reg_year_r_y

February 2, 2024

0.1 Importing

[ ]:
```
import xarray as xr
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.ensemble import BaggingRegressor
from sklearn.tree import ExtraTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn import preprocessing

from sklearn.metrics import mean_squared_error as mse

import os

from time import sleep
from tqdm import tqdm
```

0.2 Datasets Preparation

[ ]:
```
def datasets_preparation (i):

    # Dataset and date
    ds_name = ('/results2/SalishSea/nowcast-green.202111/' + i + '/
 ↪SalishSea_1d_' + '20' + str(i[5:7]) + str(dict_month[i[2:5]])+str(i[0:2]) +␣
 ↪'_' + '20' + str(i[5:7]) + str(dict_month[i[2:5]]) + str(i[0:2]) + '_grid_T.
 ↪nc')

    ds_bio_name = ('/results2/SalishSea/nowcast-green.202111/' + i + '/
 ↪SalishSea_1d_'  + '20' + str(i[5:7]) + str(dict_month[i[2:5]])+str(i[0:2]) +␣
 ↪'_' + '20' + str(i[5:7]) + str(dict_month[i[2:5]]) + str(i[0:2]) + '_biol_T.
 ↪nc')

    ds = xr.open_dataset (ds_name)
    ds_bio = xr. open_dataset (ds_bio_name)
```

```
    temp_i1 = (ds.votemper.where(mask==1)[0,0:15] * ds.e3t.where(mask==1)
               [0,0:15]).sum('deptht', skipna = True, min_count = 15) / mesh.
↪gdepw_0[0,15]
    temp_i2 = (ds.votemper.where(mask==1)[0,15:27] * ds.e3t.where(mask==1)
               [0,15:27]).sum('deptht', skipna = True, min_count = 12) / (mesh.
↪gdepw_0[0,27] - mesh.gdepw_0[0,14])
    saline_i1 = (ds.vosaline.where(mask==1)[0,0:15] * ds.e3t.where(mask==1)
                 [0,0:15]).sum('deptht', skipna = True, min_count = 15) /␣
↪mesh.gdepw_0[0,15]
    saline_i2 = (ds.vosaline.where(mask==1)[0,15:27] * ds.e3t.where(mask==1)
                 [0,15:27]).sum('deptht', skipna = True, min_count = 12) /␣
↪(mesh.gdepw_0[0,27] - mesh.gdepw_0[0,14])

    diat_i = (ds_bio.diatoms.where(mask==1)[0,0:27] * ds.e3t.where(mask==1)
               [0,0:27]).sum('deptht', skipna = True, min_count = 27) / mesh.
↪gdepw_0[0,27]
    flag_i = (ds_bio.flagellates.where(mask==1)[0,0:27] * ds.e3t.where(mask==1)
               [0,0:27]).sum('deptht', skipna = True, min_count = 27) / mesh.
↪gdepw_0[0,27]

    return (temp_i1, temp_i2, saline_i1, saline_i2, diat_i, flag_i)
```

0.3 Dataset Presentation (Yesterday)

[ ]:
```
def datasets_preparation_y(i):

    # Dataset and date
    ds_name = ('/results2/SalishSea/nowcast-green.202111/' + i + '/
↪SalishSea_1d_' + '20' + str(i[5:7]) + str(dict_month[i[2:5]])+str(i[0:2]) +␣
↪'_' + '20' + str(i[5:7]) + str(dict_month[i[2:5]]) + str(i[0:2]) + '_grid_T.
↪nc')

    ds = xr.open_dataset (ds_name)

    temp_i1_y = (ds.votemper.where(mask==1)[0,0:15] * ds.e3t.where(mask==1)
    [0,0:15]).sum('deptht', skipna = True, min_count = 15) / mesh.gdepw_0[0,15]

    temp_i2_y = (ds.votemper.where(mask==1)[0,15:27] * ds.e3t.where(mask==1)
    [0,15:27]).sum('deptht', skipna = True, min_count = 12) / (mesh.
↪gdepw_0[0,27] - mesh.gdepw_0[0,14])

    saline_i1_y = (ds.vosaline.where(mask==1)[0,0:15] * ds.e3t.where(mask==1)
    [0,0:15]).sum('deptht', skipna = True, min_count = 15) / mesh.gdepw_0[0,15]

    saline_i2_y = (ds.vosaline.where(mask==1)[0,15:27] * ds.e3t.where(mask==1)
```

```
    [0,15:27]).sum('deptht', skipna = True, min_count = 12) / (mesh.
↪gdepw_0[0,27] - mesh.gdepw_0[0,14])

    return (temp_i1_y, temp_i2_y, saline_i1_y, saline_i2_y)
```

0.4 Regressor

[ ]:
```
def regressor (inputs, targets, variable_name):

    inputs = inputs.transpose()

    # Regressor
    scale = preprocessing.StandardScaler()
    inputs2 = scale.fit_transform(inputs)
    X_train, X_test, y_train, y_test = train_test_split(inputs2, targets)

    extra_tree = ExtraTreeRegressor(criterion='poisson')
    regr = BaggingRegressor(extra_tree, n_estimators=10, max_features=8).
↪fit(X_train, y_train)

    outputs_test = regr.predict(X_test)

    m = scatter_plot(y_test, outputs_test, variable_name + ' (Testing dataset)')
    r = np.round(np.corrcoef(y_test, outputs_test)[0][1],3)
    rms = np.round(mse(y_test, outputs_test),4)


    return (r, rms, m, regr)
```

1 Printing

[ ]:
```
def printing (targets, outputs, m):

    print ('The amount of data points is', outputs.size)
    print ('The slope of the best fitting line is ', np.round(m,3))
    print ('The correlation coefficient is:', np.round(np.corrcoef(targets,␣
↪outputs)[0][1],3))
    print (' The mean square error is:', np.round(mse(targets,outputs),5))
```

1.1 Scatter Plot

[ ]:
```
def scatter_plot(targets, outputs, variable_name):

    # compute slope m and intercept b
    m, b = np.polyfit(targets, outputs, deg=1)
```

```
    printing (targets, outputs, m)

    fig, ax = plt.subplots()

    plt.scatter(targets,outputs, alpha = 0.2, s = 10)
    plt.xlabel('targets')
    plt.ylabel('outputs')

    lims = [
        np.min([ax.get_xlim(), ax.get_ylim()]),  # min of both axes
        np.max([ax.get_xlim(), ax.get_ylim()]),  # max of both axes
    ]

    # plot fitted y = m*x + b
    plt.axline(xy1=(0, b), slope=m, color='r')

    ax.set_aspect('equal')
    ax.set_xlim(lims)
    ax.set_ylim(lims)

    ax.plot(lims, lims,linestyle = '--',color = 'k')

    fig.suptitle(str(year) + ', ' + variable_name)

    plt.show()

    return (m)
```

1.2 Plotting

[ ]:
```
def plotting (variable, name):

    plt.plot(years,variable, marker = '.', linestyle = '')
    plt.legend(['diatom','flagellate'])
    plt.xlabel('Years')
    plt.ylabel(name)
    plt.show()
```

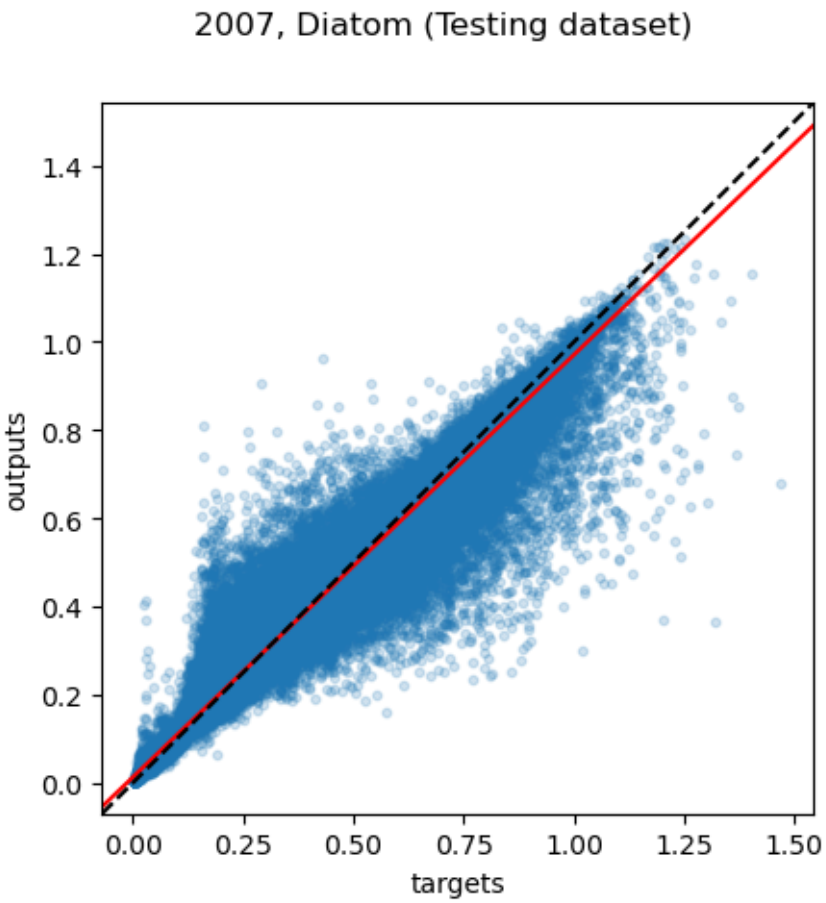
1.3 Main Body

[ ]:
```
dict_month = {'jan': '01',
         'feb': '02',
         'mar': '03',
         'apr': '04',
         'may': '05',
```

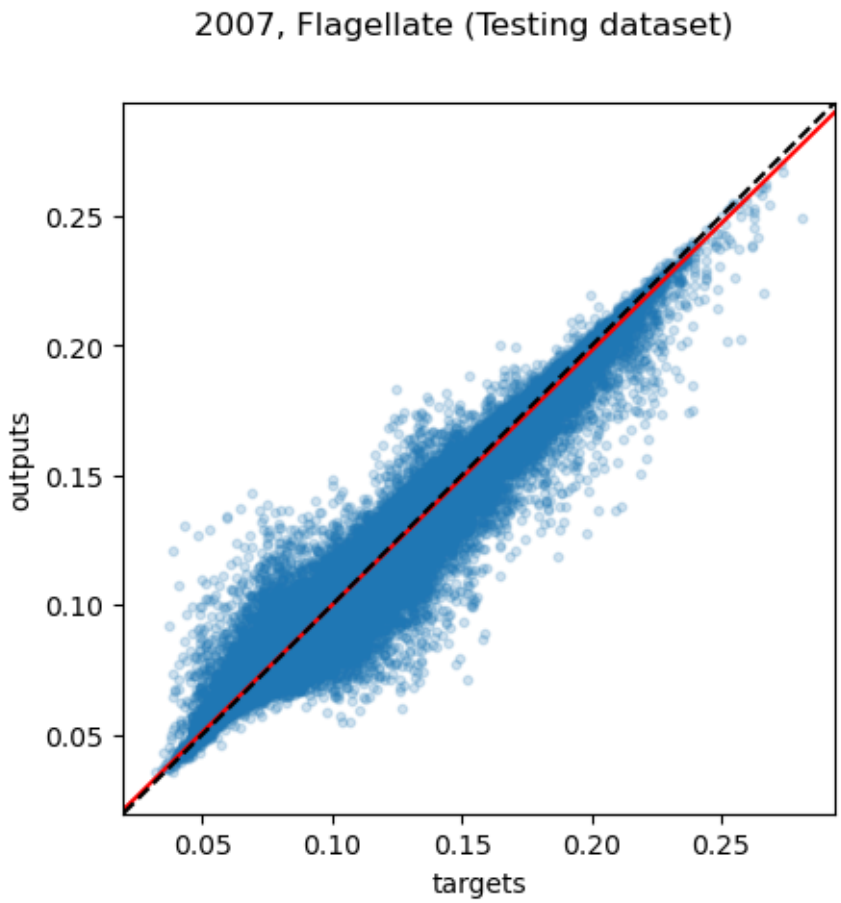
```
          'jun': '06',
          'jul': '07',
          'aug': '08',
          'sep': '09',
          'oct': '10',
          'nov': '11',
          'dec': '12'}

path = os.listdir('/results2/SalishSea/nowcast-green.202111/')

years = range (2007,2024)

# Open the mesh mask
mesh = xr.open_dataset('/home/sallen/MEOPAR/grid/mesh_mask202108.nc')
mask = mesh.tmask.to_numpy()

r_all = [],[]
rms_all = [],[]
slope_all = [],[]

for year in tqdm(range(2007,2024)):

    year_str = str(year)[2:4]

    folders = [x for x in path if ((x[2:5]=='mar' or x[2:5]=='apr' or (x[2:
↪5]=='feb' and x[0:2] > '13')) and (x[5:7]==year_str))]
    indx_dates=(np.argsort(pd.to_datetime(folders, format="%d%b%y")))
    folders = [folders[i] for i in indx_dates]

    drivers_all = np.array([[],[],[],[],[],[],[],[]])
    diat_all = np.array([])
    flag_all = np.array([])

    print ('Gathering days for year ' + str(year))

    for i in tqdm(range(1, len(folders)), position=0, leave=True):

        temp_i1, temp_i2, saline_i1, saline_i2, diat_i, flag_i =␣
↪datasets_preparation(folders[i])
        temp_i1_y, temp_i2_y, saline_i1_y, saline_i2_y =␣
↪datasets_preparation_y(folders[i-1])

        drivers = np.stack([np.ravel(temp_i1), np.ravel(temp_i2), np.
↪ravel(saline_i1), np.ravel(saline_i2), np.ravel(temp_i1_y), np.
↪ravel(temp_i2_y), np.ravel(saline_i1_y), np.ravel(saline_i2_y)])
        indx = np.where(~np.isnan(drivers).any(axis=0))
        drivers = drivers[:,indx[0]]
```

```
        drivers_all = np.concatenate((drivers_all,drivers),axis=1)

        diat = np.ravel(diat_i)
        diat = diat[indx[0]]
        diat_all = np.concatenate((diat_all,diat))

        flag = np.ravel(flag_i)
        flag = flag[indx[0]]
        flag_all = np.concatenate((flag_all,flag))

        sleep(0.1)

    print ('Done gathering, building the prediction models')
    print ('\n')

    r, rms, m, regr = regressor(drivers_all, diat_all, 'Diatom')
    r_all[0].append(r)
    rms_all[0].append(rms)
    slope_all[0].append(m)

    r, rms, m, regr = regressor(drivers_all, flag_all, 'Flagellate')
    r_all[1].append(r)
    rms_all[1].append(rms)
    slope_all[1].append(m)

    sleep(0.1)

plotting(np.transpose(r_all), 'Correlation Coefficient')
plotting(np.transpose(rms_all), 'Mean Square Error')
plotting (np.transpose(slope_all), 'Slope of the best fitting line')
```

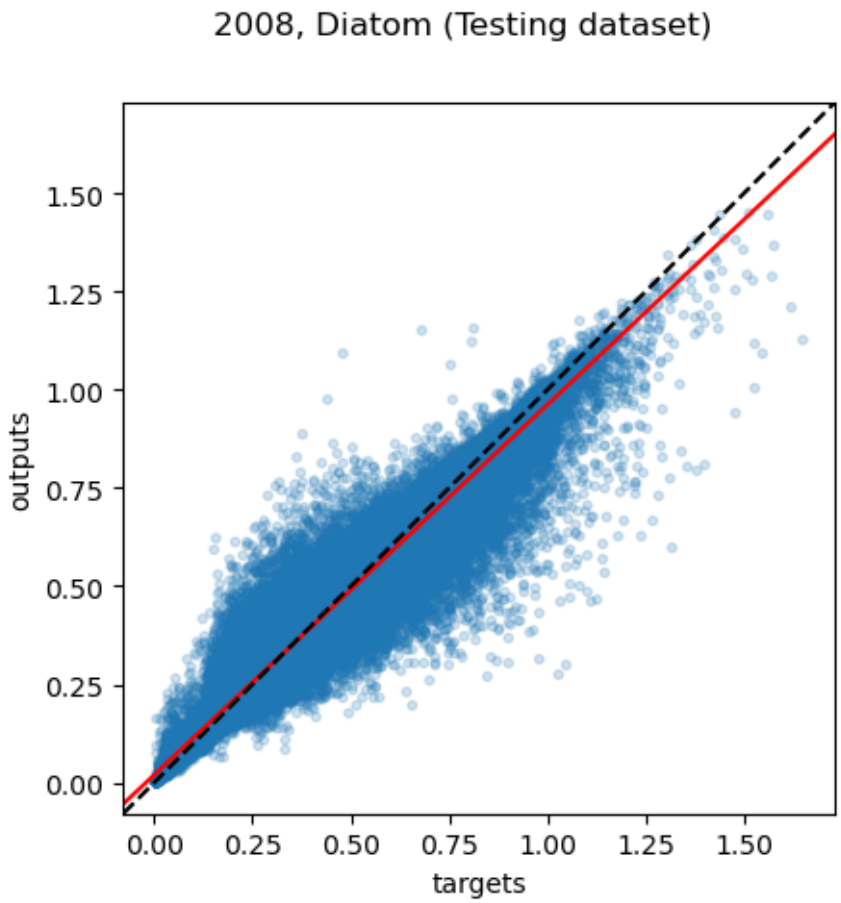
```
  0%|          | 0/17 [00:00<?, ?it/s]
Gathering days for year 2007
100%|      | 75/75 [04:14<00:00,  3.39s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.958
The correlation coefficient is: 0.987
 The mean square error is: 0.00068
```



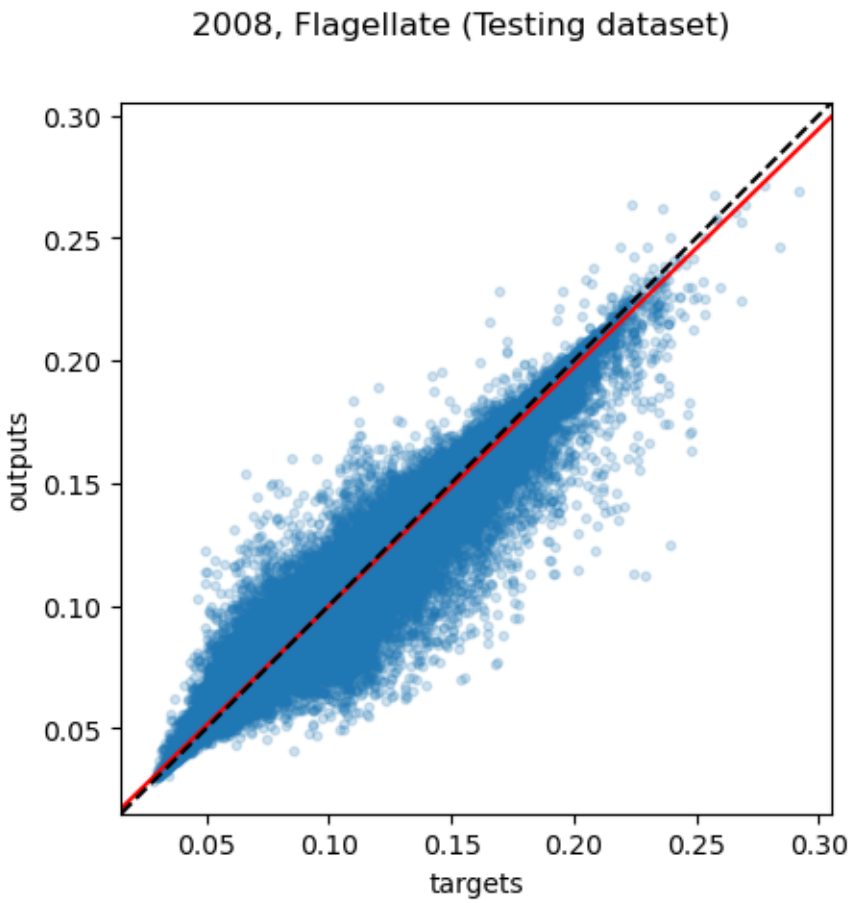

```
The amount of data points is 871482
The slope of the best fitting line is  0.981
The correlation coefficient is: 0.994
 The mean square error is: 1e-05
```



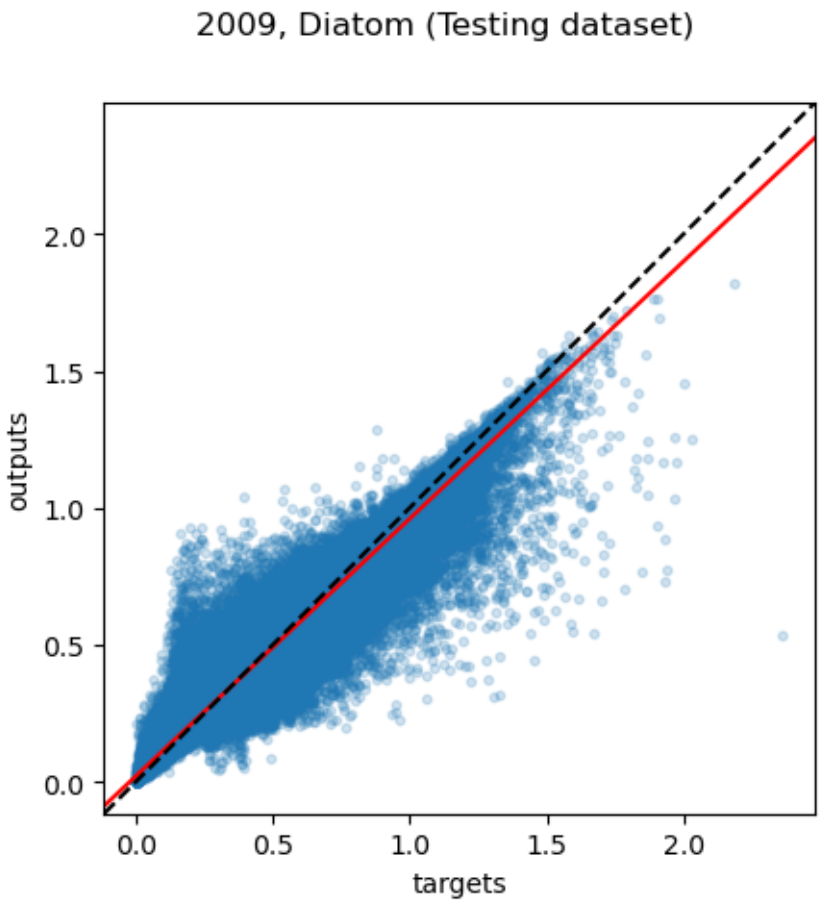

```
  6%|          | 1/17 [07:49<2:05:09, 469.33s/it]
Gathering days for year 2008
100%|      | 76/76 [04:19<00:00,  3.41s/it]
Done gathering, building the prediction models


The amount of data points is 883101
The slope of the best fitting line is  0.942
The correlation coefficient is: 0.981
 The mean square error is: 0.00079
```



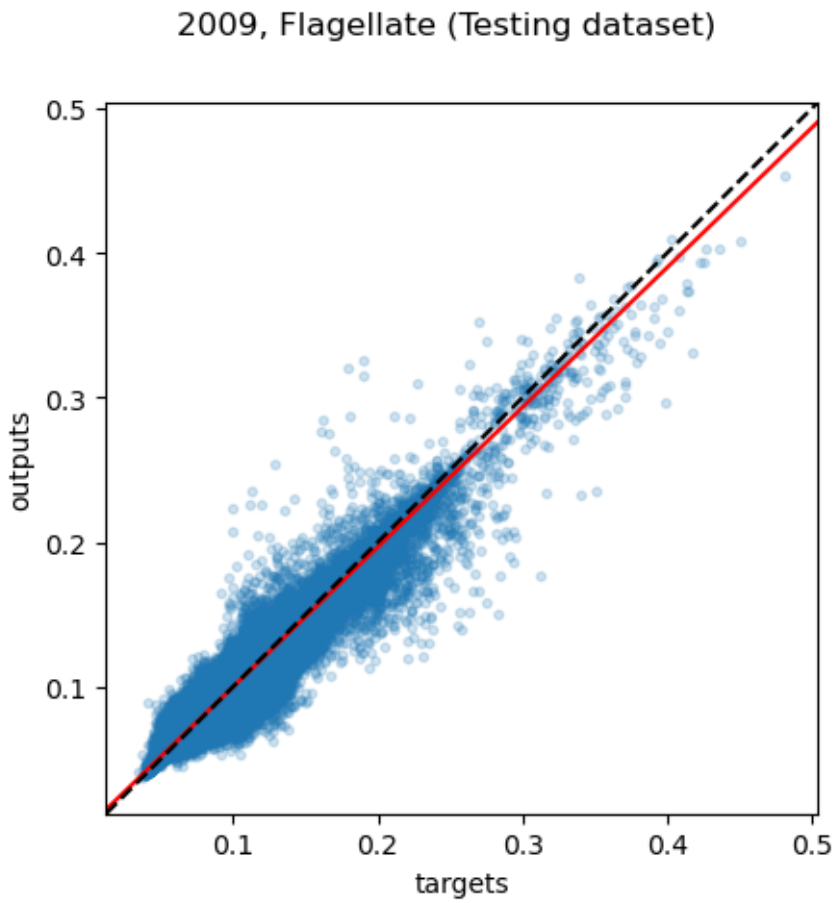

```
The amount of data points is 883101
The slope of the best fitting line is  0.973
The correlation coefficient is: 0.991
 The mean square error is: 2e-05
```



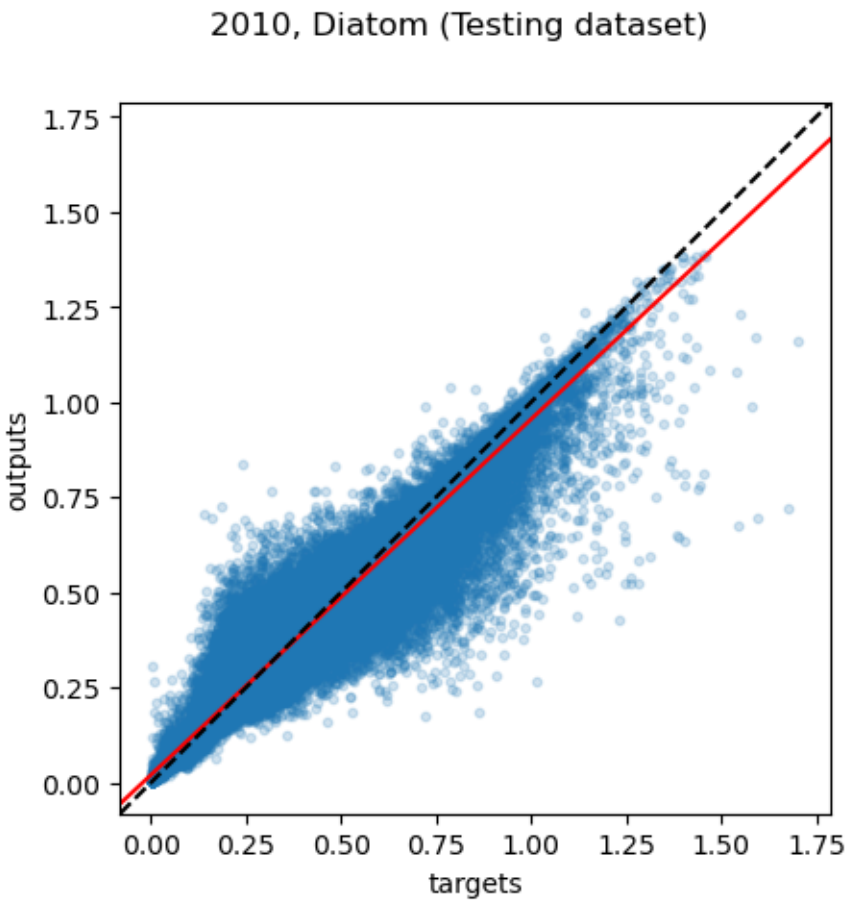

```
 12%|          | 2/17 [15:46<1:58:31, 474.13s/it]
Gathering days for year 2009
100%|      | 75/75 [04:37<00:00,  3.70s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.94
The correlation coefficient is: 0.981
 The mean square error is: 0.0015
```



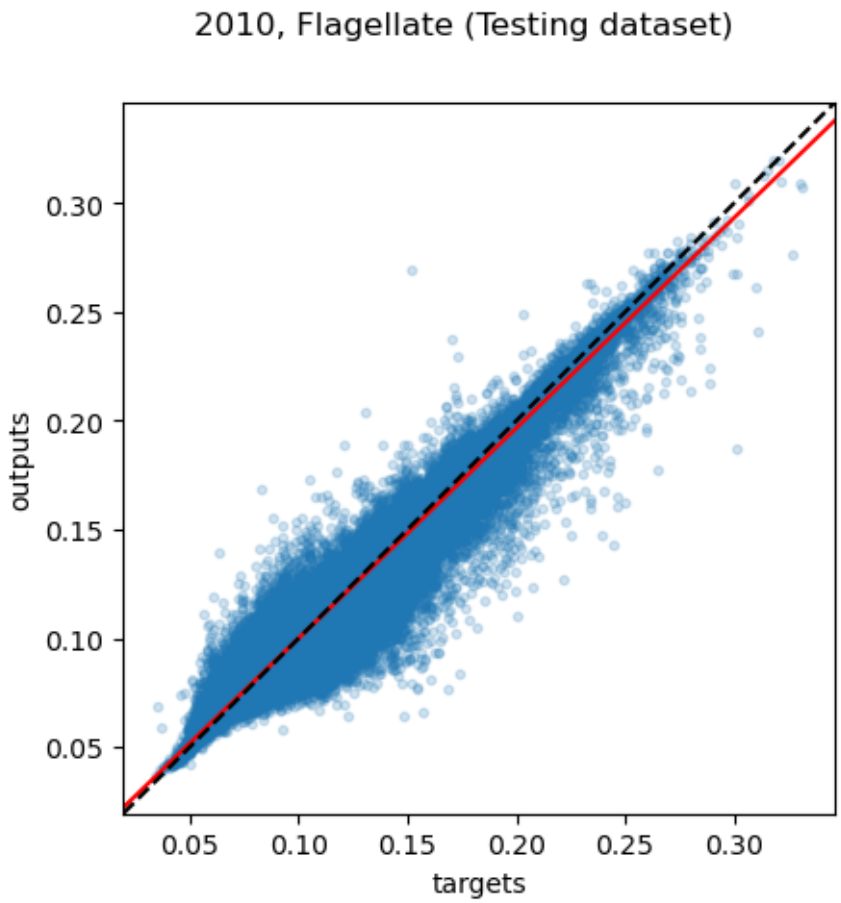

```
The amount of data points is 871482
The slope of the best fitting line is  0.967
The correlation coefficient is: 0.989
 The mean square error is: 2e-05
```



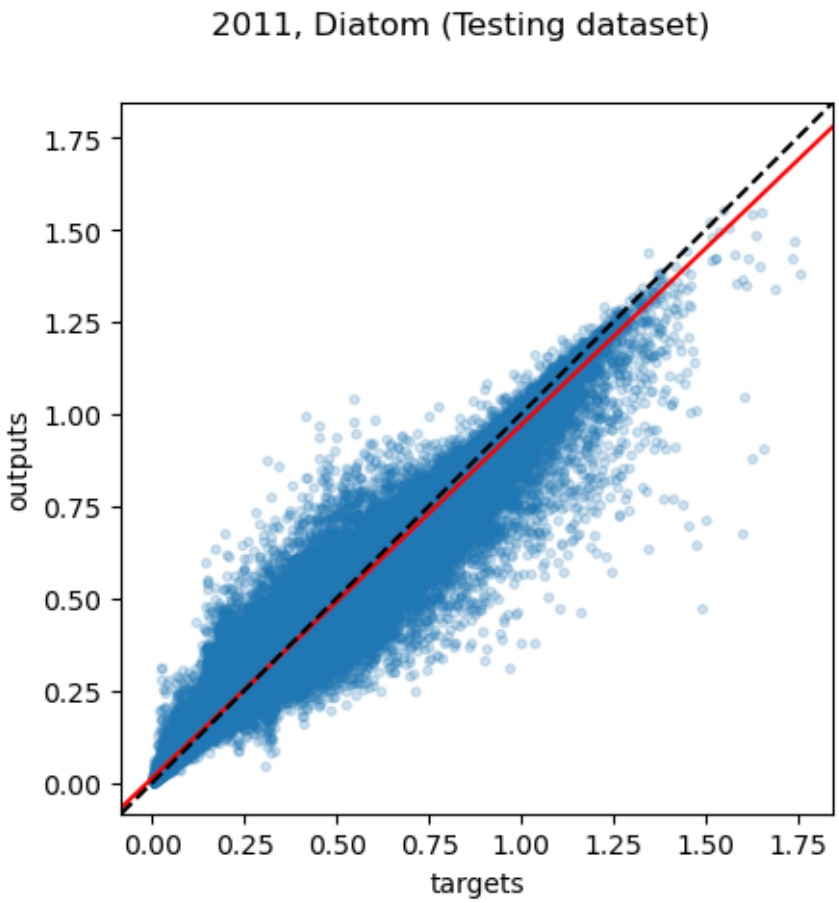

```
 18%|          | 3/17 [24:01<1:52:50, 483.64s/it]
Gathering days for year 2010
100%|      | 75/75 [04:35<00:00,  3.68s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.935
The correlation coefficient is: 0.981
 The mean square error is: 0.00083
```



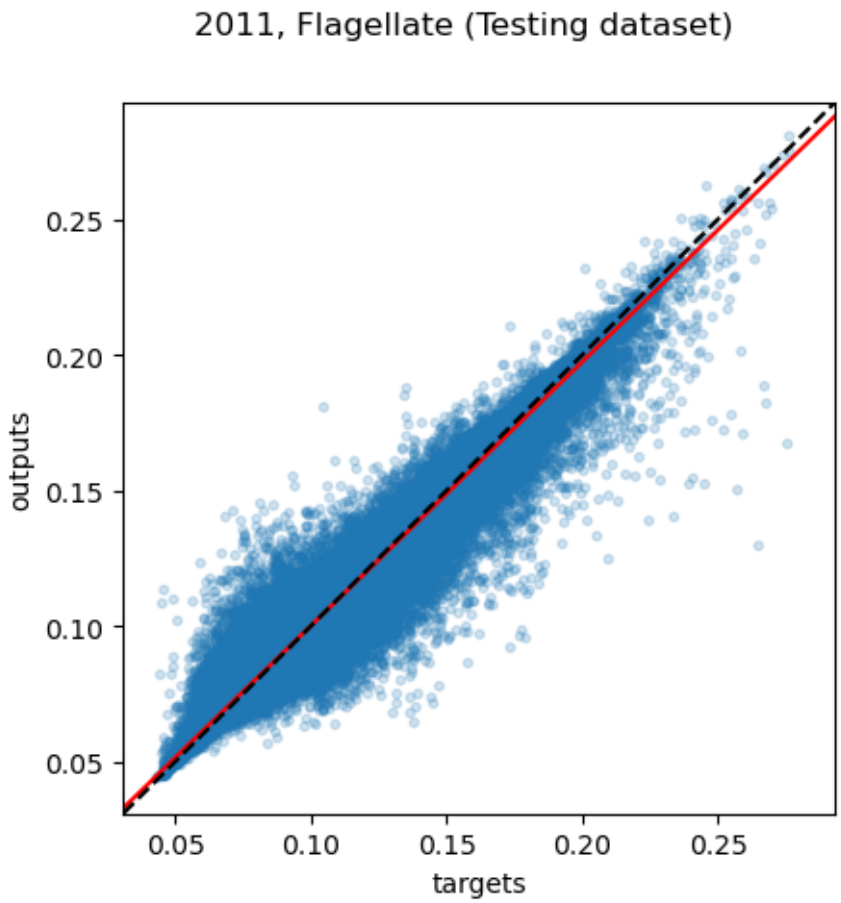

```
The amount of data points is 871482
The slope of the best fitting line is  0.966
The correlation coefficient is: 0.989
 The mean square error is: 2e-05
```



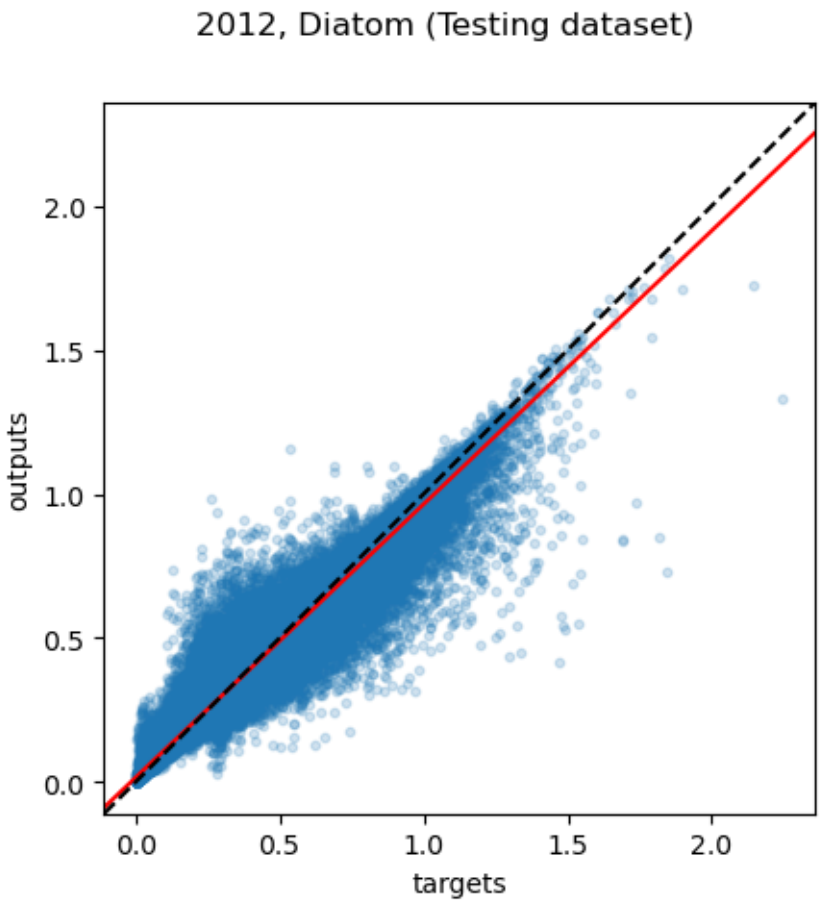

```
 24%|          | 4/17 [32:12<1:45:25, 486.58s/it]
Gathering days for year 2011
100%|      | 75/75 [04:27<00:00,  3.57s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.958
The correlation coefficient is: 0.987
 The mean square error is: 0.00063
```



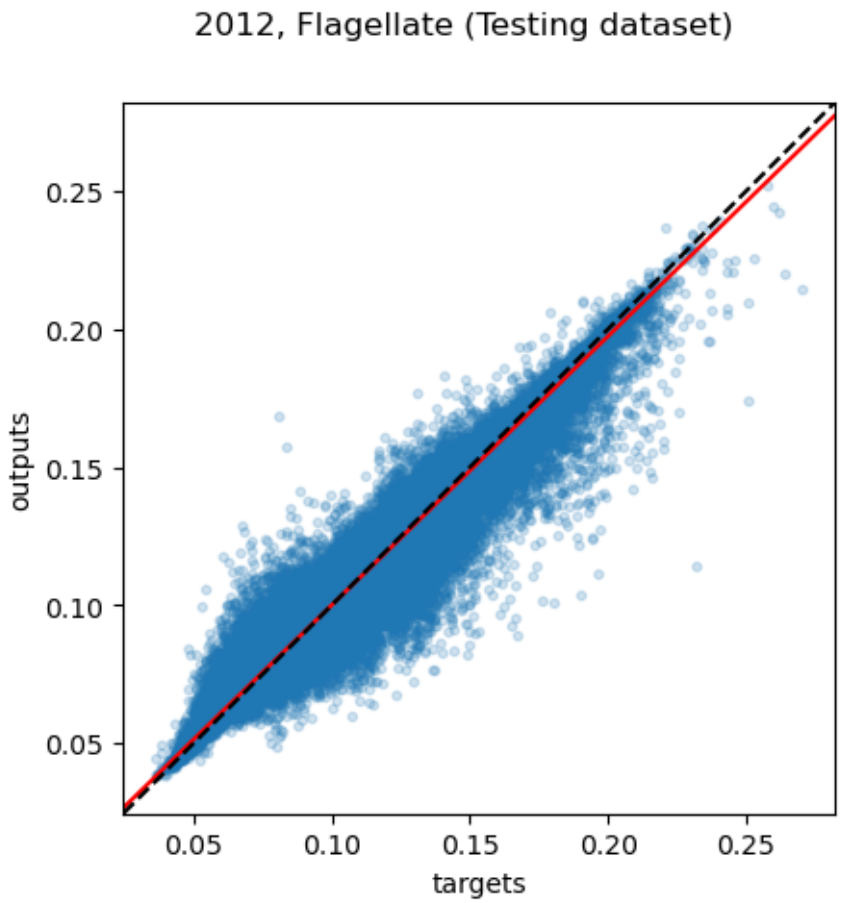

```
The amount of data points is 871482
The slope of the best fitting line is  0.973
The correlation coefficient is: 0.991
 The mean square error is: 1e-05
```



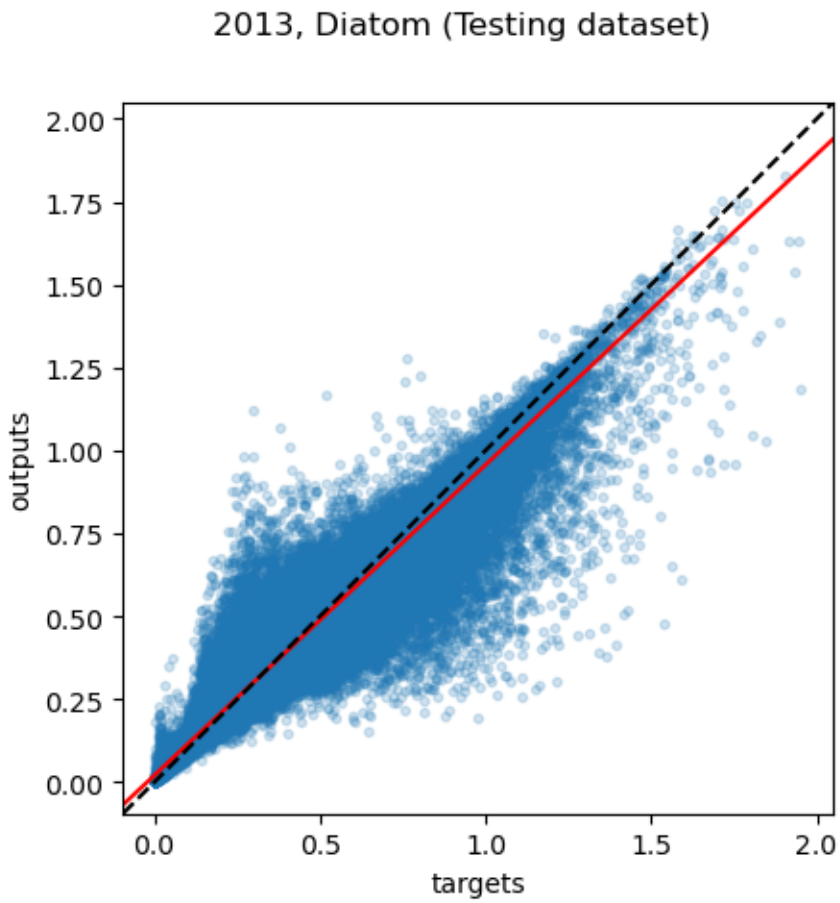

```
 29%|          | 5/17 [40:20<1:37:23, 486.98s/it]
Gathering days for year 2012
100%|      | 76/76 [04:31<00:00,  3.58s/it]
Done gathering, building the prediction models


The amount of data points is 883101
The slope of the best fitting line is  0.949
The correlation coefficient is: 0.984
 The mean square error is: 0.00078
```



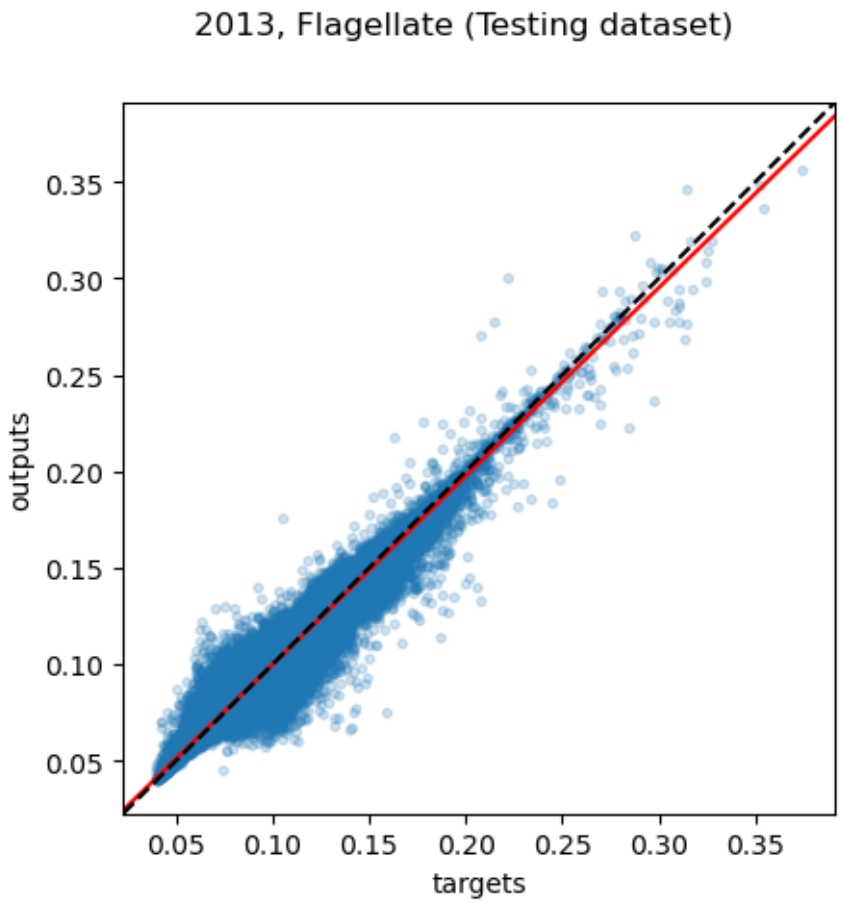

```
The amount of data points is 883101
The slope of the best fitting line is  0.974
The correlation coefficient is: 0.992
 The mean square error is: 1e-05
```



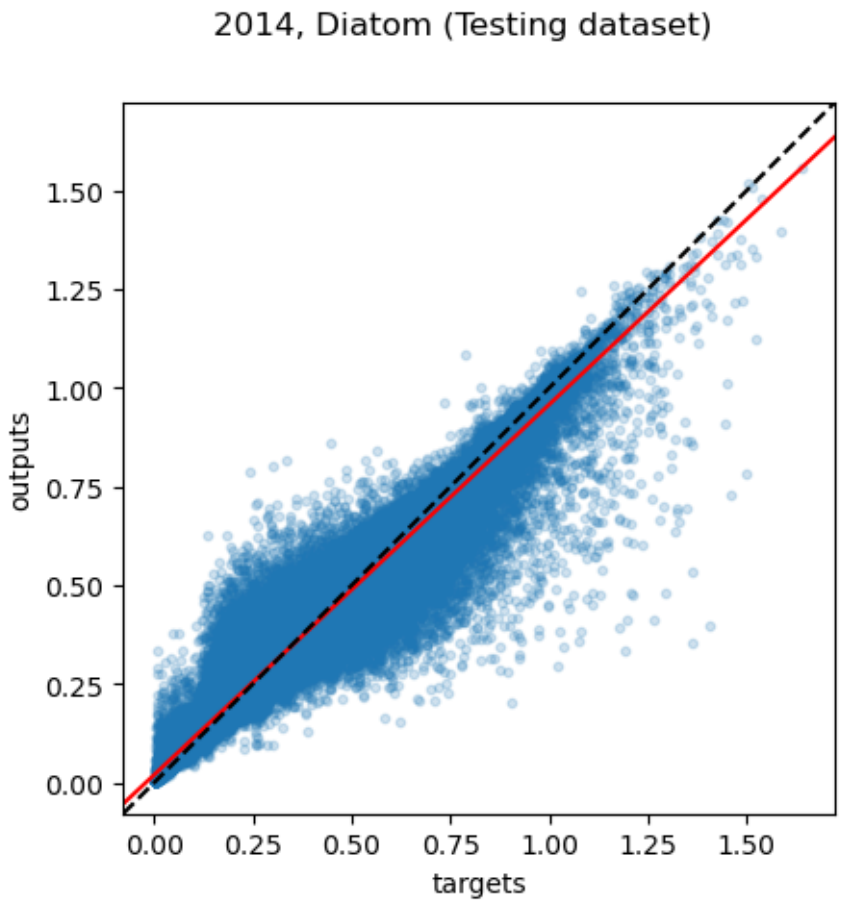

```
 35%|         | 6/17 [48:29<1:29:25, 487.80s/it]
Gathering days for year 2013
100%|      | 75/75 [04:27<00:00,  3.57s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.936
The correlation coefficient is: 0.98
 The mean square error is: 0.00127
```



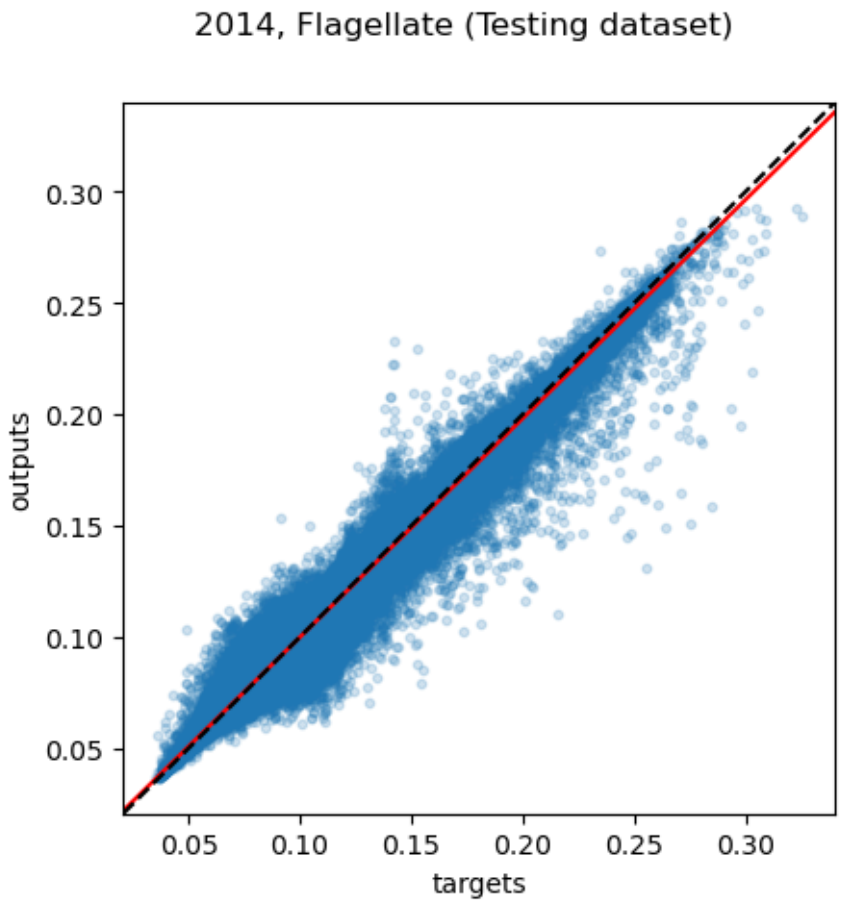

```
The amount of data points is 871482
The slope of the best fitting line is  0.976
The correlation coefficient is: 0.992
 The mean square error is: 1e-05
```



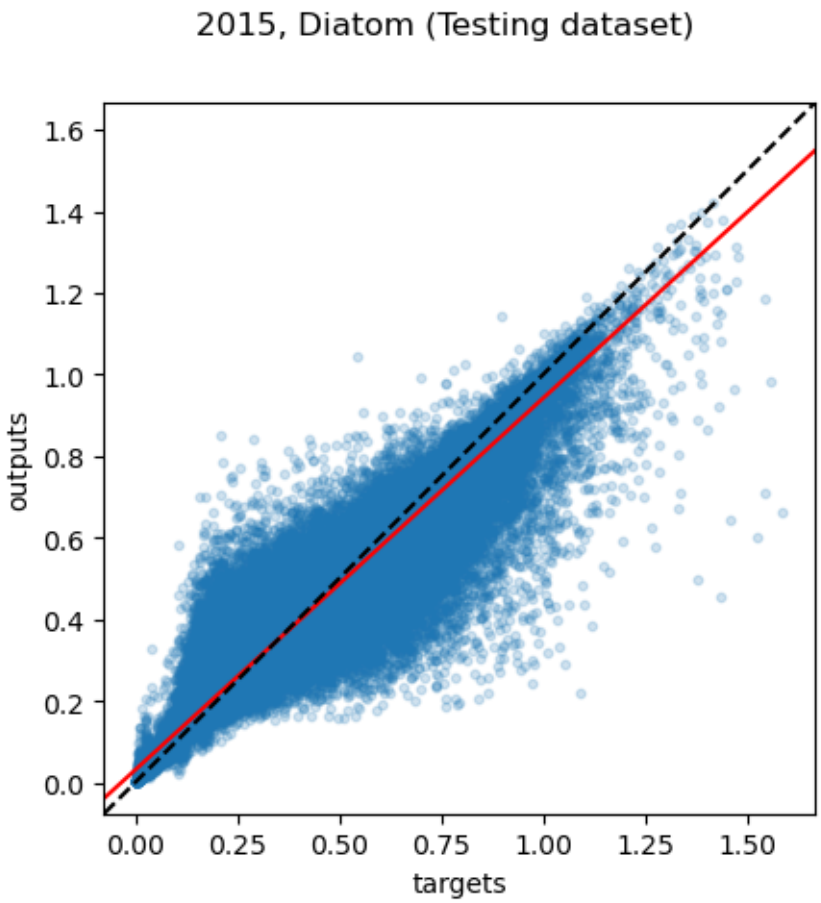

```
 41%|          | 7/17 [56:33<1:21:02, 486.25s/it]
Gathering days for year 2014
100%|      | 75/75 [04:29<00:00,  3.59s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.938
The correlation coefficient is: 0.981
 The mean square error is: 0.00079
```



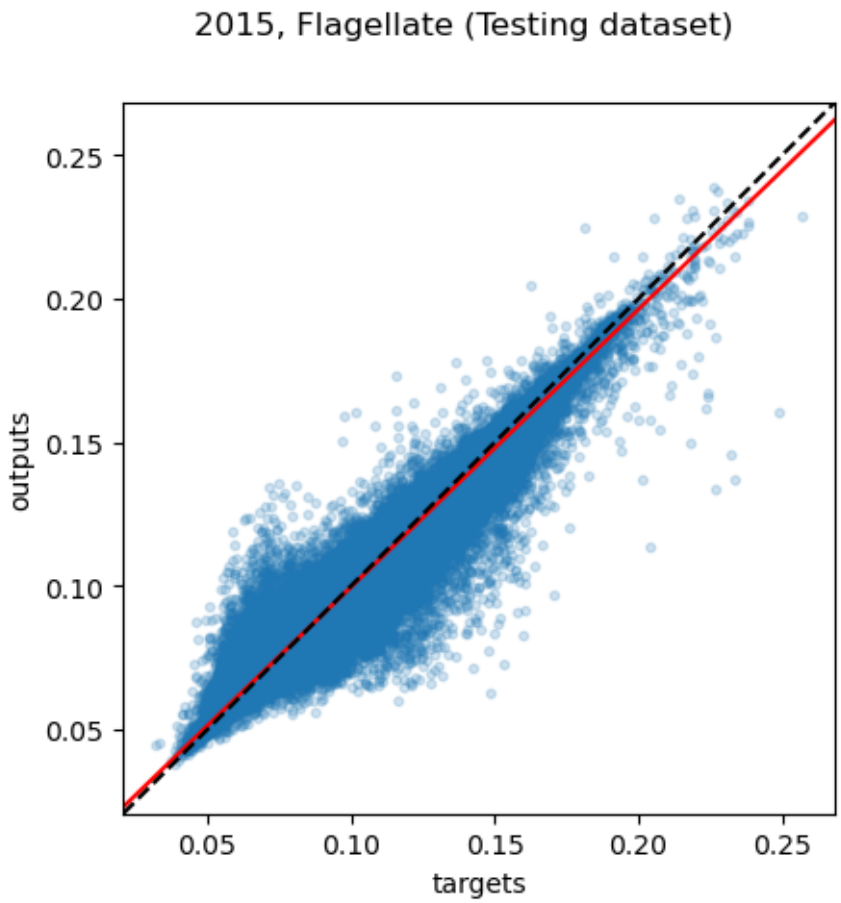

```
The amount of data points is 871482
The slope of the best fitting line is  0.982
The correlation coefficient is: 0.994
 The mean square error is: 1e-05
```



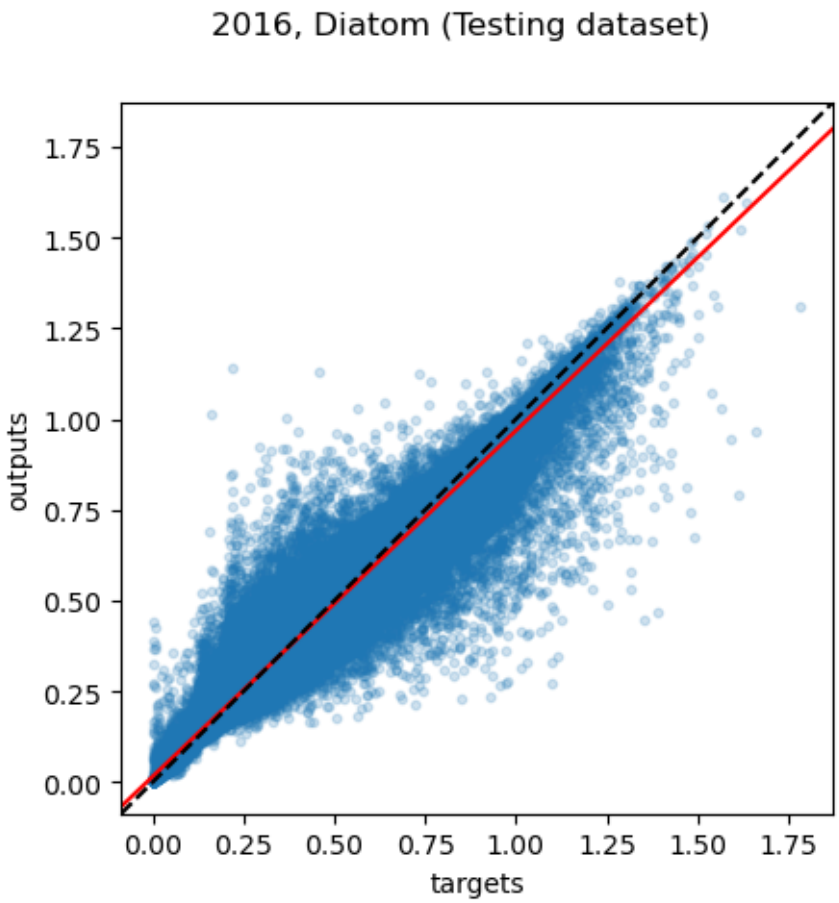

```
 47%|         | 8/17 [1:04:36<1:12:49, 485.45s/it]
Gathering days for year 2015
100%|      | 75/75 [04:25<00:00,  3.54s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.911
The correlation coefficient is: 0.972
 The mean square error is: 0.00132
```



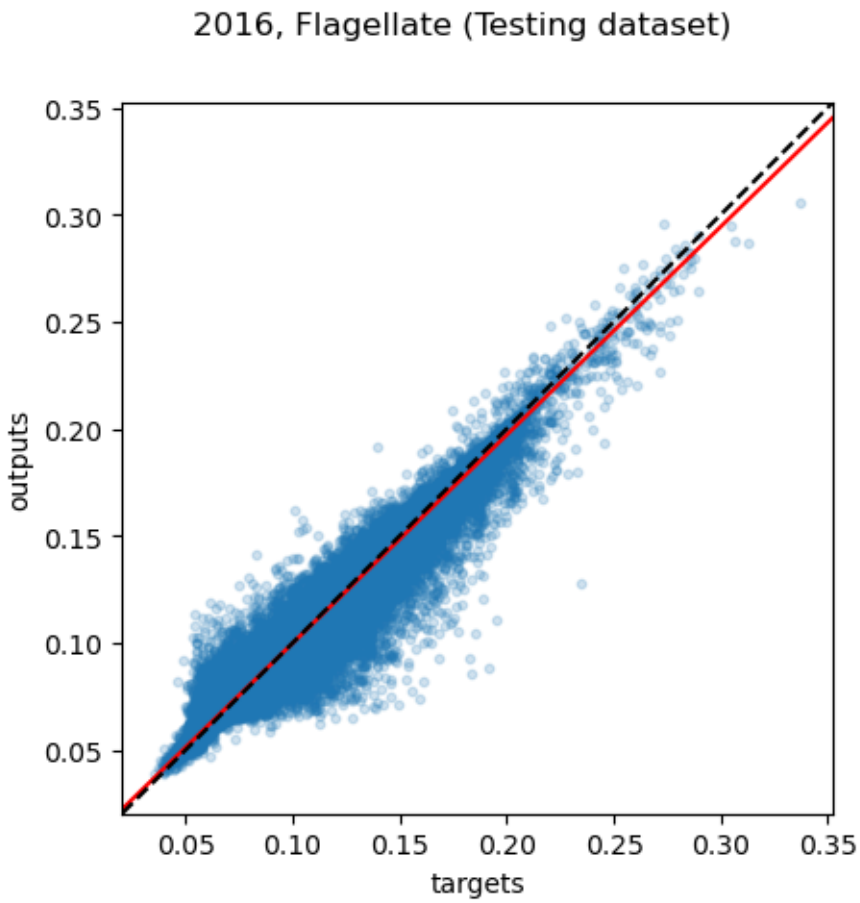

```
The amount of data points is 871482
The slope of the best fitting line is  0.967
The correlation coefficient is: 0.99
 The mean square error is: 1e-05
```



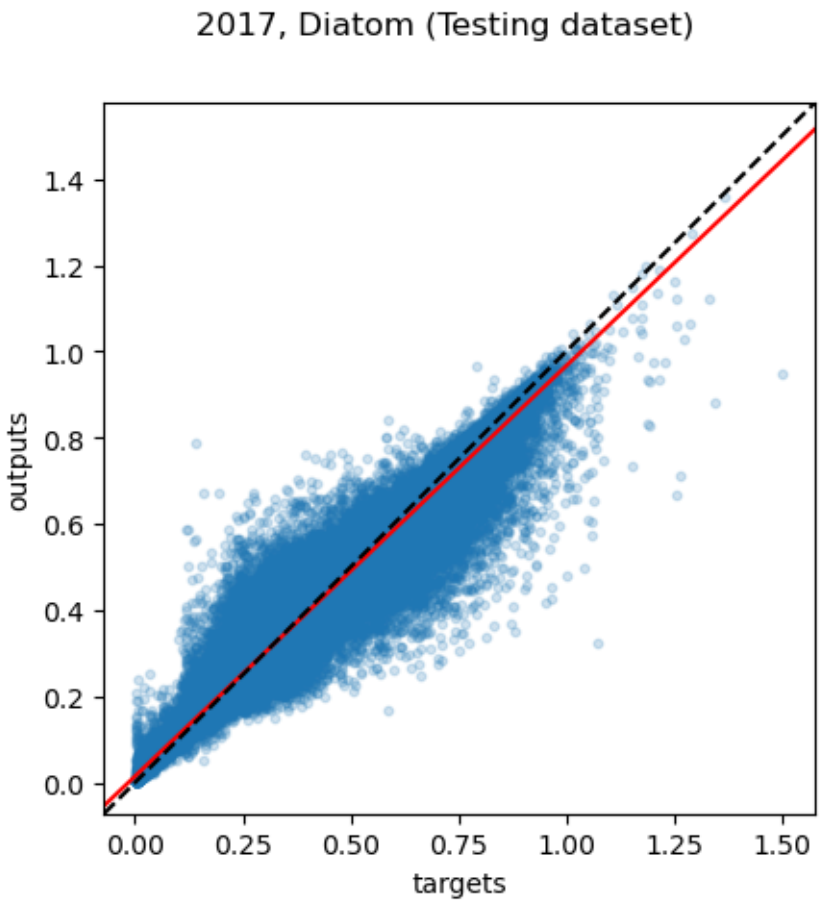

```
 53%|        | 9/17 [1:12:44<1:04:48, 486.03s/it]
Gathering days for year 2016
100%|      | 76/76 [04:34<00:00,  3.61s/it]
Done gathering, building the prediction models


The amount of data points is 883101
The slope of the best fitting line is  0.953
The correlation coefficient is: 0.985
 The mean square error is: 0.00085
```



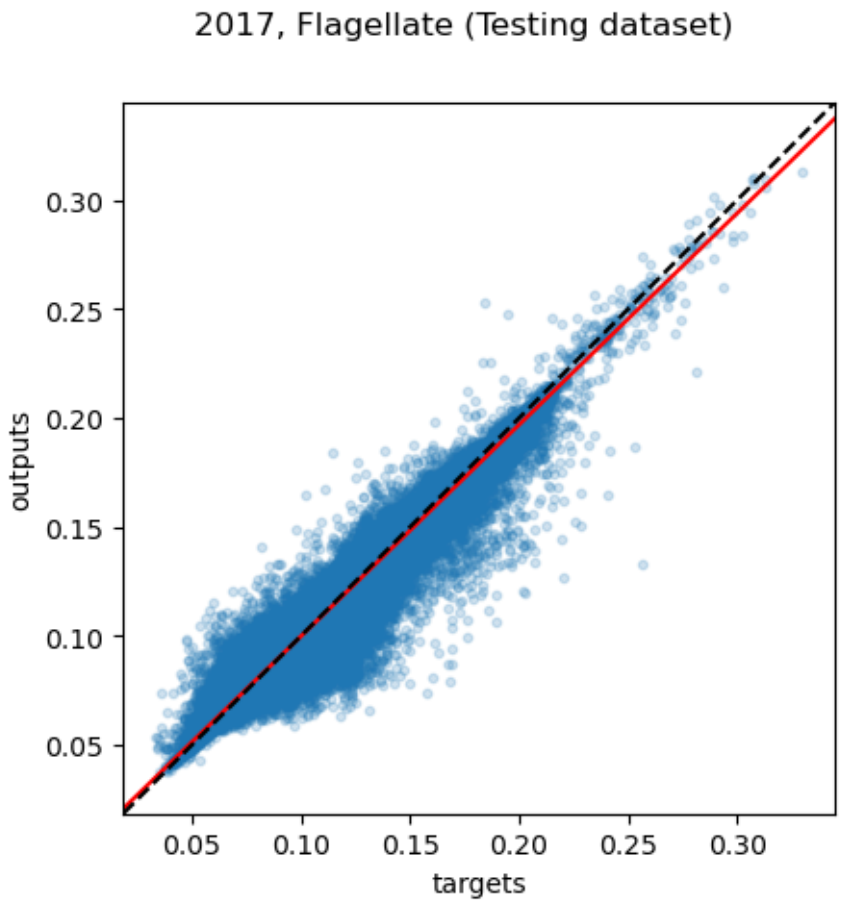

```
The amount of data points is 883101
The slope of the best fitting line is  0.973
The correlation coefficient is: 0.991
 The mean square error is: 1e-05
```



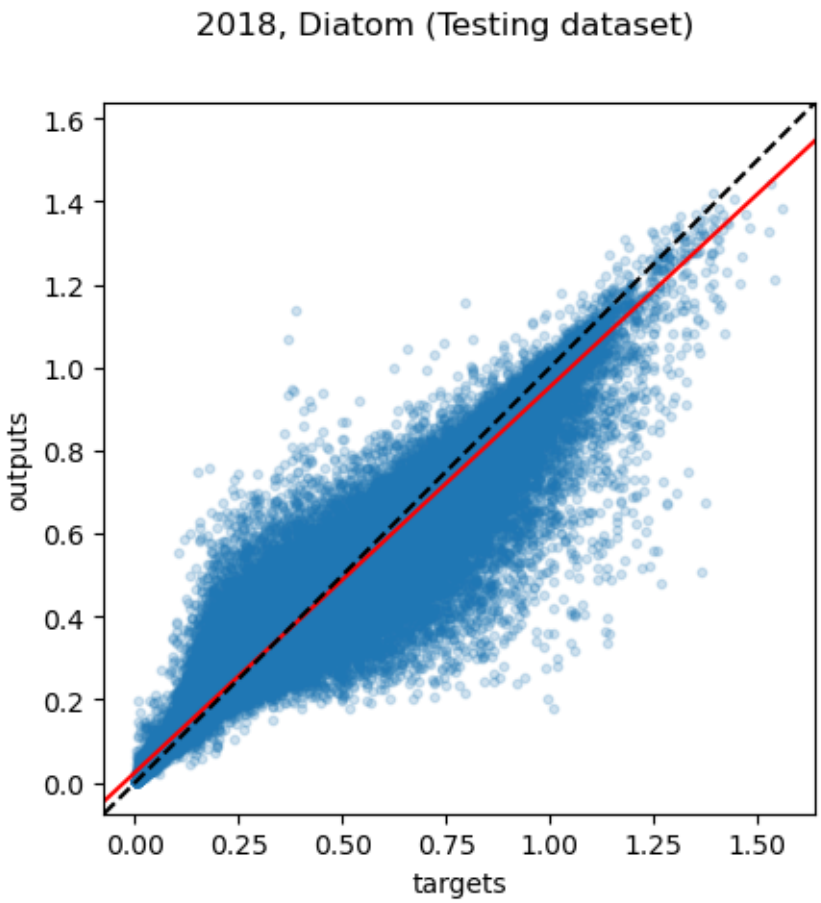

```
 59%|        | 10/17 [1:20:58<57:01, 488.75s/it]
Gathering days for year 2017
100%|      | 75/75 [04:27<00:00,  3.57s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.952
The correlation coefficient is: 0.985
 The mean square error is: 0.00053
```



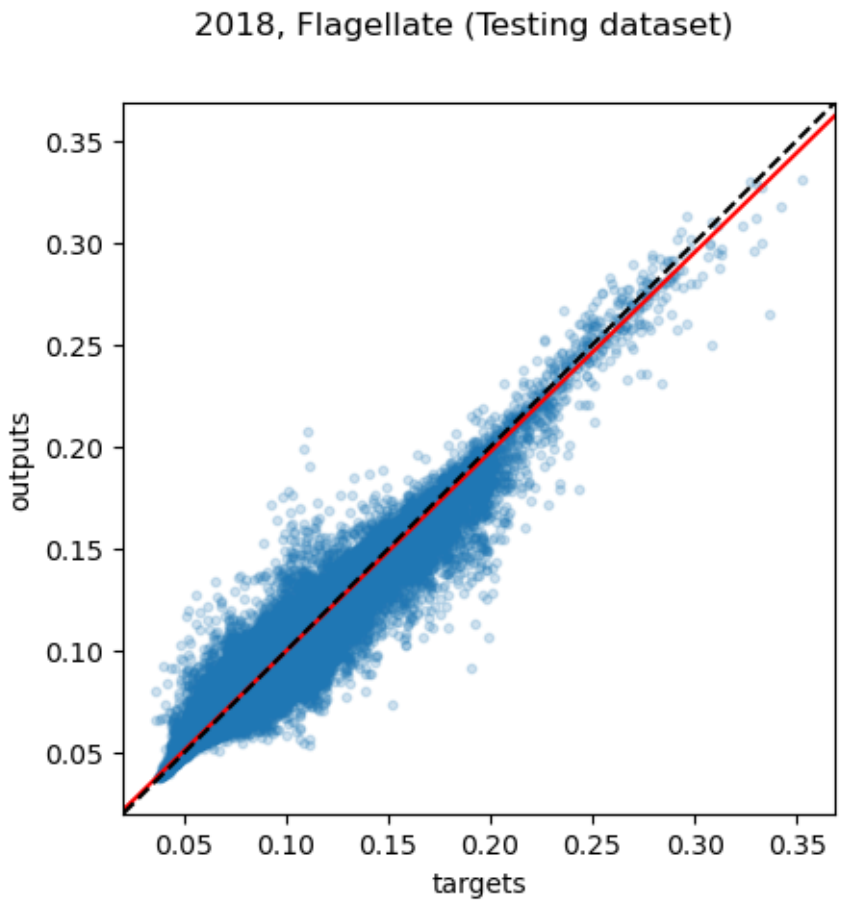

```
The amount of data points is 871482
The slope of the best fitting line is  0.971
The correlation coefficient is: 0.991
 The mean square error is: 1e-05
```



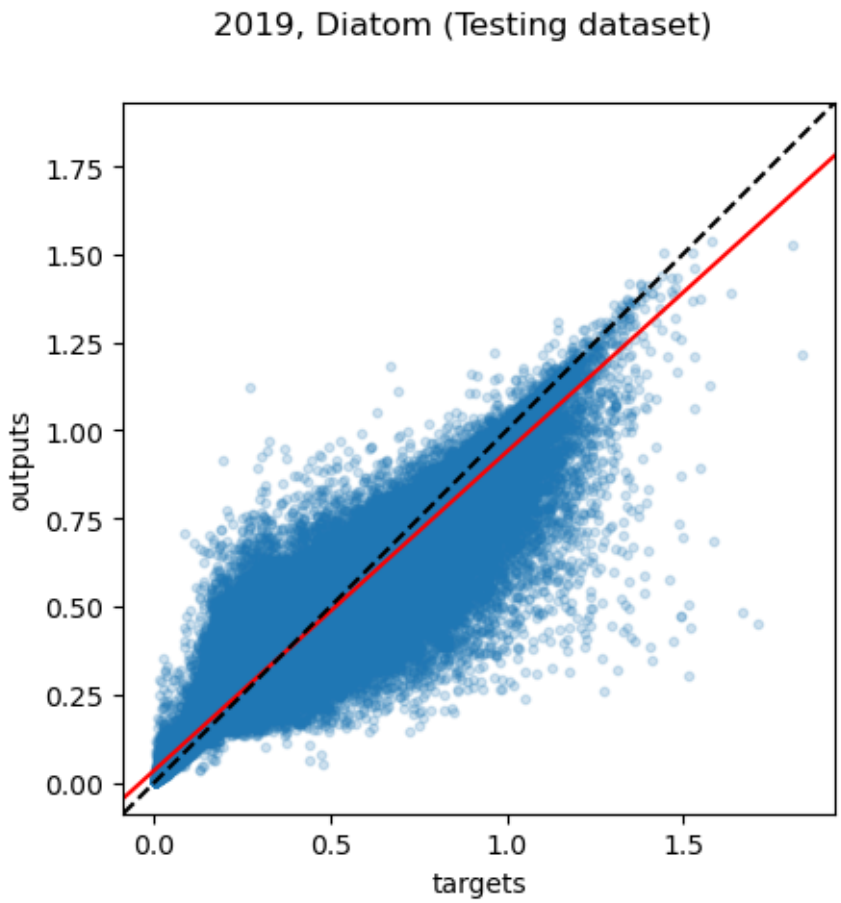

```
 65%|       | 11/17 [1:29:04<48:45, 487.66s/it]
Gathering days for year 2018
100%|      | 75/75 [04:25<00:00,  3.54s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.929
The correlation coefficient is: 0.977
 The mean square error is: 0.00113
```



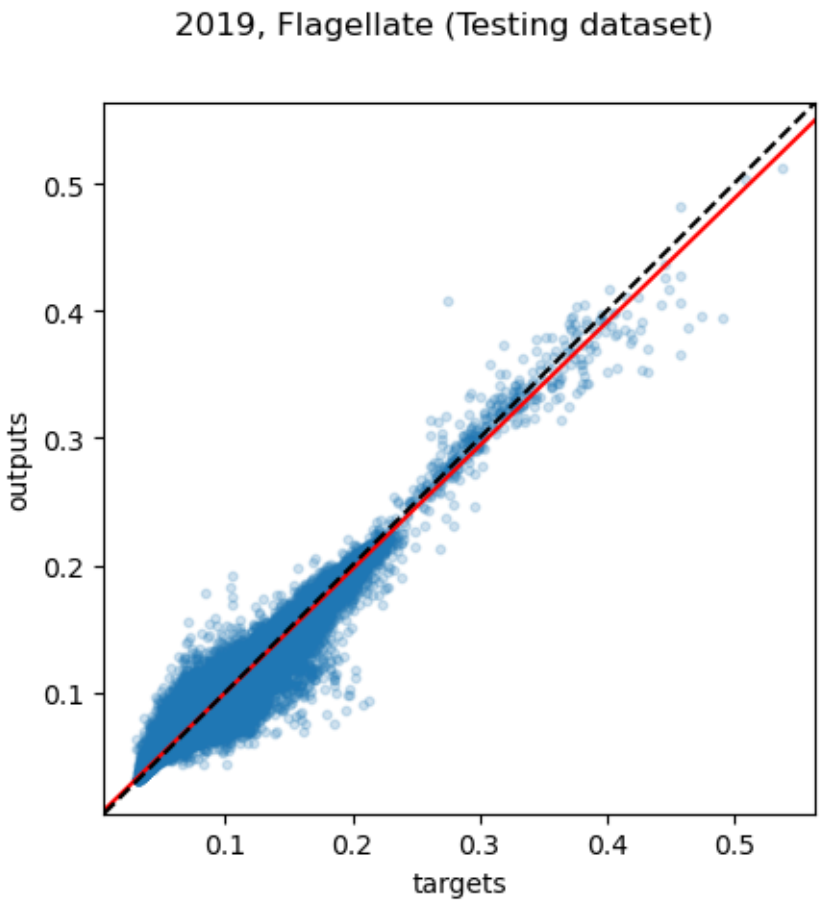

```
The amount of data points is 871482
The slope of the best fitting line is  0.976
The correlation coefficient is: 0.993
 The mean square error is: 1e-05
```



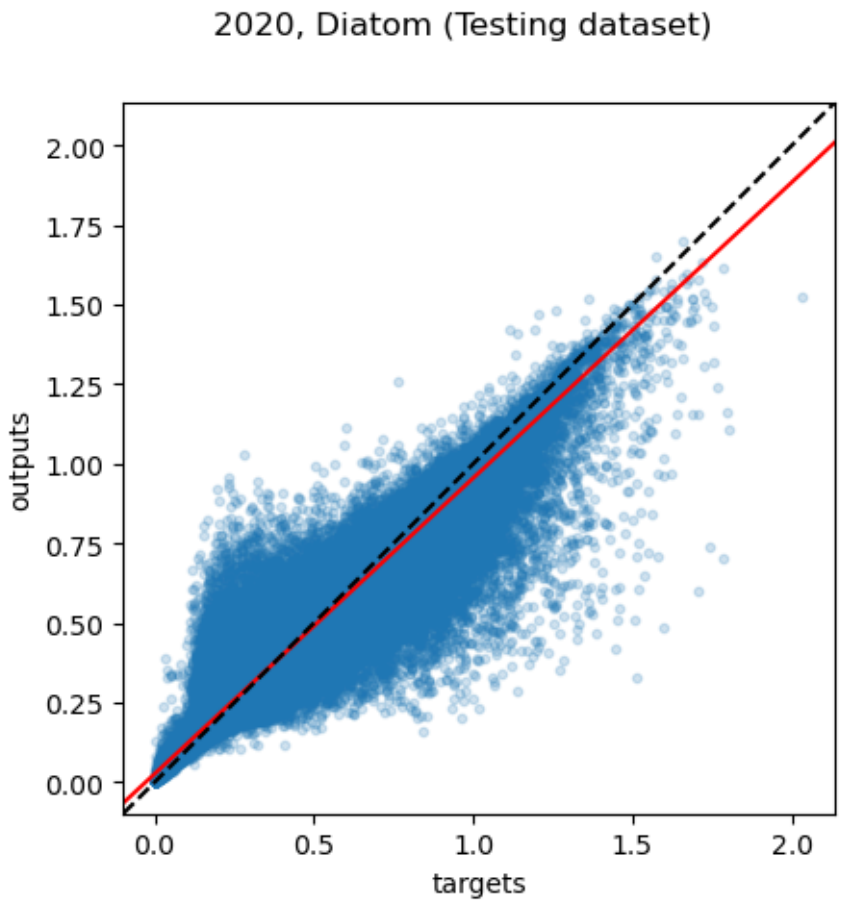

```
 71%|       | 12/17 [1:37:10<40:36, 487.39s/it]
Gathering days for year 2019
100%|      | 75/75 [04:28<00:00,  3.57s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.904
The correlation coefficient is: 0.967
 The mean square error is: 0.00201
```



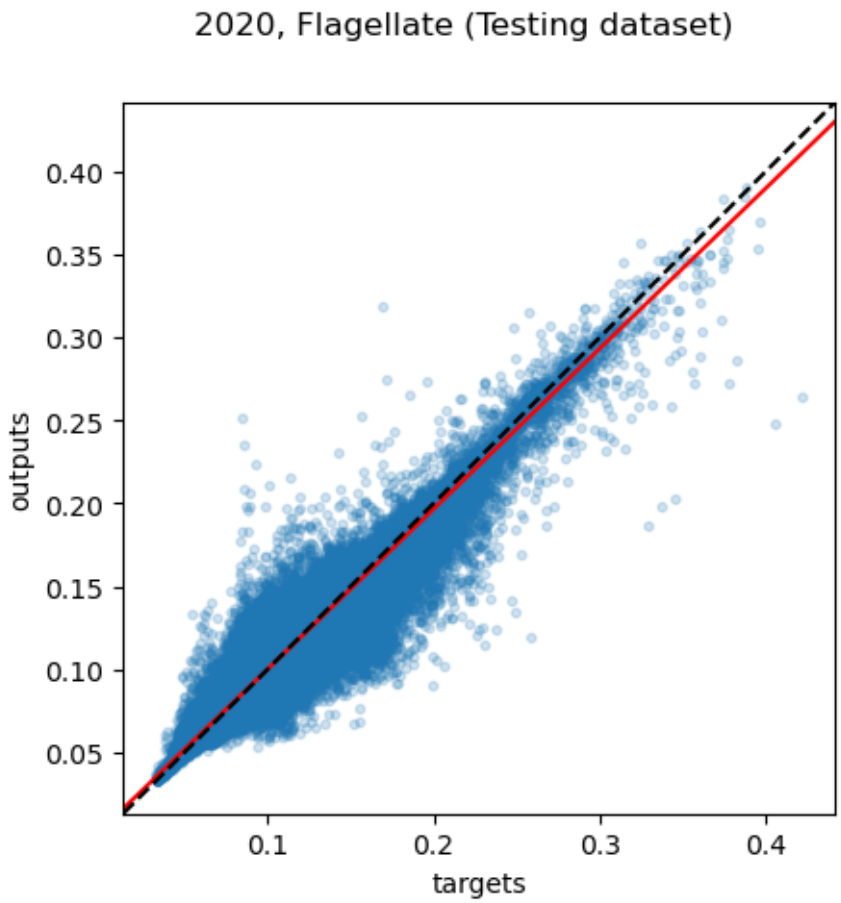

```
The amount of data points is 871482
The slope of the best fitting line is  0.971
The correlation coefficient is: 0.99
 The mean square error is: 1e-05
```



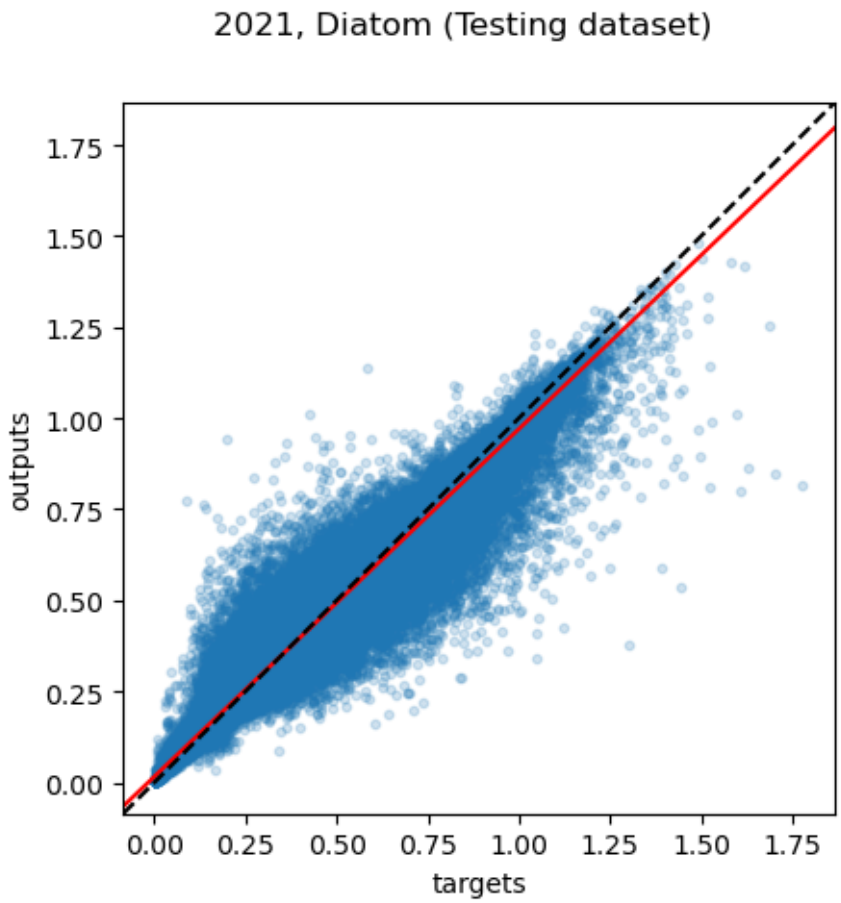

```
 76%|       | 13/17 [1:45:17<32:29, 487.29s/it]
Gathering days for year 2020
100%|      | 76/76 [04:27<00:00,  3.52s/it]
Done gathering, building the prediction models


The amount of data points is 883101
The slope of the best fitting line is  0.929
The correlation coefficient is: 0.977
 The mean square error is: 0.00187
```



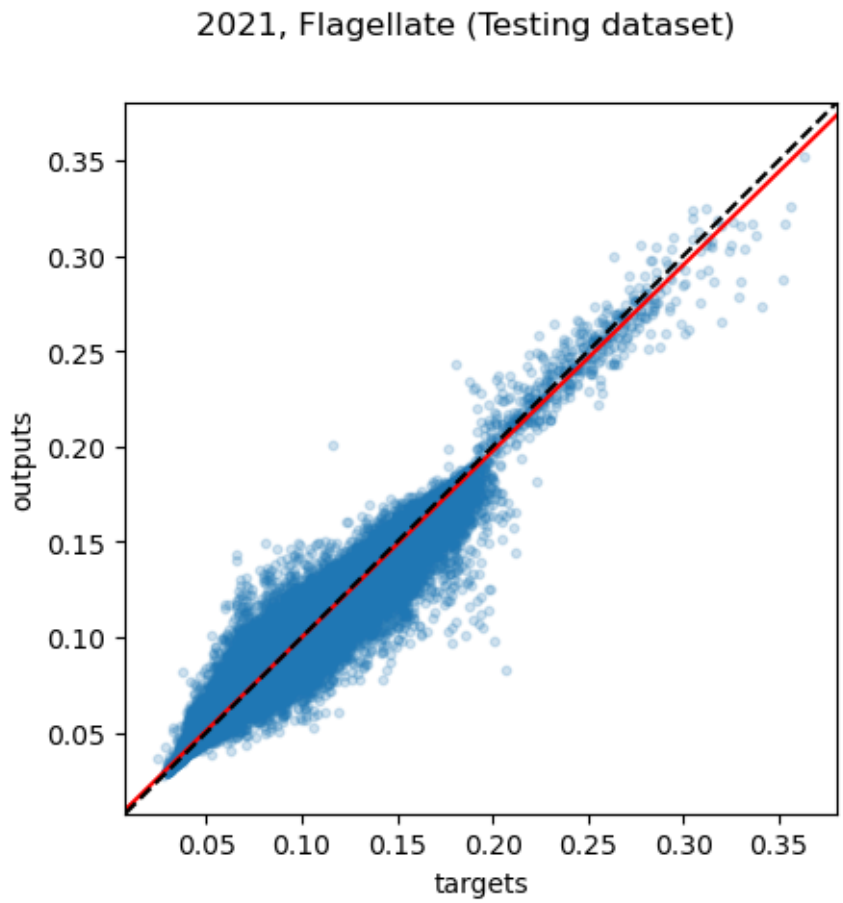

```
The amount of data points is 883101
The slope of the best fitting line is  0.966
The correlation coefficient is: 0.989
 The mean square error is: 3e-05
```



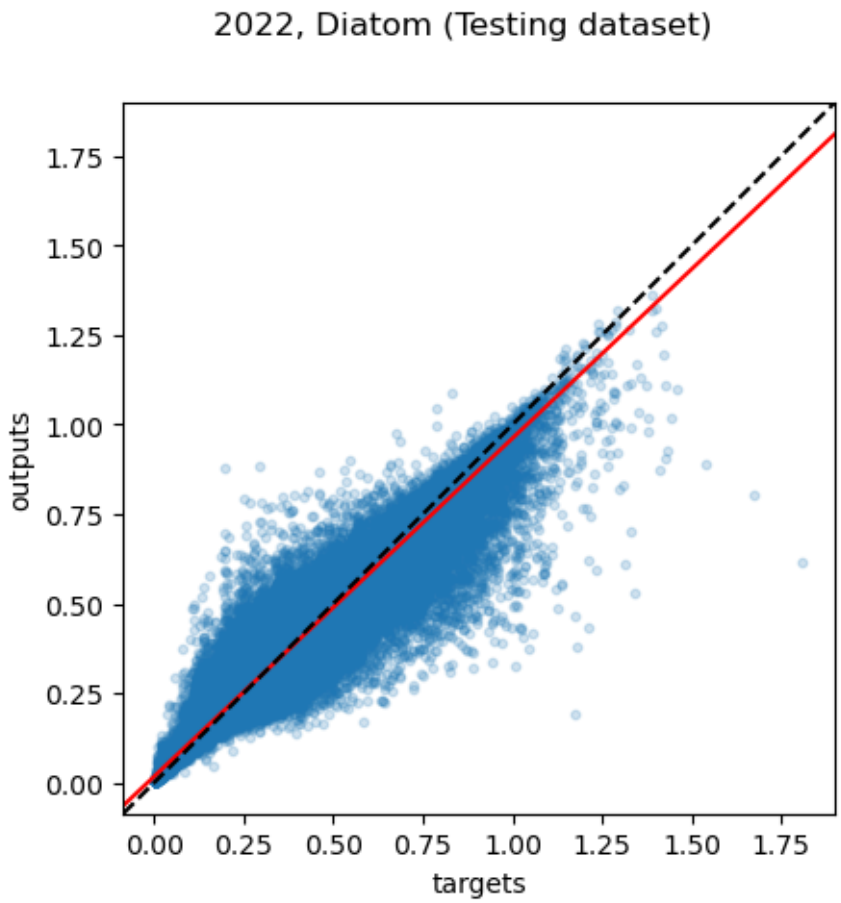

```
 82%|      | 14/17 [1:53:26<24:23, 487.79s/it]
Gathering days for year 2021
100%|      | 75/75 [04:24<00:00,  3.53s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.954
The correlation coefficient is: 0.986
 The mean square error is: 0.00089
```



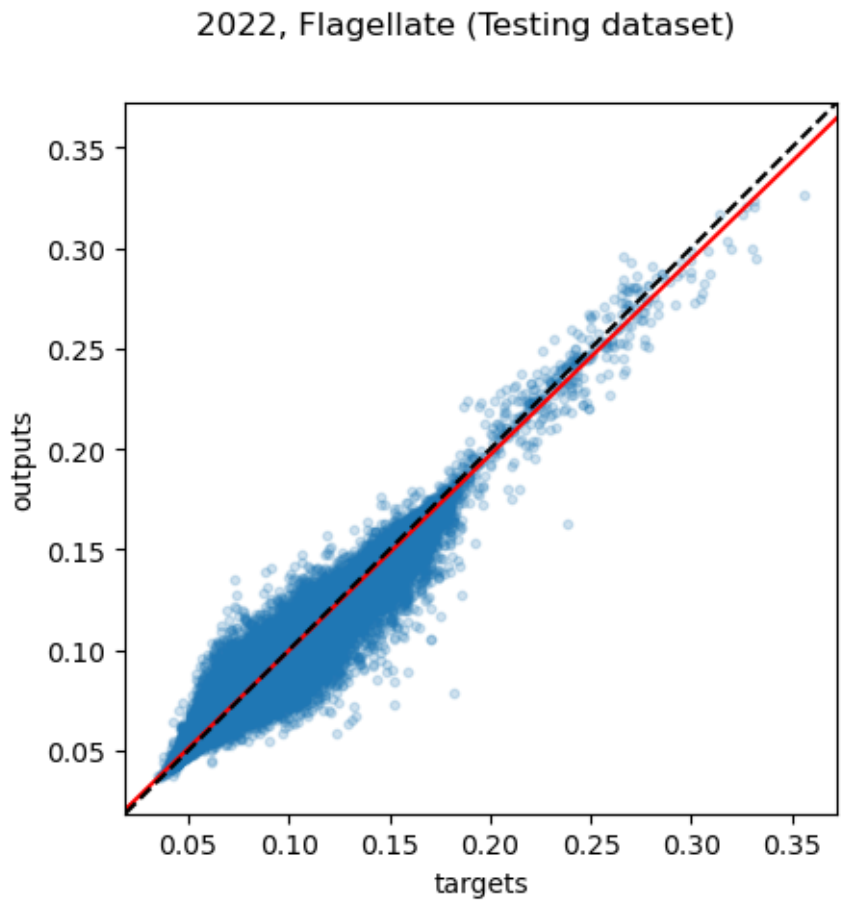

```
The amount of data points is 871482
The slope of the best fitting line is  0.977
The correlation coefficient is: 0.993
 The mean square error is: 1e-05
```



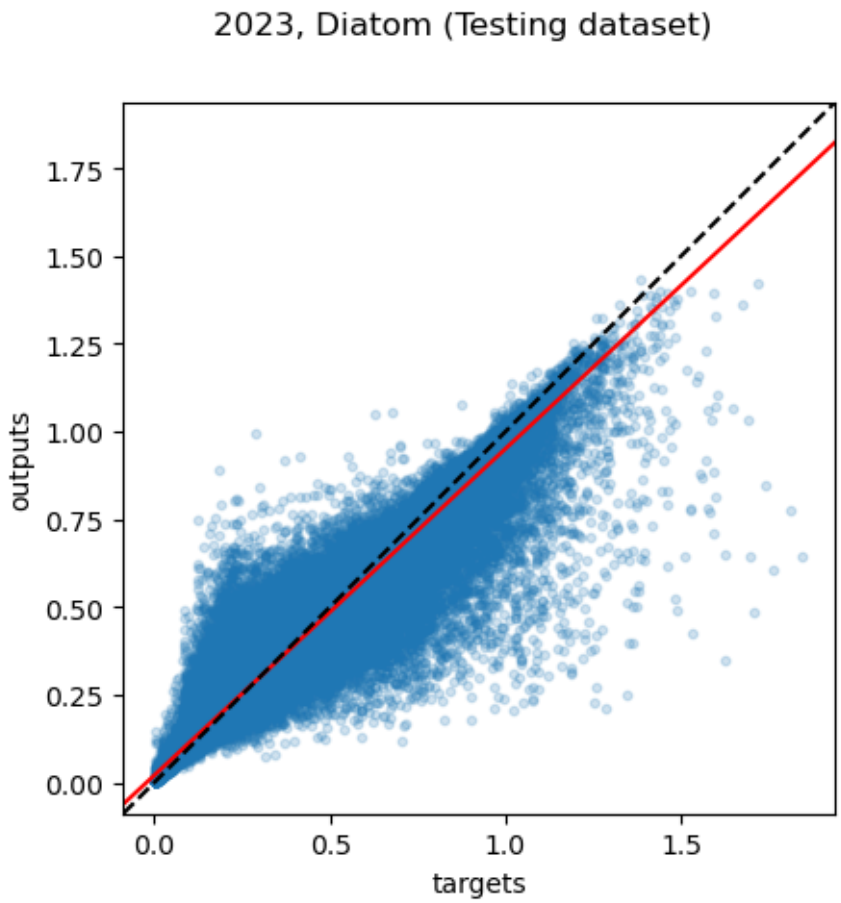

```
 88%|      | 15/17 [2:01:28<16:12, 486.06s/it]
Gathering days for year 2022
100%|      | 75/75 [04:25<00:00,  3.53s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.945
The correlation coefficient is: 0.983
 The mean square error is: 0.00074
```



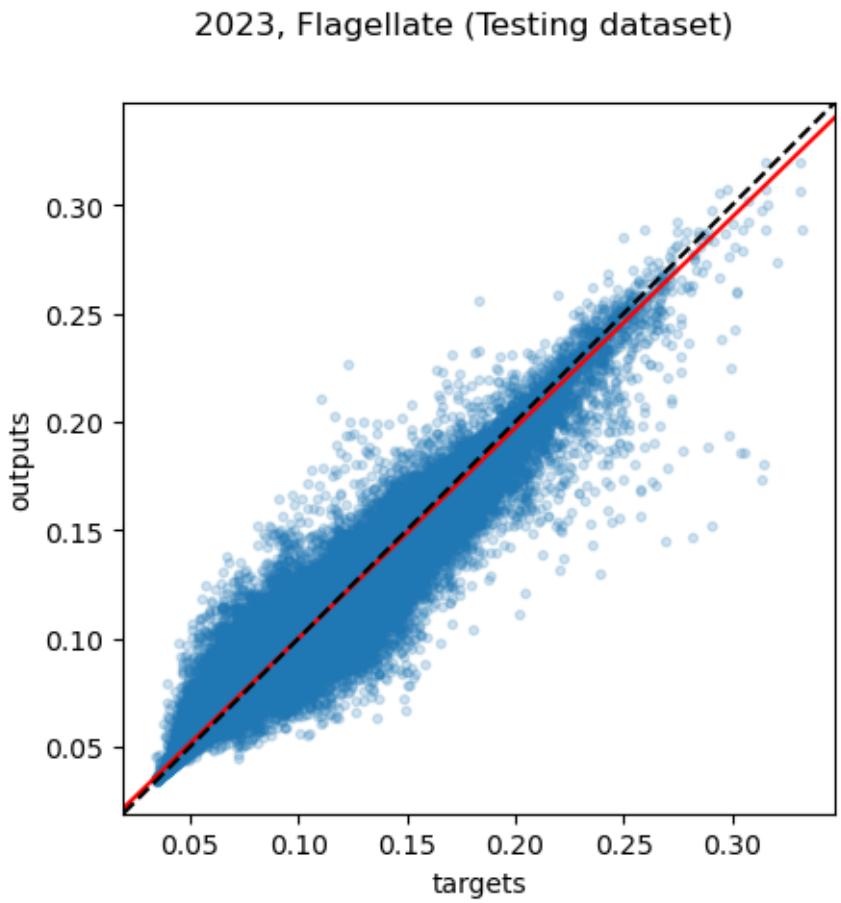

```
The amount of data points is 871482
The slope of the best fitting line is  0.973
The correlation coefficient is: 0.991
 The mean square error is: 1e-05
```



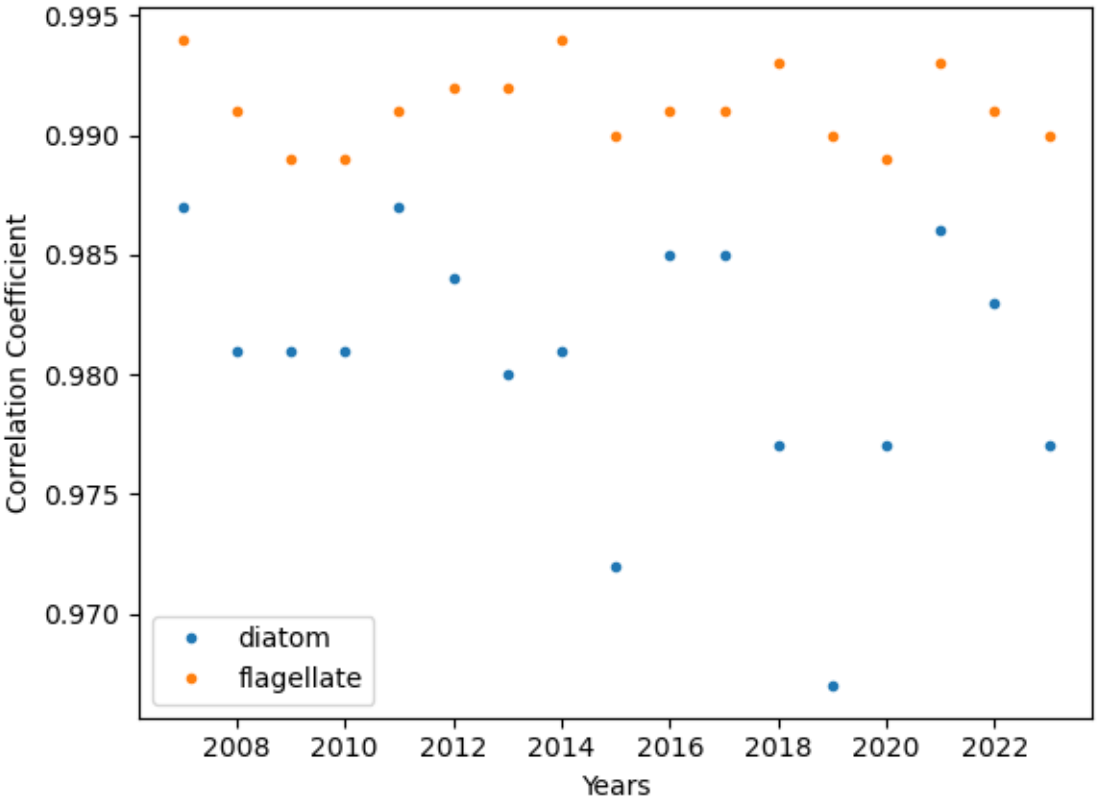

```
 94%|      | 16/17 [2:09:25<08:03, 483.15s/it]
Gathering days for year 2023
100%|      | 75/75 [04:31<00:00,  3.62s/it]
Done gathering, building the prediction models


The amount of data points is 871482
The slope of the best fitting line is  0.931
The correlation coefficient is: 0.977
 The mean square error is: 0.00122
```

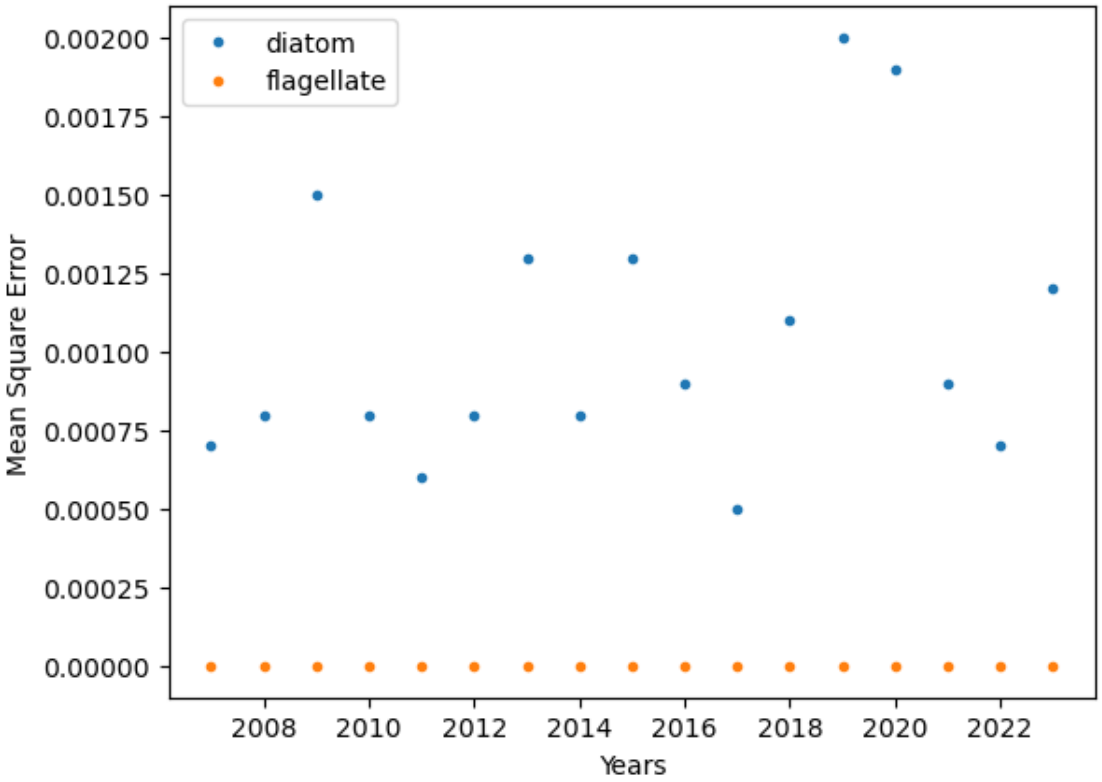


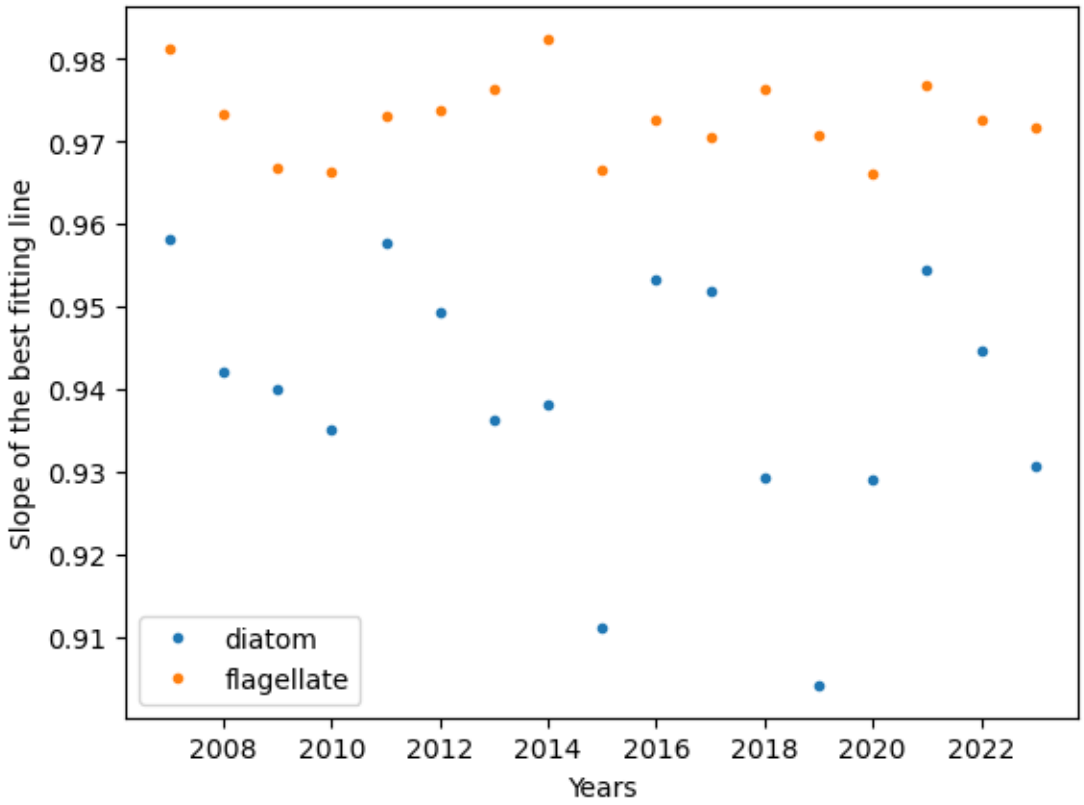

```
The amount of data points is 871482
The slope of the best fitting line is  0.972
The correlation coefficient is: 0.99
 The mean square error is: 2e-05
```



100%| | 17/17 [2:17:28<00:00, 485.23s/it]


0.995
0.990
0.985
0.980
0.975
0.970
Correlation Coefficient
diatom
flagellate
2008
2010
2012
2014
2016
2018
2020
2022
Years


diatom
flagellate
0.00200
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
Mean Square Error
2008
2010
2012
2014
2016
2018
2020
2022
Years


0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
Slope of the best fitting line
diatom
flagellate
2008
2010
2012
2014
2016
2018
2020
2022
Years

[ ]: