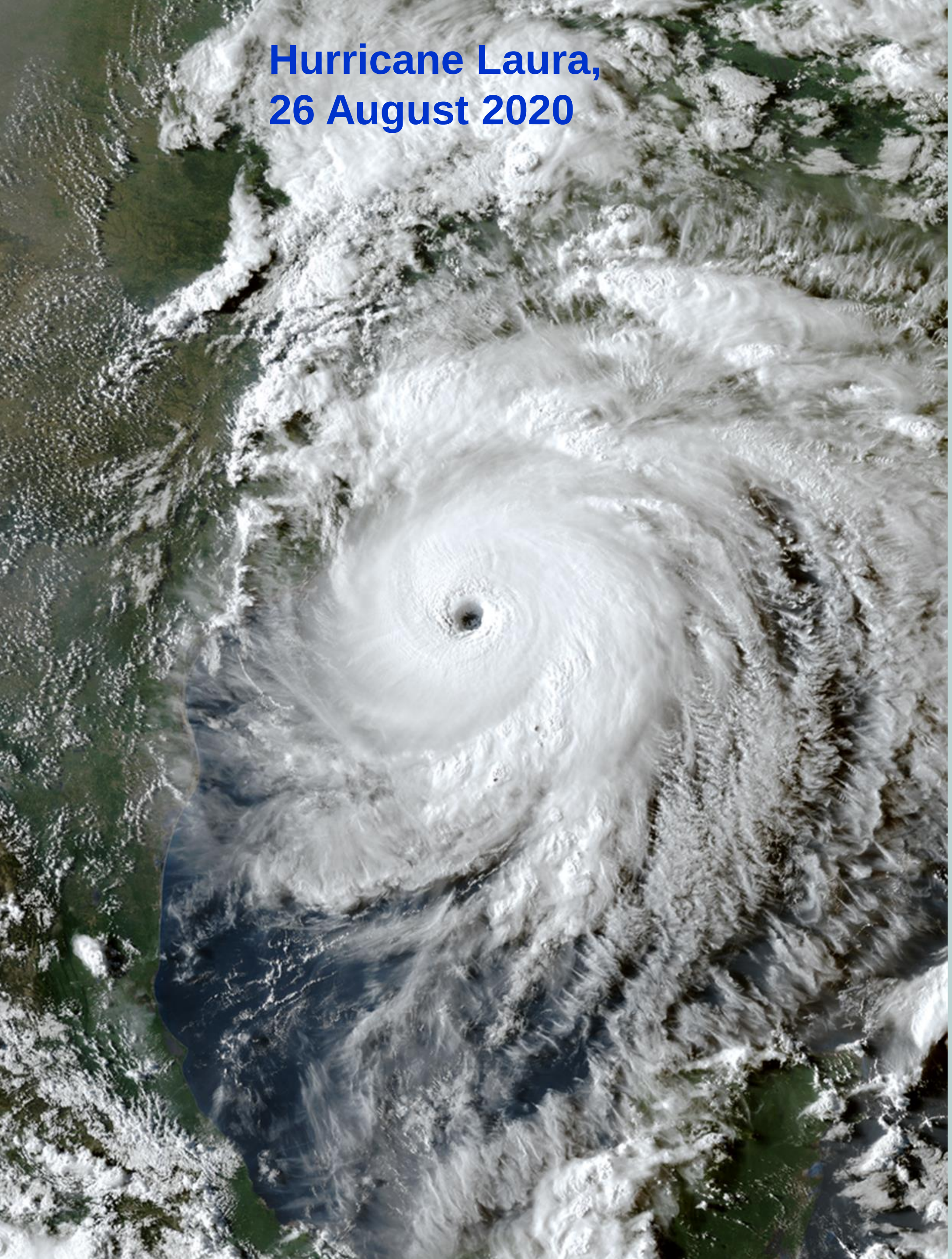



# User-driven evaluation of tropical cyclone predictions

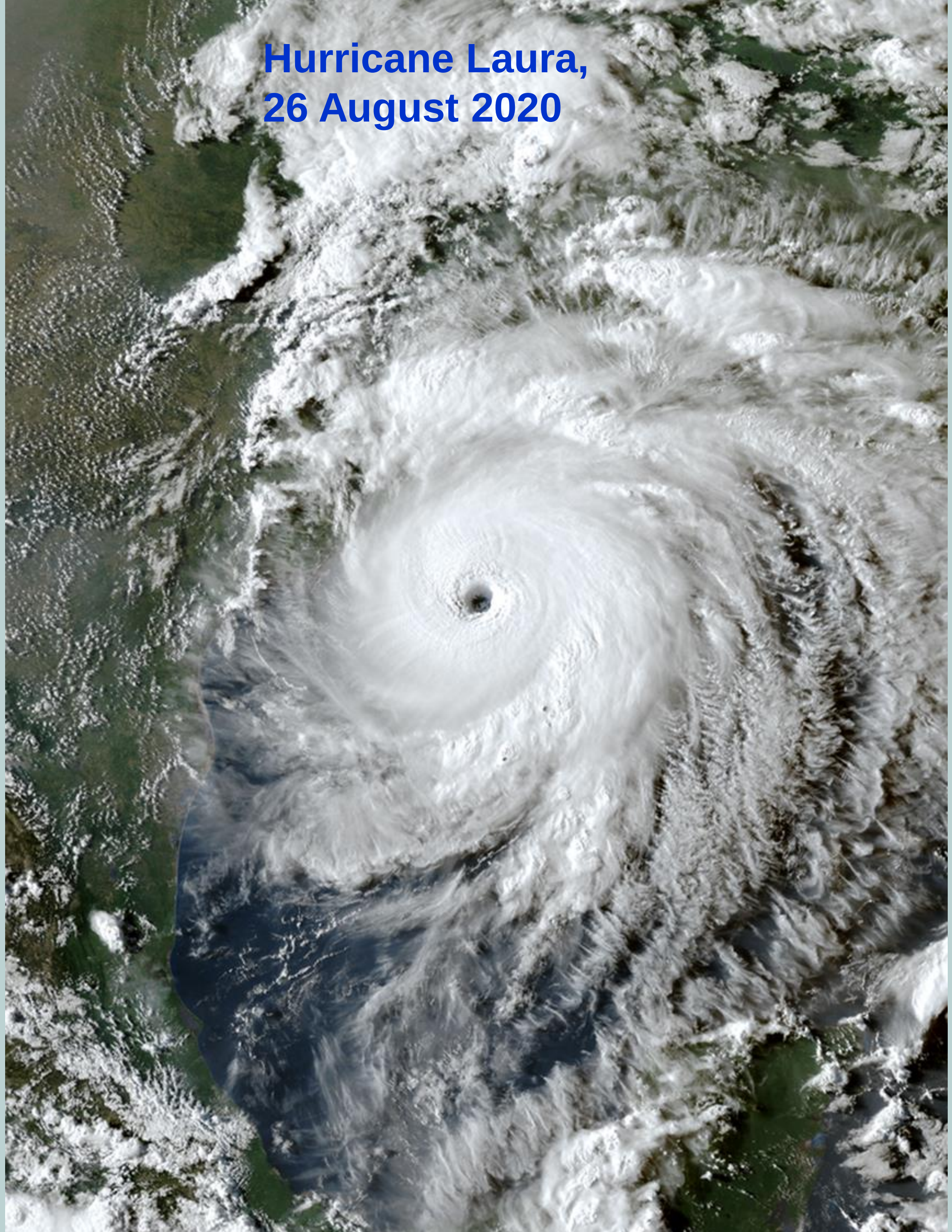
Barbara Brown[1], Louisa Nance[1], and Christopher Williams[2]

[1]National Center for Atmospheric Research, Boulder CO USA
[2]University of Florida Department of Geography

*WMO High Impact Weather Workshop*
*International Verification Methods Workshop Online*

*11 November 2020*

NCAR | RESEARCH APPLICATIONS LABORATORY



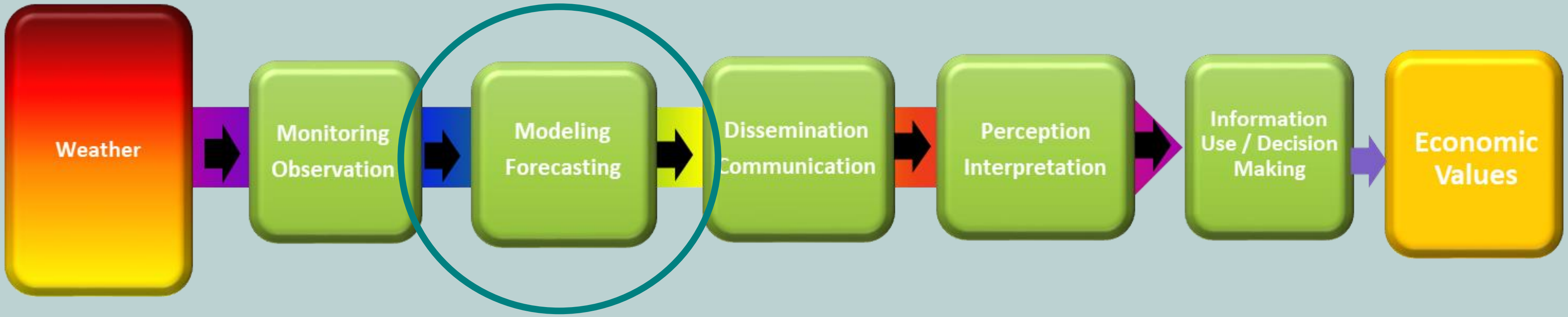

# Hurricane Laura (26 Aug 2020)

- Made landfall at Cameron, Louisiana, at near peak intensity
- Tenth-strongest U.S. hurricane landfall on record
- Led to deaths of at least 42 people in the U.S.
- $14 billion in damage in southwestern Louisiana and southeastern Texas

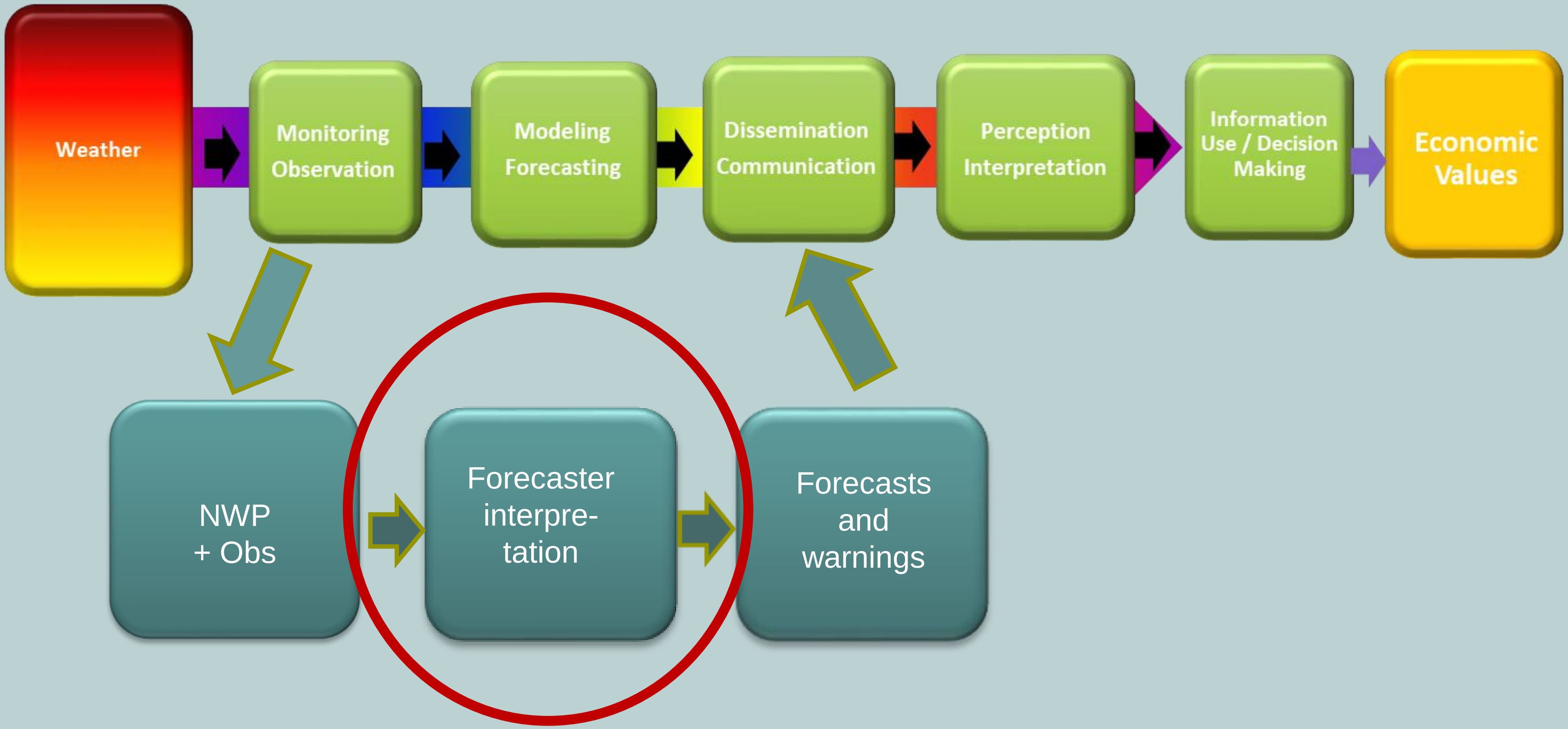
Predictions of Tropical Cyclone (TC) track and **intensity** are important for planning evacuations, protecting life and property

**Goal of this presentation is to consider meaningful – user-driven – ways of evaluating NWP guidance, to aid forecasters in making their predictions of TC intensity**

# Value chain connection (Lazo)



# Value chain connection (Lazo)




***This study considers the question, "How can verification information best inform and facilitate the use of NWP guidance by forecasters?"***

# User-relevant verification

*(Morss et al. BAMS 2008; Ebert et al. Met. Z., 2018)*

- *Level 0*: Focus on single, simple measures (one-size fits all) ("administrative" verification)
- *Level 1*: Broad diagnostic approaches (stratification, thresholds, etc.)
- *Level 2* : Features-based approaches, or more enhanced diagnostic approaches (measure many attributes of forecasts, often from a spatial or temporal perspective)
- *Level 3*: User-relevant verification approaches and measures (*User-driven*)
- *Level 4*: Forecast value estimated via conversion of forecasts to decisions (can follow through whole value chain)

# User-driven/relevant verification approaches…

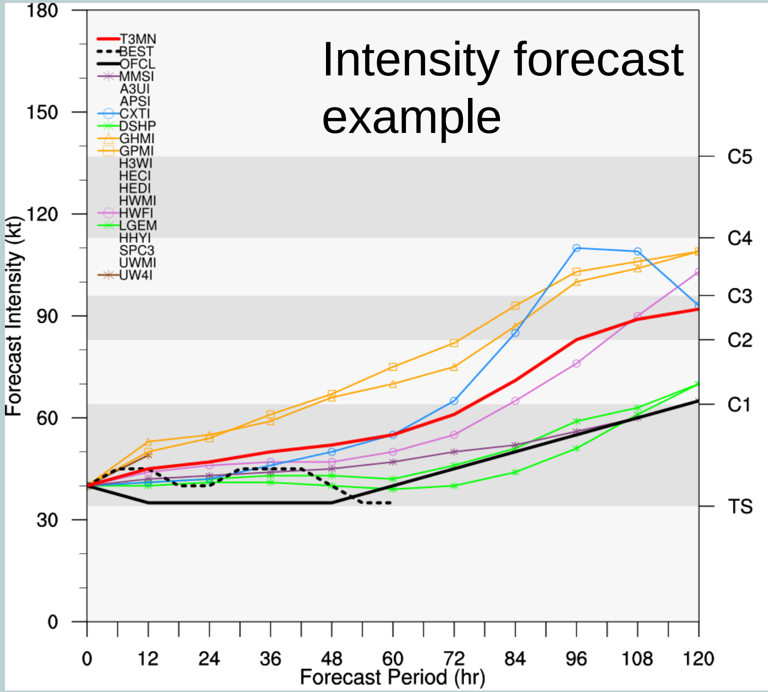
…consider information needs of *specific users* rather than applying a 1-size fits all approach to all forecasts of a specific type (e.g., RMSE or ACC for NWP)

… require understanding users' questions about the quality of the forecasts

# The Hurricane Forecast Improvement Project (HFIP)

- NOAA-funded project initiated in 2007 to significantly improve TC position and intensity predictions
  - Initial goals (first 10 years): *Significant improvements in predicted track and **intensity***
- NCAR project goal

Provide guidance to National Hurricane Center (NHC) to help select *experimental models to demonstrate to **operational forecasters** during each TC season*




***Predictions from demonstrated models must be expected to "do no harm"***

# Approach

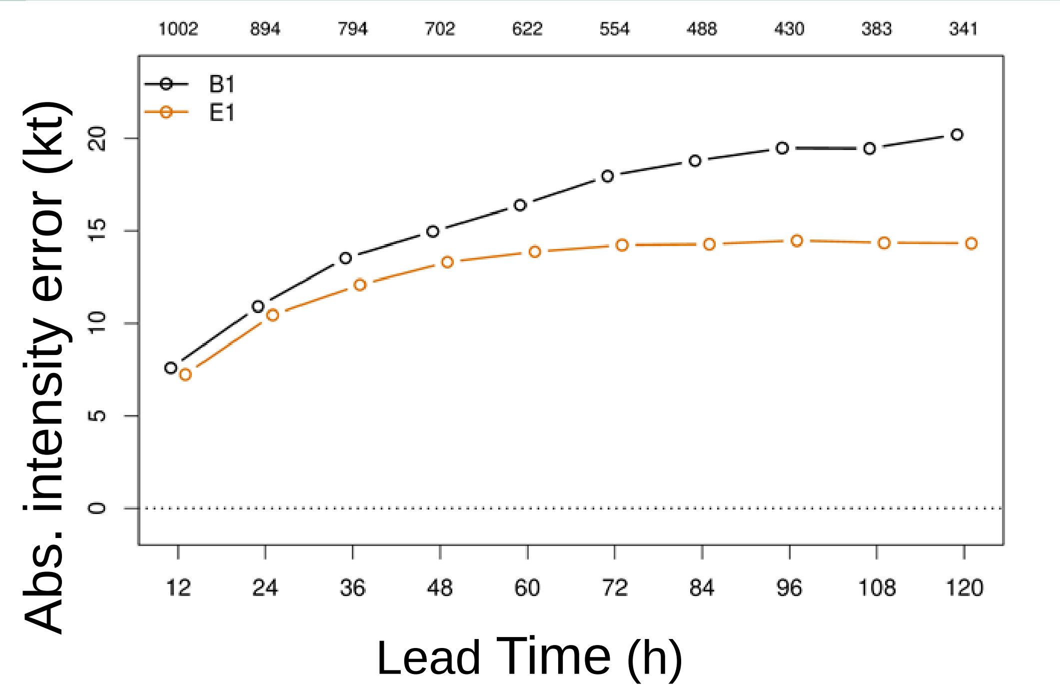
- With NHC staff, identify questions about model performance that *are relevant for their use operationally*
- Develop verification approaches to answer those questions
  - Compare *experimental* model performance to "*baseline*" model performance
  - **Data**: 3 years of retrospective forecasts produced by candidate and baseline forecasting systems
- Models evaluated in spring before start of hurricane season

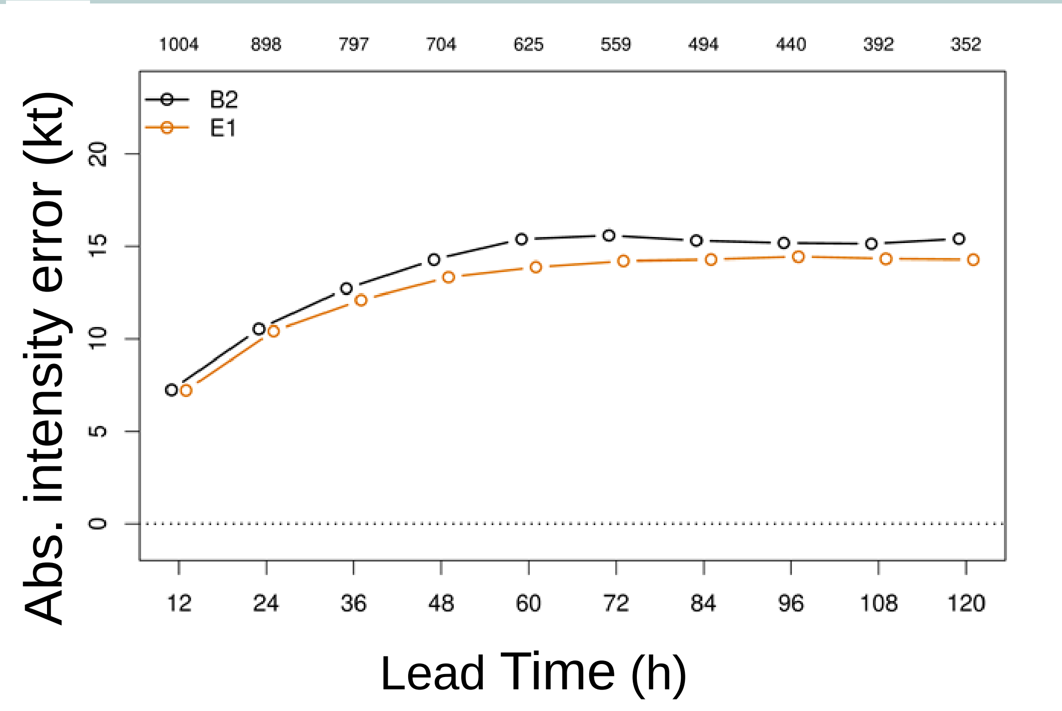
Example questions:

- *Does the experimental forecasting system perform as well or better **on average** than the baseline models?*
- *Does the experimental system have more/less outlier events?*
- *How does the candidate model "rank" with the baseline models?*

*The next slides show an example application for a single candidate model's predictions of hurricane intensity*

## *Does the candidate model (E1) have smaller errors (on average) than the baseline models (B1 and B2)?*

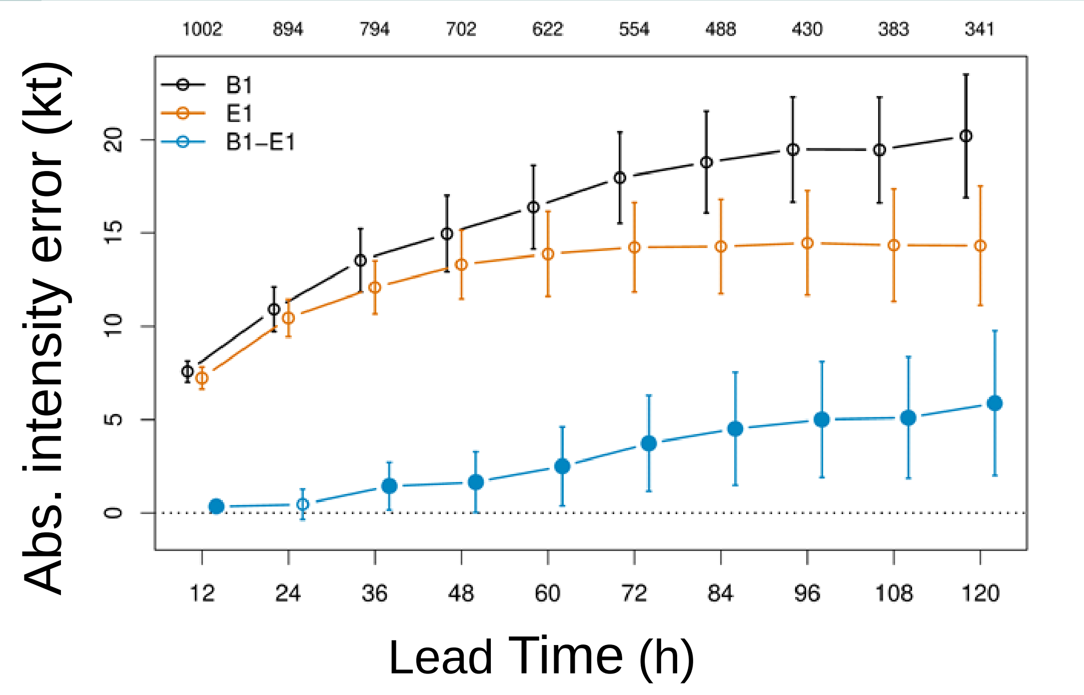


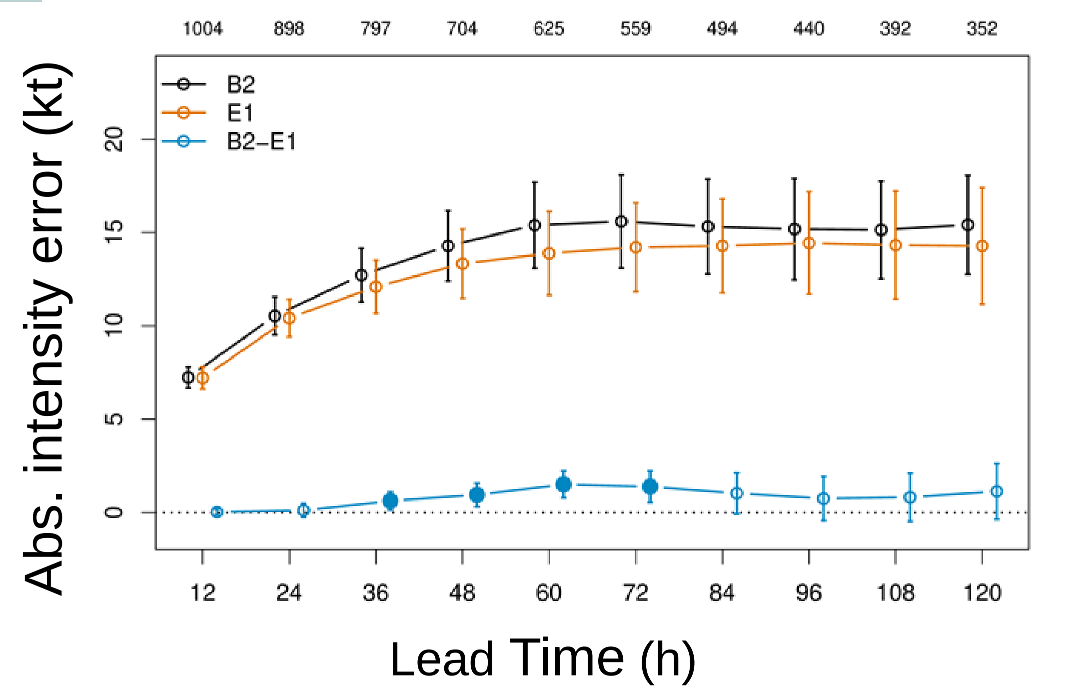




"Traditional" TC intensity verification

*Conclusion*: E1 better than B1 and B2 for all lead times

## *Does the candidate model (E1) have smaller errors (on average) than the baseline models?*

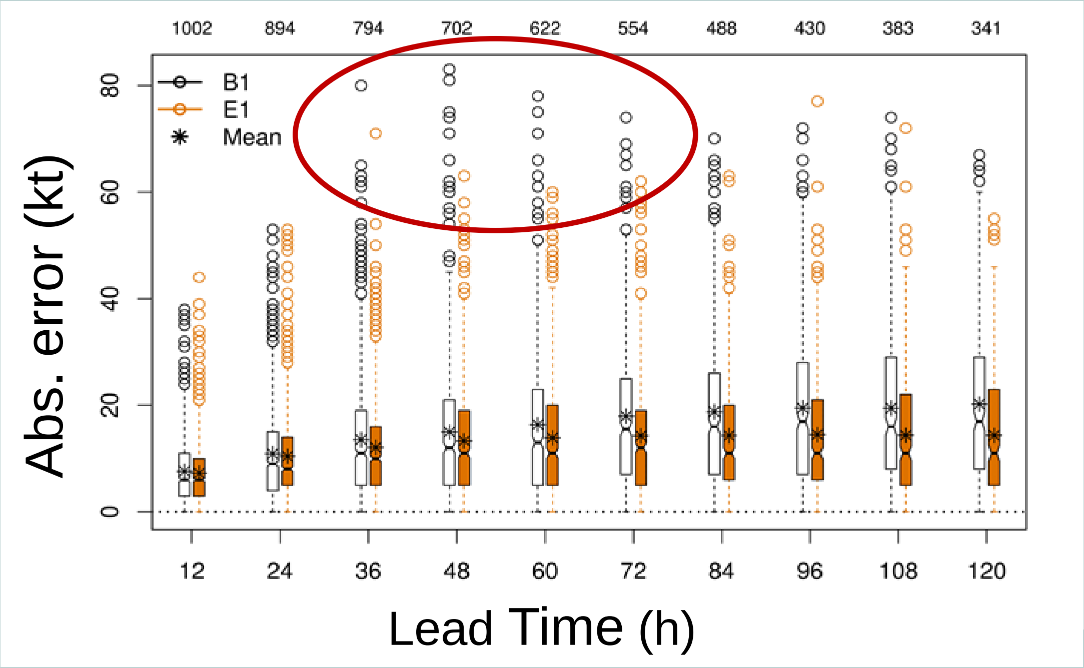


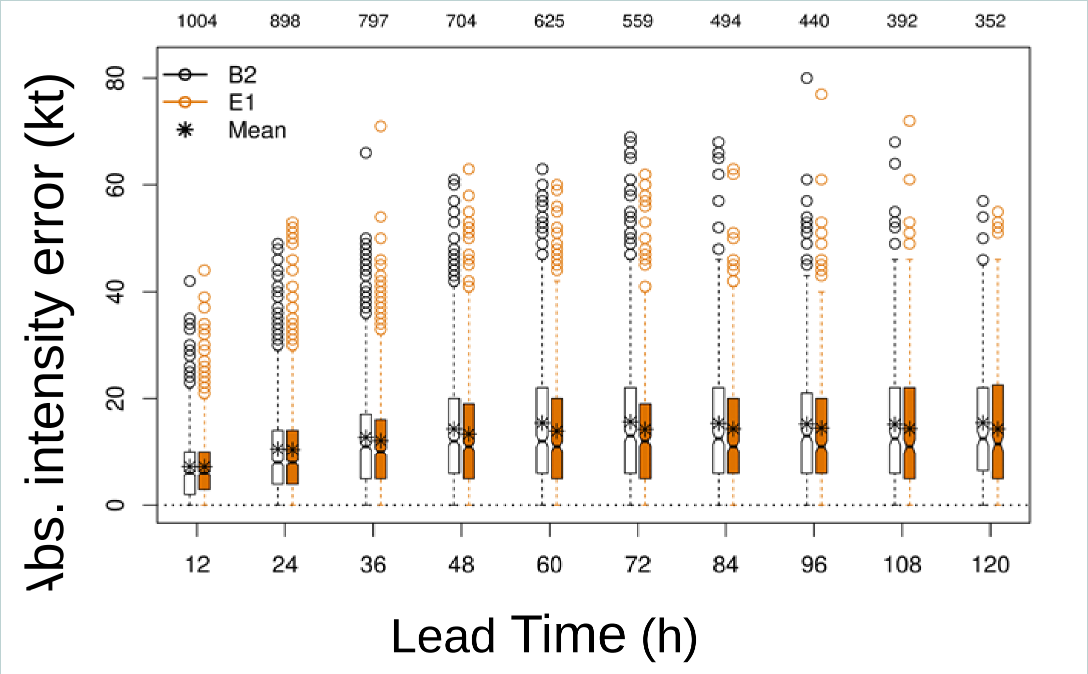




*Pairwise differences* indicate

- Significant differences for *most* lead times relative to Baseline 1
- Significant differences for *some* lead times (36 – 72 h) relative to Baseline 2

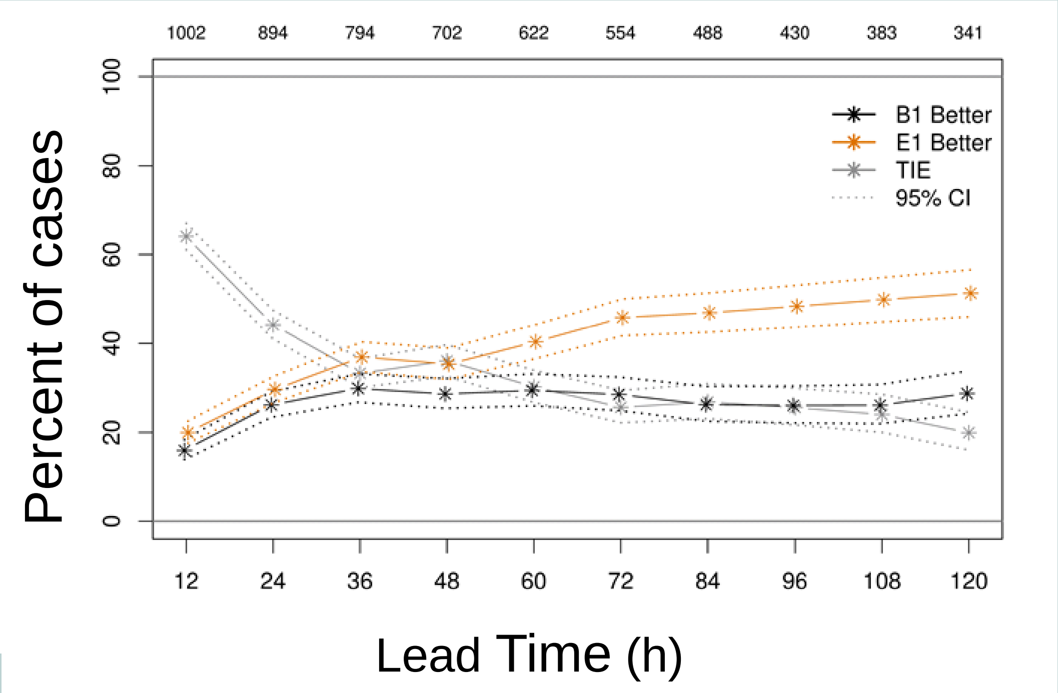
# *Does the candidate model (E1) have fewer <u>large errors</u> than the baseline models?*

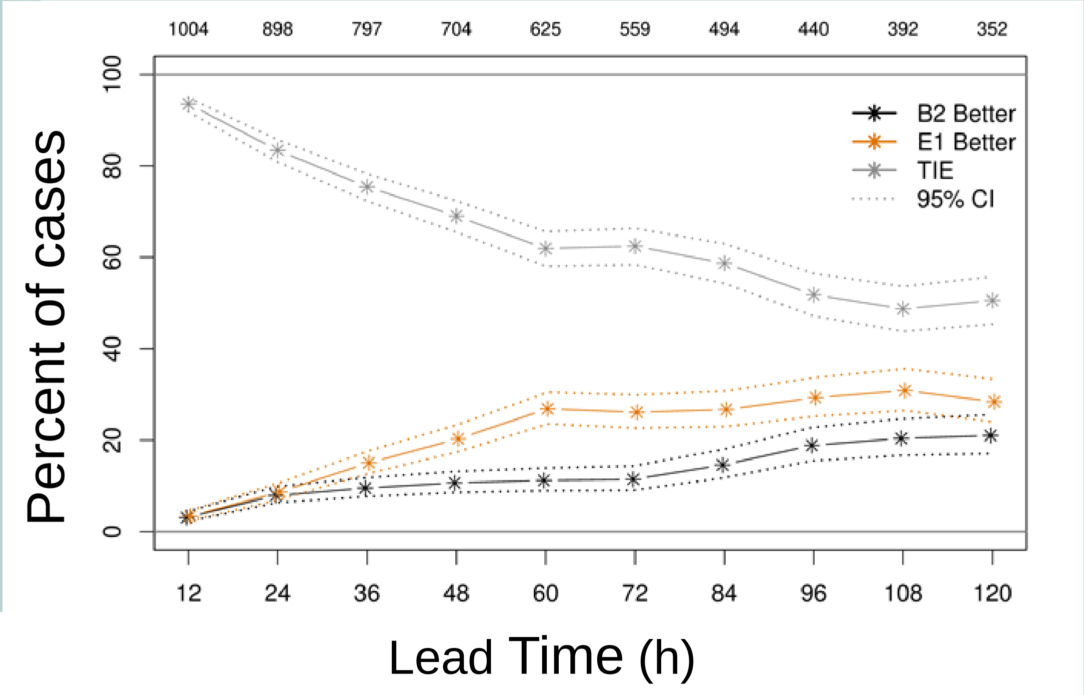



E1 has many *<u>fewer extreme errors</u>* than B1,
but about the same number as B2

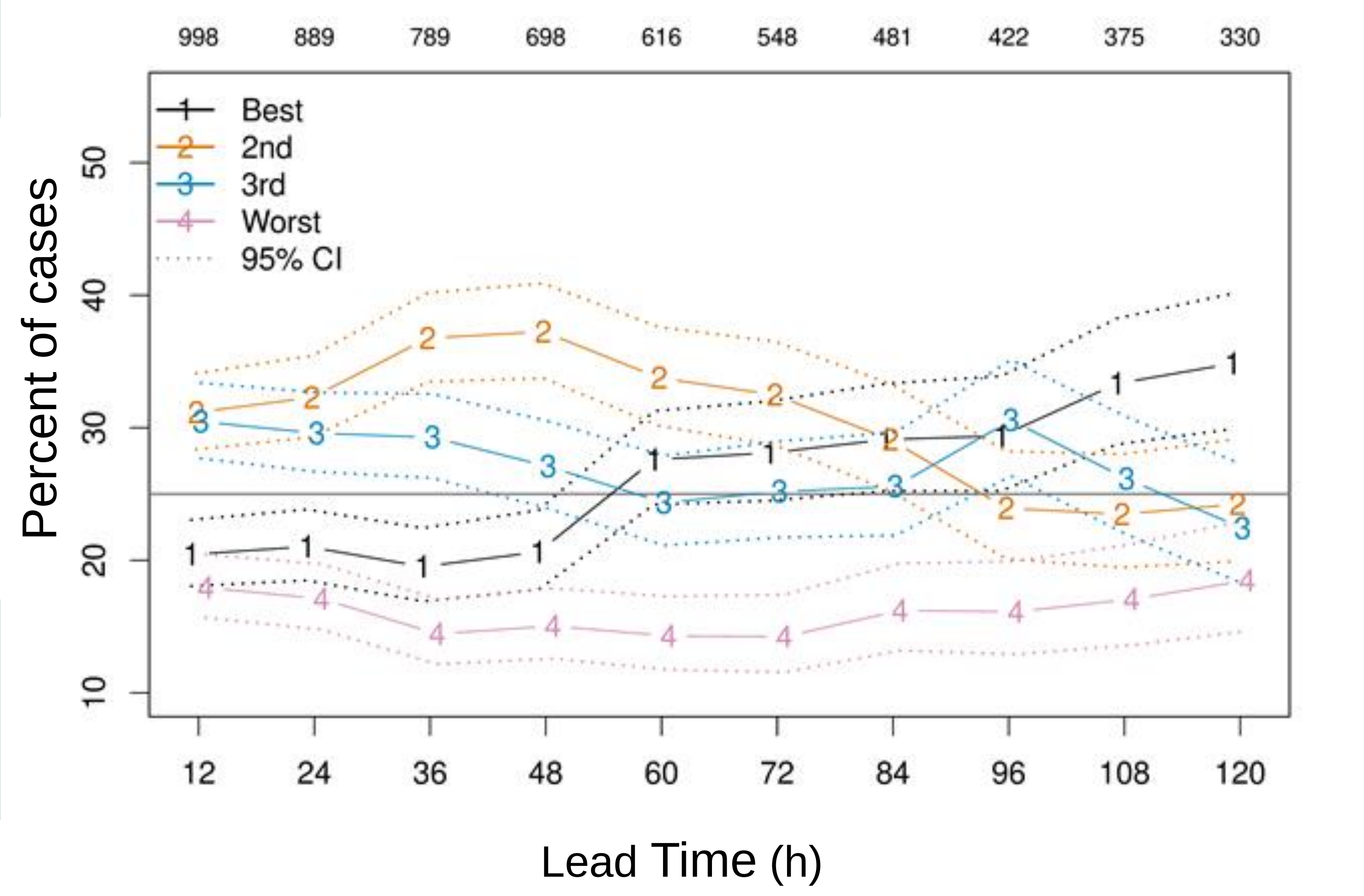
## *How often were the E1 errors smaller (by >=5 kt) than the B1 and B2 errors?*




- E1 was frequently better than B1 for leads of 60 h and longer
- E1 was better than B2 or *tied with B2* for most lead times

## ***How did the E1 forecast rank in comparison to the errors associated with three baseline models?***




- E1 was most frequently *second best* for lead times between 36 and 60 h
- E1 was significantly *best* for 120-h forecasts
- E1 was *worst* for about 15-18% of the cases

# Some conclusions…

- Evaluating uncertainty in verification measures can lead to different (more defensible) decisions
- *User-driven questions* enable strategies to make *rational and meaningful choices* among forecasting systems for specific applications (as demonstrated by this study)
- Simple/standard questions (e.g., about average behavior) may not meet user needs
  - **User-driven** approaches (e.g., *model ranking, score cards, outlier examination*) can provide information that is more meaningful and useful
- The approach applied here – working closely with decision makers – can be a model for other *user-driven verification applications of verification as a component of the value chain*

# Acknowledgements and thanks


- Contributions of many NCAR scientists who worked diligently on development and application of the evaluations for several years
- NHC forecasters and managers who took great interest in identifying the information forecasters care about
- Support from NOAA's HFIP project office, which made these developments possible
- Contributions from the many modeling groups who continuously implemented improvements to their models and provided large sets of forecasts to us for evaluation