26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

# Towards a sustainable workflow for post-processing development and evaluation

**Benjamin Owen**[1], **Esteban Abellan**[2], **Robert Johnson**[3], **Daniel Mentiplay**[4], **Belinda Trotta**[4], **Jennifer Whelan**[4]

[1] *Bureau of Meteorology, Adelaide, Australia.*
[2] *Bureau of Meteorology, Sydney, Australia.*
[3] *Bureau of Meteorology, Hobart, Australia.*
[4] *Bureau of Meteorology, Melbourne, Australia.*

*Email:* benjamin.owen@bom.gov.au

**Abstract:** In August 2024, the Bureau of Meteorology migrated to using IMPROVER (IMPROVER Collaborators, 2024; Roberts et al., 2023; Trotta, et al., 2024) as the basis for its rainfall forecasts. One of the key changes in this migration was an update to the underlying calibration methodology, adopting a new machine learning method called RainForests (Trotta, Owen, et al., 2024). This transition resulted in direct benefits including more accurate 3-hourly rainfall forecasts, better calibration of the daily rainfall accumulations at the 0.2 mm (rain/no-rain) threshold and improved consistency between daily and sub-daily rainfall. In turn, this has eliminated the need for routine intervention by forecasters to correct deficiencies that existed in the legacy rainfall guidance, resulting in significant time-savings for operational meteorologists. Despite this success, feedback from users has presented some shortcomings associated with less frequently observed events involving high-extreme rainfall amounts, warranting refinements to the calibration methodology.

One of the key challenges in establishing and then configuring IMPROVER at the Bureau was managing the complexity of both the system and interrogating the array of outputs from a system that processes 10,000's of files and produces TBs of data with each forecast cycle. Managing this complexity made it difficult to run the system exactly akin to that in the production environment. With IMPROVER now embedded as the Bureau's primary post-processing platform, considerable effort has been made to establish an efficient, reproducible workflow through which to test and interrogate changes to IMPROVER processing chains. This work includes a more explicit expression of the processing chains, shrewd filtering of the directed acyclic graphs that constitute these processing chains to limit evaluation to only the steps of interest and their predecessors, the adoption of a standard verification workflow based on the scores verification package (Leeuwenburg et al., 2024), and the use of template repositories to format experiments in a standardised way. The goal here is to build an efficient, reproducible workflow that enables rapid iterations and allows developers to focus on post-processing activities.

Here we present our recent efforts to address the shortcomings raised in RainForests calibrated guidance in a systematic way utilising this workflow. We consider three different adaptions to the RainForests approach and contrast and compare these to the current operational version. Through this lens, we showcase how this workflow has enabled us to effectively manage the large volume of experimental hindcast data and through the wide array of verification tools afforded by the scores package, tease out the characteristic differences each candidate solution affords to addressing the problem, particularly in the context of high-extreme rainfall. Through consideration of both case studies and bulk metrics, we evaluate the best approach to improving RainForests calibration that retains the existing benefits but preserves or improves upon the skill of the uncalibrated guidance for rarer extreme rainfall events.

We also present on the use of spatial based verification metrics, namely the Continuous Rain Area (CRA) metric (Ebert & Gallus, 2009), which through this work has been integrated into scores. Rather than assessing each grid point in isolation, CRA treats rainfall as spatial objects. It aligns forecasted and observed rain areas by shifting them within a search window to minimize mean squared error, then decomposes the total error into three components: location error (displacement), volume error (rainfall amount), and pattern error (structural mismatch). We show how this decomposition can provide a more interpretable and physically meaningful assessment of forecast performance.

 

## REFERENCES

Ebert, E. E., & Gallus, W. A. (2009). Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Weather and Forecasting*, *24*(5), 1401–1415. https://doi.org/10.1175/2009WAF2222252.1

IMPROVER Collaborators. (2024). *IMPROVER Software Repository*. https://doi.org/10.5281/zenodo.8410114

Leeuwenburg, T., Loveday, N., Ebert, E. E., Cook, H., Khanarmuei, M., Taggart, R. J., Ramanathan, N., Carroll, M., Chong, S., Griffiths, A., & Sharples, J. (2024). scores: A Python package for verifying and evaluating models and predictions with xarray. *Journal of Open Source Software*, *9*(99), 6889. https://doi.org/10.21105/joss.06889

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C., Trzeciak, T., Abernethy, P., Beard, L., Crosswaite, N., Fitzpatrick, B., Flowerdew, J., Gale, T., Holly, L., Hopkinson, A., Hurst, K., Jackson, S., Jones, C., Mylne, K., … Worsfold, M. (2023). IMPROVER: The New Probabilistic Postprocessing System at the Met Office. *Bulletin of the American Meteorological Society*, *104*(3), E680--E697. https://doi.org/10.1175/BAMS-D-21-0273.1

Trotta, B., Canvin, J., Gale, T., Hume, T., Johnson, R., Liu, J., Mentiplay, D., Owen, B., Schubert, A., Weymouth, G., & Whelan, J. (2024). *An initial benchmarking of IMPROVER - part 2: evaluation of precipitation diagnostics*.

Trotta, B., Owen, B., Liu, J., Weymouth, G., Gale, T., Hume, T., Schubert, A., Canvin, J., Mentiplay, D., Whelan, J., & Johnson, R. (2024). RainForests: A Machine Learning Approach to Calibrating NWP Precipitation Forecasts. *Weather and Forecasting*, *39*(11), 1715–1732. https://doi.org/10.1175/WAF-D-23-0211.1

***Keywords:*** *Rainfall calibration, ensemble post-processing, reproducible research, machine learning*

 