In [1]:
```
import pandas as pd
import numpy as np
import os, shutil
```

In [2]:
```
cols = [ 'file_id', 'phenom_cat','Sentinel1_cat','partial_id','longitude','latitu
labels_df = pd.read_csv('E:\LargeDatasets\SAR-Ocean-Images\InfoFiles\labels_and_i
```

In [3]:
```
labels_df.head()
```

Out[3]:

	file_id	phenom_cat	Sentinel1_cat	partial_id	longitude	latitude
0	H/s1a-wv2-slc-vv-20161122t035148-20161122t0351...	H	WV2	20161122t035148	-124.3430	-60.40080
1	H/s1a-wv1-slc-vv-20160406t195302-20160406t1953...	H	WV1	20160406t195302	155.0790	38.02420
2	H/s1a-wv2-slc-vv-20161102t172849-20161102t1728...	H	WV2	20161102t172849	-172.8850	24.77710
3	H/s1a-wv2-slc-vv-20161218t161205-20161218t1612...	H	WV2	20161218t161205	-156.4120	4.61075
4	H/s1a-wv2-slc-vv-20160303t220401-20160303t2204...	H	WV2	20160303t220401	-48.7943	-38.97550

In [4]:
```
labels_df.iloc[0,0]
```

Out[4]:
```
'H/s1a-wv2-slc-vv-20161122t035148-20161122t035151-014049-016a76-014.png'
```

In [5]:
```
labels_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37553 entries, 0 to 37552
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   file_id        37553 non-null  object 
 1   phenom_cat     37553 non-null  object 
 2   Sentinel1_cat  37553 non-null  object 
 3   partial_id     37553 non-null  object 
 4   longitude      37553 non-null  float64
 5   latitude       37553 non-null  float64
dtypes: float64(2), object(4)
memory usage: 1.7+ MB
```

For some reason, even though the intended files for machine learning are the .tiff files, all the file paths have a .png ending, so we will have to replace those. Additionally, we don't need any of the last four columns, so we'll drop them. Finally, we need to clean the file_id column so that it is just

the file name. Currently, its set up more like a path, but we'll handle that separately. So, we have to remove the prefix containing the category designation as well as the backslash.

In [6]:
```
labels_df['phenom_cat'].value_counts()
```

Out[6]:
```
F    4900
G    4797
I    4740
J    4709
H    4598
K    4370
N    4100
M    2160
L    1980
O    1199
Name: phenom_cat, dtype: int64
```

Fairly balanced in terms of categories. For reference, the above letters correspond to:
F - Pure Ocean Waves
G - Wind Streaks
H - Micro Convective Cells
I - Rain Cells
J - Biological Slicks
K - Sea Ice
L - Iceberg
M - Low Wind Area
N - Atmospheric Front
O - Oceanic Front
Which is saved in category_defs.txt in the directory containing the images. From the paper describing this dataset (https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/gdj3.73), we see that it is largely balanced except for the Iceberg and Oceanic Front categories due to seasonality and rarity, respectively. We may want to look at image augmentation for these categories in the future. Well, let's go about fixing up the dataframe.

In [7]:
```
labels_df = labels_df.drop(columns = ['Sentinel1_cat','partial_id','longitude',']
labels_df.head()
```

Out[7]:

	file_id	phenom_cat
0	H/s1a-wv2-slc-vv-20161122t035148-20161122t0351...	H
1	H/s1a-wv1-slc-vv-20160406t195302-20160406t1953...	H
2	H/s1a-wv2-slc-vv-20161102t172849-20161102t1728...	H
3	H/s1a-wv2-slc-vv-20161218t161205-20161218t1612...	H
4	H/s1a-wv2-slc-vv-20160303t220401-20160303t2204...	H

In [8]:
```
labels_df.iloc[0,0][2:]
```

Out[8]:
```
's1a-wv2-slc-vv-20161122t035148-20161122t035151-014049-016a76-014.png'
```

In [9]:
```
labels_df['file_id'] = labels_df.file_id.apply(lambda x: x[2:])
labels_df.head()
```

Out[9]:

	file_id	phenom_cat
0	s1a-wv2-slc-vv-20161122t035148-20161122t035151...	H
1	s1a-wv1-slc-vv-20160406t195302-20160406t195305...	H
2	s1a-wv2-slc-vv-20161102t172849-20161102t172852...	H
3	s1a-wv2-slc-vv-20161218t161205-20161218t161208...	H
4	s1a-wv2-slc-vv-20160303t220401-20160303t220404...	H

In [10]:
```
labels_df['file_id'] = labels_df.file_id.apply(lambda x: x.replace('.png','.tiff'
labels_df.iloc[0,0]
```

Out[10]:
```
's1a-wv2-slc-vv-20161122t035148-20161122t035151-014049-016a76-014.tiff'
```

In [11]:
```
labels_df.head()
```

Out[11]:

	file_id	phenom_cat
0	s1a-wv2-slc-vv-20161122t035148-20161122t035151...	H
1	s1a-wv1-slc-vv-20160406t195302-20160406t195305...	H
2	s1a-wv2-slc-vv-20161102t172849-20161102t172852...	H
3	s1a-wv2-slc-vv-20161218t161205-20161218t161208...	H
4	s1a-wv2-slc-vv-20160303t220401-20160303t220404...	H

Swag. Now we shall reorganize our images into folders containing train, validation and test sets. Here I'll note that the images are stored on my extra HDD storage, since they take up large amounts of space, and are not in this directory.

In [14]:
```
path_to_files_head = 'E:\LargeDatasets\SAR-Ocean-Images\GeoTIFF'
new_root_dir = 'E:\LargeDatasets\SAR-Ocean-Images\GeoTIFF\OrganisationForModel'
os.mkdir(new_root_dir)
subset_dir_names = ['train','val','test']
for d in subset_dir_names:
    new_dir = os.path.join(new_root_dir,d)
    os.mkdir(new_dir)

for phenom in list(labels_df.phenom_cat.unique()):
    print(f'Copying {phenom} pictures.')

    for d in subset_dir_names:
        new_dir = os.path.join(new_root_dir,d,phenom)
        os.mkdir(new_dir)

    phenom_images = labels_df[labels_df.phenom_cat == phenom]
    train, val, test = np.split(phenom_images.sample(frac=1), [int(.8*len(phenom_
    print(f'Split {len(phenom_images)} imgs into {len(train)} train, {len(val)} \

    for i, subset in enumerate([train,val,test]):
        for row in subset.index:
            filename = subset['file_id'][row]
            origin = os.path.join(path_to_files_head, phenom, filename)
            destination = os.path.join(new_root_dir, subset_dir_names[i], phenom,
            shutil.copy(origin, destination)
```

```
Copying H pictures.
Split                                                 file_id phenom_cat
0     s1a-wv2-slc-vv-20161122t035148-20161122t035151...          H
1     s1a-wv1-slc-vv-20160406t195302-20160406t195305...          H
2     s1a-wv2-slc-vv-20161102t172849-20161102t172852...          H
3     s1a-wv2-slc-vv-20161218t161205-20161218t161208...          H
4     s1a-wv2-slc-vv-20160303t220401-20160303t220404...          H
...                                                 ...        ...
4593  s1a-wv2-slc-vv-20161211t160926-20161211t160929...          H
4594  s1a-wv2-slc-vv-20160405t125833-20160405t125836...          H
4595  s1a-wv2-slc-vv-20160707t184601-20160707t184604...          H
4596  s1a-wv2-slc-vv-20160824t063547-20160824t063550...          H
4597  s1a-wv2-slc-vv-20161031t175718-20161031t175721...          H

[4598 rows x 2 columns] imgs into
file_id phenom_cat
1748  s1a-wv2-slc-vv-20160606t013225-20160606t013228...          H
4207  s1a-wv2-slc-vv-20161020t100927-20161020t100930...          H
3883  s1a-wv2-slc-vv-20161124t161531-20161124t161534...          H
```

In [ ]:
```
```