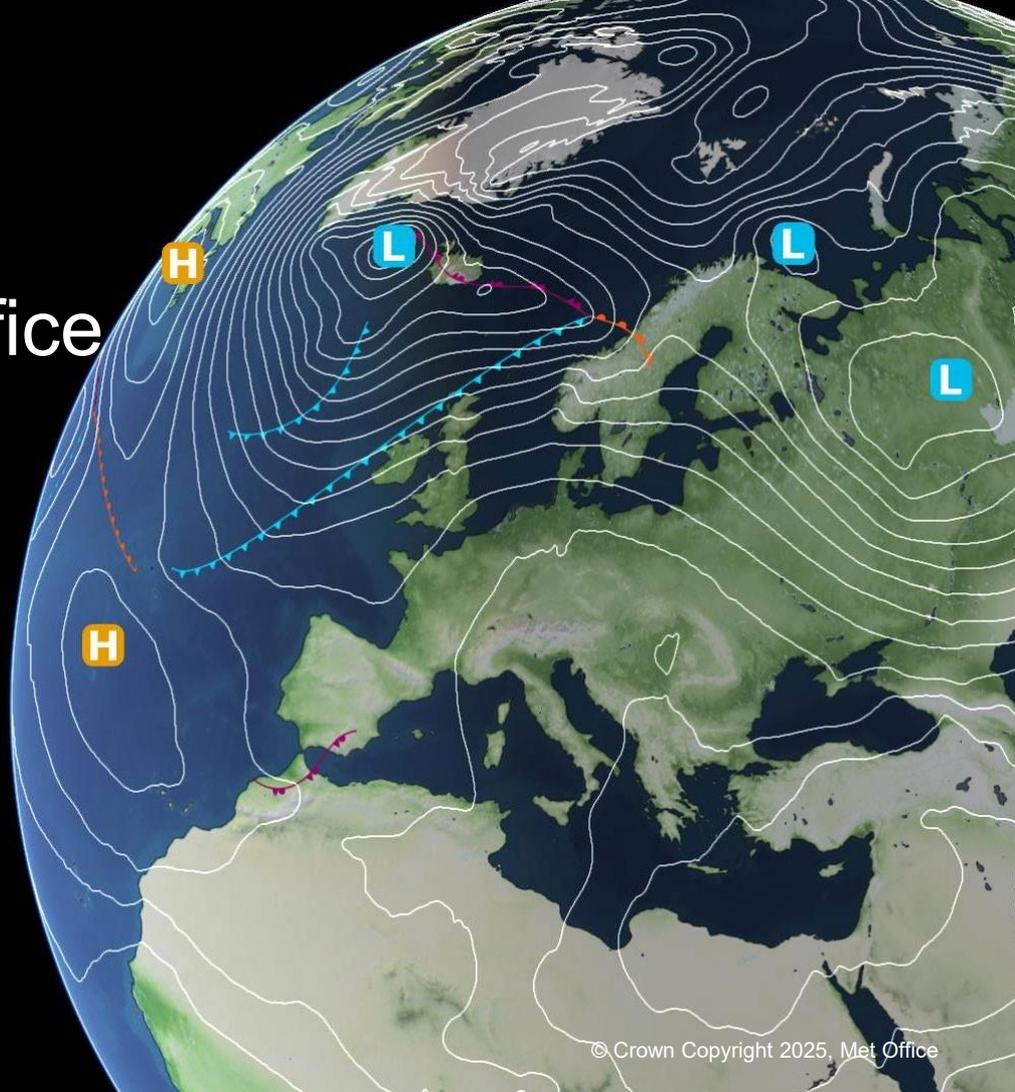# An example evaluation process in Research-to-Operations at the Met Office

Rachel North, Met Office

47th EWGLAM and 32nd SRNWP meeting
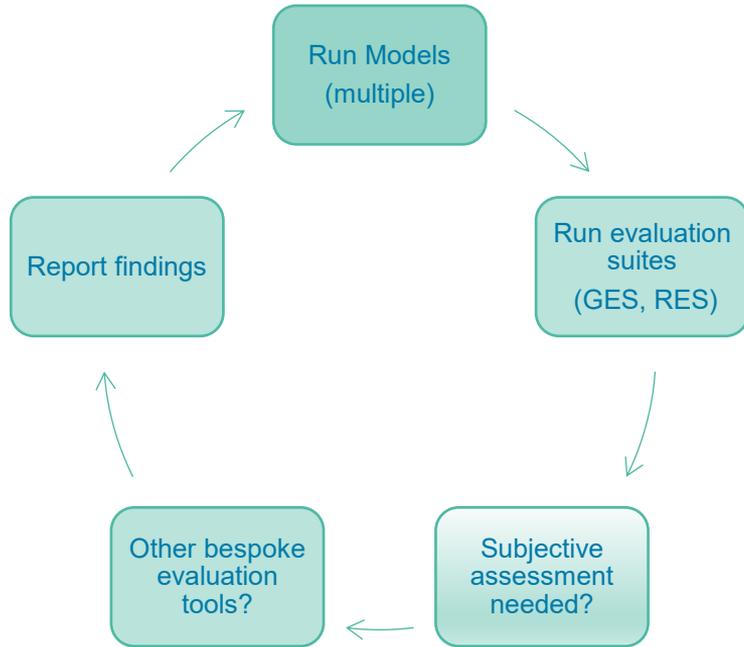
23rd September 2025

With thanks to all in R2O and beyond at the Met Office who have contributed information here

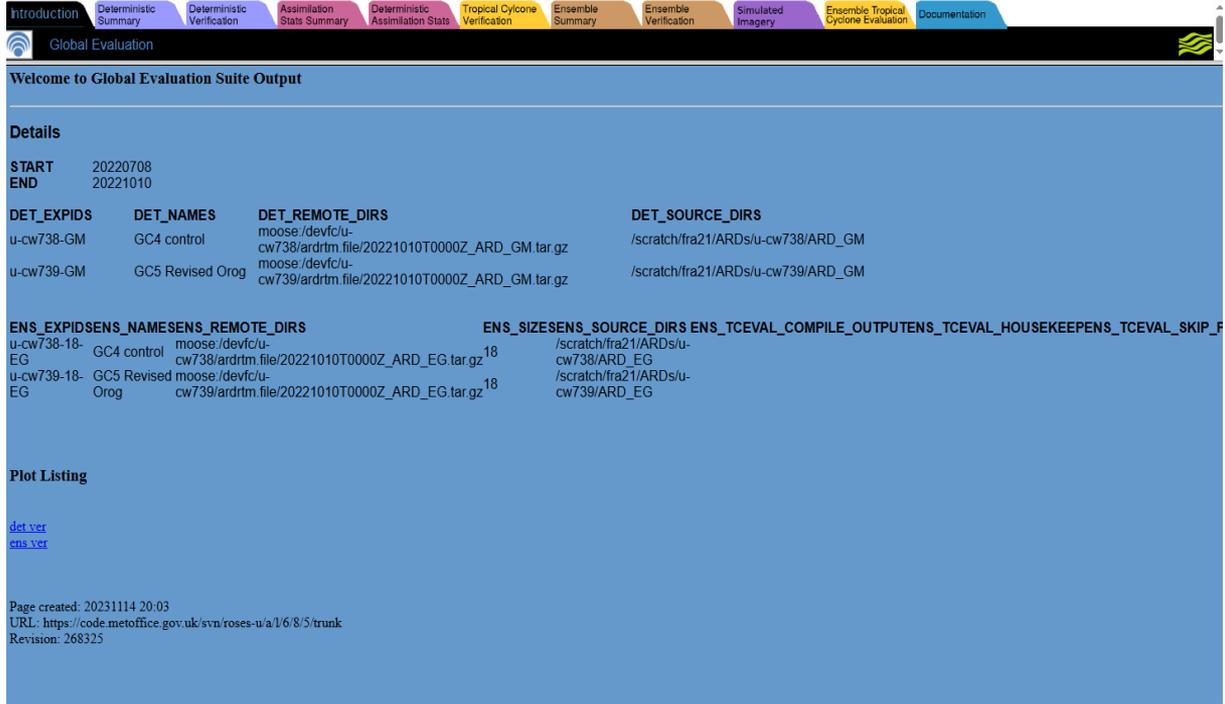# Evaluation of Parallel Suites in R2O

Met Office

Components and package trials

- Details recorded on code repository (configuration management with associated tickets and documentation)
  - Is migrating to GitHub
- Evaluation suites (to produce output)
  - GES, RES currently
  - Intend to integrate CSET into regional evaluation
- Subjective assessment done in conjunction with O2R meteorologists
  - Mainly done at final package stage

Aspects of assessment

- Use all available information (e.g., UK climate pages, ASXX charts, satellite imagery, radar data)

- Subjective assessments involve looking at model fields and corresponding observations for all model initialisations within the trial period – separate tool for visualisation

During the process for producing a parallel suite

Observation characteristics through the trial period (and relative to climate)

Objective Statistics (in 'evaluation suites')

Subjective assessment for model trial periods

Performance of models for a particular event/weather regime

Supplementary techniques

(e.g. mapped scores, scatterplots, difference tracking etc.)

Trial Evaluation (objective results)

- Global Evaluation Suite and Regional Evaluation Suite
  - Migrating to CSET for regional evaluation
- Run with trials/parallel suite
- Example GES summary information page
  - Experiment IDs
  - Location of verification results files (traceability)
  - Both deterministic and ensemble

GES plot types

- Summary scorecards
  - Deterministic & ensemble
  - Different areas
  - Different truth types
- 'Standard' set of verification metrics
  - Deterministic & ensemble
  - Against observations and analyses
  - Includes profiles
- Plus, assimilation statistics (deterministic)
- Simulated Imagery

Warren Tennant

Met Office — Regional evaluation: PS47 examples (UKV)

- RES
  - Summary of trials (as in GES)

- 'Standard' set of metrics, HiRA, FSS

- Ensemble and deterministic run separately

- Plus 'diagnostics' (descriptive statistics plus parameter-specific methods)

- Extra evaluation done alongside (e.g., fair scores, lightning, subjective assessments)

Aurore Porson, Gareth Dow, Adrian Lock, Marco Milan, Lee Hawkness-Smith, Carlo Cafaro, Seb Cole, Adrian Semple, Stephen Gallagher, Tomos Evans, Paul Field
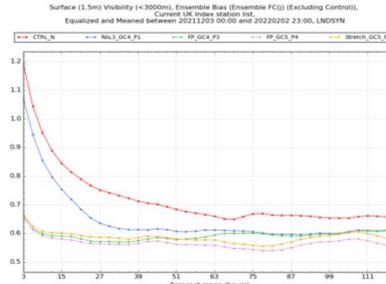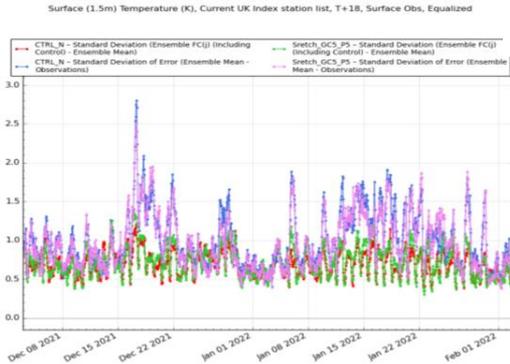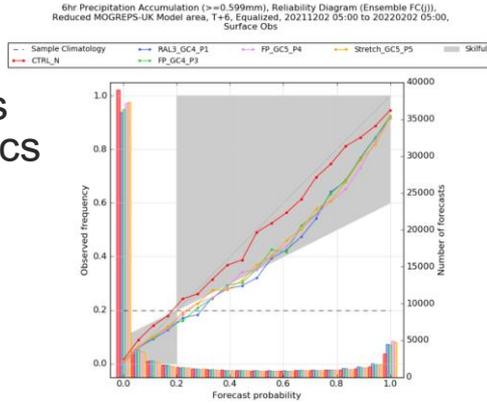
Regional Evaluation: PS47 ensemble

- RES
  - As for UKV plus ensemble metrics (e.g., attributes diagrams, rank histograms)
  - Scorecard summaries

- OVELAYS
  - For subjective comparison of model fields and observations

- Extra capability developed
  - e.g., eFSS/dFSS
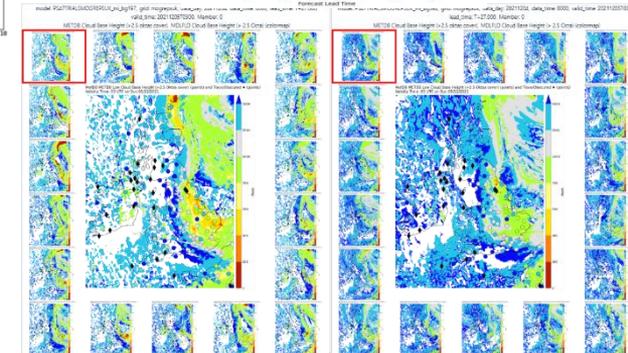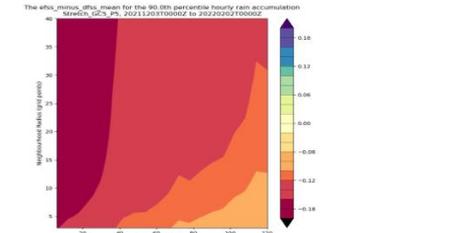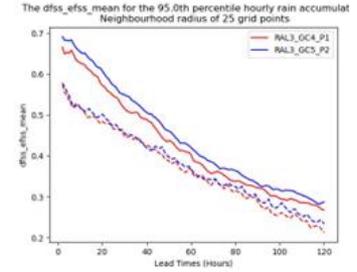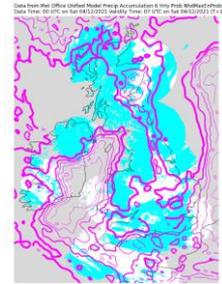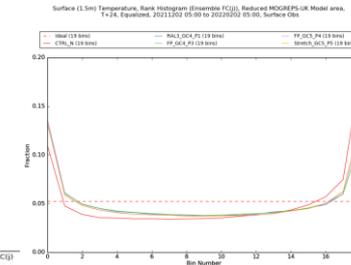
Summary

- Approach(es) continuously reviewed and evolving
  - e.g., for ensembles first approach
- New systems/methods/diagnostics/metrics always need pulling through
  - e.g., METplus, CSET, fair scores, etc.
- Are these ways of working still suitable for
  - Very high-resolution models?
  - AI models?
- Differences/Similarities to your ways of working?
- Where are the gaps?
- What will remain difficult?

Met Office

Questions?

rachel.north@metoffice.gov.uk