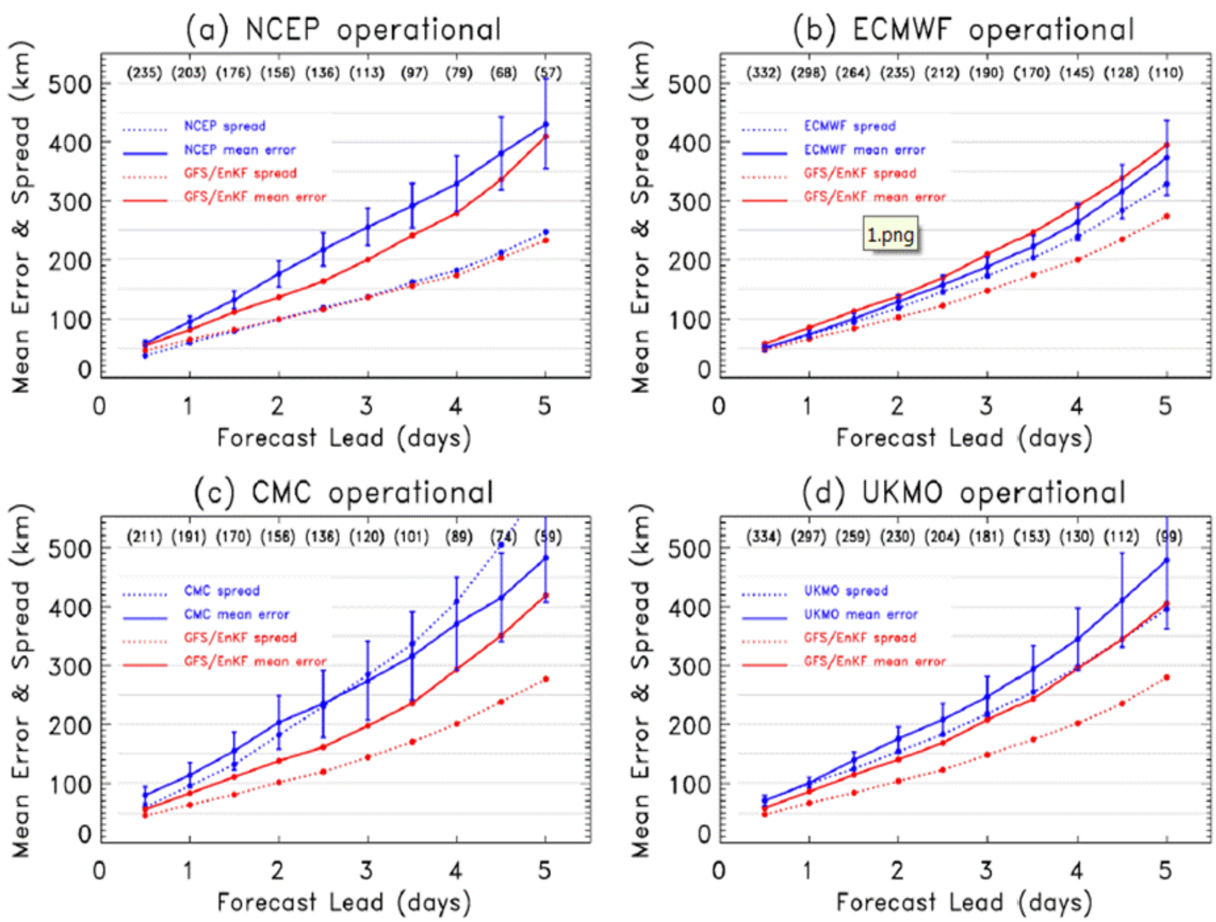


Figure 15. Comparison of track forecasts from various operational center ensembles (NCEP, ECMWF, UKMO and CMC) blue curves and the GFS/EnKF system run by HFIP (red). Solid lines are for RMSE and dashed lines are for ensemble spread. Plots are homogeneous and error bars are shown for the operational models. Number of cases is indicated in parentheses.

Based on these results HFIP expects to meet its 5 year 20% goal by using more advanced data assimilation systems than GSI (or 3DVAR). This goal can be met using already existing global models at NCEP by changing to an EnKF or Hybrid (combined GSI and EnKF) system. The GFS system at NCEP when run with high resolution and a hybrid data assimilation system is more than adequate to meeting HFIP track goals

5.3. The Impact of High Resolution Data

Last year we reported for the first time incorporation of tail Doppler radar data taken by the NOAA P3s provided dramatic improvement in forecasting hurricane intensity out to 5 days though with an adjustment period for the first day or so of the forecast where the skill was lower. This was illustrated in Figure 9. The results for last year were from the PSU group using the WRF ARW model. That result was again confirmed by them using the storms from this year and in addition the PSU results were confirmed by results from AOML, blue bars in Figure 9. Red bars are the PSU result. Both show very similar results even though the two models are different (HWRF used by AOML and ARW used by PSU) and each used different versions of the EnKF data assimilation system

In addition, AOML used not only the tail Doppler radar data but all other data transmitted from the aircraft in real-time (including flight level, dropsonde, SMFR surface wind data) as well as dropsonde and flight level data collected by the Air Force. They tried experiments with all aircraft data including the radar data and all aircraft data excluding the radar data. Both gave similar results with strong improvement over not using any aircraft data. A 10% additional improvement was seen using radar data with other aircraft data as compared to all data excluding radar. This suggests that it is not only the radar but also the high spatial resolution data near the center that is making the difference. The radar improves on the other data because it is of higher spatial resolution.

It was somewhat of a surprise that the impact of the inner core data appears to last out to 5 days and perhaps beyond. Many would have expected that the impact of this data to have decayed with the predictability timescales of the convective processes in the core. However it is likely that, in addition to initializing at least some of the convective scales near the core, the data near the core is also better specifying the larger scale components that is part of the environment in the region near the hurricane and which have longer predictability than convective process and which determine the evolution of the storm in structure and intensity as well as it’s track.

We note here that these results stress the need to use initialization schemes that can better define the storm environment. Initialization schemes that use some sort of bogus initial vortex probably don’t do this well if at all. In addition it also stresses the need to find better ways to use available satellite information at higher resolution near the hurricane for those storms when there is no additional aircraft data.

One limitation regarding using data from various sources in GSI is that the data has to be in PrepBUFR format. To facilitate the community in converting data to PrepBUFR format, HFIP and DTC have made available a website with extensive documentation, examples and tools.

5.4. A 3 km Version of HWRF Showing Significant Skill Improvement

During the 2011 HFIP demo project, a parallel model to the operational HWRF was run. The operational model had two domains with grid resolutions of 27km and 9km respectively. The parallel model had three domains resolutions of 27km, 9km and 3km respectively. Figure 16 shows the error statistics from retrospective runs from 2008-2010 and compares the operational

HWRF, and GFDL and the SHIFOR5 statistical models to the 3 km HWRF. Note that the 3 km model was very similar to the operational HWRF but produced better intensity forecasts out to 84 hours as compared to GFDL and HWRF. Beyond 84 hours the performance was substantially better than the operational model and about the same as the GFDL operational model. The triply nested HWRF will be made operational for the 2012 hurricane season.

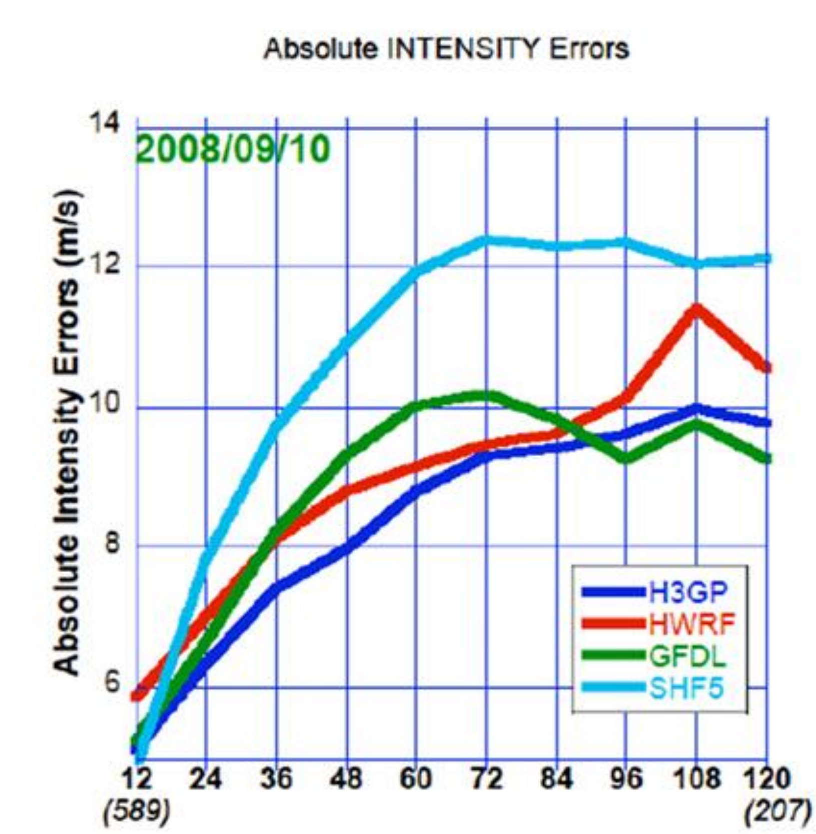


Figure 16. Absolute intensity errors for the 2008-2010 hurricane seasons for the 3 km HWRF (H3GP), operational HWRF (HWRF), operational GFDL model (GFDL) and SHIFOR5 (SHF5) statistical model. The sample is homogeneous and the number of cases is outlines in parentheses.

5.5. Initialization of the operational HWRF has been substantially improved

In past years we have noted that there are major problems with the initialization of all the dynamical models including the operational HWRF. One of the most dramatic problem was a tendency for HWRF to spin up weak storms and greatly spin down strong storms. This is illustrated in Figure 17 by the red bars which show the wind speed bias after 60 hours as a function of starting initial intensity. Note that the initialization scheme adjusts the initial intensity to match the initial observed intensity so the initial intensity bias is zero. Thus an increase by 60 hours represents a spin up and a decrease represents a spin down.

During this last year, the EMC HWRF team spent considerable effort on improving the initialization of the model including adding size parameters (from the NHC observations) in constructing the initial vortex. The result of this effort is shown in Figure 17 by the blue bars.

Spin up vs. spin down has the same meaning as for the red bars but except for the very strongest storms (where the sample size is small) the sign of the blue bars is basically randomly distributed over the various storm strengths. The bias in the storms is greatly reduced over the 2010 version of the HWRF operational model.

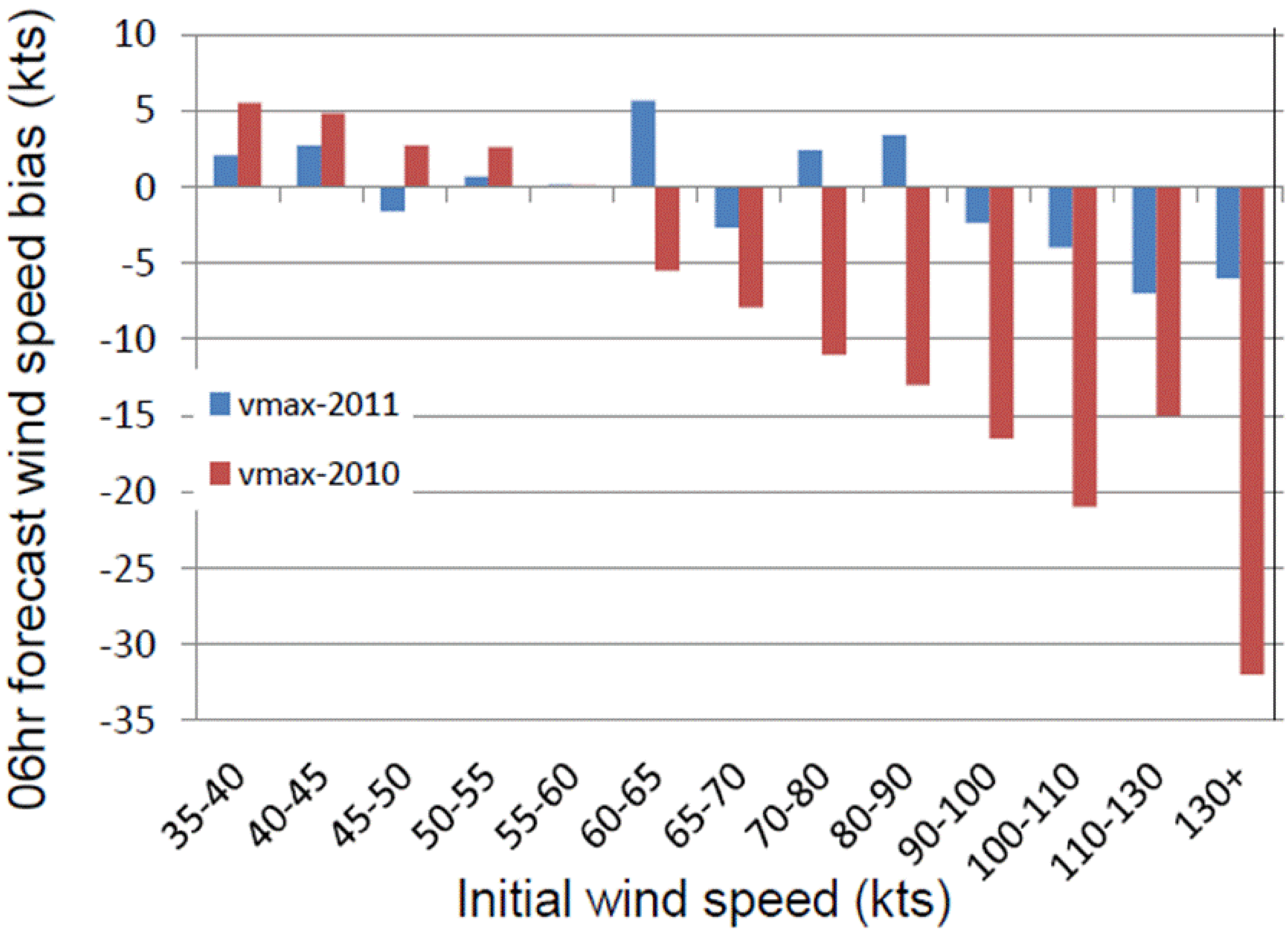


Figure 17. Forecast wind speed bias after 60 hours for the HWRF operational models in 2010 and 2011 as a function of initial wind speed.

5.6. Physical parameterizations in HWRF

An obvious statement is that there are three ways to improve models: better initialization, better model core and better model physics. It is generally felt in HFIP that improvement in the model core (the finite differencing, model grid etc.) gives only marginally improved skill but initialization (like discussed in the earlier sections) and model physics can make big differences.

The HWRF model team has also made various changes in the HWRF physics package and some of those changes are illustrated in Figure 18. The changes shown in the upper panels of the figure are related to the surface heat and momentum flux coefficients. The gray dots are that the model computed for various wind speeds and the color lines or dots are observed values. The model is now much closer to the observed values than in the past. The lower panels show the mixing coefficients in the PBL from the original GFS physics package used previously in HWRF and the modified scheme used this last summer in HWRF. Note the closer comparison with the observed values (colored).

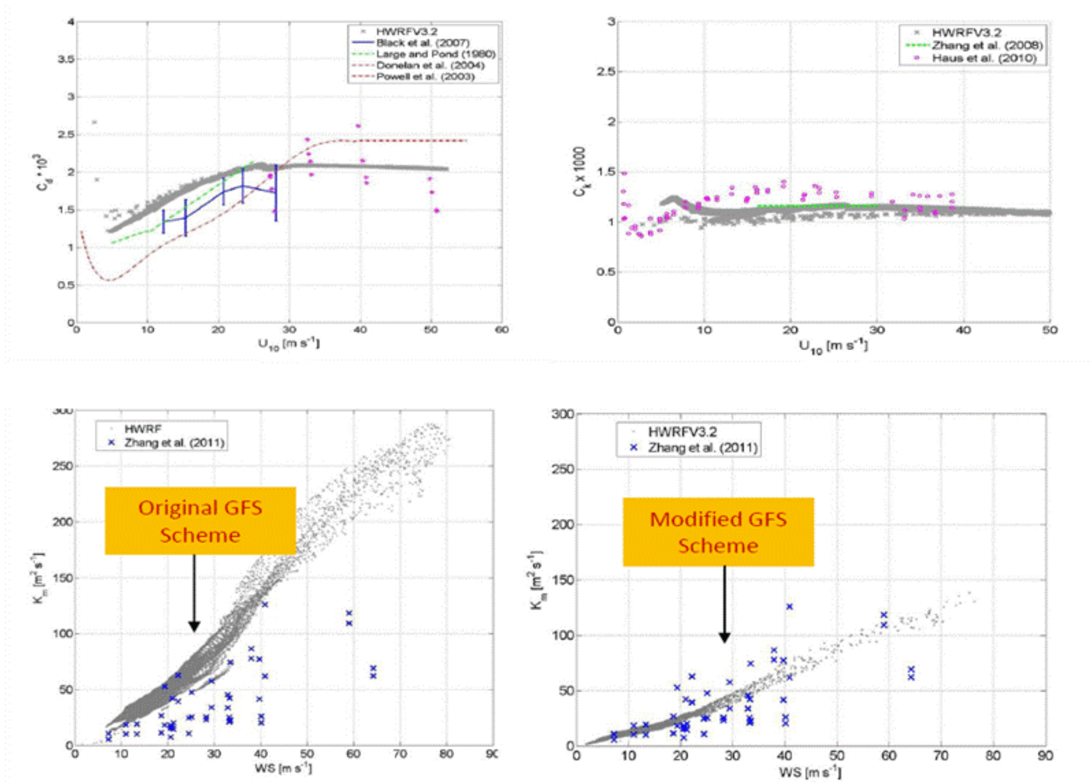


Figure 18. Comparison of Cd (drag coefficient), Ck (surface heat flux coefficient) and Km (momentum mixing in the boundary layer) used in the high resolution HWRF 27-9-3. Gray dots are computed values from the model and color dots or lines show various observation data sets.

5.7. Statistical Models

In previous reports we demonstrated the skill of model output can be further improved using statistical post processing. Figures 19 and 20 give two separate examples of this.

Figure 19 compares skill of various models compared to SHIFOR5, a standard used by NHC for guidance comparison. GFDL and HWRF operational dynamical models are shown. Also shown is the Intensity Consensus (ICON) prediction. Of interest here is the SPC3 (Statistical Prediction of Intensity with a Consensus Ensemble (SPICE), 3 parent model version) an HFIP experimental statistical model. This is a combination of 6 different models, SHIPS using results from the operational GFDL, HWRF and GFS dynamical models and LGEM using output from the same three dynamical models giving an ensemble of 6 which is then averaged. In this diagram higher numbers are better. Beyond 36 hours the SPC3 exceeds the skill of all the guidance systems shown, including what is considered the best of the guidance, LGEM, especially at the longer lead times. The SPC3 system was an attempt to see if using model output data in the statistical models would give an improvement and that clearly is the case for intensity at the longer lead times. The goal for next year is to use additional models from the HFIP stream 1.5 models in the statistical models. Since those models have been better than HWRF and GFDL it is expected we will see an even greater increase in skill than is shown in Figure 19.

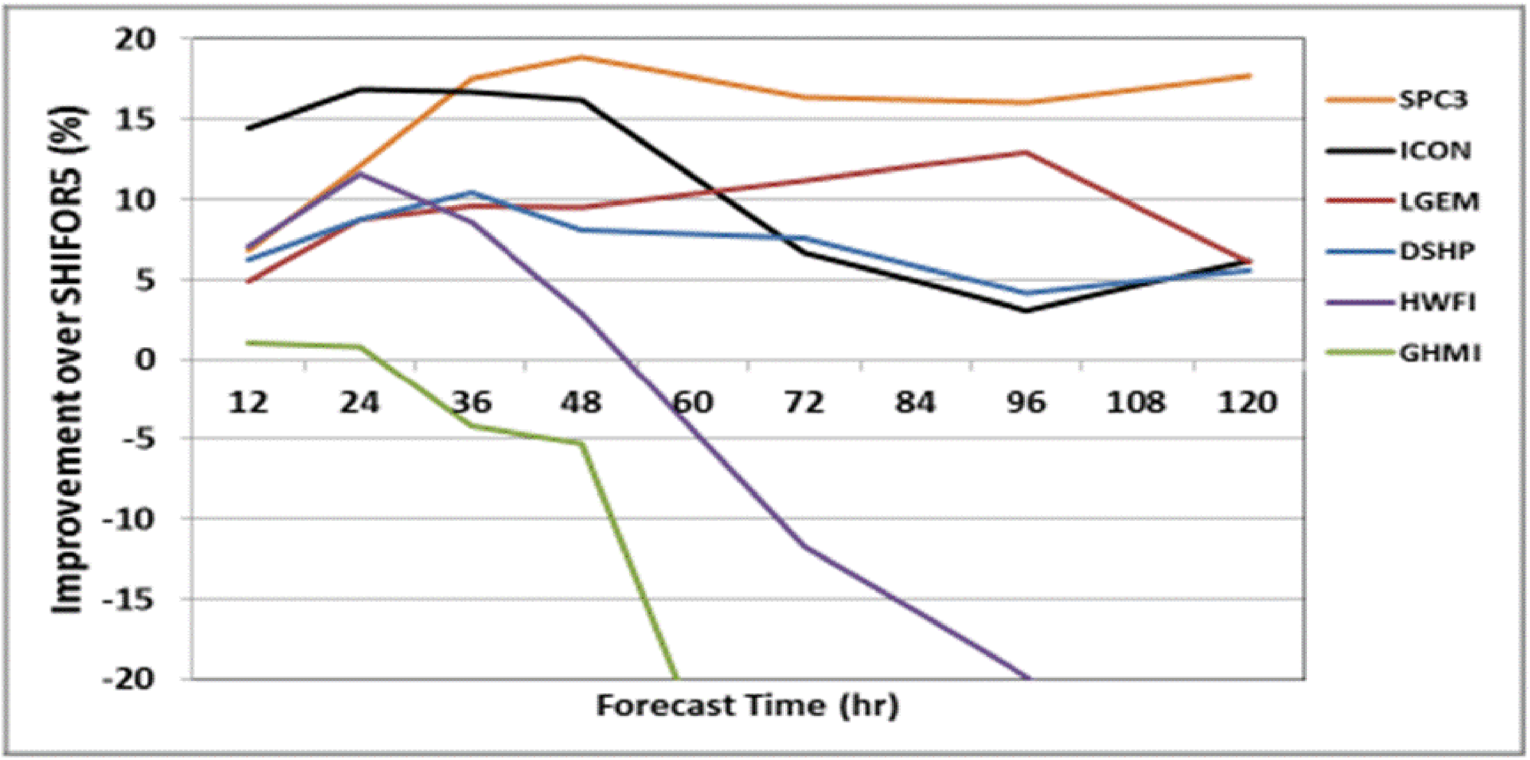
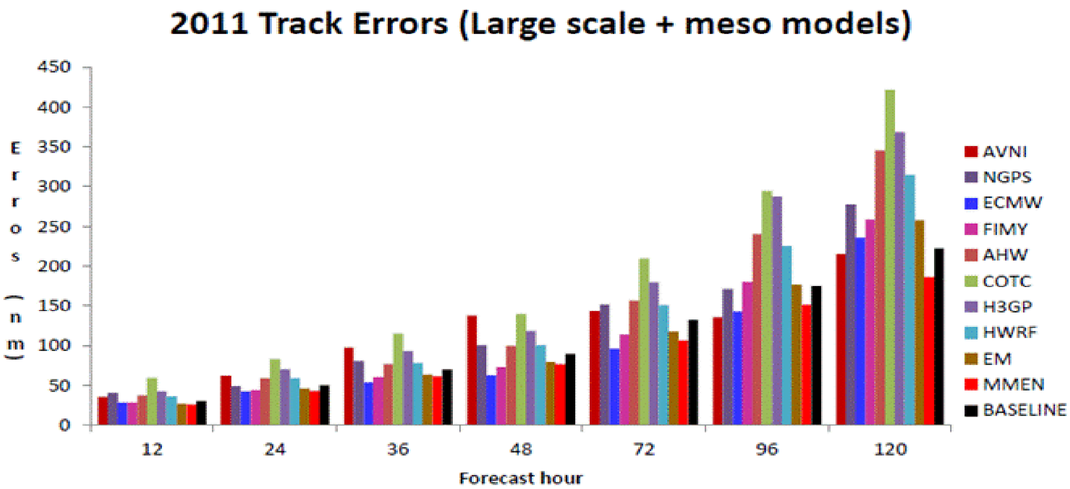


Figure 19. Intensity skill of various models relative to SHIFOR5. SHIFOR5 is an operational statistical and improvements relative to that model show increased skill relative to that model. Higher numbers are better. This type of comparison takes into account the relative difficulty of a particular forecast since that measure is contained in the SHIFOR5 statistics. SPC# is the HFIP Spice model that uses both DSHIPS and LGEM each with input parameters from the operational GFDL, HWRF and GFS for a total of 6 different statistical estimates. These are then averaged for the SPC3 prediction. ICON is the intensity consensus that NHC produces from those models it has available in real-time, LGEM and DSHP (Decay Ships) are two operational statistical models, HWFI and GHMI are the operational HWRF and GFDL models respectively.

Figure 20 shows another example of statistical post processing of model output. The system shown in Figure 20 was also shown last year and this year it includes even more models including both regional and global models. It is the FSU Multi-model Ensemble, the various models are indicated in the figure and both track and intensity are shown. The sample isn't homogeneous but includes most cases from 2011. EM is the mean of the ensemble of all the models shown; in other words an equally weighted mean. MMEN is a weighted mean of the various model members (similar to the Correlation Based Consensus discussed last year) where the weights are determined based on historical performance of the various models. Like the SPC3 system shown in Figure 19, the multi model ensemble (MMEN) is the best performer of all the models shown in Figure 20 and is better than the HFIP baseline for Intensity at the longer lead times, up to 30% better.

These results from statistical post processing of the model emphasize that the best strategy to meet the HFIP goals is to use this technique to enhance the skill of the dynamical models that are the core of the HFIP strategy. It is expected that the skill of the post processing is dependent on individual model components that go into the statistical process. Hence, the HFIP strategy is to emphasize improvement in the dynamical models, both global and regional, along with development of the best strategies for post processing. In this section we have shown two different post processing strategies. This also emphasizes the value of ensembles in enhancing the skill of any component in the forecast guidance process.



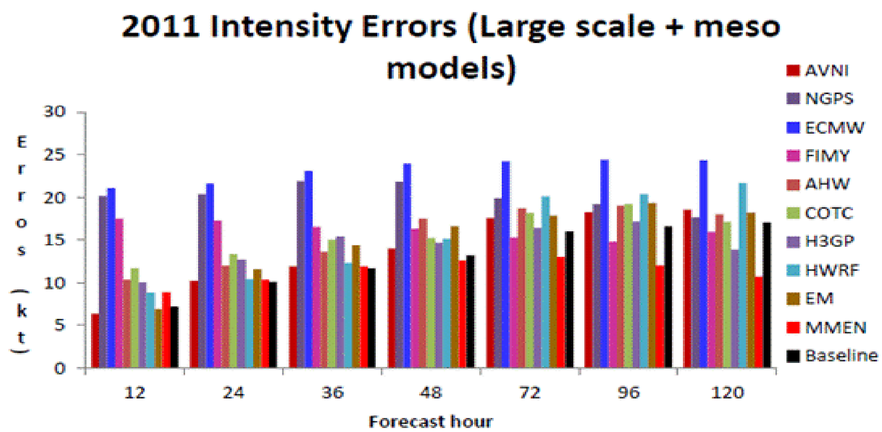
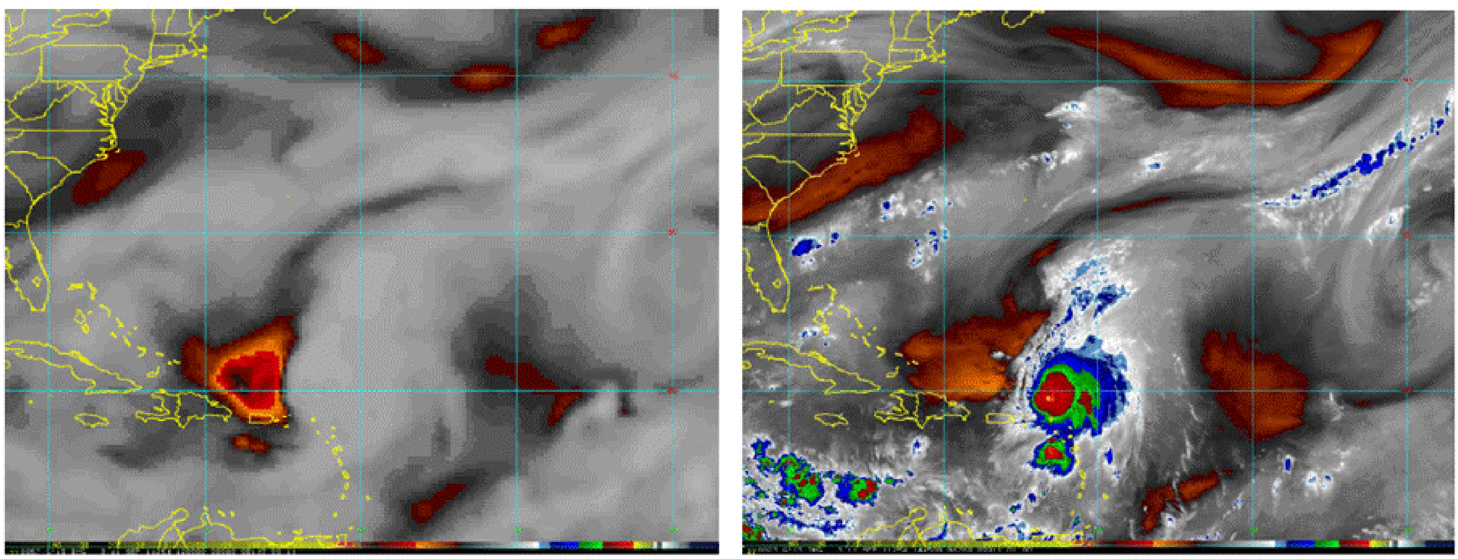


Figure 20. Non homogeneous comparison of mean track (upper figure) and RMS intensity errors from various models compared to the weighted ensemble mean of all the models (MMEN—red bar). The various models are: AVNO—operational global model, NGPS—Navy global model, ECMW—ECMWF global model, FIMY—experimental FIM global model, AHW—NCAR hurricane regional model, COTC—Navy regional hurricane model, H3GP—Experimental 3km HWRF, HWRF—operational HWRF, EM is the ensemble mean and baseline is the HFIP baseline.

5.8. New Model Output Products

It is not sufficient to simply increase the skill of the dynamical model and statistical model guidance but also to improve methods to display the information coming out of these systems to the hurricane forecasters. Ultimately it is the human forecasters who turn the guidance from all the data sources including various models described above into a forecast of the hurricane. The need for new ways to display the model output is especially critical for output from ensemble systems. In Figures 21-23, we display a few of the products that are being developed and used by the HFIP community.

Figure 21 shows a simulated water vapor image derived from the operational HWRF in the left hand panel. The right hand panel shows the observed image. This type of product is not new in forecasting but has just been implemented in the last year or so in the hurricane forecast system. The value of this type of product is both for forecasting and analysis of the model. For forecasting, it can reveal how the model is evolving the larger scale fields that influence the hurricane and hence whether the forecaster can put trust in the model results since it can be compared to the observed images that may have been taken after the model was initialized. It also can also be used to assess the initialization and spin-up processes in the model simulations. For example, in the left panel of the Figure 21, there are very few cold cloud tops near the center of synthetic image compared to the observed image in the right panel. These cold cloud tops appear in the synthetic images a few hours into the model integration.



Synthetic GOES-East Ch 3 from Maria HWRF real time run 9/11/11 12 UTC

Observed GOES-East Ch 3

Figure 21. Synthetic satellite water vapor (channel 3) image from the operational HWRF, left panel. The right panel shows the observed water vapor (channel 3) image.

Figure 22 provides a different way to display information from an ensemble. The upper part of the figure shows an example from the GFS/EnKF ensemble and depicts the cumulative probability of storm force winds for the storms/hurricanes being tracked by NHC. This example is for Ophelia and Philippe. For comparison the equivalent operational product from NHC is shown in the bottom part of the figure. In comparing these figures note that the color scales are very different and that the map backgrounds are also different. The probabilities in the current operational product are derived from climatology and hence show a wider distribution, lower probabilities and a probability that decreases with time along the storm track that may not be due to weakening of the storm. It drops because of reduced confidence with time. The probabilities in the upper panel are derived from the individual ensemble members so if all members at some point in the forecast period have storm force winds at a particular point then the probability is 100%. Hence the probabilities depend on the spread of the ensemble and not climatology probabilities. Thus the distribution of probabilities about the mean storm track is tighter and the probabilities much higher. Whether this adds value to a forecaster will depend on verification of the results illustrated in Figure 22 which has not been done yet.

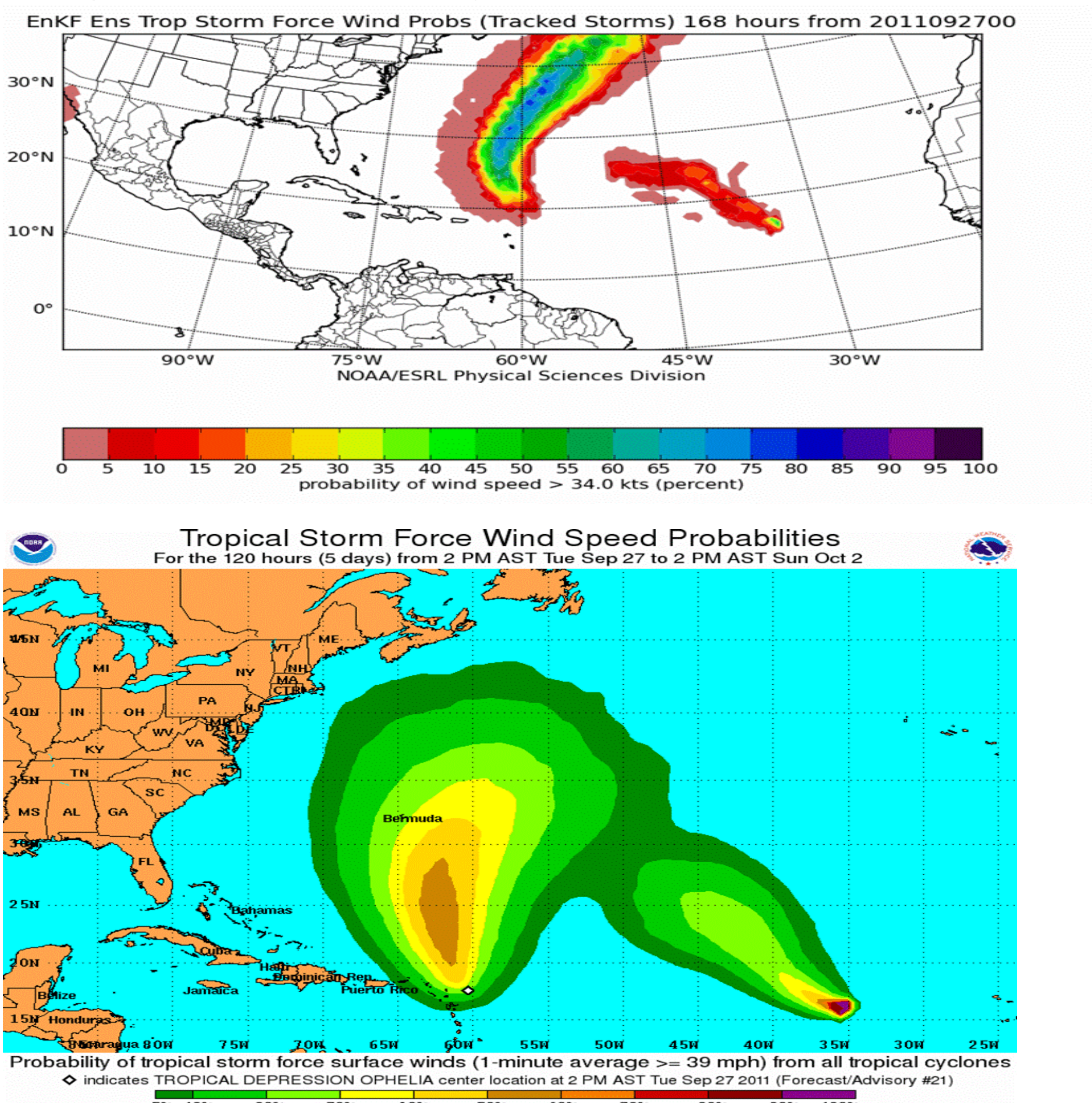


Figure 22. Upper panel shows the cumulative probability over a 5 day period of tropical storm force winds as determined from the GFS/EnKF 20 member ensemble. Lower panel similar operational figure where probabilities are determined from historical error distributions. Note that the color scales in the two figures are different.

Figure 23 is an actual example of the type of product discussed earlier in section 5.1 (Forecasting Tropical Storm Genesis out to 5 Days or more) and illustrated schematically in Figure 14. The initial time of the forecast is prior to the development of Ophelia and Philippe in the Atlantic and Hillary in the East pacific. The left hand panel shows the tracks of storms from the NCEP ensemble derived from the tracker developed at GFDL. On the right the areas of expected tropical cyclogenesis during the first 24 hours of the forecast period are shown. Note that the three areas with the highest probability of genesis were regions where hurricanes Ophelia, Philippe and Hillary eventually formed and that the probability of genesis indicated in between 80% and 90%. The area near Cuba with a probability of genesis less that 40% did not develop into a storm.

When Figure 23 is compared with Figure 14, it is very close to what was proposed in section 5.1.

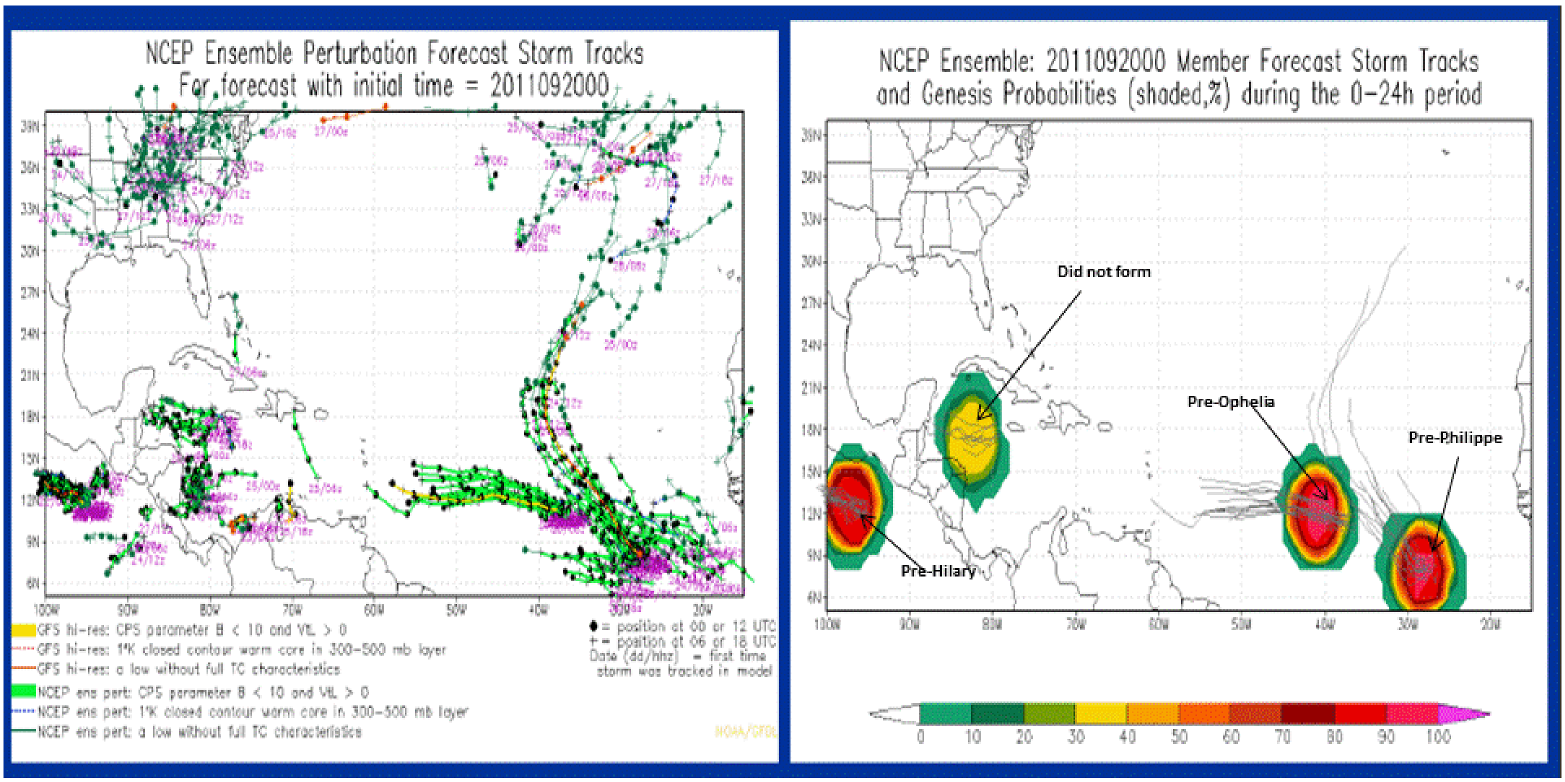


Figure 23. An experimental tropical cyclogenesis product (right panel) derived from the GFDL tracker system (left panel) from the operational GFS high resolution deterministic model and the operational GEFS ensemble. In the right panel the colors denote probability of genesis within the first 24 hours of the forecast period and the gray lines show the ensemble. This is an actual figure to be compared with the conceptual figure shown in Figure 12.

5.9. *COAMPS-TC Ensemble Forecasts*

While research is ongoing to improve deterministic atmospheric forecasts through advancements to the forecast model and more accurate estimates of the initial state, simultaneously there has been interest in obtaining probabilistic information derived from ensemble forecasts. A new COAMPS-TC ensemble system that is capable of providing probabilistic forecasts of TC track, intensity, and structure has been developed. This system makes use of the community-based Data Assimilation Research Testbed (DART) developed at the National Center for Atmospheric Research, which includes various options for Ensemble Kalman Filter (EnKF) data assimilation. The COAMPS-TC DART system constitutes a next generation data assimilation system for tropical cyclones that uses flow dependent statistics from the ensemble to assimilate observational information on the mesoscale.

A real-time COAMPS-TC ensemble system was run in a demonstration mode as part of the HFIP Stream 2 tests in 2011 for the W. Atlantic, E. Pacific, and W. Pacific regions. The system comprised an 80-member COAMPS-TC cycling data assimilation ensemble on three nested grids with horizontal spacing of 45, 15, and 5 km. For each new warning message issued by either NHC or JTWC, the COAMPS-TC ensemble was initialized by interpolating global forecast fields from the HFIP 80-member GFS-EnKF system to the three nested grids, which were centered on the storm. Six-hour forecasts were made four times daily to provide background estimates for the assimilation of observations from surface stations, ship data, radiosonde accents, cloud-track wind retrievals, and aircraft data with the DART EnKF. In addition to these conventional data sets, the NHC and JTWC TC position estimates were directly assimilated with the EnKF system. Under-sampling issues in the data assimilation procedure associated with the finite ensemble size were controlled by limiting the spatial influence of observations with a static localization radius of 1000 km, as well as applying a spatially and temporally varying inflation factor to the prior ensemble perturbations. To most effectively use the high-resolution capability of COAMPS-TC, a two-way interactive data assimilation procedure was implemented. In this algorithm, the innovations were defined using the highest resolution nest that contained the observation. These innovations were used to update the COAMPS-TC fields on all three grids. Furthermore, observations contained outside of a nest were allowed to update the fields within the nest.

Ten-member forecasts were performed twice daily to five days using the same three nested grid configuration as the data assimilation ensemble. The first 10 members of the data assimilation ensemble were used to define the forecast ensemble. Lateral boundary conditions for the forecast ensemble were drawn from the GFS-EnKF ensemble forecast. Examples of probabilistic products for Hurricane Irene are shown in Figure 24 for both track (top panel) and intensity (bottom panel). This is a real time forecast initialized at 1200 UTC 23 August, which is four days prior to landfall. The probabilistic track product shows the TC position from the individual ensemble members every 24 h and ellipses that encompass 1/3 and 2/3 of the ensemble member forecast positions. Note that the observed landfall location of the eye was within the ensemble distribution, although the ensemble mean landfall was approximately 12 h later than observed. The probabilistic intensity product (lower panel) shows a considerable spread among the members, particularly beyond 84 h, just prior to landfall. These products can be extremely valuable to assess the uncertainty in both track and intensity forecasts, and NRL is currently developing these capabilities and products further.

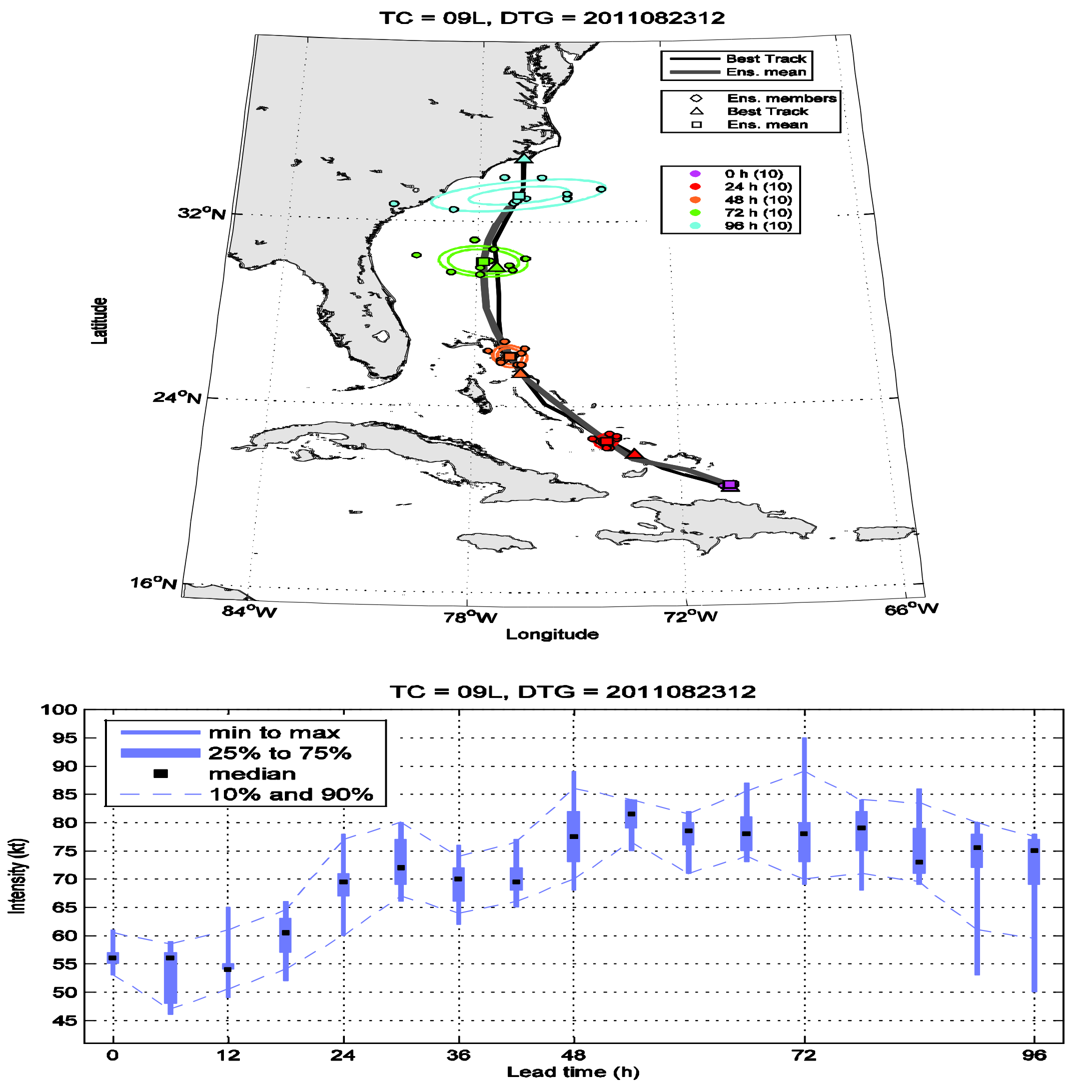


Figure 24. Probabilistic products from the COAMPS-TC ensemble for Hurricane Irene corresponding to the track (top panel) and hurricane intensity (bottom panel). This real time forecast was initialized at 1200 UTC 23 August, which is approximately four days prior to landfall. The probabilistic track product shows the TC position from the individual ensemble members every 24 h and ellipses that encompass 1/3 and 2/3 of the ensemble members. The intensity (knots) distribution is shown as a function of forecast lead time (hours) with the minimum value, maximum value and various quantiles of the ensemble distribution shown as denoted by the legend.

5.10. New Diagnostic Verification Tools

A variety of methods were used in the evaluation of forecast models this year, which provided useful information about the performance of a candidate model that goes beyond the mean or median performance as a function of lead. An example of the additional evaluation tools is shown in Figure 25. These plots show example rank histograms for difference performances. The top panel is an example when the candidate model is outperforming baseline used in comparison. The middle plot shows when the candidate model performs similar to the baseline (e.g. flat distribution). The bottom panel shows when the candidate model either performs better (rank 1) or worse (rank 4) compared to the baseline. This type of result examines the frequency of model performance and is useful for investigating how the candidate model is performing against the baseline.

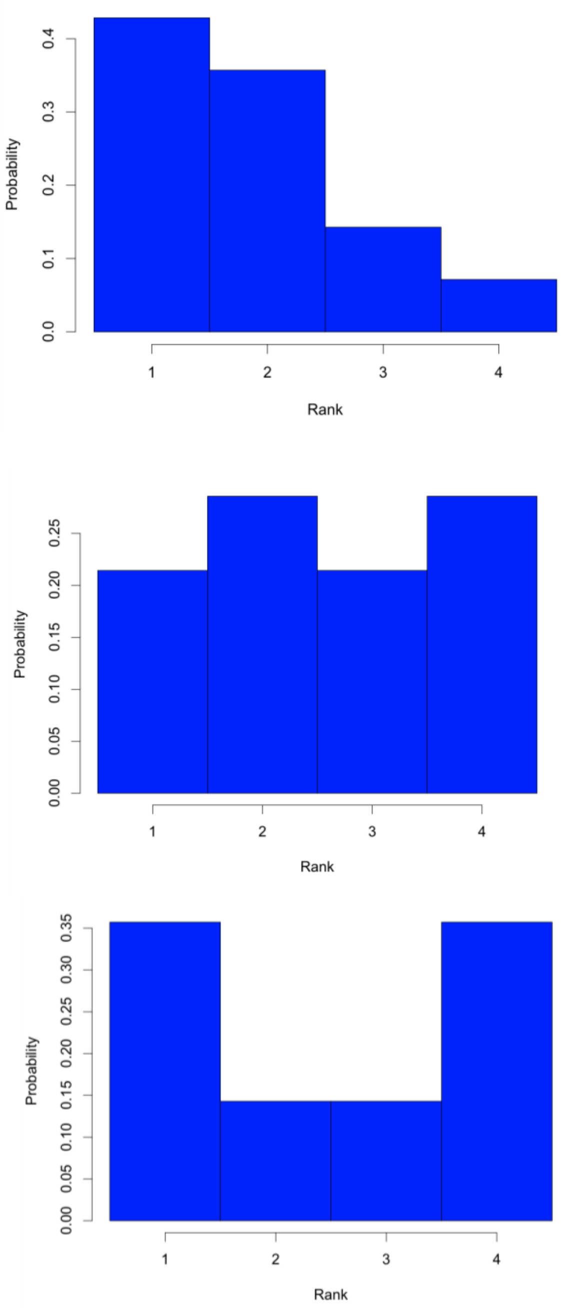


Figure 25: Rank histogram plots showing examples when a candidate mode is performing well (top), equally (middle), or often better or worse (bottom) than the baseline.

New methods are being developed to evaluate track and intensity errors of ensemble forecast systems. Figure 26 shows an example of a "wind rose" plot of intensity and track errors evaluated for the GFDL ensemble forecast system. The plot on the left shows the frequency of intensity errors as a function of track errors. The results show that the track errors tend to be east of the analyzed best track. Intensity errors tended to be positive with most errors < 39 knots. A few larger errors > 39 knots were observed for track errors to the west of the best track. The right hand plot in Figure 26 shows that most errors were < 200 n mi with a small percentage exceeding 200 n mi.

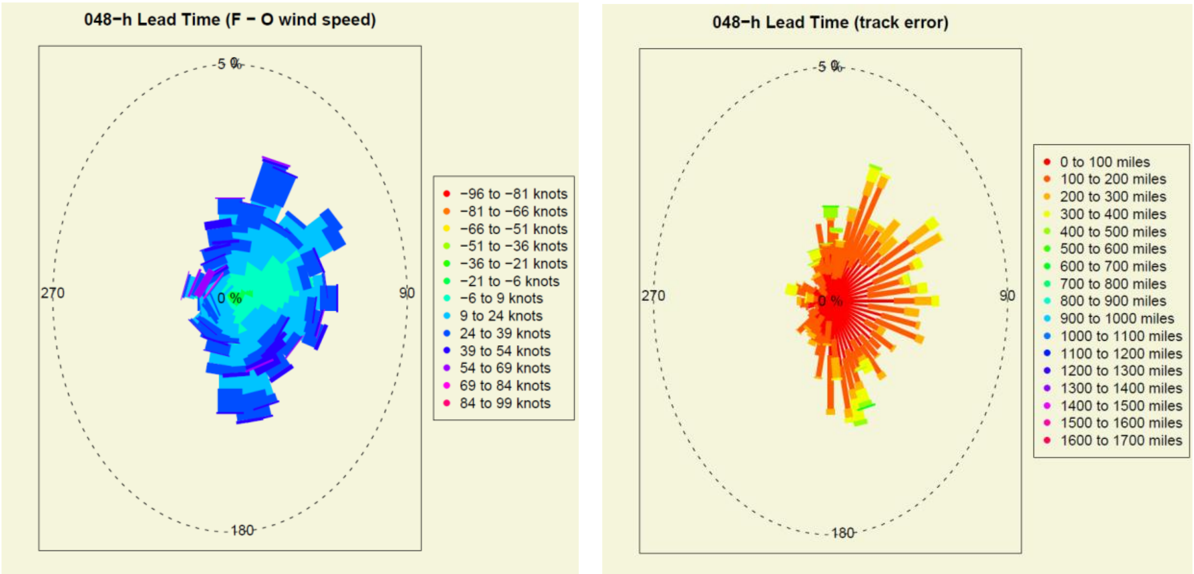


Figure 26: Example "wind rose" plots showing the frequency of intensity errors as function of track error location (left) and frequency of direction and magnitude of track errors (right).

6. Lessons Learned

The HFIP has gotten to the point were the main development foci have been identified and so the list below is mostly a repeat from earlier reports.

- Initialization of the regional models has improved but it still remains a major problem. It will need to be solved before we can make much progress on forecasting rapid intensification
- Physics packages suitable for models with resolutions of 3 km need to be improved

6.1. Initialization of the Regional Models Remains a Major problem

This is clearly shown in Figure 9 and has been pointed out many times in other annual reports. HFIP has made considerable progress by emphasizing ensemble data (or hybrid) data assimilation systems combined with the use of high resolution data near the hurricane center. NCEP is currently planning to implement hybrid data assimilation for their global model in the next year and development has started at EMC on a hybrid DA system for HWRF.

6.2. Physics Packages 3 km Models need to be improved

There has been little effort to find the optimum physics package appropriate for models with 3 km resolution. This is especially true for the hurricane models because of the high winds. Progress is being made (Figure 18), but there needs to be a careful assessment of the various options for convection, PBL and microphysics followed by a careful tuning of the entire package so the components work well together.

7. Challenges

- The biggest challenge to the results from HFIP will be for NOAA to identify the computer resources necessary to run the operational system that will be required to meet its goals
- The program needs a focus on developing appropriate physics packages for high resolution (3km) regional models
- Find ways to better use satellite data at resolutions appropriate for the hurricane

7.1. Computer Resources to run the HFIP Operational System

Our biggest concern in HFIP is that the computer resources will not be available within the National Weather Service to run the forecast models and ensembles at the resolution necessary to meet the HFIP goals. Current projections of the computing power available through the next several years are well below what will be required to meet the HFIP goals. This is an issue not just for HFIP but for US operational forecasting in general. One possible solution might be to turn the u-jet/t-jet computers into an operational machine and use it to focus on a few critical models on which much of the rest of the forecast system relies, in particular the high resolution global ensembles. This could probably be accomplished with a relatively low additional cost.

7.2. Physics Packages for High Resolution (3km) Regional Models

In past years HFIP has noted that the biggest challenges facing the program is initialization of both the regional and global models. Ensemble techniques have clearly improved the initialization of the global models (Figures 4 and 5) and NCEP has set as a goal to introduce the hybrid data assimilation (which uses an ensemble approach as well as a 3DVAR approach) within the next year. A similar effort with regional models has shown similar promise and plans are underway to test hybrid data assimilation in the operational regional models within a year. Even using the more traditional methods for regional model initialization, significant progress has also been demonstrated for the operational HWRF (Figure 16) and COAMPS-TC (Figure 7).

Since HFIP has been focused on the initialization problem it has put less emphasis on physical parameterizations. It is widely felt that along with initialization, a most promising area for improving model guidance will be improving the physical parameterization packages. Indeed, it is felt that better physics in combination with improved data assimilation techniques is a key part of improving model initialization in addition to the subsequent forecast. In particular a physics package appropriate for 3km models including surface flux, PBL, microphysics and convection representations need to be advanced for models at high resolution.

HFIP sponsored a Regional Modeling Physics Workshop in 2011 where the research and operational scientists came together to discuss the main deficiencies and routes for improvement. Follow-up discussions led to the decision to focus in three main areas: planetary boundary layer, microphysics, and convective parameterization. Work is underway at the DTC to expand the HWRF capability to be connected with a variety of physics packages and to assess the HWRF performance with a variety of cumulus schemes.

7.3. Better use satellite data

We noted above that high resolution data taken near the center of the hurricane by aircraft and incorporated into model initialization with an appropriate data assimilation system will greatly reduce intensity forecast errors out to several days. A problem with this result is that aircraft are only in the storm for perhaps 10% of all times models are initialized (but the majority when the storm is near land). We have some very preliminary results that suggest using satellite data (especially high resolution cloud track winds) provide a similar improvement. This result needs further development including consideration of other high resolution satellite data near the hurricane.

8. Final HFIP Numerical Model Hurricane Forecast System

In the discussion above we showed evidence that it is likely that HFIP will meet or exceed its 5 year intensity and track goals. It is less clear whether we can or cannot make the RI goals.

While it appears the use of aircraft data will likely help HFIP meet its intensity goals for storms for which such data is available, it will not be available for storms for a large majority of model initializations. For those we will need to rely on better use of satellite data taken in the near vicinity of the hurricane. Another major focus for HFIP is on improving satellite data assimilation in regional model initialization systems.

Except for the RI issue, we can now say with considerable confidence what a final end state operational configuration of the hurricane numerical prediction system should look like in 2014, the end of the initial 5 years of HFIP.

The longer range predictions, out to one week, of both track and intensity will be accomplished by global models run as an ensemble and initialized with a Hybrid data assimilation system and post processed with various statistical models. Resolution of these global models needs to be at least 20 km and the results will be improved if more than one global model is used in the ensemble.

The earlier forecast periods—out to 48-72 hours will be accomplished with regional models run with at least 3 km resolution as a multi- model ensemble. All models will use all available aircraft and satellite data. These will also be post processed with statistical models. The focus with the regional models will be on intensity and with the high resolution there is confidence in the community that the RI goals can be met with the regional models. More specifically the end system might include:

Global model ensemble with Hybrid Data Assimilation
20 members at 20 km
Multi Model (at least two—e.g.: (FIM, GSF)

Regional model ensemble
20 members at 3 km
Multi model (at least two—e.g.: (HWRF, AHW, TC-COAMPS)
Using all available aircraft and satellite data in core and near environment of hurricane

Statistical Post processing
Bias correction, LGEM, SHIPS

The ability to run this system will however require at least a 10-30 times increase in computer resources in operations with a major emphasis on the ability to run the high resolution ensembles.

9. Concluding Remarks

By reaching out across the hurricane research, development and operational communities, HFIP has promoted the cooperative effort necessary to make rapid improvements in hurricane forecast numerical guidance. The focus on improving the data assimilation system in the global models and the use of ensembles from the global models is likely to lead to substantial improvements in hurricane track forecasts in operations in the near future. Use of high resolution regional models with advanced data assimilation systems such as EnKF or the hybrid system, and the use of airborne radar data and satellite data at the scales of the hurricane will likely lead to improved forecasts of intensity. An initial GSI-hybrid capability for HWRF is being developed by ESRL/EMC as an extension of the system used for the GFS. The DTC is working with ESRL, EMC, and various research partners to create a system that can be used by and built on to receive developments to the community. In intensity and track, HFIP expects to easily reach its five year goals of improving both by 20% over the HFIP baseline.

Appendix

Table A.1: 2011 Stream 1.5 Evaluations: Summary of NHC Decisions

Technique Name	Model info	Contacts	Parameter	Application at NHC	Comments
AHWI	NCAR/MMM – SUNY 4-km AHW. Early version of AHW4.	Chris Davis Wei Wang Sherrie Fredrick Ryan Torn	Track	- Display explicitly in ATCF - Include in TV15*	3-4% improvement when added to Atlantic consensus w/EMXI from 24-96 h, and no degradations at other times. Track errors comparable to top-flight models. Not tested in east Pacific.
AHWI	NCAR/MMM – SUNY 4-km AHW. Early version of AHW4; GHMI interpolator.	Chris Davis Wei Wang Sherrie Fredrick Ryan Torn	Intensity	- Display explicitly in ATCF - Include in IV15	3-9 % improvement when added to operational Atlantic consensus from 24-96 h w/o serious degradations at other times.
COTI	NRL COAMPS-TC. Early version of COTC.	Jim Doyle Hao Jin	Track	None	Errors larger than top-flight models. Improvements when added to consensus w/EMXI were > 3% only at 48-72 h in the Atlantic, and it degraded the east Pacific consensus w/EMXI.
COTI	NRL COAMPS-TC. Early version of COTC; GHMI interpolator.	Jim Doyle Hao Jin	Intensity	- Include in IV15	Improved intensity consensus by 3-11% in both basins at most forecast times.
A1PI	PSU 1 km; TDR assimilated. Early version of A1PS	Fuqing Zhang Yonghui Weng	Track/Intensity	None	No systematic improvement over A4PI
A4PI	PSU 4.5 km; TDR assimilated. Early version of A4PS	Fuqing Zhang Yonghui Weng	Track	None	Improved Atlantic consensus that did not have EMXI by 3-9%, but had virtually no impact on the consensus w/EMXI (the latter restricted to 2008-10 only). Very small sample of forecasts.

A4PI	PSU 4.5 km; TDR assimilated Early version of A4PS; GHMI interpolator.	Fuqing Zhang Yonghui Weng	Intensity	- Include in IV15 when radar data assimilated	Small sample. Large SS improvements over consensus for land/water cases, but limiting to water almost completely eliminated the positive impact. For 2008-10 cases only, a sample that might be more representative of the expected 2011 performance, about 9% improvement to operational fixed consensus for all cases, 8% improvement for water only cases.
A4NI	PSU 4.5 km; No TDR. Early version of A4NR	Fuqing Zhang Yonghui Weng	Track	None	Inadequate sample for evaluation.
A4NI	PSU 4.5 km; No TDR. Early version of A4NR	Fuqing Zhang Yonghui Weng	Intensity	None	Inadequate sample for evaluation. Mostly positive impact on consensus for all cases but only 48 h SS. Mixed impact for water only cases. For 2008-10 cases only, 9% improvement to operational fixed consensus for both water/land and water only.
FIMI	GSD FIM; Early version of FIMY	Mike Fiorino	Track	- Display explicitly - Include in TV15	Lower errors than GFSI in both basins. Improved Atlantic track consensus by 3-7% from 24-120 h. Improved east Pacific consensus by up to 4% through 108 h.
FIMI	GSD FIM; Early version of FIMY	Mike Fiorino	Intensity	None	Did not perform as well as GFDL; did not improve consensus.
UWNI	U. Wisconsin UW-NMS 8 km Early version of UWN8	Will Lewis Greg Tripoli	Track	None	Track errors larger than top-flight models. Improvements to the Atlantic variable track consensus w/o EMXI were only about 1-2% and mostly not SS. For NHC fixed consensus w/EMXI improvements were about 2-3%.

UWNI	U. Wisconsin UW-NMS 8 km Early version of UWN8; GHMI interpolator.	Will Lewis Greg Tripoli	Intensity	- Include in IV15	Erratic behavior wrt/top-flight models. SS improvements to intensity consensus.
SPC3	CSU-CIRA SPICE 6-member statistical consensus	Mark Demaria Kate Musgrave	Intensity	- Display explicitly	SS improvements of up to 25% relative to DSHP, up to 12% over LGEM for the Atlantic. Improvements in eastern Pacific were much smaller but almost always positive.
H3GI	NCEP/EMC-AOML/HRD; HWRF 3 km Early version of H3GP	Vijay Tallapragada Sam Trahan Thiago Quirino	Track	- Display explicitly	Improvements of 3-12% over operational HWRF in the Atlantic, making its errors comparable to the top-flight models. Mixed impact in east Pacific. Replacing HWFI improved the variable consensus w/o EMXI. Slight (2%) improvements when added to fixed consensus w/EMXI, although made negligible contribution when other Stream 1.5 candidates were also in the consensus.
H3GI	NCEP/EMC-AOML/HRD; HWRF 3 km Early version of H3GP; GHMI interpolator.	Vijay Tallapragada Sam Trahan Thiago Quirino	Intensity	None	Even with the improvements over the operational HWRF, still lags the better operational intensity guidance. No systematic positive impact on consensus: when added to DSHP/LGEM/GHMI consensus, slight improvements in Atlantic and slight degradations in east Pacific. When the more promising Stream 1.5 candidates are also included in the consensus, H3GI had negligible impact.
GPMI	GFDL Ensemble mean Early version of GPMN	Tim Marchok Morris Bender	Track	- Display explicitly	Better than GFDL control by 2-4%. Negligible impact on consensus.

GPMI	GFDL Ensemble mean Early model version of ; GHMI interpolator.	Tim Marchok Morris Bender	Intensity	- Display explicitly	Better than GFDL control by up to 6% through 72 h. Slight degradation at 96 h. No systematic impact on consensus.
G01I	GFDL Ensemble member GP01 Early model version	Tim Marchok Morris Bender	Track	- Display explicitly	Up to 10% better than the control GFDL (except 4% worse at 72 h). Up to 3% improvement in consensus in place of control.
G01I	GFDL Ensemble member GP01 Early model version; GHMI interpolator.	Tim Marchok Morris Bender	Intensity	None	Worse than control.
G01I-G16I	GFDL Ensemble members GP01-GP16 Early model versions	Tim Marchok Morris Bender	Track	- Display spread	Spread and characteristics of regional model ensemble provides subjective information for forecasters.
G01I-G16I	GFDL Ensemble members GP01-GP16 Early model versions; GHMI interpolator.	Tim Marchok Morris Bender	Intensity	- Display spread	Spread and characteristics of regional model ensemble provides subjective information for forecasters.

10. References

Aksoy A., S. Lorsolo, T. Vukicevic, K.Sellwood, S. Aberson, F. Zhang, 2012: The HWRF Hurricane Ensemble Data Assimilation System (HEDAS) for high-resolution data: The impact of airborne Doppler radar observations in an OSSE. Submitted to Monthly Weather Review

Benjamin, S., G. Grell, J. Brown, T. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the ruc hybrid isentropic-terrain-following coordinate model. *Mon. Wea. Rev.*, **132**, 473–494.
Aksoy A., S. Lorsolo, T. Vukicevic, K.Sellwood, S. Aberson, F. Zhang, 2012: The HWRF Hurricane Ensemble Data Assimilation System (HEDAS) for high-resolution data: The impact of airborne Doppler radar observations in an OSSE. Submitted to Monthly Weather Review

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y.Z.33, Mozheng Wei, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems, Monthly Weather Review, Volume 133, Issue 5 (May 2005) pp. 1076-1097

Buizza, R., 1997: Potential Forecast Skill of Ensemble Prediction and Spread and Skill Distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 99–119.

Chang, E. K. M., 1993: Downstream development of baroclinic waves as inferred by regression analysis. J. Atmos. Sci., 50, 2038–2053.

Cohn, S. E., and D. F. Parrish, 1991: The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Wea. Rev.*,**119,** 1757–1785

Daley, R. and E Barker, 2001: NAVDAS: Formulation and Diagnostics, *Mon. Wea. Rev.*, 129, 869-883

Doyle, J.D., Y. Jin, R. Hodur, S. Chen. H. Jin, J. Moskaitis, A. Reinecke, P. Black, J. Cummings, E. Hendricks, T. Holt, C. Liou, M. Peng, C. Reynolds, K. Sashegyi, J. Schmidt, S. Wang, 2012: Real time tropical cyclone prediction using COAMPS-TC. *Advances in Geophysics* (in press).

Emanuel.K., , C. DesAutels, C. Holloway, R. Korty, 2004: Environmental Control of Tropical Cyclone Intensity, Journal of the Atmospheric Sciences, Volume 61, Issue 7 (April 2004) pp. 843-858

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99** (C5), 10 143–10 162.

Hakim, G. J., 2003: Developing Wave Packets in the North Pacific Storm Track. Monthly Weather Review, Volume 131, Issue 11 (November 2003) pp. 2824-2837

Hamill, T.M., J. S. Whitaker, M. Fiorino, S. G. Benjamin, 2011: Global Ensemble Predictions of 2009's Tropical Cyclones Initialized with an Ensemble Kalman Filter. Monthly Weather Review Volume 139, pp. 668-688

Hoskins, B. J.,, T. Ambrizzi, 1993: Rossby Wave Propagation on a Realistic Longitudinally Varying Flow. Journal of the Atmospheric Sciences, Volume 50, Issue 12 (June 1993) pp. 1661-1671

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

Keppenne, C.L., 2000: Data Assimilation into a Primitive-Equation Model with a Parallel Ensemble Kalman Filter. Monthly Weather Review Volume 128, pp. 1971-1981

Langland, R., M. Shapiro, and R. Gelaro, 2002: Initial condition sensitivity and error growth in forecasts of the 25 January 2000 East Coast snowstorm. *Mon. Wea. Rev.,* **130,** 957–974.

Palmer, T. N., R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular vectors, metrics, and adaptive observations. *J. Atmos. Sci.,* **55,** 633–653.

Parrish, D.F. and J. C. Derber, 1992: The National Meteorological Center's Spectral Statistical-Interpolation Analysis System. Monthly Weather Review, 120, 1747-1763.

Purser, J., Wan-Shu Wu, David F. Parrish, and Nigel M. Roberts, 2003: Numerical Aspects of the Application of Recursive Filters to Variational Statistical Analysis. Part I: Spatially Homogeneous and Isotropic Gaussian Covariances. Monthly Weather Review, 131, 1524-1535.

Purser, J., W.-S. Wu, D. F. Parrish, and N. M. Roberts, 2003: Numerical Aspects of the Application of Recursive Filters to Variational Statistical Analysis. Part II: Spatially Inhomogeneous and Anisotropic General Covariances. Monthly Weather Review, 131, 1536-1548.

Rabier, F., E. Klinker, P. Courtier, and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Quart. J. Roy. Meteor. Soc.*, **122,** 121–150.

Reynolds, C. A., M. S. Peng, J.-H. Chen, 2009: Recurving Tropical Cyclones: Singular Vector Sensitivity and Downstream Impacts. Monthly Weather Review
Volume 137, Issue 4 pp. 1320-1337

Snyder, C., and F. Zhang, 2003: Assimilation of Simulated Doppler Radar Observations with an Ensemble Kalman Filter. Monthly Weather Review, Volume 131, pp. 1663-1677.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, J. S. Whitaker, 2003: Ensemble Square Root Filters. *Monthly Weather Review* **131**:7, 1485-1490

Weng, Y. and F. Zhang, 2011: Assimilating airborne Doppler radar observations with an ensemble Kalman filter for cloud-resolving hurricane initialization and prediction: Katrina (2005). *Mon. Wea. Rev.*, in press.

Wu, W-S., J. Purser, and D. F. Parrish, 2002: Three-Dimensional Variational Analysis with Spatially Inhomogeneous Covariances. Monthly Weather Review, 130, 2905-2916.
.
Zhang, F., Y. Weng, J. Gamache and F. Marks, 2011: Performance of convection-permitting hurricane initialization and prediction during 2008-2010 with ensemble data assimilation of inner-core airborne Doppler radar observations. *Geophys. Res. Lett.*, **38**, L15810, doi:10.1029/2011GL048469