TOAR Data workshop 1.04

Jülich, 25-28 April 2016

SUMMARY

Note: this document is primarily an attempt to summarize decisions and action items with respect to the data processing. It does not reflect the full breadth of discussions – in particular not those that were related to chapter content. For the larger picture, please see Owen's forthcoming summary for the IGAC newsletter.

Participants: Dene Bowdalo, Owen Cooper, Ruth Doherty, Stephen Edmundson, Sara Fenech, Gina Mills, Martin Schultz, David Simpson, Erika von Schneidemesser

1. Database status

As of 28 April 2016, the TOAR database lists 9497 ozone time series at 9489 stations worldwide. There has been a steady inflow of new data series even during the workshop – this being the final week before the database will be closed for new submissions. The database freezing is needed in order to be able to QA the data series, finalize the calculation of metrics and statistics and produce aggregated datasets for tables and figures that will go into the chapters.

Until just before the workshop all QA activities in Jülich had focused on individual time series. Starting to plot data from many stations on maps or correlating various variables within summertime or growing season aggregated output actually revealed various remaining QA issues which have to be addressed in the coming weeks. Several errors were corrected during the workshop, and the Jülich staff knows that they have to reprocess the AQS data, because the quality flags that were included in this dataset had not been taken over into the TOAR database.

Based on discussions about data selection and filtering, a few additional metadata items have been defined and were/will be added to the database:

station_relative_alt: a measure of the station altitude above its surrounding landscape. This is used to distinguish "mountain stations" from "elevated" stations on a plateau. David Simpson provided an algorithm and dataset to compute this station_relative_alt based on the minimum etopo altitude in a 5 km radius around the station location.

station_toar_category: a TOAR specific classification of stations into "rural", "urban", and "unclassified" based on a combination of filters using the proxy information:

1: "rural, low elevation": (omi_no2_column <= 8) & (nightlight_5km <= 20)
& (population_density_5km <= 100)
& (altitude <= 1500) & (relative_alt <= 500)

2: "rural, high elevation": (omi_no2_column <= 8) & (nightlight_5km <= 20)
& (population_density_5km <= 100)
& (altitude > 1500)

3: "urban": (population_density_5km >= 500) & (nightlight_5km >= 60)

0: "unclassified": everything else

Note that the station_toar_category does not aim to be all-inclusive, but focuses on avoiding false classifications. Visual inspection of filter results looked highly promising that this filter set provides a globally applicable classification (see excel spreadsheet station_metadata_post2005.xlsx).

Addendum:

http://unstats.un.org/unsd/demographic/sconcerns/densurb/densurbmethods.htm:

No clear definition of rural vs urban – depends on country or even region within country.

http://www.nationmaster.com/country-info/stats/Geography/Rural-population-density/Rural-population-per-sq.-km-of-arable-land:

finds population densities between 4.86 (Australia) and 11488 (Djibouti) for arable land. Most countries range below 2000. (year 2003 data)

http://www.hrsa.gov/healthit/toolbox/RuralHealthITtoolbox/Introduction/defined.html:

US definition: rural is less than 2500 people / km2

New criteria:

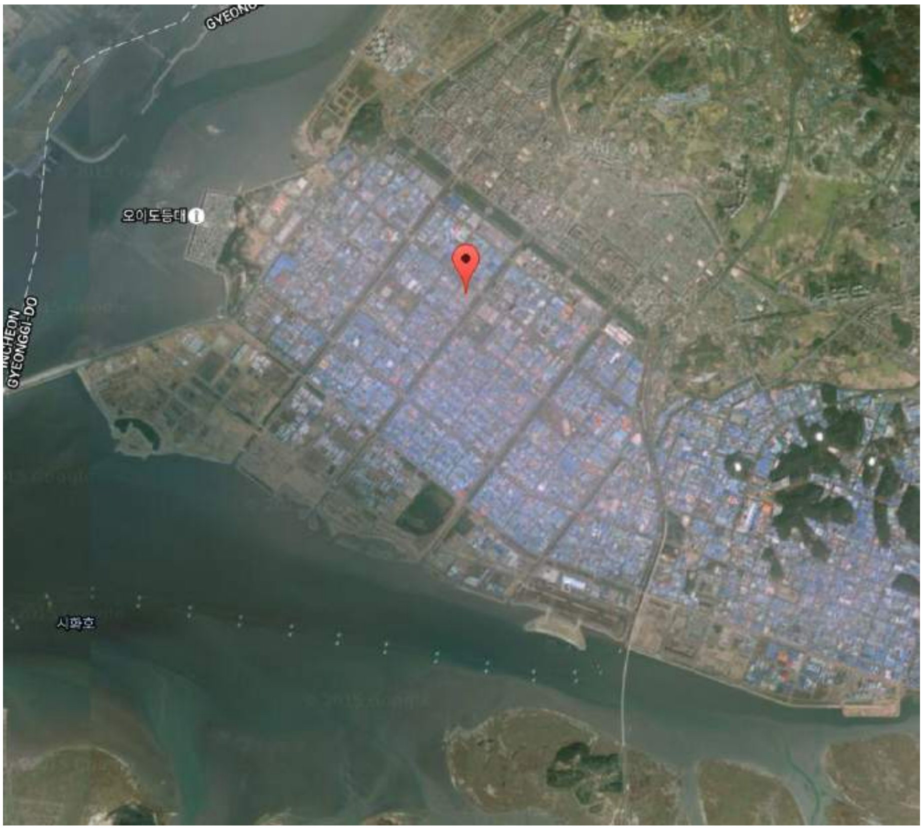
Rural: omi <= 8; light_1km <= 25; pop <= 2500 --- low alt: topo_alt <= 1500; rel alt. < 500 --- high alt: topo_alt > 1500

Urban: nightlight_1 km >= 60; pop_1 km > 2500

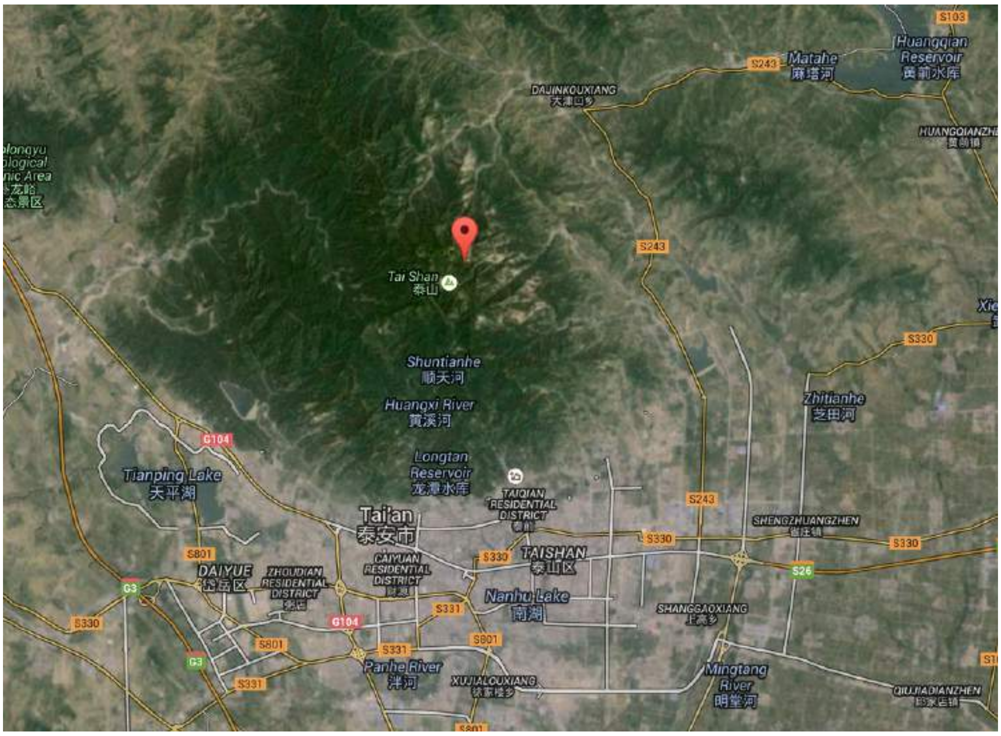
Evaluation:

Out of 12369 stations in total, 2199 are "rural, low altitude", 212 are "rural, high altitude", and 5482 are "urban".

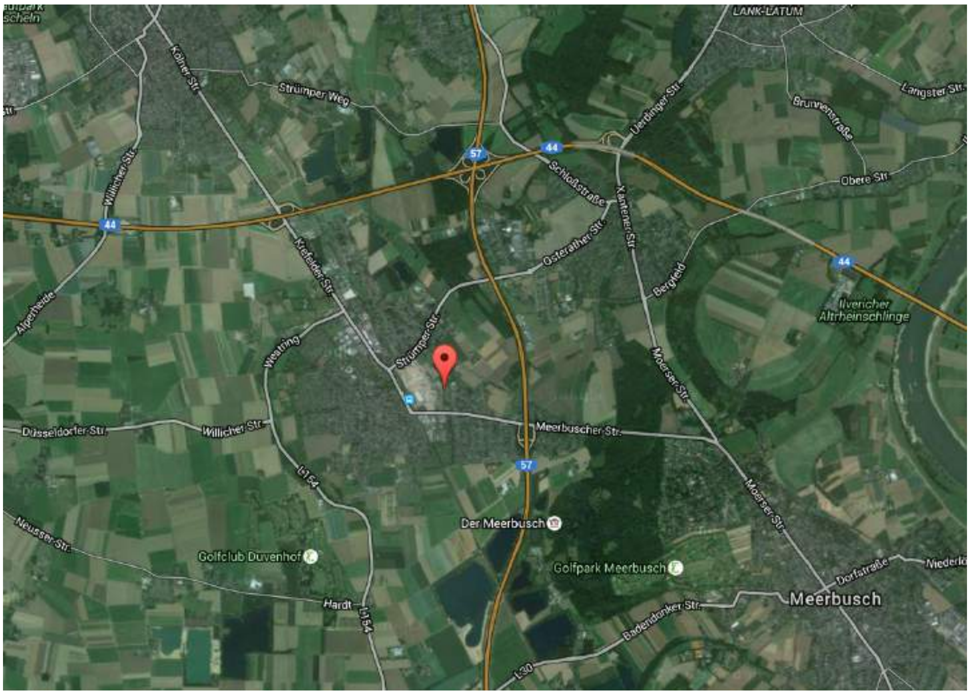
When stations are sorted by descending OMI_NO2 column, the first non-urban station (uncategorized) appears at OMI NO2 column of 15.626 (“Sihwa-IndustrialComplex_Si” in Korea). Here, nightlight_1km is 63, but population density is only 1647 km^{-2}. The next unclassified station is Mt. Tai, China, where OMI NO2 column is 12.9, population density 13227, but nightlight_1km is only 10. Here, the new proxy information “maximum nighttime light within 25 km” (63) and “max. population density within 25 km” (32478) would help to characterize this site as “under urban influence”, although the station altitude of ~1500 m might lead to less polluted signals. The first unclassified non-Asian station is Meerbusch, Germany, with OMI NO2 column = 11.9, population density 10434, and nighttime lights 46. It may be questionable not to classify this site as urban, although the satellite image indicates a rather rural landscape in the vicinity.



Google maps image of the “Sihwa-IndustrialComplex_Si” area in South Korea. The area shown measures approximately 11 x 9 km^2. The blueish houses are all industrial plantations. The MODIS landcover classification labels this site as “urban and built-up”. The 25 km radius around this site is characterized as Water: 30.1 %, Urban and built-up: 25.2 %, Croplands: 21.3 %, Grasslands: 8.8 %, Mixed forest: 8.4 %, Cropland/Natural vegetation mosaic: 2.1 %, Permanent wetlands: 1.3 %, Woody savannas: 1.1 %, Barren or sparsely vegetated: 1.0 %.



Surroundings of the Mount Tai station in China. The city of Tai'an to the south of the station is approximately 8 km away and has about 5.5 mio. Inhabitants. The landcover at the site is "mixed forest". The 25 km radius around the station is composed of Croplands: 77.8 %, Cropland/Natural vegetation mosaic: 9.7 %, Mixed forest: 4.2 %, Urban and built-up: 3.6 %, Grasslands: 2.1 %, Evergreen Needleleaf forest: 1.1 %, Closed shrublands: 1.1 %.



Environment of the German station Meerbusch. Shown is an area of approximately 10 x 7 km^2. The station is classified as "urban and built-up". Its surroundings are Urban and built-up: 43.3 %, Croplands: 30.1 %, Cropland/Natural vegetation mosaic: 11.4 %, Mixed forest: 9.9 %, Evergreen Needleleaf forest: 4.1 %.

station_htap_region: region number of the HTAP "tier1" region definitions, based on a country list and spatial mapping. This is purely informational.

The existing parameter_status flag will be extended to allow identification of data series that are suitable for TOAR analyses. A flag value of 2 (more precisely, bit "2" set) will indicate "insufficient data quality: do not use this data series for TOAR analyses". A flag value of 4 will indicate "time series too short for TOAR analyses (i.e. less than 2 years)". Criteria for omitting data series are semi-objective:

- Too many data gaps (not precisely defined, but basically, if it is impossible to determine a seasonal pattern from looking at the data summary plot)
- Obvious massive problems, i.e. data series just doesn't look like any other ozone series and doesn't seem to make any sense, or it has baseline shifts or jumps > 10 ppb

In addition to taking out entire series, the QA work will flag parts of individual time series as "questionable" or "erroneous" if these portions are clearly wrong (in other words "highly unlikely to represent correct ozone data").

The "TOAR score" that is evaluated with the data summary plots shall be included in the parameter_series table. It will be used to help identifying potentially flawed data series. Data selection by TOAR investigators shall not use this score value but instead rely on the parameter_status flag.

All data series in the TOAR database can be individually accessed through the JOIN web interface (https://join.fz-juelich.de). A few data series are, however, "embargoed", which means you can see the metadata, but not access the data or any statistics derived from these data. These series will be included in the bulk analyses for TOAR, though.

2. Mass data extractions

Just prior to the workshop, Martin started to generate aggregated csv data files containing either monthly, summertime, or annual statistics for all stations, or for stations satisfying certain criteria (e.g. "station_wheat_production > 0.01"). These files are being used for some initial analyses, map plots, etc. Since these data files contain embargoed data, they may not be passed to other colleagues. All workshop participants agreed explicitly to use these data solely for the purpose of performing TOAR analyses and producing the figures for the TOAR chapters. In the event that additional persons will need access to these aggregated data (specifically, we identified Chris Malley and Sverre Solberg), they will have to consent to these rules as well.

The aggregated csv files contain one row per month (monthly extraction) or one row per year (seasonal, summertime, or annual extraction). Each row has the station_id, station_name, several station metadata fields, the datetime, and columns for either relevant health or vegetation statistics. Each aggregate data file will contain a header line showing the data extraction date.

The workshop participants discussed the statistics that shall be included and concluded:

Health metrics:

data_capture, average_values, dma8epax, **drmdmax1h**, **drmdmax1h_day**, somo35, somo10, w90, perc05, perc10, perc25, median, perc75, perc90, perc95, perc98, daytime_avg, daylight_avg, nvgt050, nvgt060, **nvgt070**, nvgt080, nvgt090, nvgt100, nvgt120

The bold red metrics are new additions or still have to be coded. Note that perc98 will not be available in seasonal or monthly aggregates.

Sampling times: annual, summer, seasonal, monthly

Vegetation metrics:

data_capture, average_values, dma8eu, daytime_avg, **m7_avg**, nighttime_avg, daylight_avg, perc05, perc10, perc25, median, perc75, perc90, perc95, w126, aot40, daylight_aot40, dark_aot40

Sampling times: annual, summer (equivalent to 6-months growing season [forest impacts]), wheat growing seasons, rice growing seasons

It is important to preserve the direct link from the database extraction to the final TOAR figures. Therefore, the data analysts agreed to report all their data selection procedures to Jülich and Jülich will then generate data aggregates with these selection criteria directly from the database. This avoids a situation where attempts to re-create TOAR plots from the database fail because of "hand-selected" data, and it ensures full documentation of the data selection process in all TOAR analyes.

3. Trend analyses

Three periods have been fixed which shall be used in all analyses:

"present day": 2010-2014 (end points included); minimum 3 years within this time window. Note that for some summary plots we may include additional sites with less data if there is no other way of showing these data. However, all quantitative analyses will have to enforce these rules.

"trend": 1995-2014 (end points included); minimum 16 years, and no more than two years missing at either end.

"decadal change": 2005-2014 (end points included); data capture not yet fixed. This, shorter, "trend" analysis may allow to include regions which would otherwise not have any trend information. It has to be tried out which sites and/or regions are missing from the trend analysis and could be included in a decadal change analysis without relaxing data capture rules too much. Clearly, the analysis of decadal change will have only little room in the chapters.

The statistics to be applied for trend analyses are Mann-Kendall test with a threshold of pval = 0.05 to indicate a significant change, and Sen-Theil trend estimates using a confidence limit of 90% (alpha = 0.1). The comparability of results between different groups remains to be verified.

4. Example results

4.1 Station classification

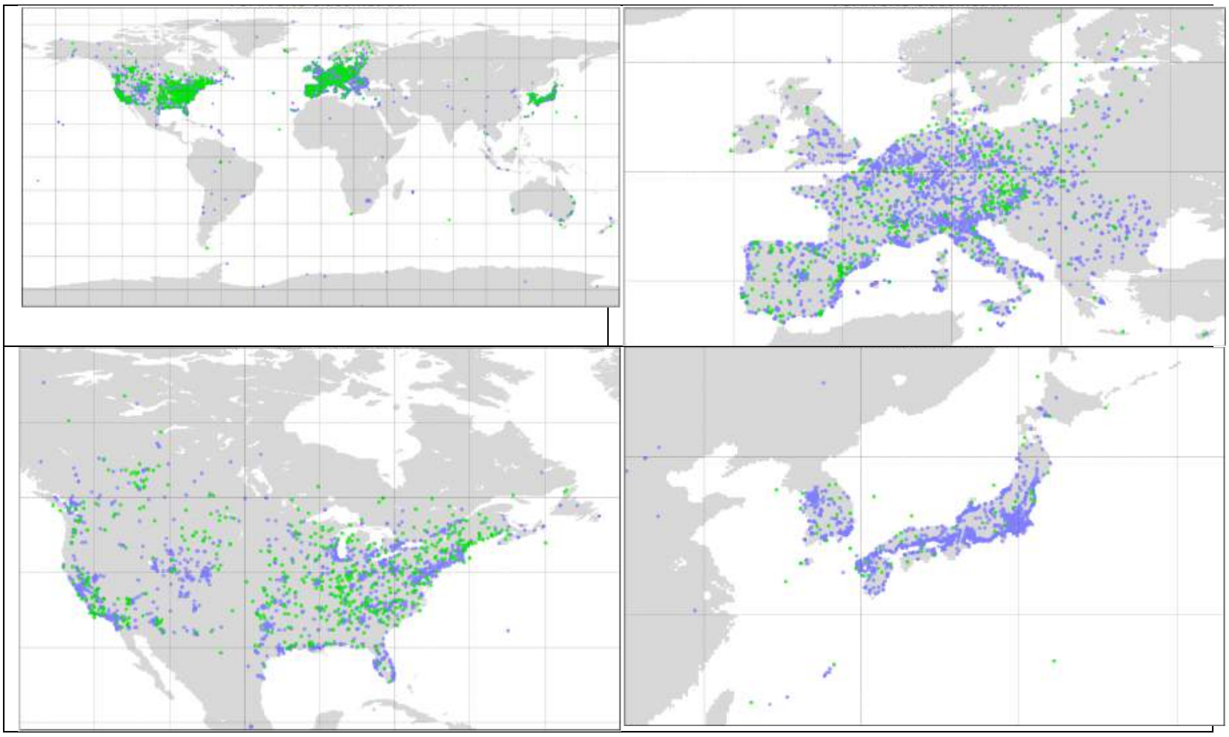


Figure 1: Stations classified as “rural, low altitude” (station_toar_category = 1) based on the proxy filters described in the text (green symbols) on the background of all stations in the TOAR database with data after 2005.

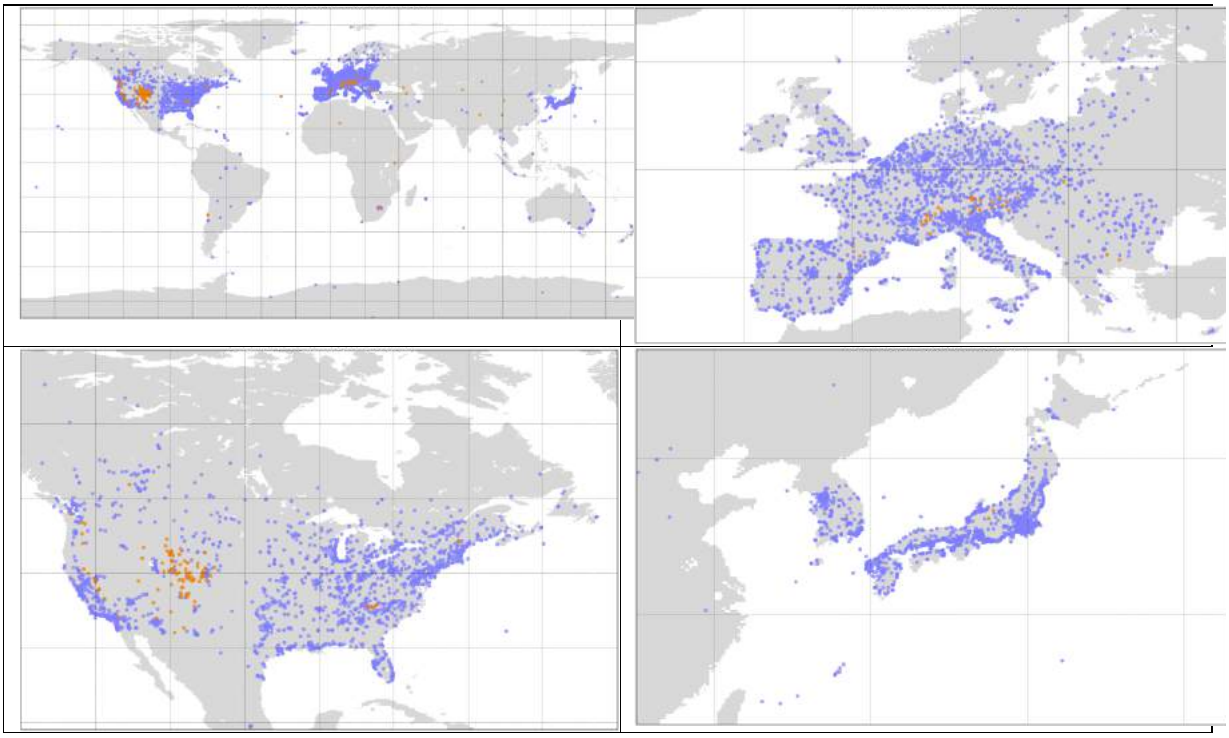


Figure 2: Stations classified as “rural, high altitude” (station_toar_category = 2) based on the proxy filters described in the text (orange symbols) on the background of all stations in the TOAR database with data after 2005.

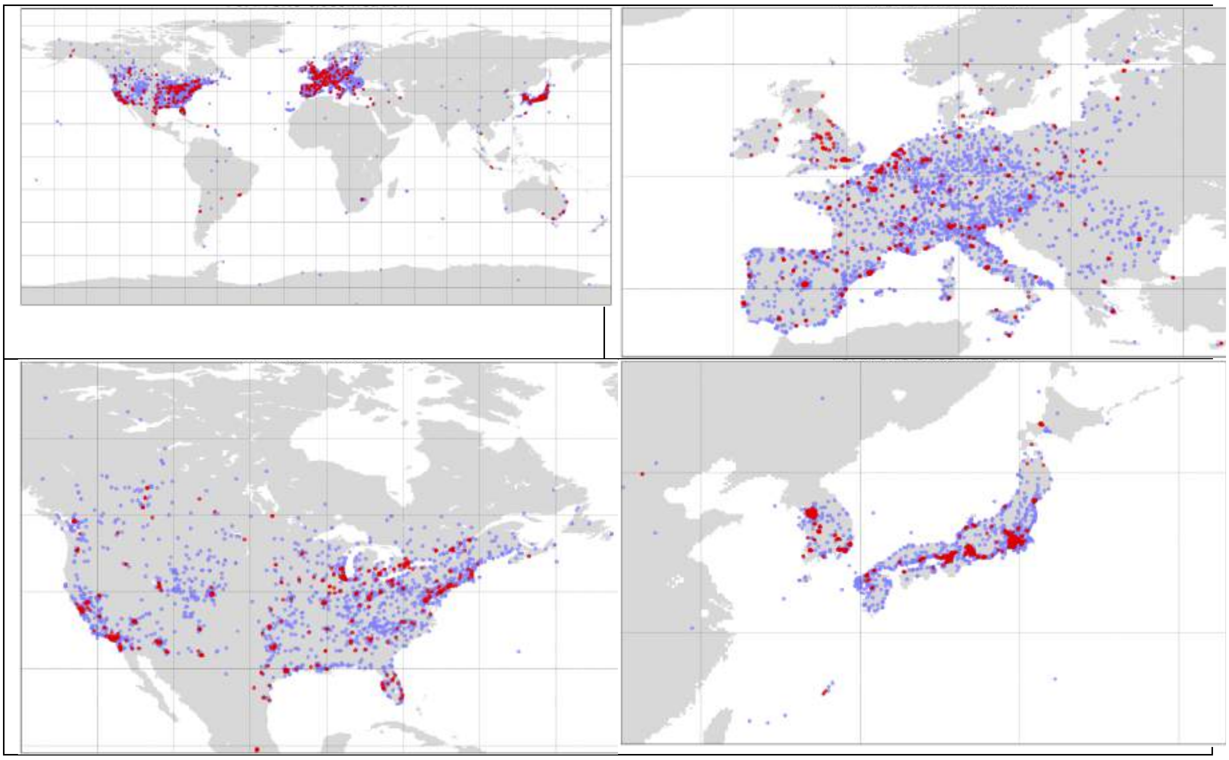


Figure 3: Stations classified as “urban” (station_toar_category = 3) based on the proxy filters described in the text (red symbols) on the background of all stations in the TOAR database with data after 2005.

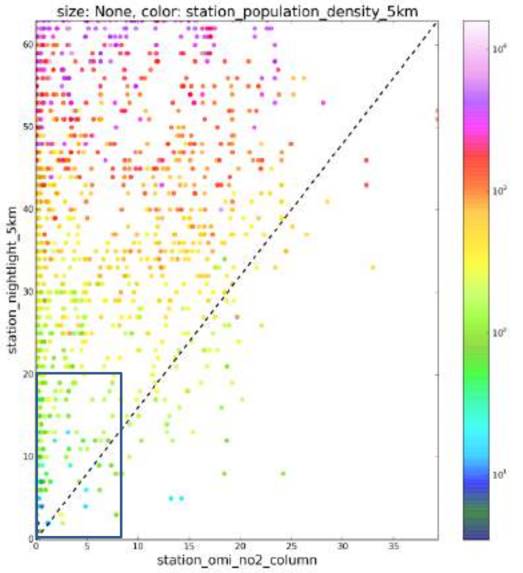
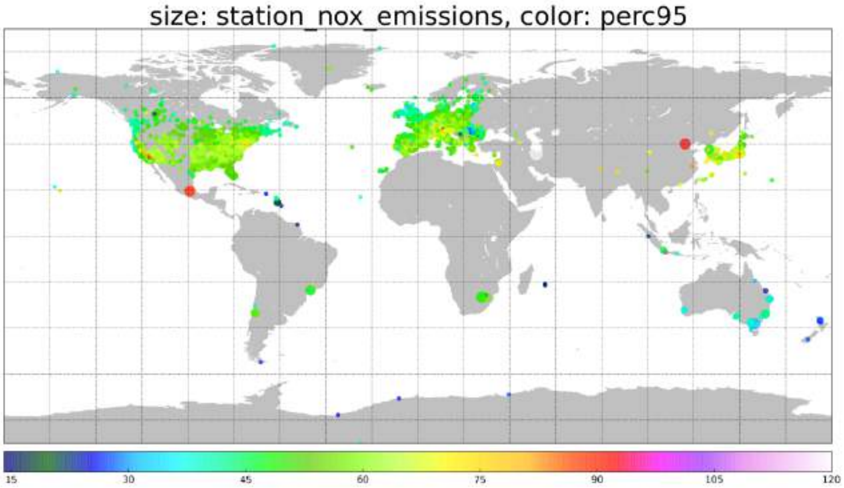


Figure 4: correlation between the proxies “station_omi_no2_column” and “station_nightlight_5km”, colored by the logarithm of “station_population_density_5km”. The rectangle marks the range for filtering “rural” stations.

4.2 Demo maps of an ozone metrics



Shown are summertime 95-percentile ozone mixing ratios of the year 2010. Summer in the northern hemisphere is defined as April-September, in the southern hemisphere it is October-March.

Figure 5: Global year 2010 95-percentiles of summertime ozone mixing ratios. Color scale is ppb. Symbol sizes vary according to log(nox_emissions). All stations with data in 2010 (most recent submissions not yet included).

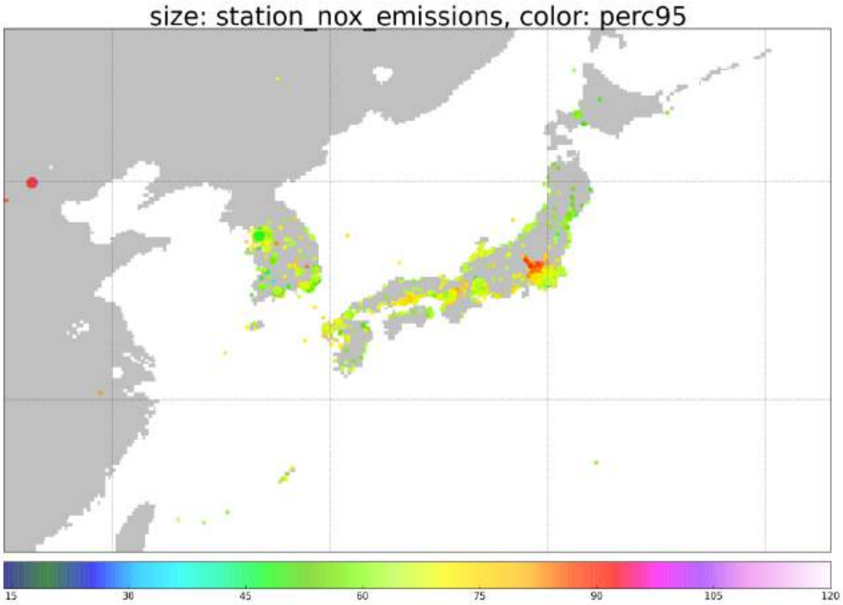


Figure 6: same as Figure 5, but for East Asia. Note: final region boundaries for plots in the TOAR chapters need to be defined.

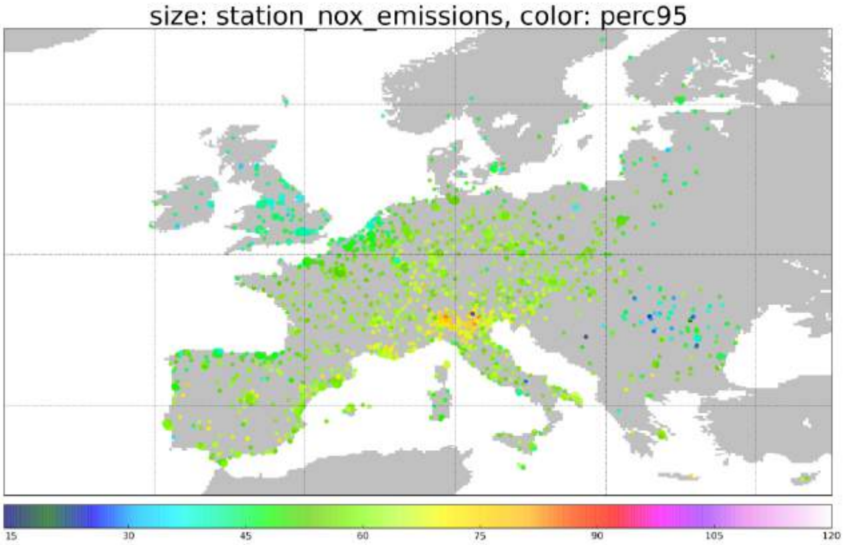


Figure 7: same as Figure 5, but for Europe. Note: final region boundaries for plots in the TOAR chapters need to be defined.

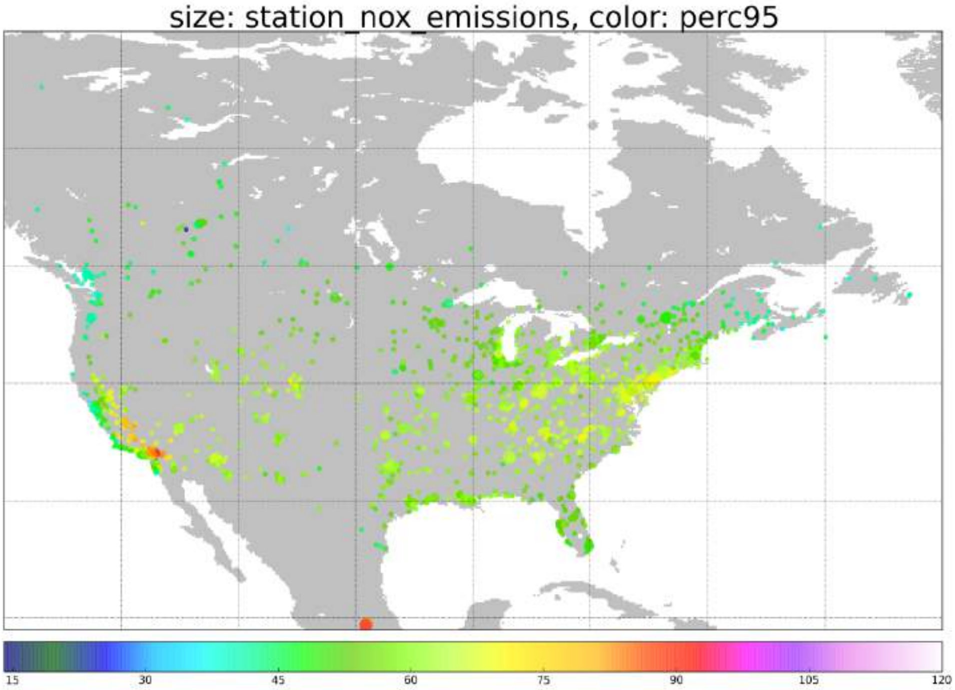


Figure 8: same as Figure 5, but for North America. Note: final region boundaries for plots in the TOAR chapters need to be defined.