EGU23-12242, updated on 25 Apr 2024
https://doi.org/10.5194/egusphere-egu23-12242
EGU General Assembly 2023

Exploring empirical and statistical approaches for determining an appropriate sample size for aggregate scores

Marion Mittermaier[1] and Eric Gilleland[2]
[1]Met Office, Exeter, United Kingdom (marion.mittermaier@metoffice.gov.uk)
[2]National Centers for Atmospheric Research, Boulder, CO, United States (ericg@ucar.edu)

Spatial sampling remains a conundrum for verification. The observations that are required are rarely on a grid, nor are they homogenously spaced. They are often located where there are people, easy access and do not sample the variable in a representative way. In an aggregate sense, scores derived from such observation locations, will give areas with greater observation density more weight in the aggregate if the variations in network density are not accounted for. Furthermore the performance in some parts of the domain may not be represented at all if there are no observations there. Gridded analyses on the other hand often provide complete coverage, and offer great ease of use, but adjacent grid boxes are not independent. Given this relative wealth of coverage and uniform sampling, we tend to use all available grid points for computing aggregate scores for an area or region, despite knowing that this is likely to produce too-narrow confidence intervals and inflate any statistical significance that may be present.

In this presentation a variety of approaches, both empirical and statistical, are explored to establish what we ought to include when computing aggregate scores. Three different empirical sampling approaches are compared to selections from statistical coverage or network design algorithms. The empirical options include what is termed “strict” sub-sampling, whereby a sample is taken from the full grid and the reduction in sample size is explored by systematically continually taking a sub-sample from the sub-sample. The second is a systematic reduction in sample size from the original grid whereby each sample is drawn from the original grid, taken every other grid point, then every 3rd grid point, every 4th etc. The third is a mean computed from N random draws of reducing sample size. These empirical options do not respect land or sea locations. They are purely intended at looking at the behaviour and stability of the sample score. The coverage design algorithms provide a methodology for deriving homogeneous samples for irregularly spaced surface networks over land, and regularly spaced sampling of grids over the ocean, to achieve an optimal blend of sampling for regions that cover both land and sea. These sample sizes and sample scores are compared to a statistically computed effective sample size.

Some interesting and surprising results emerge. One of which is that as little as 1% of the total number of grid points may be sufficient for measuring the performance of the forecast on a grid, though the proportion of the total will always be dependent on (to varying degrees) the variable, the threshold or event of interest, the metric or score, and the characteristics of the geographical

region of interest.