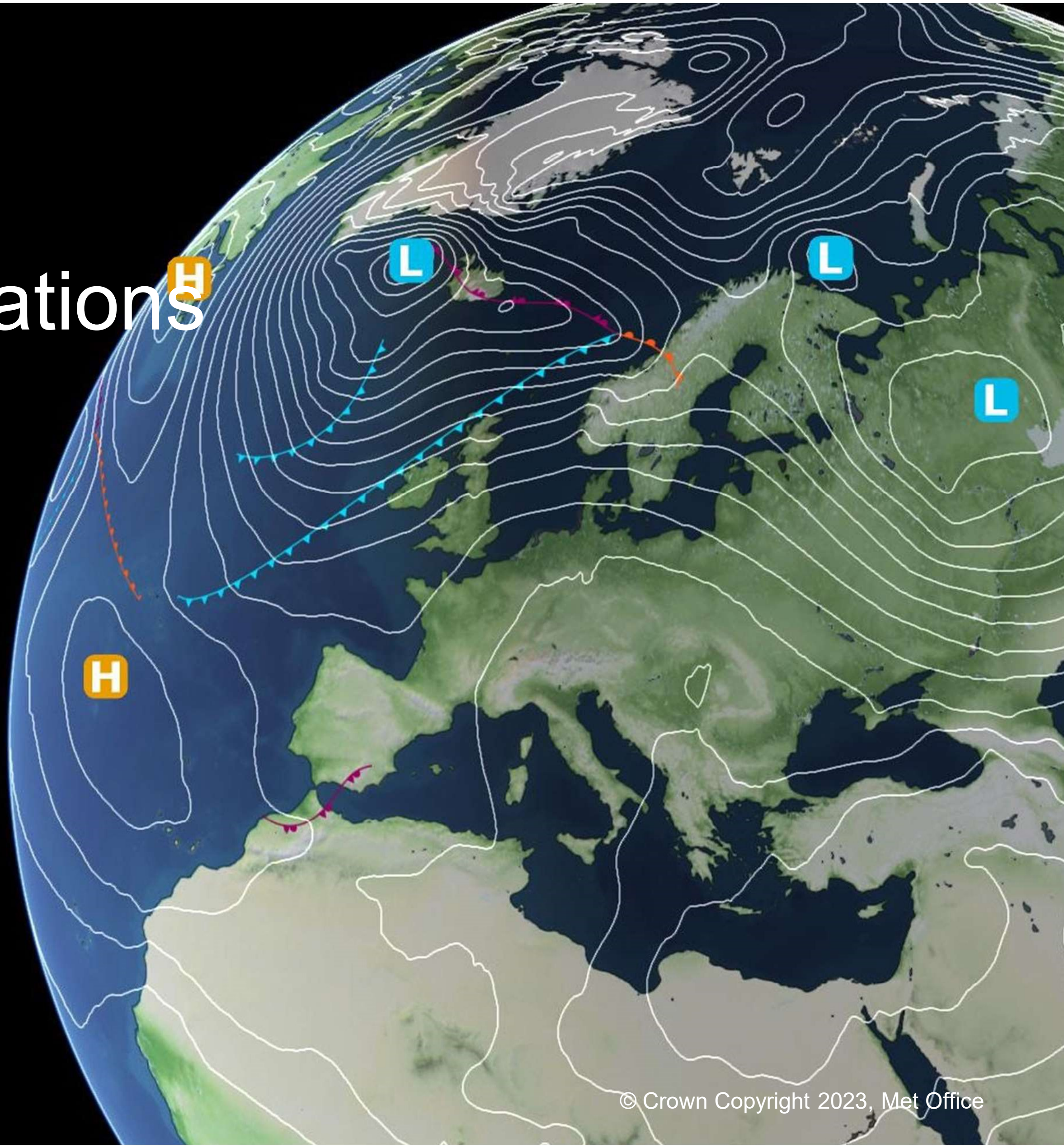
Met Office

How metrics and observations affect model comparison

Marion Mittermaier, Rachel North,
Aurore Porson, Nigel Roberts

September 2023

UK Met Office



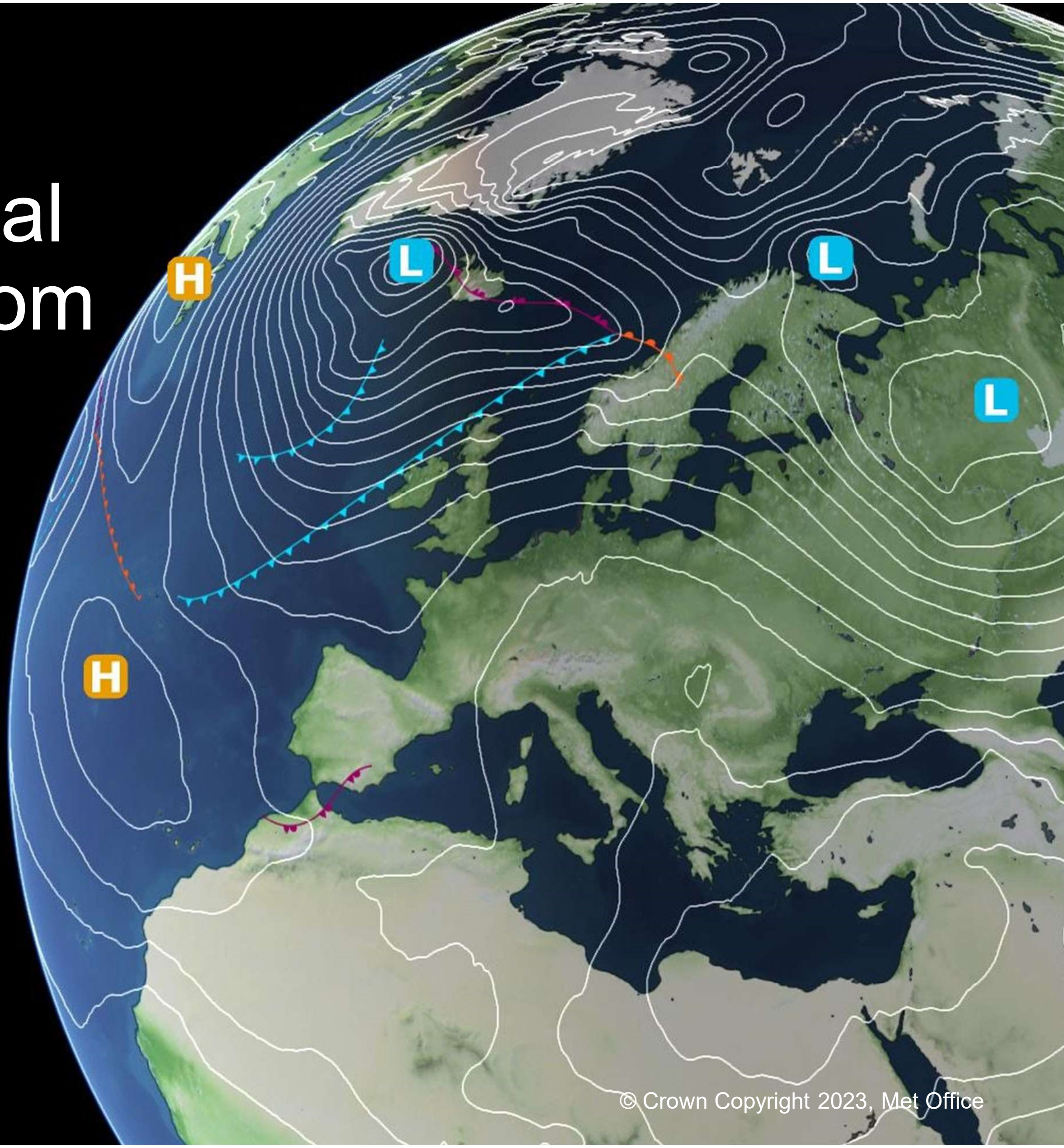
Met Office

Interpretation of RAL3 trial results and the impact from verification scores

Rachel North, Marion Mittermaier, Aurore Porson, Nigel Roberts

September 2023

rachel.north@metoffice.gov.uk

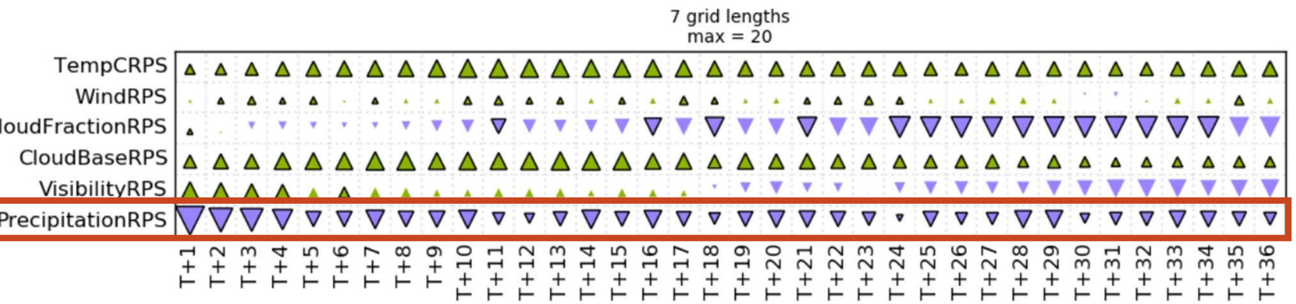


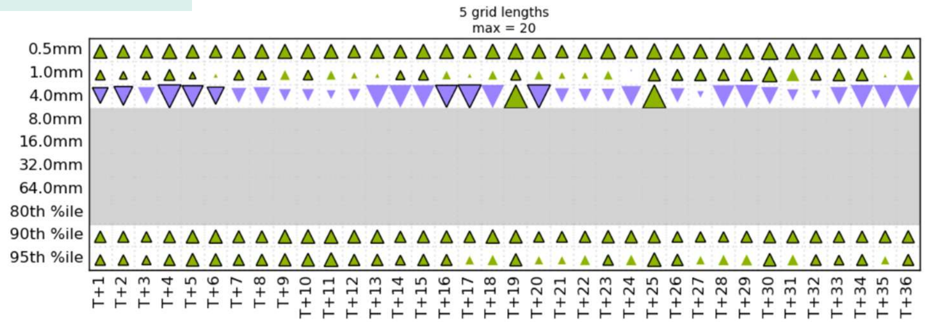
Initial verification scorecards

HiRA scorecards: hourly precipitation 7x7 neighbourhood

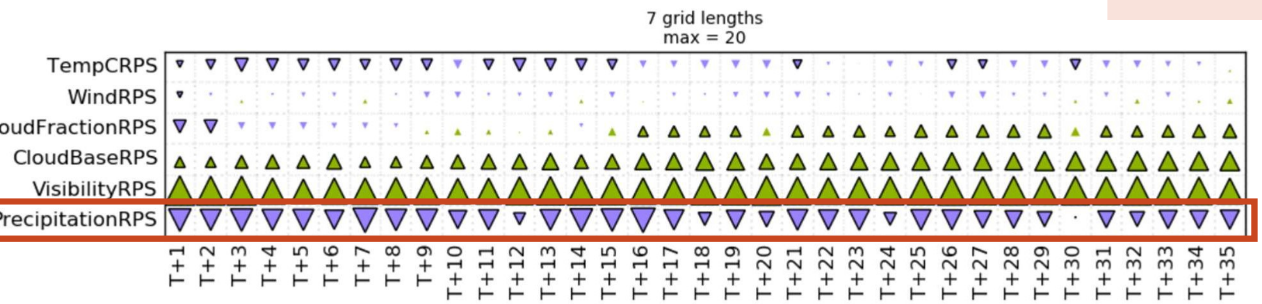
FSS scorecards: hourly precipitation 5x5 neighbourhood

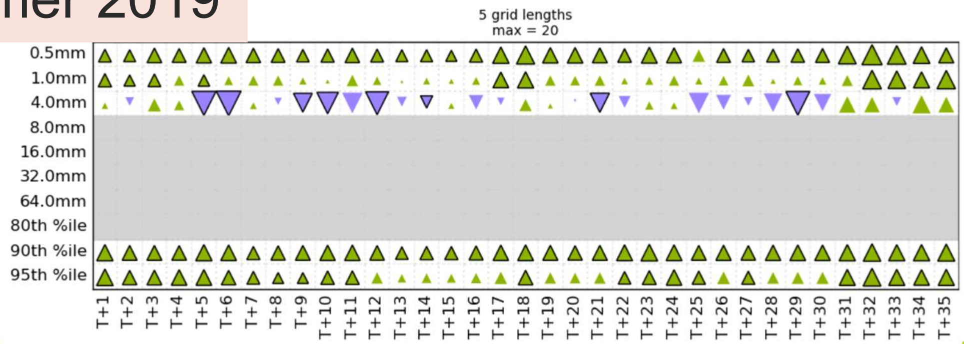
Winter 2021/22





Summer 2019





Met Office

Further Investigation

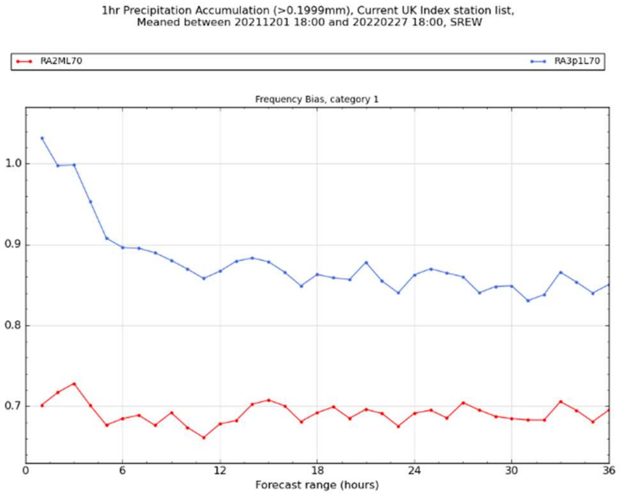
- Small team to dig into results to provide an explanation
- Found gaps in current evaluation process
- Produced recommendations to help with future trial evaluation

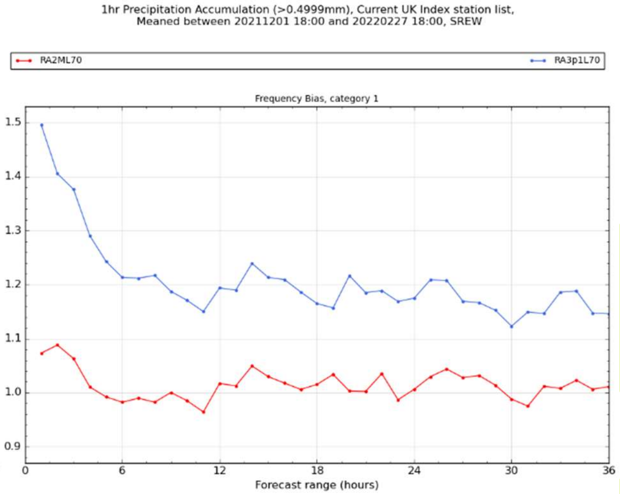
Aspects examined:

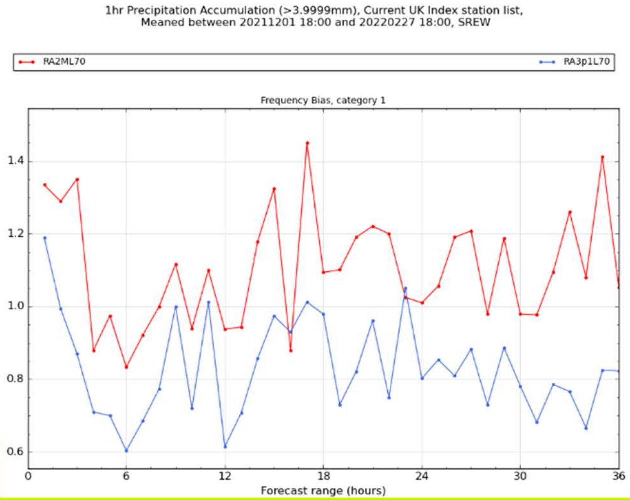
1. Making sure we believe what we see (checking the calculations for the scores)
2. Looking deeper at the scores (split by threshold)
   - ➢ Reasons for the differences
3. Making the Hinton plot triangles more appropriate for the FSS
4. Looking at the effect of bias on the verification metrics
   - ➢ Introducing the AFSS
5. Considering differences from radar error and disparity of gauges that affect the FSS
6. Considering thresholds for the FSS and bins for the RPS
7. Looking at updated trial results for any change in signal

Met Office

Further investigation: Frequency Bias







Winter 2021/22

At lowest threshold RAL3 improves the (dry) bias

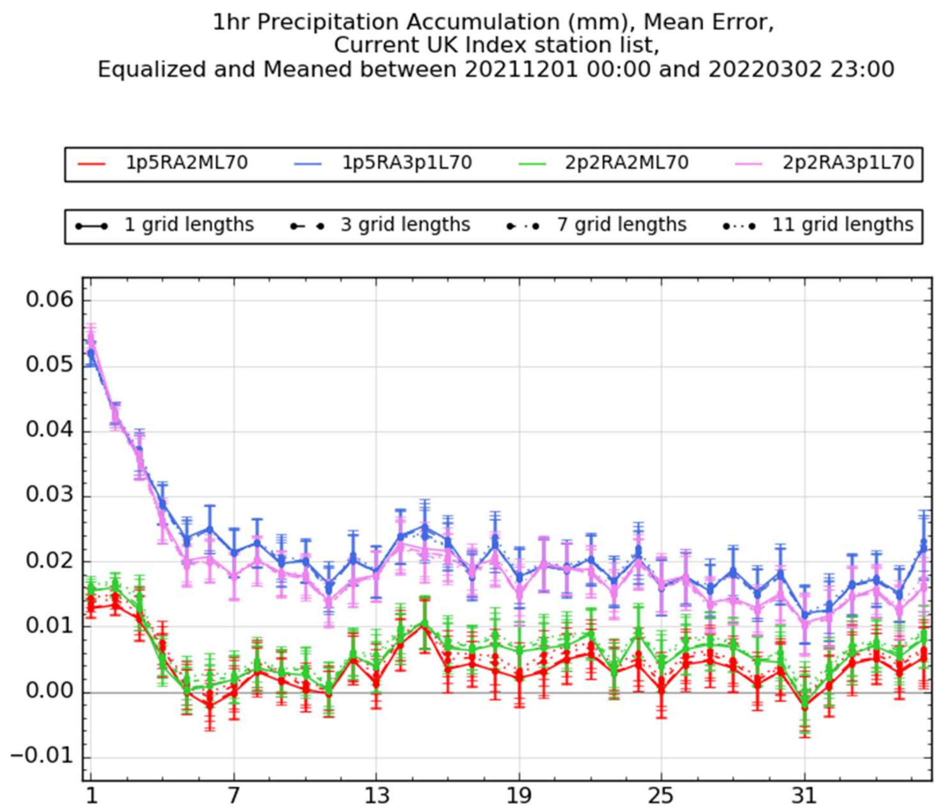
At the low thresholds RAL3 over-forecasts

At the high thresholds RAL3 under-forecasts

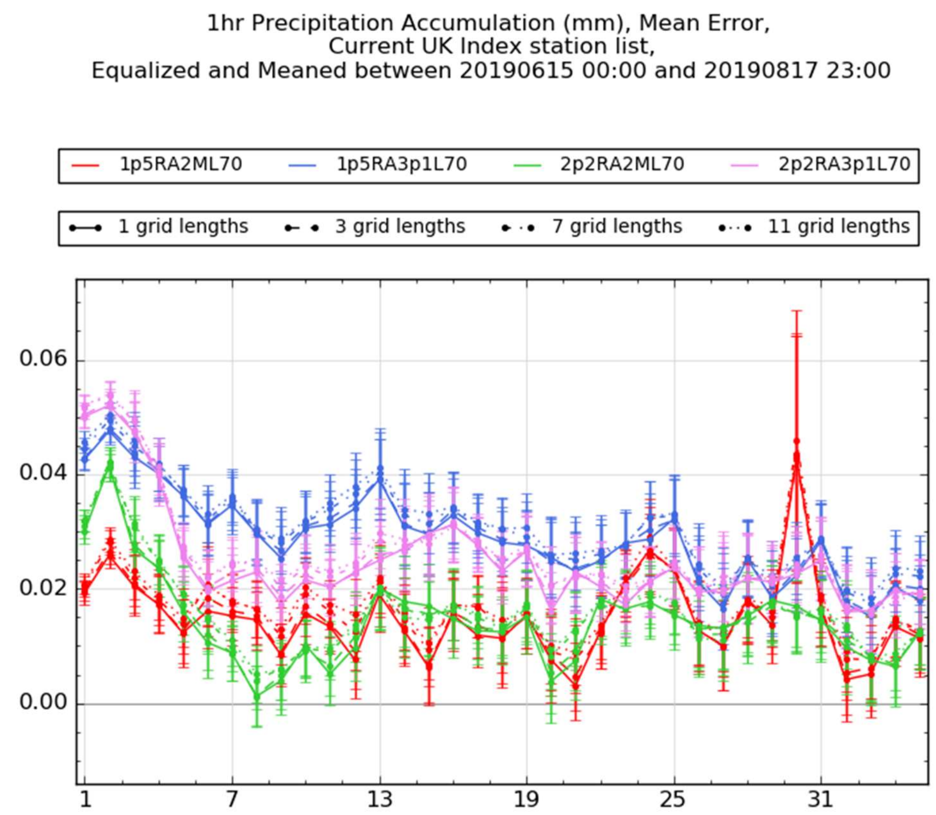
Sampling Errors

Further investigation: HiRA Bias

Winter 2021/22



Summer 2019

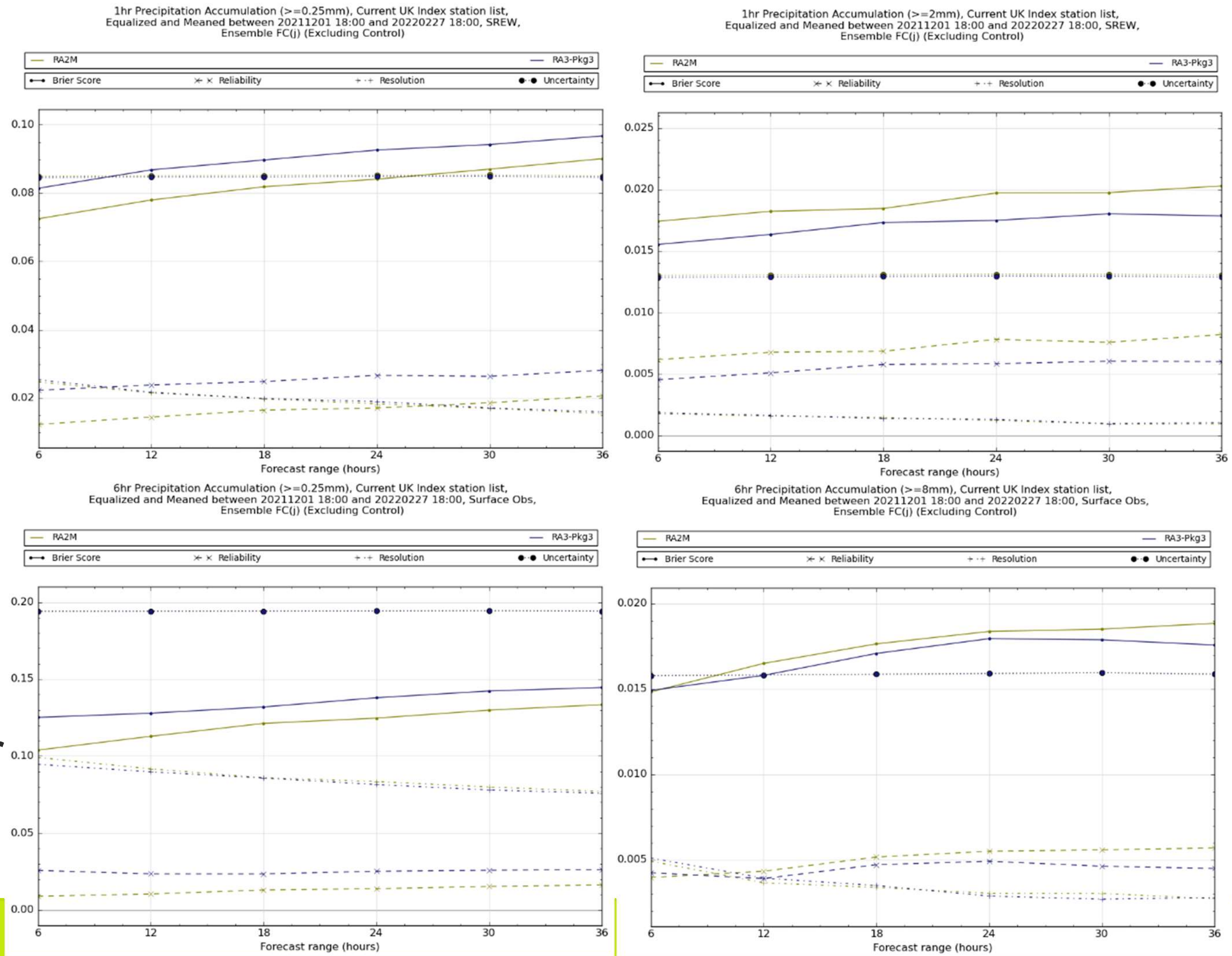


Met Office

Further investigation: Brier Score

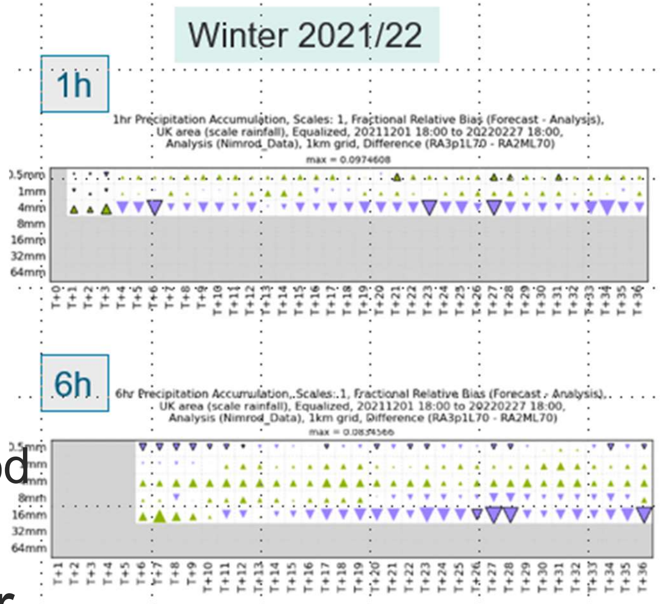
Winter 2021/22

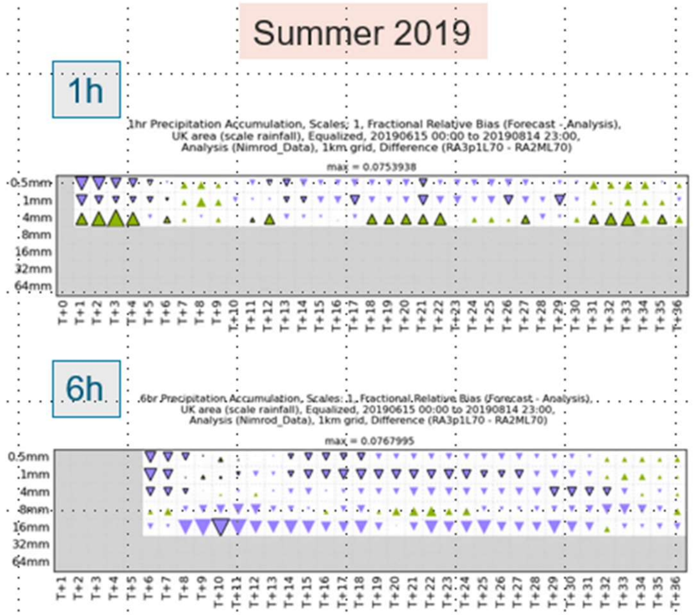
- At lower thresholds score is worse for RAL3
- Higher thresholds:
  - RAL3 better
  - But observation issues likely to dominate (e.g., gauge missing localised heavy rain)
  - In addition to event sampling issues
- Behaviour swaps at different thresholds for winter/summer and accumulation period



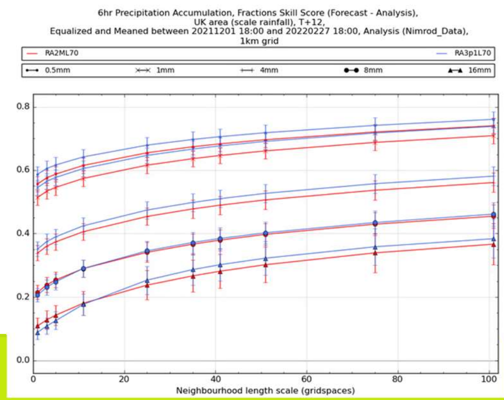
Further investigation: AFSS

- Scorecard to see change in Asymptotic FSS
- Scale series to see where the bias asymptotes
    - Led to recommendation on changing routine neighbourhood size settings for trials
- Mixed picture from AFSS for different accumulations and seasons

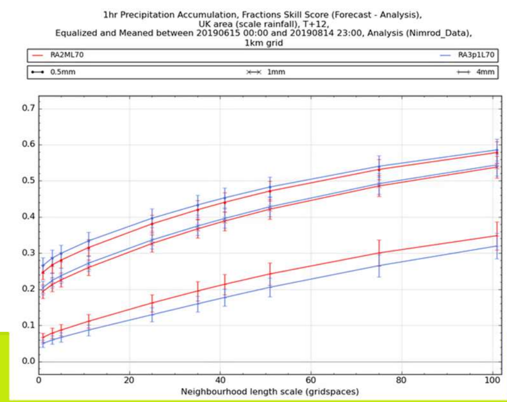




6h



1h

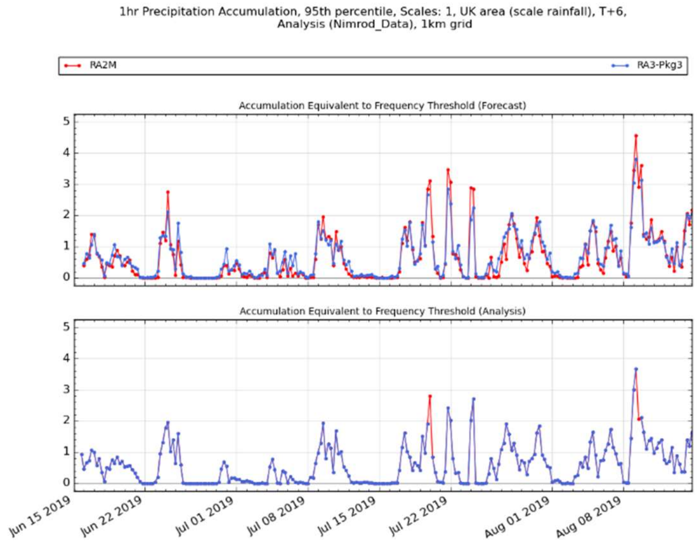


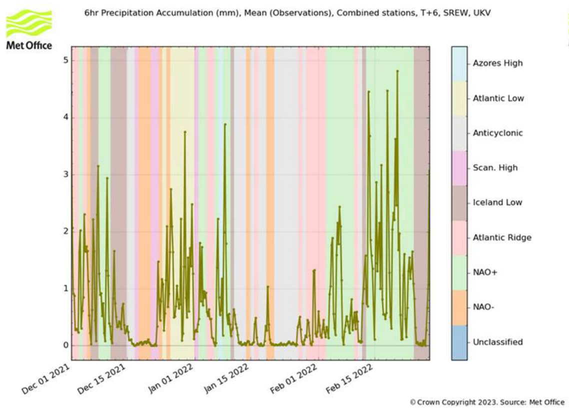
Winter 2021/22

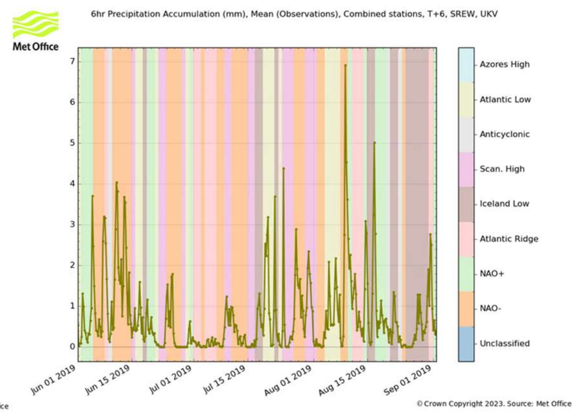
Met Office

Other points to note

- 95th centile for hourly accumulations ~0.5-2mm for both summer and winter trial
- 95th centile for 6-hourly accumulations ~1-10mm.
- Time series plots help to see when differences may be expected
- Useful to have context of weather within the trial period







So, what have we concluded…

For RAL3 trials

- The increased light rain seen in RAL3.1 leads to a too high fractional coverage bias at the larger spatial scales in the FSS
- For the low thresholds, at small spatial scales there is improvement which we think is because the higher coverage might lead to a greater likelihood of more intersection of the objects from the model forecast and radar (in the FSS)
- The model wet bias has increased with RAL3.1 (HiRA bias, frequency bias at lower thresholds)
- Might expect existing configuration to be favoured by HiRA due to RAL3 increasing coverage.
    - Since SO-NF can favour under-forecasting coverage

So, what have we concluded?…

Generally, for the evaluation process:

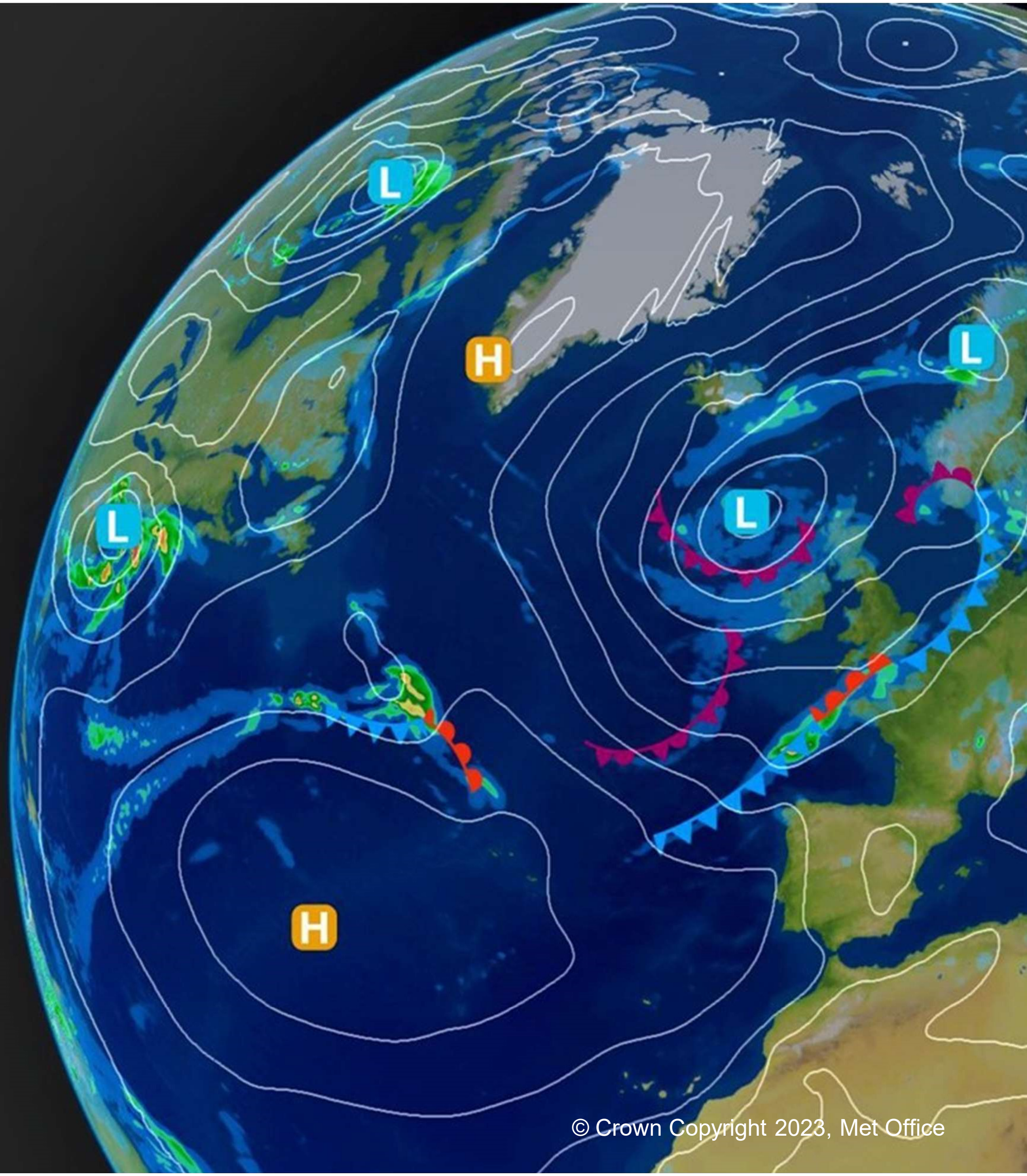
- Important to have weather/climate context for trials
  - For example, regime time series, comparison to climate, type of weather
- Look at the distributions!
- Observations need monitoring alongside
  - Radar and gauge monitoring for the time period of the trial
  - Observation uncertainty needs further investigation
- Need to ensure we don’t just take the ‘summary’ view
  - Look at the underlying verification score time series behaviour
- Ensure that FSS scorecards are presented with both absolute and percentage difference forms
- Being mindful of the skilful spatial scales when interpreting results

Met Office

(L)FSS-RPS comparison: A comparison of observations and neighbourhood methods

Marion Mittermaier

EWGLAM 2023



Met Office

Observations

- Essential for verification.
- No observation is perfect.
- **Characteristics need to be understood.**
- **QC is important**.
- **Forecasts ought to be well posed** to facilitate matching with observations.
- **Observational uncertainty** should be incorporated in whatever way possible.
- **Representativeness error is <u>only</u> <u>one</u> component** of observation uncertainty

So, what is it?

Error/uncertainty sources

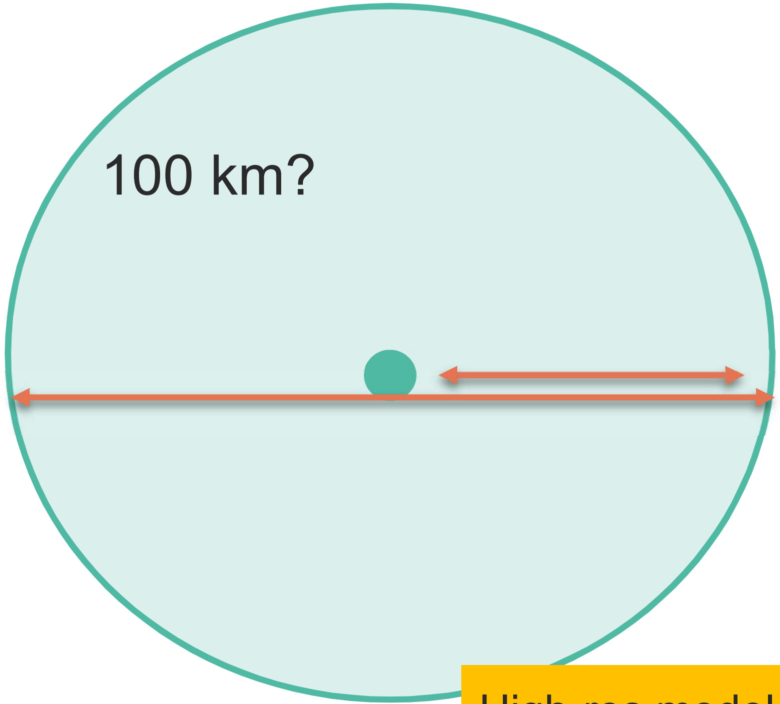
- Biases in frequency or value
- Instrument error
- Random error or noise
- Reporting errors
- Reporting of errors
- Precision error
- Conversion/derivation error
- ***Representativeness error***
- Analysis error
- Forecast error

JWGFVR training notes

Met Office

Representativeness

There are two sources… First:

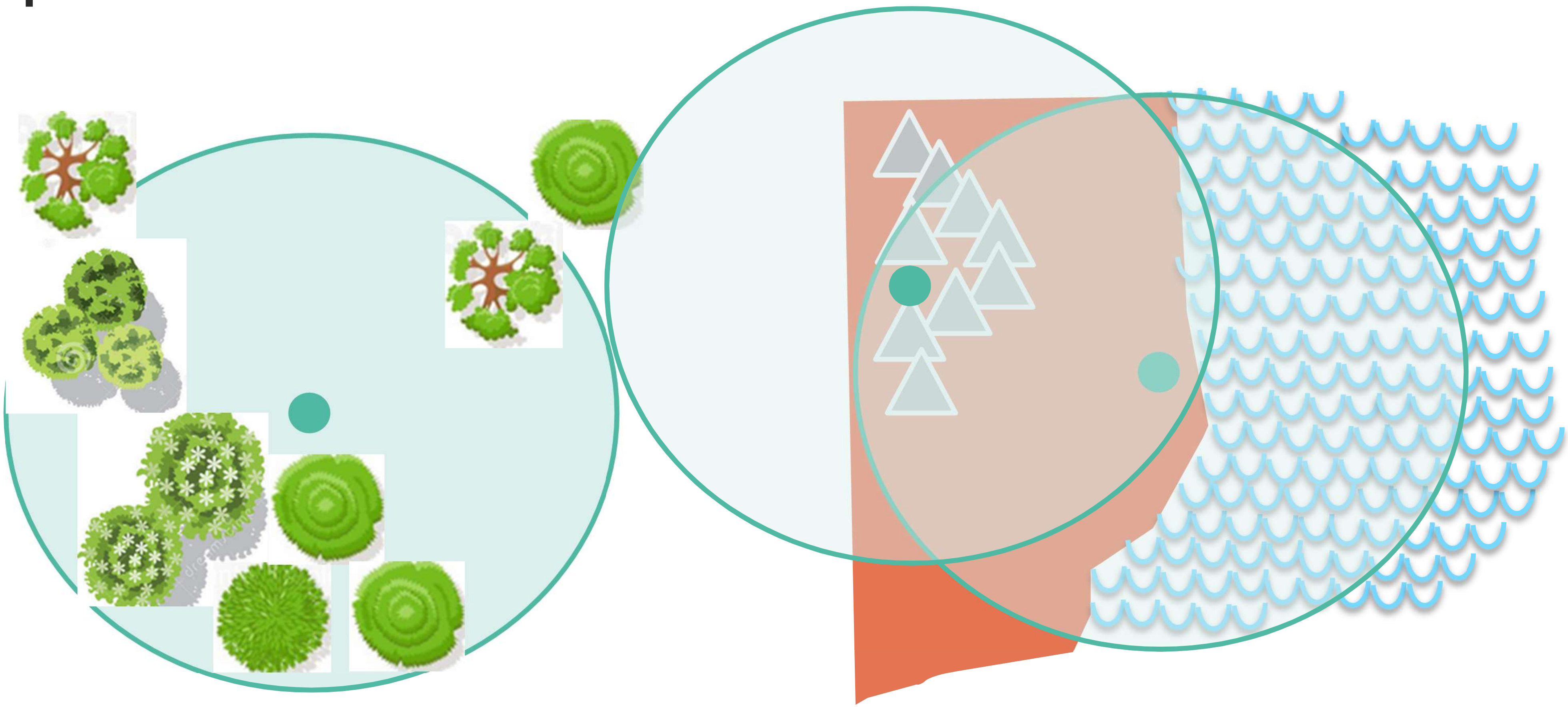


WMO-No.8 (2021) says: *"The **representativeness of an observation** is the degree to which it accurately describes the value of the variable needed for a specific purpose. Therefore, it is not a fixed quality of any observation, but results from **joint appraisal of instrumentation, measurement interval and exposure** against the requirements of some particular application. For instance, **synoptic observations** should typically be representative of an area up to *100 km around the station, but for small-scale or local applications the considered area may have dimensions of 10 km or less."***

High-res models have taught us a lot about how variable surface parameters can be

*This is not well written because it does not define the units properly.

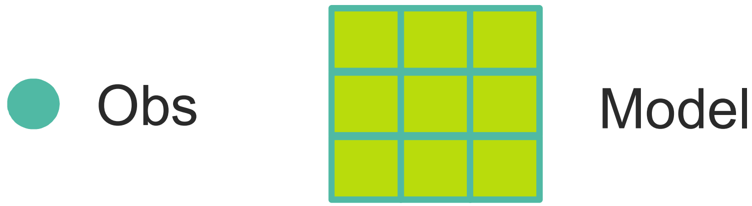
Representativeness



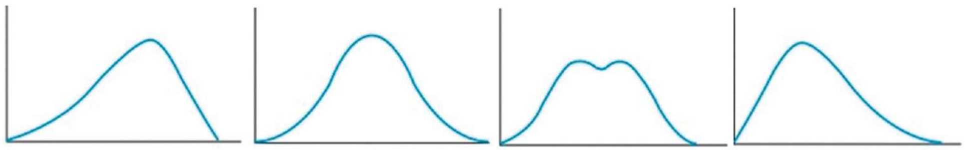
WMO definition is true in *only* the broadest / vaguest sense. **Any anisotropy** (coasts, mountains, vegetation changes, urban etc) negates this very quickly….

Met Office

Representativeness



Second… a **model grid box is not a point**.



It is always some form of *area* average
of an unresolved sub-grid-scale distribution, which depends on the variable.

Model resolution is one of the factors that determines the width of the distribution.
For some variables, e.g., cloud fraction, as the area of the grid box → 0

the cloud fraction → 0 or 1, i.e., a binary response. Either there is cloud or there isn't.

[*Aside:* a human observer takes a hemispheric view of cloud to derive a cloud fraction, but in an automated cloud fraction derived from a vertically pointing ceilometer can only be derived as a temporal aggregate. An instantaneous reading will either be cloudy or cloud free. The way an observation is made/taken is another component of observation uncertainty.]

Representativeness errors arise because a model grid box value, which is a grid-box mean, and a point observation are not really sampling the same thing, and are inappropriately matched, for whatever reason.

Why does observation error/uncertainty matter?

It can influence our ability to identify an event in observation space, thus affecting our ability to diagnose the **predictability of such an event.**

For ensemble forecast, and probabilities in particular: **where does forecast uncertainty *end* and observation uncertainty *begin*?** The observations are <u>not</u> absolute in detecting an event.

Different observation types of the same parameter (e.g., manual or automated) can provide very different results

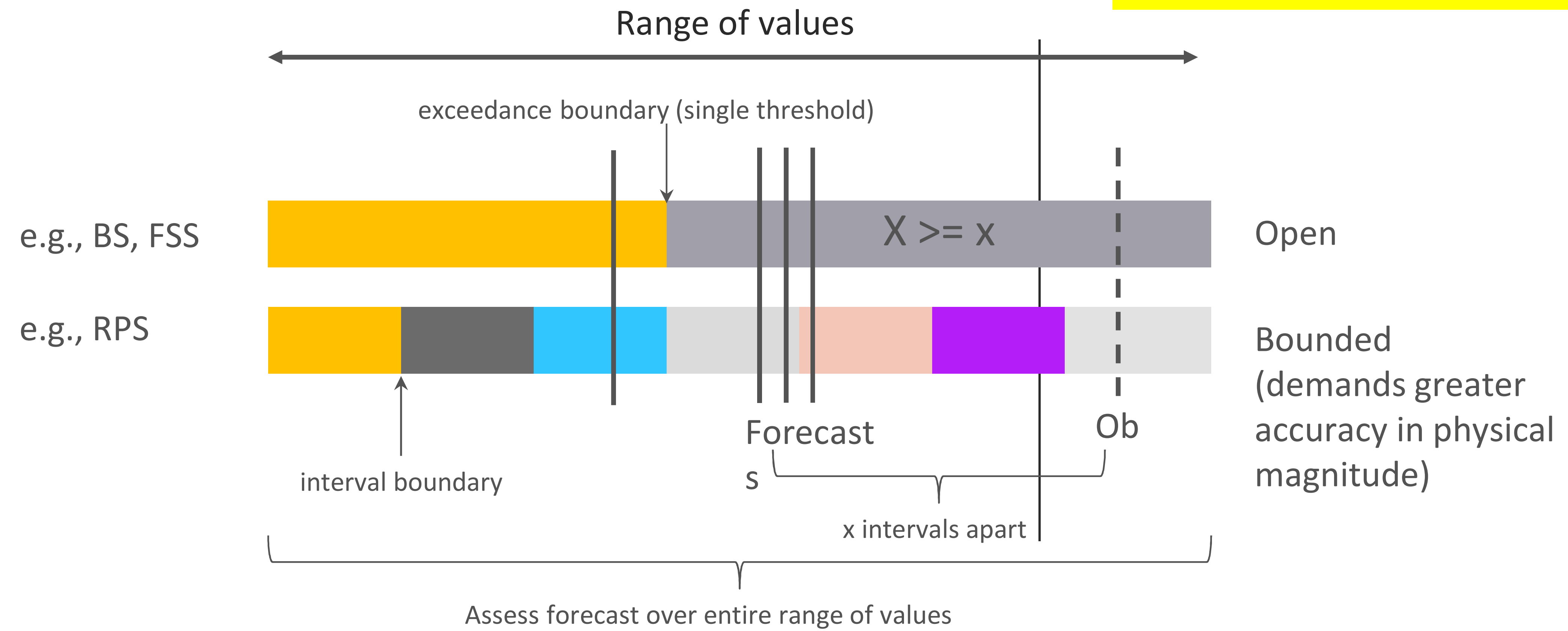
In some instances, **forecast errors <= known instrument errors**. *Should the forecast get the blame? This is a real problem that is hampering our ability to use observations for verification at the very short range, e.g., sondes.*

Open-ended or distribution

Much tougher if forecast is assessed within a bounded range

Why can the RPS be worse and the FSS better?

- They don't measure the same thing; RPS measures a distribution error
- Use different observations; different biases, characteristics, representativeness.

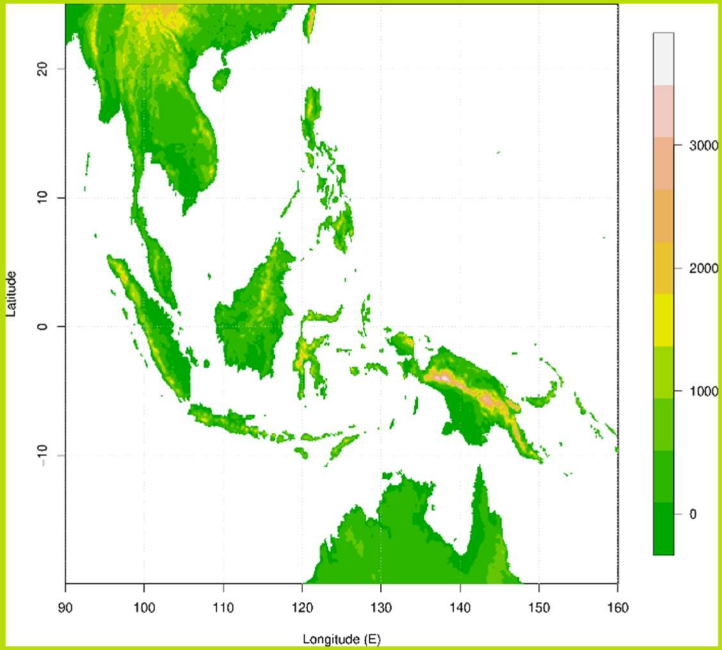


Illustrating concepts over the Maritime Continent

Using Global UM operational forecasts from December 2021 (GA7, N1280), GPM IMERG and LNDSYN stations in the region.*

Use the model orography (over land), to stratify by height.

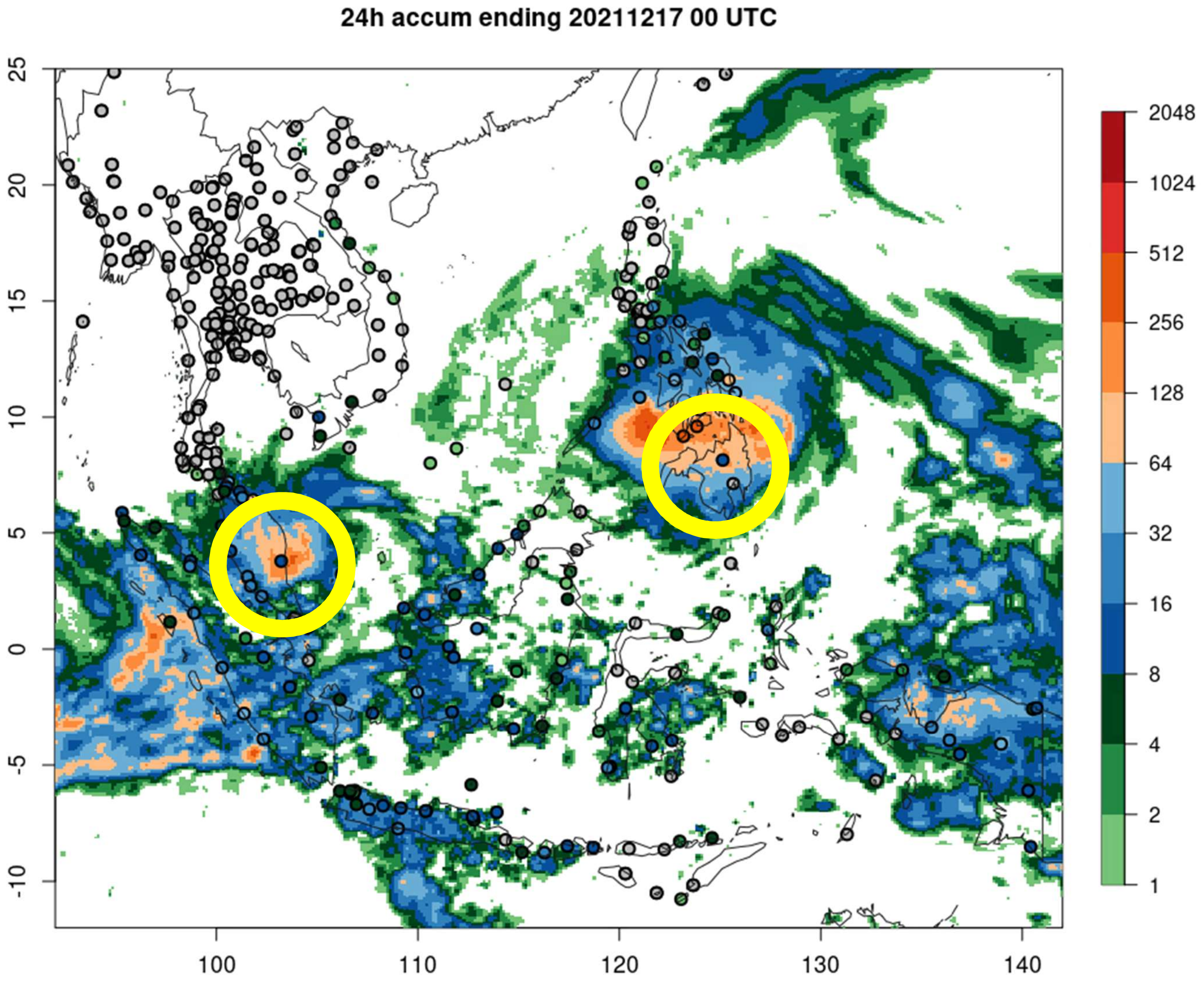
Use daily totals to maximise skill and understanding of bias.



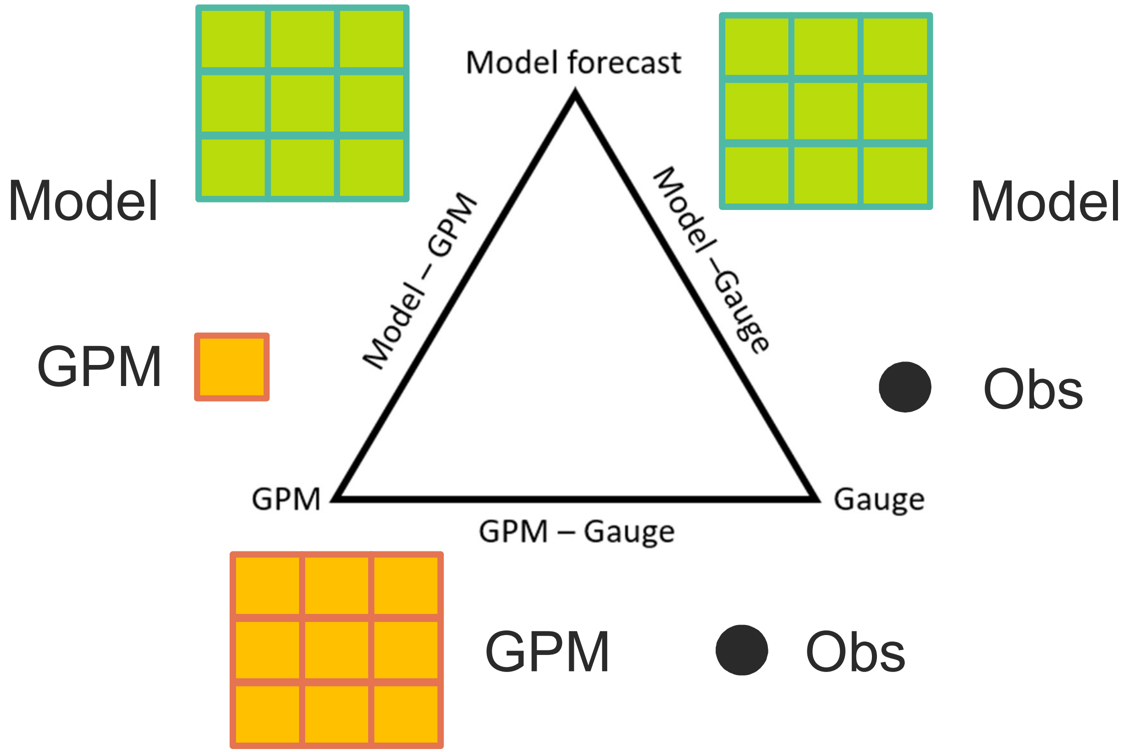
*GM oper is used due to ease of access and broad scale dynamics being unaffected by boundary conditions.

GPM vs gauge

- Example of comparison of daily accumulations.
- Some large mismatches are possible.
- When comparing the same forecast to these two observation types, one should expect the results to be different!



Comparing different precipitation datasets

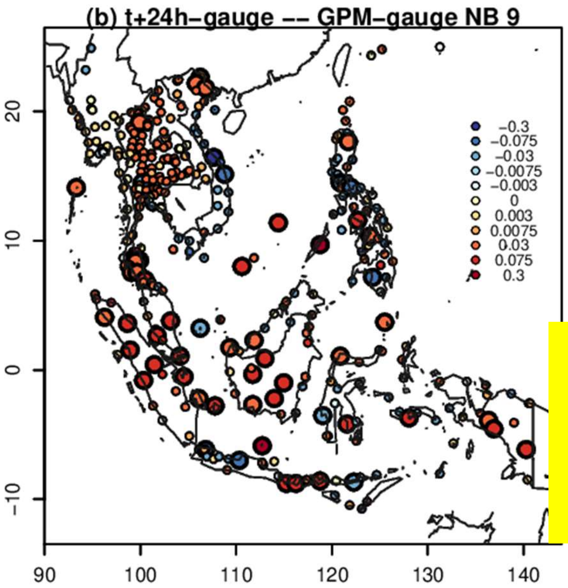


- We are often faced with the dilemma that **observation datasets** at our disposal **do not agree**.
- A common methodology in DA is to **use the model as the benchmark** to compare two observation data sets which measure/estimate the same thing.
- Here the GM forecasts and the neighbourhood-based RPS are used together to compare GPM at gauges to understand what effect **point representativeness** errors have by utilising the single nearest GPM grid point as a pseudo “gauge”.

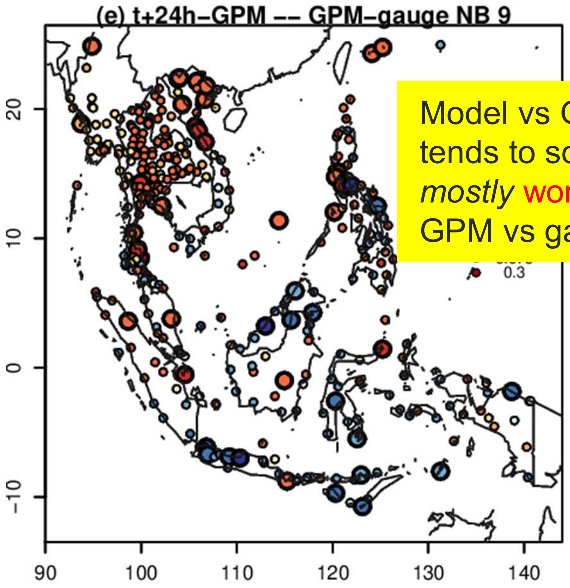
Mittermaier (QJ, in prep.)

RPS differences

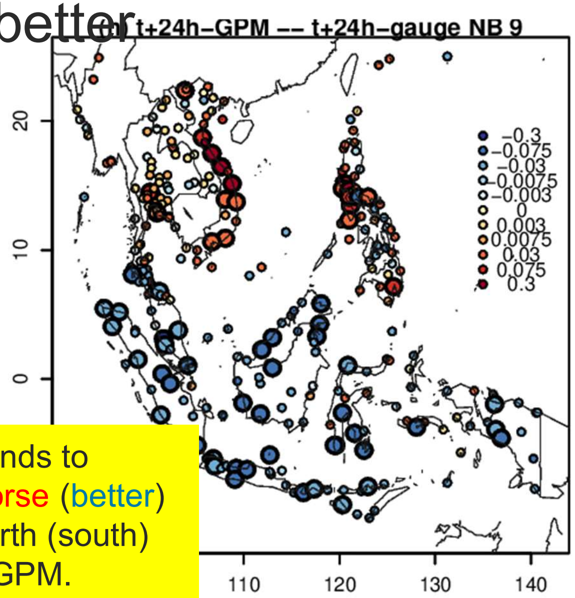
RPS → smaller is better



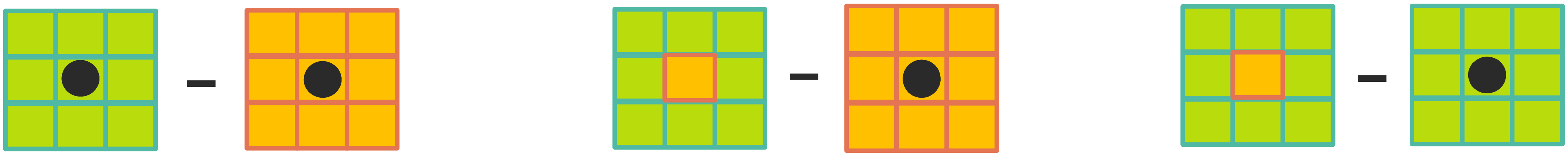
Model scores worse than GPM most of the time



Model vs GPM tends to score *mostly* worse than GPM vs gauge



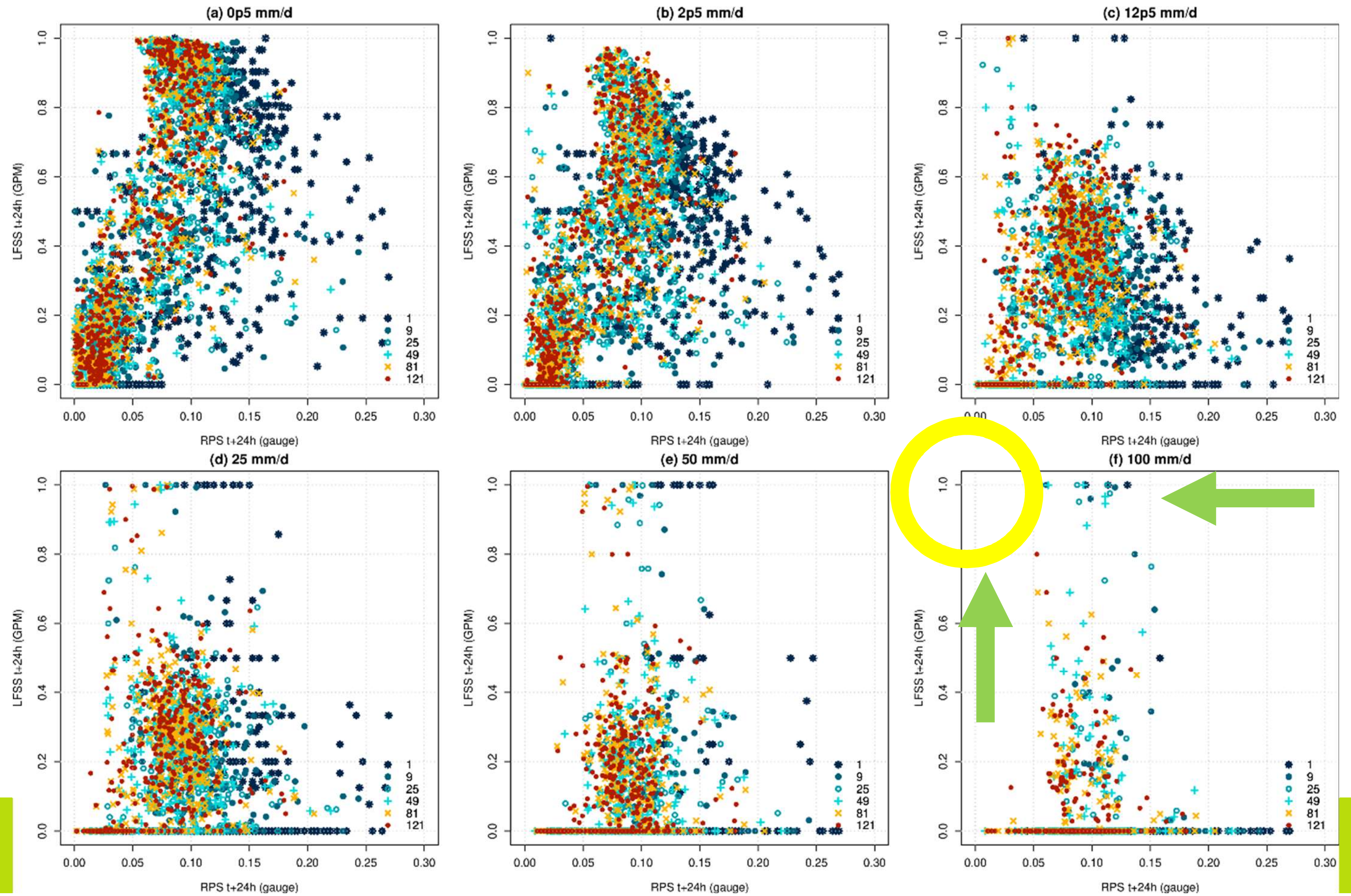
Model tends to score worse (better) in the north (south) against GPM.



Mittermaier (QJ, in prep.)

Larger symbols indicate locations where the differences are significant at the 5% level using a paired t-test for dependent samples and a Markov Chain Monte Carlo method for computing the effective sample size.

RPS vs LFSS comparison of GPM vs gauge



Neighbourhoods act to increase the LFSS (perfect = 1) and decrease the RPS (perfect = 0)

Note RPS values are the same for all panels

Mittermaier (QJ, in prep.)

So,

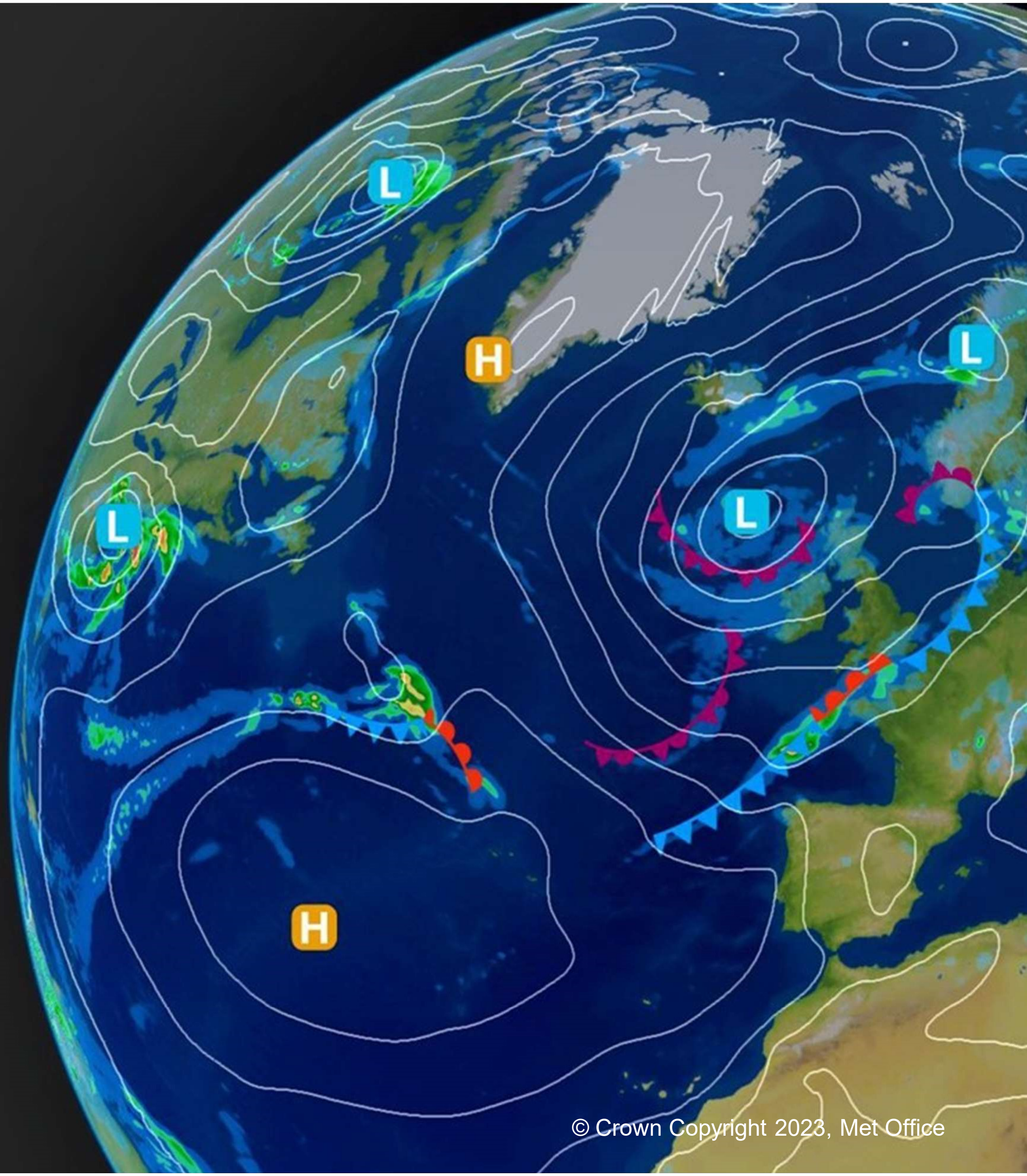
- There are **significant differences** in scores computed **for the same forecast against two different observation datasets**. Most of these differences are down to the representativeness characteristics associated with each of them.
- Radar-vs-gauge RPS can be similar in magnitude to the model scores. → **It is hard to differentiate the model-vs-gauge and the radar-vs-gauge results from each other.** Grid-to-point representativeness dominates the result.
- **Local characteristics** play an important part in determining the size of the error.

Met Office

Using RPS with neighbourhoods

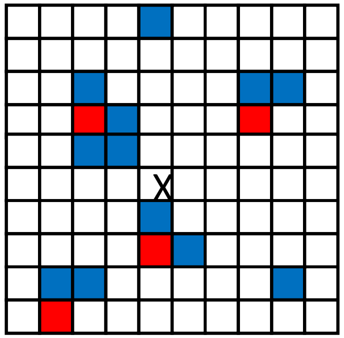
Nigel Roberts

EWGLAM 2023



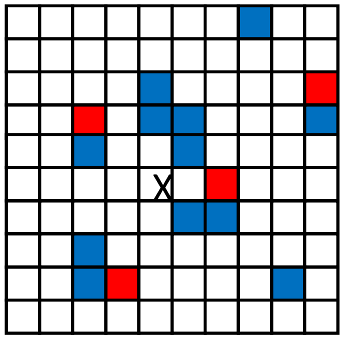
Met Office Application of a neighbourhood to the forecast, but not observation

Observed

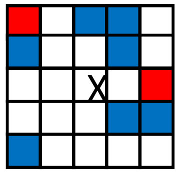


Observation

Forecas
t

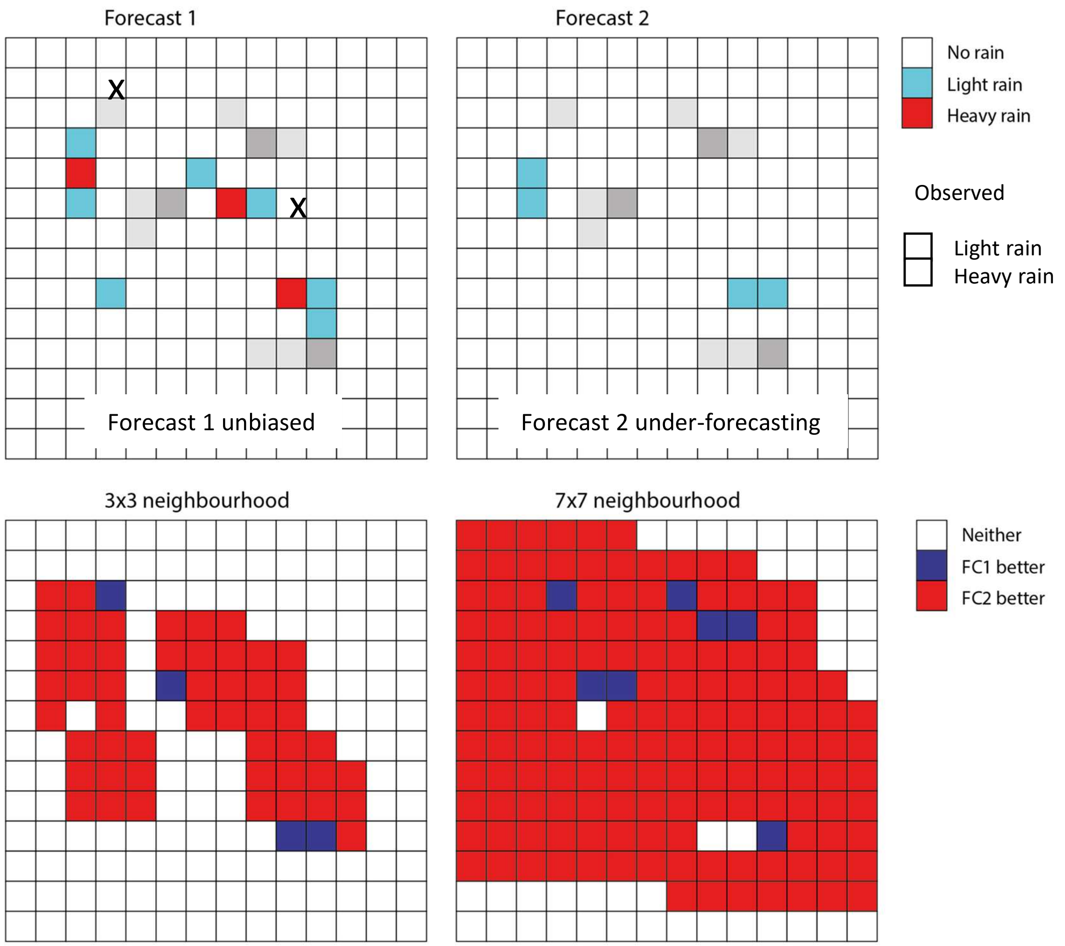


5x5 neighbourhood



Neighbourhood gives increased chance of rain categories in the forecast (at that location) -> score gets worse! Is this general or just at that location?

Idealised situation using the RPS with observed points and forecast neighbourhoods



Forecast 1 unbiased

Forecast 2 under-forecasting (biased)

Similar spatial error

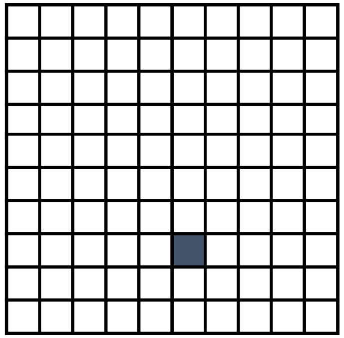
Both completely wrong at the grid scale (double penalty)

Relative RPS score for each grid square to account for all possible rain gauge locations

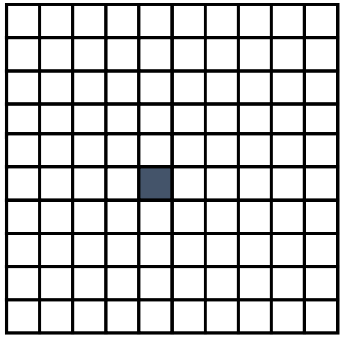
The biased forecast scores better in many more possible gauge locations (more red in 3x3 and 7x7)

The total RPS over all locations is better (lower) for the biased forecast
- More gauges (better sampling) won't help!

Idealised random rain pixel



Observed



Forecas
t

Examine an idealised random placement of 1 rain pixel in 100 pixels (possible gauge locations).

Have the randomly positioned rain pixel in both the observed and forecast grids (no bias).

Here we only have two categories, so RPS = Brier Score.

For any given forecast, the chances of the four possible outcomes, at a gauge location are:

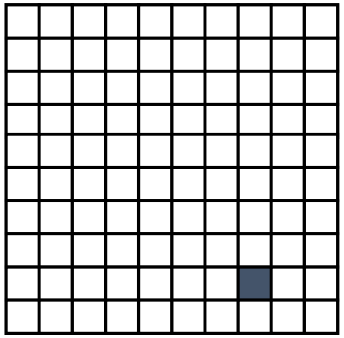
Chance of rain and rain (RPS=0) = 0.01 x 0.01 = 0.0001
Chance of rain and no rain (RPS=1) = 0.01 x 0.99 = 0.0099
Chance of no rain and rain (RPS=1) = 0.99 x 0.01 = 0.0099
Chance of no rain and no rain (RPS=0) = 0.99 x 0.99 = 0.9801

98.02% chance that RPS = 0 (perfect forecast)
1.98% chance that RPS > 0 (not perfect forecast)

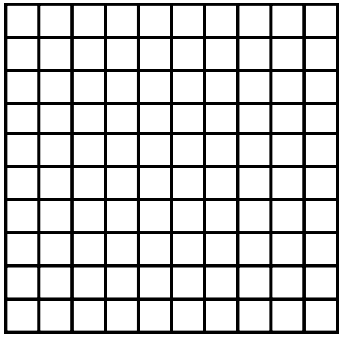
1.0000

Therefore, the expected total RPS over 100 events = **1.98**

RPS for a zero-rain forecast and observed field with 1% coverage



Observed



Forecast

Now consider a forecast system that never forecasts any rain

For any given forecast the chances of the possible outcomes are:

Chance of rain and rain (RPS=0) = 0.00
Chance of rain and no rain (RPS=1) = 0.01 x 1.00 = 0.01
Chance of no rain and rain (RPS=1) = 0.00
Chance of no rain and no rain (RPS=0) = 0.99 x 1.00 = 0.99
99.0% chance that RPS = 0
1.0% chance that RPS > 0

1.00

Expected total RPS over all permutations = **1.00**

Compare with **1.98** for an unbiased forecast (lower value is better skill)

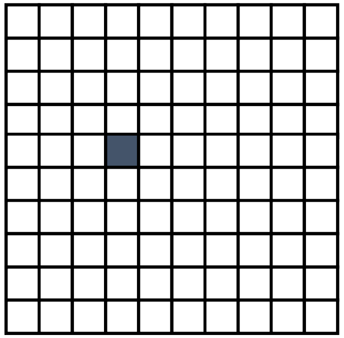
Forecasting nothing means less chance of a wrong forecast and improves skill (double penalty)

Now apply a 3x3 neighbourhood to the forecast rain

This makes no difference to the forecast with zero rain

Met Office

Apply a 3x3 neighbourhood to the forecast with rain

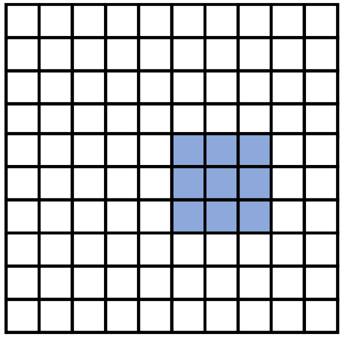


A neighbourhood is not applied to the observed field because we only know the value at the square being sampled (where the rain guage is) and can't construct a neighbourhood

For any given forecast the chances of the four possible outcomes are:

Chance of rain and rain (RPS=64/81)= 0.01 x 0.09 = 0.0009
Chance of rain and no rain (RPS=1) = 0.01 x 0.91 = 0.0091
Chance of no rain and rain (RPS=1/81) = 0.99 x 0.09 = 0.0891
Chance of no rain and no rain (RPS=0) = 0.99 x 0.91 = 0.9009

90.09% chance RPS = 0
9.91% chance RPS > 0



Nine forecast squares can have probability > 0 (probability = 1/9)

Chance of rain and rain RPS=64/81 x 0.0009 = 0.00071
Chance of rain and no rain RPS=1 x 0.0091 = 0.00910
Chance of no rain and rain RPS=1/81 x 0.0891 = 0.00110
Chance of no rain and no rain RPS=0 x 0.9801 = 0.00000

1.0000

Total RPS = **1.00**

0.01091

www.metoffice.gov.uk

Overall findings

Zero rain forecasts	Rain no neighbourhood	3x3 neighbourhood	5x5 neighbourhood
1.0% chance that RPS > 0	**1.98%** chance that RPS > 0	**9.91%** chance that RPS > 0	**25.75%** chance that RPS > 0

The use of a neighbourhood greatly increases the chance of a forecast with rain scoring worse than a no-rain forecast

Zero rain forecasts	Rain no neighbourhood	3x3 neighbourhood	5x5 neighbourhood
Total RPS = **1.0**	Total RPS = **1.98**	Total RPS = **1.091**	Total RPS = **1.02**

The use of a neighbourhood scores worse on average than a no-rain forecast, but less than with no neighbourhood

Overall – the use of a neighbourhood means that favouring under-forecasting is more likely

Met Office

Final comments

Idealised scenarios suggest the use of a neighbourhood can favour under-forecasting (using RPS / Brier Score)

Other idealised configurations show the same (if coverage < 50%)

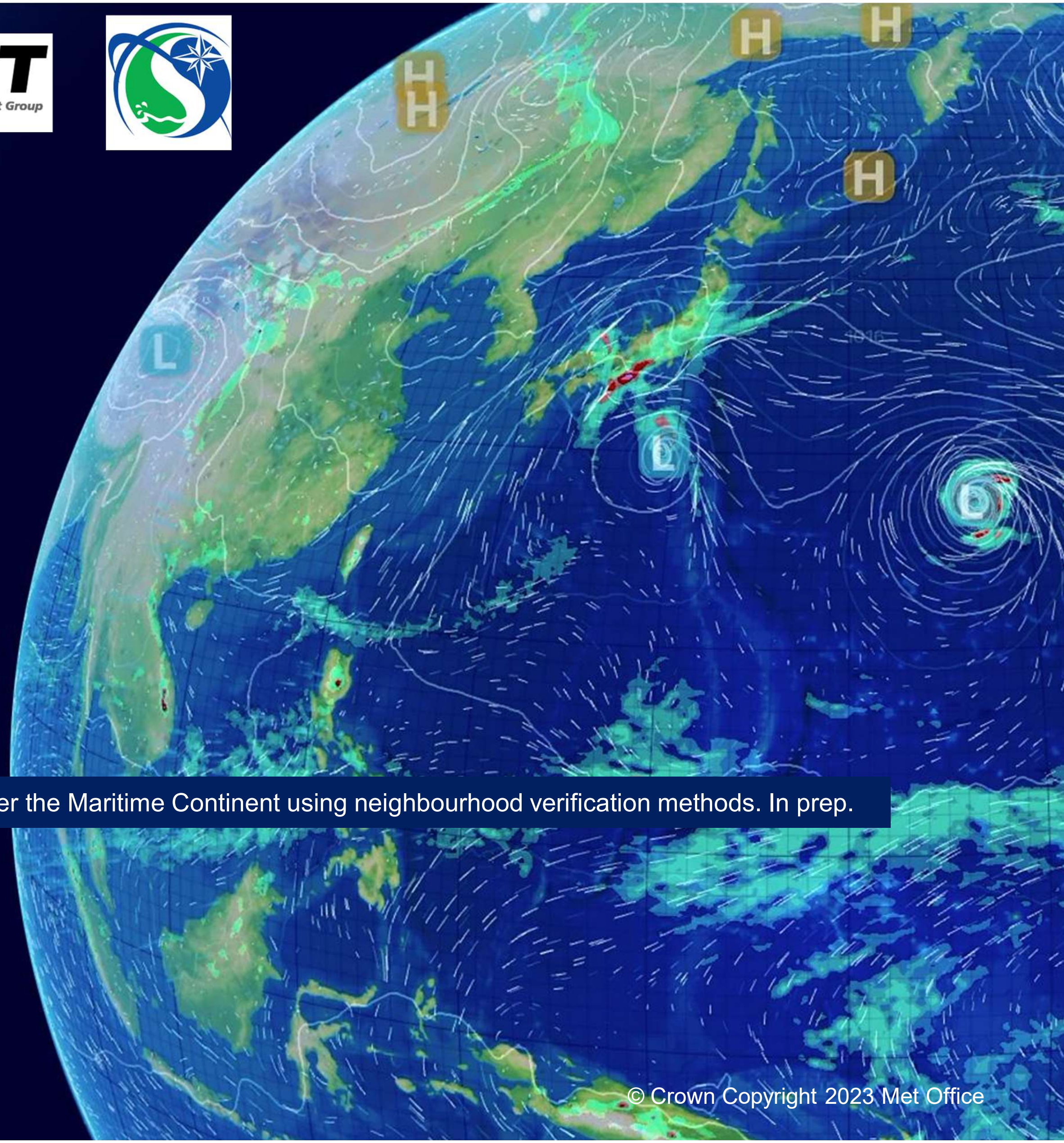
Worse if rain coverage is small or the neighbourhood does not span the spatial error

An ensemble will have the same effect as a neighbourhood (because it increases forecast coverage)

These are idealised studies and further investigation using more realistic or real cases is needed to confirm whether there is an issue

Met Office

Thanks for listening!
Questions?



Mittermaier, M.P., 2023: Comparing point- and gridded observation types over the Maritime Continent using neighbourhood verification methods. In prep.

Newton Fund

www.metoffice.gov.uk