Benchmarking homogenisation algorithm performance against test cases

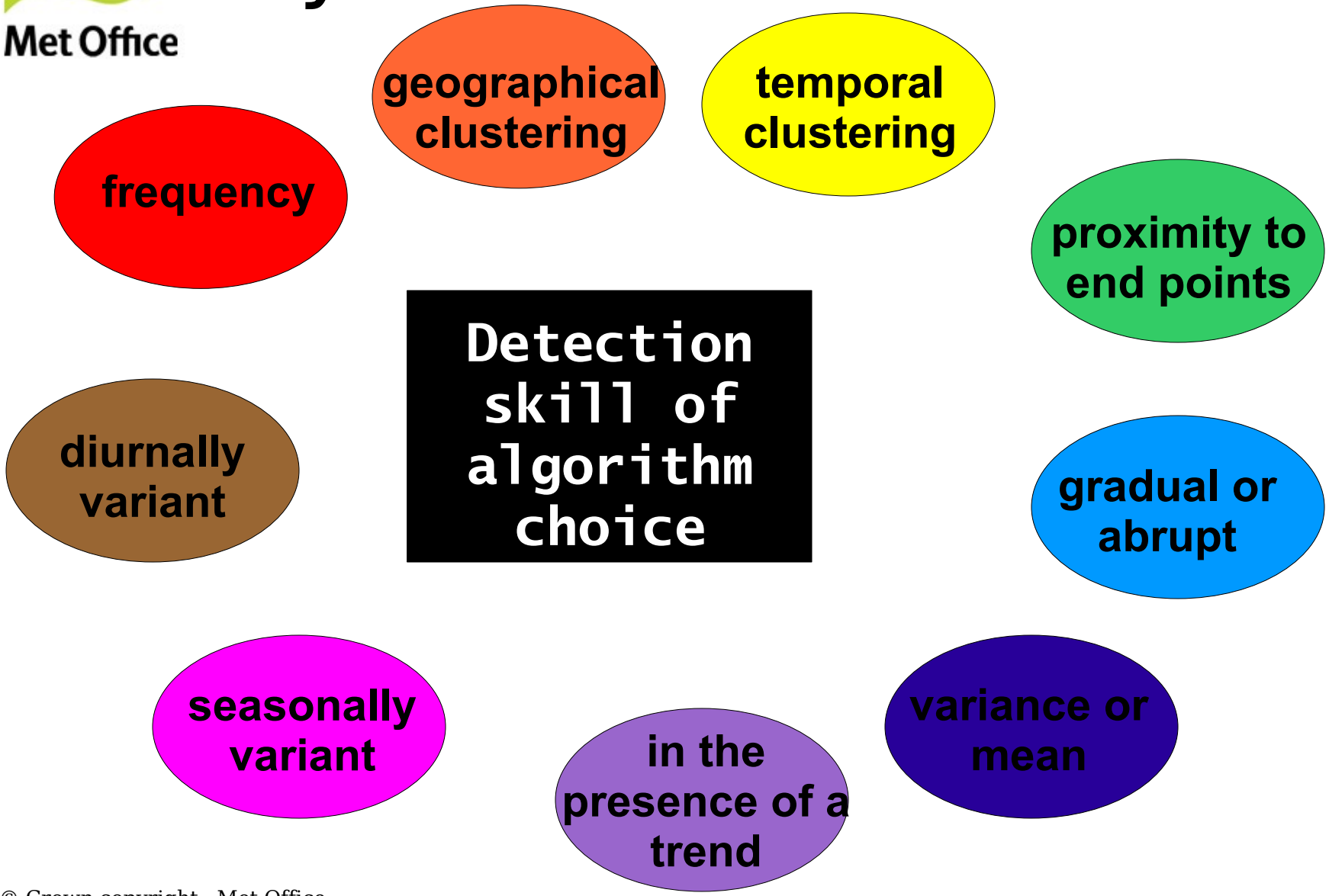
Kate Willett, Matt Menner, Peter Thorne, Stefan Brönnimann, Ian Jolliffe, Lucie Vincent and Xiaolan L. Wang

Contents

- Why benchmark?
- Importance of spatial and temporal sampling of the test data
- Source data for creation of homogeneous test data
- Optimum exploration of all discontinuity eventualities
- Avoiding over-tuning of algorithms to discontinuity test cases

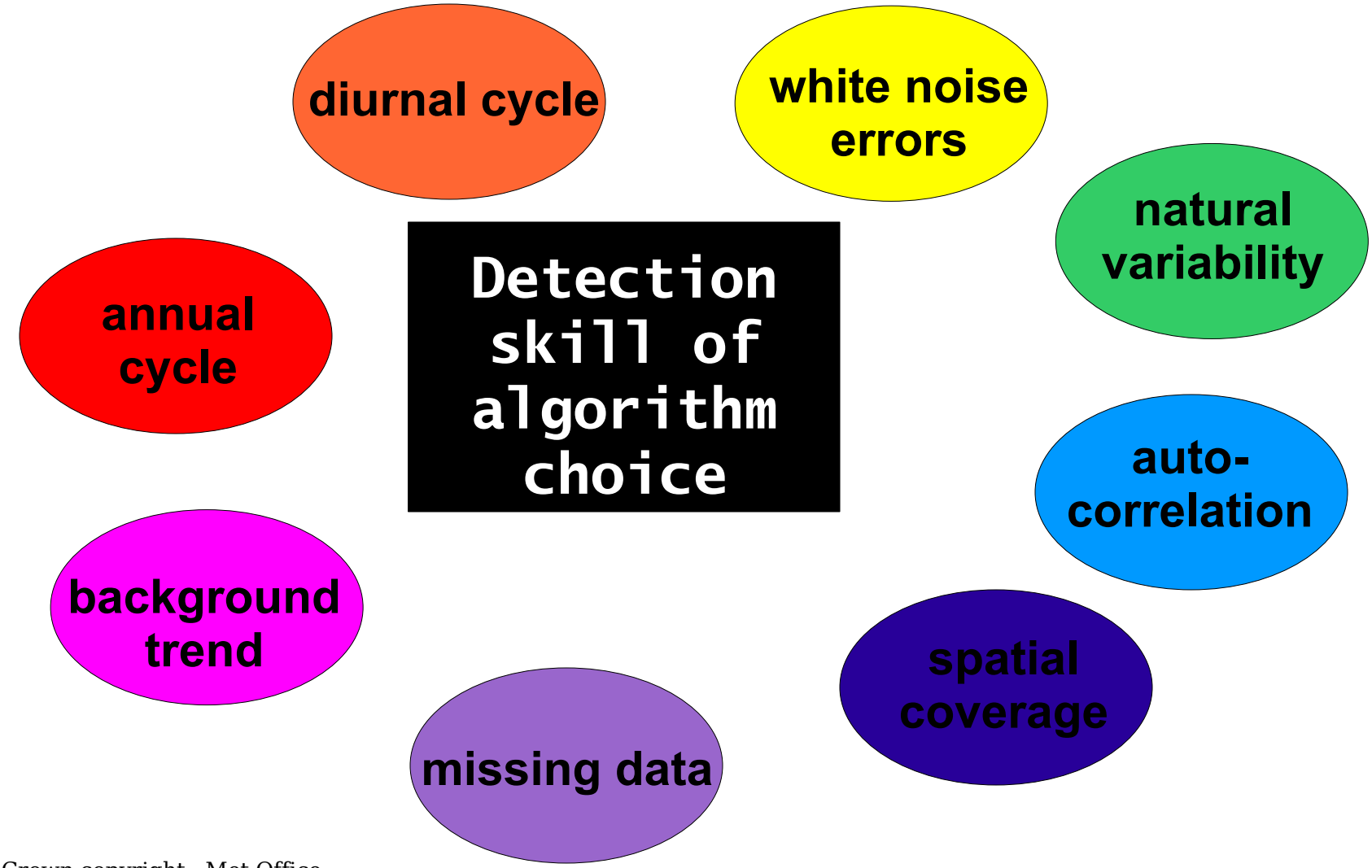
Why benchmark?

Why benchmark?



Importance of spatial and temporal sampling of the test data

Reconstructing full space and time sampling of the observational network

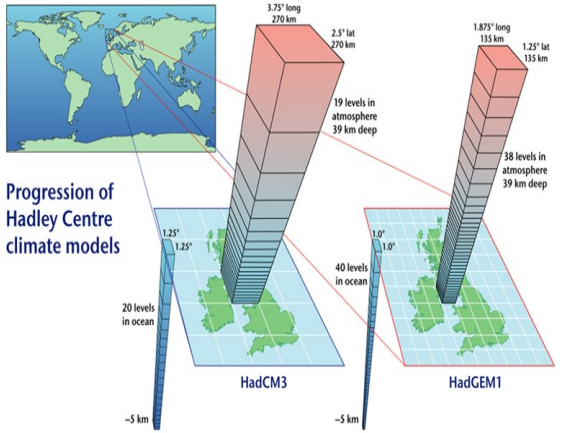


Source data for creation of homogeneous test data

Source data for creation of homogeneous test data

Regionally homogenised observational data

4th generation REANALYSES



High Resolution Global Climate Models – historically forced, natural only, all forcings etc.

Nudge gridded fields with real station characteristics: climatology, variance, autocorrelation, white noise to create test data for the globe

Optimum exploration of all discontinuity eventualities

Exploring all eventualities with optimum discontinuity test case design

1) Review of metadata and state of understanding

2) Create test cases with physically plausible effects of changes in station location, instrument, recording practice, surrounding environment

3) Benchmark i.e.
A set of 10 different discontinuity test cases and the homogeneous test data

Run data product algorithms on the test cases

Benchmark performance by proximity to the 'truth' (homogeneous test data) and record percentage of successful, missed and incorrect detection and adjustments

Avoiding over tuning of algorithms to discontinuity test cases

Avoiding prior knowledge of test cases and resulting over-tuning of algorithms to discontinuity test cases

Creation of test data and test cases and benchmarking should be done independently from data product producers

Benchmarking should be published in journals – peer reviewed – but withholding the solutions

Low likelihood of over-tuning as long as a wide range of discontinuity types are included that fully represent real world eventualities

Recommendations summary

Recommendation summary

- Global test data with real world characteristics

- GCM/Reanalyses data should be used as source base with real spatial, temporal and climatological characteristics applied

- Review of inhomogeneity causes and characteristics finalised via a session at an international conference to ensure plausibility of discontinuity test cases

- Suite of 10 test cases, physically based on real world inhomogeneities and orthogonally designed to maximise answerable objective science questions

- Benchmarking to rank homogenisation algorithm skill in terms of performance using climatology, variance and trends calculated from homogeneous test data and inhomogeneous data (test cases applied)

- Independent test data and test case creation and benchmarking

- Peer-reviewed publication of benchmarking methodology but with test case solutions withheld

Questions?