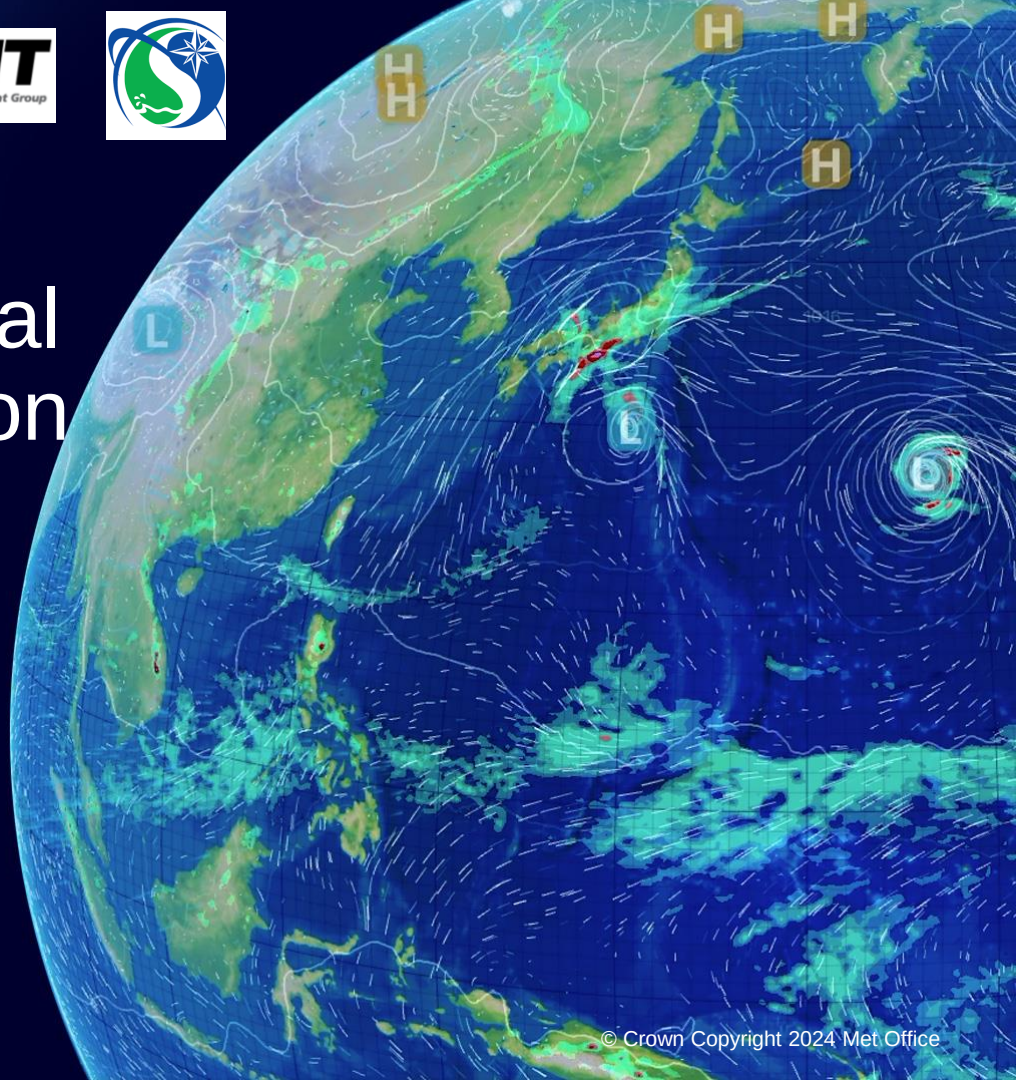
An exploration of spatial and temporal correlation on sample size

Marion Mittermaier and Eric Gilleland

9th international verification methods workshop

Cape Town, May 2024

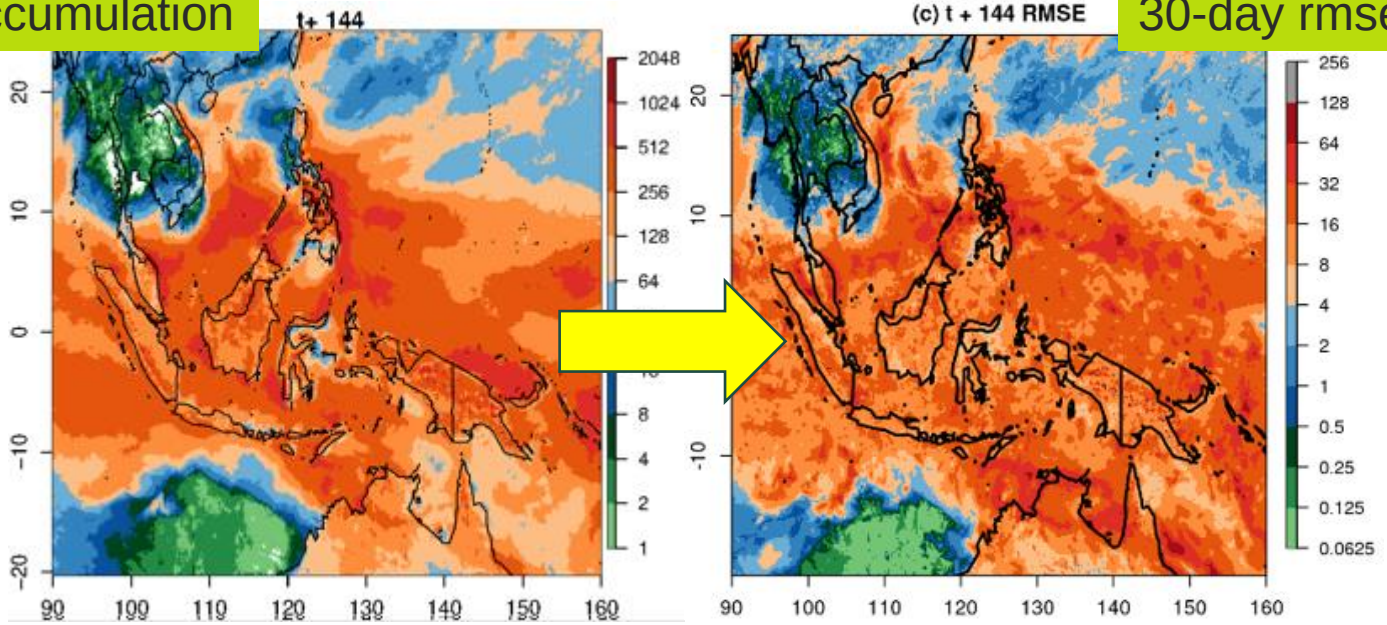
Mittermaier and Gilleland, 2024, in prep.



www.metoffice.gov.uk

The fields drive the correlations

30-day accumulation



30-day rmse

Strong gridpoint-to-gridpoint correlation in the precipitation fields which is translated to the rmse

Note: each grid point is an aggregate of *something* over time

Points to note:

The spatial patterns are very apparent when looking at a 2-D field.

2-D fields of scores *only* collapse the time dimension. This means:

→ time series behaviour / temporal evolution and similarity (correlation) between *adjacent time* points is still hidden with this view when using a map.

→ correlation will be worse the closer the time points are together, e.g. adjacent hourly fields will generally be more correlated than 6h apart or daily

→ time series / temporal evolution of adjacent grid points *in space* is likely to be very similar in many instances

Adjacent grid points are not independent pieces of information, either at a single time point t or over a time window (e.g. a month) → contemporaneous correlation

However, …

We often “consume” metrics (like the rmse) as **a single number representing the whole domain or some region**.

When we compute a single number, **we collapse *both* time and space dimensions**.

In our example, the domain has 239040 grid points (individual scores).

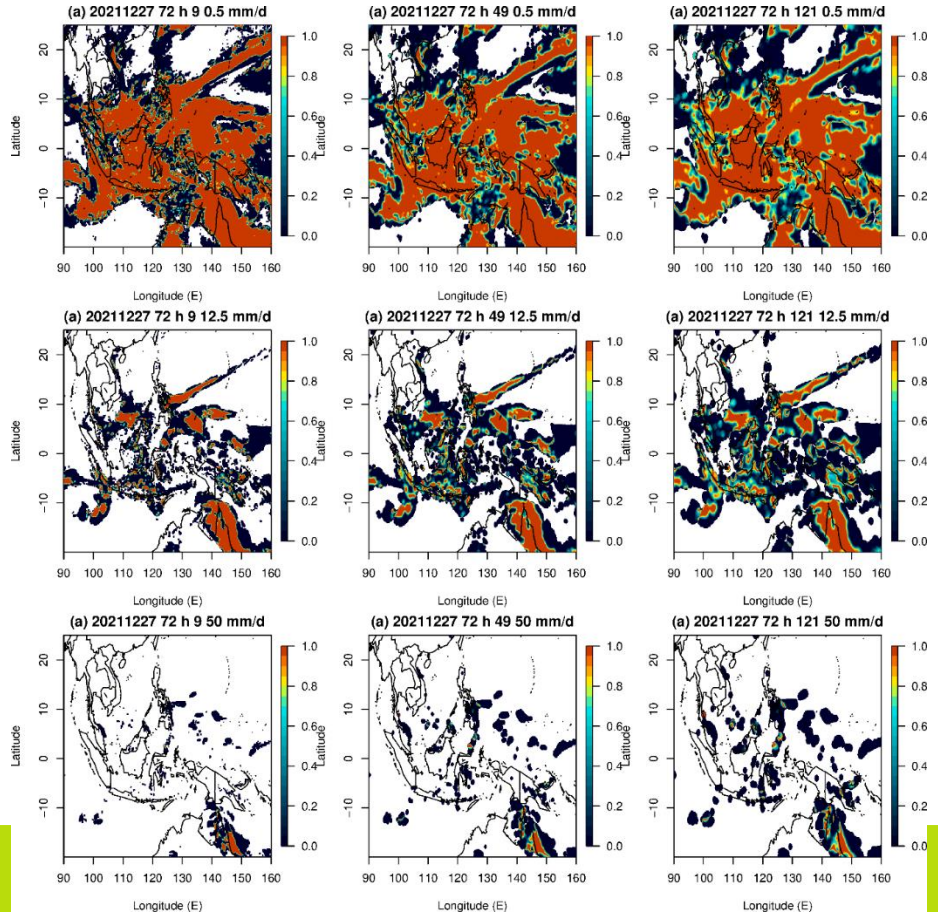
Each grid point *rmse* is in turn computed from 30 (daily) time points.

Given we can strongly suspect that adjacent points (in time and space) are correlated, what is the sample size for computing CIs?

What is the effective sample size for temporal and/or spatial aggregation?

LFSS maps for Dec 2021

Maps as a function of threshold & n'hood size.



Large amount of similarity in the scores, which increases with increasing neighbourhood size (smoothing)

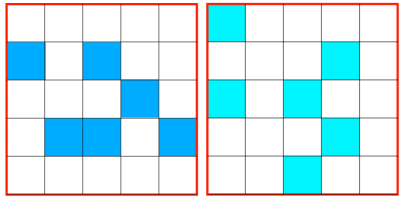
Mittermaier (QJ, 2024)

Neighbourhood based metrics

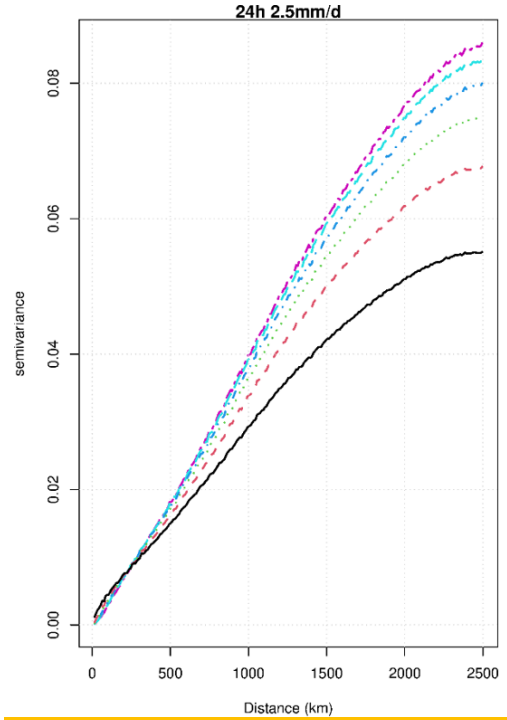
Most (not all) collapse information in a neighbourhood down to a single number.

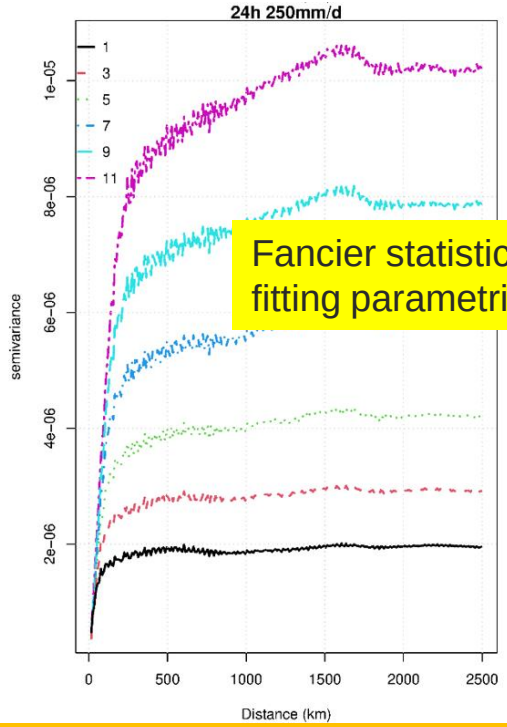
For the FSS, a **neighbourhood fraction is calculated at each grid point**, ***increasing*** the dependency (and thus the correlation) between adjacent grid points (you're making the grid points more similar to each other).

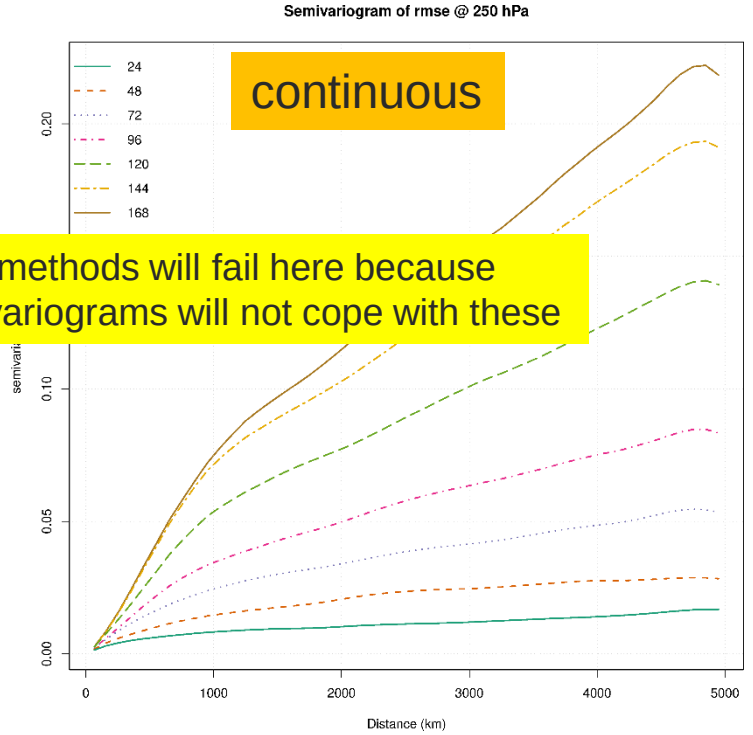
This dependency increases with increasing neighbourhood size.



Semi-variogram







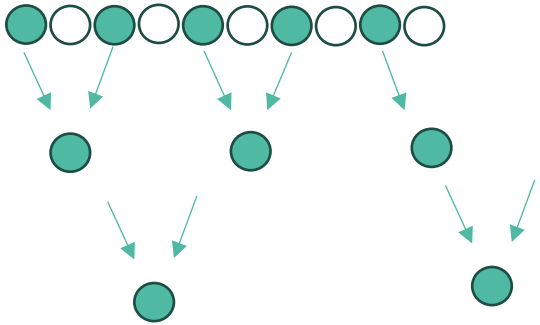
continuous

Fancier statistical methods will fail here because fitting parametric variograms will not cope with these

FSS categorical as a function of nb size @ t+24h

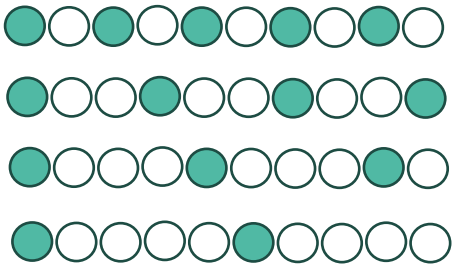
Empirical sampling techniques

Strict



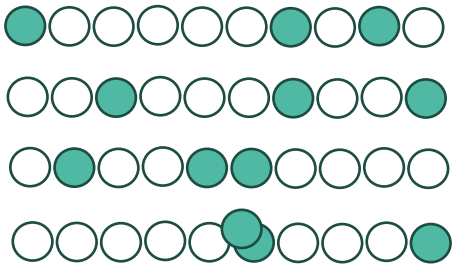
Extract every other grid point from the full grid
Continue to sample every other grid point from each sub-sample.
Each reduced sample is a subset.

Systematic



Always start from the full grid but first take every other grid point, then every 3rd, 4th, 5th etc.
This means there is some overlap in samples.

Random

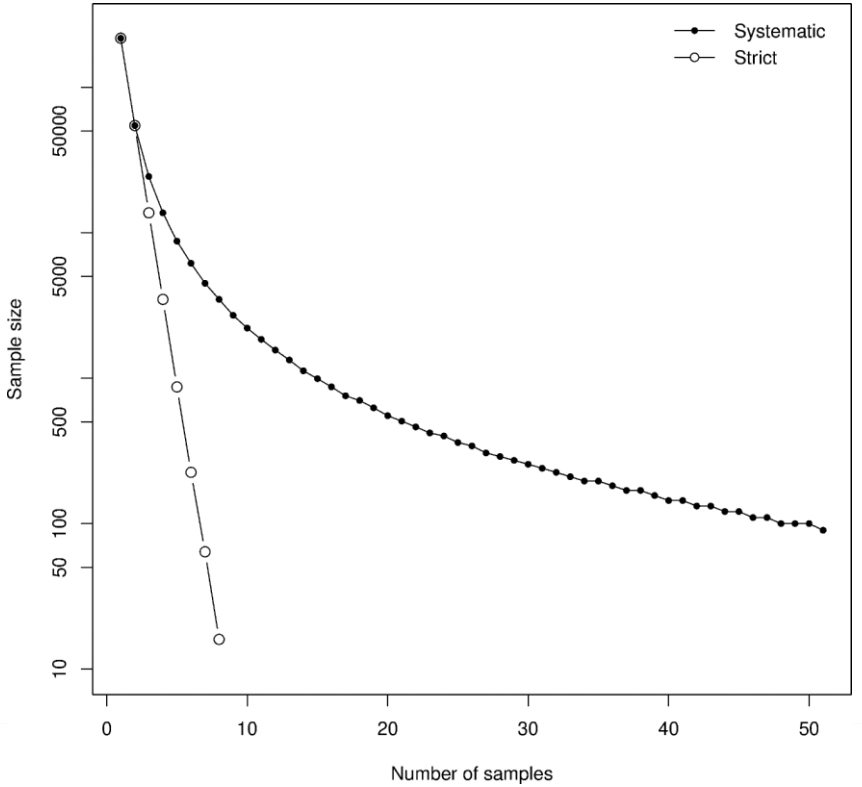


Draw random samples from the grid WITH replacement.
This means there could be duplicates.
Repeat M times.

Empirical sampling: what does this look like?

- Given the 498 x 480 grid this gives the following progressing of sample sizes

- For the random sampling the same progression as the systematic sample size was used.

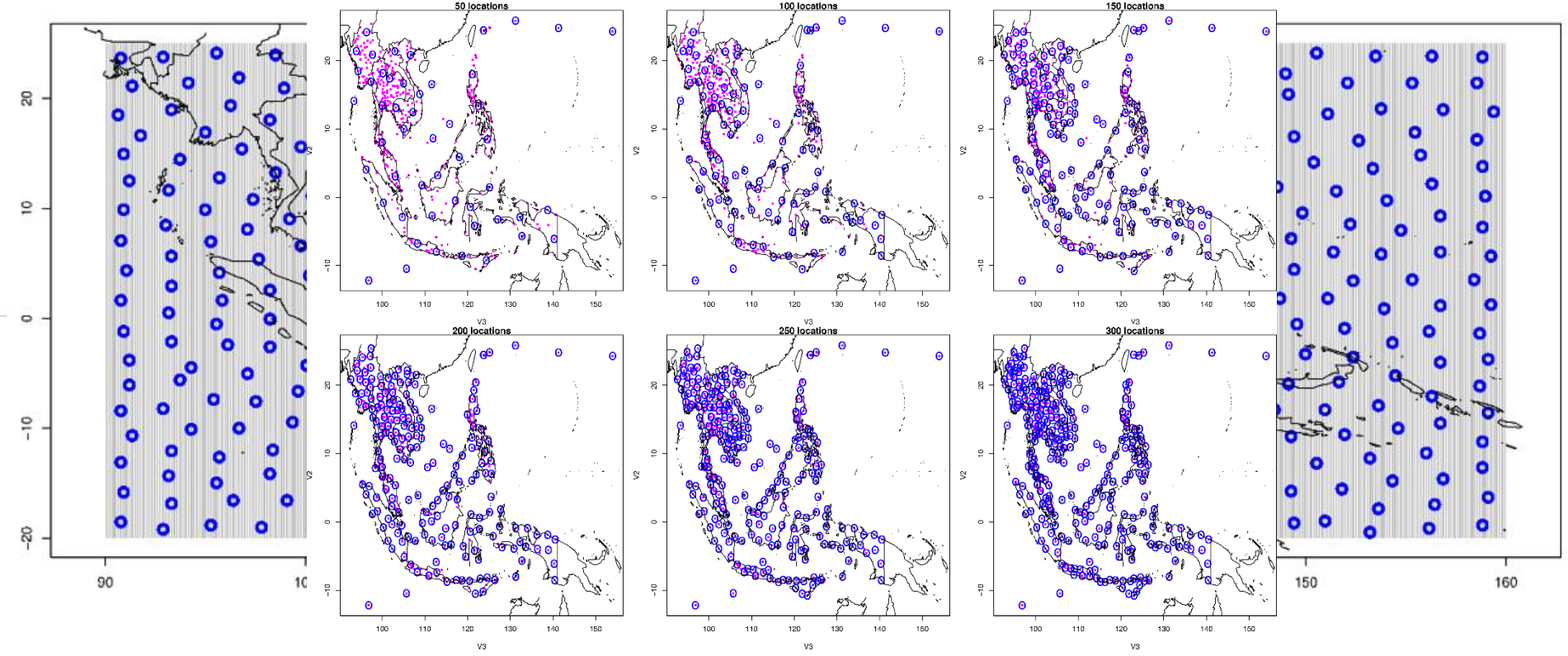


Statistical sampling techniques

- **Coverage or network design** attempts to answer the question whether the observations of a given variable in an area are sufficient for describing or characterising the performance of the forecast in that area (e.g. Cressie et al., 1990, Angulo et al., 2005). It can be used **to identify gaps** (in observation networks), or **identifying redundancy**, often referred to as **network thinning**, to **ensure more optimal or uniform spatial sampling** whilst reducing the density.
- The **primary objective of using such an algorithm is to increase distance between the grid points and decrease the similarity** and spatial correlation whilst preserving the ability of the subset of points to provide an uncompromised representation of performance.
- Used the `cover.design` function in R *fields* (see also Gilleland and Fowler, 2006) to define such designs.

Met Office BMKG MiGHT Malaysian Industry-Government Group for High Technology Department for Science, Innovation & Technology

Statistical sampling: what does this look like?



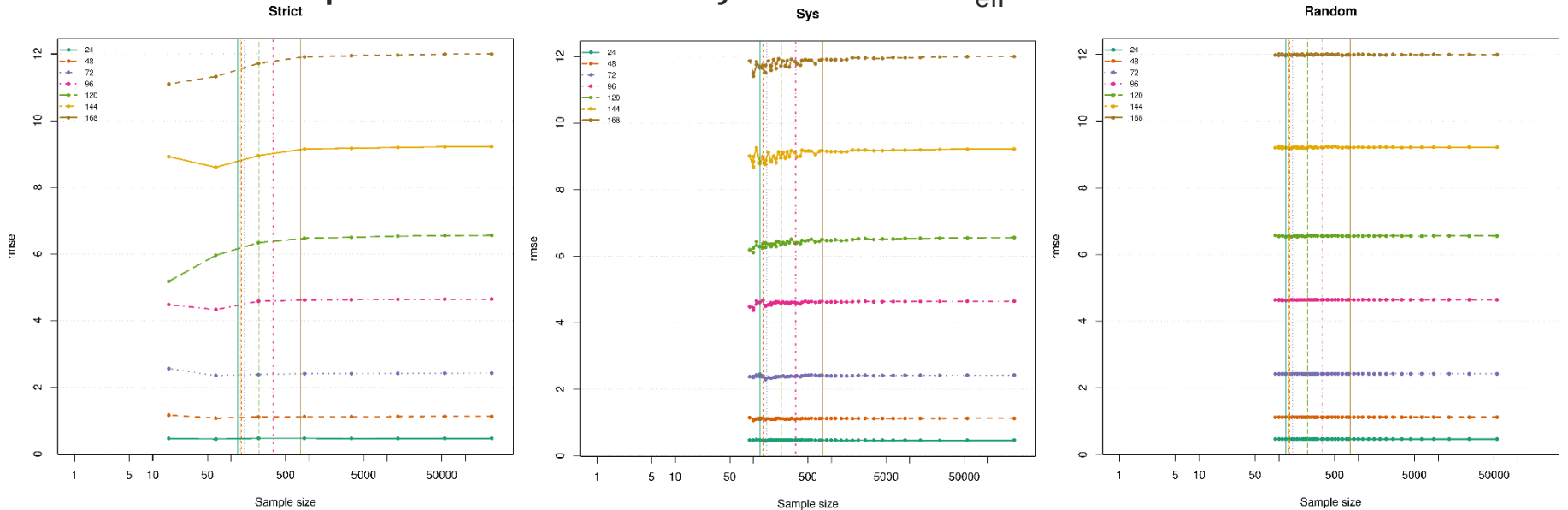
One-sample case: 250 hPa Temperature

Loss function: squared error

Empirical sampling options

Aggregate rmse based on sample size.

Vertical lines represented theoretically calculated N_{eff} at each lead time.

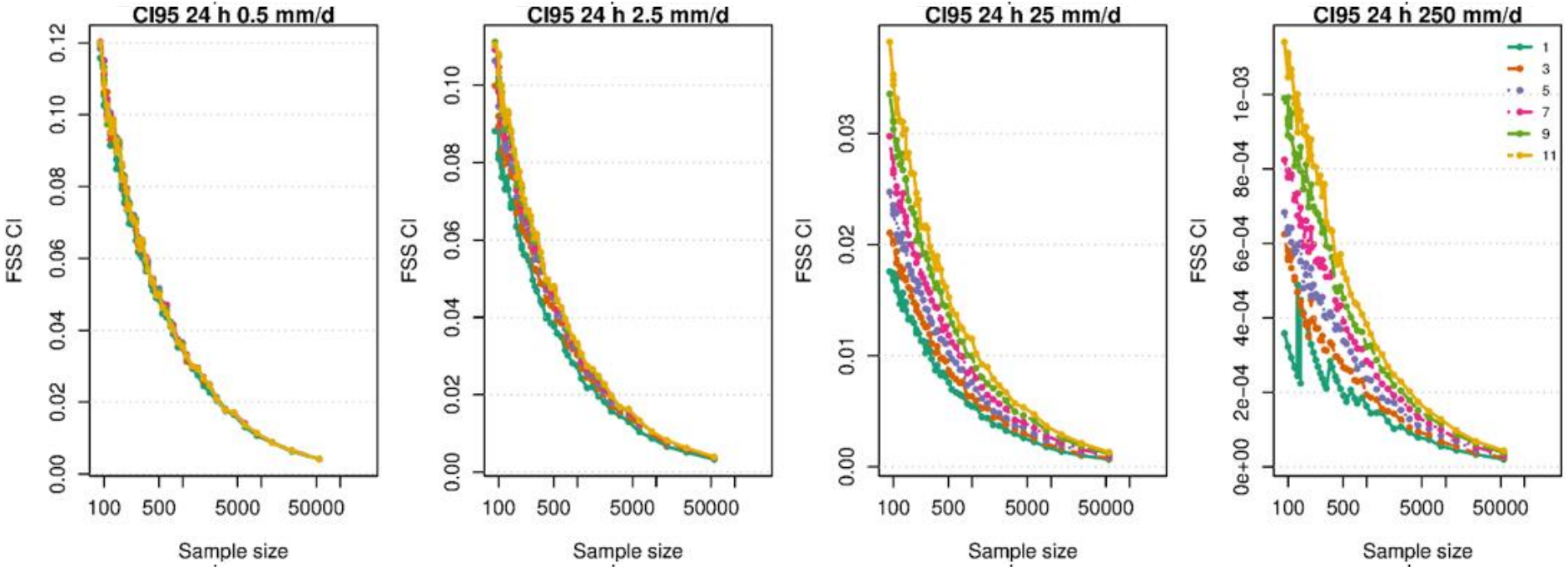


Two-sample case: LFSS forecast vs persistence

Impact of sample size on spatial aggregate

$N_{eff} < 500$ (2%)

Vertical lines represented theoretically calculated N_{eff} at each lead time.



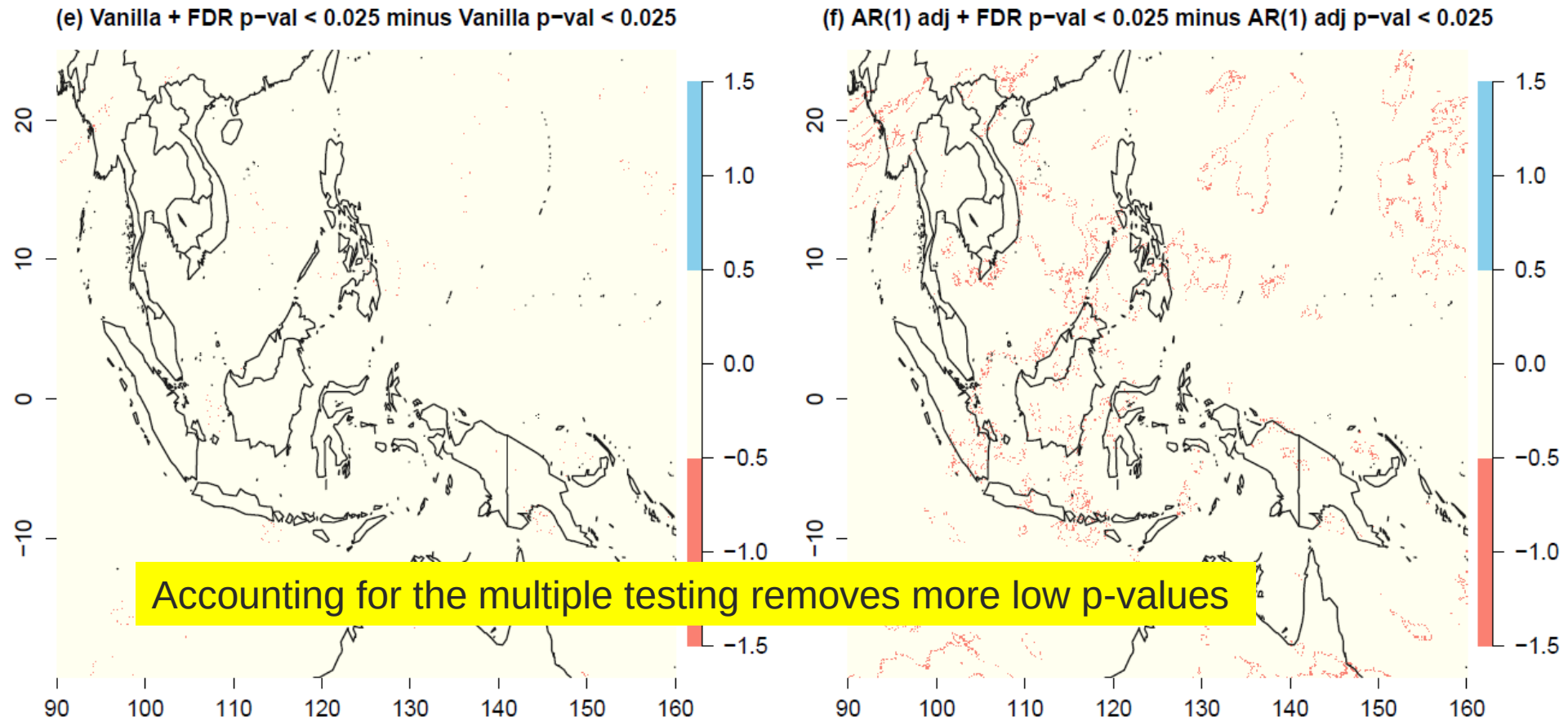
Statistical inference

- **One sample case (of the difference)**→ e.g., tracking whether the forecast is significantly different from the observations. Here a two-sided test is generally used. H_0: F – O = 0; H_1: F – O ≠ 0
- **Two-sample (paired difference) comparison** → is model A better than model B? Both are measured against the same observation, run over the same time period (i.e. they are paired/dependent). Here a one-sided test may be best because you want to be clear whether A is better than B (not just “different”). H_0: A – B = 0; H_1: A – B < 0 or > 0 depending on metric
- Best established by computing CIs such that if 0 ∈ CI, the difference is *not* significant.

Two theoretical examples

- **One-sample:** A t-test with AR(1) compared to a vanilla t-test to look at the effect of autocorrelation of 30-day aggregate rmse
- **Two-sample:** the spatial prediction comparison test (`spct` in *SpatialVx,* Gilleland 2013) applied to the difference in the daily forecast and persistence LFSS aggregates over the grid using the `cover.design` derived sample sizes.

Effect of auto-correlation on 30-day aggregate rmse values on the grid

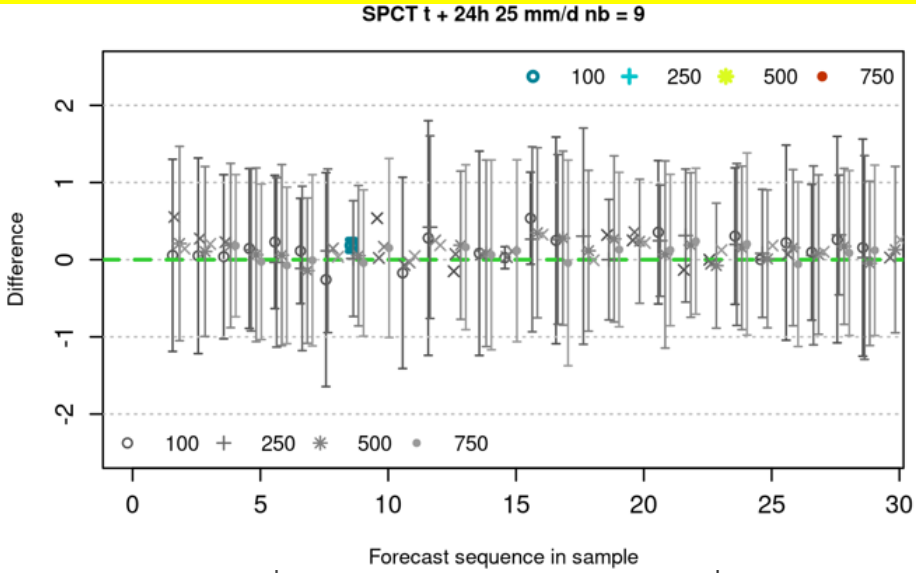


Accounting for the multiple testing removes more low p-values

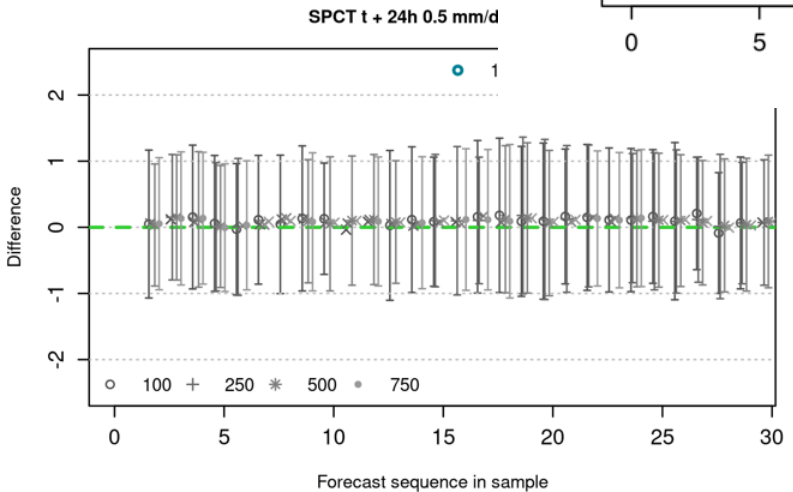
Effect of spatial correlation at each time slice on the FSS aggregate over the domain

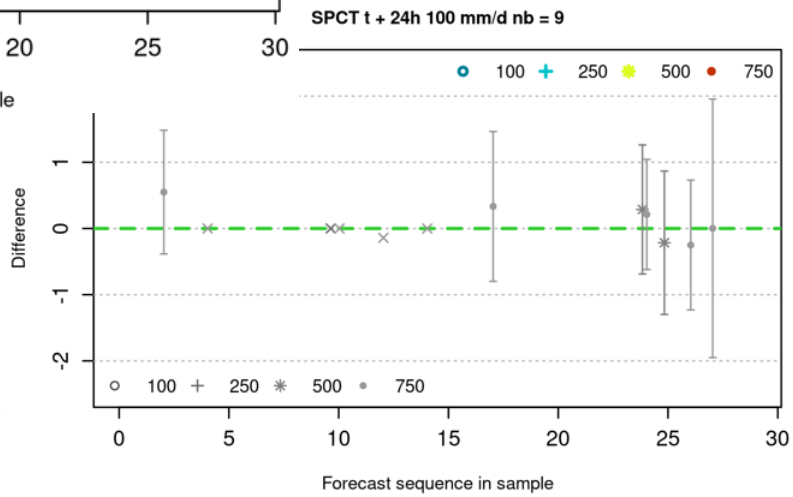
Checking each day's domain-wide score to see whether CI contains 0

Using the `cover.design` sample sizes



One day (spatial aggregate) in the time series which was significant at one threshold!





Why is all this important?

- **Grids are getting denser and ensemble members are increasing → the cost of verification is increasing**
- The complexity (and cost) of the methods required to provide better verification guidance is also increasing
- How do we constrain this?
- One option, demonstrated here is thinning the data.

This work shows that:

- making this choice may not only **reduce computational cost** *but* also **improve robustness**.
- **for the examples the aggregate scores change only slowly as the grid is thinned** → domain-wide can be represented by a subset of grid points.
- **statistical significance should be considered rare (unless or *even if* you have done everything possible to reduce the impact of temporal and spatial correlations by whatever means)** → we *should* be suspicious of results which show excessive statistical significance

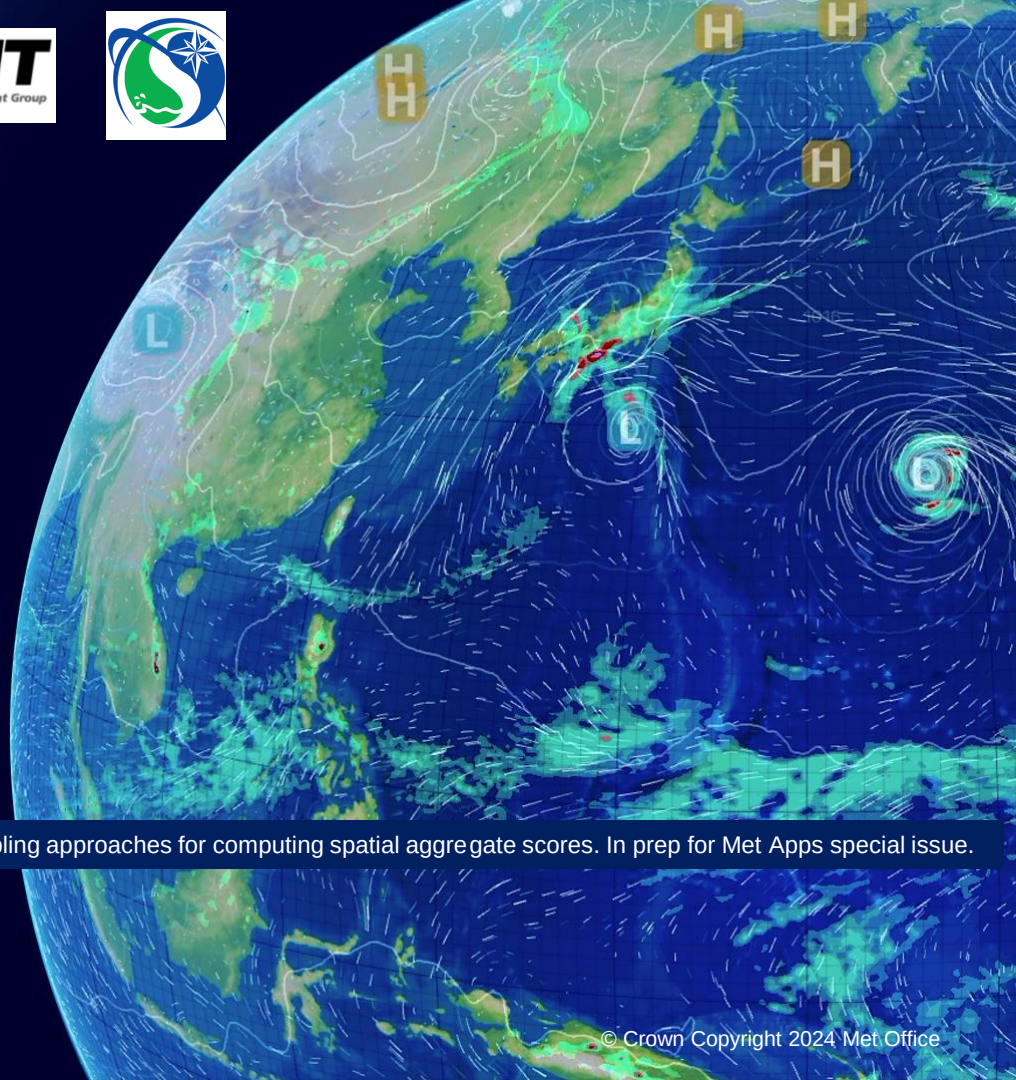
Mittermaier and Gilleland, in prep. 2024)

Final thoughts:

- **Sample sizes will vary for different variables and potentially lead times as well as thresholds (and neighbourhood sizes)** → further statistical analysis will be needed to establish some baseline sample sizes to preclude the need for computing these every time (cost to do this would be prohibitive operationally)
- **A pragmatic approach, compared to a more fancy statistical (and computationally expensive approach), can be just as effective** at mitigating the worst of the correlation effects. Fancy statistics may not be necessary all of the time.

Mittermaier and Gilleland, in prep. 2024)

Thanks for listening!
Questions?



Mittermaier, M.P. and E. Gilleland, 2024: Comparing empirical and statistical sampling approaches for computing spatial aggregate scores. In prep for Met Apps special issue.