

Developing a Scientific Data Search Engine

Part 2: Features & Results

*Sabira Duishebaeva¹,
Teagan Johnson¹, Nathan Hook²,
Saquib Aziz-Khan², Eric Nienhouse²*



SIParCS Intern¹ - NCAR Mentor²

July 27, 2022



Table of Contents

- Introduction to **3 new features** implemented
- Background for Metadata Validation
- Results & Future work for Metadata Validation
- Background for Sitemap/JSON -LD feature
- Results & Future work for JSON-LD feature
- Background for Faceting feature
- Results & Future work for Faceting

BACKGROUND

Metadata

Validation



Feature#1: Metadata Validation

Why need?

- Improves searchability and findability
- Keeps data consistent & reliable

DSET Rules:

Concepts	Definition
- Minimum Required	
- Author	The person(s)/institution(s) receiving credit, as in a citation.
- Title	A name given to the data set, model, software or other asset.
- ISO Asset Type	Type of asset.
- Landing page	Web accessible landing page.
- Publication Date	Date asset was first made available.
- Metadata date	Date stamp when metadata record created or last updated
- Publisher	The lab (or smaller group) that made asset available.
- Resource Support Contact	Person, group, or institution to contact for support on asset.
- Metadata point of Contact	Party responsible for the metadata
- Description	A summary of data set content, or description of asset.

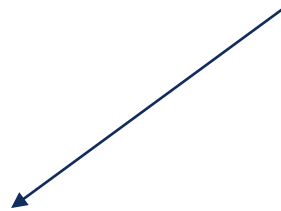
Metadata Validation: How Does It Work?



Data Providers



Record their data

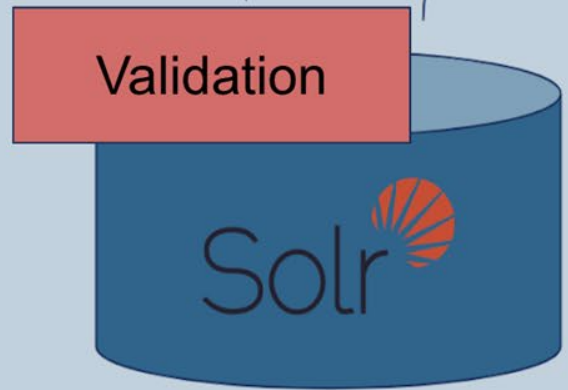
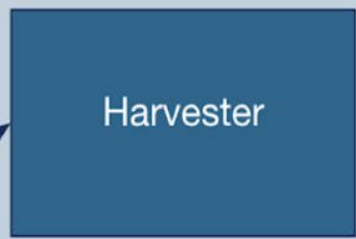
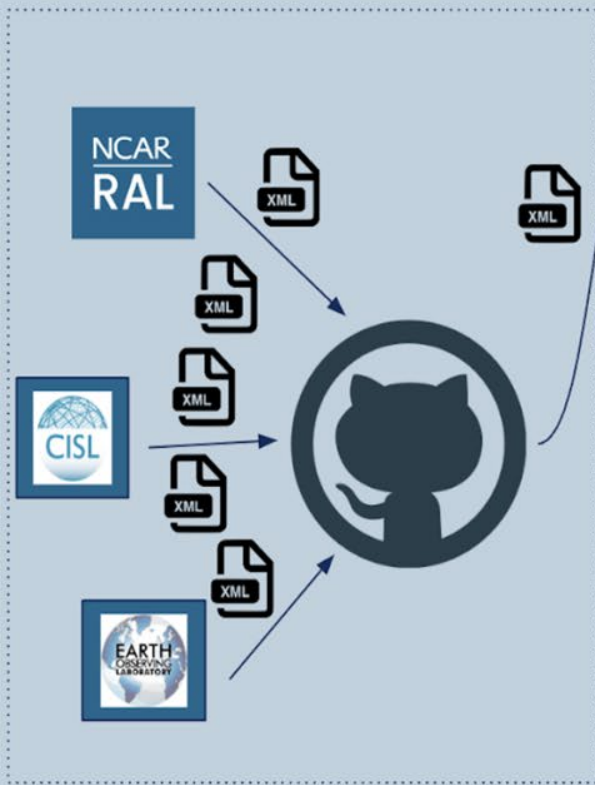


Put them in xml files



XML Files in GitHub

```
1 <?xml version="1.0" encoding="UTF-8" ?>
2 <gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:gco="http://www.isotc211.org/2005/gco" xmlns:gml="http://www.opengis.net/gml" xmlns:xsi="htt
3 <gmd:fileIdentifier>
4   <gco:CharacterString>edu.ucar.rda::ds010.0</gco:CharacterString>
5 </gmd:fileIdentifier>
6 <gmd:language>
7   <gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/" codeListValue="eng; USA">eng; USA</gmd:LanguageCode>
8 </gmd:language>
9 <gmd:characterSet>
10   <gmd:MD_CharacterSetCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.xml#MD_CharacterSetCode" codeListValue="utf8">utf8</gmd:MD_C
11 </gmd:characterSet>
12 <gmd:hierarchyLevel>
13   <gmd:MD_ScopeCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.xml#MD_ScopeCode" codeListValue="dataset">dataset</gmd:MD_ScopeCode
14 </gmd:hierarchyLevel>
15 <gmd:contact>
16   <gmd:CI_ResponsibleParty>
17     <gmd:organisationName>
18       <gco:CharacterString>NCAR Research Data Archive</gco:CharacterString>
19     </gmd:organisationName>
20     <gmd:contactInfo>
21       <gmd:CI_Contact>
22         <gmd:phone>
23           <gmd:CI_Telephone>
24             <gmd:facsimile>
25               <gco:CharacterString>303-497-1291</gco:CharacterString>
26             </gmd:facsimile>
27           </gmd:CI_Telephone>
28         </gmd:phone>
29         <gmd:address>
30           <gmd:CI_Address>
31             <gmd:deliveryPoint>
32               <gco:CharacterString>National Center for Atmospheric Research</gco:CharacterString>
33             </gmd:deliveryPoint>
34             <gmd:deliveryPoint>
35               <gco:CharacterString>CISL/DSS</gco:CharacterString>
36             </gmd:deliveryPoint>
37             <gmd:deliveryPoint>
```



NCAR | NATIONAL CENTER FOR UCAR | ATMOSPHERIC RESEARCH

NCAR is sponsored by National Science Foundation

Welcome to Metadata Search Engine!

This is a search engine to access digital assets from all NCAR/UCAR.

XML Files in Solr

```
{
  "id": "edu.ucar.rda:ds061.0",
  "title": "NMC 47 by 51 Northern Hemisphere Stratospheric Analyses, daily 1964-1980",
  "description": "This dataset of daily gridded analyzed geopotential height and temperature",
  "doi": "https://doi.org/10.5065/Y7MH-0127",
  "keywords": ["EARTH SCIENCE > ATMOSPHERE > ALTITUDE > GEOPOTENTIAL HEIGHT"],
  "resource_type": "dataset",
  "authoritative_source_url": "https://doi.org/10.5065/Y7MH-0127",
  "authoritative_source_location_on_disk": "/Users/sduishebaeva/Java/xml/test-pull-method/t",
  "authoritative_source_md5": "fa845f9e23cacf2923e89416c11de82f",
  "github_xml_url": "https://github.com/dsabira/test-pull-method.git/blob/main/test11.xml",
  "is_valid": true,
  "index_timestamp": "2022-07-15T17:11:34.690Z",
  "_version_": 1738439492569661440},
{
  "id": "10.5065/9n3z-7x72",
  "description": "The PyConform package is a Python-based package for converting model time-
  "doi": "https://doi.org/10.5065/9n3z-7x72",
  "keywords": ["Software"],
  "resource_type": "dataset",
  "authors": ["Paul, Kevin",
    "Mickelson, Sheri",
    "Dennis, John"],
  "author_emails": ["",
    "",
    ""],
  "authoritative_source_url": "https://doi.org/10.5065/9n3z-7x72",
  "authoritative_source_location_on_disk": "/Users/sduishebaeva/Java/xml/test-pull-method/t",
  "authoritative_source_md5": "8e5492f804fe8744a516c8a71ba64cbb",
  "github_xml_url": "https://github.com/dsabira/test-pull-method.git/blob/main/test9.xml",
  "is_valid": false,
  "validation_messages": ["Title must not be empty"],
  "index_timestamp": "2022-07-15T17:11:36.370Z",
  "_version_": 1738439494331269120},
{
```


Results & Future Work Validation



Results & Future Work for Metadata Validation

Your metadata is invalid for these reasons:

- Title must not be empty.

Id:

edu.ucar.opensky::articles:17812

Description:

The authors present a new method to diagnose the middle-atmosphere climate sensitivity by extending the climate feedback–response analysis method (CFRAM) for the coupled atmosphere–surface system to the middle atmosphere. The middle-atmosphere CFRAM (MCFRAM) is built on the atmospheric energy equation per unit mass with radiative heating and cooling rates as its major thermal energy sources. MCFRAM preserves CFRAM’s unique feature of additivity, such that partial temperature changes due to variations in external forcing and feedback processes can be added to give a total temperature change for direct comparison with the observed temperature change. In addition, MCFRAM establishes a physical relationship of radiative damping between the energy perturbations associated with various feedback processes and temperature perturbations associated with thermal responses. In this study, MCFRAM is applied to both observations and model output fields to diagnose the middle-atmosphere climate sensitivity. The authors found that the largest component of the middle-atmosphere temperature response to the 11-yr solar cycle (solar maximum vs solar minimum) is the partial temperature change due to the variation of the solar flux. Increasing CO₂ cools the middle atmosphere, whereas the partial temperature change due to changes in O₃ can be either positive or negative. The application of MCFRAM to model dynamical fields reconfirms the advantage of introducing the residual circulation to characterize middle-atmosphere dynamics in terms of the partial temperature changes. The radiatively driven globally averaged partial temperature change is approximately equal to the observed temperature change, ranging from -0.5 K near 25 km to -1.0 K near 70 km between solar maximum and solar minimum.

DOI:

<http://n2t.net/ark:/85065/d7sf2xmr>

Results:

- Consistent & reliable metadata search results
- Error messages in the Solr

Future Work:

- More complex validation
- Notifications to data providers

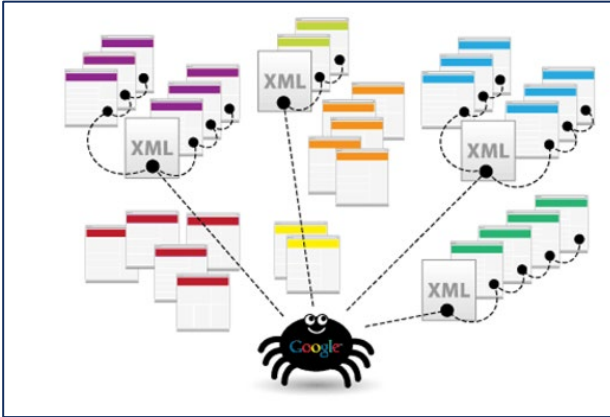
BACKGROUND Sitemap/JSON -LD



Feature#2: Sitemap/JSON -LD

Why need?

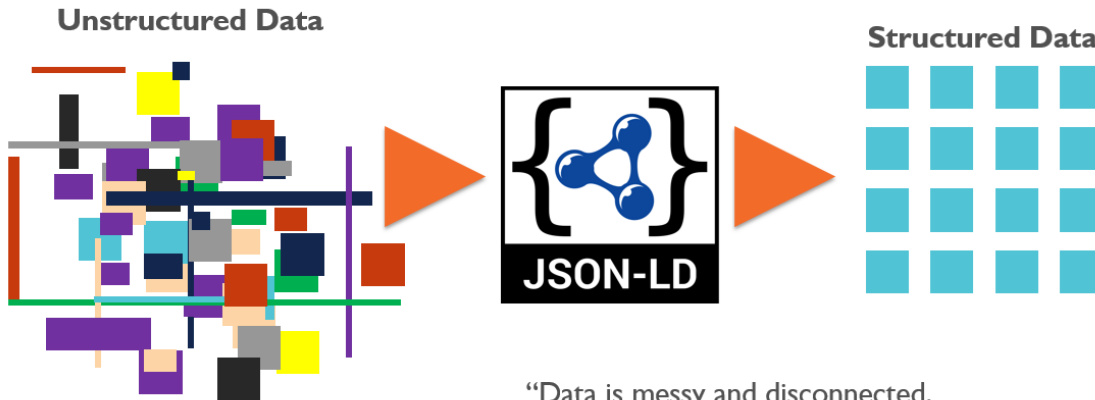
- A good XML sitemap acts as a roadmap of your website that leads Google to all your important pages and JSON-LD describes it.



```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/org.earthsystemgrid.www:0c156aac-ba21-4092-8c9d-0e4e62ebf933</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/f2dfa88e-3abb-11e9-8f53-b808cf016134</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/7f1a2b80-8976-42bb-a2b4-b3d1fbf335c7</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/ca38361d-3aaf-11e9-8f53-b808cf016134</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/edu.ucar.cisl:cloud-collection</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/1f325eee-0d1a-428a-b2f8-6e5eba3652d2</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/dcu-UDI-1958-20150320113628.0-ocm01332775</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/n1-OLA-1968-20150320161613.0-ocm09338459</loc>
  </url>
  <url>
    <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/ncu-GPO-1982-20150313184310.0-ocm10283357</loc>
  </url>
</urlset>
```

Feature#2: Sitemap/JSON -LD

What does JSON-LD do?



“Data is messy and disconnected.
JSON-LD organizes and connects it,
creating a better Web.”

```
"age" : {
  "value" : 37,
  "units" : "years"
},
"children" : [
  {
    "@type" : "Person",
    "first name" : "Josephine",
    "last name" : "Schmoe",
    "age" : {
      "value" : 1,
      "units" : "years"
    }
  }
],
"@context" : "http://schema.org",
"@type" : "WebPage",
"breadcrumb" : "Homepage > Category > Wonderful Things",
"Store" : "http://example123.com/store",
"Product" : "http://example123.com/product",
"<script type='application/ld+json'>
{
  random random random ...
}
"</script>
```

Results & Future Work

Sitemap/JSON -LD



Results: JSON -LD on Our Web Application



High Resolution Historical and Future Simulations Over Hawaii

Id:

org.earthsystemgrid.www::0c156aac-ba21-4092-8c9d-0e4e62ebf933

Description:

To better understand the rainfall climatology and its impacts on hydrological cycle over the Hawaiian Islands under historical and future climates, regional climate simulations over the main Hawaiian islands have been conducted for two 10-year periods using the Weather Research and Forecasting (WRF) model in a configuration of two nested domains. The historical 10-year simulation was driven by the ERA-Interim global reanalysis data and observed sea surface temperature from Oct. 2002 to Sep. 2012 (historical simulation). A high-resolution vertical coordinate was employed to better resolve the trade wind inversion (TWI). Results show that the historical simulation reproduces the mean surface temperature, relative humidity and winds with low biases (+/- 1 degree C, +/- 4% and +/- 1 m s⁻¹, respectively) and high spatial correlations ($r > 0.80$). Additionally, for the historical simulation WRF accurately reproduced aggregated daily and hourly rainfall probability density functions (PDFs) and rainfall spatial-temporal distributions, likely because WRF captured the TWI properties well. The historical simulation outputs are available at hourly resolution for near surface (2-dimensional) fields and for the 3-dimensional atmosphere.

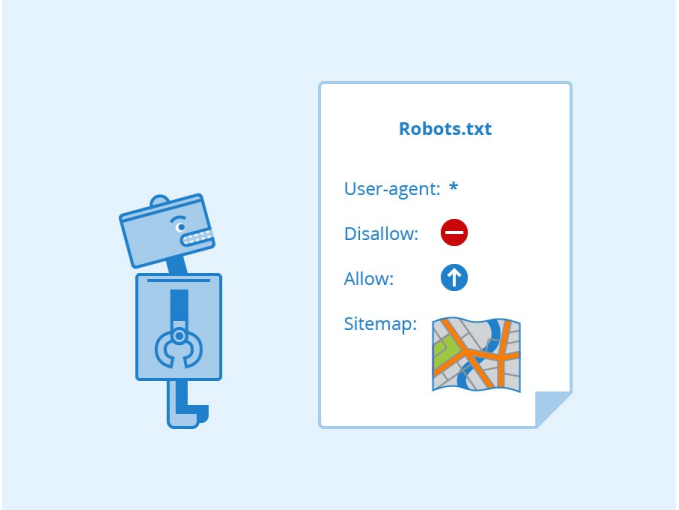
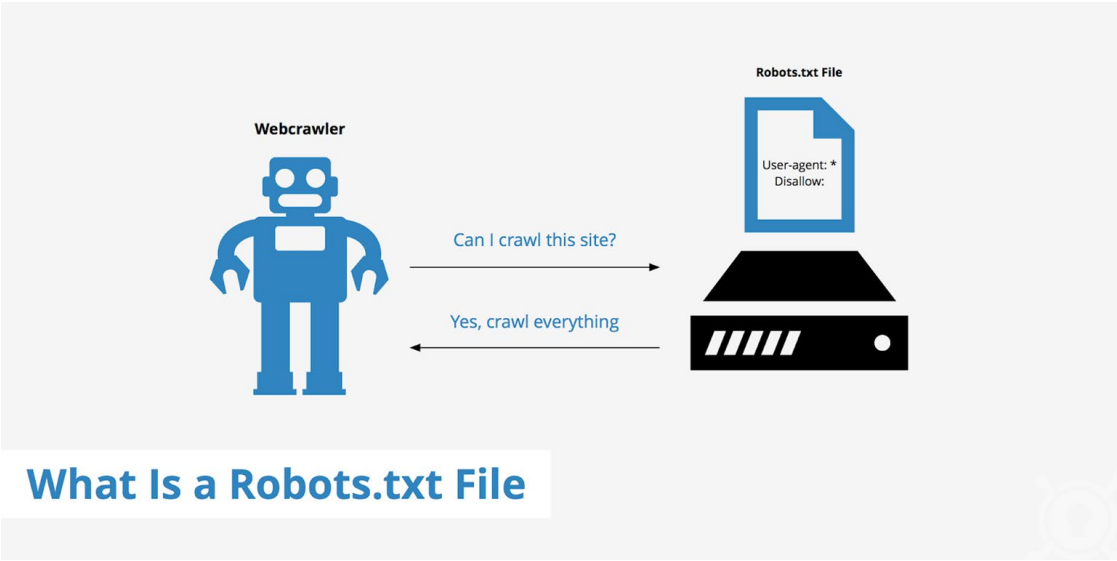
DOI:

<https://doi.org/10.5065/7c2d-bg23>

```
{
  "@context" : "http://schema.org",
  "@type" : "Dataset",
  "@id" : "https://doi.org/10.5065/7c2d-bg23",
  "identifier" : "https://doi.org/10.5065/7c2d-bg23",
  "name" : "High Resolution Historical and Future Simulations Over Hawaii",
  "description" : "To better understand the rainfall climatology and its impacts on hydrological cycle over the Hawaiian Islands under historical and future climates, regional climate simulations over the main Hawaiian islands have been conducted for two 10-year periods using the Weather Research and Forecasting (WRF) model in a configuration of two nested domains. The historical 10-year simulation was driven by the ERA-Interim global reanalysis data and observed sea surface temperature from Oct. 2002 to Sep. 2012 (historical simulation). A high-resolution vertical coordinate was employed to better resolve the trade wind inversion (TWI). Results show that the historical simulation reproduces the mean surface temperature, relative humidity and winds with low biases (+/- 1 degree C, +/- 4% and +/- 1 m s-1, respectively) and high spatial correlations (r > 0.80). Additionally, for the historical simulation WRF accurately reproduced aggregated daily and hourly rainfall probability density functions (PDFs) and rainfall spatial-temporal distributions, likely because WRF captured the TWI properties well. The historical simulation outputs are available at hourly resolution for near surface (2-dimensional) fields and for the 3-dimensional atmosphere.",
  "url" : "https://doi.org/10.5065/7c2d-bg23",
  "sameAs" : "https://doi.org/10.5065/7c2d-bg23",
  "isAccessibleForFree" : "true"
}
```

Future Work for Sitemap/JSON -LD

Adding Robots.txt to Direct Search Engines to our sitemap.xml



BACKGROUND

Search Engine Faceting





Feature #3: Faceting

Why need?

- We want to be able to filter metadata search results by their resource type

```
{
  "id": "ea0f3e13-8ec9-4424-bb8c-8fc4d79edce3",
  "title": "MetPy",
  "description": "MetPy is a collection of tools in Python for reading, visualizing, and performing calculations with weather data. The space MetPy a",
  "doi": "https://www.unidata.ucar.edu/software/metpy",
  "keywords": ["Software"],
  "resource_type": "software",
  "authors": ["Ryan May"],
  "author_emails": ["rmay@ucar.edu"],
  "authoritative_source_url": "https://www.unidata.ucar.edu/software/metpy",
  "authoritative_source_location_on_disk": "/Users/sduishebaeva/Java/xml/test-pull-method/test12.xml",
  "authoritative_source_md5": "d12d65916718529bc4e9e8edle9e7c28",
  "github_xml_url": "https://github.com/dsabira/test-pull-method.git/blob/main/test12.xml",
  "is_valid": true,
  "index_timestamp": "2022-07-15T17:11:34.267Z",
  "_version_": 1738439492126113792},
{
  "id": "edu.ucar.rda::ds010.0",
  "title": "Daily Northern Hemisphere Sea Level Pressure Grids, continuing from 1899",
  "description": "The 5-degree latitude/longitude grids contained in this dataset make up the longest continuous set of daily gridded Northern Hemisp",
  "doi": "https://doi.org/10.5065/7NB6-RJ33",
  "keywords": ["EARTH SCIENCE > ATMOSPHERE > ATMOSPHERIC PRESSURE > SEA LEVEL PRESSURE"],
  "resource_type": "publication",
  "authoritative_source_url": "https://doi.org/10.5065/7NB6-RJ33",
  "authoritative_source_location_on_disk": "/Users/sduishebaeva/Java/xml/test-pull-method/test7.xml",
  "authoritative_source_md5": "84df1f65ac71e181212602691f04f0d1",
  "github_xml_url": "https://github.com/dsabira/test-pull-method.git/blob/main/test7.xml",
  "is_valid": true,
  "index_timestamp": "2022-07-15T17:11:34.987Z",
  "_version_": 1738439492881088512},
```

Results & Future Work Search Engine Faceting



Faceting: Resource Type in the UI

Welcome to Metadata Search Engine!

This is a search engine to access digital assets from all NCAR/UCAR.

Resource Type Clear

[publication](#) 8631

[dataset](#) 8072

[software](#) 18

[fieldSession](#) 4

[collectionSession](#) 1

[model](#) 1

[service](#) 1

16728 Results Found

[Cloud-Optimized Datasets](#)

The datasets in this collection are in Zarr format and hosted by various cloud-based computing providers. See dataset documentation for links to example Jupyter notebooks. The notebooks show how to process the data in parallel within the cloud using Python-based tools.

[COLLECTIONSESSION](#)

[Climatological data for Arctic stations, 1957-1958](#)

This resource is a booklet summarizing polar meteorological observations from the International Geophysical Year, July 1957-Dec. 1958. Information provided includes measurements from two drifting stations, and radiation data. DRIFTING STATION A: Introduction; Map of track and area traversed; Surf...

[DATASET](#)

[Climate summaries of the year in New Caledonia and Dependencies](#)

This resource contains meteorological data from New Caledonia, 1965-1967, 1969-1989. Summary: Annual summary of climate data collected from 64 observation points in New Caledonia and its territories. Measured quantities include: I - annual rainfall deviation from average; II - [total monthly] prec...

Faceting: It Works!

Resource Type **Clear** 8631 Results Found

publication 8631

[On the sources and sinks of atmospheric VOCs: An integrated ...](#)
We apply a high-resolution chemical transport model (GEOS-Chem CTM) with updated treatment of volatile organic compounds (VOCs) and a comprehensive suite of airborne datasets over North America to (i) characterize the VOC budget and (ii) test the ability of current models to capture the distribut...
PUBLICATION

[Aircraft-based aerosol sampling in clouds: Performance chara...](#)
Interaction of liquid cloud droplets and ice particles with aircraft aerosol inlets can result in the generation of a large number of secondary particles and contaminate aerosol measurements. Recent studies have shown that a sampler designed with a perpendicular subsampling tube located within a ...
PUBLICATION

[Middle atmosphere temperature trends in the twentieth and tw...](#)
We use Whole Atmosphere Community Climate Model simulations made under various climate change scenarios to study the evolution of the global-mean temperature trend in the late twentieth century and the twenty-first century. Results are compared with available satellite observations, including new...

Future Work:

- More faceting options such as labs, keywords, authors etc.
- Multi-select option

Resource Type **Clear** 8072 Results Found

dataset 8072

[Climatological data for Arctic stations, 1957-1958](#)
This resource is a booklet summarizing polar meteorological observations from the International Geophysical Year, July 1957-Dec. 1958. Information provided includes measurements from two drifting stations, and radiation data. DRIFTING STATION A: Introduction; Map of track and area traversed; Surf...
DATASET

[Climate summaries of the year in New Caledonia and Dependencies](#)
This resource contains meteorological data from New Caledonia, 1965-1967, 1969-1989. Summary: Annual summary of climate data collected from 64 observation points in New Caledonia and its territories. Measured quantities include: I - annual rainfall deviation from average; II - [total monthly] prec...
DATASET

[Monthly normals of temperature, precipitation, and heating a...](#)
This resource presents climatological normals based on records for the 30-year period 1951-1980. A normal of a climatological element is the arithmetic mean computer over a time period spanning three consecutive decades. Contents include: temperature normals (maximum, minimum, mean) averaged by m...

ACKNOWLEDGEMENTS



Thank You

NCAR and CISL

SIParCS Mentors:

- Nathan Hook
- Saquib Aziz Khan
- Eric Nienhouse

Project Partner:

- Teagan Johnson

SIParCS Program Leads:

- Virginia Do
- Jerry Cyccone
- AJ Lauer
- Francesgladys Pulido

And everyone else involved with the SIParCS program this summer.

Resources

- <https://thenounproject.com/icons/> - Most of the images/icons used
- <https://www.searchenginejournal.com/upgrade-to-json-ld-structured-data/319327/> - Information about JSON-LD
- <https://www.seobility.net/en/wiki/Robots.txt> - Images/information about the Sitemap



QUESTIONS?

