Geophysical Research Abstracts
Vol. 21, EGU2019-18538, 2019
EGU General Assembly 2019

The Use of Large-Scale Datasets for Optimizing Monitoring Networks

Joana Soares (1), Paul Andrew Makar (1), Yayne Aklilu (2), and Ayodeji Akingunola (1)

(1) Air Quality Modelling and Integration Section, Air Quality Research Division, Environment and Climate Change, Toronto (ON), M3H 5T4, Canada, (2) Environmental Monitoring and Science Division, Alberta Environment and Parks, Edmonton (AL), T5J 5C6, Canada

Associativity analysis is a powerful tool to deal with large-scale time series datasets, and can be helpful to assess the extent to which monitoring network locations are representative of a larger region. However, clustering algorithms tend to be very demanding computationally: as the number of data records to be compared (n) increases, the required processing time and memory requirements both scale as O(n2). However, hierarchical clustering of gridded model output has been shown to be useful for the design of air-quality modelling networks (Soares et al., 2018), implying the need for a highly optimized code for carrying out clustering for very large numbers of model-generated data records (n∼300,000, for the grid cells contained in a typical air-quality model domain).

Our underlying methodology makes use of gridded chemical reaction-transport model output as in the place of observation station data - using hierarchical clustering of the time series from each model grid cell, we estimate the similarity of the records using metrics designed to capture (dis)similarities associated with concentration magnitude and temporal variation. This methodology was first applied to hourly observation time series and to hourly model results extracted at grid cells containing station locations to determine the extent to which the synthetic clustering captured the behaviour of the observations. It was then applied to the larger computational problem considering each model grid-cell as a potential station. The first dissimilarity analysis shows comparable results for model and observations with some restrictions. This indicated that hierarchical clustering of gridded model results can be used to generate maps describing sub-regions, within each of which a single station will represent the entire sub-region, to a given level of dissimilarity: these sub-regions may also be thought of as an area of representativeness for a single station (AoR). These maps may be combined with other georeferenced data to assist in monitoring network design. This initial combined analysis provided a “proof of concept” that the methodology can be applied to gridded data, but there is a limit of how many data points can be clustered with convential clustering algorithms. Code optimization efforts were therefore required.

We took the approach of creating a parallel implementation of a hierarchical clustering algorithm. Our overall goal was to achieve effective utilization of parallel hardware that enables the use of this methodology for large datasets to be carried out in a reasonable amount of processing time. The second part of this study will demonstrate that the algorithm is capable of clustering very large air-quality datasets. We demonstrate this methodology as a screening tool to infer the AoR for single stations in a large area, and assist on the decision-making process for the reduction, expansion, or reorganization of a monitoring network.