AI Topics in Verification and Testbeds
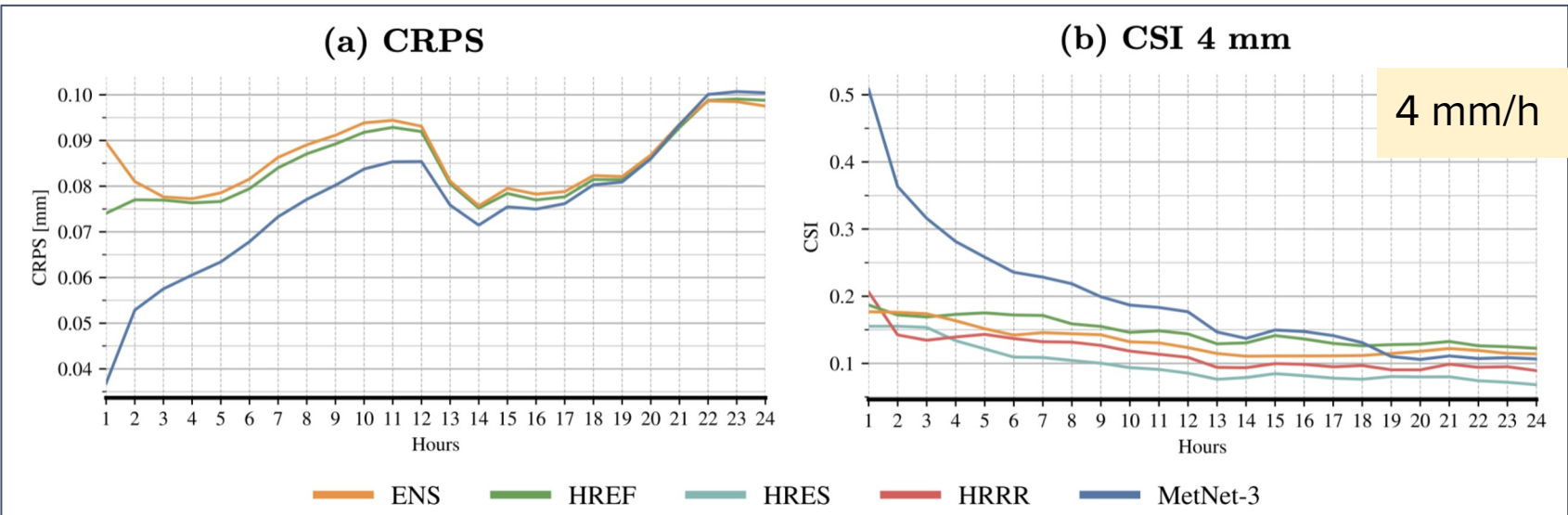
User-driven

AI can do impressive things

We just need to make sure it is what we need!

Need user-driven development

AI Impressive Things

CONUS 1-h precip

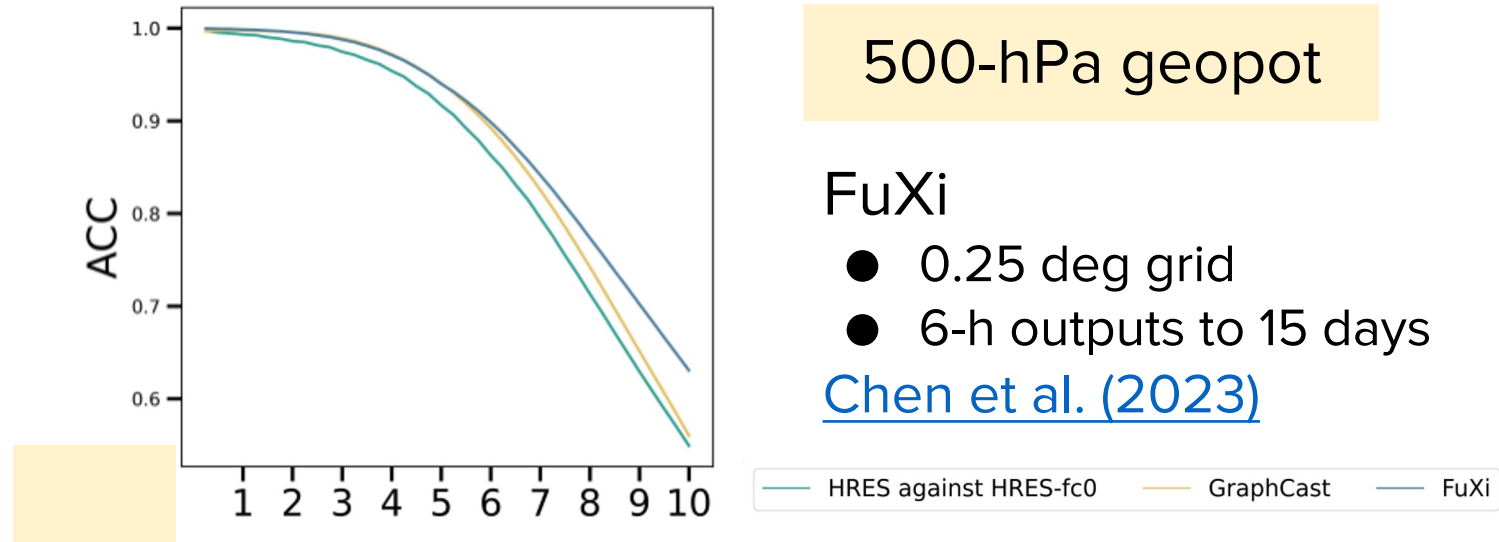### MetNet-3
- 1-km grid
- 2-min outputs to 24h

Andrychowicz et al. (2023)

500-hPa geopot

FuXi
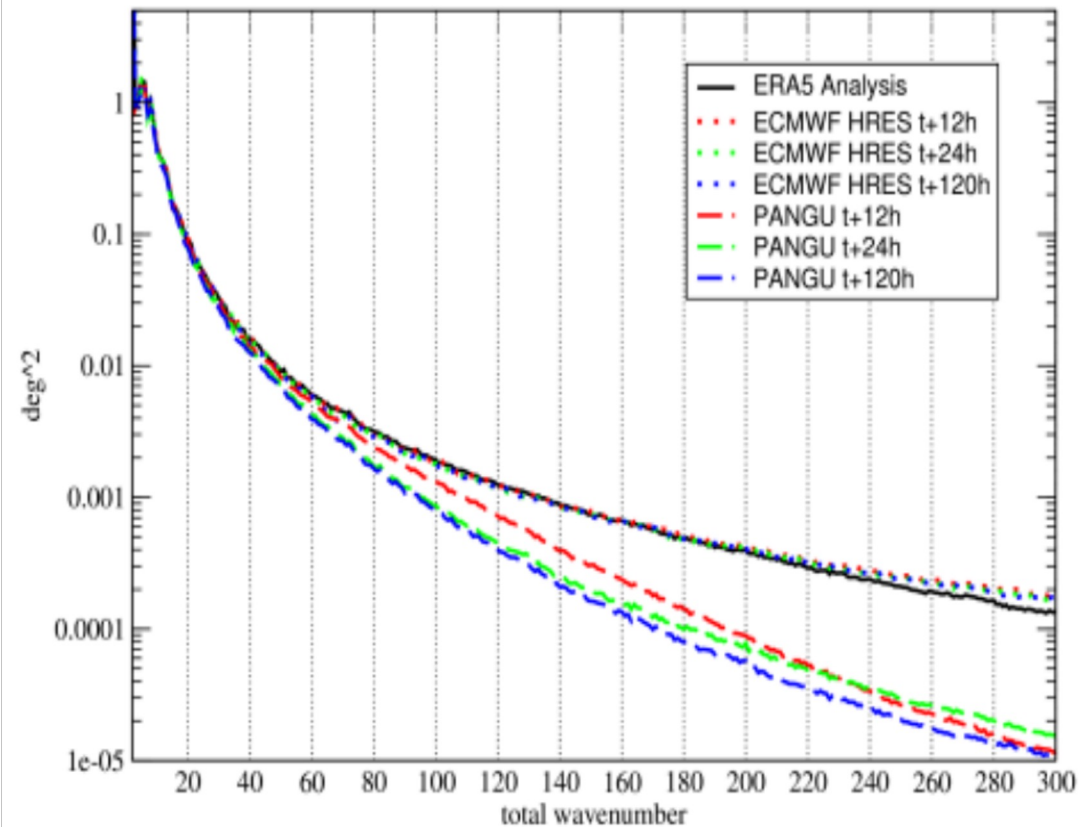- 0.25 deg grid
- 6-h outputs to 15 days

Chen et al. (2023)

Physical Representation

Power spectral density T 850 hPa

(…) Pangu-Weather (…) is not a general-purpose atmosphere simulator or (…) ML-driven atmospheric digital twin.

Bonavita (2023)

Do we care?

Is the Model Trained to Do What **You** Need?

World Climate Service @WorldClimateSvc

A significant change today at last – all three models now showing landfall in Nova Scotia next weekend. If correct, GraphCast will be the winner in this particular case.
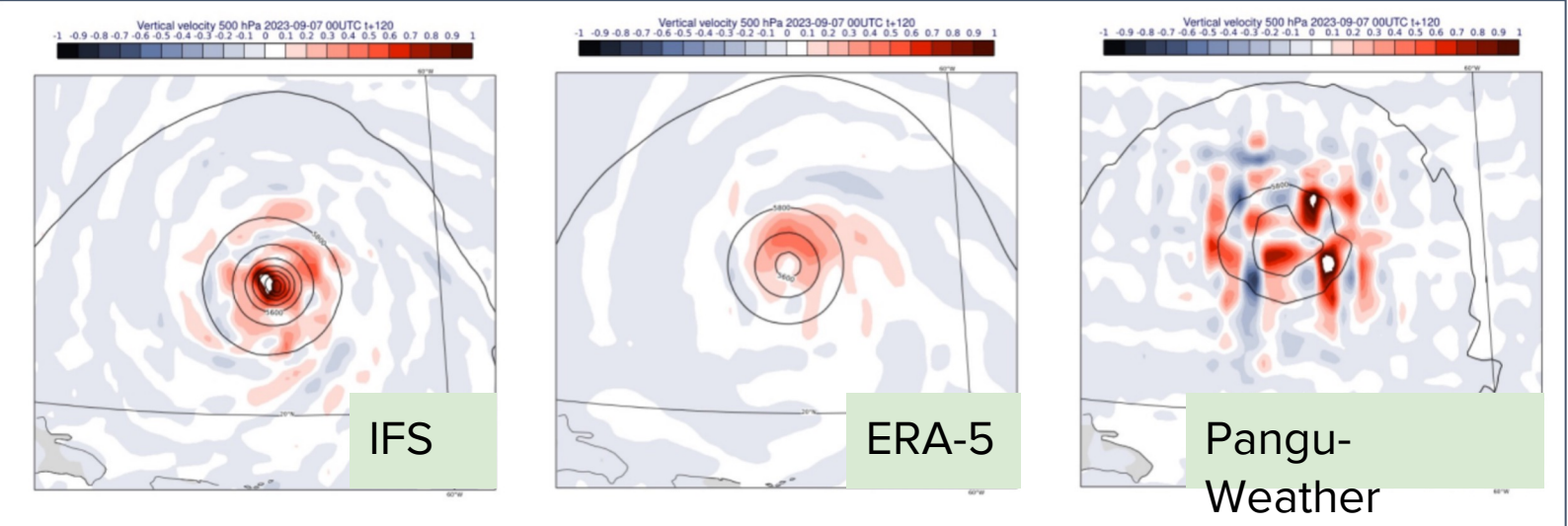
MSLP and 850-hPa winds

Pangu-Weather, GraphCast, FourCast Net

<u>Good</u> predictions for Hurricane Lee location

IFS

ERA-5

Pangu-Weather

Vertical Motion

<u>Problematic</u> (derived) Pangu-Weather field for Hurricane Lee

Bonavita (2023)

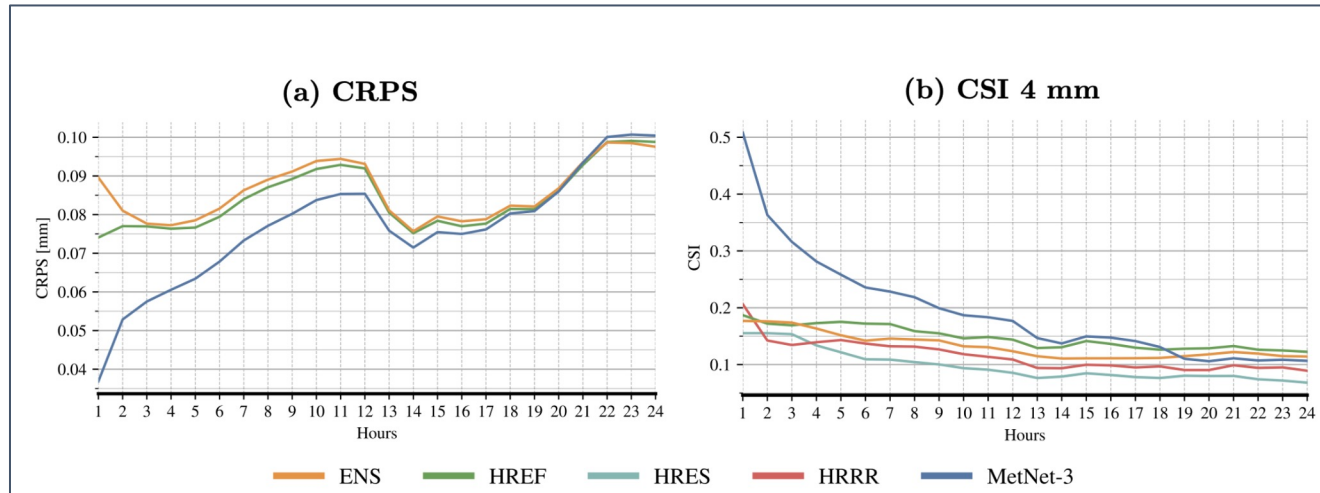Deriving non-predicted quantities leads to questionable results

Other Issues

- **Temporal discontinuity**
  - If models for different lead times are trained independently

- **Low effective resolution (blurry)**
  - Minimize RMSE
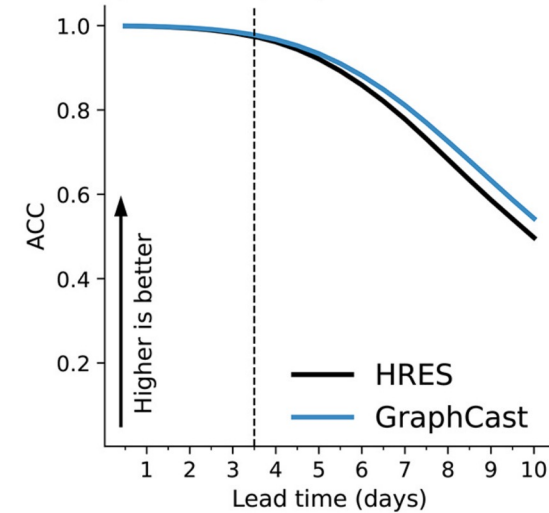
Lead time: Short- and Medium-Range

1-h precip over CONUS
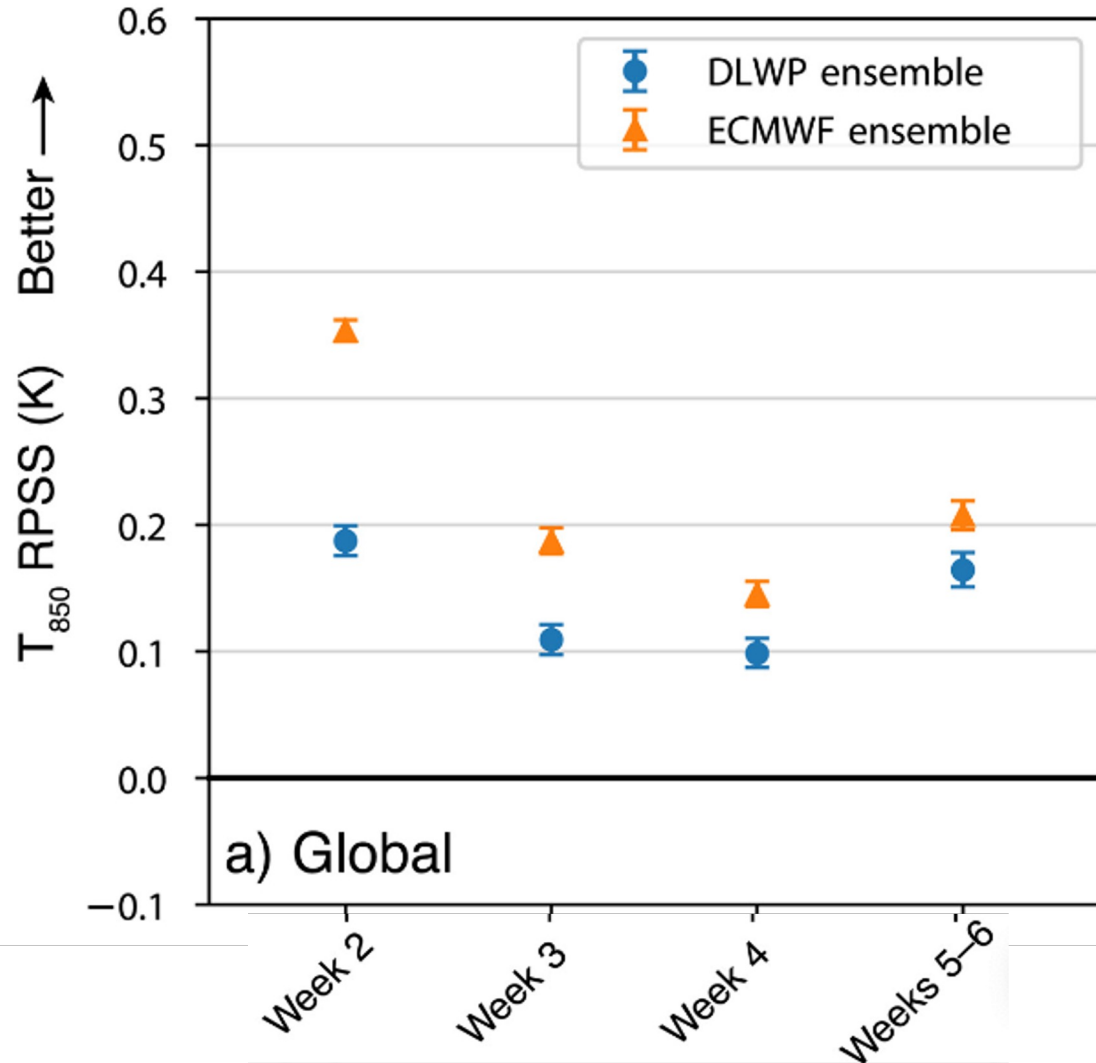
500-hPa geopot

MLWP prediction improves over physical models for the first ~20h

MetNet-3 Andrychowicz et al. (2023)

MLWP improves Z500 ACC 4-10 days

GraphCast Lam et al. (2023)
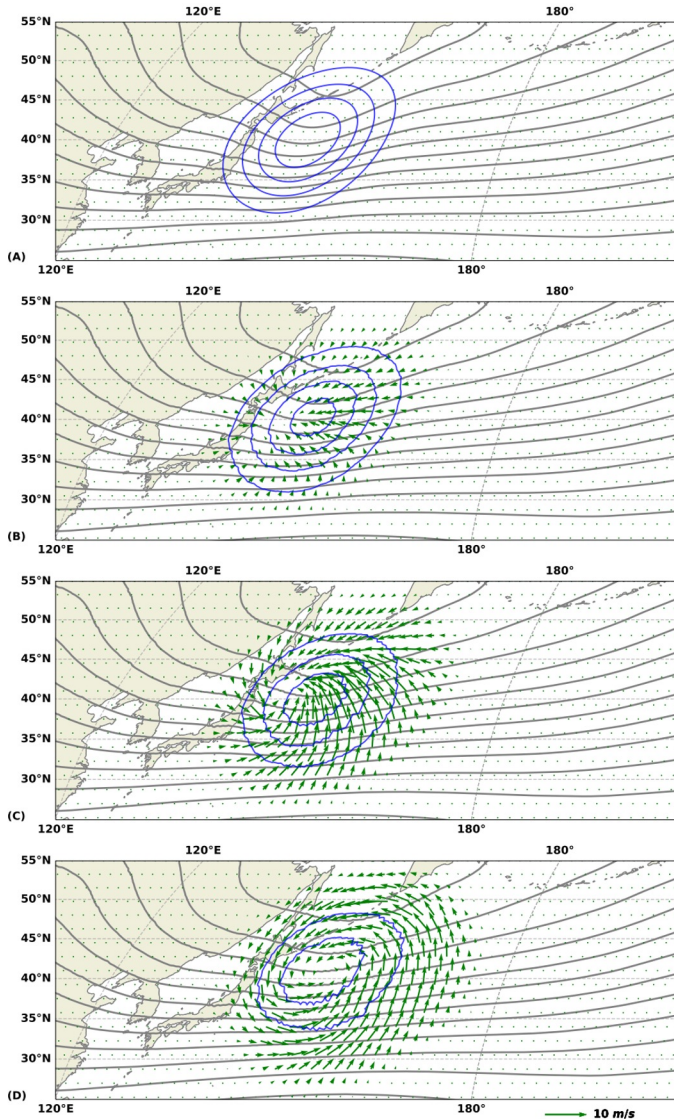
Lead-Time: S2S

a) Global

[Weyn et al. (2021)](#)

This model
- Is worse than ECMWF
- But does not lose skill as fast

No SST: What does it know about predictability that we don't?

AI is not just for operations!

[Hakim and Masanam (2023](#))

 (Submitted to AGU Advances)

Pangu-Weather is used to predict the response to artificially-introduced anomalies

Low 500 geop pert -> geostrophic adjustment

The response is found to be physically consistent

-> Pangu Weather can be used for idealized tests

Verification of MLWP

- Each group verifies their own

- WeatherBench (Rasp et al. 2023) good standardization (but insufficient)

- A few groups verifying others (ECMWF, CIRA, GSL)

- Often using basic metrics (RMSE, ACC)

- Additional metrics are needed

  - Probabilistic forecasts

  - Distribution (extremes)

  - Case studies

  - Etc.

Powered by
**Global Systems Laboratory**
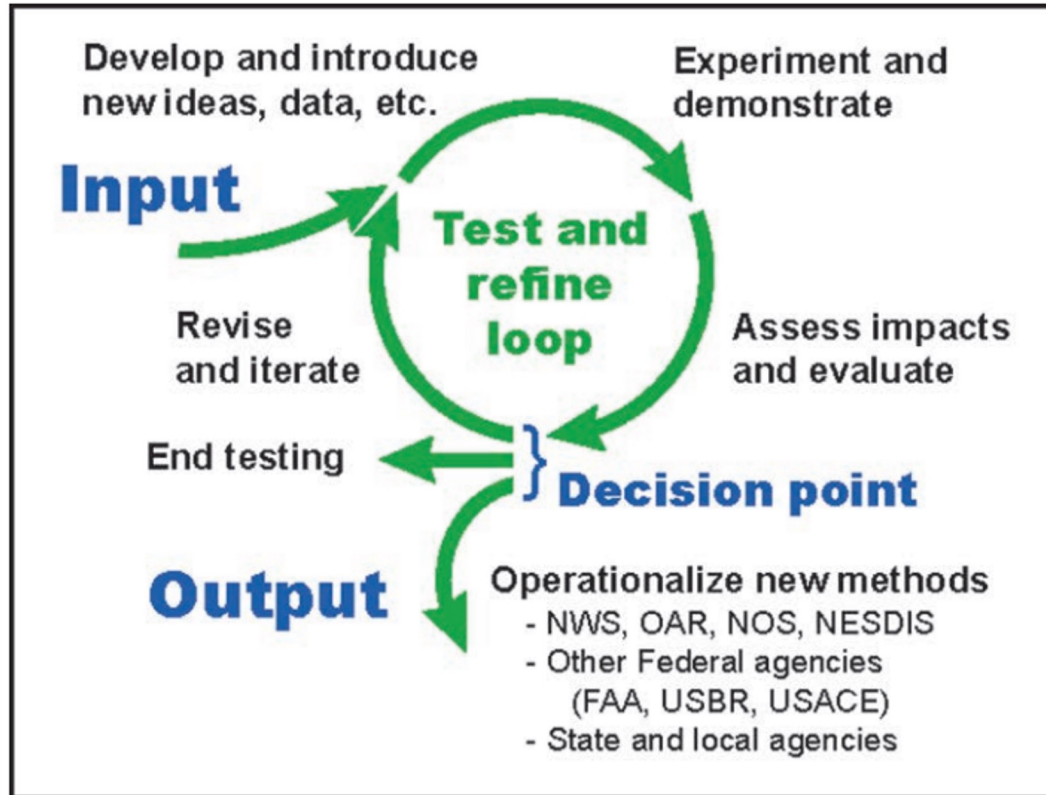
- Independent broker
- Testing at various readiness levels
- Connecting with the community to expand ability to test
- Connection with NWS
- Expanding vx software

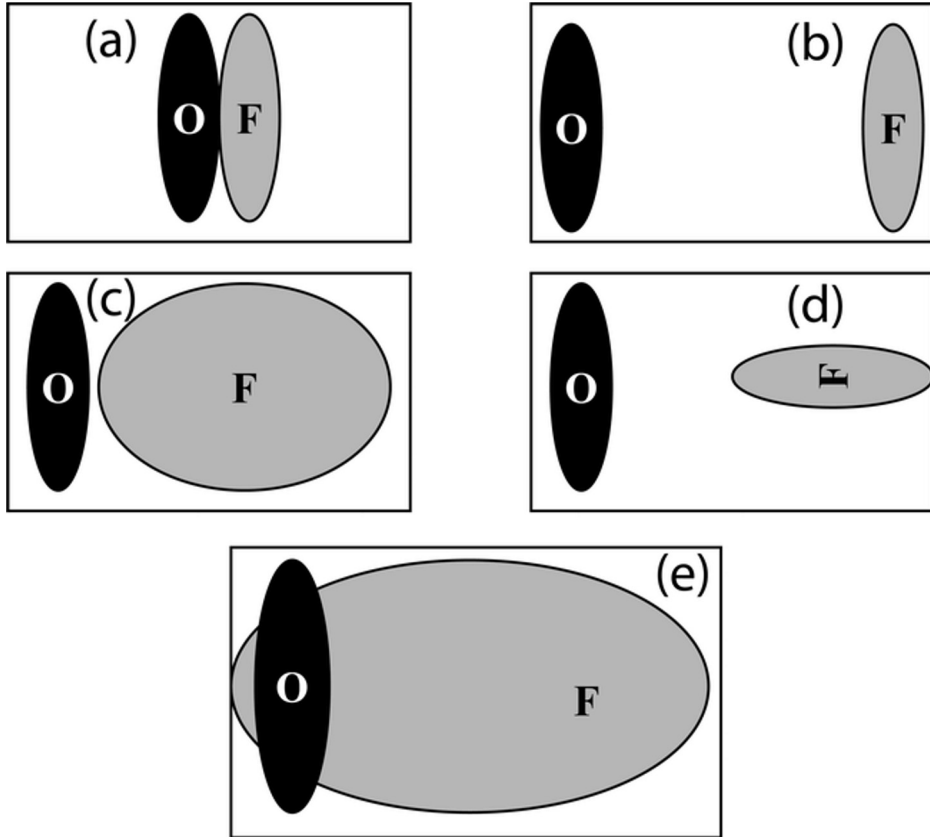Ralph et al. 2013

Examples of Impacts

Test of initial/boundary conditions perturbations for ensemble -> RRFS

Test of various ways of constructing ensemble (time lagging) -> RRFS

Community involvement in development and testing of radiation and cloud representation -> HWRF
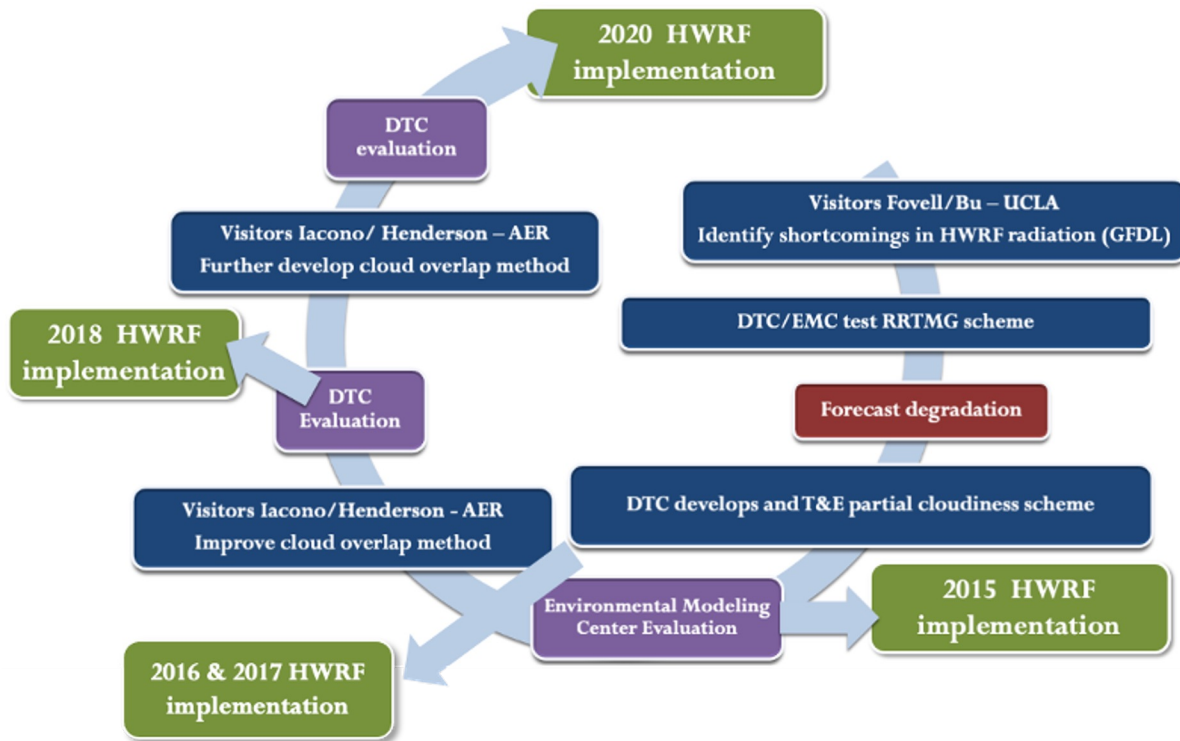
ML Potential within Verification

Humans can easily eyeball the *goodness* of a forecast

Can DL do that without requiring users to set many parameters (like in METplus-MODE)?

Davis et al. (2006)

Wrap Up - Big Need for Vx of MLWP Models

- Verification is key to advancing NOAA's ML efforts

- Requires independent evaluation

- Using user-driven metrics

- Testbeds have a role to play
  - **Developmental Testbed Center (DTC)** is following up on a recommendation from its Science Advisory Board and considering a **workshop on metrics/methods/tools for verifying MLWP**
  - Connection with NOAA AI groups will be important

Shameless advertising: **Two federal positions opening at GSL** soon for Scientific Computing (AI, or HPC, GPUs, etc.)