

User-Responsive Diagnostic Forecast Evaluation Approaches: Application to Tropical Cyclone Predictions

BARBARA G. BROWN¹,^a LOUISA B. NANCE,^a CHRISTOPHER L. WILLIAMS,^b KATHRYN M. NEWMAN,^a
JAMES L. FRANKLIN,^c EDWARD N. RAPPAPORT,^c PAUL A. KUCERA,^d AND ROBERT L. GALL^e

^a National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado

^b Department of Geography, University of Florida, Gainesville, Florida

^c NOAA/National Weather Service, National Hurricane Center, Miami, Florida

^d COMET, University Corporation for Atmospheric Research, Boulder, Colorado

^e NOAA/National Weather Service, Silver Spring, Maryland

(Manuscript received 21 April 2023, in final form 16 August 2023, accepted 21 August 2023)

ABSTRACT: The Hurricane Forecast Improvement Project (HFIP; renamed the “Hurricane Forecast Improvement Program” in 2017) was established by the U.S. National Oceanic and Atmospheric Administration (NOAA) in 2007 with a goal of improving tropical cyclone (TC) track and intensity predictions. A major focus of HFIP has been to increase the quality of guidance products for these parameters that are available to forecasters at the National Weather Service National Hurricane Center (NWS/NHC). One HFIP effort involved the demonstration of an operational decision process, named Stream 1.5, in which promising experimental versions of numerical weather prediction models were selected for TC forecast guidance. The selection occurred every year from 2010 to 2014 in the period preceding the hurricane season (defined as August–October), and was based on an extensive verification exercise of retrospective TC forecasts from candidate experimental models run over previous hurricane seasons. As part of this process, user-responsive verification questions were identified via discussions between NHC staff and forecast verification experts, with additional questions considered each year. A suite of statistically meaningful verification approaches consisting of traditional and innovative methods was developed to respond to these questions. Two examples of the application of the Stream 1.5 evaluations are presented, and the benefits of this approach are discussed. These benefits include the ability to provide information to forecasters and others that is relevant for their decision-making processes, via the selection of models that meet forecast quality standards and are meaningful for demonstration to forecasters in the subsequent hurricane season; clarification of user-responsive strengths and weaknesses of the selected models; and identification of paths to model improvement.

SIGNIFICANCE STATEMENT: The Hurricane Forecast Improvement Project (HFIP) tropical cyclone (TC) forecast evaluation effort led to innovations in TC predictions as well as new capabilities to provide more meaningful and comprehensive information about model performance to forecast users. Such an effort—to clearly specify the needs of forecasters and clarify how forecast improvements should be measured in a “user-oriented” framework—is rare. This project provides a template for one approach to achieving that goal.

KEYWORDS: Forecast verification/skill; Numerical weather prediction/forecasting; Operational forecasting; Model evaluation/performance; Decision support


1. Introduction

Every year, tropical cyclones (TCs) cause significant property damage and human impacts (e.g., death, injuries, loss of livelihoods) around the world (e.g., Pielke and Pielke 1997; Rappaport 2000; Pielke et al. 2008; Gall et al. 2013). To mitigate these impacts, weather prediction centers across the globe provide forecasts of TC movement (i.e., track) and intensity; information based on these forecasts is provided to

emergency managers and the public to aid in decision-making and actions related to lessening the impacts of TCs, such as evacuating homes and businesses, and protecting property from damage (e.g., Lazo and Waldman 2011; Bostrom et al. 2018). In response to the needs for better predictions of TC track and intensity (with a major focus on intensity) to aid in providing warnings, the U.S. National Weather Service (NWS) established the Hurricane Forecast Improvement Project (HFIP; now called the Hurricane Forecast Improvement “Program”; <https://hfip.org/about>) in 2007. The goal of HFIP has been to significantly improve predictions of TC’s track (location) and intensity (maximum wind speed at 10 m, averaged over 1 min; Landsea and Franklin 2013) in both the Atlantic and eastern North Pacific basins (Gall et al. 2013).

The NWS’s official TC track and intensity forecasts for the Atlantic and eastern North Pacific basins are produced by forecasters at the NWS’s National Hurricane Center (NHC) who use output from numerical weather prediction (NWP)

Franklin, Rappaport, and Gall: Retired.

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Barbara G. Brown, bgb@ucar.edu

DOI: 10.1175/WAF-D-23-0072.1

© 2023 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

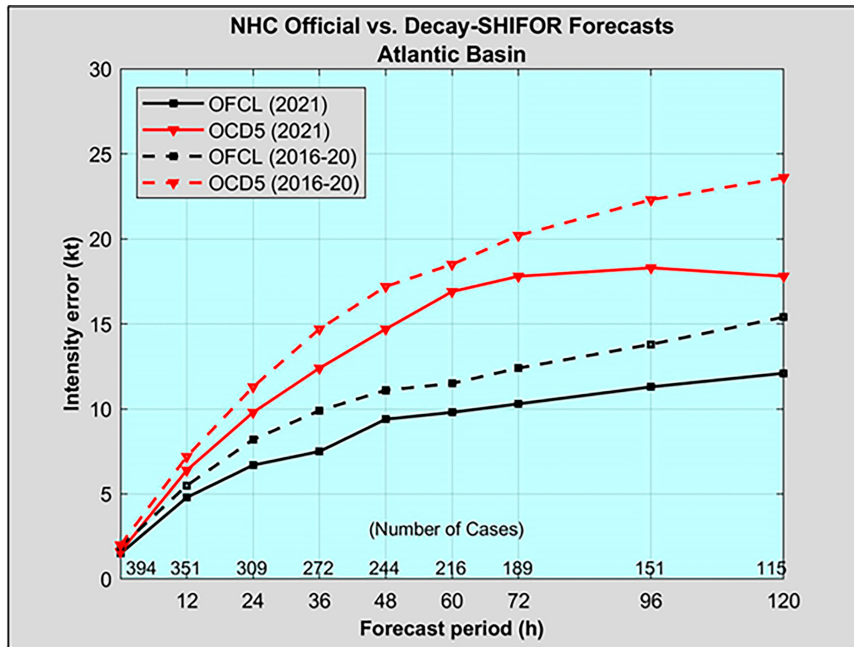


FIG. 1. Example operational evaluation of 2021 official (“OFCL”) TC intensity forecasts at NHC (Cangialosi 2022). The average performance of official forecasts is compared to OCD5 (a no-skill reference based on climatology and persistence) along with a comparison of the average performance of official forecasts and OCD5 for the previous 5 years.

and statistical models (as well as other information, including observations from aircraft, buoys, and other sources) as guidance for creating their forecasts. To achieve meaningful improvements in TC predictions, a major HFIP effort focused on improvement of TC guidance from NWP and statistical models. This effort engaged mesoscale and global NWP model developers at universities, research laboratories and government organizations to increase the skill of TC guidance.

To promote and monitor model improvements, HFIP established an annual intercomparison of models to select experimental model guidance to be made available to NHC forecasters during the subsequent hurricane season. Independent scientists and forecast verification experts at the National Center for Atmospheric Research (NCAR) were tasked with collecting retrospective forecasts from the various research and development groups and evaluating the performance of these experimental forecast systems relative to predictions from operational forecast models and guidance products.

TC forecasts have been subject to evaluation/verification for many decades (Powell and Abernson 2001; Franklin et al. 2003; Rappaport et al. 2009), and NHC performs its own annual evaluations of the forecasts produced by NHC forecasters as well as predictions produced by operational NWP and statistical models. Traditionally, these evaluations provide an overall representation of year-to-year performance of NHC’s forecasts as well as forecast guidance products (e.g., statistical, NWP), with a focus on basic verification statistics summarizing errors in predictions of track (total track, along-track, and cross-track errors) and intensity (e.g., Cangialosi 2022; Cangialosi et al. 2020; Franklin et al. 2003; Rappaport et al. 2009). While case

studies are also undertaken, the annual summary verification statistics often are presented in a bulk form (i.e., aggregated across all storms in a given year and ocean basin or aggregated by individual storm) using summary measures that ignore some important aspects of performance, such as the underlying variability associated with the statistics, which are computed using a finite sample (Wilks 2019, p. 470). For example, Cangialosi (2022) presents overall statistics for the operational NHC forecasts from 2021 in comparison to performance in previous years and relative to a no-skill climatology and persistence forecast (“OCD5”) that provides a representation of the difficulty of the forecast situation (Fig. 1).

While useful (e.g., for NHC forecasters) for monitoring year-to-year changes in performance, basic summary verification statistics provide limited diagnostic information that forecasters can apply to their interpretation and use of model guidance to improve their subjective forecasts. Obtaining this kind of information requires more in-depth analysis methods. The bulk statistics also are of limited use for applications such as those related to model development and meeting specific user needs. For example, forecasters may be interested in understanding a variety of attributes that go beyond average performance, such as the size and frequency of large errors, a performance aspect that cannot be measured using traditional verification approaches. Such basic statistics also do not provide meaningful information about variability in performance—information that can help forecasters and other users gain confidence in the forecasts and understand the frequency and circumstances with which the forecasts are especially good or poor. When faced with an unfamiliar model, such information

about the variability in performance would be especially important for developing understanding regarding how to optimally apply the model guidance. Finally, the number of storms included in TC verification studies can sometimes be relatively small, which limits the confidence that can be placed on the resulting verification statistics and their representation of overall performance (WMO 2013). The evaluations described in this study mitigate this issue through verification of TC predictions across multiple years, as NHC typically does when selecting a new operational model. Methods were also developed or selected to respond to other issues mentioned (e.g., measuring and understanding variability in performance, examining large errors and relative performance of several models).

Over the past 15 years, work by the World Meteorological Organization (WMO; e.g., WMO 2013) and various researchers (e.g., Marchok et al. 2007; G. M. Chen et al. 2013; Moskaitis 2008; Yu et al. 2013; Chen et al. 2018) have demonstrated the increased information about the performance of TC predictions that can be attained through the application of diagnostic verification methods. Diagnostic methods provide more in-depth information regarding forecast performance than can be obtained using traditional verification approaches that rely solely on summary measures (e.g., Murphy et al. 1989; Marchok et al. 2007; Moskaitis 2008; G. M. Chen et al. 2013; Y. E. Chen et al. 2013; WMO 2013; Ebert et al. 2015). The development of user-responsive approaches for forecast evaluation can also lead to improved decision-making by forecast users and can guide forecast developers toward increasing the usefulness of the forecasts (e.g., Morss et al. 2008; Ebert et al. 2018).

The user-responsive evaluations of experimental HFIP forecasts conducted by NCAR (in collaboration with HFIP, NHC, and researchers at multiple universities and research and operational centers) endeavored to focus on specific questions of interest to forecasters and managers, who were interested in understanding particular behaviors of the forecasts provided by each candidate forecasting system, relative to a set of baseline prediction systems. The user-responsive questions of interest were defined in collaboration with these individuals and groups, and the statistics generated through the evaluations were specifically designed to answer these questions. The goal of this effort was to provide NHC with as much information as possible about the quality of the forecasts produced by the candidate forecasting systems so that meaningful choices could be made about their potential usefulness to NHC forecasters. Examples of questions that were posed and answered via the evaluations include “Was the overall performance of the experimental model at least as good as the performance of a baseline model?”; “What is the uncertainty associated with the verification results?”; and “How frequently did the experimental model perform better (or worse) than a baseline model?” Specific approaches were developed to measure and communicate the results of these evaluations.

Each of the five yearly evaluations during the period 2010–14 considered experimental forecasts over an extensive retrospective period (covering at least two and usually three hurricane seasons). This approach led to large sample sizes (which allowed robust estimation of the verification statistics) and made it possible to represent forecast performance across a variety of

meteorological and forecast circumstances. The large sample sizes also increased the stability of the metrics that were computed. In addition, the forecast samples were homogeneous; that is, all samples were generated by the same experimental model version, and model comparisons between experimental and individual baseline models utilized the same model initializations and lead times for each pair of models. The evaluations took place prior to the beginning of each hurricane season using the output from updated or new modeling systems. They considered extreme behavior as well as the average behavior traditionally examined in standard evaluations, by making note of outliers; outlier information can help forecasters develop confidence in new forecasting systems.

This paper describes the user-responsive verification methods for TC forecast evaluation that were developed through close collaborations between the NCAR team and those making the decision as to which experimental models would be made available to NHC forecasters during the upcoming hurricane season. Section 2 provides more details regarding the HFIP model evaluation and demonstration project and gives an overview of general aspects of the yearly evaluation approach for TC prediction models. A hierarchy of evaluation approaches, starting with the methods that have traditionally been used for TC forecast evaluation and moving to the advanced user-responsive approaches applied in the HFIP evaluations, is described and demonstrated in section 3. A discussion and some concluding remarks (including lessons learned) are presented in section 4.

2. HFIP background: The model demonstration and evaluation project

To meet the need for track and intensity forecasts with higher accuracy and reliability, HFIP has focused on improving the numerical guidance provided to NHC forecasters by NOAA’s operational modeling suite. In this context, “forecast model” refers to any objective tool used to generate a prediction of a future event. These investments have focused on both near-term development and testing directed at yearly upgrades to the operational NWP capabilities, for which HFIP adopted the term “Stream 1,” and longer-term efforts that take multiple years to enhance operations, for which HFIP adopted the term “Stream 2.” From 2010 to 2014, HFIP and NHC included an intermediate path to operations known as “Stream 1.5.” The Stream 1.5 models consisted of predictions from vetted experimental models and/or techniques to which NHC forecasters were provided real-time access during a particular hurricane season (Gall et al. 2013).

The driving force behind HFIP’s Stream 1.5 was to provide NHC forecasters the opportunity to have access to promising guidance for one or more hurricane seasons prior to what was possible without Stream 1.5. In particular, Stream 1.5 provided a temporary path that bypassed the budgetary and technical bottlenecks associated with traditional operational implementation via the use of nonoperational computing resources and a real-time data feed that was separate from the operational data feed (Gall et al. 2013).

Stream 1.5 prediction systems were run systematically as part of HFIP's annual "Demonstration Project" on nonoperational computing platforms. The resulting experimental guidance was provided to NHC in the form of "A-deck" files, which are text files containing forecasted storm properties (e.g., center location, maximum wind speed, storm size) that conform to the Automated Tropical Cyclone Forecast (ATCF) format specifications (Sampson and Schrader 2000) used in forecast operations. The introduction of experimental guidance into the operational environment without thorough vetting can negatively impact the forecast process by distracting and possibly wasting valuable time as forecasters work to synthesize their forecast products, and could possibly make their operational forecasts worse. Knowledge about the strengths and weaknesses of a particular forecast model plays an important role in the forecasting process and can help forecasters develop confidence in the model predictions. Hence, user-responsive evaluations of the models prior to operational demonstration were a critical component of the Stream 1.5 concept.

The Stream 1.5 qualification process for each upcoming hurricane season involved an extensive evaluation of each experimental forecast model. This evaluation focused on retrospective forecasts for a diverse set of tropical cyclones in the Atlantic and eastern North Pacific basins that were selected in December, with the cases spanning two to three hurricane seasons. The nominal schedule for the selection process included delivery of retrospective forecast samples by each of the development teams to the NCAR team for evaluation by mid-April, and completion of the evaluations by the NCAR team by the end of May. Following delivery of the evaluation reports, NHC selected systems to be included in Stream 1.5 with a goal of the Stream 1.5 data feed being ready by 1 August, traditionally considered the start of the "peak" of the hurricane season in the Atlantic basin.

The retrospective forecasts, delivered as A-deck files,¹ were generated by each modeling group by applying their own method of storm tracking to their model output. The model development groups were required to submit a homogeneous sample of retrospective forecasts in the form that would be used in real-time if their model was selected. To be consistent with NHC's procedures, the experimental models that were considered to be "late guidance"² were converted to early model guidance by applying an interpolator package with the same functionality as the software used by NHC (Cangialosi 2022). That is, the evaluation process was configured to capture all aspects of the real-time data processing that would occur during the hurricane season.

Over the years, the experimental candidates included deterministic global and regional models, as well as multimodel

ensembles, ensembles based on perturbations to a single-model, and experimental configurations of statistical models. Performance statistics for track and intensity were considered individually, such that an experimental model might be selected to provide track guidance but not intensity guidance and vice versa. Experimental candidates, when appropriate, also had the potential to be selected to be included in an experimental consensus forecast or simple multimodel ensemble, which was created by simply averaging the forecasts from a select set of operational and experimental model forecasts (e.g., Simon et al. 2018; see also <https://www.nhc.noaa.gov/verification/verify2.shtml>).

Each Stream 1.5 evaluation focused on comparing the performance of the experimental models with the performance of the previous year's top-flight forecast models; these models are defined as the three operational models with the best performance during the previous hurricane season. Note that the top-flight forecast models for track are not necessarily the same as those for intensity. These operational baselines were, for the most part, based on the version of the model run at the time of the storm (i.e., the real-time operational guidance). When appropriate, an evaluation of the Stream 1.5 candidate's impact on operational consensus forecasts was also conducted. For this evaluation, the baseline was the operational consensus forecasts and an experimental consensus was created by adding the experimental forecasts to the operational consensus.

The performance guidelines put forth by NHC for their decision process served as the starting point for developing the evaluation plan with an eye toward providing the necessary information in the most concise yet informative format possible. The evaluation plans evolved over the years as outcomes from the previous year's evaluation stimulated more questions and an interest in gaining a better understanding of nuances in the datasets. Once a question was identified by NHC staff and discussed with NCAR staff, the evaluation group endeavored to identify a statistically valid analysis approach that would provide a meaningful answer. The planning and selection guidelines also took into account the NHC's extensive experience with respect to the performance trends of the operational models.

Differences between the performance trends for track and intensity forecasts meant that the selection guidelines for track and intensity were not identical:

Tropical cyclone track forecast guidelines: At the time of the Stream 1.5 project, NHC's track forecast errors decreased by about 3%–4% per year on average, which paralleled gains made in the operational numerical guidance. Based on these trends, NHC put forth the following guidelines for experimental models to be selected for Stream 1.5 track guidance:

- Projected improvement of 3%–4% over the average error of the previous year's top-flight models
- Techniques that improve the conventional model consensus track error by at least 3%–4%
- Schemes that otherwise enhance the operational forecast by providing better "guidance on guidance" (e.g., Rappaport et al. 2012)

¹ A file format used by NHC and other entities that includes a TC's location and intensity.

² Late model guidance refers to guidance that would not be available during the forecast cycle, which starts at a synoptic time (e.g., 1200 UTC) and ends with the release of the official forecast 3 h after the synoptic time (e.g., 1500 UTC). Late model guidance is converted to early model guidance for the next cycle by applying an adjustment based on the current synoptic time and analyzed position and intensity of the TC.

- An especially high “frequency of superiority”
- High run-to-run consistency in combination with acceptable performance

Tropical cyclone intensity forecast guidelines: Leading up to the Stream 1.5 project, little to no improvement in NHC tropical cyclone intensity forecast accuracy had occurred over at least 20 years. Model guidance had improved but on average was no better than the NHC forecast. Hence, techniques that improved upon existing guidance for tropical cyclone intensity and rapid intensification³ received special consideration in the selection process.

While model developers were not required to submit retrospective forecasts for every storm or forecast cycle laid out in the test plan, shorter test periods or smaller samples were strongly discouraged. Participants were made aware that smaller samples would likely necessitate larger improvements in performance in order for a model to be selected for Stream 1.5. Ultimately, to participate in Stream 1.5, a candidate project had to be approved by HFIP management, the developer or developer’s home institution, and the NHC, with NHC having the final authority to select which candidate prediction systems qualified for the Stream 1.5 real-time activity.

Over the five years during which the Stream 1.5 concept was implemented, the number of experimental configurations evaluated each year varied from 4 to 10, where some developers submitted multiple configurations for evaluation in a particular year. Stream 1.5 candidates selected in a particular year ranged from a little more than 40% of those submitted to almost 90%. Candidates had the possibility to be selected to have their forecasts displayed explicitly for track and/or intensity, or to be a member of an experimental consensus for track and/or intensity, or both. At least one candidate was selected for each of these four categories each year, where the number selected per category varied from one to as many as four.

While the selection guidelines appear to be relatively straightforward, the outcomes of the evaluations were not always clear. Often a candidate would outperform the baselines for certain lead times while either being indistinguishable from the baselines as assessed via statistical significance evaluations or performing worse than the baselines for other lead times; or they could exhibit strong performance in one basin but not the other. To be selected to be displayed explicitly during the demonstration, NHC looked for candidates that demonstrated the required degree of improvement for a majority of the lead times with little to no degradations for the remaining lead times for at least one of the basins. Performance for the other basin could be somewhat mixed (e.g., degradations for a limited number of lead times) or simply not distinguishable from the baselines, but a strong signal of poor performance for the other basin led to a candidate not being selected. Candidates with smaller sample sizes could be selected to be included in Stream 1.5, but only if the candidate’s

³ NHC defines rapid intensification as an increase in the maximum sustained winds of a tropical cyclone of at least 30 kt (35 mph; 55 km h⁻¹) in a 24-h period (<https://www.nhc.noaa.gov/aboutgloss.shtml>).

performance was especially strong. Simply meeting the selection guidelines for small samples was not sufficient. Candidates for which the metrics revealed inconsistent or erratic performance (outperformed baselines about as often as performing substantially worse) were not selected, nor were candidates associated with forecasts with a notable number of cases with exceptionally large errors. These two characteristics can only be determined by looking beyond the traditional bulk statistics, as will be demonstrated in the examples discussed in [section 3](#).

3. Hierarchy of user-responsive evaluations

Evaluations of retrospective TC intensity predictions produced by two experimental models (E1 and E2) compared to “best track” estimates of actual intensity⁴ are presented in the following subsections with the goal of demonstrating the concepts and ideas behind the hierarchy of analyses developed for the HFIP retrospective evaluations. The description of each example starts with approaches that were applied to answer very basic questions about the performance of the experimental model, in comparison to the performance (also relative to the best track estimates) of the three baseline models (B1, B2, and B3; [Table 1](#))⁵ and then progress to more complex evaluations that were undertaken to respond to questions that have greater specificity. These methods include approaches traditionally used in TC forecast evaluations (both operationally and in research), along with several new approaches developed during the project. The retrospective forecasts in these examples were produced by each experimental modeling group for TCs that occurred in three prior years. While multiple experimental models were considered for HFIP’s demonstration project, it is important to note that the purpose of the retrospective evaluations was not to directly compare the experimental models to each other. Rather, NHC was interested in how well the individual models performed relative to the baseline models, which represent the top performing models available and have traditionally been used to predict storm movement and intensity. The examples consider intensity forecasts, but the concepts presented equally apply to predictions of track.

a. Example 1

This example was selected to demonstrate how the hierarchy of analyses provided NHC with in-depth information on the performance of an experimental model (E1) that was selected for Stream 1.5. The information was able to provide NHC forecasters with the confidence they needed to feel comfortable adding the experimental model guidance to the products they use to create their forecasts.

⁴ The “best track” is a subjectively smoothed representation of a TC’s location and intensity over its lifetime based on a post-storm assessment by experts (<https://www.nhc.noaa.gov/aboutgloss.shtml#ra>). Best track intensity estimates have a precision of 5 kt.

⁵ The experimental and baseline model-based intensity predictions have a precision of 1 kt.

TABLE 1. Baseline models used in the HFIP evaluations of intensity predictions (NHC 2019).

Model No.	Model ID	Name	Description	Reference
B1	GHMI	Interpolated-dynamical Geophysical Fluid Dynamics Laboratory (GFDL) model	Previous cycle of the NWP model, GFDL, adjusted using a variable intensity offset correction	Bender et al. (2007)
B2	LGEM	Logistic Growth Equation Model	Statistical-dynamical model based on a logistic growth equation	DeMaria (2009)
B3	DSHP	Decay-Statistical Hurricane Intensity Prediction Scheme	Statistical-dynamical model based on multiple regression techniques, that considers relationships between storm behavior and environmental conditions estimated from dynamical model predictions, climatology, and persistence	DeMaria and Kaplan (1994)

1) AVERAGE PERFORMANCE

Figures 2 and 3 consider average performance of the forecasts, measured as the mean of the absolute values of the intensity errors (i.e., MAIE). In particular, Fig. 2 presents comparisons of the average performance of the B1, B2, and B3 forecasts to the average errors for the three years of retrospective forecasts produced by E1 (using homogeneous samples), with sample sizes for each lead time presented across the top of the chart.⁶ The plots in Fig. 2 are similar in form to those shown in Fig. 1. That is, the plots show verification results for the forecasts of interest relative to a standard of comparison. In contrast to Fig. 1, however, the standards of comparison are the three baseline models selected as the standards by which to evaluate the experimental models (note that these models—denoted by NHC as “top-flight” models—are different from the OCD5 no-skill forecasts applied in Fig. 1). The results presented in Fig. 2 suggest that (i) the E1 forecasts had much better average performance than the B1 forecasts, for almost all lead times, and especially for later times, and (ii) the E1 predictions are slightly better than the B2 and B3 predictions for most lead times. Based on this information, one might conclude that E1 has better performance overall than all three of the baseline models.

Figure 3 displays the same scores as Fig. 2; however, Fig. 3 also considers the question of whether the average differences are meaningful from a statistical perspective. A natural first step toward answering this question is to estimate statistical confidence intervals (CIs; Chambers et al. 1983; McGill et al. 1978; Wilks 2019) for the average intensity error values shown

in Fig. 2. The 95% CIs, computed using standard methods based on the t distribution and incorporating a variance inflation factor (VIF; Wilks 2019) to account for temporal autocorrelation, are shown in Fig. 3 using black and orange vertical lines.

The black and orange CIs in Fig. 3 represent a measure of the statistical uncertainty associated with the average intensity errors estimated for E1 and each type of baseline forecast. Examining Fig. 3 indicates that the CIs for E1 overlap the CIs for all three baseline forecasts for all lead times. Based on this information, one would conclude that E1 performance is not significantly better than the performance of any of the baseline models for any lead times—quite a different conclusion from that reached via Fig. 2.

Further examination of Fig. 3 provides more insights. The values in blue at the bottom of each graph are based on “paired comparisons,” of the intensity errors associated with E1 and each of the baseline forecasts. That is, the errors for E1 and each baseline model forecast initialization (also commonly called “issue time”) and lead time are directly compared; for example, in Fig. 3a, the E1 error for a particular date, issue, and lead time has been subtracted from the corresponding B1 error for the same date, issue, and lead time, and the same has been done for each date, issue, and lead time in the sample. This process yields a set of error differences for each model pair (i.e., E1 and B1, E1 and B2, and E1 and B3) and for each lead time. From these difference samples, average paired differences and their CIs were computed using the t distribution, taking into account first-order temporal correlations (Chambers et al. 1983; McGill et al. 1978; Wilks 2019). The blue dots represent the average paired differences between errors associated with the E1 intensity predictions and corresponding errors associated with each of the baseline model predictions; the vertical lines associated with each blue point are the 95% CIs on these differences. The paired differences can be considered significantly different from zero if the CIs do not overlap the horizontal “zero” lines; these significant differences are represented by filled blue dots.

Examination of the paired differences in Fig. 3 yields still different conclusions from the information provided by comparing the average errors in Fig. 2 and the black and orange lines in Fig. 3. In particular, the paired comparisons suggest

⁶ Note that the sample sizes in this and later figures differ across the lead times and the baselines. This variation occurs because individual times are not included in the verification unless observed and forecast intensity are present in the sample for both the baseline and the experimental model being compared. Tropical cyclones have a finite lifetime. The samples include forecasts for the same storm throughout its lifetime, which naturally leads to smaller samples for longer lead times because the observed and/or forecasted disturbance will no longer be classified a tropical cyclone for longer lead times as the forecast cycles approach the end of the tropical cyclone’s lifetime. The same would be true for evaluations of track location.

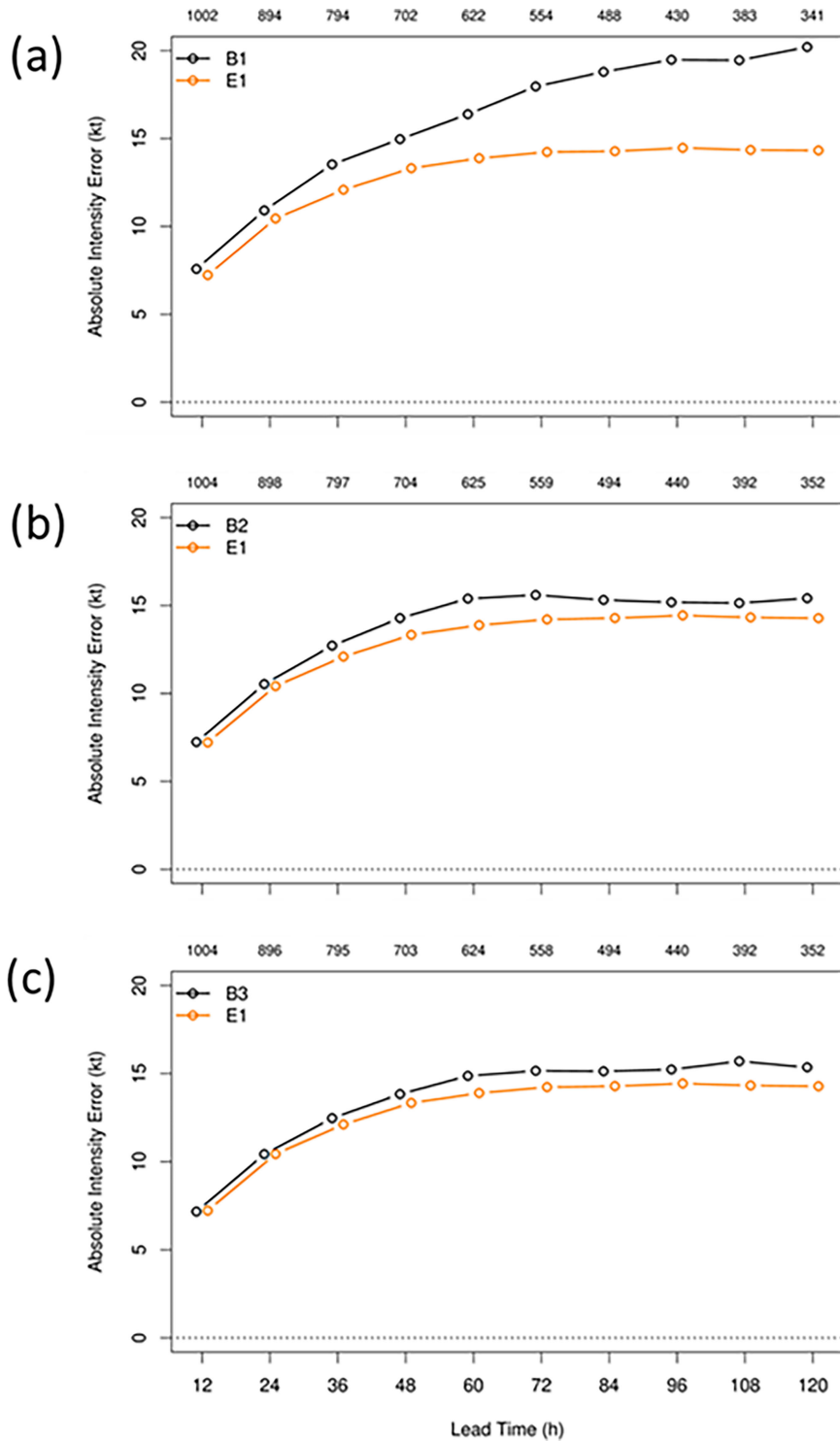


FIG. 2. Mean absolute intensity error (MAIE) values for E1 (orange) compared to the (a) B1, (b) B2, and (c) B3 baseline models (black). The sample size for each lead time and model comparison is shown along the top of each graph.

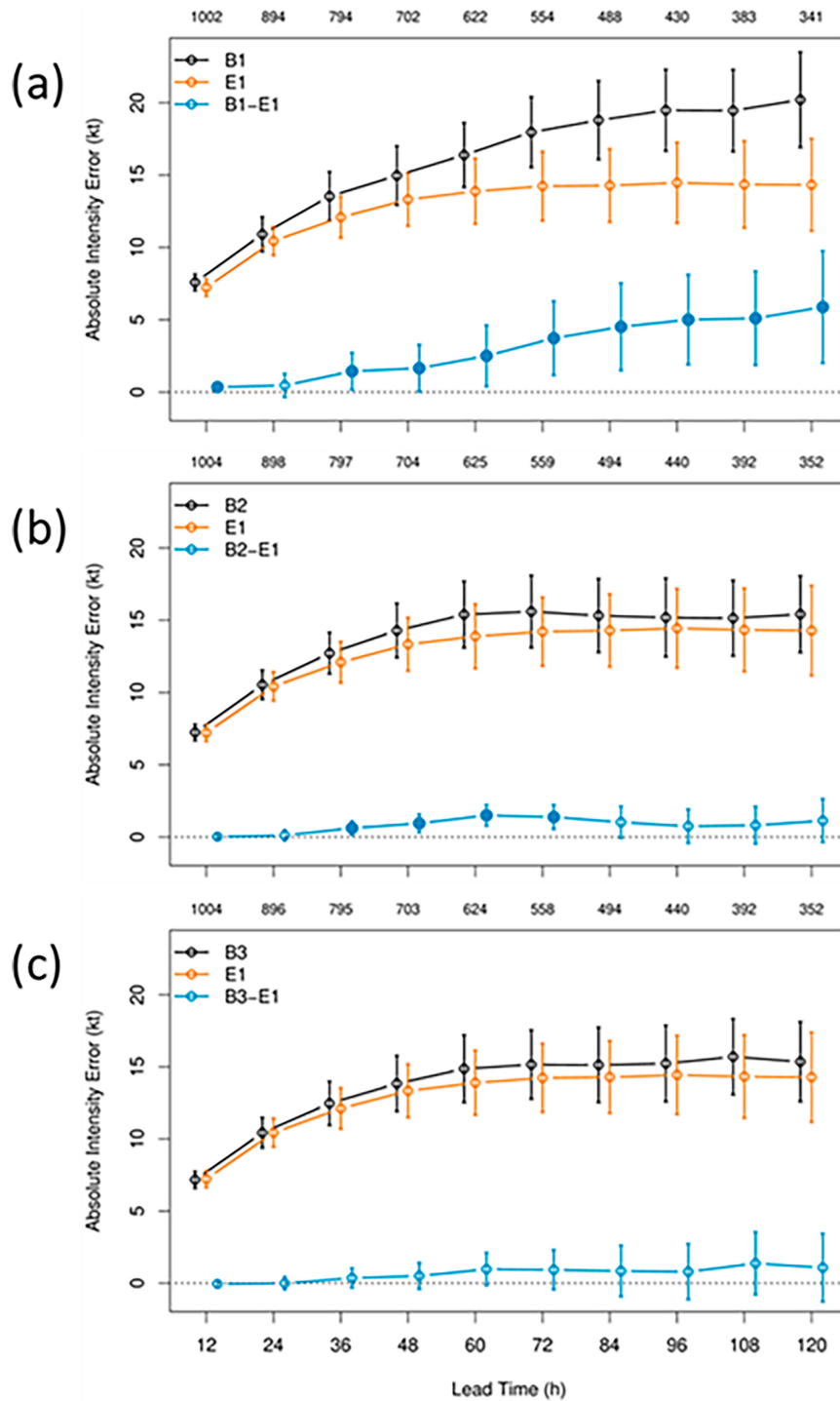


FIG. 3. Black and orange dots show the MAIE values for E1 (orange) and each of the three baseline models (black), as in Fig. 2. Blue dots show average differences between the paired absolute intensity values for E1 and the baseline models (baseline model error minus E1 model error). Vertical lines represent the 95% CIs on the mean errors and differences. A filled blue dot occurs whenever a CI for the error differences does not intersect zero, indicating a statistically significant mean error difference.

Forecast Hour	12	24	36	48	60	72	84	96	108	120
B1	1002	894	794	702	622	554	488	430	383	341
	0.3	0.5	1.4	1.7	2.5	3.7	4.5	5.0	5.1	5.9
	5%	4%	11%	11%	15%	21%	24%	26%	26%	29%
	0.970	0.735	0.973	0.954	0.980	0.996	0.997	0.998	0.998	0.997
B2	1004	898	797	704	625	559	494	440	392	352
	0.0	0.1	0.6	1.0	1.5	1.4	1.0	0.7	0.8	1.1
	0%	1%	5%	7%	10%	9%	7%	5%	5%	7%
	0.257	0.453	0.988	0.997	0.999	0.999	0.933	0.789	0.786	0.862
B3	1004	896	795	703	624	558	494	440	392	352
	-0.1	0.0	0.4	0.5	1.0	0.9	0.8	0.8	1.4	1.1
	-1%	0%	3%	4%	7%	6%	6%	5%	9%	7%
	0.373	0.023	0.698	0.719	0.905	0.814	0.648	0.581	0.785	0.628

FIG. 4. Summary table for comparisons of E1 performance to the performance of the three baseline models. The four elements in each cell, from top to bottom, are the sample size (i.e., the total number of paired comparisons), the mean error difference, the percent improvement (or reduction in performance), and the probability ($\times 100$) of having a smaller error difference (i.e., a difference closer to 0) based on a paired comparison. Blue (red) lettering in the second row of each table cell indicates the candidate model had better (worse) performance than the baseline model. Red/blue shading is used to indicate whether significant differences are $\geq 5\%$ (lighter shading) or $\geq 10\%$ (darker shading).

that (i) E1 performance is significantly better than B1 performance for almost all lead times and (ii) E1 performance is significantly better than B2 performance for lead times between 36 and 72 h, inclusive. However, in general, E1 performance is not shown to be significantly better than B3 performance for any lead time. The differences in results between the unpaired and paired comparisons in Fig. 3 occur because the paired comparison—evaluating the error differences—is a more powerful statistical test of the differences between the errors (e.g., Wilks 2019).

The lessons learned from Figs. 2 and 3 include the importance of considering sampling variability (as represented here via confidence intervals) when making comparisons between forecasting systems. Unjustified conclusions can be made when this sampling uncertainty is ignored; for example, it would be inappropriate to conclude that E1 forecast performance was significantly better than B3 performance (as might be concluded from Fig. 2). Moreover, unpaired comparisons (represented by the black and orange vertical lines in Fig. 3) lead to conclusions that are different from those obtained by more powerful comparisons using a paired statistical test, which demonstrate that the E1 forecasts were significantly better than the B1 predictions for most lead times, and they were better than the B2 predictions for some lead times. It is worth noting that even this more efficient comparison approach was not able to identify significant differences between the E1 predictions and those provided by B3. However, another important conclusion from these plots in the context of HFIP goals is the fact that E1 did not perform significantly worse than the baseline forecasts.

Many of the results in Figs. 2 and 3 can be summarized in a table, to provide a quick-look summary of the overall performance of the candidate model relative to the baseline models. Over several years, NCAR worked closely with NHC staff to develop and enhance a table to provide a useful summary of

the information that can be gleaned from the figures, including further information about the relevance of estimated performance differences. The display that resulted from this evolutionary process is demonstrated in Fig. 4 for the E1 model evaluation. This type of quick-look summary of multiple comparisons is often referred to as a scorecard.

Each cell in Fig. 4 includes key information for a given comparison, which succinctly summarizes the results of each comparison in a rich and easily understandable overview of the model’s performance. The numbers, from top to bottom in each cell, are the sample size, the average error difference (with blue text indicating the candidate model had smaller average errors, and red text indicating that the baseline model had smaller average errors), the percentage difference in average errors, and the significance level for the paired statistical comparison. This table is essentially a summary of the graphs (e.g., Fig. 3) showing paired comparisons between the candidate and baseline models, with explicit results of the application of a paired t test, including an autocorrelation-based adjustment. Colored shading is used to represent the strength of the differences for comparisons that are statistically significant, as defined in the table legend. The Stream 1.5 selection guidelines described in section 2 served as the guiding principles for the difference ranges associated with gradations in the shading. These guidelines provided clear guidance for track—improvement of 3%–4%. Since the intensity guidelines did not specify a threshold, the NCAR team consulted with NHC to determine what shading scheme they felt would be most useful for their decision-making process.

A quick look at Fig. 4 indicates (as also shown in Fig. 3) that many of the differences in performance between E1 and B1 and some of the differences between E1 and B2 are statistically significant. Moreover, many of the (statistically significant) comparisons of E1 and B1 errors indicate average error improvements of 10% or greater. In contrast, the comparisons

between E1 and B2 performance suggest improvements (for lead times with statistically significant results) between 5% and 10%. While the average differences estimated for the comparison between E1 and B3 suggest better performance by E1, the differences are not statistically significant.

These results gave NHC confidence that forecasts provided by model E1 would on average be at least as good as—and clearly no worse than—the baseline forecasts, and would improve upon the forecasts provided by two of the baseline forecasts (when considering average performance) for at least some of the lead times. While the Stream 1.5 guidelines focused on the candidates' demonstrating improvements over the top-flight models, NHC did not require this improvement to be universal for all top-flight models for all lead times. Hence, based on these metrics, E1 was a promising candidate.

2) LARGE ERRORS

Very large individual errors in track or intensity are often of greater importance to NHC than average errors. Such large errors could be associated with consequential incorrect decisions made by forecasters. For example, large errors in track location forecasts could have important implications for decision-making (e.g., regarding regions to be warned), as could large under- or overprediction of TC intensity, which could lead, for example, to nonoptimal estimates of life-threatening surge or potential damage by end-users and have other impacts on end-users' decision-making (e.g., about evacuation), particularly if forecast credibility is damaged. Thus, NHC was interested in characterizing the frequency and size of large errors associated with forecasts provided by the experimental models.

One approach to evaluating the occurrence of large errors involves examining the full distributions of absolute errors, as in Fig. 5. In this figure, boxplots show the quantiles of the distributions of E1 absolute errors in comparison to the absolute error distributions for the three baseline models. The plots for individual lead times also show the mean values (denoted by the * within each box), the upper limits of values not considered to be outliers (the end of the upper bar—also called the “whisker”—above each box) and the points considered to be outliers⁷ (open circles above the upper whiskers).

Examination of Fig. 5 suggests that the central portions of the error distributions (as represented by the “boxes”) for E1 and the three baseline models are fairly similar, with considerable overlap, but some differences can be observed with respect to the locations of the medians (horizontal line inside each box) and the heights of the boxes, which represent the central portions of the distributions of values. The boxplot approach to comparing the errors also brings to light that the B1 errors are more variable than the E1 errors, as can be seen from the fact that the boxes for B1 tend to be notably taller

⁷ Here outliers are defined as data values that are located above the “whisker” located above the central box. The length of the whisker is estimated as 1.5 times the interquartile range (IQR). (IQR is the difference between the 0.75th and 0.25th quantiles.) See Tukey (1977), Wilks (2019).

than the boxes for E1. In addition, errors considered to be outliers (represented by the points above the boxes and whiskers) suggest some differences in performance. In particular, B1 tended to have more large errors than E1 (Fig. 5a). In contrast, outlier points in Figs. 5b and 5c (comparing the upper points for E1 to those for B2 and B3) indicate the E1 outlier errors are fairly similar to those from these baseline models. This result suggests that E1 is associated with similar or smaller outlier errors than would arise from use of the baseline models—that is, E1 performs as well or better than the baseline models in terms of this attribute. Hence, the NHC forecasters could expect to encounter large errors that are not outside the norm associated with the baseline models when using E1 in their forecasting processes.

3) RELATIVE PERFORMANCE

In considering differences in performance (i.e., differences in errors between models) for particular storms and times, NHC was interested in the frequency with which the errors produced by the candidate model were larger or smaller than the corresponding (paired) errors produced by the baseline models. If the experimental model typically produced smaller or equivalent errors, it might be a good candidate to present to the forecasters for consideration in their forecasting process (depending on other aspects of performance).

The analyses in Fig. 6 show the frequencies with which the experimental and baseline models were “better” than the alternative (i.e., the percent of times when the baseline was better, when the experimental forecast was better, and when the models were essentially tied). In this example, an error difference threshold of 1 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$) (i.e., the precision of the model-based intensity forecasts), was applied to define “better.”⁸ When the difference in errors was 1 kt or smaller, the model results in these diagrams were categorized as a tie. To represent uncertainty, 95% confidence intervals are shown for the frequency values, based on application of the binary distribution and taking into account first-order autocorrelations (Wilks 2019).

In the comparison of E1 to B1 (Fig. 6a), it is apparent that the experimental forecasts more frequently had smaller errors for longer (>48 h) lead times. For the comparison between E1 and B2 (Fig. 6b), E1 tended to perform better than B2, particularly for lead times between 36 and 84 h. As in other analyses, the comparison of E1 and B3 provides a more nuanced result: while E1 tended to be associated with superior performance for many lead times, the frequency differences were not large and were not statistically significant (as indicated by the overlap of the CIs) for many lead times. For all comparisons, ties were very infrequent.

It is interesting to note that some of the results in Fig. 6 seem to be inconsistent with the results shown in Fig. 4. For example, while Fig. 4 suggests E1 performance is significantly better than B1 performance for almost all lead times, the confidence intervals in Fig. 6 overlap somewhat for early lead

⁸ For track, a difference of 6 n mi ($1 \text{ n mi} = 1.852 \text{ km}$) was used as the threshold for evaluations of superior performance.

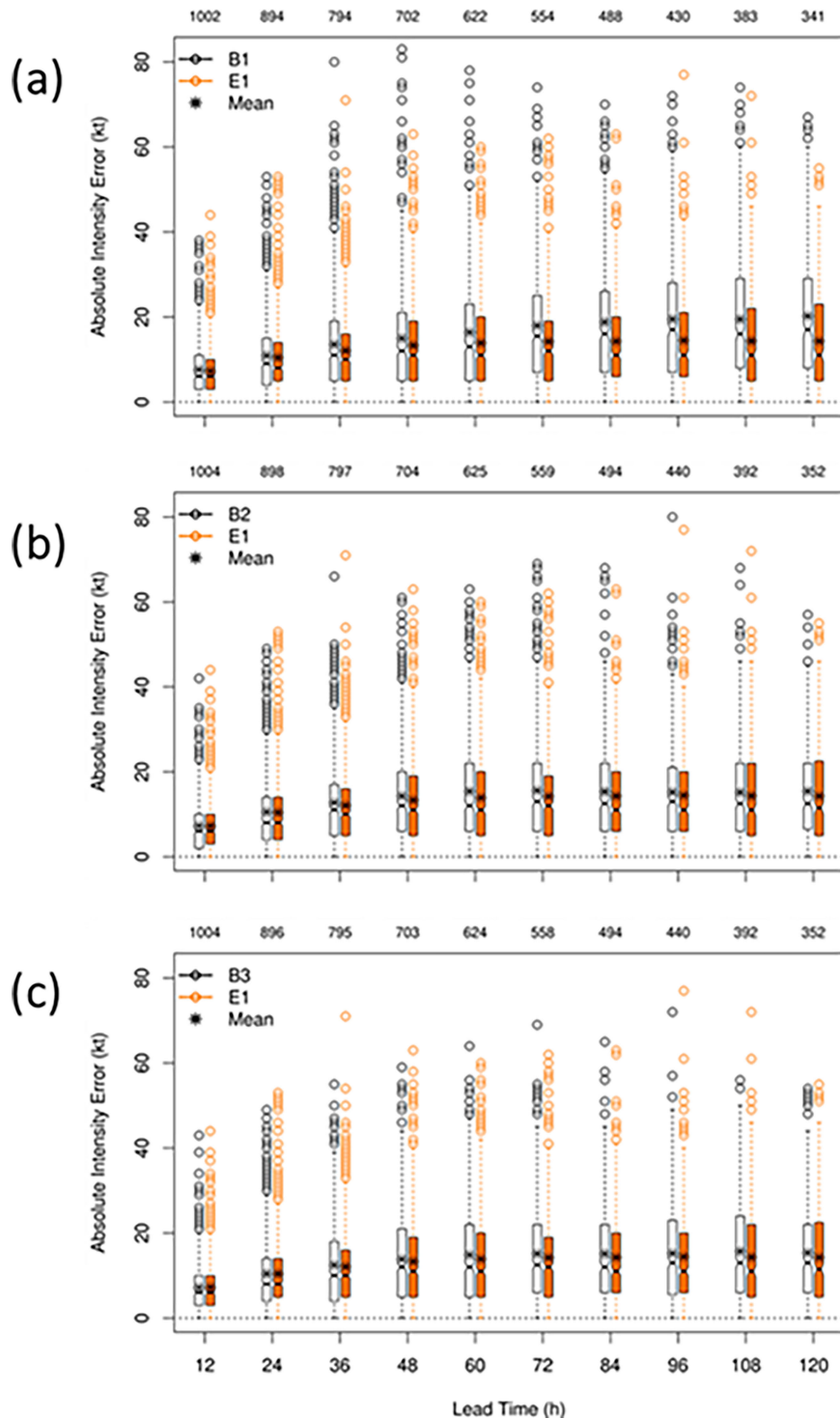


FIG. 5. Boxplots showing distributions of absolute intensity errors for predictions by E1 (in orange) and each of the baseline models (in black). Boxes represent (from top to bottom) the 0.75th (top horizontal line), 0.50th (middle horizontal line), and 0.25th (bottom horizontal line) quantile values of the error distributions. “Whiskers” above the boxes represent the nonextreme values (greater than the 0.75th quantile) included in the distributions, and whiskers below the boxes extend from the 0.25th quantile to the minimum values (0, in this case), representing the smallest errors. The mean is denoted by an asterisk. Outlier values are represented by the circular points near the tops of the plots.

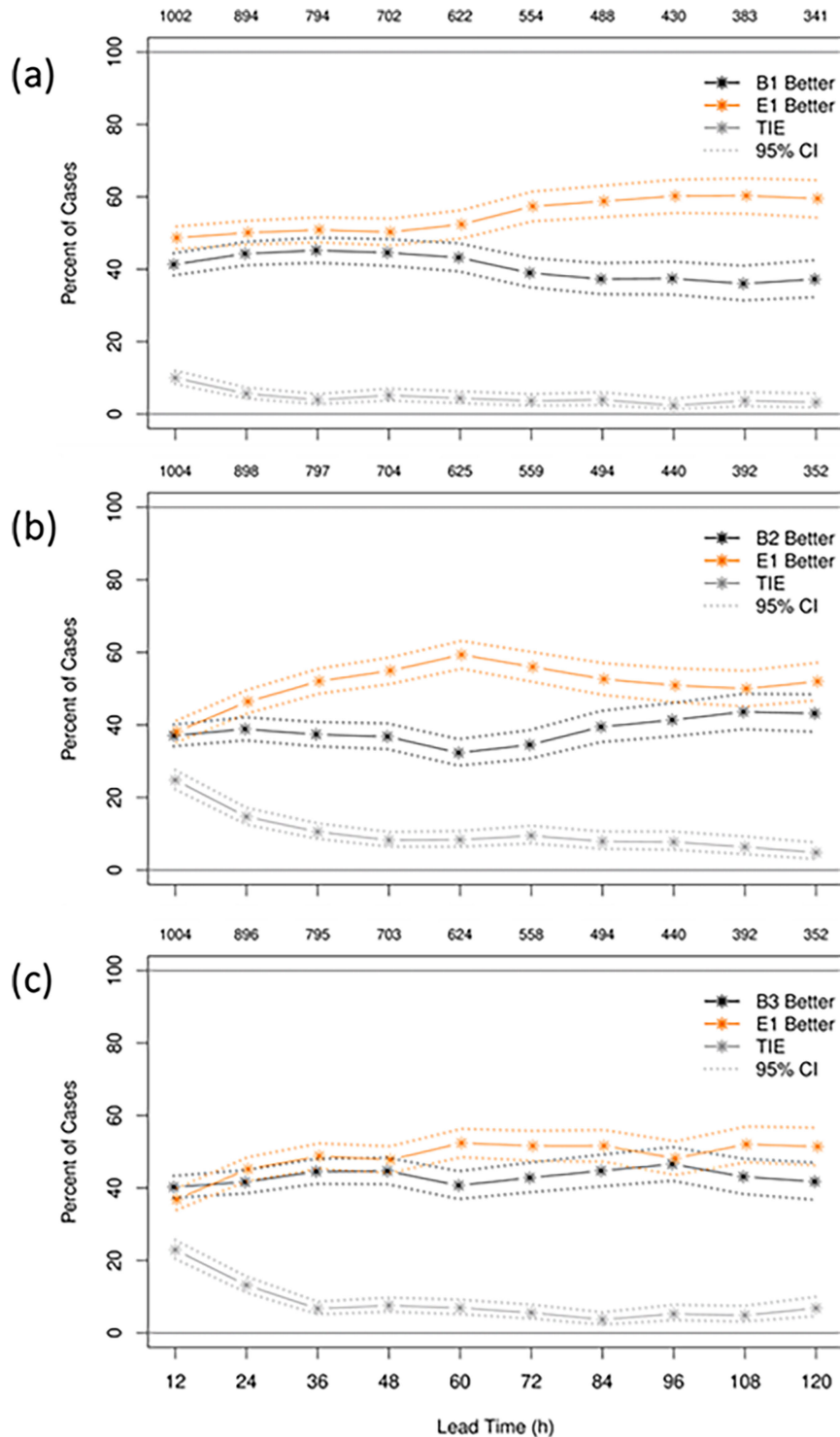


FIG. 6. Frequency (%) with which the experimental model (E1) had larger or smaller errors than the baseline models (a) B1, (b) B2, and (c) B3, using an error difference threshold of 1 kt to distinguish improvement of one model over another. A “tie” was assigned when the absolute error difference was less than or equal to 1 kt. The bands around the lines represent 95% confidence intervals on the frequencies, based on application of the binary distribution, including consideration of first-order autocorrelations.

times, giving a less strong indication of differences. However, in both Figs. 4 and 6, the largest differences between B1 and E1 performance are associated with longer lead times (72–120 h). Similar apparent discrepancies can be seen in the comparisons of E1 with B2 and B3. Specifically, the E1 versus B2 comparison in Fig. 6b suggests significant improvements of E1 over B2 for lead times between 36 and 84 h (and marginally, 24 and 96 h) lead times, whereas Fig. 4 shows significant differences for lead times of 36–72 h. For the B3 comparison, Fig. 4 does not identify any significant differences between B3 and E1 results, whereas the results in Fig. 6 suggest significant differences at 60 h and marginally at 72 and 108 h.

While the results in Figs. 4 and 6 seem inconsistent, the two analyses represented in these diagrams demonstrate the benefits of examining performance from more than one perspective. In Fig. 4, the CIs focus on the average magnitude of the differences (e.g., $E1 - B1$) and their variability, while the analyses shown in Fig. 6 are not particularly focused on the magnitude of the differences. In fact, in the Fig. 6 analysis, a small (2-kt) difference counts equally as much as a large one (e.g., 10 or 20 kt). In contrast, in Fig. 4, the magnitude and variability of the differences has more impact. Hence, the Fig. 6 analysis answers a very different question from the one considered in Fig. 4, and provides a different nuance that is meaningful when considering how often a model exhibits an error larger than that of other models.

Although not applied during the original HFIP evaluations, one benefit of many of the approaches that were employed is the ability to vary thresholds and other aspects of the analysis to target specific questions of interest. For instance, the examination of “superior” performance can be adapted to apply a threshold larger than 1 kt. As an example, Fig. 7 shows the frequency of superior performance percentages for E1 versus B1, B2, and B3 using a difference threshold of 5 kt to define “superior performance,” which corresponds to the precision of the intensity estimates in the best track and NHC forecasts. Comparing Fig. 7a to Fig. 6a suggests that the overall patterns are mostly similar between the two thresholds, but some of the nuances are different. For example, the number of ties is much more frequent with the 5-kt threshold compared to the 1-kt threshold. In fact, for longer lead times (72 h and greater) for the E1 versus B1 comparison, ties occurred with the same frequency as superior performance for B1. In the E1 versus B2 and E1 versus B3 comparisons, ties were more common than the experimental or baseline frequencies for all (E1 versus B2) or most (E1 versus B3) of the lead times. Hence, one could conclude that, with a 5-kt difference threshold, E1 performed more similarly to B2 and B3, and most frequently had smaller errors than B1, particularly for longer lead times. Finally, the curves for the frequencies of ties in Figs. 6 and 7 all have a decreasing trend with lead time. This result mirrors the fact that the forecasts tend to be more similar for shorter lead times and to depart from each other as the lead-time increases.

To summarize, the frequency of superior performance analyses provided information to NHC about relative performance and gave them confidence regarding whether the experimental system would produce forecasts that would be

as good as or better than the baseline forecasts a majority of the time.

4) PERFORMANCE RANKING

All of the approaches presented earlier compare the experimental forecasts to an individual baseline. These individual comparisons do not provide any insight into whether cases where E1 outperformed B1 were the same cases as those where E1 outperformed B2 or B3. Discussions with NHC staff revealed that they also were interested in understanding how the experimental models performed relative to the top-flight models as a group. In this example, forecasts from the experimental model are compared to forecasts from all three baselines. For each forecast, the models’ errors were ordered from smallest to largest and each model was assigned a rank value. These rank values can be summarized by examining how frequently the candidate model achieved each rank. Note that for this approach, the same samples of forecast dates and times are used for all of the forecast systems included in the comparisons (i.e., a homogeneous sample was used for all baselines and the experimental model).

Figure 8 summarizes the information gleaned from this approach applied to candidate model E1. The frequency with which E1 performance was best (i.e., ranked 1), second best (i.e., ranked 2), third best (i.e., ranked 3), and last (i.e., ranked 4) is presented, with 95% confidence intervals on the frequencies shown using dashed lines. The horizontal line intersecting the y axis at the 25% point represents the expected percentage of time that any of the models would be expected to be best if the frequencies were random.

As shown in Fig. 8, the results for this comparison are somewhat mixed. For shorter lead times (through 72 h), E1 most commonly ranked second, whereas for very long lead times (108 and 120 h) it ranked first/best. It was uncommon for E1 to rank last (i.e., 4th) in this comparison. It is notable that (i) the frequency of E1 ranking best is above the 25% line for lead times greater than 48 h and (ii) the frequency of E1 ranking worst is less than 25% for all lead times. These results gave NHC confidence that E1 could contribute meaningfully to the forecasting process.

5) NHC DECISION-MAKING

These analyses gave NHC a wide variety of insights into the potential contributions of E1 to their forecasting process. After examining these results, NHC made a decision to explicitly include E1 in the forecasting process during the subsequent hurricane season. This decision was made because E1 demonstrated promise for being able to “improve upon existing guidance for tropical cyclone intensity” and therefore met the selection guidelines.

b. Example 2

This example was selected to demonstrate how the hierarchy of analyses provided NHC with in-depth information on the performance of an experimental model (E2) that was not selected for Stream 1.5. This experimental model showed promise for some aspects of the retrospective sample but also

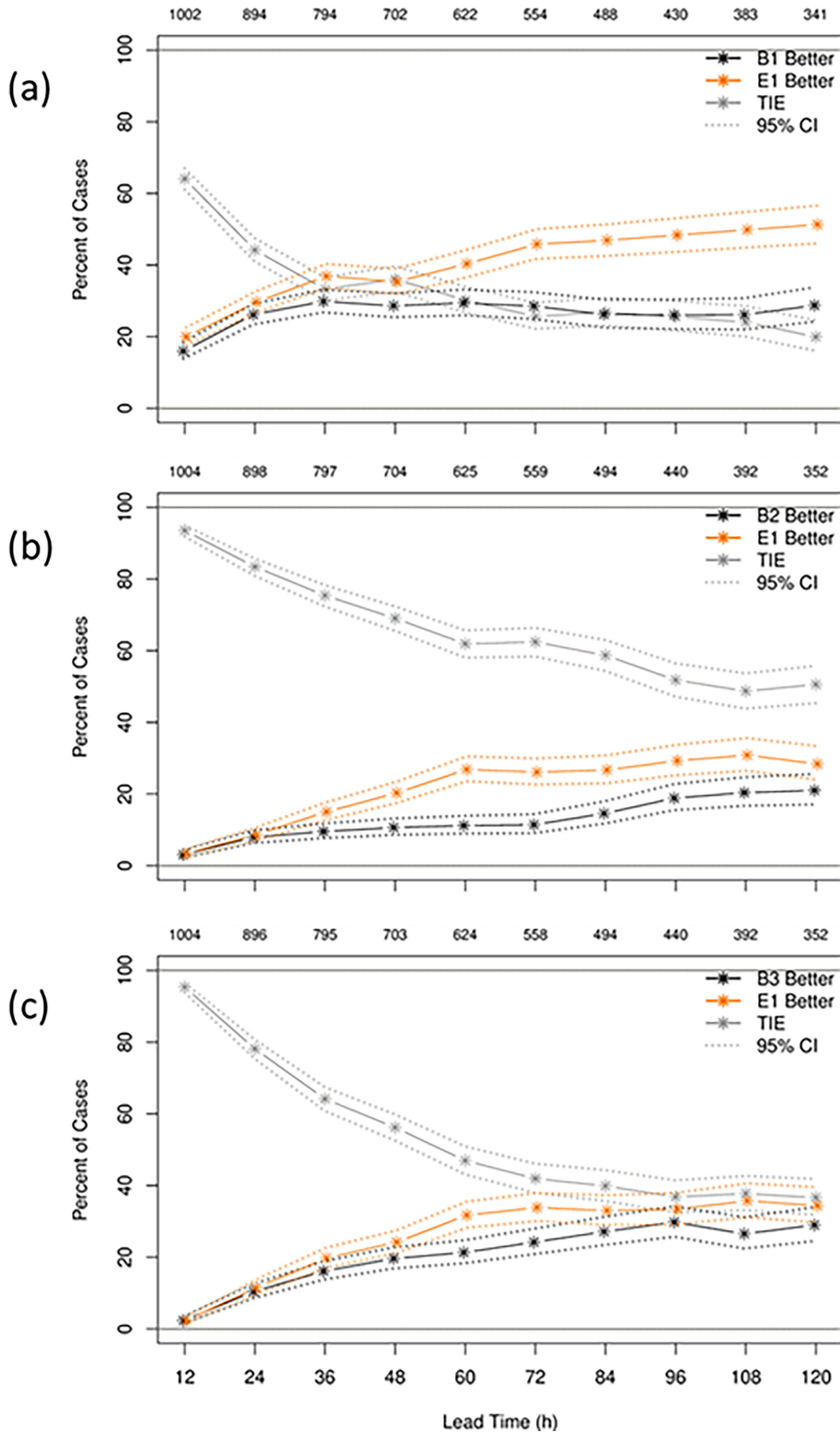


FIG. 7. As in Fig. 6, but a tie is assigned to differences less than or equal to 5 kt.

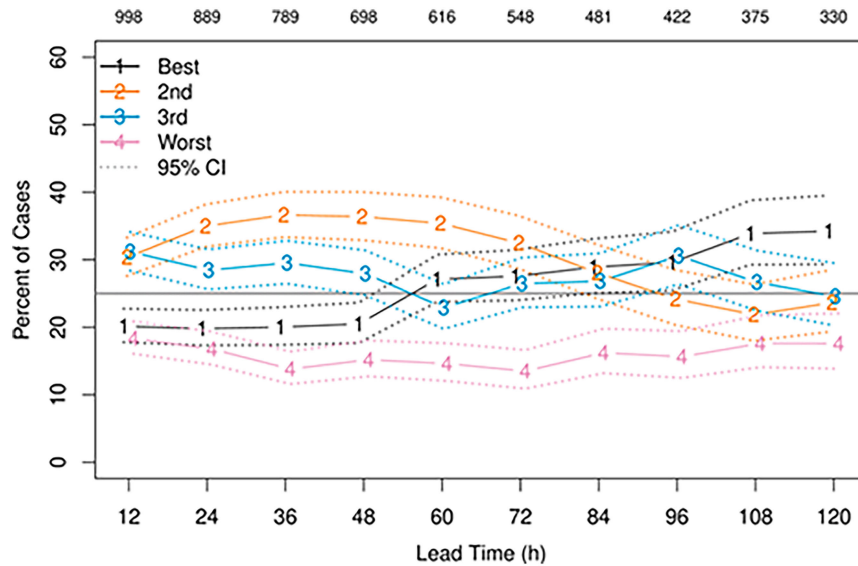


FIG. 8. Ranking of E1 performance in predicting TC intensity relative to three top-flight forecasting models in use by NHC, showing the percent of cases for which E1 was best (1), second best (2), third best (3), and worst (4). The 95% confidence intervals for the frequencies, based on the binomial distribution, are shown using dotted lines surrounding each solid line. The solid horizontal line intersecting the y axis at a value of 25% represents the frequency with which a forecast would be expected to be “best” if the frequencies were random.

behaviors that raised concerns about a potential for inconsistent performance.

Figure 9 shows average performance results for E2 compared to B1, B2, and B3. Without considering the CIs, it appears that the E2 average errors are (i) notably smaller than the errors associated with B1 for nearly all lead times and (ii) approximately equivalent to the errors associated with B2 and B3 for lead times less than 96 h, but with larger average errors for longer lead times. The paired average differences between E2 and the baseline system errors are shown in blue, with 95% CIs, as in Fig. 3 for E1. The CIs around the MAIE for the three model pairs (the top parts of the three plots) suggest there are no significant differences between the performance of E2 and the three baseline models. In contrast, the results in the bottom portion of Fig. 9a (blue line) indicate that E2 performance is significantly better than B1 performance for 72-h forecasts, with no significant differences for other lead times. In contrast, E2 performance is significantly worse than B2 performance for lead times of 12, 108, and 120 h; and significantly worse than B3 performance for 12- and 120-h lead times.

Figure 10 quantifies these observed differences. The results in Fig. 10 also indicate that the percentage improvement of E2 over B1 at 72 h amounts to an 18% reduction in error. In contrast, the significant percentage degradations of E2 relative to B2 and B3 range from -10% (12-h predictions) to -32% (120-h predictions from B2).

The boxplots in Fig. 11 summarize the distributions of absolute intensity errors for E2 compared to B1, B2, and B3. As in Example 1, the results in Fig. 11 are useful for examining both the variability in the errors and the magnitudes of outliers. In

general, the boxes in Fig. 11 are relatively similar for the baseline models and E2, particularly for shorter lead times. It is interesting to note somewhat greater central variability in B1 errors relative to E2 errors for intermediate (36–96 h) lead times, as well as greater central variability in E2 errors compared to B2 and B3 errors for the longest lead times. However, the largest differences in performance are associated with the large errors for all lead times greater than 24 h. In particular, the E2 forecasts had larger outlier values than the outlier values for the baseline systems (especially B2 and B3) for these lead times, as indicated by the many orange points with values between 60 and 100 kt. In contrast, most of the outlier points associated with B3 are less than 60 kt.

The frequency of superior performance results in Fig. 12 provide a different take on the performance of E2. In particular, this figure suggests that E2 may have performed better than B1 for many lead times. The comparisons of E2 to B2 and B3 (Figs. 9b,c) suggest E2 was superior for some of the middle lead times (36–72 h), but the differences are generally not significant statistically. However, for the longest lead times (108–120 h), B2 and B3 tended to have superior performance more frequently than E2. This result may be related to the differences in variability in the distributions of errors for these lead times (as noted in the discussion of Fig. 11).

These results are consistent with the results shown in Fig. 13, which shows the ranking of E2 performance versus the performance of the three top-flight models. As shown in Fig. 13, E2 had the best performance for several lead times between 36 and 84 h, and the percent of cases for which E2 was best was notably larger than 25%. However, E2 also frequently ranked fourth (worst) for these lead times. Thus,

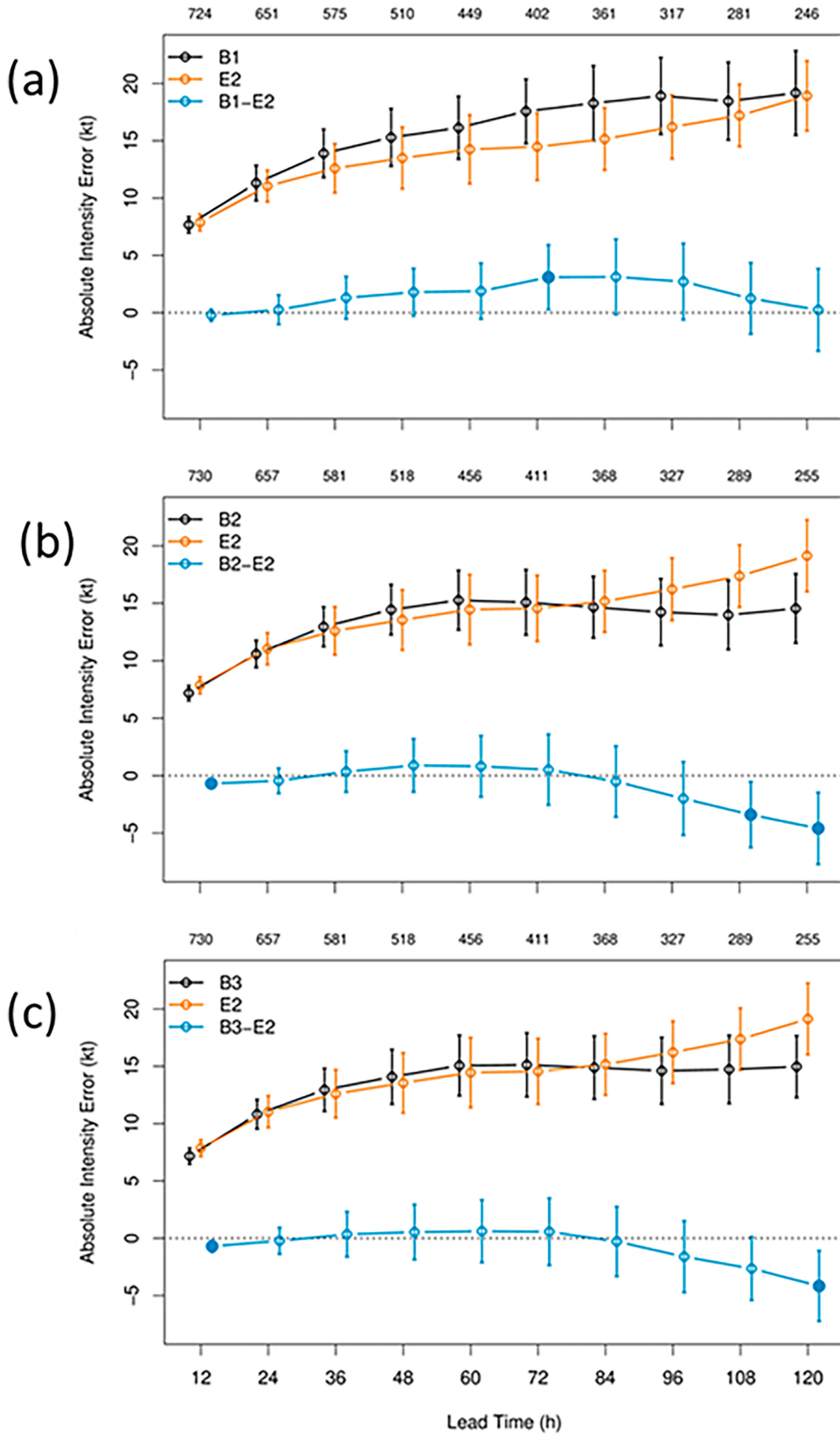


FIG. 9. As in Fig. 3, but for E2 comparisons with the three baseline models.

Forecast Hour	12	24	36	48	60	72	84	96	108	120
B1	724	651	575	510	449	402	361	317	281	246
	-0.2	0.3	1.3	1.8	1.9	3.1	3.1	2.7	1.2	0.2
	-3%	2%	9%	12%	12%	18%	17%	14%	7%	1%
	0.570	0.309	0.831	0.909	0.870	0.968	0.937	0.890	0.568	0.108
B2	730	657	581	518	456	411	368	327	289	255
	-0.7	-0.4	0.4	0.9	0.8	0.5	-0.5	-2.0	-3.4	-4.6
	-10%	-4%	3%	6%	5%	3%	-3%	-14%	-24%	-32%
	0.999	0.575	0.300	0.552	0.451	0.261	0.255	0.778	0.980	0.996
B3	730	657	581	518	456	411	368	327	289	255
	-0.7	-0.2	0.3	0.5	0.6	0.6	-0.3	-1.6	-2.6	-4.2
	-10%	-2%	3%	4%	4%	4%	-2%	-11%	-18%	-28%
	0.991	0.293	0.269	0.337	0.342	0.297	0.146	0.689	0.939	0.992

FIG. 10. Summary table showing comparisons of E2 performance to the performance of baseline models B1, B2, and B3, as in Fig. 4.

E2 was most commonly best or worst, rather than in the middle range of performance.

Based on the set of results for E2, NHC decided to not include explicit E2 intensity predictions as part of the forecast demonstration during the subsequent TC forecasting season. This determination was based on the inconsistent behavior of E2 intensity guidance. While E2 did frequently perform the best in comparison to other baseline models, especially for middle lead times, E2 was also frequently the worst performer and had an error profile that included outliers larger in magnitude than all the baseline models. These factors were key indicators that led to NHC’s decision.

4. Discussion and conclusions

The effort described in this paper was truly collaborative, involving NHC forecasters and managers; university and agency-based modeling groups; and statisticians, atmospheric scientists, and analysts at NCAR. Each year, the following activities and collaborations occurred:

- NHC forecasters and managers initiated the process of defining which forecast attributes should be evaluated and clearly specified their ideas, needs, and wishes.
- Statisticians and the evaluation team developed approaches to respond to questions raised by NHC. Each year, they worked with NHC to ensure that the methods would provide meaningful information to aid in NHC selection of models for the summer demonstration, and they implemented and redesigned versions of the targeted approaches until the full team was satisfied with the methods to be applied. Moreover, additional analyses were developed during the evaluation phase and implemented in response to questions or insights that arose.
- The modeling groups created and shared hundreds of retrospective forecasts from their experimental modeling systems, which made the evaluations possible.

The annual HFIP evaluations provided meaningful and comprehensive information about model performance to

NHC. Tailoring the verification to answer questions of relevance to NHC helped ensure that the model improvements would be relevant to the forecasters. Such an effort—to clearly specify the needs of forecasters and clarify how forecast improvements should be measured in a “user-oriented” framework—is rare. This project provides a template for one approach to achieving that goal.

Through the iterative process of method development, the team was able to tailor verification approaches and displays to respond to specific questions by NHC forecasters and managers. This process ensured that the evaluation approaches provided meaningful responses to NHC’s evolving questions and needs, and made it possible to provide enhanced performance information to NHC across the lifetime of the project. Examples of enhancements during the project include tailoring the attributes of the summary tables (e.g., Fig. 4), the application of meaningful statistical inference procedures, and the development of the ranking plots (e.g., Fig. 13). Of course, the methods used here could also be expanded and improved upon. For example, the boxplots presented in Figs. 5 and 11 show the extreme values as individual points and it would be possible for one point to overlap (and hide) another; a display of the extreme points using a distribution curve would provide more information for comparison.

The breadth and depth of the HFIP evaluations were critical to the process of building forecaster confidence in the new modeling capabilities. They broke new ground through the application of some new approaches but also built on methods previously applied by NHC or other researchers. The evaluations utilized a holistic approach, bringing together a body of techniques (some old, some developed through this project) that enabled verification of TC forecasts from multiple perspectives in ways that are consistent with operational objectives. Best practices were applied, including the use of statistical uncertainty measures and the evaluation of paired differences. The evaluations demonstrate the benefits of “going beyond the basics” of verification approaches commonly applied by identifying and addressing questions posed by forecasters, looking beyond the evaluation of mean values, examining the frequency of superior performance, developing

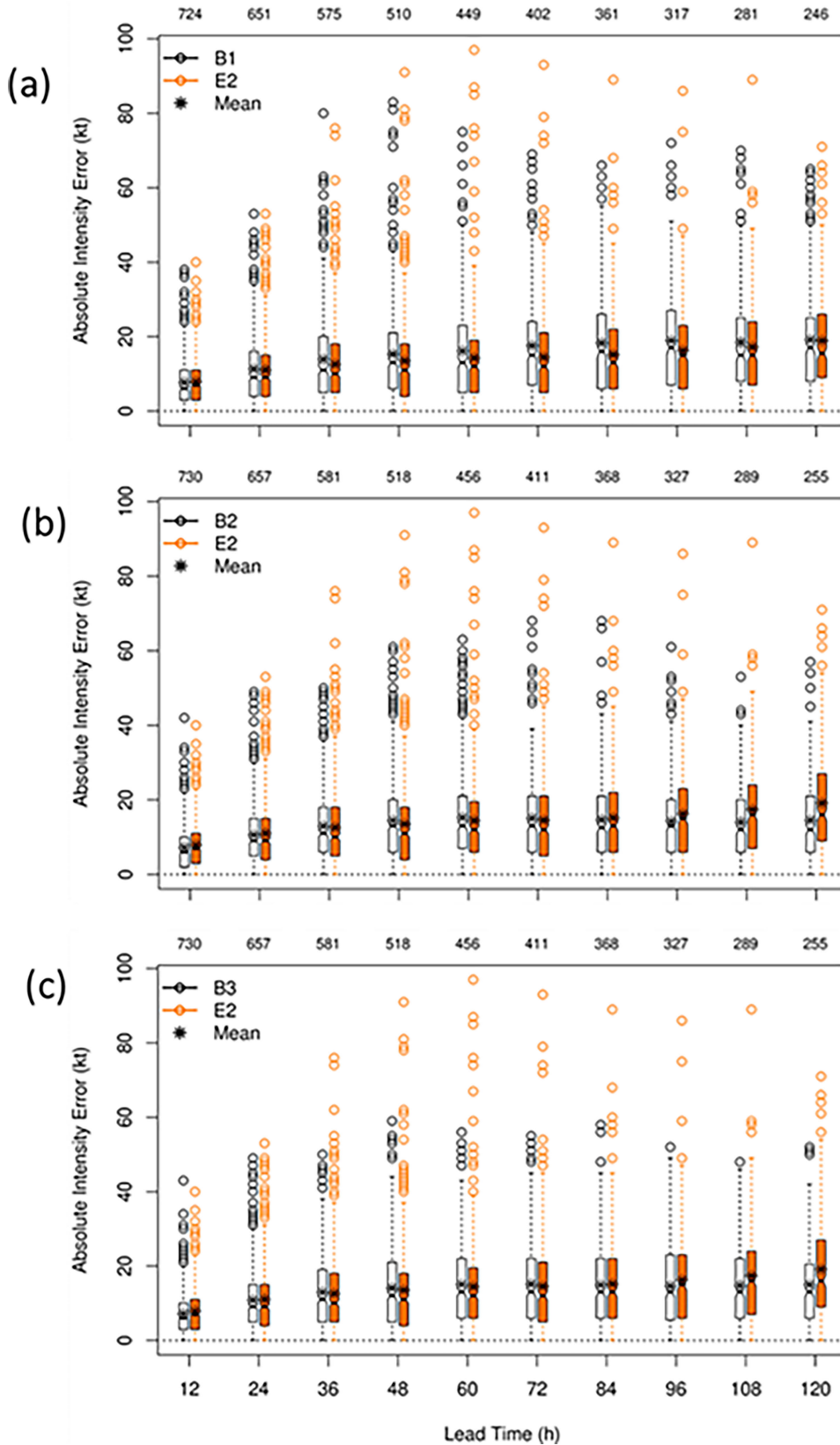


FIG. 11. Boxplots of absolute intensity error results for E2 and B1, B2, and B3, as in Fig. 6 for E1.

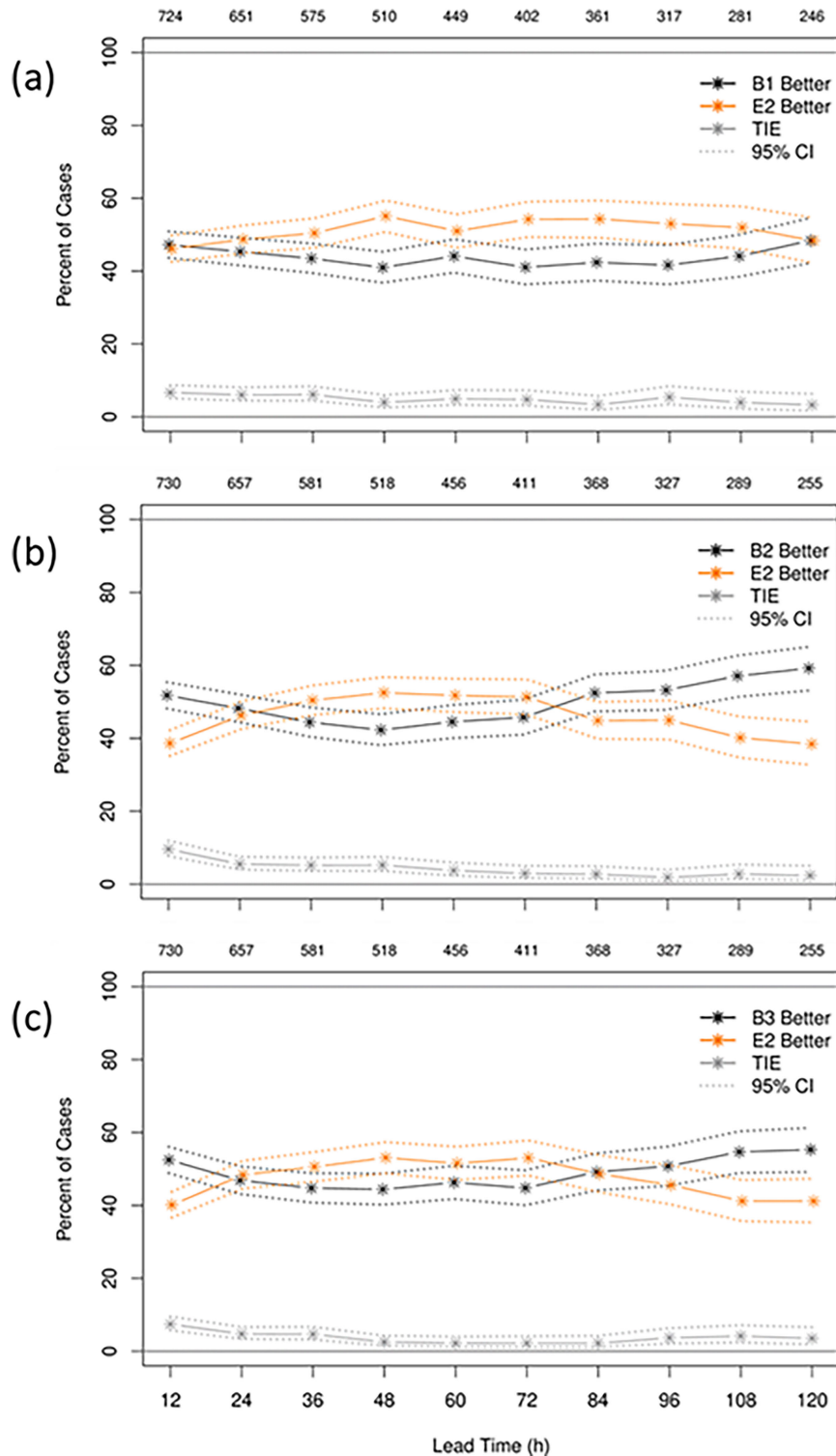


FIG. 12. As in Fig. 7, but for frequency of superior performance for E2 relative to the baseline models.

methods to evaluate the ranking of models, and applying statistical confidence intervals to all comparisons.

This collection of methods was tailored to answer a set of specific questions, rather than focusing on a single measure

(i.e., average performance); these questions were formulated, and specific analyses were developed, to provide meaningful guidance about which models could provide useful information and in what circumstances they could be expected to

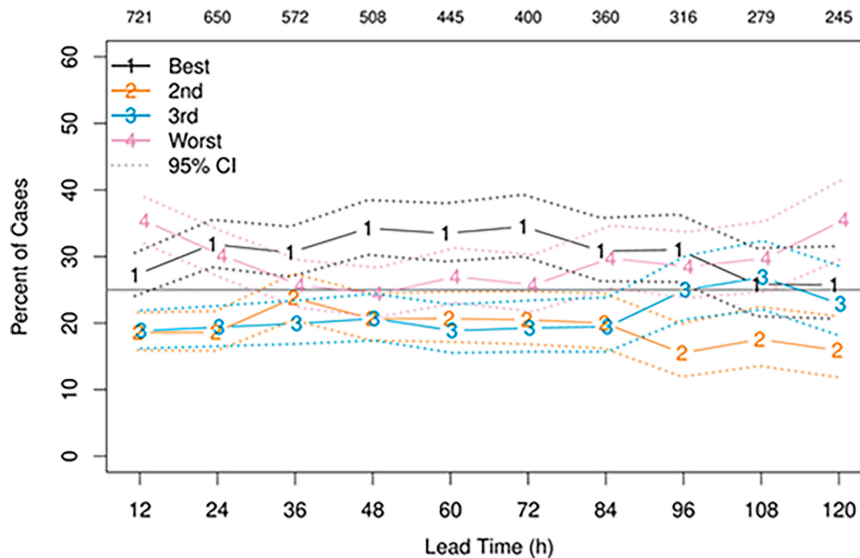


FIG. 13. As in Fig. 8, but for ranking of E2 performance relative to performance of three top-flight forecasting systems.

contribute positively to a forecast. The analyses and results of the evaluations enable the possibility to look more deeply into the circumstances leading to certain kinds of errors (e.g., outliers, poor performance at certain lead times relative to baseline models, etc.). In addition to informing NHC, the evaluations provided extensive feedback to the modeling teams regarding the performance of their models, with specific information regarding aspects of the predictions needing improvement; this information could also be applied by the modeling teams to develop strategies to further improve their modeling systems.

The flexibility inherent in many of the tools that were developed and applied made it possible to adjust aspects of the evaluations—thresholds, significance levels, definition of “ties”—to represent the needs of NHC (and they could easily be adjusted for a different user group). One example shown in section 3 is associated with the frequency of superior performance assessments. While a threshold of 1 kt—a small difference in forecast errors—was identified by NHC as representing a meaningful difference in performance for that evaluation, other thresholds could easily be applied for other applications to, for example, assess the frequency of larger differences, as shown in Fig. 7. Other aspects of the HFIP evaluations could easily be adapted to answer additional questions about the models’ performance. Importantly, the philosophy embraced in this study is readily extendable to other types of forecasts and evaluations, from short-range severe weather predictions to climate outlooks.

Many of the evaluation methods that were developed and applied during the Stream 1.5 verification project have been incorporated into a set of tools (MET-TC) that is included in the enhanced Model Evaluation Tools (METplus; <https://dtcenter.org/community-code/model-evaluation-tools-met/documentation>), an important outcome of the project. METplus is a freely available, community-supported forecast evaluation software package that is developed and supported by the Developmental Testbed Center (DTC; Brown et al.

2021). As a result, the methodologies applied in this study are widely available to the community for application to TC predictions in other contexts.

While the statistical approaches applied in the HFIP evaluations were advanced in comparison to many traditional approaches, a few analyses could be improved or enhanced. For example, a *t* statistic was applied in many of the comparisons. The *t* statistic assumes the underlying data are normally distributed, which generally is not a good assumption for the experimental and baseline error distributions due to their skewed nature, as illustrated by the boxplots in Fig. 5. In contrast (though not shown here) the distributions of paired differences between the E1 and E2 model errors and the baseline model errors are well-fit by a normal distribution, and thus application of the *t* distribution is appropriate. In a simple evaluation, as represented in the top part of Fig. 3, the skewness of the distributions might easily be ignored and the basic *t* distribution applied.

Although the unpaired *t* test is a commonly applied (sometimes naïve) approach that requires the data to satisfy several assumptions (e.g., a normal distribution), a bootstrapping approach does not require such assumptions (Efron and Tibshirani 1994). The use of bootstrap techniques for computing confidence intervals would have avoided the assumption of normality and in some cases perhaps would have provided somewhat more meaningful confidence intervals—especially for the “raw” errors—than those computed using an assumption of normality. Finally, a more detailed examination of extreme errors (e.g., the large errors at the top of the boxplots in Fig. 5) and their characteristics (e.g., which storm they are associated with) could provide greater insight into differences among the models.

Another statistical concern is the large number of comparisons made during the project, which raises the question of “multiplicity” (e.g., Wilks 2006, 2019). This issue focuses on

the fact that with many tests and comparisons, some fraction of the test statistics is likely to be statistically significant simply by chance, and the true statistical significance level is not the assumed value [i.e., when many tests are undertaken in the same study, the true significance level is larger than the assumed value (e.g., 0.05)]. While this consideration is incorporated into many confirmatory statistical studies (e.g., in medical trials), it was ignored in the exploratory HFIP model evaluations.

The kinds of approaches applied here—and user-relevant verification approaches in general—could be used beneficially in many other applications and for evaluations and comparisons of NWP predictions and other types of forecasts (e.g., hydrologic, road weather, and climate predictions) where specific users are identified and involved in defining the verification questions of interest. As in HFIP, these approaches could provide a framework for decision-making and selection of models, as well as the assessment of the benefits associated with many kinds of forecasting systems. However, this kind of framework has not commonly been applied in the development and improvement of weather/climate forecasting systems. As presented and described in this paper, some factors required to apply these types of approaches include

- identification of and coordination with users/decision-makers who can provide information on important factors in their application of the forecasts;
- selection of specific questions that can guide evaluation of these factors;
- identification of valid statistical or other approaches that can be used to answer the selected questions;
- iterative discussions between forecasters, forecast users, and an independent evaluation team.

In summary, this paper presented some new ideas related to the development and application of user-relevant verification approaches, along with statistically valid evaluation methods designed to answer specific questions about forecast performance. The benefits of the approach described include the ability to obtain more meaningful information about forecast quality in the context of users' decision-making processes. Moreover, many of the types of questions and evaluation methods applied here (e.g., boxplot analyses, frequency of superior performance, and so on) would be appropriate for application to many other types of forecasts. Finally, while it is not possible to quantify the contribution of this HFIP project to NHC's forecast operations, NHC was encouraged that improvements in NHC's forecast accuracy occurred during the course of the project.

Acknowledgments. HFIP's Stream 1.5 forecast evaluations involved many individuals and groups, including NHC staff, NWP model development teams at government agencies, universities, and NCAR; their contributions were critical to the success of the project. We are grateful for consistent funding by NOAA's HFIP project office, which provided financial support to the modeling groups and to the NCAR Research Applications Laboratory, Joint Numerical Testbed, Tropical Cyclone

Modeling Team. NOAA funding to NCAR was provided via three grants (NA13AANWG0251, NA19AANWG0219, and NA14AANWG0247). NCAR is sponsored by the National Science Foundation. We are particularly appreciative of Fred Toepfer's vision and leadership of HFIP and foresight in supporting the retrospective evaluations each year. We also greatly appreciate the efforts of NHC staff to identify questions to be addressed about forecast performance that are relevant for TC predictions by NHC and other agencies. In addition to the authors of this paper, many individuals contributed to the forecast evaluations, including Tressa Fowler Smith, Mrinal Biswas, Jonathan Vigh, Chunhua Zhou, Michelle Harrold, and John Halley Gotway. Tressa Fowler Smith played a very important role in the project by leading efforts to define statistically meaningful, often novel, approaches to apply in the statistical evaluations of the forecasts, which made it possible to respond to a variety of questions about forecast performance posed by NHC staff. We also are grateful to the team of METplus scientists and engineers who have incorporated the methods applied in this study into the METplus verification package (<https://dtcenter.org/community-code/metplus>). Finally, we are very appreciative of the efforts by the many numerical modeling groups who contributed experimental forecasts to the project, including Colorado State University, Florida State University, NOAA's Environmental Modeling Center, NOAA's Geophysical Fluid Dynamics Laboratory, the Naval Research Laboratory, NCAR's Mesoscale and Microscale Meteorology Laboratory, NOAA's Global Systems Laboratory, The Pennsylvania State University, State University of New York at Albany, and University of Wisconsin–Madison. Individuals at those institutions made the HFIP forecast evaluation effort possible through their annual provision of experimental forecasts.

Data availability statement. Data used in this paper are available at <https://doi.org/10.5281/zenodo.7804230>.

REFERENCES

- Bender, M. A., I. Ginis, R. Tuleya, B. Thomas, and T. Marchok, 2007: The operational GFDL coupled hurricane–ocean prediction system and summary of its performance. *Mon. Wea. Rev.*, **135**, 3965–3989, <https://doi.org/10.1175/2007MWR2032.1>.
- Bostrom, A., R. E. Morss, J. Demuth, H. Lazrus, and J. K. Lazo, 2018: Eyeing the storm: How residents of coastal Florida see hurricane forecasts and warnings. *Int. J. Disaster Risk Reduct.*, **30**, 105–119, <https://doi.org/10.1016/j.ijdr.2018.02.027>.
- Brown, B. G., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, <https://doi.org/10.1175/BAMS-D-19-0093.1>.
- Cangialosi, J. P., 2022: National Hurricane Center forecast verification report: 2021 hurricane season. NHC Tech. Rep., 76 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2021.pdf.
- , E. Blake, M. DeMaria, A. Penny, A. Latta, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecasting*, **35**, 1913–1922, <https://doi.org/10.1175/WAF-D-20-0059.1>.

- Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey, 1983: *Graphical Methods for Data Analysis*. Wadsworth and Brook/Cote Publishing Company, 410 pp.
- Chen, G. M., H. Yu, Q. Cao, and Z. H. Zeng, 2013: The performance of global models in TC track forecasting over the western North Pacific from 2010 to 2012. *Trop. Cyclone Res. Rev.*, **2**, 149–158, <https://doi.org/10.6057/2013TCRR03.02>.
- Chen, Y., E. E. Ebert, K. J. E. Walsh, and N. E. Davidson, 2013: Evaluation of TRMM 3B42 precipitation estimates of tropical cyclone rainfall using PACRAIN data. *J. Geophys. Res. Atmos.*, **118**, 2184–2196, <https://doi.org/10.1002/jgrd.50250>.
- , —, N. E. Davidson, and K. J. E. Walsh, 2018: Application of contiguous rain area (CRA) methods to tropical cyclone rainfall forecast verification. *Earth Space Sci.*, **5**, 736–752, <https://doi.org/10.1029/2018EA000412>.
- DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.*, **137**, 68–82, <https://doi.org/10.1175/2008MWR2513.1>.
- , and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, [https://doi.org/10.1175/1520-0434\(1994\)009<0209:ASHIPS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2).
- Ebert, E. E., and Coauthors, 2015: Numerical prediction of the earth system: Cross-cutting research on verification techniques. *Seamless Prediction of the Earth System: From Minutes to Months*, WMO 1156, G. Brunet, S. Jones, and P. M. Ruti, Eds., World Meteorological Organization, 403–418.
- , and Coauthors, 2018: The WMO challenge to develop and demonstrate the best new user-oriented forecast verification metric. *Meteor. Z.*, **27**, 435–440, <https://doi.org/10.1127/metz/2018/0892>.
- Efron, B., and R. J. Tibshirani, 1994: *Introduction to the Bootstrap*. Chapman and Hall, 456 pp.
- Franklin, J. L., C. J. McAdie, and M. B. Lawrence, 2003: Trends in track forecasting for tropical cyclones threatening the United States, 1970–2001. *Bull. Amer. Meteor. Soc.*, **84**, 1197–1204, <https://doi.org/10.1175/BAMS-84-9-1197>.
- Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.*, **94**, 329–343, <https://doi.org/10.1175/BAMS-D-12-00071.1>.
- Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, <https://doi.org/10.1175/MWR-D-12-00254.1>.
- Lazo, J. K., and D. M. Waldman, 2011: Valuing improved hurricane forecasts. *Econ. Lett.*, **111**, 43–46, <https://doi.org/10.1016/j.econlet.2010.12.012>.
- Marchok, T., R. Rogers, and R. Tuleya, 2007: Validation schemes for tropical cyclone quantitative precipitation forecasts: Evaluation of operational models for U.S. landfalling cases. *Wea. Forecasting*, **22**, 726–746, <https://doi.org/10.1175/WAF1024.1>.
- McGill, R., J. W. Tukey, and W. A. Larsen, 1978: Variations of box plots. *Amer. Stat.*, **32**, 12–16, <https://doi.org/10.2307/2683468>.
- Morss, R. E., J. K. Lazo, B. G. Brown, H. E. Brooks, P. T. Ganderton, and B. N. Mills, 2008: Societal and economic research and applications for weather forecasts: Priorities for the North American THORPEX program. *Bull. Amer. Meteor. Soc.*, **89**, 335–346, <https://doi.org/10.1175/BAMS-89-3-335>.
- Moskaitis, J. R., 2008: A case study of deterministic forecast verification: Tropical cyclone intensity. *Wea. Forecasting*, **23**, 1195–1220, <https://doi.org/10.1175/2008WAF2222133.1>.
- Murphy, A. H., B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting*, **4**, 485–501, [https://doi.org/10.1175/1520-0434\(1989\)004<0485:DVOTF>2.0.CO;2](https://doi.org/10.1175/1520-0434(1989)004<0485:DVOTF>2.0.CO;2).
- NHC, 2019: NHC track and intensity models. NHC, accessed 31 October 2023, <https://www.nhc.noaa.gov/modelsummary.shtml#:~:text=The%20term%20%22forecast%20model%22%20refers,official%20track%20and%20intensity%20forecasts>.
- Pielke, R. A., Jr., and R. A. Pielke Sr., 1997: *Hurricanes: Their Nature and Impacts on Society*. John Wiley and Sons, 301 pp.
- , J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Musulin, 2008: Normalized hurricane damage in the United States: 1900–2005. *Nat. Hazards Rev.*, **9**, 29–42, [https://doi.org/10.1061/\(ASCE\)1527-6988\(2008\)9:1\(29\)](https://doi.org/10.1061/(ASCE)1527-6988(2008)9:1(29)).
- Powell, M. D., and S. Aberson, 2001: Accuracy of U.S. tropical cyclone landfall forecasts in the Atlantic basin (1976–2000). *Bull. Amer. Meteor. Soc.*, **82**, 2749–2768, [https://doi.org/10.1175/1520-0477\(2001\)082<2749:AOUSTC>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<2749:AOUSTC>2.3.CO;2).
- Rappaport, E. N., 2000: Loss of life in the United States associated with recent Atlantic tropical cyclones. *Bull. Amer. Meteor. Soc.*, **81**, 2065–2074, [https://doi.org/10.1175/1520-0477\(2000\)081<2065:LOLITU>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<2065:LOLITU>2.3.CO;2).
- , and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting*, **24**, 395–419, <https://doi.org/10.1175/2008WAF2222128.1>.
- , J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The joint hurricane test bed: Its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380, <https://doi.org/10.1175/BAMS-D-11-00037.1>.
- Sampson, C. R., and A. J. Schrader, 2000: The automated tropical cyclone forecasting system (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, [https://doi.org/10.1175/1520-0477\(2000\)081<1231:TATCFS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2).
- Simon, A., A. B. Penny, M. DeMaria, J. L. Franklin, R. J. Pasch, E. N. Rappaport, and D. A. Zelinsky, 2018: A description of the real-time HFIP Corrected Consensus Approach (HCCA) for tropical cyclone track and intensity guidance. *Wea. Forecasting*, **33**, 37–57, <https://doi.org/10.1175/WAF-D-17-0068.1>.
- Tukey, J. W., 1977: *Exploratory Data Analysis*. Addison-Wesley, 688 pp.
- Wilks, D. S., 2006: On “field significance” and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, <https://doi.org/10.1175/JAM2404.1>.
- , 2019: *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier, 840 pp.
- WMO, 2013: Verification methods for tropical cyclone forecasts. WMO WWRP Rep. 2013-7, 98 pp., https://filecloud.wmo.int/share/s/Fotf-7H1RLyK1kSC_onBA.
- Yu, H., G. M. Chen, and B. Brown, 2013: A new verification measure for tropical cyclone track forecasts and its experimental application. *Trop. Cyclone Res. Rev.*, **2**, 185–195, <https://doi.org/10.6057/2013TCRR04.01>.