

M. G. Schultz et al.

Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations

SUPPLEMENT 1: Documentation of TOAR surface ozone data products

Introduction

The datasets described in this document constitute the central information resource for all analyses in the TOAR report that are related to surface observations of ozone. There are 3 ways to access TOAR data and metadata:

1. The TOAR surface ozone data products (metrics files, statistics, and graphics) are available through the PANGAEA data portal at <https://doi.org/10.1594/PANGAEA.876108>. These data products form the basis of many analyses described in other parts of the TOAR, i.e. other publications in this special feature. They are “frozen data”, which means that all analyses from TOAR should be reproducible with these data products.
2. The JOIN web interface <https://join.fz-juelich.de> allows for easy data selection and display of time series and metrics at individual stations from the “live database”. All data except hourly values can also be downloaded in simple ASCII files. See JOIN user guide (supplement 2 to this article) for details. Note that the TOAR database continues to be further developed, and new data are being added. This implies that users may find data that have not been available for analyses described in TOAR, or that individual datasets may differ from those that were used in TOAR. We maintain a frozen version of the TOAR database, which is consistent with the data products on PANGAEA; however, this version is currently not accessible for outside users.
3. The JOIN interface also provides Representational State Transfer (REST) services. These services return metadata or data in response to a query URL (web address) thereby allowing the inclusion of TOAR data in your own programs or web applications. See <https://join.fz-juelich.de/services/rest/surfacedata/> and the JOIN user guide (supplement 2 to this article) for details. Currently, the REST services are connected to the “live database”. We are planning to also provide access to the frozen database via REST services.

If you make use of TOAR data products obtained through any of these means, we request that you acknowledge TOAR as follows:

“[We thank] the Tropospheric Ozone Assessment Report (TOAR) initiative for providing the surface ozone data [*and/or*: analysis] shown in [*or*: used by] this report [*or*: publication].”

Also, please include a reference to the TOAR Database and Metrics Data paper and the PANGAEA TOAR data repository:

Schultz, MG, et al.: Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elem. Sci. Anth., September 2017.

<https://doi.org/10.1525/elementa.244> .

Schultz, MG, et al.: Tropospheric Ozone Assessment Report: Global Surface Ozone Data Products, 2017. <https://doi.org/10.1594/PANGAEA.876108>.

This document describes the content of the TOAR data portal on PANGAEA (as far as surface ozone data products are concerned) and provides additional information on the methods that were employed to derive the aggregated statistics and trends. For questions that go beyond the content of this report, please contact Martin Schultz directly (m.schultz@fz-juelich.de).

Contents:

1. Content of the TOAR data portal.....	4
2. Overview.....	5
3. Common analysis periods and sampling rules	6
4. Metrics file names and definitions of metrics.....	10
5. Data file formats.....	15
6. Station metadata.....	17
7. TOAR site classification.....	21
8. The data series merging procedure.....	24
9. Plot gallery.....	30

1. Content of the TOAR surface ozone data portal

The TOAR data portal can be freely accessed at <https://doi.org/10.1594/PANGAEA.876108>. The parent node of this data collection directs you to the different data products provided by TOAR:

Surface ozone data:

- Pre-compiled metrics data sets (ASCII)
- Gridded ozone data sets (NetCDF)
- Plots of TOAR ozone metrics
- Software that was used to generate TOAR data products and calculate ozone metrics

The pre-compiled metrics data sets consist of comma separated values (CSV) formatted text files which can readily be imported in Microsoft Excel© or other data analysis programs (the separator character is ‘;’. The files contain different numbers of header lines. Each header line starts with ‘#’, and the total number of header lines is listed at the top of the file. The column headers do not count as header lines. Details on the file format(s) can be found in section 4.

The main repository of metrics files contains “aggregated_statistics”, “trend_statistics”, and “yearly_statistics”. Aggregated statistics are mean values over the respective multi-year intervals denoted in the file names, i.e. one data line per station (for details see sections 2 and 3). If you are interested in the state of surface ozone during one of these intervals, these files will be your first choice. “Trend statistics” contain the trends statistics of the non-parametric Mann-Kendall test and Sen-Theil trend estimates for various time intervals (again one line per station), while “yearly statistics” summarize the calculated metrics of each individual year, i.e. the number of lines varies by station. Aggregated statistics and trend statistics have been derived from the “yearly statistics” files. Details on the calculated metrics, the data selection, metadata, etc. can be found in the following sections.

Gridded data sets are supplied in the form of CF compliant NetCDF files in resolutions of $10^{\circ} \times 10^{\circ}$, $5^{\circ} \times 5^{\circ}$, and $2^{\circ} \times 2^{\circ}$. We recommend use of the 5° longitude $\times 5^{\circ}$ latitude products as they provide a reasonable compromise between global coverage and regional differentiation.

The collection of “plots” contains “box-whisker-plots present-day”, “box-whisker-plots trends”, “maps gridded”, “maps present-day”, and “maps trends” figures. See section 9 for examples and detailed links.

2. Overview

Table 1 lists all (ASCII) metrics files that have been produced. Detailed explanations can be found in the following sections of this document. It is important that you familiarize yourself with the data selection criteria applied in each metrics file so that you can perform the right analyses on the right dataset and draw correct conclusions.

Table 1: Overview of available metrics files on <https://doi.org/10.1594/PANGAEA.876108>. The cell values denote the number of stations included in each dataset

	“present” (2010-2014)	“trend” (1995-2014)	“decadal” (2005-2014)	“long-term” (1970-2014)	“maximum coverage” (2008-2015)
monthly	4812	1689	3755	913	--
seasonal	4812	1689	3755	913	6136
summer	4812	1689	3755	913	6136
annual	4812	1689	3755	913	6136
rice	1131	286	862	175	--
wheat	2918	1047	2249	541	--

In order to ensure reproducibility of TOAR results, you are kindly requested to perform your analyses with as little extra filtering and processing as possible. For example, do not use the monthly statistics if you want to show summertime values, but use the pre-calculated summer statistics instead. This way we can ensure consistency between your analyses and the analysis presented in the TOAR papers of the Elementa special issue. If you detect errors in the statistics or if you need an additional filter set, please contact Martin Schultz, and we can expand the list of pre-calculated metrics files.

Feel free to explore different options to filter the results according to the extensive metadata that are provided for each station (for example, `nighttime_lights`, `population_density`, `dominant_landcover`, etc.). However, please refrain from inventing your own categorization of stations as “urban” or “rural” in the context of the TOAR analyses. We have developed a *TOAR_category* classification based on the combination of several metadata elements (see main paper and section 6 of this document). This classification again ensures consistency with TOAR analyses.

Before you go ahead and start analyzing this fascinating dataset, please make sure to carefully read through the following sections of this document so that you understand how the database is built, how the data extraction works, and how the metrics are calculated.

3. Common analysis periods and sampling rules

In Table 1 above you saw the common TOAR analysis periods as column headings and various sampling intervals as row labels.

The common analysis periods were defined at the TOAR data workshop in Jülich in April 2016 and slightly expanded afterwards. By defining the rules that are explained below we wanted to achieve a “reasonable statistical representation” of data while at the same time trying to maximize the number of sites that can be included in each analysis. Furthermore, these common periods shall ensure comparability of results across the globe. Table 2 lists the specifics for each common analysis period and the rationale behind the choices made.

Table 2: TOAR common analysis periods and associated data requirements

Name	Year range	Data requirements	Comments
<i>present</i>	2010-2014	at least three “ozone seasons” during the analysis interval, i.e. 2.5 years	selection criteria optimized to include data from recently opened stations as well as data from stations with slow delivery; compromise between robust statistics (≥ 5 years) and inclusion of sites with sparse data
<i>trend</i>	2000-2014	at least 12 “ozone seasons” during the analysis interval, i.e. 11.5 years, and not more than 2 years missing from either end (i.e. data must commence no later than April 2002 and stop no sooner than September 2012)	provide a robust set of metrics for trend analyses. The interval 2000-2014 was chosen to allow inclusion of the vast Korean data set which commences only in the year 2000.
<i>trend(2)</i>	1995-2014	at least 16 “ozone seasons” during the analysis interval, i.e. 15.5 years, and not more than 2 years missing from either end (i.e. data must commence no later than April 1997 and stop no sooner than September 2012)	alternative period for trend analysis. Statistically more robust, but contains fewer stations than trend 2000-2014.
<i>decadal</i>	2005-2014	at least 8 “ozone seasons” during the analysis interval, i.e. 7.5 years	many stations, particularly in Asia, have no observations prior to the early 2000’s. The “decadal” dataset shall allow derivation of some change signal for such stations which are not included in the “trend” dataset. Any analysis based on these data should not use the word “trend”, however

Table 2, continued

<i>long-term</i>	1970-2014	at least 25 years of data with some data available before 1990	the set of stations which allow analysis of long-term changes. Note that start and end dates within this long interval may still vary considerably
<i>maximum coverage</i>	2008-2015	at least 9 months of data	limited set of statistics (e.g averages, median and 25 th and 75 th percentile) to allow at least some presentation of stations which would otherwise not appear in the TOAR report, and which are often located in world regions where no information has been available until now.

Figure 1 shows the set of stations included in the “present day” analyses (grey dots) and highlights stations which are additionally included in the 2008-2015 “max coverage” datasets, because they have too little data for the 2010-2014 analysis (red dots). Note that the “max coverage” data only includes simple averages, medians, and percentiles. Other statistics would not be robust for short timeseries and have therefore not been evaluated.

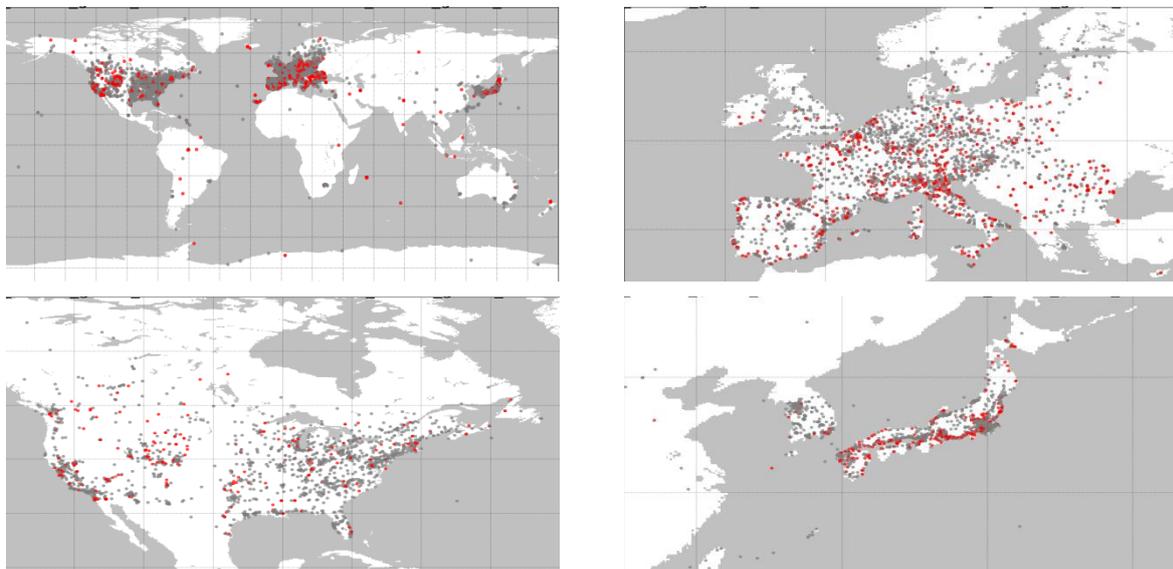


Figure 1: Maps of stations which are included in the TOAR present-day analyses (grey dots) and the “max coverage” datasets (red dots and grey dots). See Table 1 for station counts.

Be careful in your selection of the appropriate dataset if you wish to analyze or display data from sub-periods of the common analysis periods listed above. For example, to compare maps of ozone data between 2011 and 2013 with data between 2006 and 2008, you should use the 2005-2014 dataset and not the 1995-2014 or 1970-2014 ones, because these will miss many sites which began measuring only after 2000. Even then you will miss stations which stopped operation shortly after 2008, because of the data series length requirements. Therefore, if such needs arise, please contact Martin Schultz who will run a custom extraction query for you.

The sampling intervals are largely straightforward:

- *monthly*: one row of data for every month within the analysis period,
- *seasonal*: one row of data for every year within the analysis period; each statistics is evaluated for the four standard meteorological seasons DJF, MAM, JJA, and SON,
- *summer*: one row of data for every year within the analysis period; summer is defined as April-September in the northern hemisphere and October-March in the southern hemisphere,
- *annual*: one row of data for every year within the analysis period,
- *rice*: one row of data for every year within the analysis period; the statistics are evaluated during the growing period of rice in the climatic zone of each station (see Table 3),
- *wheat*: like rice, but for growing seasons of wheat (see Table 4).

The *rice* and *wheat* seasons correspond to “sampling=vegseason, crop=rice”, and “sampling=vegseason, crop=wheat”, respectively, when you are using the TOAR python scripts for data extraction or the REST service provided by the [JOIN web interface](#) (see documentation in Supplementary Material 2).

Table 3: Rice growing seasons defined as 3-months intervals for the TOAR analyses (courtesy of Gina Mills, CEH, Edinburgh)

Growing season name	month list (January = 1)
rice-cool_temperate_moist-NH	[5, 6, 7]
rice-cool_temperate_dry-NH	[5, 6, 7]
rice-warm_temperate_moist-NH	[6, 7, 8]
rice-warm_temperate_dry-NH	[6, 7, 8]
rice-tropical_moist-NH	[7, 8, 9]
rice-tropical_dry-NH	[8, 9, 10]
rice-tropical_wet-NH	[7, 8, 9]
rice-tropical_montane-NH	[7, 8, 9]
rice-tropical_dry-SH	[1, 2, 3]
rice-tropical_wet-SH	[12, 1, 2]
rice-tropical_moist-SH	[12, 1, 2]
rice-warm_temperate_moist-SH	[11, 12, 1]
rice-warm_temperate_dry-SH	[1, 2, 3]
rice-cool_temperate_moist-SH	[12, 1, 2]
rice-cool_temperate_dry-SH	[12, 1, 2]

Table 4: Wheat growing seasons defined as 3-months intervals for the TOAR analyses
(courtesy of Gina Mills, CEH, Edinburgh)

Growing season name	month list (January = 1)
wheat-boreal_moist-NH	[6, 7, 8]
wheat-boreal_dry-NH	[6, 7, 8]
wheat-cool_temperate_moist-NH	[4, 5, 6]
wheat-cool_temperate_dry-NH	[4, 5, 6]
wheat-warm_temperate_moist-NH	[3, 4, 5]
wheat-warm_temperate_dry-NH	[3, 4, 5]
wheat-tropical_montane-NH	[1, 2, 3]
wheat-tropical_wet-NH	[1, 2, 3]
wheat-tropical_moist-NH	[1, 2, 3]
wheat-tropical_dry-NH	[1, 2, 3]
wheat-tropical_moist-SH	[7, 8, 9]
wheat-tropical_dry-SH	[8, 9, 10]
wheat-warm_temperate_moist-SH	[8, 9, 10]
wheat-warm_temperate_dry-SH	[8, 9, 10]
wheat-cool_temperate_moist_le30-SH	[2, 3, 4]
wheat-cool_temperate_dry_le30-SH	[2, 3, 4]
wheat-cool_temperate_moist_gt30-SH	[11, 12, 1]
wheat-cool_temperate_dry_gt30-SH	[11, 12, 1]

4. Metrics file names and definitions of metrics

The pre-compiled metrics files on the TOAR data portal follow the naming convention:

TOAR_sfc_ozone_SAMPLING_REGION_DATERANGE_FILETYPE.csv, where *SAMPLING* denotes a sampling interval for data aggregation (see bullet list in section 2), *REGION* will always be “global” at present, *DATERANGE* is the range of years for which the extraction took place (see Table 2), and *FILETYPE* is either “aggregated”, “trends”, or empty (i.e. “yearly”). Example filenames are thus:

- *TOAR_sfc_ozone_annual_global_2005-2014.csv*: annual statistics with yearly values for the time period 2005 to 2014. A list of variables contained in “annual” files can be found in Table 5.
- *TOAR_sfc_ozone_rice_growing_season_global_2010-2014_aggregated.csv*: present day data, aggregated over the 2010 to 2014 time period, for stations in rice growing regions ($\text{station_rice_production} > 0.001 \text{ tons/km}^2$), averaged over the respective rice growing seasons (Table 3) depending on the *station_climatic_zone*.
- *TOAR_sfc_ozone_seasonal_global_2000-2014_trends.csv*: trend statistics from 2000 to 2014 calculated for individual seasons (DJF, MAM, JJA, and SON). Table 5 list sthe statistics that are available in such files.

Gridded data files contain data in monthly time resolution and adopt the following naming convention:

TOAR_monthly_TIMEPERIOD_METRIC_gridded_RESOLUTION.nc

Example: *TOAR_monthly_2010-2014_daytime_avg_gridded_5x5.nc*

Table 5 lists the various metrics sets that have been defined, and Table 6 explains each individual metrics in more detail.

Table 5: Metrics sets (collection of metrics) used in the pre-compiled metrics files

Name of set	Metrics included
annual, summer	data_capture, average_values, dma8epa, dma8epa_strict, dma8epax, dma8epax_strict, dma8eu, dma8eu_strict, avgdma8epax, somo10, somo10_strict, somo35, somo35_strict, w90, nvgt050, nvgt060, nvgt070, nvgt080, nvgt090, nvgt100, nvgt120, median, perc05, perc10, perc25, perc75, perc90, perc95, perc98, aot40, daylight_aot40, w126, w126_24h, daytime_avg, nighttime_avg, drmdmax1h, day_of_max_drmdmax1h
seasonal	data_capture, average_values, dma8epax, w90, w126, median, perc05, perc10, perc25, perc75, perc90, perc95, daytime_avg, nighttime_avg, daylight_avg, nvgt050, nvgt060, nvgt070, nvgt080, nvgt090, nvgt100, nvgt120
rice and wheat growing seasons	data_capture, average_values, dma8epa, dma8epa_strict, dma8epax, dma8epax_strict, dma8eu, dma8eu_strict, avgdma8epax, perc05, perc10, perc25, perc75, perc90, perc95, perc98, aot40, daylight_aot40, w126, w126_24h, daytime_avg, nighttime_avg
2008-2015 data	data_capture, average_values, daytime_avg, avgdma8epax, median, perc25, perc75, perc95, perc98, dma8epax

Table 6: Definition of metrics used in the TOAR analyses. Note that -DJF, -MAM, -JJA, -SON is appended to these variable names in case of seasonal statistics.

Name	Description
data_capture	Fraction of valid (hourly) values available in the aggregation period.
average_values	Daily, monthly, ... average value. No data capture criterion is applied, i.e. a daily average is valid if at least one hourly value of the day is present.
daytime_avg	Daytime average is defined as average of hourly values for the 12-h period from 08:00h to 19:59h solar time. All hourly values in the aggregation period are averaged, and the resulting value is valid if at least 75% of hourly values are present.
nighttime_avg	Same as daytime_average but accumulated over the daily interval from 20:00 h to 07:59 h solar time.
median	Median mixing ratio over the aggregation period. At least 10 valid values must be present to accept a median value as valid.
perc05	Fifth-percentile of hourly values in the aggregation period. At least 10 valid values must be present to accept a percentile value as valid.
perc10	As perc05, but for the 10 th -percentile.
perc25	As perc05, but for the 25 th -percentile.
perc75	As perc05, but for the 75 th -percentile.
perc90	As perc05, but for the 90 th -percentile.
perc95	As perc05, but for the 95 th -percentile.
perc98	As perc05, but for the 98 th -percentile. This percentile is only calculated for "summer" or "annual" aggregation periods.

Table 6, continued

dma8epa	Daily maximum 8-hour average statistics according to the US EPA definition. 8-hour averages are calculated for 24 bins starting at 0 h local time. The 8-h running mean for a particular hour is calculated on the concentration for that hour plus the following 7 hours. If less than 75% of data are present (i.e. less than 6 hours), the average is considered missing. When the aggregation period is “seasonal”, “summer”, or “annual”, the 4th highest daily 8-hour maximum of the aggregation period will be computed. Note that in contrast to the official EPA definition, a daily value is considered valid if at least one 8-hour average is valid.
dma8epa_strict	As dma8epa, but additionally, a diurnal 8-hour maximum value is only saved if at least 18 out of the 24 8-hour averages are valid. This is the official dma8epa definition.
dma8epax	As dma8epa, but using the new US EPA definition of the daily 8-hour window from 7 h local time to 23 h local time.
dma8epax_strict	As dma8epax, but additionally, a diurnal 8-hour maximum value is only saved if at least 13 out of the 17 8-hour averages are valid. This is the official dma8epax definition.
dma8eu	As dma8epa, but using the EU definition of the daily 8-hour window starting from 17 h of the previous day. When the aggregation period is “seasonal”, “summer”, or “annual”, the 26th highest daily 8-hour maximum of the aggregation period will be computed.
dma8eu_strict	As dma8eu, but additionally, a diurnal 8-hour maximum value is only saved if at least 18 out of the 24 8-hour averages are valid. This is the official dma8eu definition.
avgdma8epax	Average value of the daily dma8epax statistics during the aggregation period.
drmdmax1h	Maximum of the 3-months running mean of daily maximum 1-hour mixing ratios during the aggregation period.
day_of_max_drmdmax1h	Julian day in the year when the maximum value of the 3-months running mean of daily maximum 1-hour concentrations occurred.
somo10	Sum of excess of daily maximum 8-h means (EU Airbase standard with relaxed criterion: dma8eu) over the cut-off of 10 ppb, i.e. 20 µg/m ³ calculated for all days in the aggregation period. SOMO10 will be set to missing if less than 75% of days are available. The quantity will be weighted by the number of theoretical days over the number of available days.
somo10_strict	As somo10, but using dma8eu_strict for data capture.
somo35	As somo10, but accumulating ozone values above 35 ppb.
somo35_strict	As somo10_strict, but accumulating ozone values above 35 ppb.

Table 6, continued

w90	<p>Daily maximum W90 5-h Experimental Exposure Index: $EI = \text{SUM}(w_i \cdot C_i)$ with weight $w_i = 1/[1 + M \cdot \exp(-A \cdot C_i/1000)]$, where $M = 1400$, $A = 90$, and where C_i is the hourly average O3 mixing ratio in units of ppb (Lefohn et al., 2010). For each day, 24 W90 indices are computed as 5-hour sums, requiring that at least 4 of the 5 hours are valid data (75%). If a sample consists of only 4 data points, a fifth value shall be constructed from averaging the 4 valid mixing ratios.</p> <p>For aggregation periods “month”, “season”, “summer”, or “annual”, the 4th highest W90 value is computed, but only if at least 75% of days in this period have valid W90 values.</p>
aot40	<p>Daily 12-h AOT40 values are accumulated using hourly values for the 12-h period from 08:00h until 19:59h solar time interval. AOT40 is defined as cumulative ozone above 40 ppb. If less than 75% of hourly values (i.e. less than 9 out of 12 hours) are present, the cumulative AOT40 is considered missing. When there exist 75% or greater data capture in the daily 12-h window, the scaling by fractional data capture ($n_{\text{total}}/n_{\text{valid}}$) is utilized.</p> <p>For monthly, seasonal, summer, or annual statistics, the daily AOT40 values are accumulated over the aggregation period and scaled by ($n_{\text{total}}/n_{\text{valid}}$) days. If less than 75% of days are valid, the value is considered missing.</p>
daylight_aot40	As aot40, but using solar elevation > 5 degrees to identify “daytime” hours.
w126	<p>Daily W126 index is accumulated using hourly values for the 12-h period from 08:00h until 19:59h solar time interval. $W126 = \text{SUM}(w_i \cdot C_i)$ with weight $w_i = 1/[1 + M \cdot \exp(-A \cdot C_i/1000)]$, where $M = 4403$, $A = 126$, and where C_i is the hourly average O3 mixing ratio in units of ppb. If there are less than 9 valid hourly values in the 12 hour window, the daily value is considered missing. When there exist 75% or greater data capture in the daily 12-h window, the scaling by fractional data capture ($n_{\text{total}}/n_{\text{valid}}$) is utilized.</p> <p>Seasonal, summer, or annual statistics are calculated as sum over the daily W126 values. Results are marked as missing if less than 75% of daily values are valid.</p>
w126_24h	As w126, but using all 24 hours of a day.
nvgt050	Number of days with exceedance of the dma8epax value above 50 ppb. The value is marked as missing if less than 75% of days contain valid data.
nvgt060	Number of days with exceedance of the dma8epax value above 60 ppb. The value is marked as missing if less than 75% of days contain valid data.
nvgt070	Number of days with exceedance of the dma8epax value above 70 ppb. The value is marked as missing if less than 75% of days contain valid data.
nvgt080	Number of days with exceedance of the dma8epax value above 80 ppb. The value is marked as missing if less than 75% of days contain valid data.
nvgt090	Number of days with exceedance of the daily max1h_values above 90 ppb. The value is marked as missing if less than 75% of days contain valid data.

Table 6, continued

nvgt100	Number of days with exceedance of the daily max1h_values above 100 ppb. The value is marked as missing if less than 75% of days contain valid data.
nvgt120	Number of days with exceedance of the daily max1h_values above 120 ppb. The value is marked as missing if less than 75% of days contain valid data.

5. Data file formats

5.2 ASCII files – metrics by station

The ASCII metrics files are semicolon-separated csv file. Figure 2 presents a snapshot of one metrics file when loaded into MS-Excel for illustration. The first line contains the number of header lines excluding the column header (i.e. variable names). Each header line starts with “#” and contains a key: value pair of metadata. The first block contains general information on TOAR and the surface ozone datasets, the second block has variable definitions. The data block (in the example starting from line 80) begins with the column headers (variable names).

	A	B	C	D	E	F	G	H	I	J					
1	#number_of_comment_lines: 79														
2	#title: Tropospheric Ozone Assessment Report (TOAR) aggregated metrics file														
3	#filename: TOAR_sfc_ozone_annual_global_2010-2014_aggregated.csv														
4	#license: CC-BY4 (https://creativecommons.org/licenses/by/4.0/legalcode)														
5	#author: Schultz, MG, Forschungszentrum Juelich GmbH, Juelich, Germany (m.schultz@fz-juelich.de)														
6	#citation: Schultz, MG et al. (2017): Tropospheric Ozone Assessment Report, Global surface ozone datasets. doi:10.1594/PANGAEA.876108														
7	#project: Tropospheric Ozone Assessment Report (TOAR), an IGAC-sponsored activity														
8	#data_source: Surface ozone database at Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany														
9	#data_source_description: The TOAR surface ozone database contains hourly surface ozone measurements from over 9,600 stations around the world made between 1970 and														
10	#data_source_reference: Schultz, MG et al. (2017): Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations. Elementa - Scier														
11	#data_use_policy: If you make use of this dataset in your own presentations or publications, please do three things: 1. cite the dataset (see citation above), 2. cite the data_so														
12	#dataset_description: This file contains several statistical data were aggregated as average annual values during this period														
13	#data_aggregation_period: annual														
14	#date_range: 2010-01-01 to 2014-12-31														
15	#column01: numid	internal station number (unique identifier)													
16	#column02: network_name	name of the measurement network/data center from which the data originated													
17	#column03: station_id	alphanumeric identifier of station according to network conventions													
18	#column04: station_name	name of the station as provided by network data center or station PI													
19	#column05: station_country	country operating the station (ISO names with few exceptions)													
20	#column06: station_type	characterization of station as "traffic", "industry", or "background" according to data provider													
80	numid	network_name	station_id	station_name	station_country	station_type	station_latitude	station_longitude	station_altitude	station_population	station_max_population	station_max_elevation	station_high_statistics		
81	3338	UBA	DEUB028	Zingst	Germany	background rural	54.43667	12.72528	1	1	1	349	386	6619	9
82	3339	UBA	DENW094	Aachen-Burt	Germany	background urban	50.754704	6.093923	205	200	66	14514	20125	26839	48
83	3340	UBA	DENW067	Bielefeld-Os	Germany	background urban	52.023169	8.548389	102	106	29	16176	16191	16191	60
84	3341	UBA	DENW081	Borken-Gem	Germany	background rural	51.862	6.874554	45	45	8	3566	3569	12635	48
85	3342	UBA	DENW021	Botrop-Wel	Germany	industrial urban	51.525963	6.976908	40	31	6	20282	34286	39682	61
86	3343	UBA	DENW008	Dortmund-Ei	Germany	background urban	51.536911	7.457463	75	68	6	27057	27082	39682	62
87	3347	UBA	DENW058	HÄxirith	Germany	industrial suburban	50.876121	6.873789	90	92	20	9046	13039	31661	58
88	3348	UBA	DENW042	Krefeld-Linn	Germany	background urban	51.337669	6.640242	32	34	5	16721	22153	36502	55
89	3351	UBA	DENW079	Leverkusen-	Germany	background suburban	51.028879	7.005063	50	50	9	28743	31605	34559	61

Figure 2: Snapshot of a metrics file (here TOAR_sfc_ozone_annual_global_2010-2014_aggregated.csv) when opened in MS-Excel. Note that in “yearly statistics” files the number of data lines may vary among stations depending on the length of time series within in the analysis period.

Most variable names should be self-explanatory if you consult the metrics table (Table 6). The *datetime* variable is always in column 26 and marks the beginning of the statistical data. For the “RICE” and “WHEAT” metrics, *datetime* appears in column 27, because there is one additional station metadata column included in the data files. In case of seasonal statistics, each variable is quadrupled and carries the season label (“-DJF”, “-MAM”, “-JJA”, or “-SON”) as suffix.

If you wish to read these data into python, we recommend that you use the *pandas* package and read the data into a dataframe with:

```
data = pd.read_csv(filename, sep=';', header=79, index_col=None,
                  skipinitialspace=True, parse_dates=False)
```

You can then use `data.groupby('numid')` to group all data belonging to one station and one of the aggregation methods of `groupby` to calculate mean values over the analysis period etc.

5.2 NetCDF files – gridded data products

The gridded surface ozone data files are NetCDF binary files (see <http://www.unidata.ucar.edu/software/netcdf/>). NetCDF is a self-documenting format that uses attributes to describe the dataset and its variables. The gridded TOAR data files follow the Climate and Forecasting (CF) conventions, version 1.6 (see <http://cfconventions.org/>).

Each gridded data file contains the variables described in Table 7. Note that only data from stations at altitudes less than 2000 m were used in order to obtain a consistent “surface ozone” dataset.

Table 7: Variables contained in gridded TOAR data products. Each variable has dimensions (month/time, latitude, longitude).

Variable name	Description
all_count	number of valid data (stations) included in each grid box at every output interval (month); “all stations”, i.e. no selection by TOAR category is applied
all_mean	mean value of metrics in grid box; “all stations”, i.e. no selection by TOAR category is applied. Explanation: all_mean of TOAR_monthly_2010-2014_daytime_avg_gridded_5x5.nc contains the grid box average values of the daytime mean values in a given month, temporally averaged over the 2010-2014 period. Note that a minimum of 3 valid years of data is also required before data from a station are added and counted.
all_median	median value of metrics in grid box; “all stations”, i.e. no selection by TOAR category is applied
rural_count	number of valid data (stations) included in each grid box at every output interval (month); “rural stations”, i.e. only stations with TOAR category = 1 are considered
rural_mean	mean value of metrics in grid box; “rural stations”
rural_median	median value of metrics in grid box; “rural stations”
urban_count	number of valid data (stations) included in each grid box at every output interval (month); “urban stations”, i.e. only stations with TOAR category = 3 are considered
urban_mean	mean value of metrics in grid box; “urban stations”
urban_median	median value of metrics in grid box; “urban stations”

6. Station metadata

All stations in the TOAR database carry extensive metadata information which has been obtained from the data submissions, extensive lookups in google maps etc., the python module geocoder, and from various global gridded datasets. The station metadata information that is contained in the metrics files (and in the stations_ files) is summarized in Table 8. Additional metadata information in the database primarily concerns QA measures. For example, the database stores the source of the altitude information that is eventually used as station_alt, documents the verification status of station coordinates, and allows commentary on a station or its coordinates.

Table 8: Explanation of station metadata variables contained in the metrics data files.

Variable name	Description
numid	the (internal) numeric id of the station
network_name	the network_name; in case of a merged series, two or more network_names may be combined with a '/'
station_id	the station identifier; in case of a merged series, two or more station_ids may be combined with a '/'
station_name	the station name; in case of a merged series the first name will be used
station_country	the country to which the station belongs, and in most cases where the station resides. Country names are standardized throughout the database, but they do not strictly follow the ISO naming. Be aware that, for example, stations in Antarctica are associated with different countries.
station_type	a characterization of the station type according to the EEA Airbase classification. Airbase distinguishes between "background", "industrial", "traffic", and "other". The TOAR database also has "unknown" as a potential value if no metadata information on station type was provided. Note that, although there are rules for designating a station type within Airbase, this information is somewhat ad-hoc and has to be used with caution.
station_type_of_area	a characterisation of the station's surrounding according to the EEA Airbase classification scheme. Airbase distinguishes between "rural", "suburban", and "urban". We added "remote" and "unknown" as additional values. The same caveat as for station_type also applies here.
station_lat	the latitude (in degrees_north) of the station location. Latitudes range from -90 to +90.
station_lon	the longitude (in degrees_east) of the station location. Longitudes range from -180 to +180.
station_alt	the best value of the station altitude in metres above sea level. By default this is the altitude reported by the data provider. However, we identified several hundred cases with wrong altitude information. In these cases station_alt is usually changed to the elevation obtained from google.

Table 8, continued

station_google_alt	the elevation at the location of the station as obtained from the google maps API.
station_etopo_relative_alt	a measure for the orographic variability around the station location. A third altitude value for the station has been obtained from the ETOPO1 dataset (in 0.1 deg resolution), and the minimum altitude within a 5 km radius around the station has been extracted as well. Station_etopo_relative_alt gives the difference between these two values. Stations with station_etopo_relative_alt > 500 can normally be regarded as “mountain stations”. This criterion is also used for the derivation of the TOAR category (see section 6).
station_population_density	population density extracted from the gridded CIESIN product (http://sedac.ciesin.columbia.edu/gpw) at about 5 km resolution at the equator. Values are of type integer and range from 0 to $\sim 2 \cdot 10^6$ people per km ² .
station_max_population_density_5km	the maximum population density in a 5 km radius around the station location
station_max_population_density_25km	the maximum population density in a 25 km radius around the station location
station_nightlight_1km	nighttime lights extracted from the NOAA DMSP product (http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html) at 1 km grid resolution. Values are integers from 0 (very dark) to 63 (very bright).
station_nightlight_5km	the average nighttime_light value in a 5 km radius around the station location.
station_max_nightlight_25km	the maximum nighttime_light value in a 25 km radius around the station location.
station_climatic_zone	integer value indicating the climatic zone of the station location according to http://eusoils.jrc.ec.europa.eu/projects/RenewableEnergy . The codes are: -1 unknown (no value assigned) 0 unclassified 1 Warm Temperate Moist 2 Warm Temperate Dry 3 Cool Temperate Moist 4 Cool Temperate Dry 5 Polar Moist 6 Polar Dry 7 Boreal Moist 8 Boreal Dry 9 Tropical Montane 10 Tropical Wet 11 Tropical Moist 12 Tropical Dry

Table 8, continued

station_dominant_landcover	integer value of the dominant landcover type at the station location according to the IGBP classification and derived from the MODIS MD12C1 dataset (https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd12c1). Codes are: -1 no value assigned 0 Water 1 Evergreen Needleleaf forest 2 Evergreen Broadleaf forest 3 Deciduous Needleleaf forest 4 Deciduous Broadleaf forest 5 Mixed forest 6 Closed shrublands 7 Open shrublands 8 Woody savannas 9 Savannas 10 Grasslands 11 Permanent wetlands 12 Croplands 13 Urban and built-up 14 Cropland/Natural vegetation mosaic 15 Snow and ice 16 Barren or sparsely vegetated 255 Fill Value/Unclassified
station_landcover_description	a text field describing the major landcover types in a 25 km radius around the station location
station_wheat_production	annual wheat production of the year 2000 extracted from http://gaez.fao.org/Main.html . The result is a float value in units of thousand tonnes per year. This keyword accepts a single value or a range (list) of two values specifying the minimum and maximum value for which metadata shall be extracted. If no value has been assigned for a station, the value of -999. is returned. The maximum value stored in the database is 22.03.
station_nox_emissions	annual NOx emissions of the year 2010 from EDGAR HTAP inventory V2 (gridded data in 0.1 degrees resolution and in units of grams of NO ₂ m ⁻² yr ⁻¹ obtained from http://edgar.jrc.ec.europa.eu/htap_v2/index.php?SECURE=123). Values range from 0 to ~1000.
station_omi_no2_column	5-year average (2011-2015) high-resolution NO ₂ column value from the OMI instrument in units of 10 ¹⁵ molec cm ⁻² . The product has been obtained from Chris McLinden at Environment Canada. Values are in the range of 0 to 20.80.
station_toar_category	an integer value describing a station as either: 0 unclassified 1 rural, low elevation 2 rural, high elevation or mountain 3 urban For details see section 7.

Table 8, continued

station_rice_production (only in "RICE" metrics)	annual rice production of the year 2000 extracted from http://gaez.fao.org/Main.html . The value range is from 0 to ~35.
station_wheat_production (only in "WHEAT" metrics)	annual wheat production of the year 2000 extracted from http://gaez.fao.org/Main.html . The value range is from 0 to ~25.

7. TOAR site classification

Through careful combination of the station metadata information derived from the high-resolution gridded products it has become possible for the first time to develop an objective classification scheme for ozone monitoring stations globally. The *station_toar_category* variable (in short TOAR category) distinguishes between “rural, low elevation”, “rural, high elevation or mountain”, and “urban” sites. Roughly one half of all stations in the database are characterized by one of these labels. For the other half, the categorization is not robust, therefore these stations are labeled as “unclassified”. The primary intention of the TOAR category is to provide robust station sets with certain characteristics. Table 8 lists the criteria that have been applied for the station classification. The resulting sets of classified stations have been carefully inspected and it was verified that important monitoring sites known to be typical of urban, rural or high elevation environments were accurately categorized.

Successful station classification relies on the availability of precise station coordinates. Great effort was applied to cleaning up wrongly reported station locations and ensuring that stations listed under different networks are located at the same position, there are still cases where the exact coordinates are not known. Usually, the error is less than about 1 km, such that the station classification should not be affected by these errors. In a few cases, however, the error may be much larger, and then the automated classification scheme may assign an incorrect station characterization. This issue is also relevant when we think ahead and wish to make use of much finer resolved geo information (e.g. Landsat). We will then have to find a way to estimate the potential error of the station location.

Figures 3-5 show maps of rural and urban stations in different world regions.

Table 9: Criteria for the characterization of stations according to the objective TOAR category

TOAR category name	Numerical value	Criteria ¹
rural, low elevation	1	omi_no2_column <= 8 AND nightlight_5km <= 25 AND population_density <= 3000 AND max_population_density_5km <= 30000 AND google_alt <= 1500 AND etopo_relative_alt <= 500
rural, high elevation or mountain	2	omi_no2_column <= 8 AND nightlight_5km <= 25 AND population_density <= 3000 AND (google_alt > 1500 OR (google_alt > 800 AND etopo_relative_alt > 500))
urban	3	population_density > 15000 AND nightlight_1km >= 60 AND max_nightlight_25km == 63
unclassified	0	all other stations

¹ the prefix “station_” is omitted from the variable names for readability

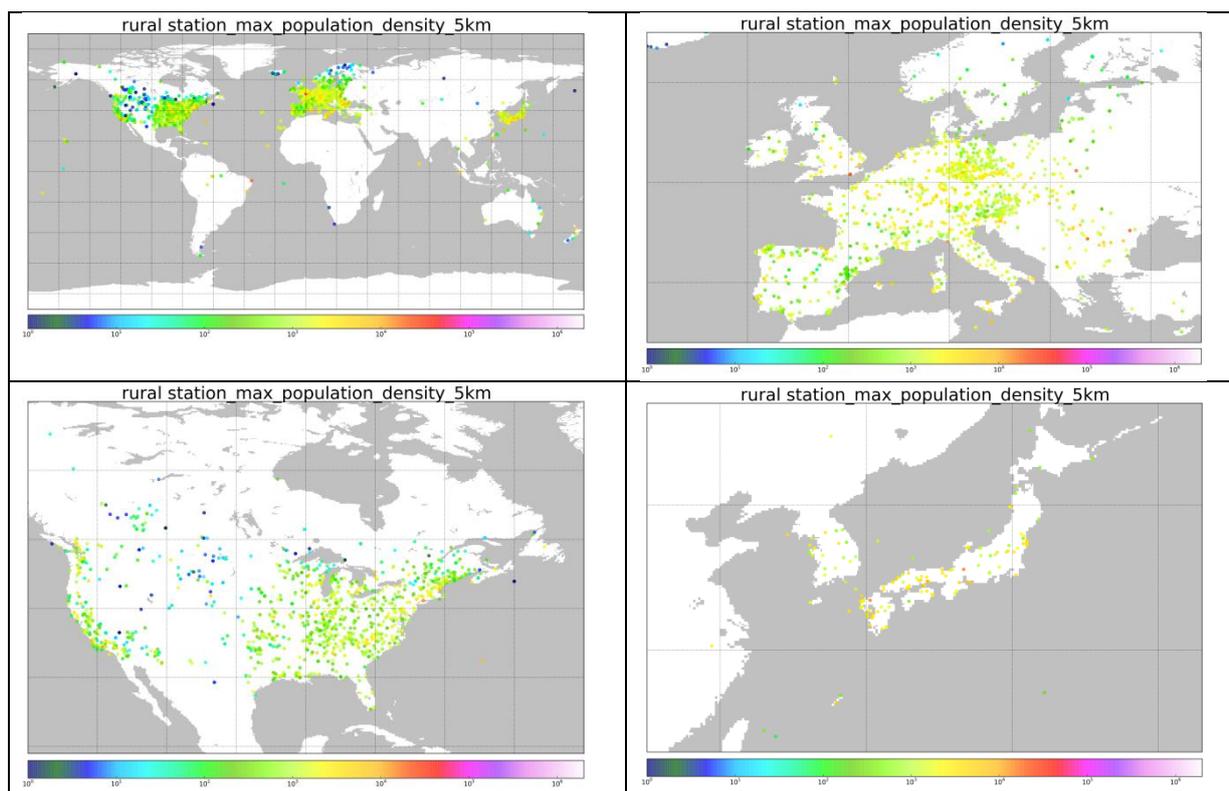


Figure 3: Location of stations classified as „rural, low elevation“ (TOAR category = 1). The symbol color shows the maximum population density in a 5 km radius around the station location.

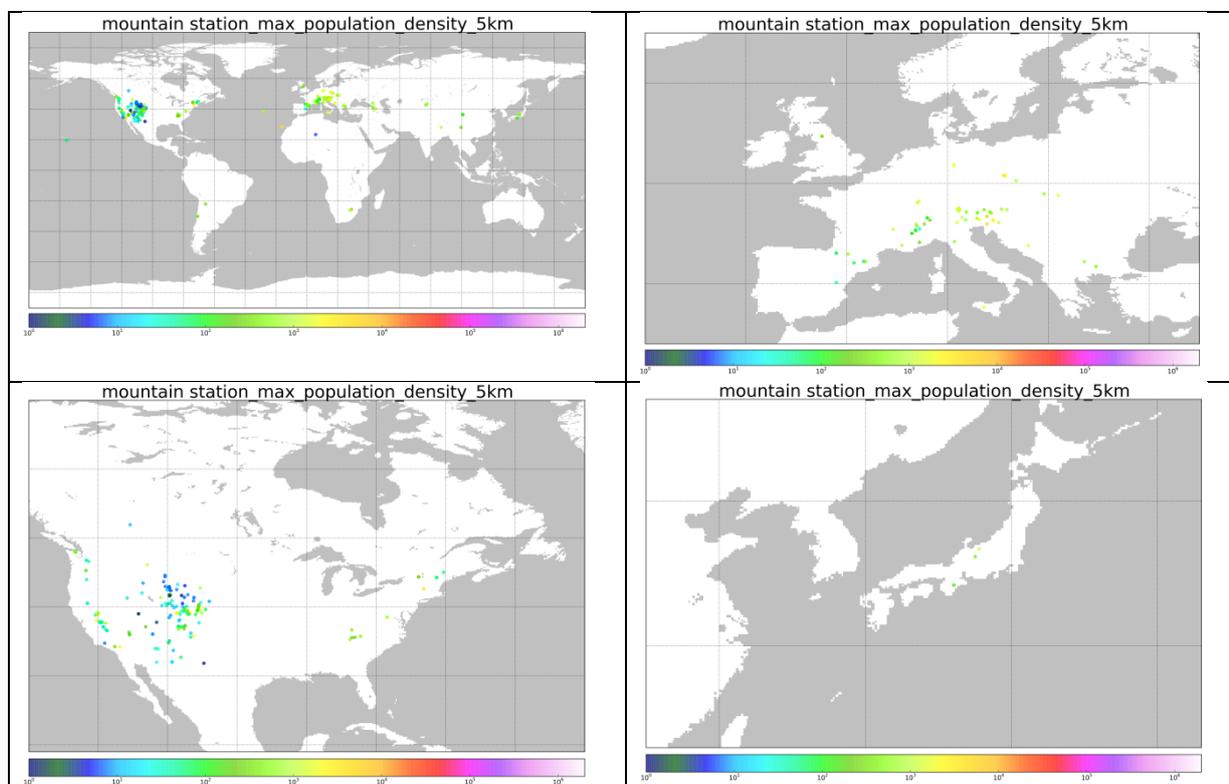


Figure 4: Location of stations classified as „rural, high elevation or mountain“ (TOAR category = 2). The symbol color shows the maximum population density in a 5 km radius around the station location.

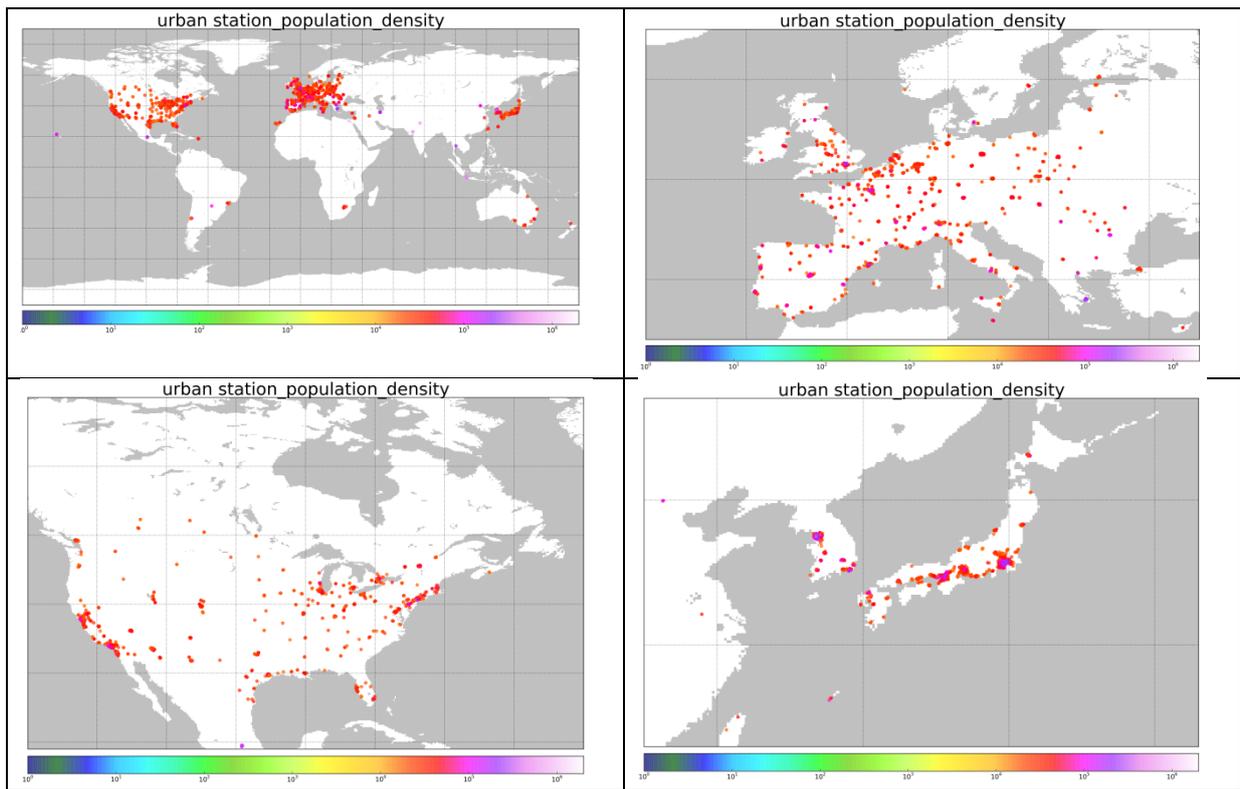


Figure 5: Location of stations classified as „urban“ (TOAR category = 3). The symbol color shows the population density at the station location.

8. The data series merging procedure

Some ozone series consist of two or more partial series, and in other cases, stations with long records were relocated within a few hundred meters so that the combined series could be regarded as continuous, but the parts are stored individually in the database. Yet, in other cases, a station reported data to different networks (in Europe you can find data from the same station reported up to four times, namely as UBA dataset, EMEP dataset, Airbase dataset, and finally as GAW dataset). For example, ozone data from the station “Westerland, Sylt” are available under station_id “DEUB001” (UBA and Airbase), “DE0001R” (EMEP), and “WES654N00” (GAW). Each of these series has different start and end dates:

AIRBASE: 1984-03-01 - 2012-12-31

EMEP: 1984-02-29 - 2013-12-31

UBA: 1990-01-01 - 2016-06-30

GAW: 1989-12-31 - 2013-12-31

We visually inspected these data series and identified the most suitable series for use during the different periods. The results of this visual inspection are coded in a `toar_dataset_groups.csv` table, which contains the necessary information how the data series shall be merged in a machine-readable form. For the example of Westerland, you will find the following two lines in this table:

```
[16601; 25758; 21935; 23065],19,DE0001R,EMEP,25758,29.02.1984 23:00,31.12.1989 23:00
```

```
[16601; 25758; 21935; 23065],19,DEUB001,UBA,16601,01.01.1990 01:00,03.03.2016 06:00
```

The first column contains a list of data series ids (unambiguous numerical identifiers of a data series) to which the following merge statement applies. Here, we have the ids for the Westerland series from UBA, EMEP, GAW, and Airbase (in this order). The next column is a “statement number” – all lines with the same statement number belong together. Then we see the station_id and network name of the data series that shall be used during one period of the time series. This is purely for human readability and not evaluated by the merging code. Finally, we see the data series id that shall be used (corresponding to the station_id and network_name listed before), and the date range in which this series shall be used.

In this example, the result of any query asking for data from Westerland will always return EMEP data during the time period from February, 29th, 1984 to December, 31st, 1989, and UBA data beginning with January 1st, 1990. In consequence, the data from either GAW or AIRBASE are not included in the TOAR metrics files, and there is only one ozone record from Westerland, Sylt.

Two other, prominent examples of merging are the GAW time series from Cape Grim, Australia, and Gosan, South Korea. Both stations have two ozone series associated with them which are consecutive in time. In the case of Cape Grim this denotes a change in the calibration scale applied, whereas in the case of Gosan the data provider changed in 2012. Without merging, these stations would not appear in the long-term trend dataset (Cape Grim) or in the present day dataset (Gosan), respectively.

As mentioned above, the merging table has been constructed by hand, so please be aware that there may be errors. We will appreciate any feedback on series that are merged while they shouldn't be or

series that are not merged although they should be. Below we present a few examples to highlight the potential difficulties and raise your awareness on issues that can be associated with the merging process (or issues arising from not-merging).

Example 1: GAW data from Cape Point, South Africa

The GAW station Cape Point in South Africa provides two ozone data series to the GAW data archive: one series containing all measurements, and one with “filtered” data where local or regional pollution influences from the nearby city of Capetown are removed. For details on the filtering procedures, see Brunke et al. (2004). Figure 6 displays the two time series, which show distinct differences. Clearly, for an analysis of “background” ozone trends, the filtered data would give the more appropriate results. In the context of TOAR, however, which focuses on metrics for impact assessments, the polluted episodes are of course relevant. Therefore, the unfiltered data from Cape Point has been included in the TOAR metrics files.

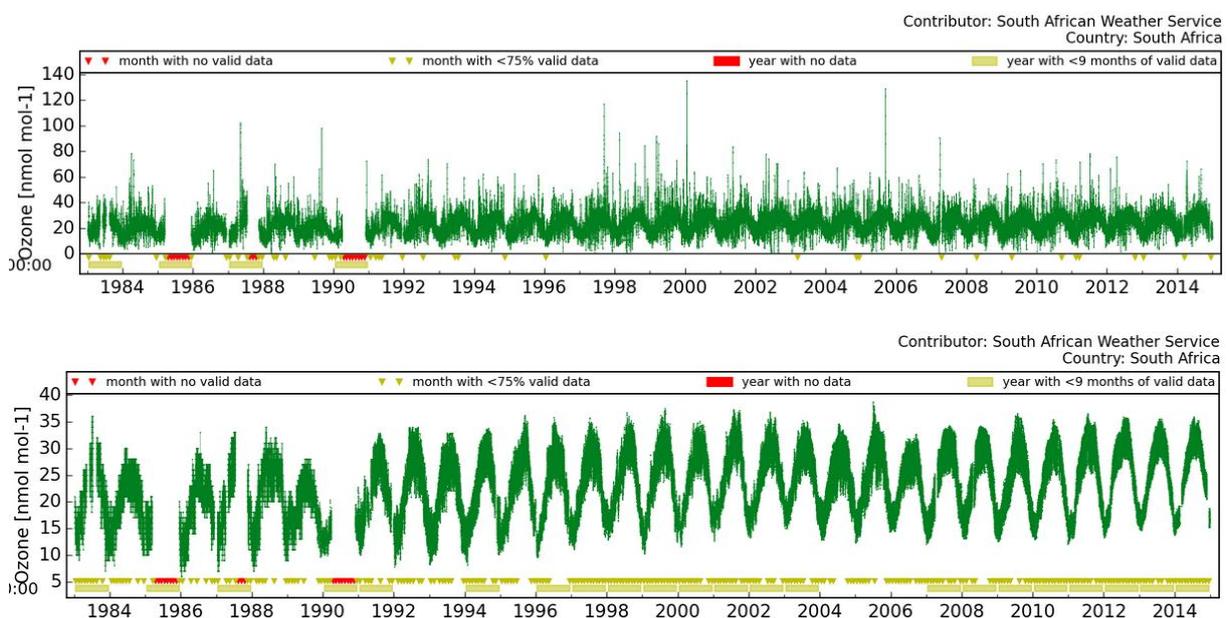


Figure 6: Time series of ozone data from Cape Point, South Africa showing the unfiltered data in the top row and the filtered series (prepared by the station operators) on the bottom. For TOAR analyses, the unfiltered data are used.

Example 2: Paris, Eiffel tower

Figure 7 shows a snapshot from the [JOIN web interface](#) with three markers of stations on, at, or near the Eiffel tower in Paris. All three “stations” are associated with ozone series from Airbase. The station names are “Tour Eiffel 1er étage”, “Tour Eiffel 3e etage”, and “PARIS 7eme”. All three series commence on January, 1st, 1999. The first and third series end in January 2003, while the second series (3e etage) ends in December 2011. Figure 8 displays monthly averages of the three series during the common period. The ground-level station (PARIS 7eme) shows a lot of similarities with the “1er etage” site, but the “3e etage” exhibits significantly higher values, because it is farther away from the local pollution sources. In order to best support impact assessments from TOAR data, we chose “PARIS 7eme” as the data series for this location and ignore the other two series in the metrics data files.

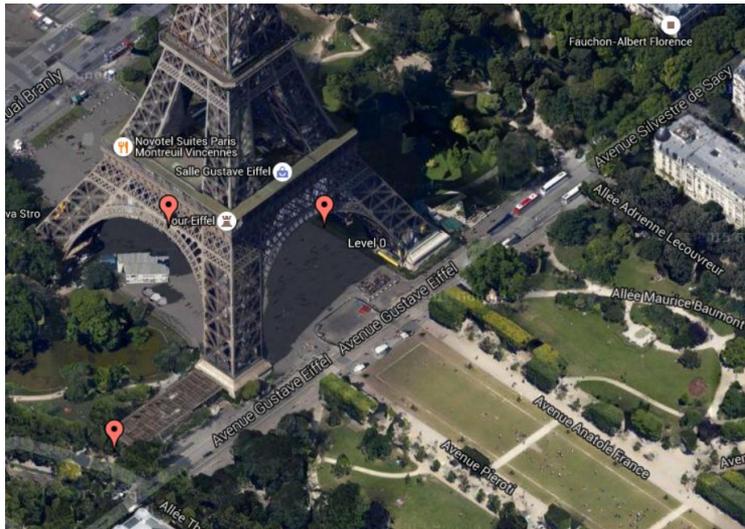


Figure 7: Snapshot from the JOIN web interface showing the markers for the three ozone monitoring stations at or on the Eiffel tower in Paris, France.

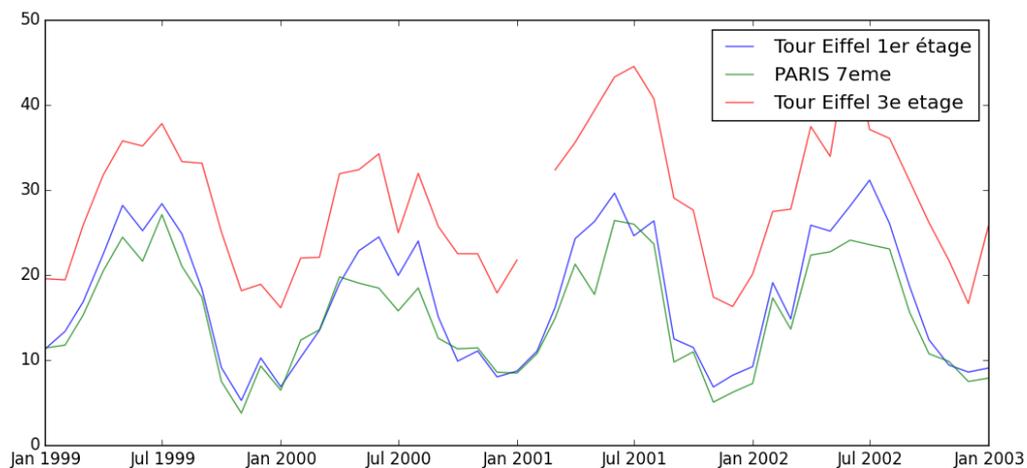


Figure 8: Monthly averages of ozone data from the three “stations” shown in Figure 7 during the common measurement period.

Example 3: Central London

Figure 9 shows the location of three sites in central London (“LONDON BRIDGE PLACE”, “LONDON WESTMINSTER”, “CENTRAL LONDON”) which were operated one after each other. All three sites show similar characteristics with respect to the station metadata (Table 10), and their data series appear to fit together well (Figure 10). Therefore, by merging these three series from three distinct sites, we obtain one data series for central London which is suitable for long-term trend analysis. Without merging, the only metrics data files that would include any data from central London would be the “decadal change” sets, and the “maximum coverage” sets.

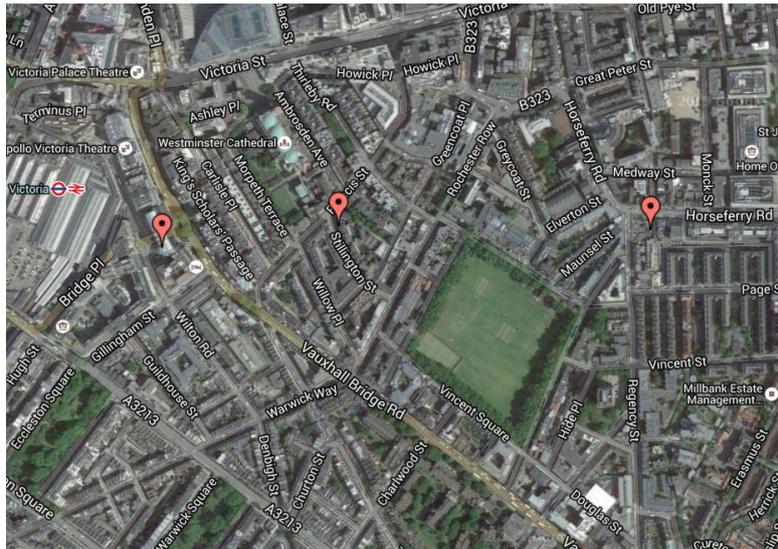


Figure 9: Locations of three stations in central London which have ozone data series from different years

Table 10: Station metadata excerpts for the three central London stations shown in Figure 9

	LONDON BRIDGE PLACE	LONDON WESTMINSTER	CENTRAL LONDON
station_lat	51.494446	51.49467	51.494724
station_lon	-0.141953	-0.131931	-0.138336
station_alt	20	5	20
station_reported_alt	20	5	20
station_google_alt	14.5	13.0	11.9
google_resolution	152.7	152.7	152.7
station_etopo_relative_alt	13	11	11
station_dominant_landcover	13	13	13
station_omi_no2_column	10.629	10.629	10.629
station_nightlight_1km	63	63	63
station_nightlight_5km	63	63	63
station_population_density	123144	123144	123144
station_max_population_density_5km	147838	147838	147838
station_nox_emissions	48.7	48.7	48.7
station_toar_category	3	3	3

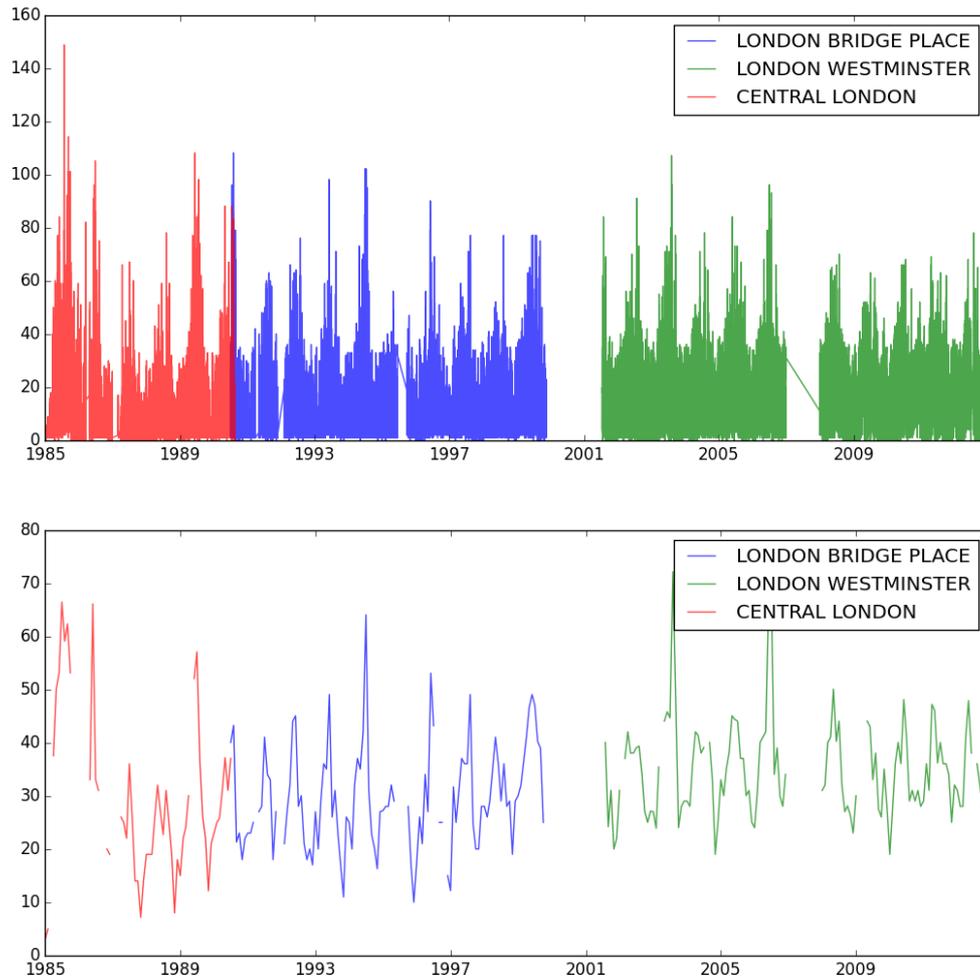


Figure 10: Time series of hourly values (top) and 95th-percentiles (bottom) of the three stations shown in Figure 9.

Example 4: Pireaus, Greece

Airbase lists two sites in Pireaus, Greece: “PIREAUS-1” (data from 1988 to 2012), and “PIREAUS-2” (data from 2000 to 2007). Again, the two stations are rather closely located to each other and share similar characteristics. Since the PIREAUS-1 series exhibits major data gaps in 2000 and 2001, it might seem a good idea to merge the two data series in order to obtain one, more complete record. However, as shown in Figure 11, the consistency between these two series is not very high. Therefore, we refrain from merging them together. These two sites are thus listed individually in the TOAR metrics datasets. Due to the period of measurements at PIREAUS-2, this station will not play a role in the TOAR analyses; in fact it does not appear in any of the metrics files. Note, however, that in other cases of a similar nature, there may be two independent data series in the metrics file which could be more different than they are supposed to be.

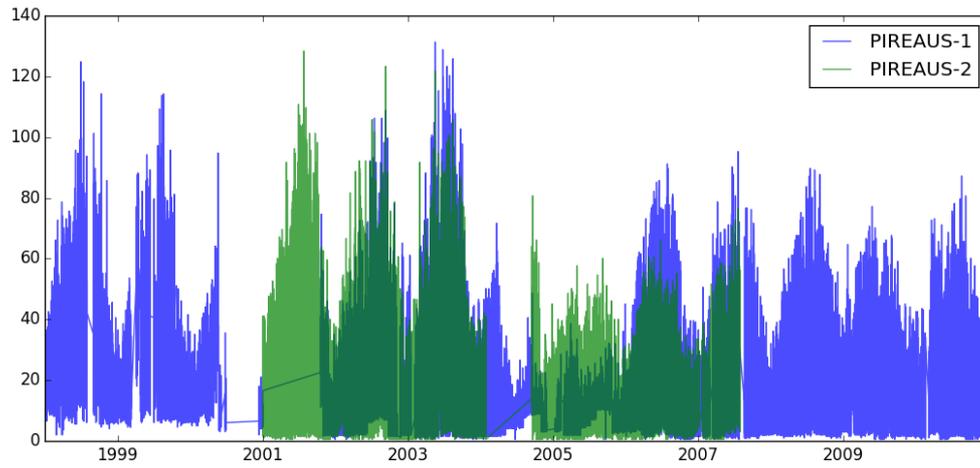


Figure 11: Hourly time series from the two ozone data series at PIREAUS-1, and PIREAUS-2 in Greece

9. Plot gallery

In this section we present examples of the different plot types which are made available on the TOAR data portal.

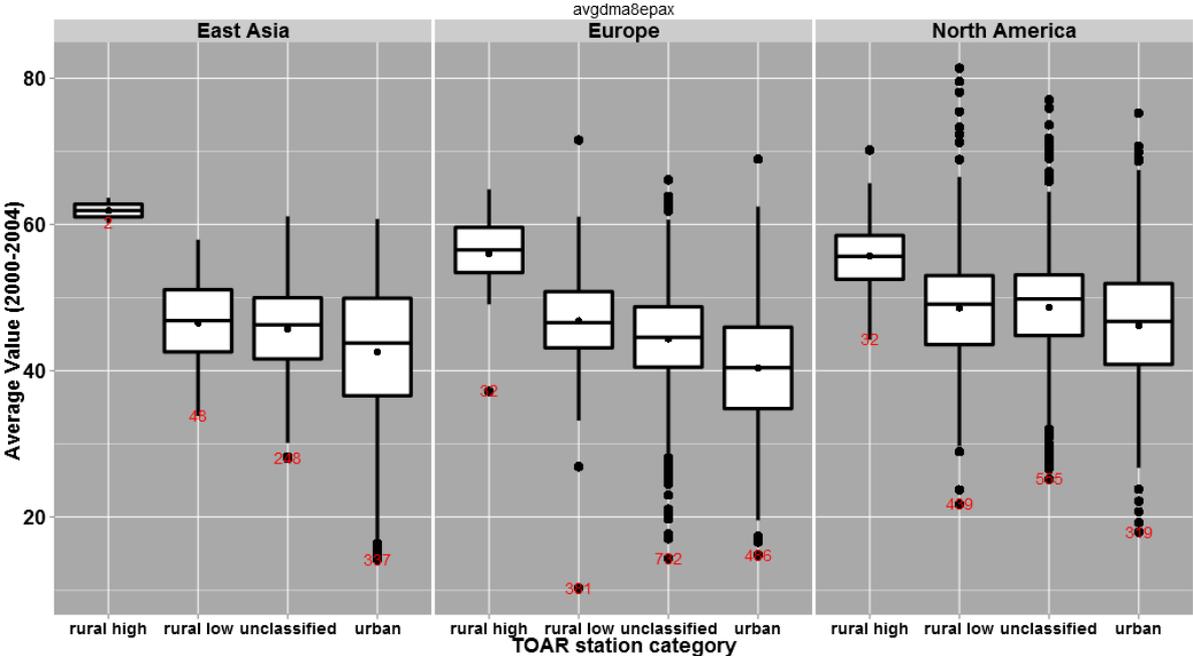


Figure 12: Example of a “present-day” box and whisker plot. Such plots are available from http://hs.pangaea.de/Projects/TOAR/Graphical_products/box-whisker-plots_present-day_1995-2014.zip

AOT40, wheat-growing season Data extracted on: 2016-10-19

AOT40 ozone, production threshold = 0.1 ktonne, 2010-2014 (minimum 3 years): all sites

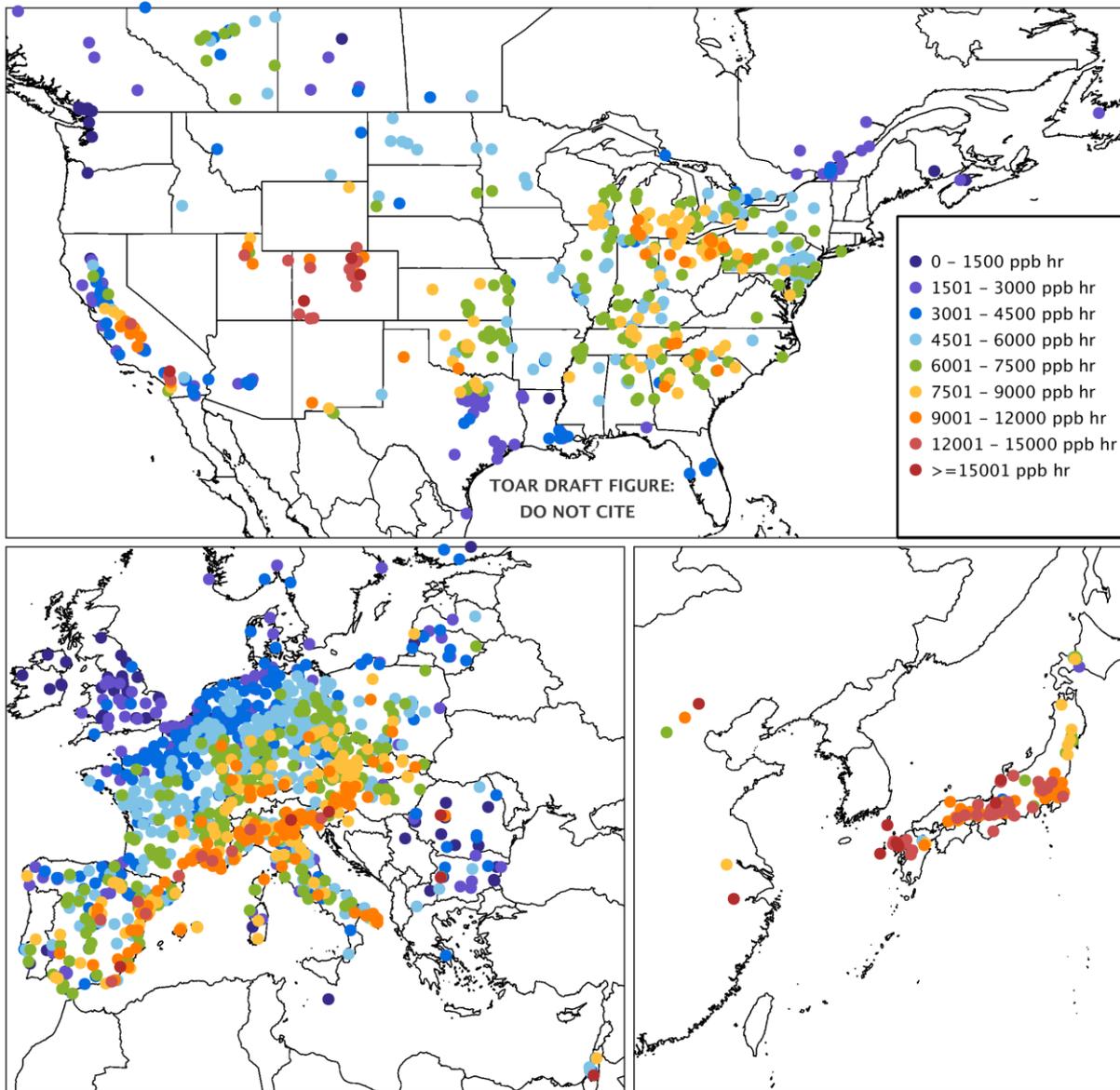


Figure 13: Example of a “present-day” map plot. Such plots are available from http://hs.pangaea.de/Projects/TOAR/Graphical_products/maps_present-day_2010-2014.zip

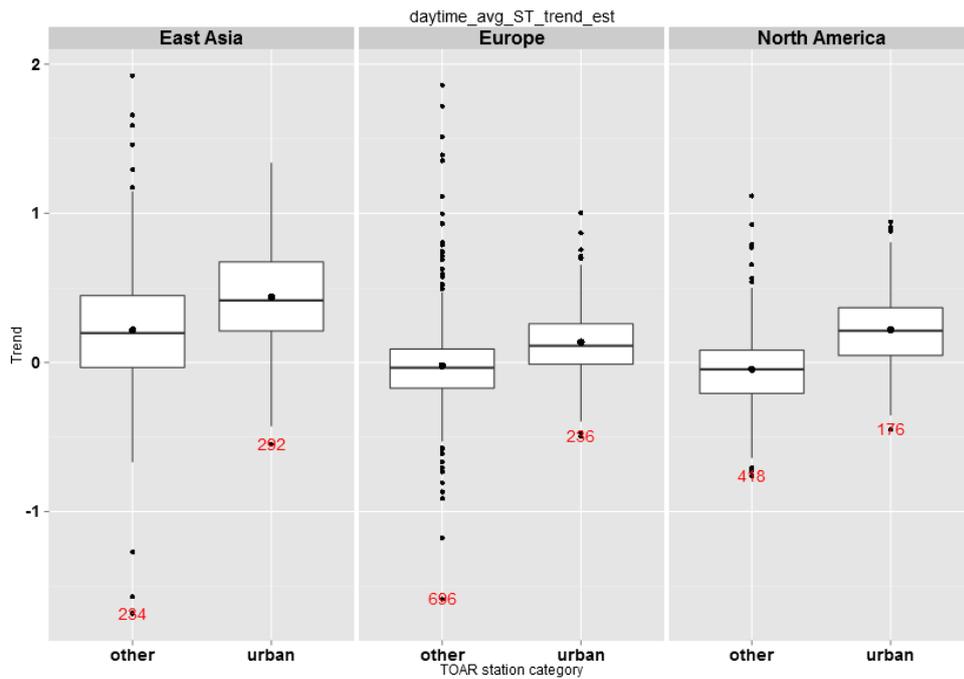


Figure 14: Example of a “trends” box and whisker plot. Such plots are available from http://hs.pangaea.de/Projects/TOAR/Graphical_products/box-whisker-plots_trends_1970-2014.zip

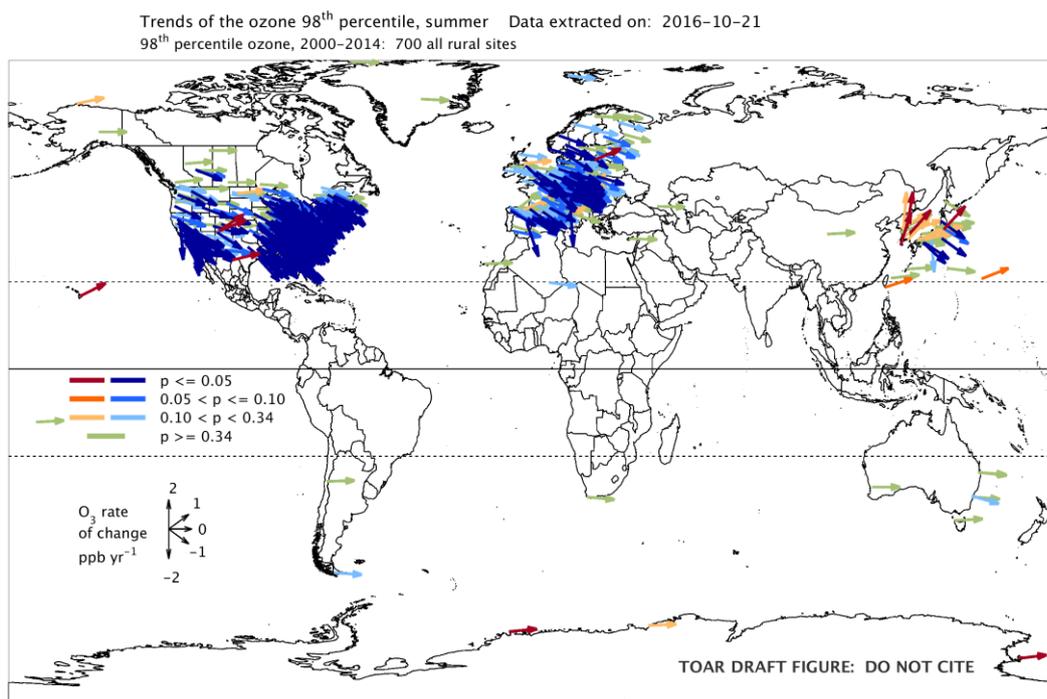


Figure 15: Example of a “trends” vector map. Such plots are available from http://hs.pangaea.de/Projects/TOAR/Graphical_products/maps_trends_1970-2014.zip

DMA8(EPAx), urban stations, July (1995-1999)

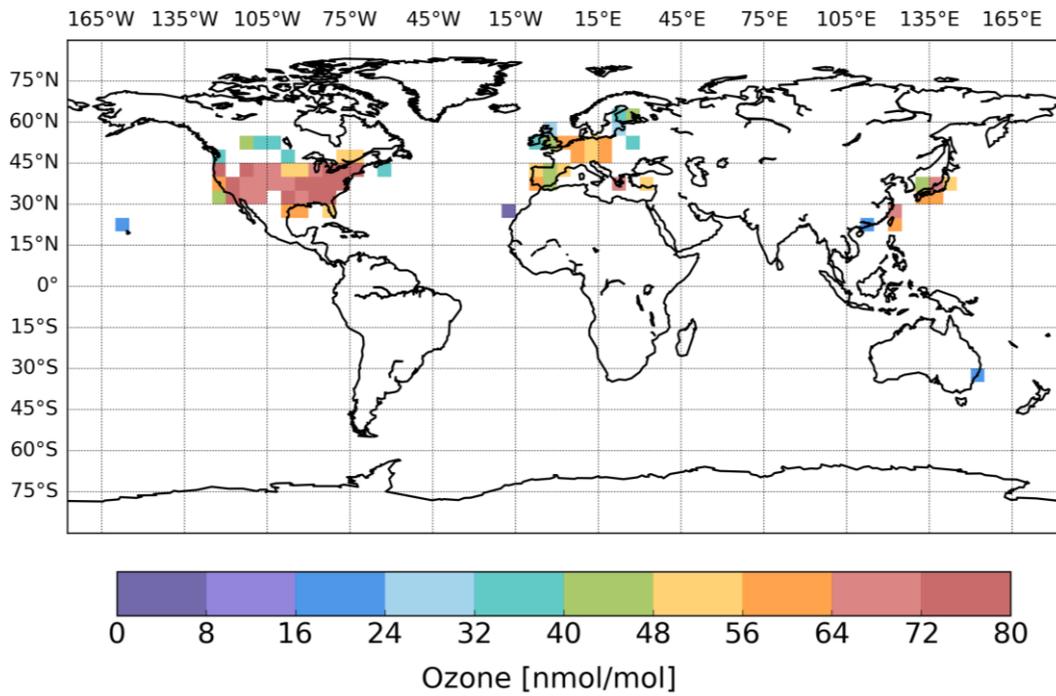


Figure 16: Example of a gridded map plot. Such plots are available from http://hs.pangaea.de/Projects/TOAR/Graphical_products/maps_gridded_1990-2014.zip