DatastructuresForAnalysis

March 24, 2023

1 Data analysis with data structures

Given a list of CMIP files for the same setup by different models the task is to perform several analysis tasks:

- Compare results based on the model or site
- Compute statistics on different selections of the data (like standard deviation)
- Compute correlations of CO2 und Sea Ice Extend

Background Global warming and the CO2 emissions: Based on a paper by Dirk Notz https://doi.org/10.1126/science.aag2345

[1]:
```
import os
datadir='../data/'
!ls ../data/CMIP*
```

```
../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.Amon.gn.nc
../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.SImon.gn.nc
../data/CMIP.MPI-M.MPI-ESM1-2-HR.historical.Amon.gn.nc
../data/CMIP.MPI-M.MPI-ESM1-2-HR.historical.SImon.gn.nc
../data/CMIP.MPI-M.MPI-ESM1-2-LR.historical.Amon.gn.nc
../data/CMIP.MPI-M.MPI-ESM1-2-LR.historical.SImon.gn.nc
../data/CMIP.NCAR.CESM2-FV2.historical.Amon.gn.nc
../data/CMIP.NCAR.CESM2-FV2.historical.SImon.gn.nc
../data/CMIP.NCAR.CESM2.historical.Amon.gn.nc
../data/CMIP.NCAR.CESM2.historical.SImon.gn.nc
../data/CMIP.NCAR.CESM2-WACCM-FV2.historical.Amon.gn.nc
../data/CMIP.NCAR.CESM2-WACCM-FV2.historical.SImon.gn.nc
../data/CMIP.NCAR.CESM2-WACCM.historical.Amon.gn.nc
../data/CMIP.NCAR.CESM2-WACCM.historical.SImon.gn.nc
../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc
../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc
```

1.1 Question: How to set up a data structure to support the tasks?

1.2 1 Idea: `list`

[2]:
```
import glob
fileList = glob.glob(f"{datadir}/CMIP*")
print(fileList)
```

```
['../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc',
'../data/CMIP.NCAR.CESM2-WACCM-FV2.historical.SImon.gn.nc',
'../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.SImon.gn.nc',
'../data/CMIP.NCAR.CESM2.historical.Amon.gn.nc',
'../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc', '../data/CMIP.HAMMOZ-
Consortium.MPI-ESM-1-2-HAM.historical.Amon.gn.nc',
'../data/CMIP.NCAR.CESM2-WACCM.historical.Amon.gn.nc', '../data/CMIP.MPI-M.MPI-
ESM1-2-HR.historical.Amon.gn.nc', '../data/CMIP.MPI-M.MPI-
ESM1-2-HR.historical.SImon.gn.nc',
'../data/CMIP.NCAR.CESM2-FV2.historical.Amon.gn.nc',
'../data/CMIP.NCAR.CESM2-FV2.historical.SImon.gn.nc',
'../data/CMIP.NCAR.CESM2.historical.SImon.gn.nc',
'../data/CMIP.NCAR.CESM2-WACCM-FV2.historical.Amon.gn.nc',
'../data/CMIP.NCAR.CESM2-WACCM.historical.SImon.gn.nc', '../data/CMIP.MPI-M.MPI-
ESM1-2-LR.historical.Amon.gn.nc', '../data/CMIP.MPI-M.MPI-
ESM1-2-LR.historical.SImon.gn.nc']
```

- What are the pros and cons?
- How to select specific data?

Start: Compute the meta information from the files

[8]:
```
d1 = {"NorESM2" :  ['../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc','../
  ↪data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc']}
print(d)
model = "NorESM2"
testDaten_ice = d1[model][0]
testDaten_co2 = d1[model][1]
d2 = {"NorESM2" : { 'ice' : '../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.
  ↪nc', 'co2': '../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc'} }
testDaten_ice = d2[model]['ice']
testDaten_co2 = d2[model]['co2']
print(testDaten_co2)
```

```
{'NorESM2': ['../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc',
'../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc']}
../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc
```

[28]:
```
!ls ../data/CMIP.*.MPI-ESM-1-2-HAM.historical.*.nc
```

```
../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.Amon.gn.nc
../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.SImon.gn.nc
```

[34]:
```
model = "NorESM2-MM"
files = glob.glob("../data/CMIP.*."+model+".historical.*.nc")
print(files)
iceFile = glob.glob("../data/CMIP.*."+model+".historical.SImon.*.nc")
co2File = glob.glob("../data/CMIP.*."+model+".historical.Amon.*.nc")
print(iceFile,co2File)
```

```
['../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc',
'../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc']
['../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc']
['../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc']
```

[36]:
```
import numpy as np

models = [f.split('.')[-5] for f in fileList]
#print(models)
models = np.unique(models)
#print(models)
dataDict = dict() # = {}
for model in models:
    print(model)
    dataDict[model] = {
        'ice': glob.glob("../data/CMIP.*."+model+".historical.SImon.*.nc")[0],
        'co2': glob.glob("../data/CMIP.*."+model+".historical.Amon.*.nc")[0]
    }

print(dataDict)
#sites  = [f.split('.')[-6] for f in fileList]
#[models,sites]
#list(filter(lambda x: (x.split('.')[-5] == 'CESM2-FV2'), fileList))
```

```
CESM2
CESM2-FV2
CESM2-WACCM
CESM2-WACCM-FV2
MPI-ESM-1-2-HAM
MPI-ESM1-2-HR
MPI-ESM1-2-LR
NorESM2-MM
{'CESM2': {'ice': '../data/CMIP.NCAR.CESM2.historical.SImon.gn.nc', 'co2':
'../data/CMIP.NCAR.CESM2.historical.Amon.gn.nc'}, 'CESM2-FV2': {'ice':
'../data/CMIP.NCAR.CESM2-FV2.historical.SImon.gn.nc', 'co2':
'../data/CMIP.NCAR.CESM2-FV2.historical.Amon.gn.nc'}, 'CESM2-WACCM': {'ice':
'../data/CMIP.NCAR.CESM2-WACCM.historical.SImon.gn.nc', 'co2':
'../data/CMIP.NCAR.CESM2-WACCM.historical.Amon.gn.nc'}, 'CESM2-WACCM-FV2':
{'ice': '../data/CMIP.NCAR.CESM2-WACCM-FV2.historical.SImon.gn.nc', 'co2':
'../data/CMIP.NCAR.CESM2-WACCM-FV2.historical.Amon.gn.nc'}, 'MPI-ESM-1-2-HAM':
{'ice': '../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.SImon.gn.nc',
```

```
'co2': '../data/CMIP.HAMMOZ-Consortium.MPI-ESM-1-2-HAM.historical.Amon.gn.nc'},
'MPI-ESM1-2-HR': {'ice': '../data/CMIP.MPI-M.MPI-
ESM1-2-HR.historical.SImon.gn.nc', 'co2': '../data/CMIP.MPI-M.MPI-
ESM1-2-HR.historical.Amon.gn.nc'}, 'MPI-ESM1-2-LR': {'ice':
'../data/CMIP.MPI-M.MPI-ESM1-2-LR.historical.SImon.gn.nc', 'co2':
'../data/CMIP.MPI-M.MPI-ESM1-2-LR.historical.Amon.gn.nc'}, 'NorESM2-MM': {'ice':
'../data/CMIP.NCC.NorESM2-MM.historical.SImon.gn.nc', 'co2':
'../data/CMIP.NCC.NorESM2-MM.historical.Amon.gm.nc'}}
```

[43]:
```
import random
def modelAnalysis(iceFile,co2File):
    return random.random()
```

[50]:
```
modelValues_list = []
modelValues_dict = {}
for model in models:
    ice = dataDict[model]['ice']
    co2 = dataDict[model]['co2']
    result = modelAnalysis(ice,co2)
    modelValues_list.append( [model,result] )
    modelValues_dict[model] = result

print(models)
print(modelValues_list)
print(modelValues_dict)
```

```
['CESM2' 'CESM2-FV2' 'CESM2-WACCM' 'CESM2-WACCM-FV2' 'MPI-ESM-1-2-HAM'
 'MPI-ESM1-2-HR' 'MPI-ESM1-2-LR' 'NorESM2-MM']
[['CESM2', 0.42967087554459493], ['CESM2-FV2', 0.9963623049668217],
['CESM2-WACCM', 0.6930955067073447], ['CESM2-WACCM-FV2', 0.7627488911486247],
['MPI-ESM-1-2-HAM', 0.34315192192508015], ['MPI-ESM1-2-HR', 0.5088914006703273],
['MPI-ESM1-2-LR', 0.0202380065038259], ['NorESM2-MM', 0.6950668696907668]]
{'CESM2': 0.42967087554459493, 'CESM2-FV2': 0.9963623049668217, 'CESM2-WACCM':
0.6930955067073447, 'CESM2-WACCM-FV2': 0.7627488911486247, 'MPI-ESM-1-2-HAM':
0.34315192192508015, 'MPI-ESM1-2-HR': 0.5088914006703273, 'MPI-ESM1-2-LR':
0.0202380065038259, 'NorESM2-MM': 0.6950668696907668}
```

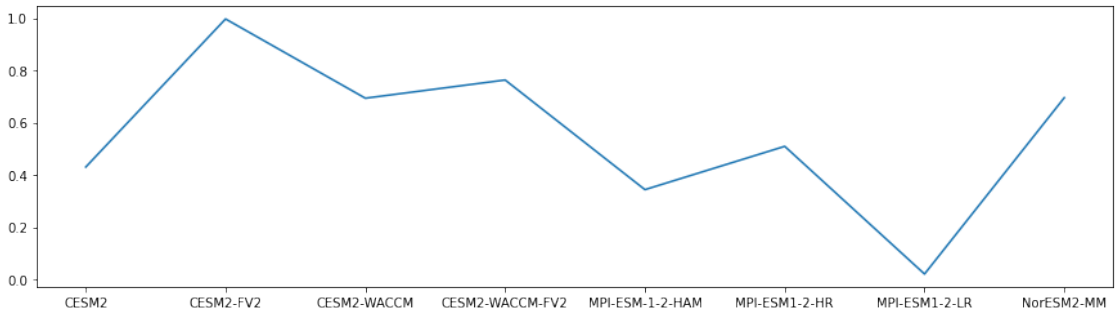
[52]:
```
%matplotlib inline
import matplotlib.pyplot as plt
```

[59]:
```
values = modelValues_dict.values()
keys   = modelValues_dict.keys()
```

[63]:
```
fig,ax = plt.subplots(figsize=(15,4))

plt.plot(keys,values)
```

[63]: [<matplotlib.lines.Line2D at 0x7fff95a6c460>]



1.3 2 Idea: `dict`

Set up a dictinary with models as keys and data files as values

```
{'CESM2-FV2' :
    ['../data/CMIP.NCAR.CESM2-FV2.historical.SImon.gn.nc',
     '../data/CMIP.NCAR.CESM2-FV2.historical.Amon.gn.nc'],
  ...
```

Target loop:

```
for model , files in dataDict.items():
    co2File, seaIceExtFile = files[0], files[1]
    corrField = computeCorrelation(co2File, seaIceExtFile)
```

1.4 3 Idea: reverse `dict`

Set up a dictionary with data files as keys and attach meta data as values

```
{ '../data/CMIP.NCAR.CESM2-FV2.historical.SImon.gn.nc' :
    { 'model': 'CESM2-FV2', 'site': 'NCAR', 'varname':'siextentn', 'component': 'ocean', ...}
```