Test Plan for NOAA/OAR/WPO-funded Testbed Projects

Forecaster Support Products for Analysis of Tropical Cyclone Intensity and Structure from Aircraft Reconnaissance Observations

Test Plan

I. What major **concepts/techniques** will be tested? What is the scope of testing (what will be tested, what won't be tested)?

Five observational processing and analysis techniques will be integrated during this project, resulting in forecaster support products. These include the following.

- Spline Analysis at Mesoscale Utilizing Radar and Aircraft Instrumentation (SAMURAI), a 3-d variational data assimilation method which has been used extensively in field campaigns to develop research-grade analyses and has recently become optimized to run in real-time (Bell et al. 2012, Boehm and Bell 2021)/
- Enhanced Vortex Data Message Dataset (VDM+): which processes Vortex Data Messages into a structured machine-readable format with over 100 parameters (Vigh et al. 2015; Vigh 2015).
- A wind center-finding technique based on the Willoughby-Chelmow method (Willoughby and Chelmow 1982).
- Extended Flight Level Dataset for Tropical Cyclones Dataset (FLIGHT+): a research-grade dataset which includes earth-relative aircraft observations processed into storm-relative radial legs at the highest temporal resolution available (e.g., 30-s, 10-s, or 1-s) using automatic leg parsing and quality control, then interpolates the data to a standardized radial grid (Vigh et al. 2021).
- TC-OBS (Vigh et al. 2016, 2018), which uses objective criteria-weighted analysis algorithms to analyze input observations to provide detailed temporal analysis of TC intensity, wind structure, and size, as well as the time- and observations-dependent uncertainty bounds for track, intensity, RMW, and wind radii; and

The concepts and techniques associated with these capabilities have already been widely used in research settings, and are, in some cases, quite mature (e.g., SAMURAI, FLIGHT+). It is tempting to think that these would not need to be tested, however, since this project will make substantial improvements to some of the techniques and components, it will be important to conduct testing on both the changed components and the overall analysis outputs of the integrated system to ensure that the quality of outputs remains high. Testing will also be needed to ensure that the codes remain robust to ensure reliability of outputs in operational settings. Most importantly, testing and evaluation of the forecaster support products that result from the integrated system will be essential to ensure that the project's outputs meet NHC's needs and this project's goals.

II. **How** will they be tested? What **tasks** (processes and procedures) and activities will be performed, what preparatory work has to happen to make it ready for NOAA testing, and what will occur during the experimental testing in the testbed?

A hierarchy of testing is envisioned, which involves testing various code components where necessary, testing outputs of major techniques, and testing the outputs of the integrated system. Where appropriate, key specific code components that undergo changes can be tested via unit tests and regression tests. Regression testing can also be performed for the major outputs of each technique.

The analysis products of this project will be evaluated for the attributes of accuracy, improved representation of rapid variations, observed improvements in models, rated usefulness by forecasters, evidence of use by forecasters, and resulting time-savings to forecasters. The full evaluation protocols will be established in the first year of the project through close coordination with NHC, HRD and HOT. There are several options for how this could be done.

Option A: Parallel testing in operations

This evaluation protocol would involve parallel testing in operations, with and without the products. One forecaster would use the project's analysis products to undertake their analysis of the TC and subsequent forecast, while the other forecaster would use NHC's current operational practices. Each forecaster would also track how long it took them to analyze the aircraft data during their forecast period. After the season concludes, the analyzed estimates on VMAX, RMW, and wind radii could then be compared to each other and to the post-season best track.

Option B: Randomized serial testing in operations

If parallel testing would require too much effort, another option would be to randomly assign forecaster shifts to use the tools, while the remaining shifts would not use the tools. While a direct comparison of impact of the tools on the human forecast process would not be possible as it would be in parallel operations, the cumulative impacts should become apparent over a large number of forecast cycles. This method would still gather robust information for the metrics of time savings, usability, and forecaster's subjective value of the tools.

Option C: Parallel testing in operational forecast aids

If NHC opts not to do a parallel forecast process test protocols, then an alternate approach would be to simply compare the forecasts of one or more forecast aids (e.g., statistical-dynamical guidance like SHIPS) initialized directly from the analyses of this project to versions initialized from the operational analysis. Improvements can then be measured through direct comparison using verification tools like METplus.

If resources are limited, a combination of approaches B and C may be optimal. Depending on which direction HOT chooses to go, forecast protocols will be set up and specific written instructions will be provided to forecasters outlining the procedures that they should follow.

The usefulness of the products will be obtained through a survey of forecasters and discussions with them. It would be helpful if the Weather Program Office can provide expertise from the Social Science Program to ensure that the survey instrument uses best practices.

The evidence of use will be obtained by observing whether forecasters mention the products in their forecast discussions. Finally, the impact to forecaster workload could be determined by analyzing the time forecasters spent interpreting reconnaissance data to see if there was a substantial time savings.

III. **When** will it be tested in coordination with the NOAA testbed? What are **schedules and milestones** for all tasks described in section II that need to occur leading up to testing, during testing, and after testing?

The project will test the integrated system's capabilities in coordination with the NOAA testbed in Year 2 (2024 hurricane season). Testing will occur in a functionally-similar operational environment in Year 3 (2025 hurricane season). Our complete list of milestones is as follows.

Milestone/Checkpoint (Goal completion date)	**Transition Activities (if any)**	**Date (MM/YY)**
9A: First NHC visit to gather requirements and establish evaluation protocols (Oct 2022)		
1A: FLIGHT+/VDM+ modified to run in real-time using operational data (Oct 2022)		
3A: Complete organization of post-processed data (Nov 2022)		

2A: Willoughby-Chelmow technique running in real-time using operational data (Dec 2022)		
3B: Complete preliminary version of surface wind reduction parameterization (May 2023)		
5A: TC-OBS implemented in real-time to provide temporal analysis of VMAX / RMW / size (May 2023)		
4A: SAMURAI implemented in real-time to provide 3-d and 2-d analyses (Jun 2023)		
7A: Outputs available to NHC forecasters via TCGP (Jun 2023)		
7B: First demo period begins (Jul 2023)		
9B: Second NHC visit to gather feedback and observe and train forecasters (Aug 2023)	Forecaster training	
7C: First demo period ends (Nov 2023)		
8A: Evaluation completed for first demo period with report to NHC (Feb 2024)		
3C: Single unified parameterization for the vertical profile of wind completed (Mar 2024)		
2B: Implement improved center-finding for weak storms (Apr 2024)		
5B: TC-OBS optimal parameter estimation techniques improved (Apr 2024)		
4B: SAMURAI updated to use unified parameterization (May 2024)		
7D: Outputs provided in formats for testing in AWIPS-2 and the NHC HFIP Display (May 2024)		
7E: Second demo period begins (Jun 2024)	Capabilities running in NHC sandbox	

7F: Second demo period ends (Nov 2024)		
2C: Center-finding improved to handle simultaneous recon flights (Jan 2025)		
6A: Uncertainty quantified in TDR / SFMR / Dropsondes (Jan 2025)		
8B: Evaluation completed for second demo period with report to NHC (Feb 2025)		
1B: Update FLIGHT+ for simultaneous recon flights (Feb 2025)		
5C: TC-OBS updated with quantified uncertainty and multiple flight capability (Feb 2025)		
9C: Complete automation to run on operational systems (Mar 2025)	Testing in a functionally-similar operational environment	
8C: Two manuscripts submitted for publication (Mar 2025)		
9D: Documentation and training materials completed (May 2025)	Documentation made available through online or internal resources	
7G: Demonstration begins of all capabilities at NHC on testbed resources (Jun 2025)	Demonstration in the operational environment	
9E: Third NHC visit to train forecasters on use of tools (Jun 2025)	In-person training, evaluation, and feedback	
9F: Codes available for operational transition (Jul 2025)	Codes begin running in operational environment (subject to acceptance)	
8D: Final report submitted (Aug 2025)		

IV. **Where** will it be tested? Will it be done at the PI location or at a NOAA testbed location?

At the end of the first year (2023 hurricane season), the capabilities will be demonstrated at the PI locations (NCAR and CSU).

At the end of the second year (2024 hurricane season), the full system will be tested at the HOT Testbed, possibly on the HOT Sandbox machine at NHC.

At the end of the third year (2025 hurricane season), the full system will be tested on resources to be determined by NWS staff.

V. Who are the key **stakeholders** involved in testbed testing (PIs, testbed support staff, testbed manager, forecasters, etc.)? Briefly what are their **roles and responsibilities**?

Stakeholders	**Roles/Responsibilities**	**Team Members**
PIs and project research staff	Principal Investigators and researchers at NCAR, CSU, and UM will coordinate with the HOT facilitator as directed by NHC for demonstration and evaluation of the analysis products. The PIs and project research staff will undertake testing at the code and component level, implementing unit tests and regression tests where appropriate. The PIs and project research staff will work closely with HOT and NHC forecasters on devising and implementing the forecaster evaluation of the system.	Jonathan Vigh (NCAR) Michael Bell (CSU) Jun Zhang (UM) Eric Hendricks (NCAR) Christopher Rozoff (NCAR) Jennifer DeHart (CSU) Alex DesRosiers (CSU)
Testbed Manager	The HOT Testbed Manager will guide the project team to ensure that testing and evaluation conforms to best practices and procedures of the HOT. The Testbed Manager	Wallace Hogsett (NHC SOO)

	will also facilitate access to NHC forecasters and coordinate with the Testbed Support Staff on the internal testing protocols to be conducted at NHC.	
Testbed Support Staff	The HOT support staff will assist in facilitating and implementing the evaluation protocols at NHC. They will work with the Hurricane Specialist Unit to guide the process of integrating the analysis capabilities into forecaster workflows during the evaluation period.	Stephanie Stevenson (NHC TSB) Alan Brammer (HOT Facilitator)
Forecasters	The forecasters will undertake the evaluation protocol described above, with the guidance and support of the HOT staff.	John Cangialosi (NHC HSU Senior Forecaster)

VI. What **testing resources** will be needed from each of the above participants (hardware, software, data flow, internet connectivity, office space, video teleconferencing, etc.), and who will provide them?

In Year 1, the PIs and research project staff will undertake testing using resources at their institutions.

In Year 2, HOT Testbed Support Staff will need to provide the following for successful testing to occur:

- A Linux workstation capable of running the integrated system:
    - multi-core processors (8, 12, or 20 cores),
    - sufficient memory (128 GB),
    - internet connectivity (>50 MBps), and
    - sufficient disk storage (1 TB).

We do not anticipate needing dedicated office space or video teleconferencing at this time, however these could be beneficial if available.

In Year 3, the final integrated system will be tested in an operationally-similar environment on resources determined by NWS staff. These resources will need to be provided by NWS.

VII. What are the **test goals, performance measures, and success criteria** that will need to be achieved at the end of testing to measure and demonstrate success to advance to higher Readiness Levels (RL) and proceed to RL 8 (ready for transitioning to NOAA operations)?

Our **test goals** will be as follows:

- Unit tests:
  - Each portion of code with unit coverage will continue to provide the expected answer after code changes occur.
- Regression tests:
  - Each major component will continue to provide the same high quality output and accuracy after code changes occur without a degradation of robustness.

Our **performance measures** are as follows:

- Techniques or major components which have undergone changes to improve the outputs need to demonstrate evidence of the improvement in a systematic and rational manner across a test set that includes a large number of cases that cover a wide variety of storm and infrastructure circumstances. Testing over an entire season would be ideal.
- The forecaster support products should demonstrate accuracy and representativeness.
  - One data point must not be given too much weight, yet, a high quality observation should be given due weight.
- The forecaster support products should demonstrate improved representation of rapid variations.
  - Vortex-scale changes should be captured with high fidelity. Fluctuations due to small-scale features can be noted, but should not sway the overall analysis. Finding the right balance on this matter will require coordination between project staff, HOT, and the NHC forecasters.
- Improvements are demonstrated in models which use the analysis products as inputs.
  - This can be measured in a straightforward fashion using standard verification tools and metrics.
- Forecasters rate the forecaster support products as useful.
- There is tangible evidence that forecasters are using the forecaster support products in their forecast process.
- The forecaster support products result in time-savings to forecasters.

Our **success criteria** for the project outputs will be[1]:

[1] Specific numerical goals for 1) and 2) will be established in coordination with, and subject to refinement, by NHC and JHT.

1. Operational estimates for VMAX, RMW,[2] and wind radii using this project's capabilities provide a reduction in error[3] over operational estimates made without these capabilities (using the final postseason best track as verification "truth"; Landsea and Franklin 2013; Cangialosi and Landsea 2016).
2. Using the observations as "truth", the analysis products possess lower mean absolute errors than the postseason best track.
3. The 24-h forecasts of a statistical-dynamical intensity forecast aid initialized with the analyses is improved (Cangialosi et al. 2020; DeMaria et al. 2005).
4. A majority of forecasters indicate in their survey responses a reduced workload for a shift when using this project's capabilities.
5. A majority of forecasters indicate that the analysis capabilities are "useful" or "very useful" in their survey responses (on a five-point Lickert scale with the middle option being "somewhat useful").
6. The new analysis capabilities are mentioned in forecast discussions in the third year of the project.
7. NWS decides to operationally implement the project's capabilities.

VIII. How will testing **results** be documented? Describe what information will be included in the **test results final report**.

The results of unit testing will be noted in GitHub issues tracking the development of features. The results of regression tests will be noted in GitHub issues when pull requests are made. These may also be captured in shared Google Docs which define the testing protocols for a specific component. The results of the testing for major component improvements will be documented in shared Google Docs.

Testing results on robustness (ability of the system to generate the desired analysis outputs) will be captured. Any specific times in which the system failed to generate the expected analysis outputs will be noted in a shared Google Doc or Sheet. As time permits, failure mode analysis will be undertaken to learn how the system can be improved.

All relevant and significant test results will be also captured in an annual testing and evaluation report. This report will summarize what testing was done over the annual period, what the results of the tests were with regard to improvements, and whether the improved component was accepted. Overall evaluation of the forecaster support products will be provided in this report according to the performance metrics and success criteria previously

[2] NHC began best-tracking the RMW in 2021 (personal communication, J. Cangialosi, 2021).

[3] Since the best track smooths out oscillations with a period of less than 24 h, it will be necessary to smooth our analyses to have similar temporal characteristics as the best track. This will be accomplished by applying a low-pass filter.

noted. HOT and NHC forecasters will have the chance to review this report and offer input. The testing report will be available to the public via the project web site.

References

Bell, M. M., M. T. Montgomery, and K. E. Emanuel, 2012: Air-sea enthalpy and momentum exchange at major hurricane wind speeds observed during CBLAST. *J. Atmos. Sci.*, **69**, 3197-3122. https://doi.org/10.1175/JAS-D-11-0276.1

Boehm, A. M., and M. M. Bell, 2021: Retrieved Thermodynamic Structure of Hurricane Rita (2005) from Airborne Multi-Doppler Radar Data. *J. Atmos. Sci.*, **78**, 1583-1605. https://doi.org/10.1175/JAS-D-20-0195.1

Cangialosi, J. P., and C. W. Landsea, 2016: An examination of model and official National Hurricane Center tropical cyclone size forecasts. *Wea. Forecast.* **31** (4), 1293-1300. https://doi.org/10.1175/WAF-D-15-0158.1

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecast.*, **35** (5), 1913-1922. https://doi.org/10.1175/WAF-D-20-0059.1

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). Wea. Forecast. **20** (4), 531-543. https://doi.org/10.1175/WAF862.1

Landsea, C. W., and J. L. Franklin, 2013: Atlantic Hurricane Database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141** (10), 3576-3592. https://doi.org/10.1175/MWR-D-12-00254.1

Vigh, J. L., 2015. VDM+: The Enhanced Vortex Data Message Dataset (Version 1.100). Tropical Cyclone Data Project, National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado. [Available online at: http://dx.doi.org/10.5065/D61Z42GH.]

Vigh, J. L., 2015: VDM+: The Enhanced Vortex Data Message Dataset (Version 1.100). Tropical Cyclone Data Project, National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado. [Available online at: http://dx.doi.org/10.5065/D61Z42GH.]

Vigh, J. L., E. Gilleland, C. L. Williams, D. R. Chavas, N. M. Dorst, J. Done, G. Holland, and B. G. Brown: 2016: A New Historical Database of Tropical Cyclone Position, Intensity, and Size Parameters Optimized for Wind Risk Modeling. Extended Abstract, 32nd Conf. on Hurricanes and Tropical Meteorology, San Juan, Puerto Rico, *Amer. Meteor. Soc.*, Paper 12C.2. https://doi:10.13140/RG.2.1.3720.5361

Vigh, J. L., E. Gilleland, C. L. Williams, D. R. Chavas, and N. M. Dorst, 2018: TC-OBS: The Tropical Cyclone Observations-Based Structure Database (version 0.42, an alpha-level release). Tropical Cyclone Data Project, National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado. [Available online at: https://doi.org/10.5065/D6BC3X95.]

Vigh, J. L., N. M. Dorst, C. L. Williams, E. W. Uhlhorn, B. W. Klotz, J. Martinez, H. E. Willoughby, F. D. Marks, Jr., D. R. Chavas, 2021: FLIGHT+: The Extended Flight Level Dataset for Tropical Cyclones (Version 1.3). Tropical Cyclone Data Project, National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado. [Available online at: https://doi.org/10.5065/D6WS8R93.]

Willoughby, H. E. and M. B. Chelmow, 1982: Objective Determination of Hurricane Tracks from Aircraft Observations. *Mon. Wea. Rev.*, **110**, 1298-1305. https://doi.org/10.1175/1520-0493(1982)110<1298:ODOHTF>2.0.CO;2