Environment and Climate Change Canada | Environnement et Changement climatique Canada

The WWRP Joint Working Group in Forecast Verification Research (JWGFVR) and recent WMO activities:

Barbara Casati (ECCC and co-chair of the JWGFVR)
Ocean Predict – 15 April 2026

- VerAIops workshop
- WP-MIP verification plan

Canada

Joint Working Group in Forecast Verification Research

https://community.wmo.int/site/knowledge-hub/governance/research-board/scientific-steering-committee-ssc/joint-working-group-forecast-verification-research

WORLD METEOROLOGICAL ORGANIZATION

Weather · Climate · Water

WWRP WORLD WEATHER RESEARCH PROGRAMME

JWGFVR

Mission: The JWGFVR aims to advance the development and application of improved diagnostics and verification methods to assess and monitor the quality and value of weather and environmental predictions, for time scales encompassing weather forecasts, sub-seasonal and seasonal predictions, and decadal and climate projections.

Promote good verification practices :

- Verification web-page
- Verification tutorials
- Verification software
- WMO recommendation reports and WIPPS verification standards for ops. centers -> INFCOM

Advance verification research:

- Spatial verification method intercomparisons
- International verification methods workshops
- Verification challenges
- Scientific Journal Special Issues & publications

Support verification activities in WWRP and WGNE/WCRP

- AvRDP2
- HIW
- SAGE
- PCAPS
- InPRHA
- TC-PFP
- URBAN
- PEOPLE
- ADVANCE
- WGNE WP-MIP
- WGSIP
- WGCM

WORLD METEOROLOGICAL ORGANIZATION | WWRP World Weather Research Programme | WCRP World Climate Research Programme | WGNE Working Group on Numerical Experimentation | JWGFVR | Environment and Climate Change Canada / Environnement et Changement climatique Canada

Workshop on

Verification of AI models in Operational Centers

Canadian Meteorological Center, 22-23 October 2025

https://hpfx.collab.science.gc.ca/~bca851/VerAIops2025/index.html

Barbara Casati barbara.casati@ec.gc.ca

Background: Robust, fair, and transparent verification methods are essential to understand the characteristics and capabilities of the new ML prediction systems in comparison to traditional NWP, to guide their effective use, and to develop optimal hybrid approaches to weather forecasting.

Aims: Advance our understanding; plan WP-MIP verif activities; feed into WIPPS operational standards

How to:

1. adapting established (operational) verification practices to the specific features of AI-based forecasts
2. testing existing (spatial) metrics and developing new diagnostics for meaningful comparisons

➡ ensure that WMO Members and forecast users can **trust and interpret** outputs from both paradigms with confidence.

Participation: more than 200 scientists from MRD+MSC (ECCC) and WGNE+JWGFVR international research network

Leveraging on the outcomes of the preceding VerAI workshops

Jochen Broecker, University of Reading, UK

- VerAI 2024 workshop: https://www.met.reading.ac.uk/~pt904209/VerAI2024/index.html
- VerAI 2025 workshop: https://www.met.reading.ac.uk/~pt904209/VerAI2025/index.html
- VerAI 2026 workshop: https://www.met.reading.ac.uk/~pt904209/VerAI2026/index.html

Three main themes were identified :

1. Statistical properties of the AI-based models and how to evaluate them;
2. Meaningful benchmarking and reliable reference datasets;
3. physical realism and explainability.

The latter relates to **the legal and ethical requirements for "explainable AI"**

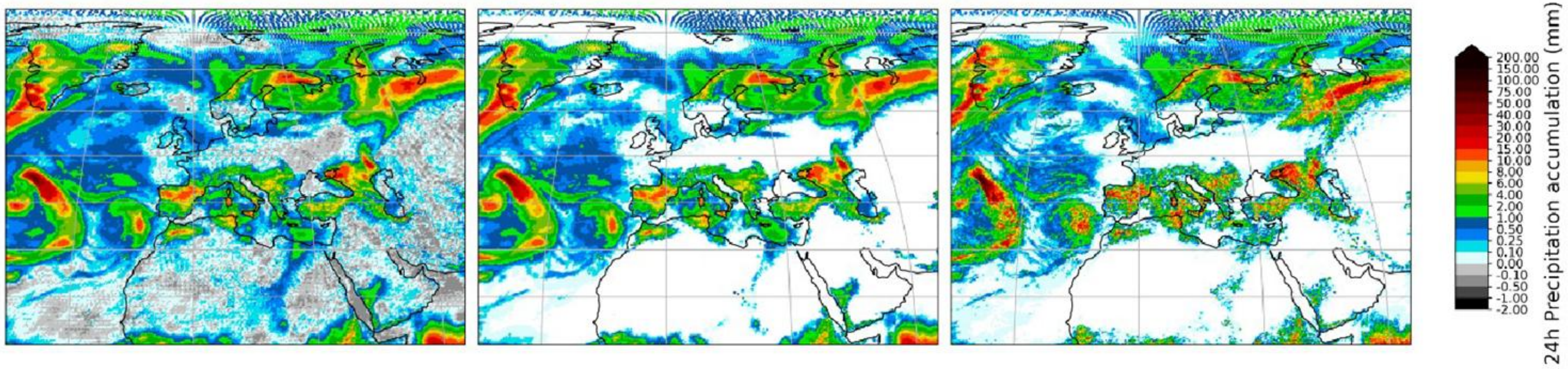
- See recommendations from the "Hiroshima Process International Code of Conduct for Advanced AI systems" and "G7 Leaders' Statement" (https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems and https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process)
- See also the EU AI Act, recital 27 https://artificialintelligenceact.eu/recital/27/

new tools that enable **transparency** with respect to AI-generated forecasts, allowing appropriate **traceability** and **explainability** of AI-based weather prediction, are needed!

Verification, Diagnostics and Falsification

From Zied Ben-Bouallègue (ECMWF): while traditional evaluation methods focus on forecast *verification* and model *diagnostics*, the emergence of ML-based systems introduces the need for *falsification* of the underlying data-driven rules.

Image from Moldovan et al (2025)



(a) AIFS previous (b) AIFS revised (c) IFS

This new conceptual framework calls for the development of tools (e.g. see M.Höver, S.Wahl talks) for testing for physical inconsistencies (e.g. negative precipitation, featuring in some data-driven models) and compliance with physical laws (such as energy and dynamical balance, e.g. Bonavita 2024 GRL), which are needed to provide confidence in AIWP and enhance explainability by aligning with physical reasoning.

Verification measures for AI training: inherent responsibilities and new meta-verification tools

Environment and Climate Change Canada
Environnement et Changement climatique Canada
JWGFVR

From B.Casati presentation:

- With the advent of ML, **verification metrics have assumed a new role**, from assessing the performance after the forecast is issued, to be cost functions for training – and hence shaping – the weather forecast.
- This comes with inherent responsibilities, since the metric used to train defines the characteristics of the produced data-driven forecast (e.g. RMSE → smooth forecast).
- Subproduct: AI training can be used as a meta-verification tool, since reveals the forecast characteristics which can be used for hedging, and artificially improve, the score.
- Examples shown at the workshop are:
  1. the very well-known smoothing if training with RMSE (cf. C.Subic presentation),
  2. precipitation over-forecasting if training with CSI (cf. D.Brunet presentation).

Benchmarking reference datasets and incestuousness

Marburg University

JWGFVR

Meaningful benchmarking and reference datasets are a concern (see S.Lerch, B.Casati, J.Broeker talks), as fair comparisons are challenging:

- AIWP models are optimized based on the RMSE as a loss function, which might put them at an unfair advantage in RMSE-based comparisons – this consists in **metric incestuousness!**
- What is the ground truth (ERA5 vs. IFS analysis vs. own analysis vs. obs)? What is the effect of fine-tuning AIWP models towards a specific observational dataset? Incestuousness between ERA5 as verification reference and ERA5-trained AI-models is concerning – this consists in **dataset incestuousness!**
- **Benchmarking**: AIWP models are usually compared to raw NWP forecasts, whereas a (more) practically relevant and fair comparison might be against post-processed NWP forecasts - S.Lerch introduces the potential CRPS: deterministic model dressed+debiased, enables the use of CRPS = proper scoring rule

➡ **WP-MIP experimental design** – verification of SIC and OIC, against IFS and own analysis + obs, stationary vs transient weather, portfolio of different metrics, …

Enhanced spatial verification diagnostics

Environment and Climate Change Canada
Environnement et Changement climatique Canada

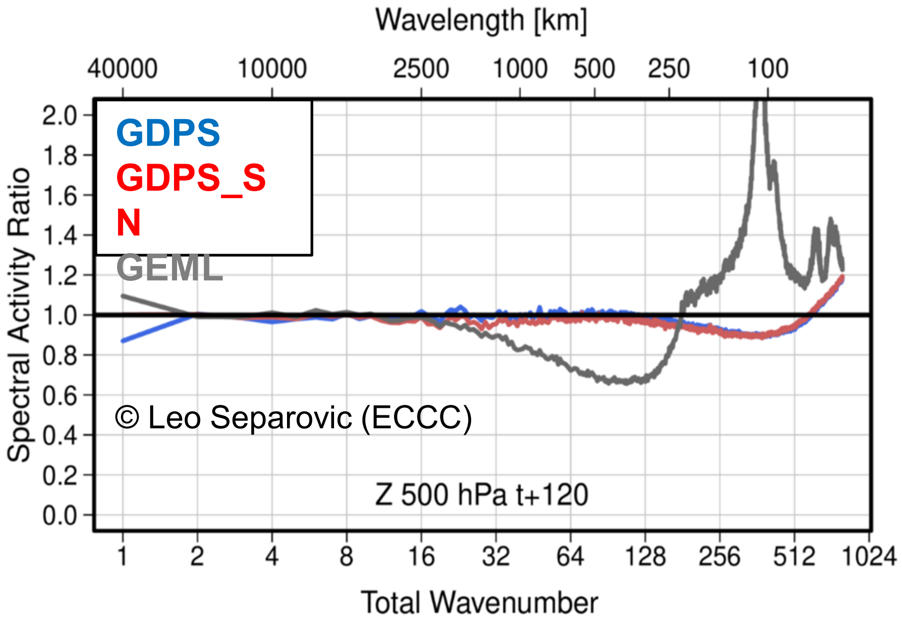
Traditional summary statistics are limited. Complementing them with spatial verification diagnostics can provide more in-depth knowledge of the forecast characteristics. Examples:

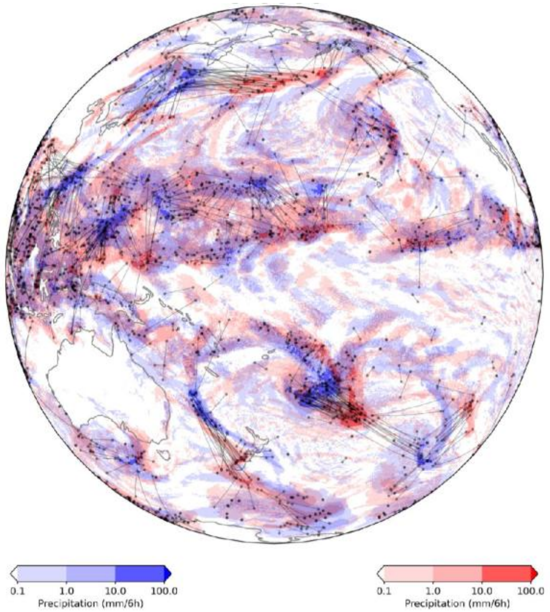
spectral decomposition diagnoses the forecast effective resolution and enables identifying unphysical compensating behaviors that are hidden when using integral metrics (cf. Leo Separovic talk).

Also other spatial metrics, such as the FSS and its modified versions, the wavelet-based SBE, and the SSMI were proposed as additional diagnostics (cf. B.Antonio talk).

UNIVERSITY OF OXFORD

Comparing AI-based models to traditional NWP models with distance metrics, such as the PAD (cf. Skok talk), is of high interest, given the positive results AI models exhibit on TC track position (cf. G.Chen talk) and correct large-scale positioning / phase error (cf. C.Subic and Leo Separovic talks).



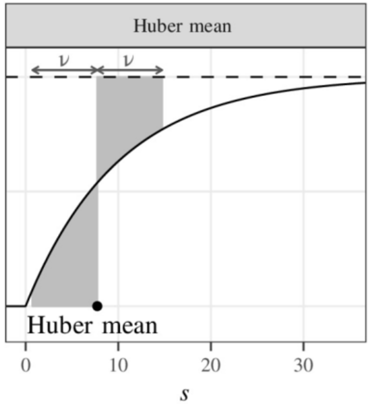


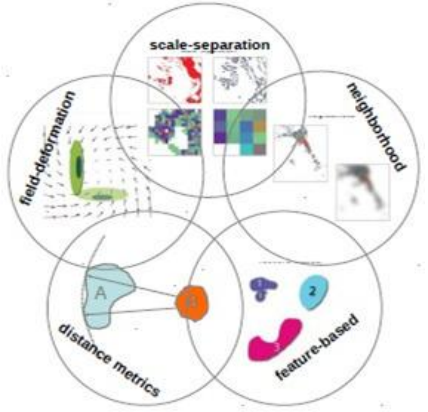
- Skok (2023): Precipitation Attribution Distance (PAD) Atm. Research
- Skok and Lledó (2025), *QJRMS.*

UNIVERZA V LJUBLJANI

Fair and Proper Scoring Rules for Ensembles

A large portfolio of **fair and proper scoring rules** has been developed in the past decade, and should be imported **into operational verification practices**, to fairly compare traditional NWP-based ensembles and the data-driven ensembles, which are advantaged by the significantly larger number of members (cf. J. Broeker, A.Dirkson talks).





Romain Pic: proper scoring rules have however some **limitation in interpretability** - need to render them more interpretable, while maintaining their statistical rigor! **Aggregation and transformation** principles can be leveraged for **bridging the gap** between **consistent and proper scoring rules** and **spatial verification methods**. Examples mentioned during the workshop:

- Verification in AI model lower dimensional "latent space" (cf. J. Broecker talk);
- Verification of deterministic forecasts in spectral space (cf. C. Subic and L. Separovic talks)
- Application of proper scoring rules to spectral components was also suggested for ensembles (Q&A following A.Dirkson talk)

Extreme Events

- **Benchmark extreme-event datasets** are being developed, for comparing AI versus NWP models by displaying verification results on common case studies:
  - ❑ Extreme Weather Bench (cf. A.McGovern talk) Brightband
  - ❑ ECMWF extreme event catalogue (L.Magnusson) ECMWF
  - ❑ EW4All "iconic" case studies WORLD METEOROLOGICAL ORGANIZATION
- The issue of gathering a sufficiently large sample for robust analysis of extreme events, without overlapping with AIWP training periods, was discussed.
  - In the previous VerAI workshop it was advanced the idea of artificially inserting into the training dataset more extreme events (e.g. storms or hurricanes), to help the data-driven models learn their dynamics; however, this would distort their climatological frequency.
  - In the VerAIops workshop, "AI Retraining Without Iconic Events (AIRWIE)" was a proposed solutio establish a leave-one-out cross-validation protocol for selected iconic events (cf. S.Nath talk). UNIVERSITY OF OXFORD
- Existing extremal dependence indices (e.g. Ferro & Stephenson, 2011, W&F) and consistent scoring functions (e.g. Taggart, 2021, QJRMS) should be adopted in operational verification practices, to avoid the forecast dilemma (Lerch et al, 2017, Stat Sci) leading to hedging (cf. ECCC D.Brunet talk).

Evolving AI-model verification practices in operational centers

Environment and Climate Change Canada
Environnement et Changement climatique Canada

- Data-driven AI/ML models must undergo the same protocol of standardized verification as the traditional physically-based NWP (cf. S.Beauregard, S.Wahl presentations)
- Additional diagnostics are being included into operational verification suites. Examples:
  - Spectral components verification (ECCC, DWD, ECMWF, …)
  - Attribution diagnostics such as the multiple correlation decomposition (cf. S.Wahl talk)
- Novel light-weight exportable verification packages (e.g. https://github.com/nci/scores) are developed in operational environments to enable rapid adaptation associated to AI/ML forecasts rapid evolution (cf. N. Loveday talk).

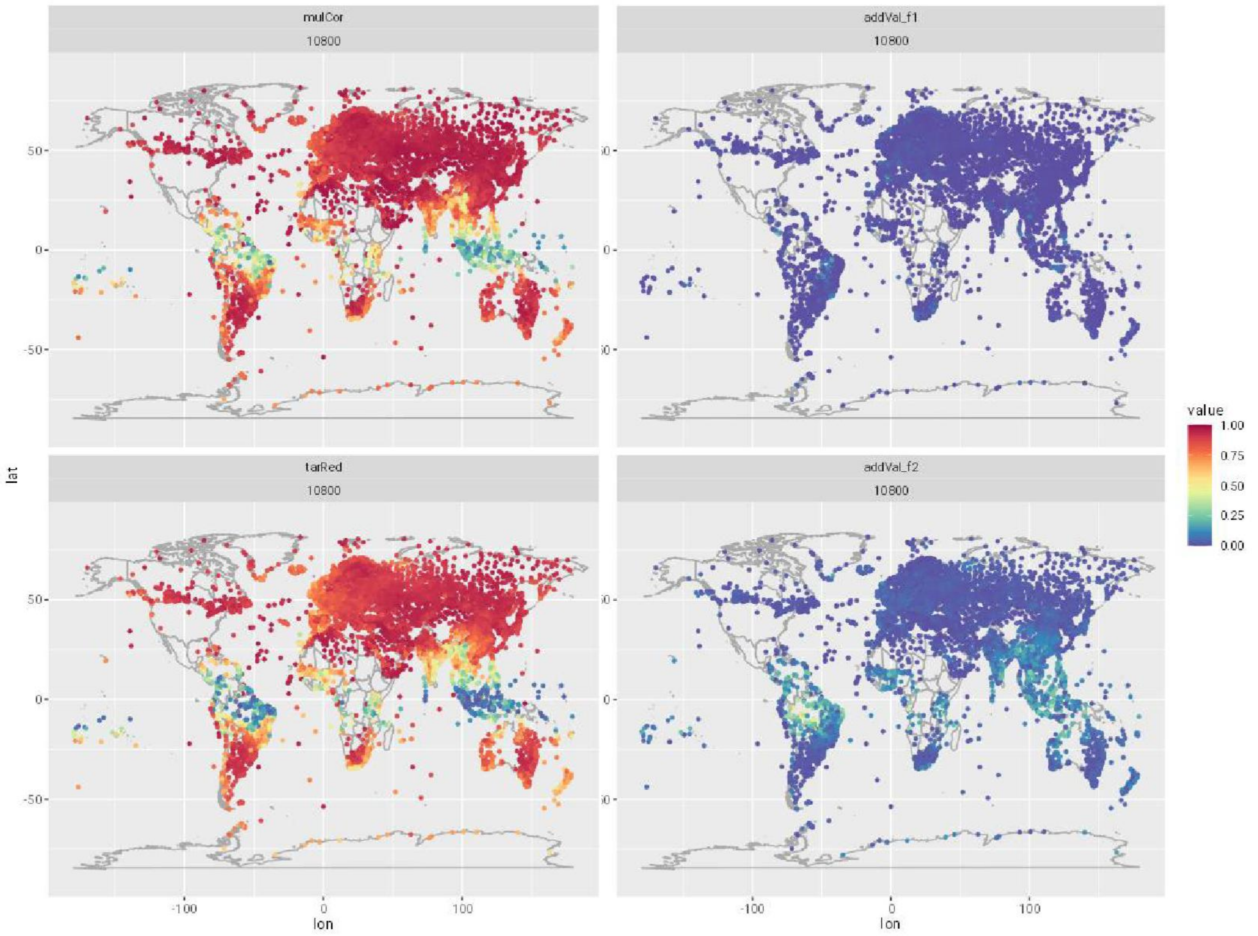


Image courtesy of Sabrina Wahl (DWD)

VerAIops workshop outcomes and ways forward

Verification activities within the WGNE Weather Prediction Model Intercomparison Project :

- WP-MIP website: https://www.wcrp-esmo.org/activities/wp-mip
- WP-MIP verification activities: https://github.com/WP-MIP/verification/wiki

Several researchers from international institutions have agreed on performing a coordinated verification exercise on the WP-MIP dataset.

Some of these studies are already shaping the new the WIPPS standardized verification practices.

WP-MIP github verification coordination table

Contributors	Research questions and plan	Theme	Github issue
…	…	Traditional verification	…
…	…	Explainability	…
…	…	Falsification	…
…	…	Spatial (scale-sep.)	…
Barbara Casati (ECCC), Marion Mittermaier (UK Met Office), Sebastian Lerch (Marburg Univ.), Joel Stein, Fabien Stoop, Maxime Taillardat (Meteo France)	Compare the neighbourhood CRPS (Stein and Stoop, 2023) versus potential CRPS (Gneiting et al 2025 arxiv) for the verification of smooth fields.	Spatial (neighbour)	neighbour-vs-potential CRPS
…	…	Extremes	…
…	…	Tropical cyclones	…
…	…	Sub-seasonal to seasonal	…
…	…	Regional studies	…

Conclusions!

Explore (to advance and improve): analyzing the behavior and characteristics of ML prediction is fundamental to understand strengths and limitation of this new prediction archetype:

- Falsification and explainability + physical coherence
- Summary scores are limited – spatial verification & novel diagnostics

New role of verification metrics: from assessment to AI training:

- this comes with inherent responsibilities, since the choice of the metric shapes the characteristics of the produced data-driven forecast!!!
- AI-training revisited as meta-verification tool: confirms stuff we knew (RMSE & smoothing), reveals stuff we did not foresee …


Thank you!