


ECMWF
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS

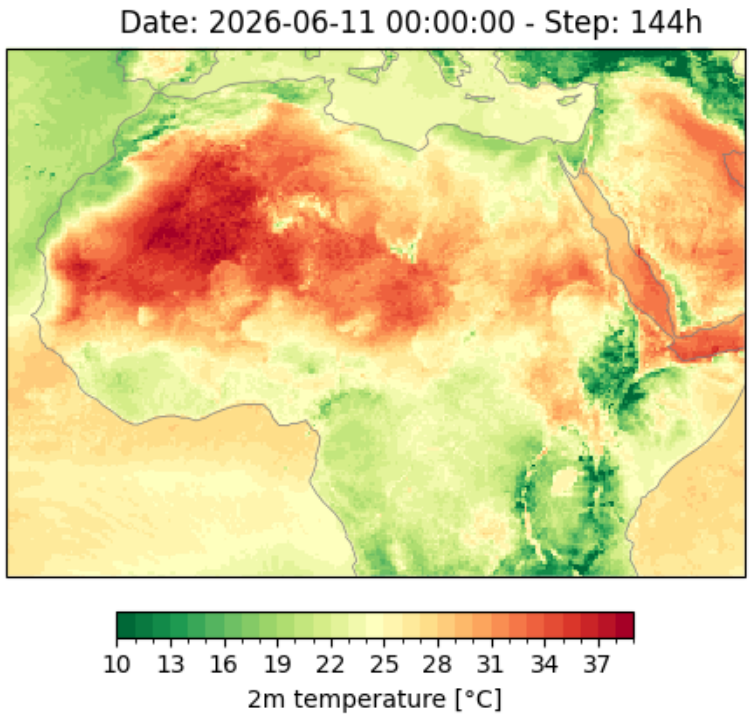
# An introduction to the assessment of ML-based forecasts

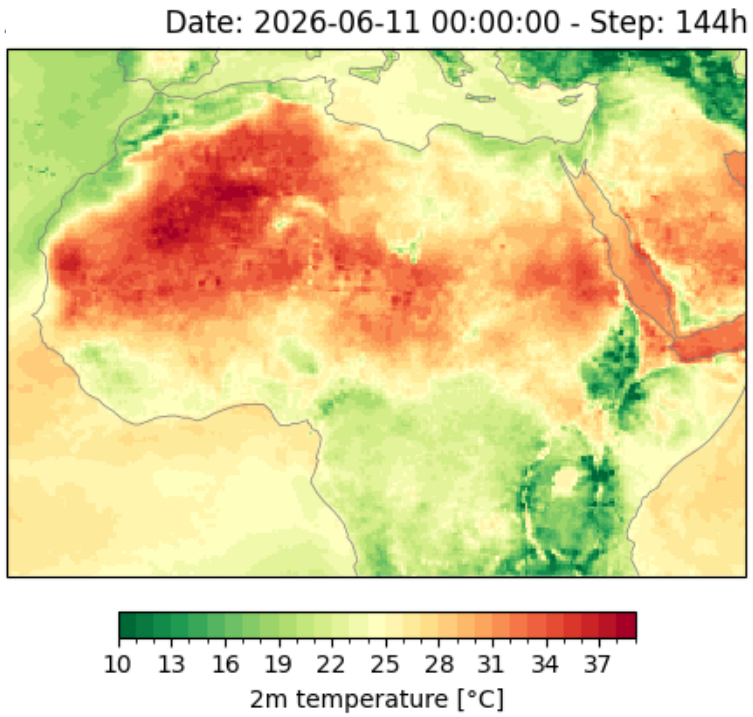
**Zied BEN BOUALLÈGUE**

**ML training course**
**November 2026**

# 0. Introduction

NWP (theory-driven) and ML (data-driven) forecasts

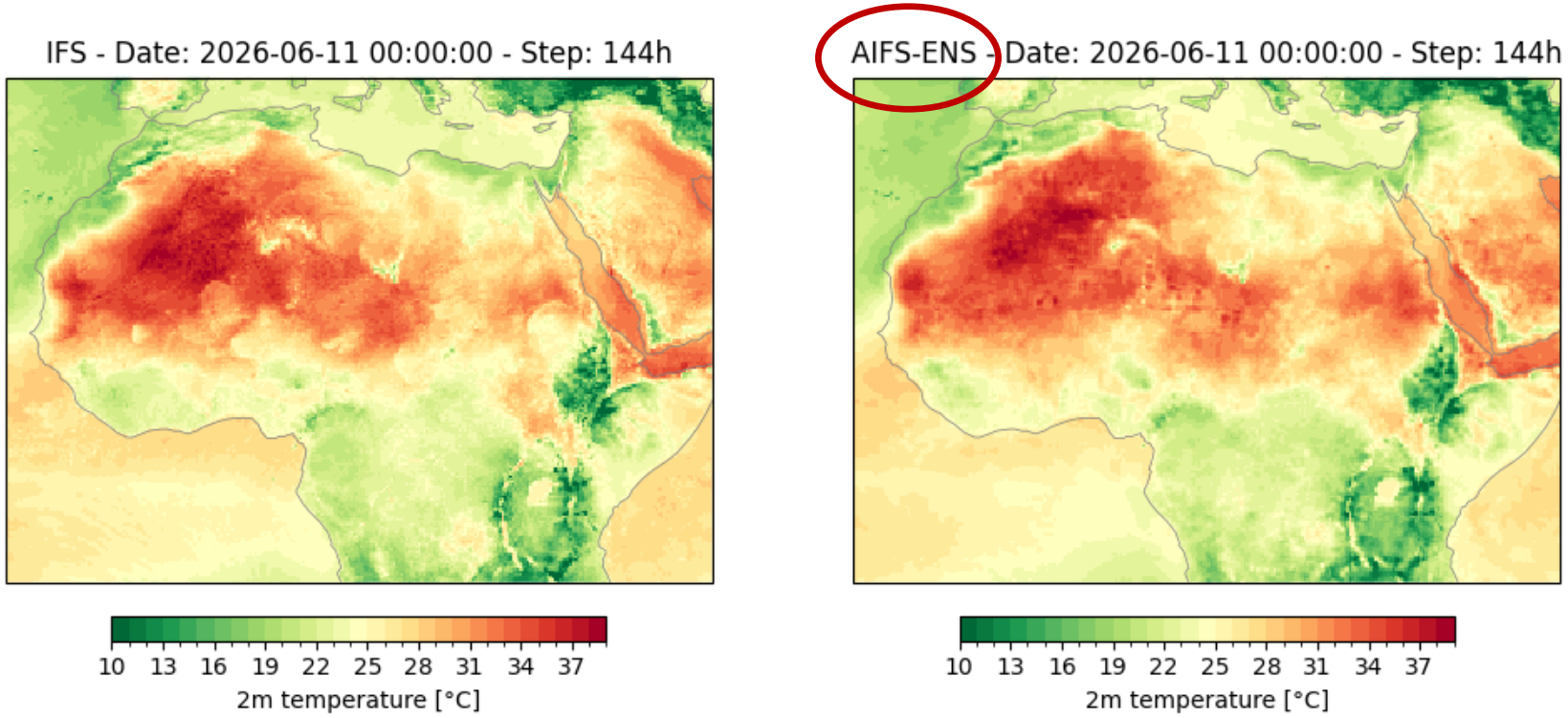



© ECMWF

# 0. Introduction

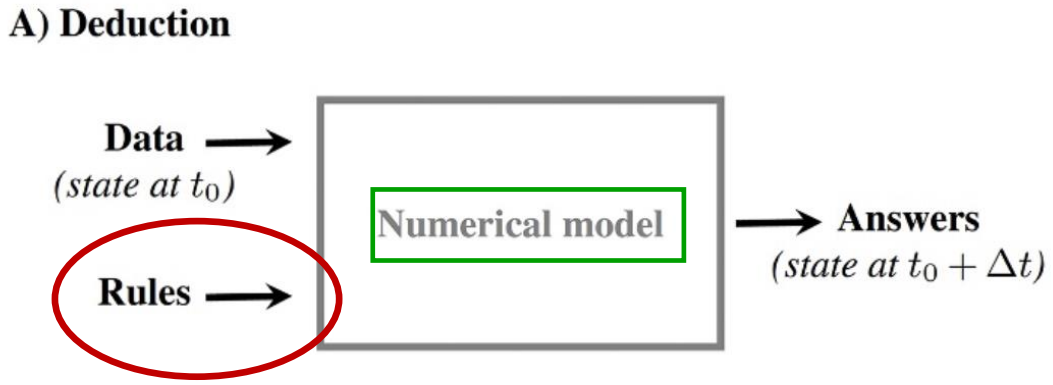
NWP (theory driven) and ML (data-driven) forecasts

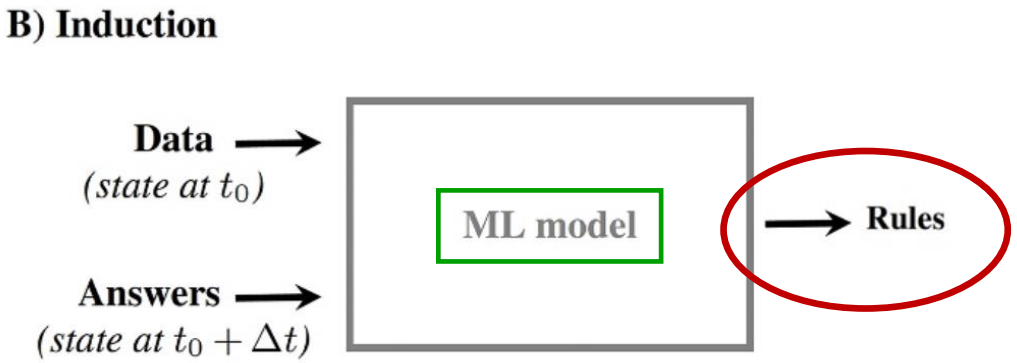



© ECMWF

# 0. Introduction

What's the difference between NWP and ML models?




Ben Bouallègue 2026 (Env. Res. Com.)

© ECMWF

# Agenda

1. Forecast falsification
- visual inspection
- conservation laws
- sensitivity tests

2. Forecast diagnostics
- bias, Q-Q plots
- forecast activity and spectra
- ensemble calibration

3. Forecast verification
- continuous variables
- binary (dichotomous) variables
- forecast value

**+ Special topics**
- data leakage
- representativeness
- benchmarking

© ECMWF

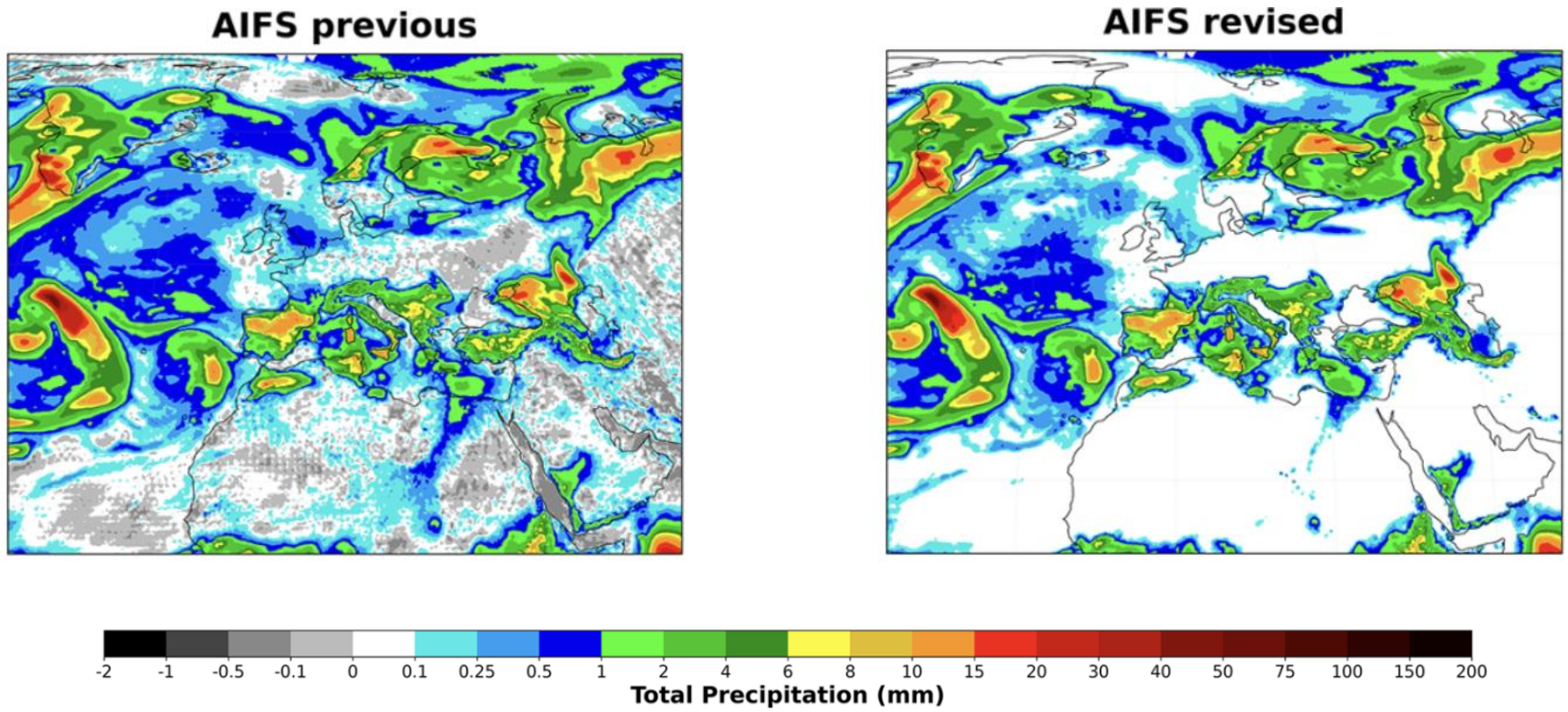
# 1. Forecast falsification

*Physical inconsistencies* in the forecast?
(between variables, in space and time, artefacts, …)

© ECMWF

# 1. Forecast falsification

**A. Visual inspection -** precipitation in AIFS v1




Moldovan et al 2026, GMD.

© ECMWF

# 1. Forecast falsification

## B. Conservation laws - horizontal wind divergence and vertical wind gradient

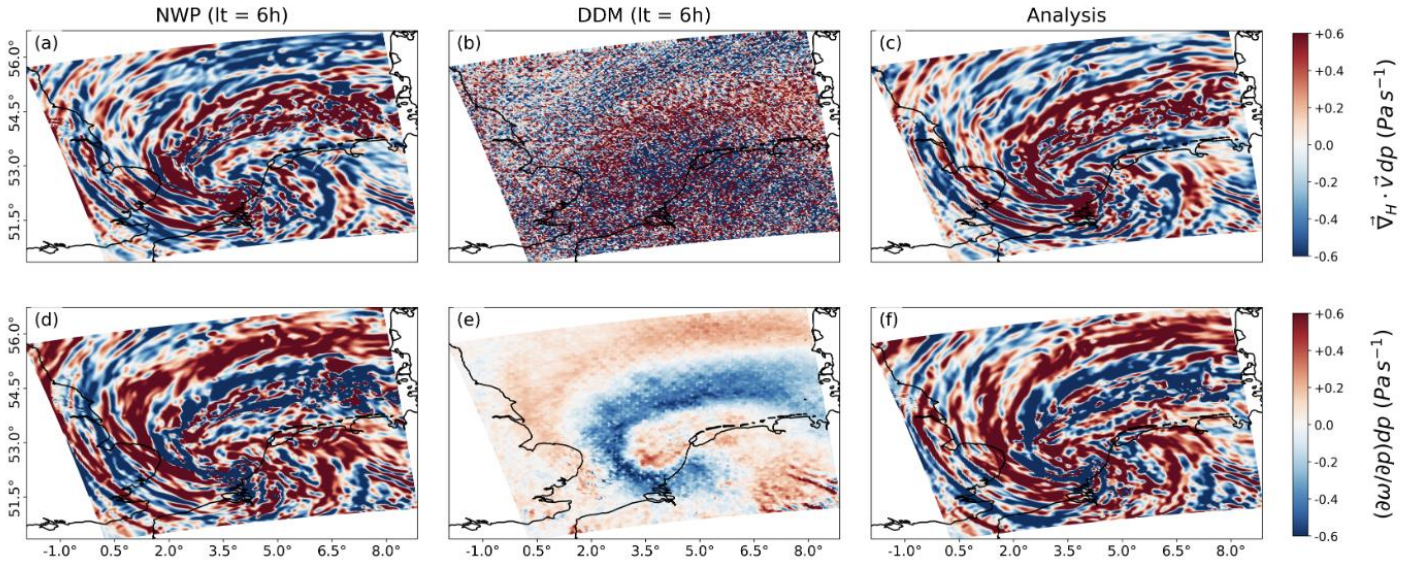



**FIGURE 8** (a–c) vertically integrated horizontal wind divergence over the 500–400 hPa layer; (d–f) vertical difference of the vertical velocity over the same 500–400 hPa layer. The maps show a 250 × 250 grid-point subdomain centered (roughly) on the storm's LPC at 06 UTC on 5 July 2023. Panels refer to NWP (left) and DDM (centre) +6 h forecasts, and the corresponding analysis field (right).

Pasquini et al 2026, QJRMS.

© ECMWF

# 1. Forecast falsification

## C. Sensitivity experiments - wave response to idealised initial conditions

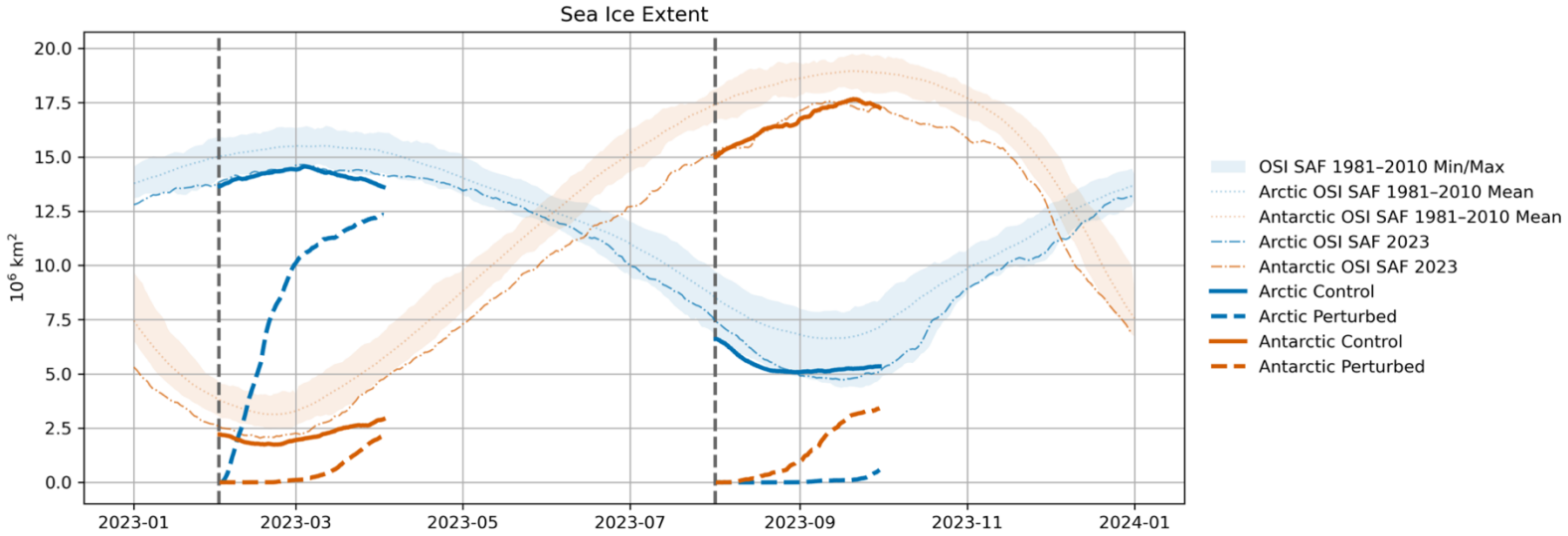



Figure 14: Sensitivity experiment in which sea ice is removed from the initial conditions. Time series of (top) sea ice extent and (bottom) sea ice volume for the Arctic (blue) and Antarctic (orange). Solid lines denote control forecasts, while dashed lines show perturbed forecasts initialised without sea ice. Vertical dashed lines indicate the two forecast initialisation dates (1 February 2023 and 1 August 2023). Shaded envelopes and dotted lines show the OSI SAF (extent) and PIOMAS (volume, Arctic only) 1981–2010 climatological range and mean, respectively, with dash-dotted lines indicating the corresponding 2023 observational estimates.

*Hahner et al 2026, arXiv.*

➢ Falsification: use of physical knowledge to assess data-driven forecasts

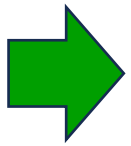
© ECMWF

# 1. Forecast falsification

*Take-home message*

Step 1:

- Inspect the forecast
- Use domain knowledge



Build trust

© ECMWF

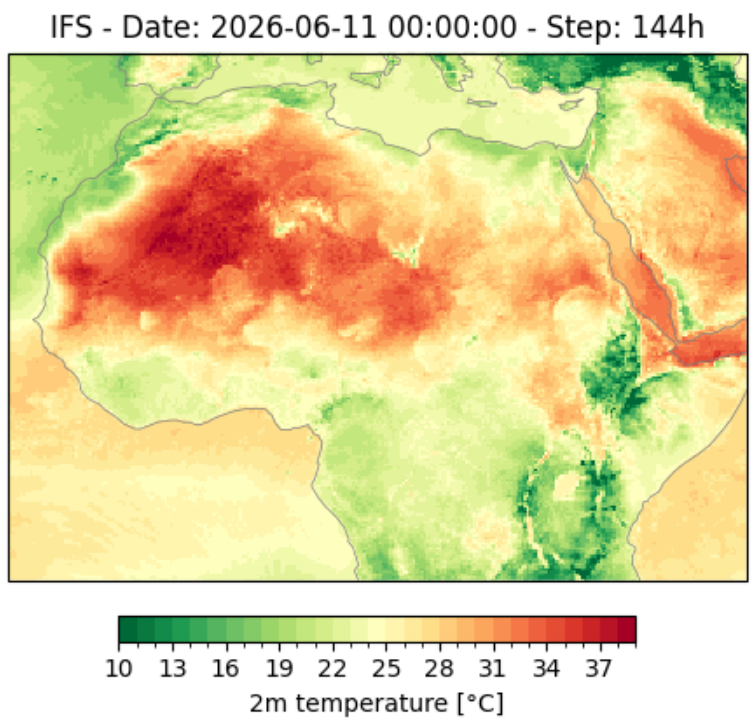
# 2. Forecast diagnostics

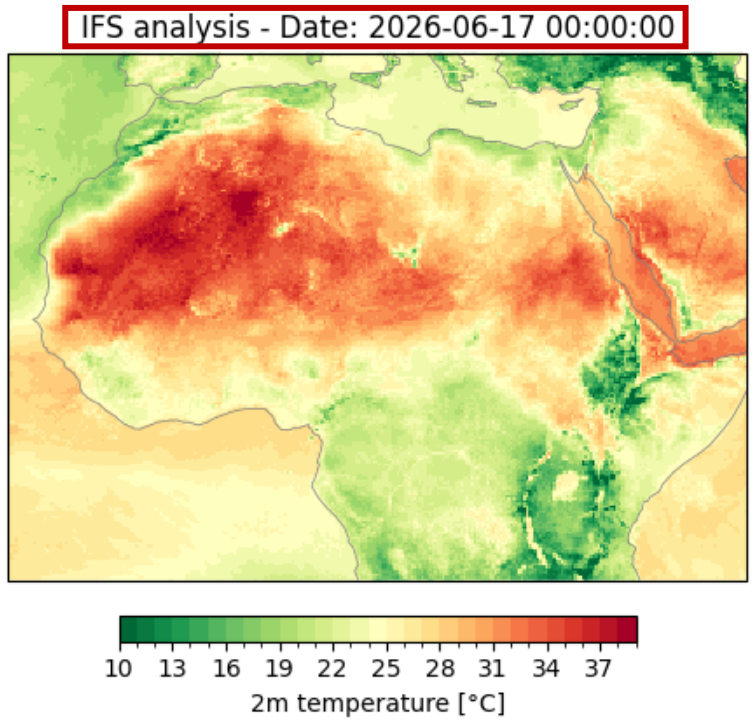
*Are there systematic* statistical *differences* between forecasts and observations?
(in texture, structure, biases,…)

© ECMWF

# 2. Forecast diagnostics

*Are there systematic* statistical *differences* between forecasts and observations?
(in texture, structure, biases,…)



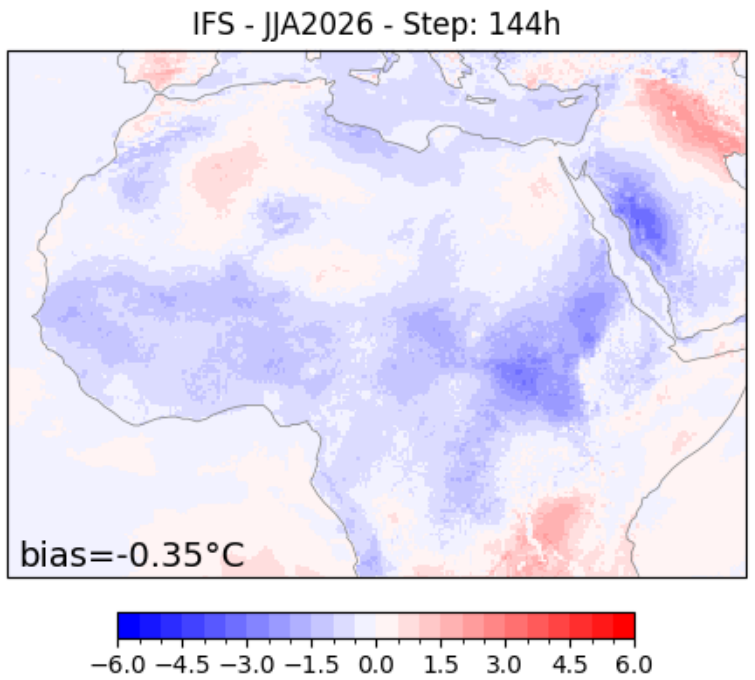

© ECMWF

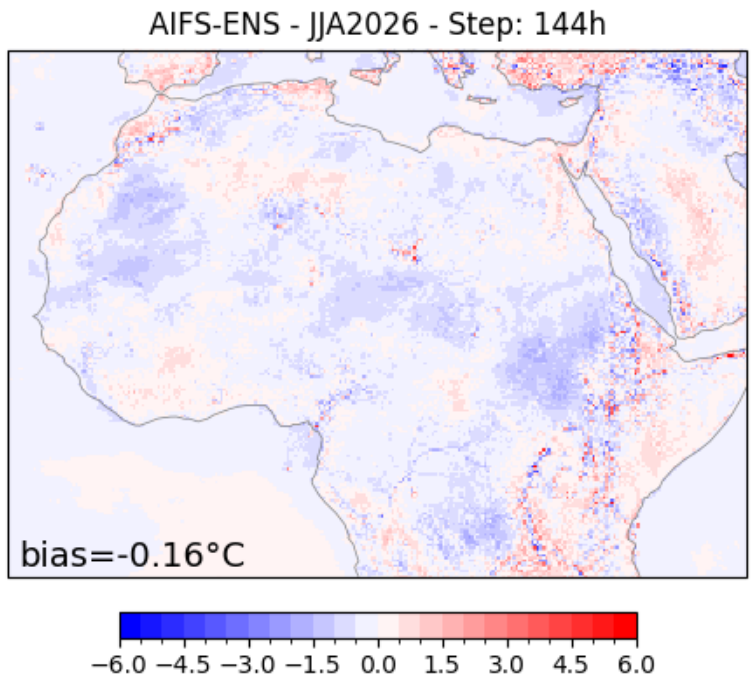
# 2. Forecast diagnostics

*Mean difference (bias)*

(forecast *systematically too warm or too cold?)*

$$\text{bias} = \overline{f - o}$$



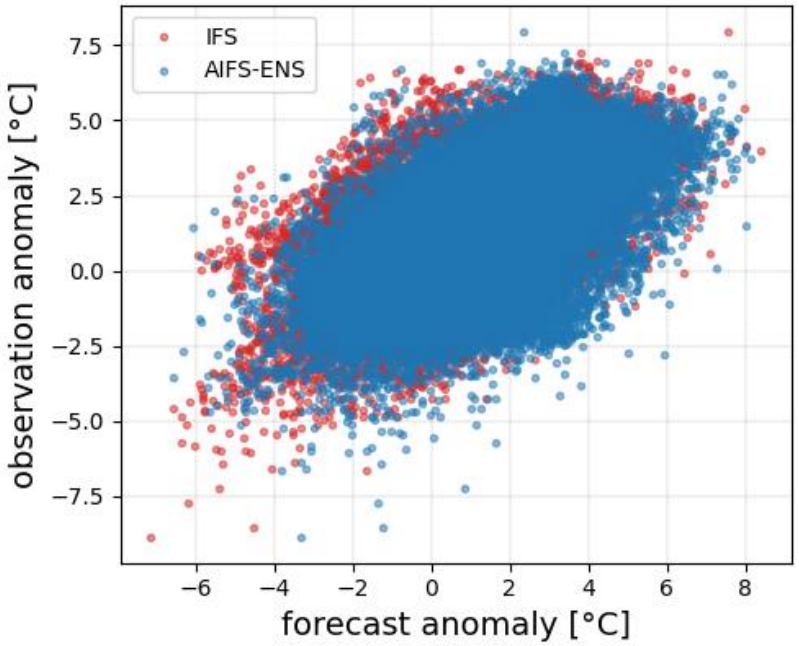

© ECMWF

# 2. Forecast diagnostics

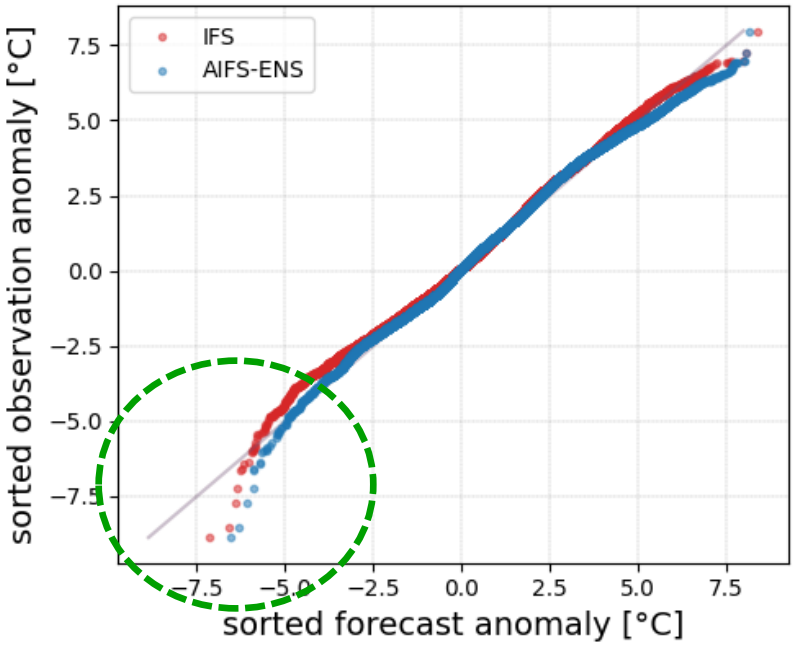
## *Conditional biases*

(forecast *systematically too warm in cold conditions?)*

scatter plot




quantile-quantile (Q-Q) plot




© ECMWF

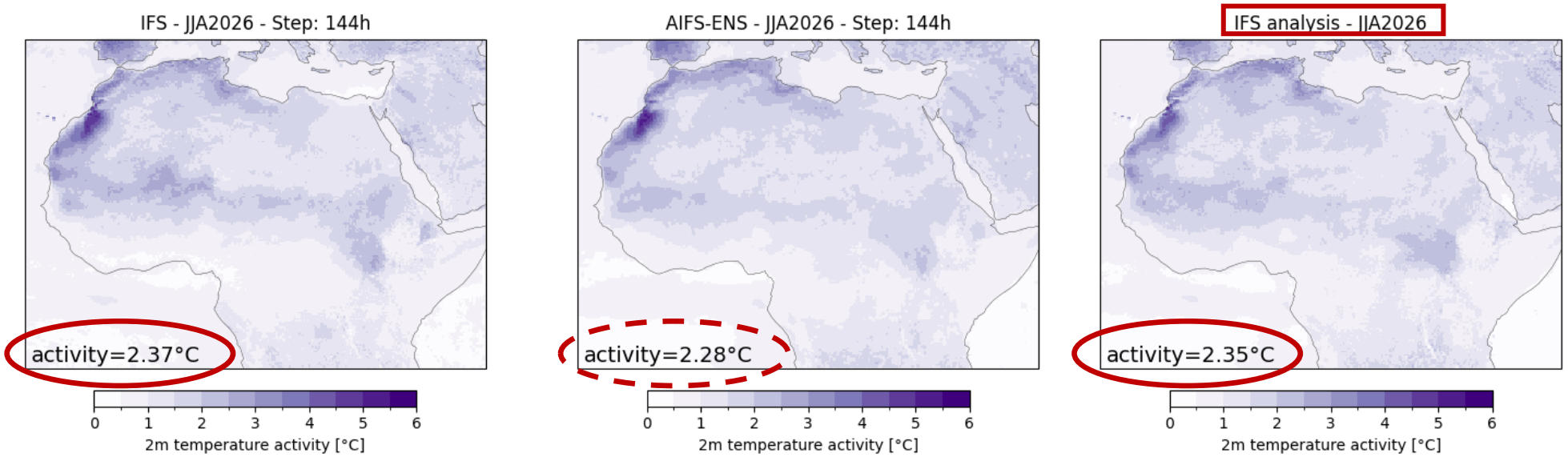
# 2. Forecast diagnostics

## Forecast and observation activity

(*how much variability in the forecast?)*

$$\text{activity}_{forecast} = \sqrt{\overline{(f_a - \overline{f_a})^2}}$$

$$\text{activity}_{observation} = \sqrt{(o_a - \overline{o_a})^2}$$

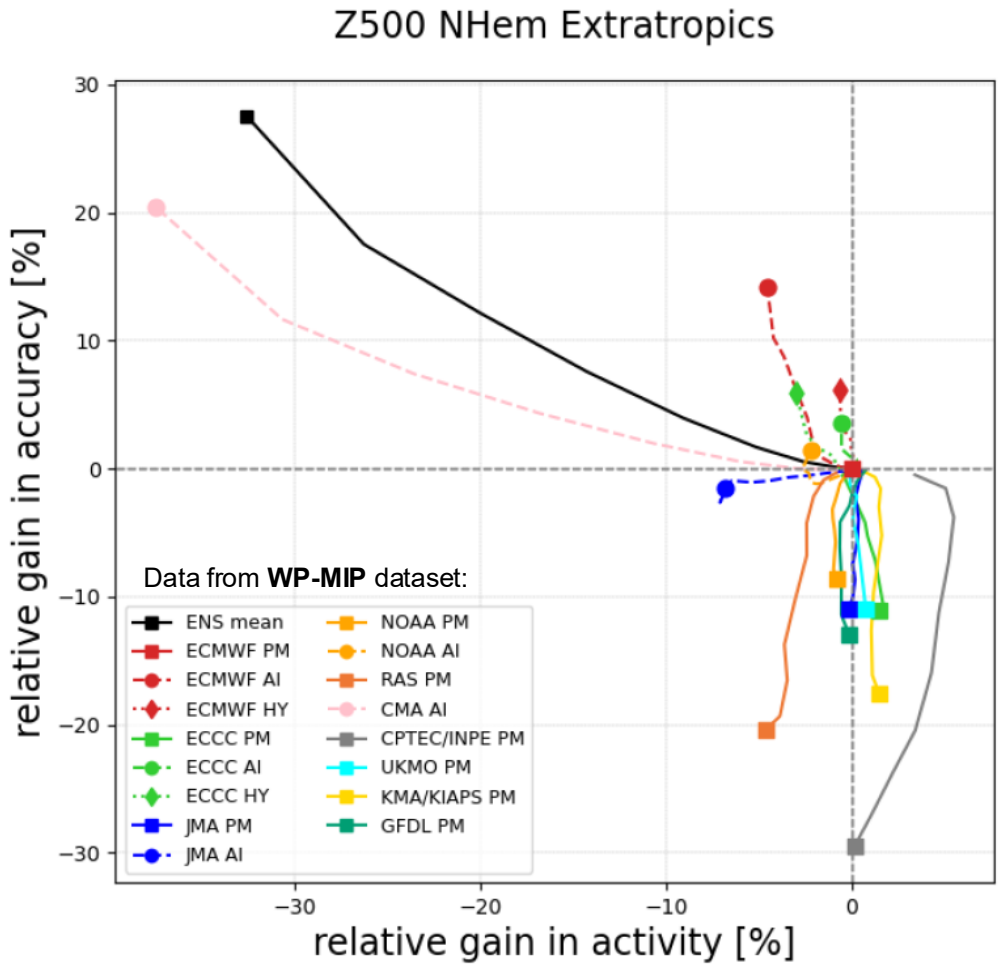



© ECMWF

# 2. Forecast diagnostics

## Activity/Accuracy trade-off

(*can we improve accuracy without smoothing?)*




For deterministic forecasts, **improved forecast accuracy** might come at the cost of a **loss in forecast activity.**

(a smoother forecast can score better in terms of RMSE)

$$\mathrm{SS}(f, \bar{x}, x) = r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\bar{f} - \bar{x})/s_x]^2. \quad (12)$$

Murphy 1988 (Mon. Wea. Rev.)

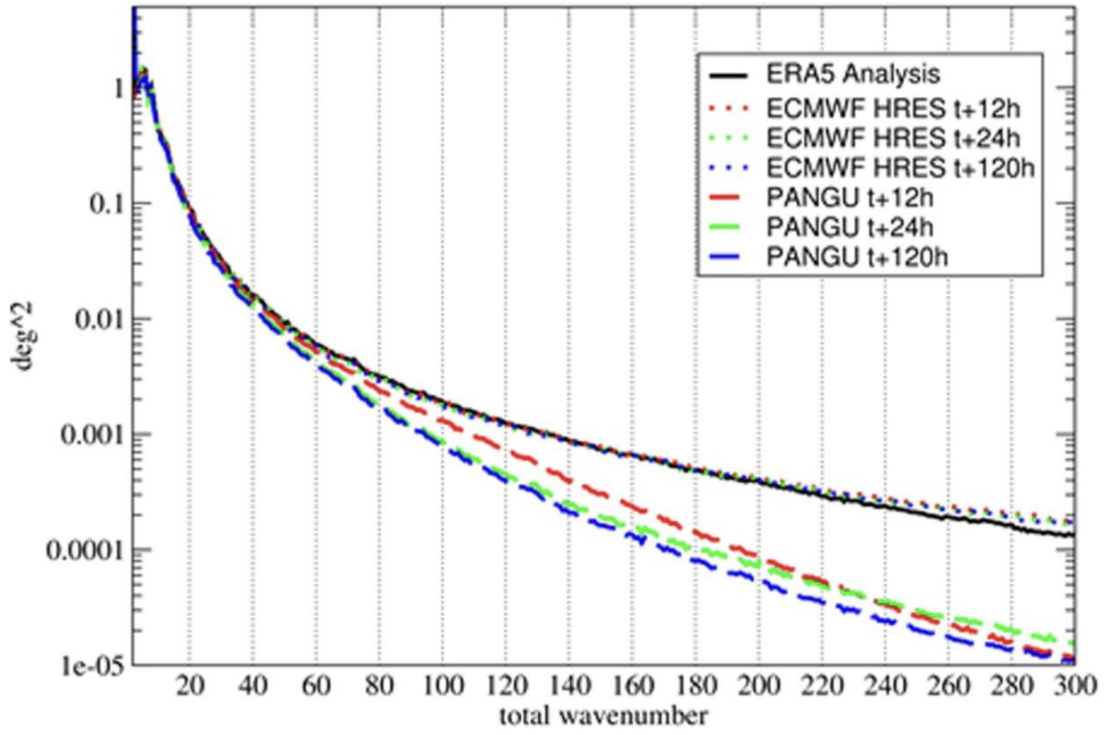
https://www.ecmwf.int/en/about/media-centre/aifs-blog/2024/accuracy-versus-activity

© ECMWF

# 2. Forecast diagnostics

## Power spectra

*(Do forecasts and observations show the same energy at all scales?)*



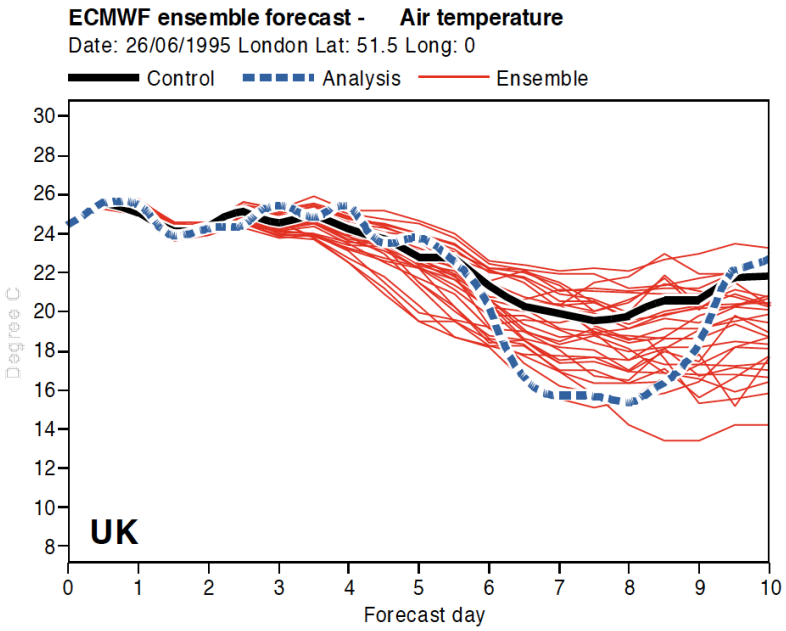

Bonavita 2024 (G. Res. Letters)

© ECMWF

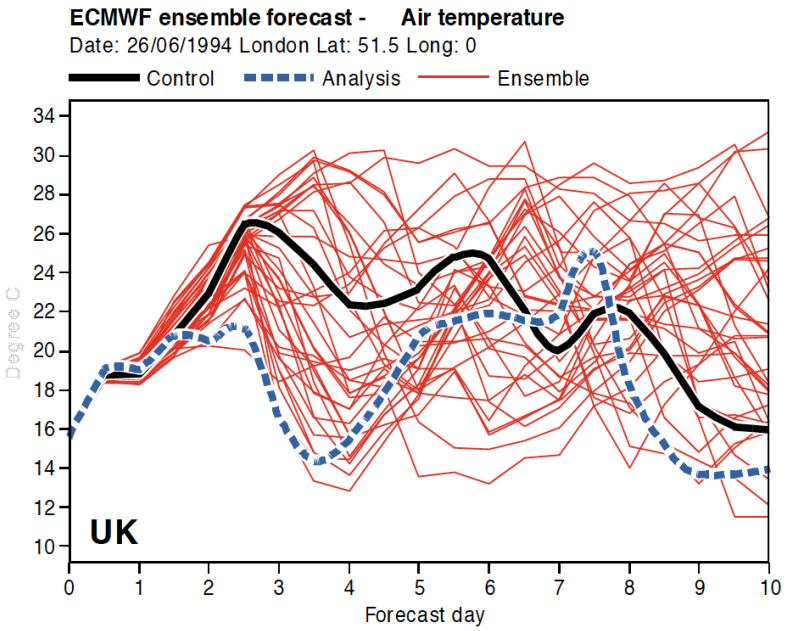
# 2. Forecast diagnostics

## Ensemble forecasting

26th June 1995




26th June 1994

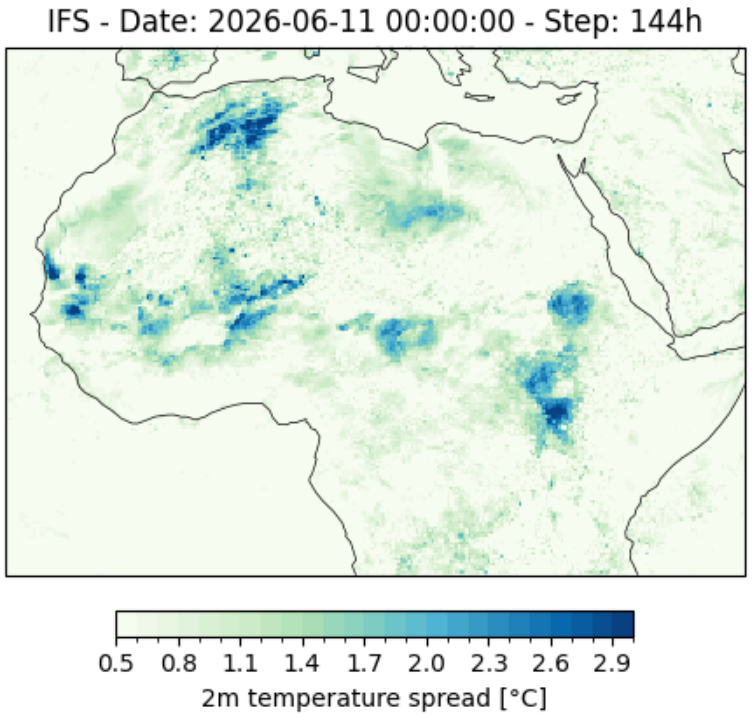


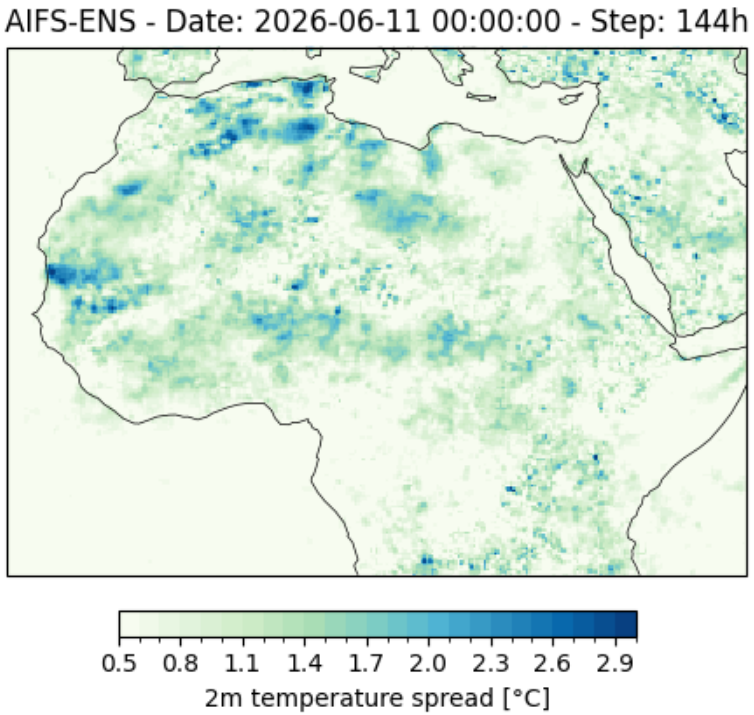

Courtesy: David S. Richardson (ECMWF)

# 2. Forecast diagnostics

## Ensemble forecasting

*(forecasting the forecast error)*



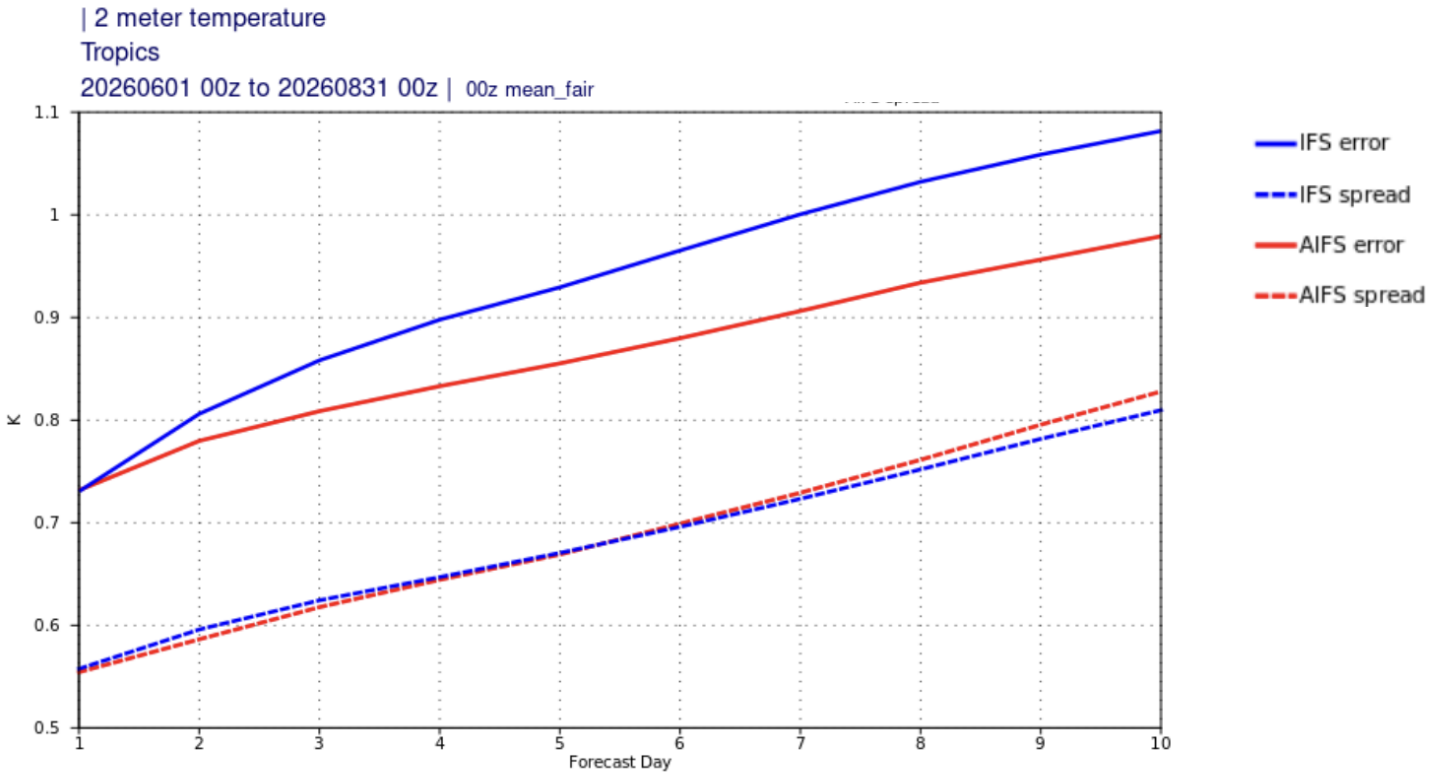




© ECMWF

# 2. Forecast diagnostics

## Ensemble forecasting

*(forecasting the forecast error)*



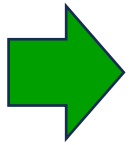

➢ Calibration: **forecasts and observations** drawn from the **same conditional distribution**

# 2. Forecast diagnostics

*Take-home message*

Step 2:

- check for *systematic differences*
- check for systematic *conditional differences*
- check for ensemble *reliability*



Diagnose model deficiencies

© ECMWF

# 3. Forecast verification

*What is the forecast's predictive performance?*

© ECMWF

# 3. Forecast verification

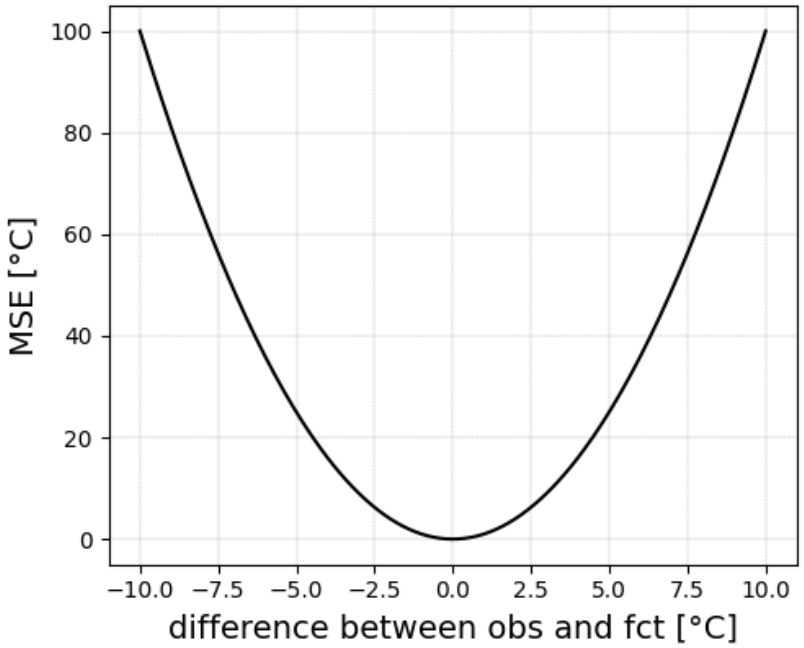
## A. *Continuous variables*

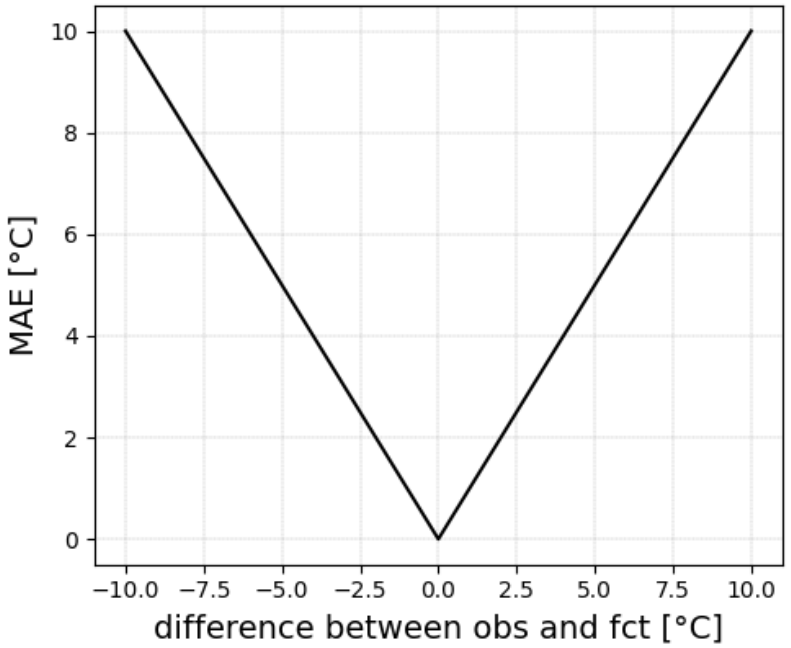
Mean squared error (mse), mean absolute error (mae), quantile score (qs)

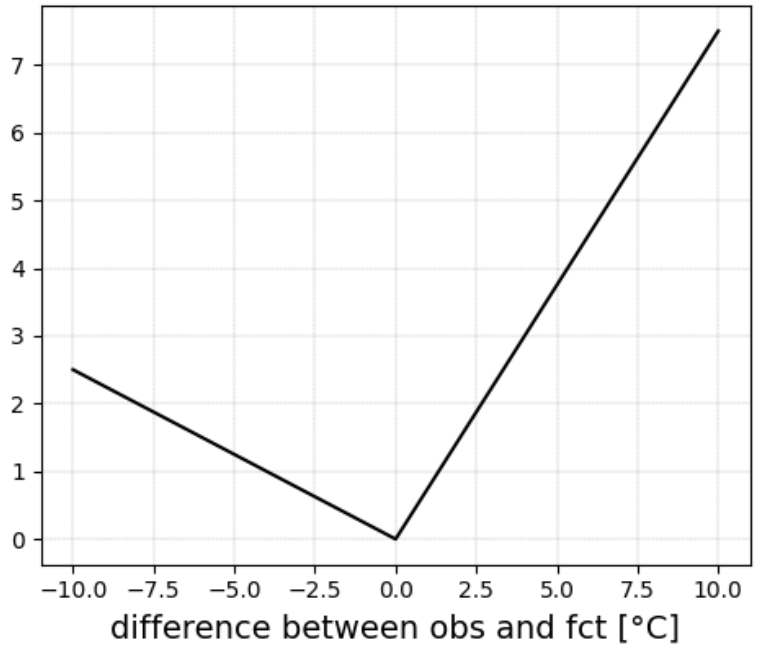
$$\text{mse} = (f - o)^2$$

$$\text{mae} = |f - o|$$

$$\text{qs} = \begin{cases} 2\alpha(q - o), & o \leq q, \\ 2(1 - \alpha)(o - q), & o > q. \end{cases}$$




© ECMWF

# 3. Forecast verification

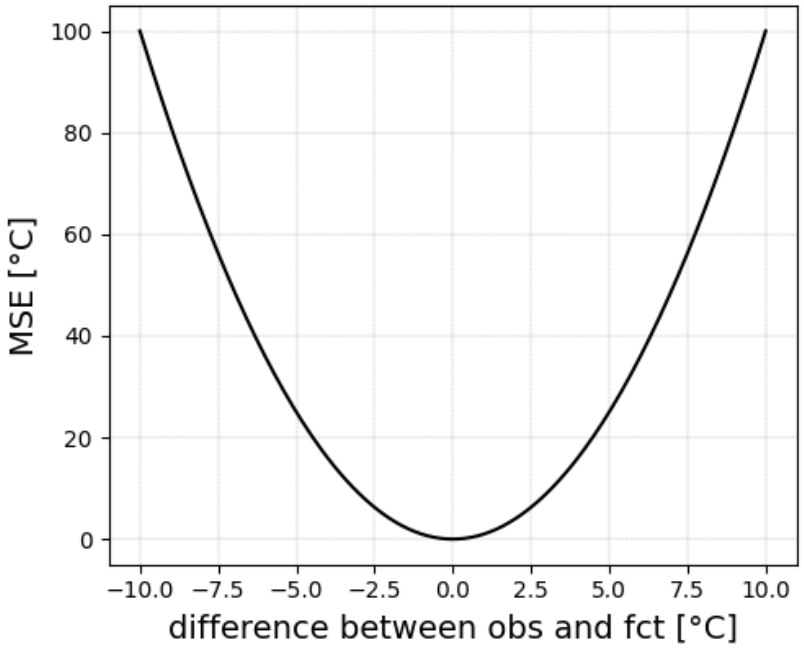
## A. *Continuous variables*

Mean squared error (mse), mean absolute error (mae), quantile score (qs)

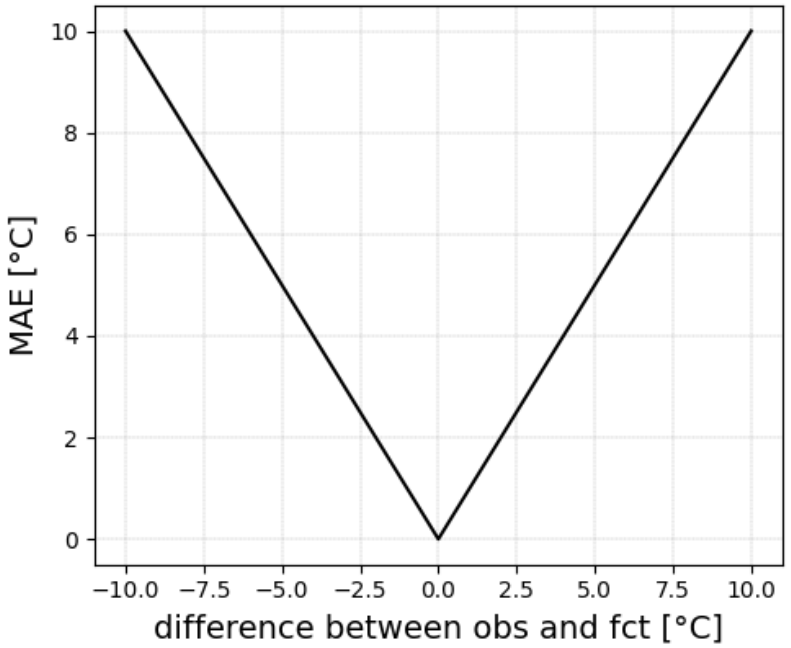
$$\text{mse} = (f - o)^2$$

$$\text{mae} = |f - o|$$

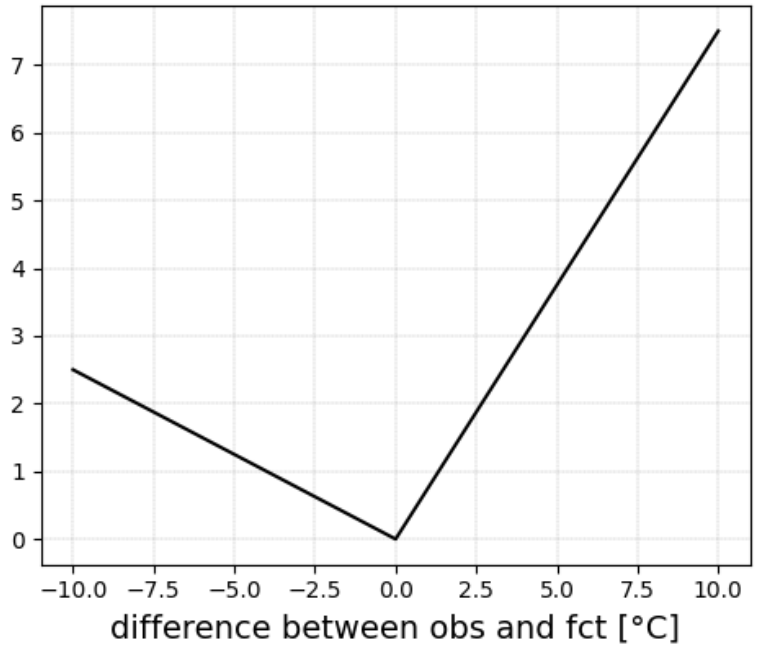
$$\text{qs} = \begin{cases} 2\alpha(q - o), & o \leq q, \\ 2(1 - \alpha)(o - q), & o > q. \end{cases}$$




consistent with the conditional ***mean***




consistent with the conditional ***median***



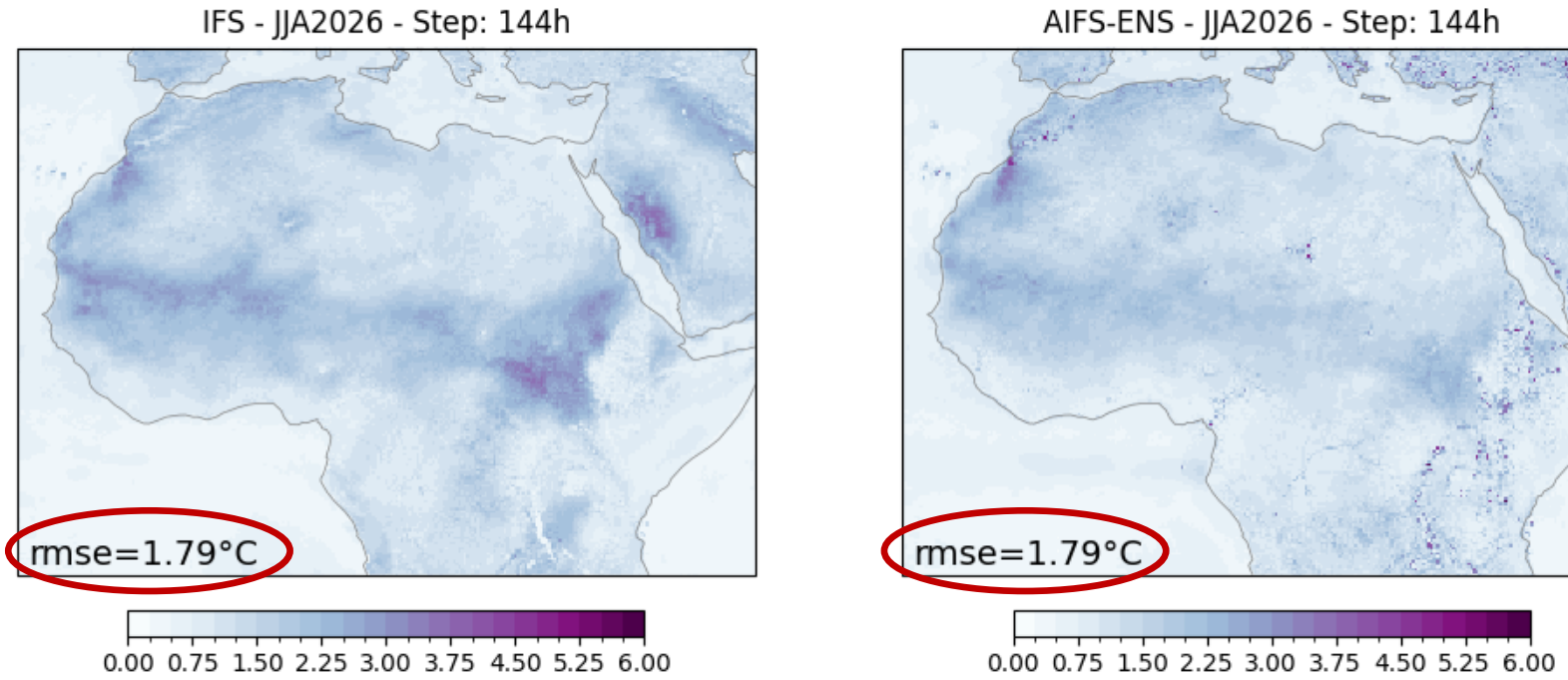

consistent with the conditional ***quantile***

© ECMWF

# 3. Forecast verification

## A. *Continuous variables*

Root mean squared error applied to the ensemble ***mean***




- Aggregated scores can hide different error patterns

© ECMWF

# 3. Forecast verification

## A. *Continuous variables – focus on extremes*

**Threshold weighted scores** to focus on more **extreme events**,
but avoiding **the forecaster's dilemma**

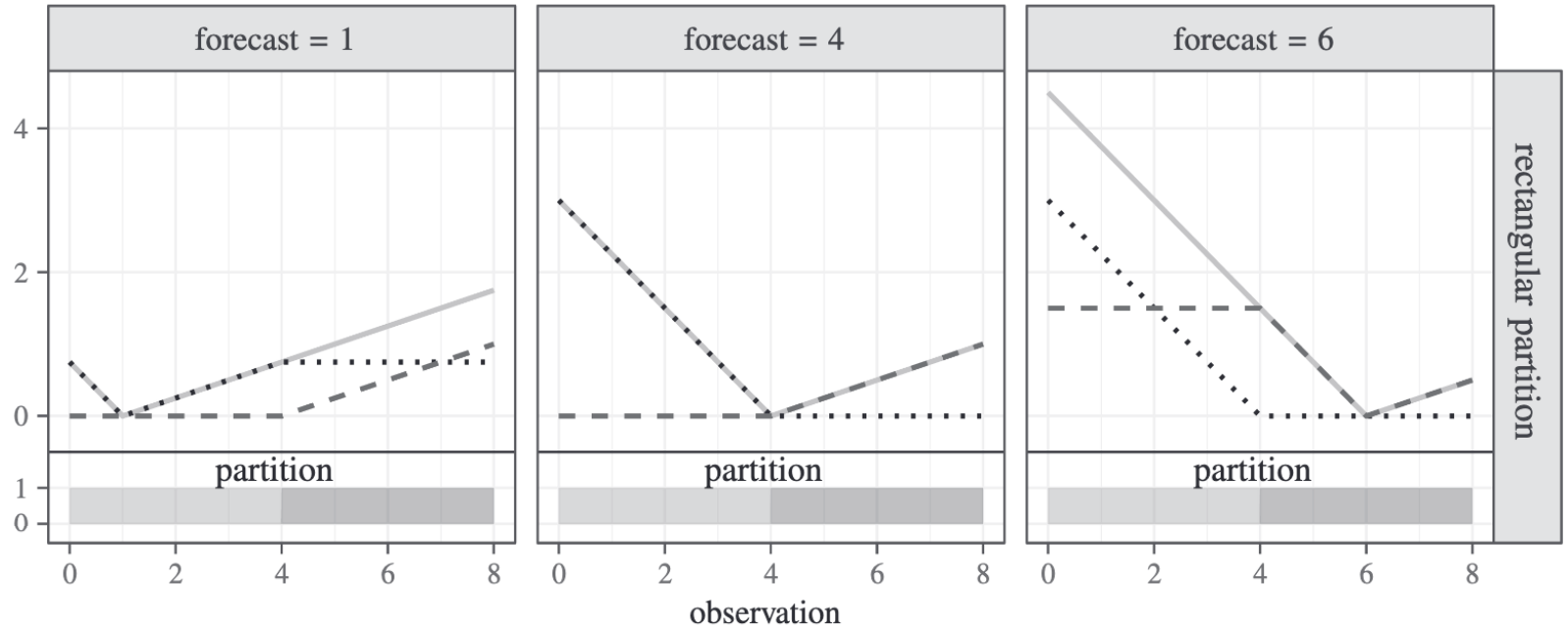



FIGURE 5 Decomposition $S = S_1 + S_2$ of the commonly used consistent scoring function for 0.25-quantile forecasts

*Taggart 2021 (QJRMS)*

© ECMWF

# 3. Forecast verification

## A. *Continuous variables – ensemble forecasts*

Continuous ranked probability score **(CRPS)** and the energy score **(ES)**

$$\text{CRPS(F,o)} = \int_0^1 qs_\alpha(F_\alpha, o)\,\mathrm{d}\alpha$$

$$\text{CRPS(F,o)} = E[|f - o|] - \tfrac{1}{2}E[|f - f\prime|]$$

**Univariate metric**

$$\text{ES(F,o)} = E[||\mathbf{f} - \mathbf{o}||] - \tfrac{1}{2}E[||\mathbf{f} - \mathbf{f}\prime||]$$

**Multivariate metric**

© ECMWF

# 3. Forecast verification

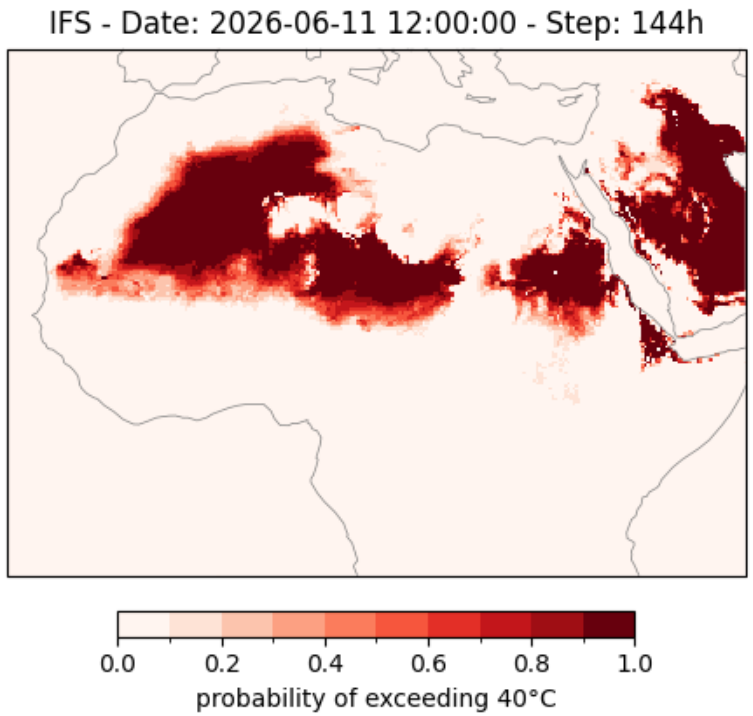
## B. *Dichotomous variables*

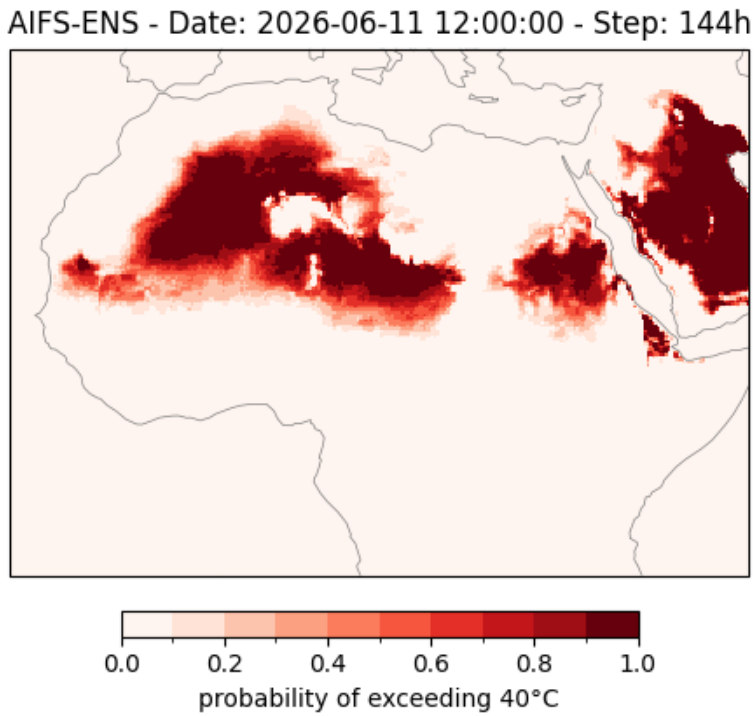
© ECMWF

# 3. Forecast verification

## B. *Dichotomous variables*

example: 2m temperature (at 12UTC) exceeding 40 °C



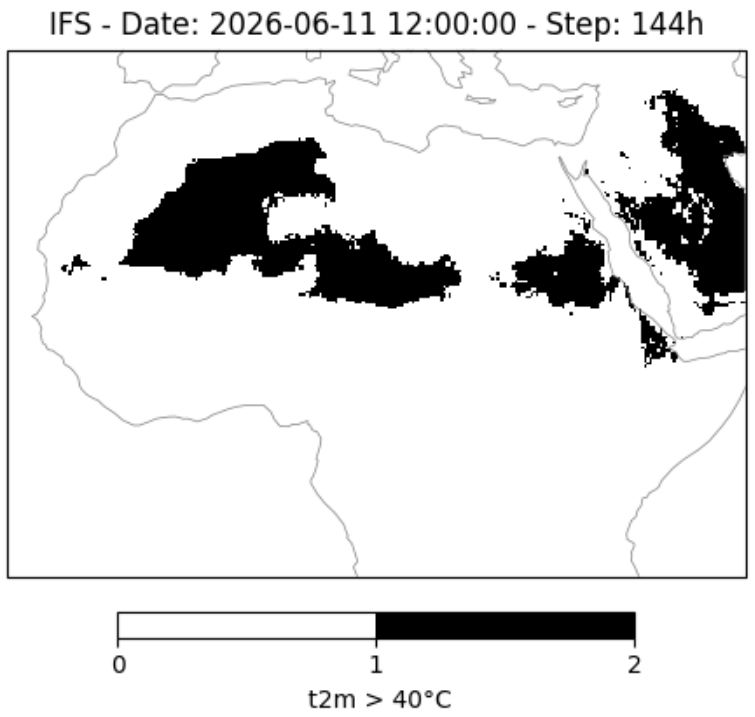



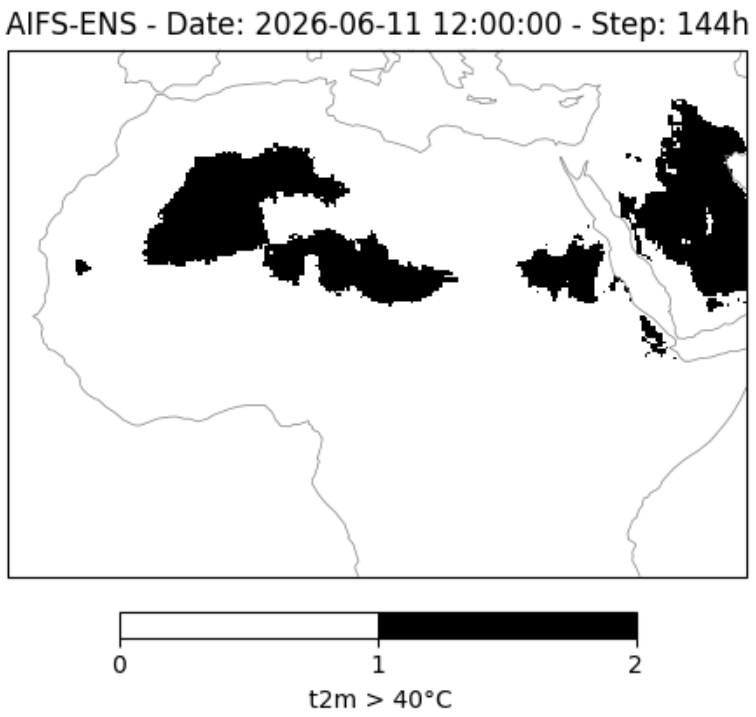

© ECMWF

# 3. Forecast verification

## B. *Dichotomous variables*

example: 2m temperature (at 12UTC) exceeding 40 °C - **probability >80%**




© ECMWF

# 3. Forecast verification

## B. *Dichotomous variables*

contingency table-based scores

*Courtesy from Sam Allen*

| | | Observed: Yes | Observed: No |
|---|---|---|---|
| Forecast | Yes | **True positive (TP)** | **False positive (FP)** |
| Forecast | No | **False negative (FN)** | **True negative (TN)** |

© ECMWF

# 3. Forecast verification

## B. *Dichotomous variables*

contingency table-based scores

*Courtesy from Sam Allen*

| Forecast \ Observed | Yes | No |
|---|---|---|
| Yes | **True positive (TP)** | **False positive (FP)** |
| No | **False negative (FN)** | **True negative (TN)** |

$$PSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$$

$$F1\ score = \frac{2TP}{2TP+FN+FP}$$

$$FBI = \frac{TP+FP}{TP+FN}$$

$$\text{POD} = \frac{TP}{TP+FN} \qquad \text{POFD} = \frac{FP}{FP+TN}$$

$$\text{SEDI} = \frac{\ln(\text{POFD}) - \ln(\text{POD}) - \ln(1-\text{POFD}) + \ln(1-\text{POD})}{\ln(\text{POFD}) + \ln(\text{POD}) + \ln(1-\text{POFD}) + \ln(1-\text{POD})}$$

*Ferro and Stephenson, 2011 (WAF)*

© ECMWF

# 3. Forecast verification

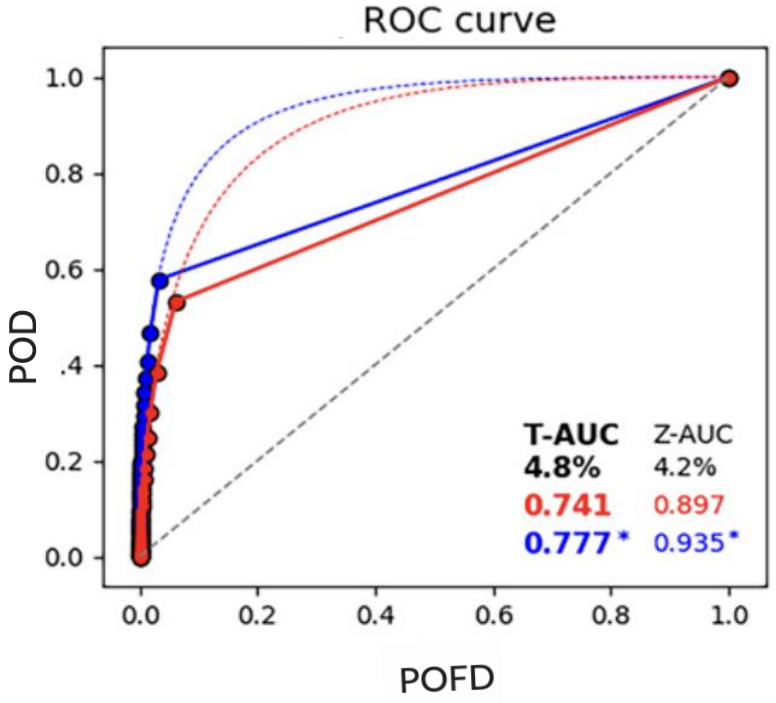
## B. *Dichotomous variables*

contingency table-based scores

*Courtesy from Sam Allen*

| Forecast \ Observed | Yes | No |
|---|---|---|
| Yes | **True positive (TP)** | **False positive (FP)** |
| No | **False negative (FN)** | **True negative (TN)** |

$$\text{POD} = \frac{TP}{TP+FN} \qquad \text{POFD} = \frac{FP}{FP+TN}$$




*Ben Bouallègue and Richardson, 2022 (WAF)*

© ECMWF

# 3. Forecast verification

## B. *Dichotomous variables*

cost-loss model and the economic value

| | | Occurs | |
|---|---|---|---|
| | | Yes | No |
| Action | Yes | C | C |
| | No | L | 0 |

*Richardson 1999 (ECMWF report)*

cost-loss ratio = $\frac{C}{L}$

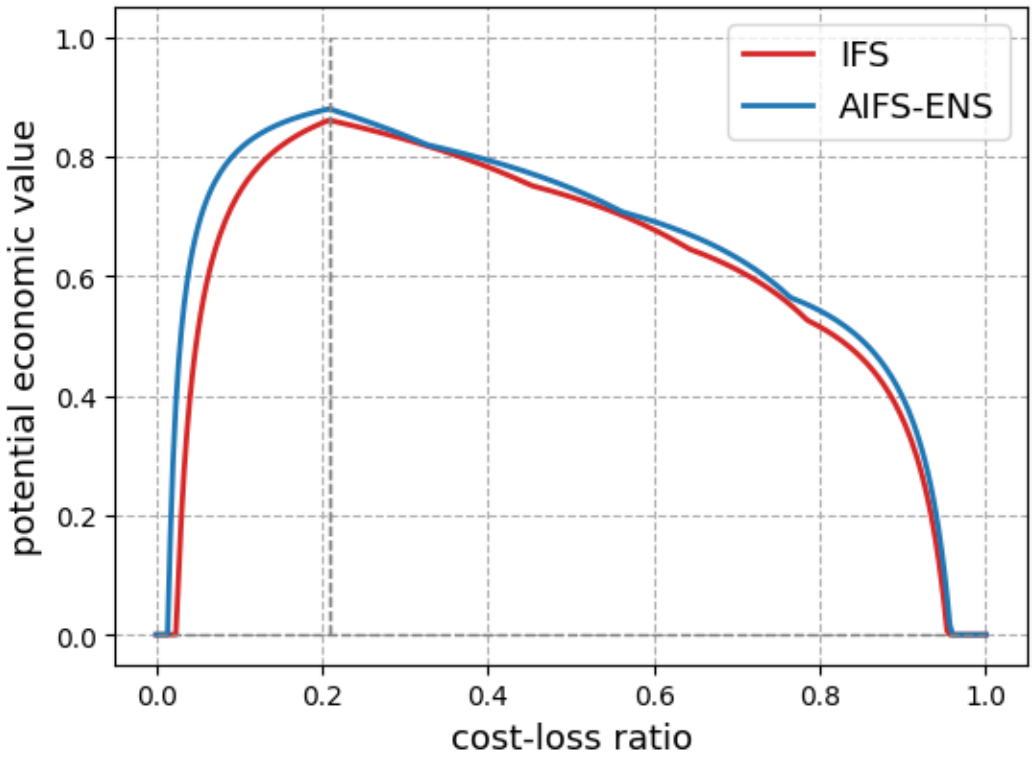
depends on the user, the application

© ECMWF

# 3. Forecast verification

## B. *Dichotomous variables*

example: 2m temperature exceeding **30 °C**, 00UTC, JJA 2026




cost-loss ratio = $\frac{C}{L}$

depends on the user, the application

$$V = \frac{\min(\alpha,s) - \alpha s(1-POD) - (1-s)(1-\alpha)POFD - \min(\alpha,s)(1-\min(\alpha,s))}{\min(\alpha,s) - s\alpha}$$

© ECMWF

# 3. Forecast verification

*Take-home message*

Step 3:

- choose a metric consistent with the *forecast*
- choose a metric meaningful for the *user*
- check *multiple scores* for a wider picture

Compare and communicate performance

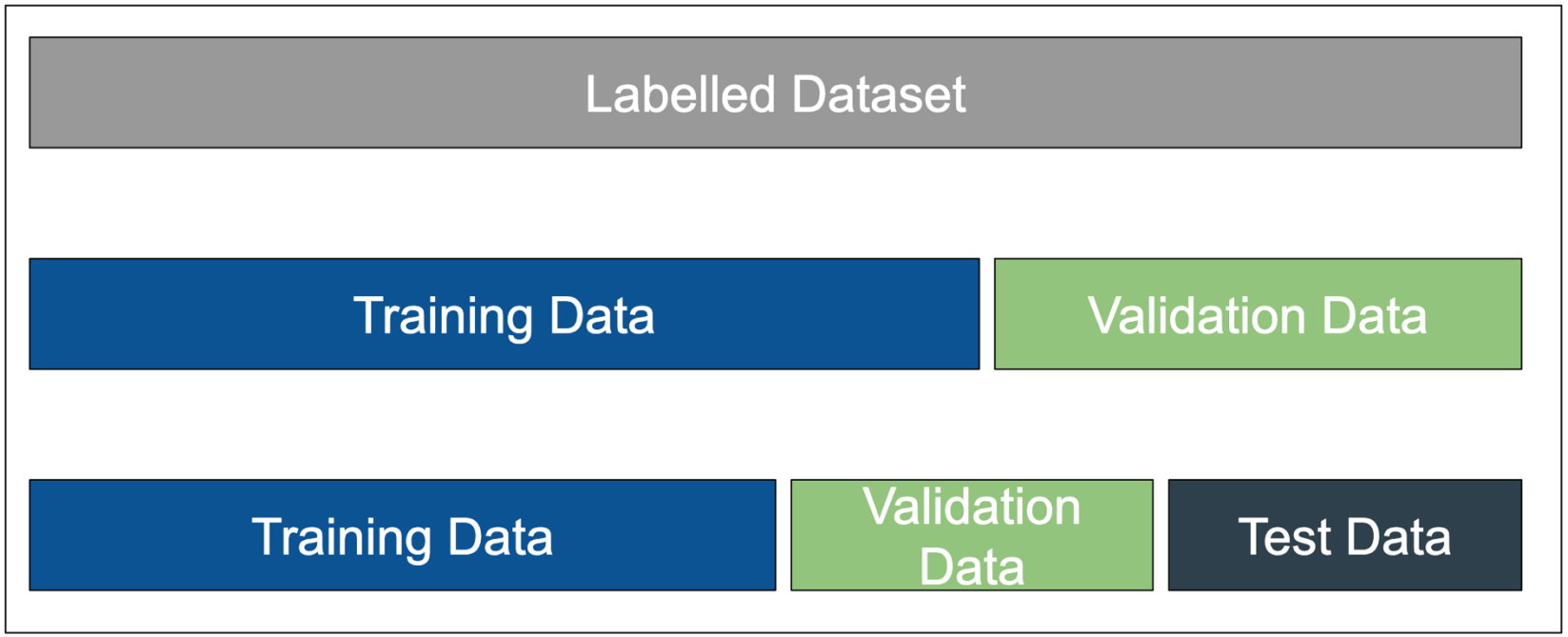
© ECMWF

# Special topics

© ECMWF

# A. Data leakage

Will models work on unseen data (in the future)?




Courtesy: Julian Kuehnert (ECMWF)

© ECMWF

# B. Representativeness

Do model output and observation represent the same thing?

Representativeness errors:

- Systematic errors




Pagano et al 2024. Figure 4.

© ECMWF

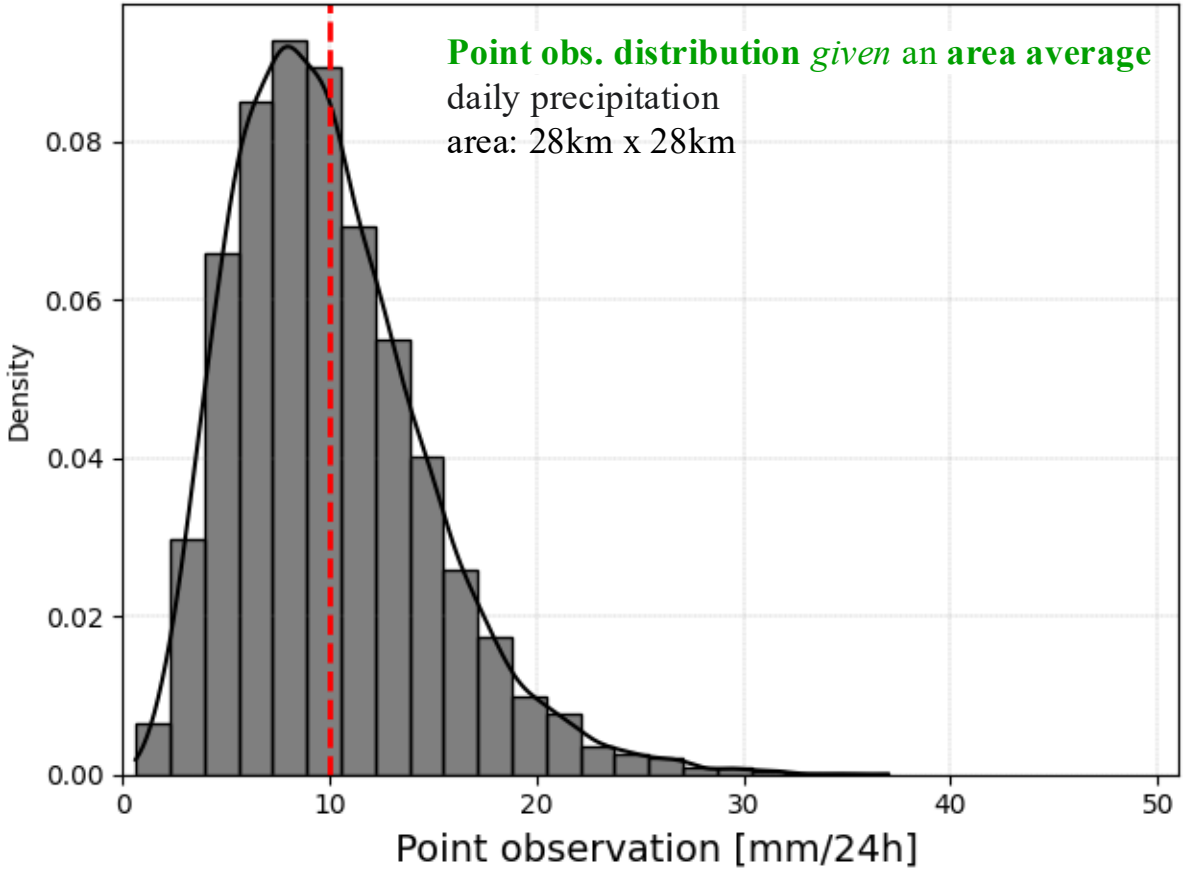
# B. Representativeness

Do model output and observation represent the same thing?

Representativeness errors:

- Systematic errors

- Random errors




Ben Bouallègue et al 2020 (MWR)

© ECMWF

# C. Benchmarking

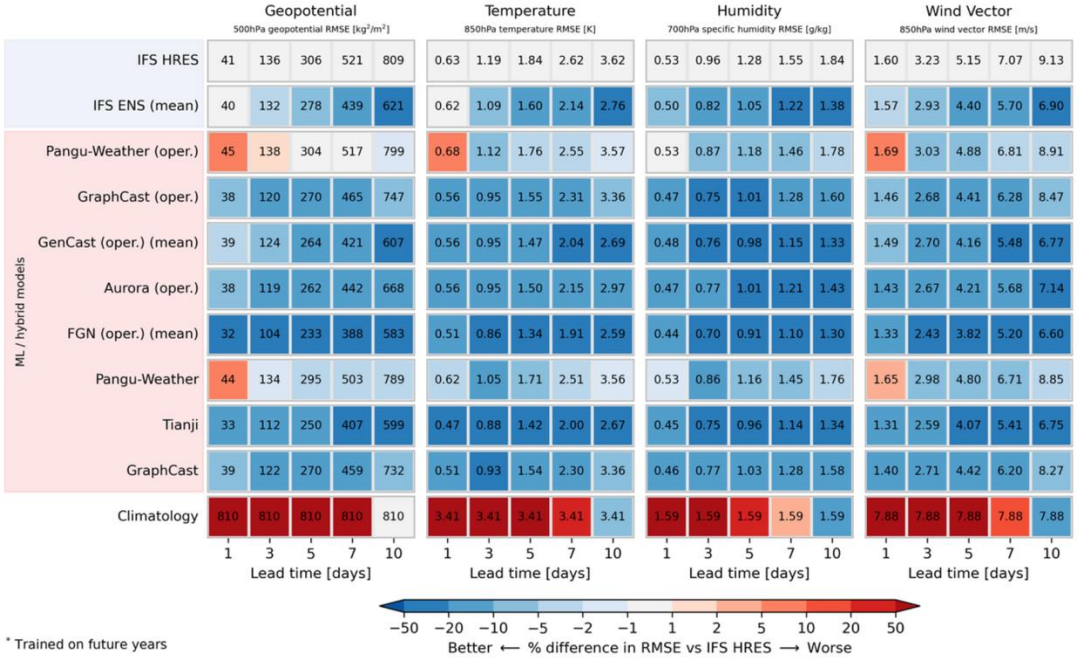
How to make a fair comparison of different forecasting systems?

## Benchmark

- a clear problem statement
- a clear definition of evaluation metrics
- a clear verification protocol

## Post-processing step?

Should we correct for systematic errors before comparing forecast performance?




*Rasp et al, 2024 (JAMES)*

© ECMWF

# Summary

*How to evaluate AI-based forecasts?*

1. Forecast falsification

Check for the physical consistency to build trust
(subjective verification, conservation laws, sensitivity tests …)

2. Forecast diagnostics

Understand forecast characteristics to pinpoint model errors
(bias, Q-Q plots, activity, spread/skill relationship,…)

3. Forecast verification

Measure and communicate forecast performance
(scoring functions, contingency table scores, economic value,…)

**Important aspects to keep in mind:**

- risk of data leakage
- representativeness issues (observation errors in general)
- fairness in benchmarking exercises

© ECMWF




- Webpage on forecast verification methods:

  https://jwgfvr.github.io/forecastverification

- EGU27 session on Verification, diagnostic, and interpretability of AI models for weather forecasting:

  https://meetingorganizer.copernicus.org/EGU27/session/60441

© ECMWF

# Incomplete list for further reading


Ferro, CAT (2017), Measuring forecast performance in the presence of observation error. Q.J.R. Meteorol. Soc, 143: 2665-2676. https://doi.org/10.1002/qj.3115

Lerch, S et al (2017) Forecaster's Dilemma: Extreme Events and Forecast Evaluation, https://doi.org/10.1214/16-STS588

Taggart, R. (2021) Evaluation of point forecasts for extreme events using consistent scoring functions, https://doi.org/10.1002/qj.4206

Ben Bouallègue Z and Richardson D S (2022) On the ROC Area of Ensemble Forecasts for Rare Events, https://doi.org/10.1175/WAF-D-21-0195.1

Dueben, P (2022) Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook, https://doi.org/10.1175/AIES-D-21-0002.1

Rasp, S et al (2024) WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models, https://doi.org/10.1029/2023MS004019

Ben Bouallègue, Z (2026) 'What is a realistic forecast?' Assessing data-driven weather forecasts, a journey from verification to falsification, https://doi.org/10.1088/2515-7620/ae9f1f

McTaggart-Cowan, R et al (2026) WP-MIP: An Artificial Intelligence, Hybrid, and Physically Based Model Intercomparison Project for Weather Prediction, https://doi.org/10.1175/BAMS-D-26-0012.1

Casati, B et al (2026) Operational Verification of Artificial Intelligence Weather Prediction, https://doi.org/10.1175/BAMS-D-26-0054.1


[…]

© ECMWF