Standardized, FAIR and CF-compliant publication of urban climate model data

K. Heinke Schlünzen[UHH], David Grawe[UHH], Vivien Voss[UHH]
Angelika Heil[DKRZ]
Anette Ganske[TIB]
Jan Kretzschmar[ULei]

AtMoDat

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

User Workshop @ EMS 2021
Tue, 07 Sep, 11:00–12:30 (CEST)
https://meetingorganizer.copernicus.org/EMS2021/session/41771#

TIB
LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UNIVERSITÄT LEIPZIG

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

On Publication of Urban Climate Model Results

AtMoDat

- Status
  - Data rarely published
  - Data not findable
  - Need for published data increasing (researchers, planners, …)
- Project aim: increase reusability of urban climate data
- Workshop objectives
  - Explain standards (FAIR, CF, NetCDF, ATMODAT)
  - Learn to use a software to check if your data fulfil a standard
  - Jointly determine model output variables that need to be standardized
  - Next steps

Why do we need data publication standards?

AtMoDat

More and more data are being published,

but

often they are not reusable because they are:

- not adequately described,
- stored in file formats that cannot be read and processed with open software,
- not findable by search engines.

What hinders a standardised data publication?

AtMoDat

- Many data producers do not know how to correctly standardise their data for publication.
- Only a few data repositories support data producers by advising them and/or by controlling the standardisation of submitted data.
- There are few incentives to standardise data.

What are the key principles of a standardised data publication?

AtMoDat

The FAIR principles

- most widely adopted guiding principles for scientific data management and stewardship (Wilkinson et al. 2016*).
- aim at improving the **F**indable, **A**ccessible, **I**nteroperable, **R**eusable of digital assets.
- can be applied within all research disciplines.
- put specific emphasisis on enhancing machine actionability, but also target improving human readibility.

* https://doi.org10.1038/sdata.2016.18

FAIR data principles

AtMoDat

Findable

- (Meta)Data have a Persistent Identifier (PID)
- (Meta)data are stored in an open repository
- Data are detailed described through metadata

Accessible

- (Meta)data can be downloaded in a standardised way
- Metadata can still be accessed even if the data was deleted

Interoperable

- Standardised wording - controlled vocabulary
- References to other (meta)data

Reusable

- Specify licence
- Adhere to community standards

Persistent Identifiers (PIDs) for Data: DOI

AtMoDat

Digital object identifier (DOI): PID used to identify objects uniquely (standardised by ISO)

DataCite: global non-profit organisation that provides Digital Object Identifier (DOI) for research data

DataCite DOI:

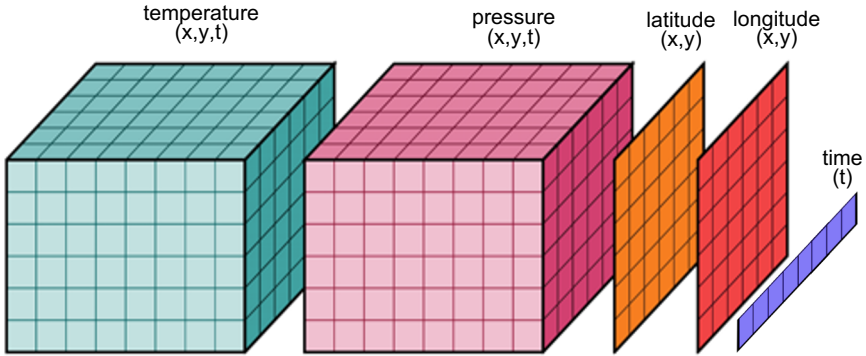
- Persistent and unique identifier for research data
- Resolves directly to a landing page which displays the metadata and download instructions
- Easy to cite, e.g. https://doi.org/10.1594/WDCC/CMAQ_CCLM_HZG_2008
- Enables machine readability (link and metadata)

Climate and Forecast (CF) Metadata Conventions

AtMoDat

- CF most widely used *data standard* for earth science data (first released in 2003).
- CF targets interoperability & reusability of information stored in *netCDF* data files*
- CF-netCDF: CF-compliant netCDF file using CVs & other self-describing metadata

A NetCDF file has dimensions, variables, and attributes.



Metadata for

– Variables

– Dimensions

– Global

Attributes

*network Common Data Form: machine-independent binary data formats for array-oriented science data.

CF Standard names

CF Standard names
– are a tool for a FAIR description of variables in NetCDF files.
– are *unique text strings* constructed using *controlled vocabulary*
– have a *precise definition* and an *associated SI unit*.
– are the value for the standard_name variable attribute, e.g.

Standard name. ≠ netCDF variable name

```
float co(time,lat, lon) ;
        co:standard_name = "mole_concentration_of_carbon_monoxide_in_air”
                    co:units  = "mol m-3”
```

The set of *permissible standard names* is contained in the *CF standard name table*, which currently encompasses 4500 entries*

* https://cfconventions.org/Data/cf-standard-names/77/build/cf-standard-name-table.html

CF Standard names

AtMoDat

Search

https://cfconventions.org/Data/cf-standard-names/77/build/cf-standard-name-table.html

tendency_of_atmosphere_mass_content_of_par | Search Standard Names | Show All Standard Names

AND OR (separate search terms with spaces)

Also search help text

Found 1 standard names matching query: tendency_of_atmosphere_mass_content_of_particulate_organic_matter_dry_aerosol_particles_expressed_as_carbon_due_to_emission_from_savanna_and_grassland_fires

View by Category

Atmospheric Chemistry	Atmosphere Dynamics	Carbon Cycle	Cloud	Hydrology
Ocean Dynamics	Radiation	Sea Ice	Surface	

Standard Name	Canonical Units	AMIP
tendency_of_atmosphere_mass_content_of_particulate_organic_matter_dry_aerosol_particles_expressed_as_carbon_due_to_emission_from_savanna_and_grassland_fires *alias:* tendency_of_atmosphere_mass_content_of_particulate_organic_matter_dry_aerosol_expressed_as_carbon_due_to_emission_from_savanna_and_grassland_fires "tendency_of_X" means derivative of X with respect to time. "Content" indicates a quantity per unit area. The "atmosphere content" of a quantity refers to the vertical integral from the surface to the top of the atmosphere. For the content between specified levels in the atmosphere, standard names including "content_of_atmosphere_layer" are used. The phrase "expressed_as" is used in the construction A_expressed_as_B, where B is a chemical constituent of A. It means that the quantity indicated by the standard name is calculated solely with respect to the B contained in A, neglecting all other chemical constituents of A. "Aerosol" means the system of suspended liquid or solid particles in air (except cloud droplets) and their carrier gas, the air itself. Aerosol takes up ambient water (a process known as hygroscopic growth) depending on the relative humidity and the composition of the aerosol. "Dry aerosol particles" means aerosol particles without any water uptake. "Primary particulate organic matter " means all organic matter emitted directly to the atmosphere as particles except elemental carbon. The sum of primary_particulate_organic_matter_dry_aerosol and secondary_particulate_organic_matter_dry_aerosol is particulate_organic_matter_dry_aerosol. The specification of a physical process by the phrase "due_to_" process means that the quantity named is a single term in a sum of terms which together compose the general quantity named by omitting the phrase. "Emission" means emission from a primary source located anywhere within the atmosphere, including at the lower boundary (i.e. the surface of the earth). "Emission" is a process entirely distinct from "re-emission" which is used in some standard names. The "savanna and grassland fires" sector comprises the burning (natural and human-induced) of living or dead vegetation in non-forested areas. It excludes field burning of agricultural residues. "Savanna and grassland fires" is the term used in standard names to describe a collection of emission sources. A variable which has this value for the standard_name attribute should be accompanied by a comment attribute which lists the source categories and provides a reference to the categorization scheme, for example, "IPCC (Intergovernmental Panel on Climate Change) source category 5 as defined in the 2006 IPCC guidelines for national greenhouse gas inventories".	kg m-2 s-1	

CF Standard names

AtMoDat

- New CF standard names can be proposed to the CF community.
- The CF community publicly discusses proposals in terms of (a) consistency with the CF rules and (b) relevance.
- AtMoDat project has recently proposed standard names for airborne pollen concentrations

Proposed CF Standard Name for Pollen

AtMoDat

float pollen_conc(time,lev,lat,lon,taxon) ;
pollen_conc:standard_name = "number_concentration_of_biological_taxon_pollen_grains_in_air“ ;
pollen_conc:units = „m-3“ ;
pollen_conc:coordinates = "taxon_lsid taxon_name“ ;
pollen_conc:long_name = „airborne pollen concentration“ ;
char taxon_name(taxon,string80) ;
taxon_name:standard_name = "biological_taxon_name“ ;
taxon_name:long_name = „pollen (Latin name)" ;
char taxon_lsid(taxon,string80) ;
taxon_lsid:standard_name = "biological_taxon_lsid“ ;
taxon_lsid:long_name = „ITIS identifier“ ;
taxon_lsid:url = „https://www.itis.gov/“ ;
char pollen_common_name(taxon,string80) ;
pollen_common_name:long_name = „pollen (common name)“ ;
pollen_common_name:description = „Common names as listed in ITIS“ ;
pollen_common_name:url = „https://www.itis.gov/“ ;

dimensions:
string80 = 80 ;
taxon = 6 ;

data:
time = 6., 12., ... ;
lat = 1., 2., ... ;
lon = 5., 6., ... ;
pollen_conc = 0.0087, 0.28367, ... ;
taxon_name = "Betula L.", "Poaceae", "Artemisia L.", "Ambrosia L.", "Secale L.", "Alnus Mill.";
taxon_lsid ="urn:lsid:itis.gov:itis_tsn:19478", "urn:lsid:itis.gov:itis_tsn:40351", "urn:lsid:itis.gov:itis_tsn:35431", "urn:lsid:itis.gov:itis_tsn:36495",
"urn:lsid:itis.gov:itis_tsn:42089", "urn:lsid:itis.gov:itis_tsn:19466", "urn:lsid:itis.gov:itis_tsn:32928", "urn:lsid:itis.gov:itis_tsn:19505",
"urn:lsid:itis.gov:itis_tsn:32989", "urn:lsid:itis.gov:itis_tsn:19461", "urn:lsid:itis.gov:itis_tsn:19276“ ;
pollen_common_name="birch", "grasses", "sagebrush", "ragweed", "rye", "alder" ;

The ATMODAT standard (Ganske et al., 2021*)

AtMoDat

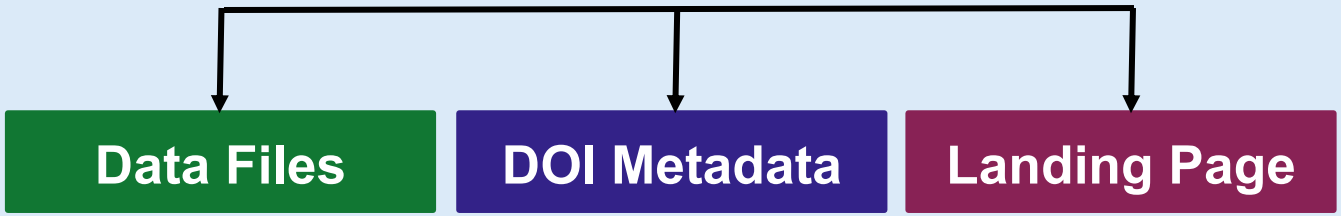
- **quality guideline** for a **FAIR publication** of atmospheric model data with **open licences**.
- **guides data producers** and **data curators**.
- **specifies requirements** for **data and metadata**.
- contains **checklists** allowing a quick and easy verification if the (meta)data are compliant with the ATMODAT standard.

* ATMODAT Standard (v3.0)
https://doi.org/10.35095/WDCC/atmodat_standard_en_v3_0

ATMODAT standard: key elements

- assumes a **data publication with a DataCite DOI**.
- defines **NetCDF** as **data format**.
- defines **adherence** to the **Climate and Forecast (CF) conventions**.
- defines **mandatory, recommended** and **optional metadata.**

ATMODAT standardisation



Requirements for Data Files

AtMoDat

Always NetCDF

Coordinate system

CF Conventions/ Controlled Vocabulary

AtMoDat is currently working on an extension of the CF Conventions

Coordinate reference system

```
netcdf CD24_base_2008_dec_1_1915785082846561030 {

dimensions:
        ...
variables:
    float gas_so2(time, z, y, x) ;
        gas_so2:coordinates = "lon lat" ;
        gas_so2:grid_mapping = "Lambert_Conformal" ;
        gas_so2:missing_value = -9.e+33f ;
        gas_so2:standard_name = "mass_concentration_of_sulfur_dioxide_in_air" ;
        gas_so2:units = "kg m-3" ;
        gas_so2:long_name = "SO2 concentration" ;
    ...
//global attributes:
        :Conventions = "CF-1.6" ;
        :institution = "Helmholtz-Zentrum Geesthacht, ....." ;
        :source = "model: CMAQ v5.0.1 cb05tump ae5; ....." ;
        :summary = "Standard CMAQ Model run over Northwestern Europe […] " ;
        :title = "Concentrations of gaseous pollutants and particulate compounds
                  over Northwestern Europe […] in 2008";
        :creation_date = "2015-04-02" ;
        :crs = "spherical earth, R = 6370 km" ;
        :history = "... abbreviated ..." ;
```

Requirements for DOI Metadata

AtMoDat

PIDs (ORCID, ROR) for persons, organisations, funders

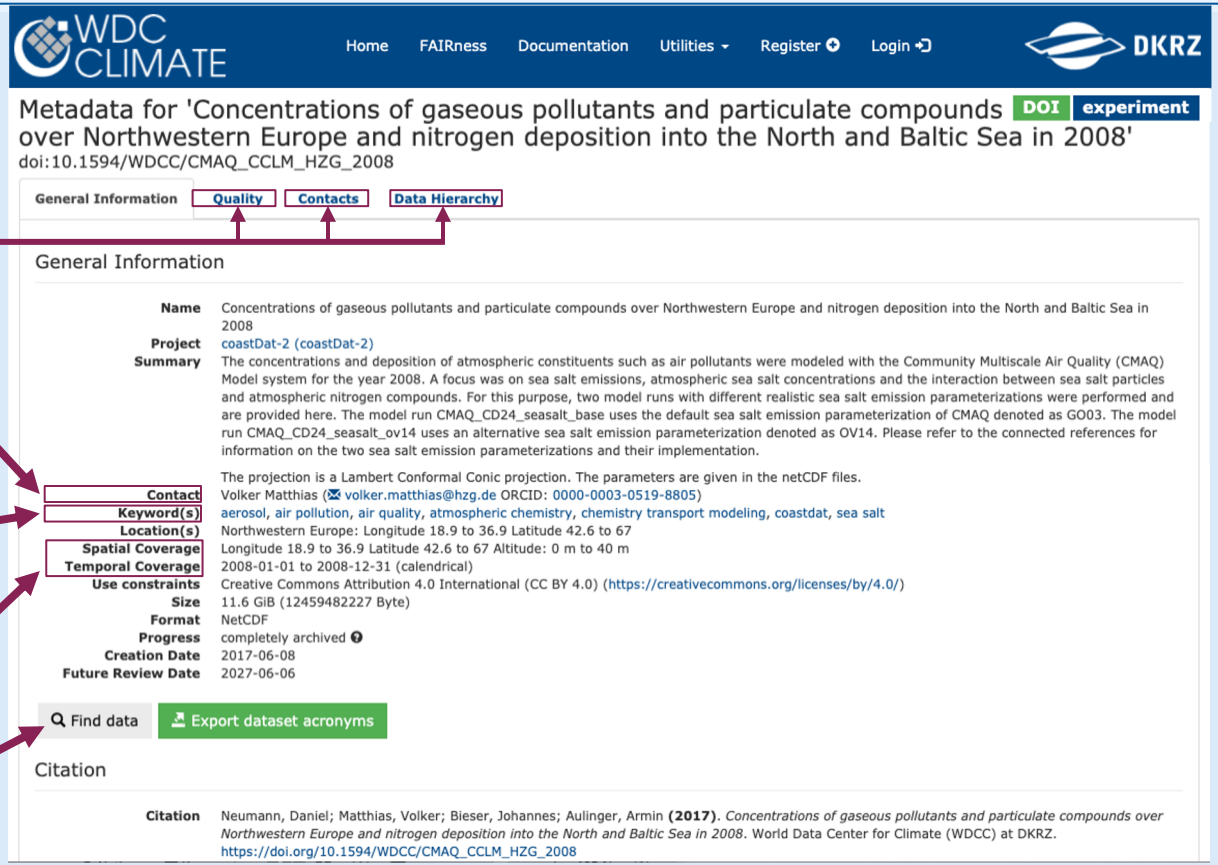
Standardized dates

"id":"10.1594/wdcc/cmaq_cclm_hzg_2008",
"doi":"10.1594/wdcc/cmaq_cclm_hzg_2008",
…...
"creators":[{
"name":"Neumann, Daniel",
“nameType":"Personal",
"nameIdentifiers":{
"nameIdentifier":"https://orcid.org/0000-0001-8574-9093",
"nameIdentifierScheme":"ORCID",
"schemeUri":"https://orcid.org"},
"affiliation":{
"name":"Leibniz -Institut fuer Ostseeforschung Warnemuende (IOW)",
"affiliationIdentifier":"https://ror.org/03xh9nq73",
"affiliationIdentifierScheme":"ROR",
"SchemeURI":"https://ROR.org",
….
"dates": "2017-06-08",
…

Landing Page (human-readable): requirements

AtMoDat

- **Sub-pages with details on datasets or variables**
- **Contact person with ORCID**
- **Use of Controlled Vocabulary**
- **Spatial and temporal coverage of the data**
- **Download access to the data**



Landing Page (machine-readable): requirements

AtMoDat

Machine interpretable language

PIDs for all persons, organisations, funders

```
{"@context":"http://schema.org",
 "type":"Dataset",
 "provider":{... ,"@id": ...},
"@id":"https://doi.org/10.26050/WDCC/...",
 "name":"...",
 "temporalCoverage":"...“,
 "spatialCoverage ":"...“,
 ...,
"author":[{"@type":"Person",
           "name":"Neumann, D. ,
           "@id":
         }]}
```

ATMODAT Checklists:

AtMoDat

Tables with **summary specifications** for

- ❑ **DataCite metadata**
- ❑ **landing page**
- ❑ **data files**

Quick overview of required specifications for data producers and curators.

Table 14: Requirements for the Data Files

Requirements	**Status**
The file format is netCDF.	M
The value of the Conventions global attribute includes the version number of the used CF convention in the form "CF-Ver".	M
The value of the Conventions global attribute includes the version number of the used ATMODAT Standard in the form "ATMODAT-Ver".	R
comment	O
contact	R
Conventions	M
creation_date	R
creator	R
crs (coordinate reference system)	R
featureType	S
frequency	R
further_info_url	O
geospatial_lat_resolution	R
geospatial_lon_resolution	R
geospatial_vertical_resolution	R
history	R

M=Mandatory, R=Recommended, O=Optional

ATMODAT Checklists: Detailed Specifications

AtMoDat

Table 2: List of all DOI metadata properties (Table for curators)

DataCite ID	Property	ATMODAT Status	Example	Description
1	Identifier (with mandatory type sub-property)	M	https://doi.org/10.1594/wdcc/cmaq_cclm_hzg_2008	the DOI itself
2	Creator (with optional family name, given name, name identifier and affiliation sub-properties)	M	Neumann, Daniel, ..., https://orcid.org/0000-0001-8574-9093	It is strongly recommended to use ORCID for persons and ROR for affiliation, see Appendix G.
3	Title (with optional type sub-properties)	M	Concentrations of gaseous pollutants and particulate compounds over Northwestern Europe and nitrogen deposition into the North and Baltic Sea in 2008	Dataset title
4	Publisher	M	World Data Center for Climate (WDCC) at DKRZ	The name of the entity that holds archives, publishes prints, distributes, releases, issues, or produces the resource.
5	PublicationYear	M	2017	Year of publication
6	Subject	M	EASYDAB, ATMODAT, meteorology and atmospheric sciences, atmosphere	Always use several keywords, which must at least include: EASYDAB, ATMODAT, the field of science and the realm of the model, which must be taken from controlled vocabularies (CVs). More than one realm is possible.
		R	atmospheric chemistry, climate,....	It is strongly recommended to add further keywords, which also should be taken from CVs, if applicable.
	Subject scheme sub-properties	R	for "atmospheric chemistry" vocabulary= GEMET, https://www.eionet.europa.eu/gemet/en/concept/623	Name and URI of the controlled vocabulary

Mandatory (M)

Recommended (R)

All reasonable recommended (R) metadata should be entered.

atmodat checker

https://github.com/AtMoDat/atmodat_data_checker

AtMoDat

- Data standardisation steps prior their publication:
  Step 1) Make data files compliant with discipline-specific standard.
  Step 2) Use a checker to control that data files are correctly standardised
- How are CMIP data standardised?
  – CMOR tool for standardising data
  – Control with PrePARE Checker and CF-checker
  – *Poblem: tools lacks flexibility to be used for datasets outside CMIP when other standardisation requirements*

→ For data that shall comply with the ATMODAT Standard, we developed the AtMoDat Standard Compliance Checker (checks global attributes + integrated the CF-checker)

atmodat checker

https://github.com/AtMoDat/atmodat_data_checker

AtMoDat

ATMODAT Standard Compliance Checker

This is a python library that contains checks to ensure compliance with the ATMODAT Standard.

Its core functionality is based on the IOOS compliance checker. The ATMODAT Standard Compliance Checker library makes use of cc-yaml, which provides a plugin for the IOOS compliance checker that generates check suites from YAML descriptions. Furthermore, the Compliance Check Library is used as the basis to define generic, reusable compliance checks. This repository is an extension of this library as it holds specific checks to ensure compliance with the ATMODAT Standard.

In addition, the compliance to the CF Conventions 1.4 or higher is verified with the CF checker.

We set up a binder where you can try out the functionalities of the ATMODAT Standard Compliance Checker:

launch binder

Installation (tested on Linux and macOS)

1. Clone this repository

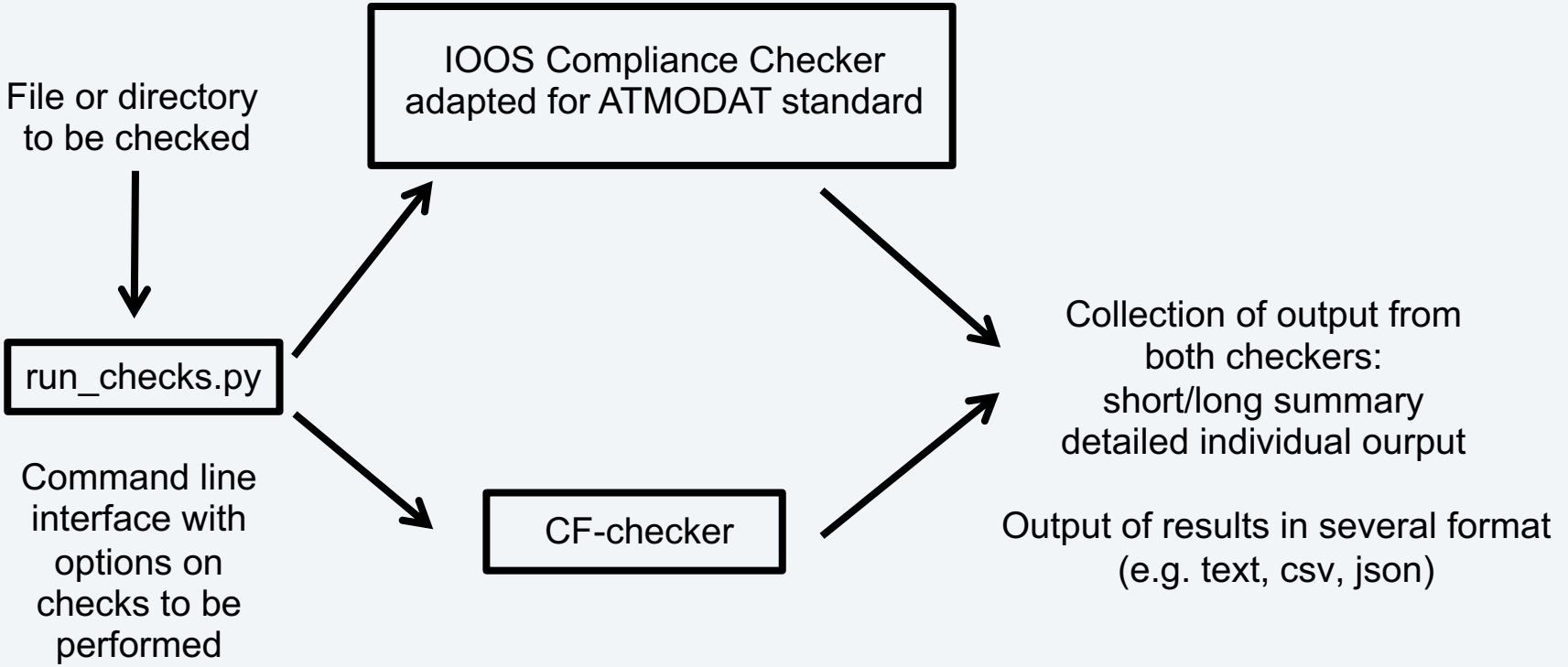
```
git clone https://github.com/AtMoDat/atmodat_data_checker.git
```

Languages

- Python 65.8%
- Jupyter Notebook 34.1%
- Shell 0.1%

atmodat checker

https://github.com/AtMoDat/atmodat_data_checker



atmodat checker

https://github.com/AtMoDat/atmodat_data_checker

AtMoDat

Checks are defined in a yaml-file which users could adjust to defined their own checks

```yaml
---
suite_name: "atmodat_standard:3.0"

checks:

  ######################################## Mandatory checks ########################################
  # Check global attributes

  - check_id: "institution_attribute_type_check"
    parameters: {"attribute": "institution", "type": "str", "status": "mandatory"}
    check_name: "atmodat_checklib.register.GlobalAttrTypeCheck"

  - check_id: "source_attribute_type_check"
    parameters: {"attribute": "source", "type": "str", "status": "mandatory"}
    check_name: "atmodat_checklib.register.GlobalAttrTypeCheck"

  # Check if Conventions version is within given range
  - check_id: "cf_conventions_version_check"
    parameters: {"attribute": "Conventions", "convention_type": "CF", "min_version": 1.4, "max_version": 1.8,
                 "status": "mandatory"}
    check_name: "atmodat_checklib.register.ConventionsVersionCheck"

  # Check if AtMoDat version matches the version against which checks should be performed
  - check_id: "atmodat_conventions_version_check"
    parameters: {"attribute": "Conventions", "convention_type": "ATMODAT", "min_version": 3.0, "max_version": 3.0,
                 "status": "mandatory"}
```

atmodat checker

https://github.com/AtMoDat/atmodat_data_checker

AtMoDat

Example short summary.txt: run_checks.py –s –f testfile.nc

```
Short summary of checks:

Checking against: atmodat_standard:3.0, CF table version: 77
Version of the AtMoDat checker: 1.1.0
Checked at: 2021-08-11T14:54:17.517485

Number of checked files: 1
Total checks passed: 4/31
Mandatory checks passed: 2/4
Recommended checks passed: 2/18
Optional checks passed: 0/9
CF checker errors: 1
```

atmodat checker

https://github.com/AtMoDat/atmodat_data_checker

AtMoDat

- If errors are reported with regard to CF conformity or global attributes, attributes need to modified
- Relatively simple to define new checks and new check suites for different applications in future
- Easy to install; accessible via github, but plans to provide packages via PyPi/Anaconda
- We will provide simple python scripts that can be used to fill global/variable attributes in netCDF files from a csv table (“atmodat attribute filler”, release in near future)

atmodat checker

https://github.com/AtMoDat/atmodat_data_checker

AtMoDat

- Let’s try it out

https://hub-binder.mybinder.ovh/user/atmodat-atmodat_data_checker-2o024vmi/tree/notebooks

→ see link posted in the chat

Open Issues

AtMoDat

Issue	many variables relevant in urban climate have no CF standard names
Examples	derived variables building variables
Solution	add names to CF standard

Derived variables

AtMoDat

Derived variables

with high spatial variability

Variable	Standard name (suggested)	Unit
PT	perceived_temperature	degree_C
UTCI	universal_thermal_climate_index	degree_C
TMRT	mean_radiant_temperature	K
PET	physiological_equivalent_temperature	degree_C

Building variables

AtMoDat

Variable	Standard name (suggested)	Unit
Building mask	volume_fraction_of_obstacles_in_air	1

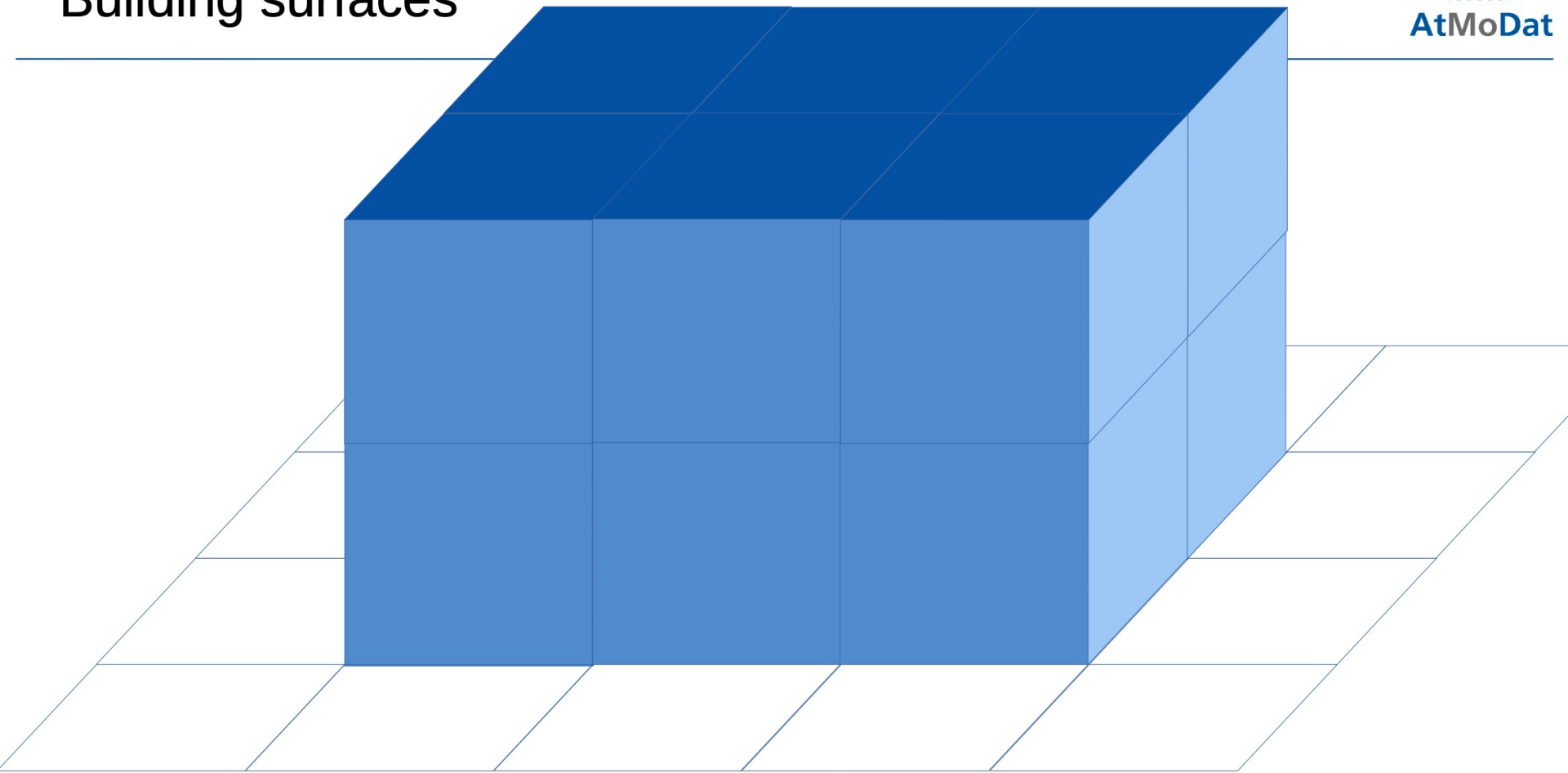
Surface variables

building cells are geometrically complex, e.g.
surface temperature needs to be stored for each surface

six geometric surfaces are possible, up to three occur per cell

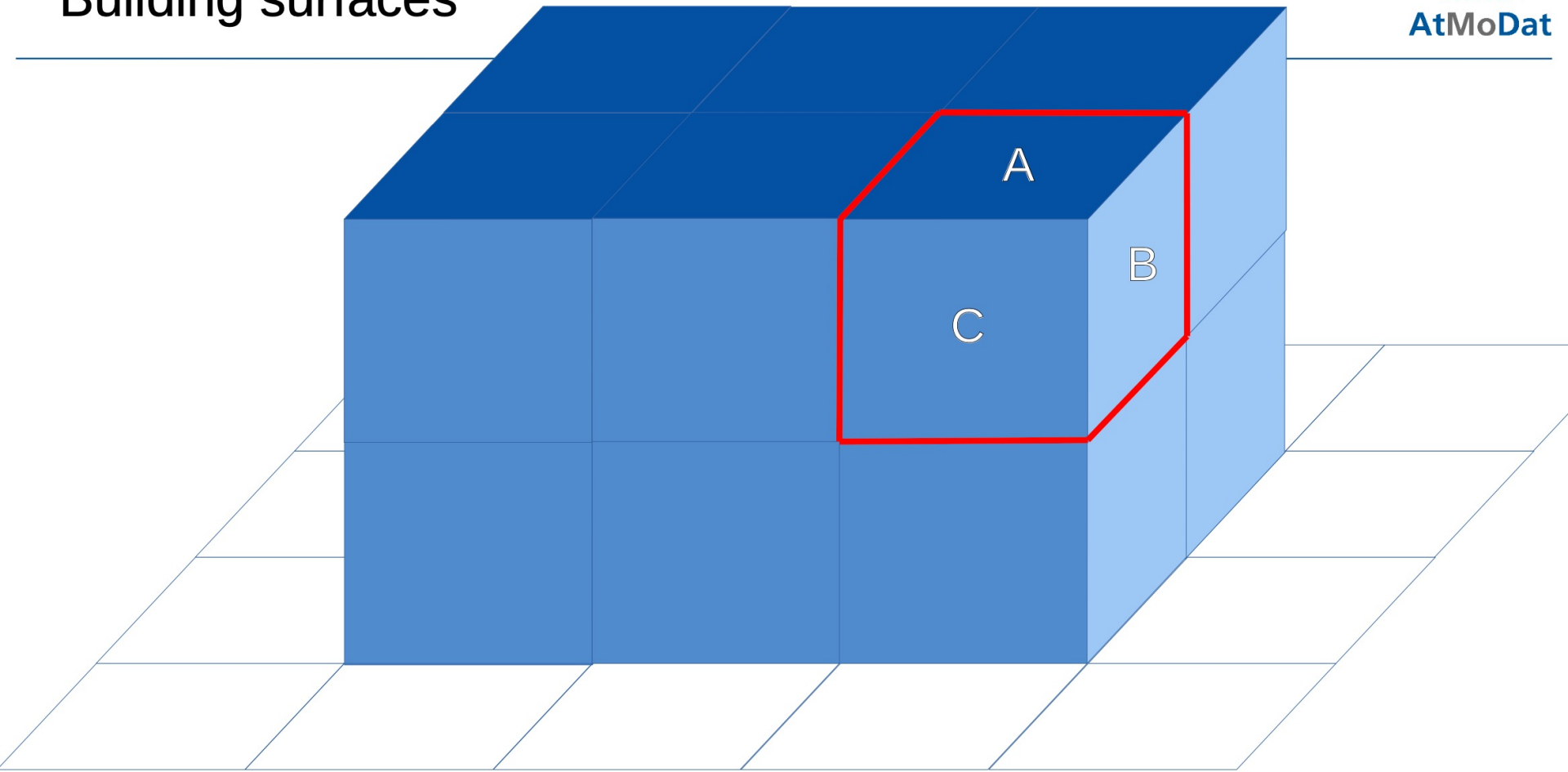
Building surfaces

AtMoDat

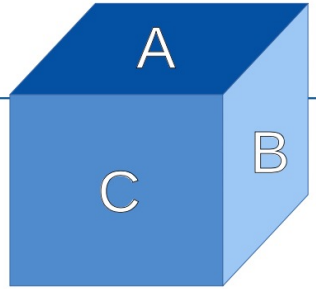


Building surfaces

AtMoDat



Building surfaces



	Standard name (suggested)	Unit
A	net_shortwave_flux_at_obstacle_top	W m-2
B	net_shortwave_flux_at_x-positive_surface	W m-2
C	net_shortwave_flux_at_y-negative_surface	W m-2
A	rainfall_rate_at_obstacle_top	m s-1

	Standard name (suggested)	Cell methods	Unit
A	net_shortwave_flux	at_obstacle_top	W m-2
B	net_shortwave_flux	at_x-positive_surface	W m-2
C	net_shortwave_flux	at_y-negative_surface	W m-2
A	rainfall_rate	at_obstacle_top	m s-1

Sparsity of data

AtMoDat

Building data is sparse

because buildings are usually attached to the ground while the top of the domain may be several building height above

store building data as 3d data or via index field?

compromise: 3d field up to a certain height

Conclusions and next steps

AtMoDat

- Data publication is not so difficult
- Software helps to check for fulfilling standards (AtMoDat checker)
- More standard names need to be defined with more urban modelers publishing their data
- For more information
  - https://www.atmodat.de/
  - Next workshop where you can **come with your own data**
    09. Nov. 2021
    Register at https://indico.dkrz.de/event/14/