In [14]:
```
from pyspark.sql.functions import udf
df = spark.read.load('hdfs://orion11:11001/Project3/part-0*'
                     , format='csv', sep='\t'
                     , inferSchema='true'
                      , header='true')
```

In [15]:
```
import pygeohash as pgh
import pyspark.sql.functions as F
geohashEncodeUDF = F.udf(lambda x, y: pgh.encode(x, y))

df = df.withColumnRenamed('1_time', 'time').withColumnRenamed('2_lat', 'lat').w
df = df.withColumn('geohash', geohashEncodeUDF(df['lat'], df['lon']))
df.take(3)
```

Out[15]:
```
[Row(time=1455440400000, lat=28.862712472612284, lon=-80.15570444411433, albed
o_surface='6.0', precipitable_water_entire_atmosphere_single_layer='null', pre
ssure_maximum_wind='17126.896', pressure_surface='102390.0', pressure_tropopau
se='23103.373', relative_humidity_zerodegc_isotherm='35.0', snow_depth_surface
='0.0', temperature_surface='295.9439', temperature_tropopause='219.42467', to
tal_cloud_cover_entire_atmosphere_single_layer='null', total_precipitation_sur
face_3_hour_accumulation='0.0', vegetation_surface='0.0', visibility_surface
='24223.668', wilting_point_surface='0.0', wind_speed_gust_surface='null', geo
hash='djph0n23kxwf'),
 Row(time=1455440400000, lat=57.697196193266976, lon=-79.3345809744617, albedo
_surface='65.0', precipitable_water_entire_atmosphere_single_layer='null', pre
ssure_maximum_wind='11126.896', pressure_surface='101925.0', pressure_tropopau
se='30303.373', relative_humidity_zerodegc_isotherm='76.0', snow_depth_surface
='0.049999997', temperature_surface='243.69392', temperature_tropopause='213.0
4967', total_cloud_cover_entire_atmosphere_single_layer='null', total_precipit
ation_surface_3_hour_accumulation='0.0', vegetation_surface='0.0', visibility_
surface='24023.668', wilting_point_surface='0.0', wind_speed_gust_surface='nul
l', geohash='f4r84xqqd2bv'),
 Row(time=1455440400000, lat=36.63995649664971, lon=-120.49956872406986, albed
o_surface='16.0', precipitable_water_entire_atmosphere_single_layer='null', pr
essure_maximum_wind='19326.896', pressure_surface='100150.0', pressure_tropopa
use='19703.373', relative_humidity_zerodegc_isotherm='26.0', snow_depth_surfac
e='0.0', temperature_surface='282.5689', temperature_tropopause='210.17467', t
otal_cloud_cover_entire_atmosphere_single_layer='null', total_precipitation_su
rface_3_hour_accumulation='0.0', vegetation_surface='20.5', visibility_surface
='24223.668', wilting_point_surface='0.1025', wind_speed_gust_surface='null',
geohash='9qd23ynghwcj')]
```

In [60]:
```
from pyspark.sql.functions import col
# Remove rows with 'null' string value in any column
df_cleaned = df.filter(~col("precipitable_water_entire_atmosphere_single_layer'
# Remove rows with 'null' string value in any column
df_cleaned = df_cleaned.filter(~col("temperature_surface").isin('null'))
```

In [16]:
```
df.createOrReplaceTempView("df_temp")
```

In [67]:
```
# Climate Change: Using two-character geohash aggregates across the entire NAM
# the past 5 years. With the regions that have experienced an increase in tempe
# using Pearson’s correlation coefficient (PCC) to determine how the variables
# whether or not the correlations are different based on the region (e.g., mayb
# with humidity in one location but not another). Analyze your results: can you
# found?
import math
from pyspark.mllib.stat import Statistics

feats = []
with open('features.txt') as f:
    feats = [line.strip() for line in f.readlines()[2:]]

df = spark.sql("""
    SELECT
    albedo_surface,
    precipitable_water_entire_atmosphere_single_layer,
    pressure_maximum_wind,
    pressure_surface,
    pressure_tropopause,
    relative_humidity_zerodegc_isotherm,
    snow_depth_surface,
    temperature_surface,
    temperature_tropopause,
    total_cloud_cover_entire_atmosphere_single_layer,
    total_precipitation_surface_3_hour_accumulation,
    vegetation_surface,
    visibility_surface,
    wilting_point_surface,
    wind_speed_gust_surface
    FROM df_temp
    WHERE geohash LIKE '9x%' and wind_speed_gust_surface != 'null' and precipit
    and total_cloud_cover_entire_atmosphere_single_layer != 'null' and temperat
""")

# Convert DataFrame to RDD of tuples
features = df.rdd.map(lambda row: tuple(row))
col_names = df.columns
scores_array = []
corr_mat = Statistics.corr(features, method="pearson")

for i in range(0, 13):
    for j in range(0, 13):
        print(corr_mat[i, j])
        scores_array.append((corr_mat[i, j], feats[i], feats[j]))

for y in range(0, len(scores_array)):
    if math.isnan(scores_array[y][0]):
        scores_array[y] = (0, scores_array[y][1], scores_array[y][2])

# sorting in reverse
scores_array.sort(reverse=True, key=lambda x: abs(x[0]))
print(scores_array)
```

```
[Stage 76:==============>                                        (5 + 15) / 2
0]
```

```
23/05/22 04:02:18 WARN TaskSetManager: Lost task 1.0 in stage 76.0 (TID 6477)
(10.0.1.29 executor 21): org.apache.spark.sql.execution.QueryExecutionExceptio
n: Encountered error while reading file hdfs://orion11:11001/Project3/part-000
30-300232b0-d488-4017-8741-efda0762a4d3-c000.tdv.gz. Details:
        at org.apache.spark.sql.errors.QueryExecutionErrors$.cannotReadFilesEr
ror(QueryExecutionErrors.scala:731)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.next
Iterator(FileScanRDD.scala:283)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.hasN
ext(FileScanRDD.scala:116)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at org.apache.spark.sql.catalyst.expressions.GeneratedClass$GeneratedI
teratorForCodegenStage1.processNext(Unknown Source)
        at org.apache.spark.sql.execution.BufferedRowIterator.hasNext(Buffered
RowIterator.java:43)
        at org.apache.spark.sql.execution.WholeStageCodegenExec$$anon$1.hasNex
t(WholeStageCodegenExec.scala:760)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at org.apache.spark.ContextAwareIterator.hasNext(ContextAwareIterator.
scala:39)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator$GroupedIterator.fill(Iterator.scala:1211)
        at scala.collection.Iterator$GroupedIterator.hasNext(Iterator.scala:12
17)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator.foreach(Iterator.scala:943)
        at scala.collection.Iterator.foreach$(Iterator.scala:943)
        at scala.collection.AbstractIterator.foreach(Iterator.scala:1431)
        at org.apache.spark.api.python.PythonRDD$.writeIteratorToStream(Python
RDD.scala:307)
        at org.apache.spark.sql.execution.python.PythonUDFRunner$$anon$1.write
IteratorToStream(PythonUDFRunner.scala:53)
        at org.apache.spark.api.python.BasePythonRunner$WriterThread.$anonfun
$run$1(PythonRunner.scala:431)
        at org.apache.spark.util.Utils$.logUncaughtExceptions(Utils.scala:206
6)
        at org.apache.spark.api.python.BasePythonRunner$WriterThread.run(Pytho
nRunner.scala:265)
Caused by: java.io.IOException: incorrect header check
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.inflateBytesDir
ect(Native Method)
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.decompress(Zlib
Decompressor.java:225)
        at org.apache.hadoop.io.compress.DecompressorStream.decompress(Decompr
essorStream.java:111)
        at org.apache.hadoop.io.compress.DecompressorStream.read(DecompressorS
tream.java:105)
        at java.base/java.io.InputStream.read(InputStream.java:205)
        at org.apache.hadoop.util.LineReader.fillBuffer(LineReader.java:191)
        at org.apache.hadoop.util.LineReader.readDefaultLine(LineReader.java:2
27)
        at org.apache.hadoop.util.LineReader.readLine(LineReader.java:185)
        at org.apache.hadoop.mapreduce.lib.input.LineRecordReader.skipUtfByteO
rderMark(LineRecordReader.java:158)
        at org.apache.hadoop.mapreduce.lib.input.LineRecordReader.nextKeyValue
(LineRecordReader.java:198)
        at org.apache.spark.sql.execution.datasources.RecordReaderIterator.has
Next(RecordReaderIterator.scala:39)
        at org.apache.spark.sql.execution.datasources.HadoopFileLinesReader.ha
```

```
sNext(HadoopFileLinesReader.scala:69)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator$$anon$17.hasNext(Iterator.scala:814)
        at org.apache.spark.sql.catalyst.csv.CSVExprUtils$.extractHeader(CSVEx
prUtils.scala:54)
        at org.apache.spark.sql.catalyst.csv.CSVHeaderChecker.checkHeaderColum
nNames(CSVHeaderChecker.scala:126)
        at org.apache.spark.sql.catalyst.csv.UnivocityParser$.parseIterator(Un
ivocityParser.scala:410)
        at org.apache.spark.sql.execution.datasources.csv.TextInputCSVDataSour
ce$.readFile(CSVDataSource.scala:104)
        at org.apache.spark.sql.execution.datasources.csv.CSVFileFormat.$anonf
un$buildReader$2(CSVFileFormat.scala:137)
        at org.apache.spark.sql.execution.datasources.FileFormat$$anon$1.apply
(FileFormat.scala:154)
        at org.apache.spark.sql.execution.datasources.FileFormat$$anon$1.apply
(FileFormat.scala:139)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.org
$apache$spark$sql$execution$datasources$FileScanRDD$$anon$$readCurrentFile(Fil
eScanRDD.scala:209)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.next
Iterator(FileScanRDD.scala:270)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.hasN
ext(FileScanRDD.scala:116)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.next
Iterator(FileScanRDD.scala:274)
        ... 20 more
```

```
23/05/22 04:02:26 WARN TaskSetManager: Lost task 16.0 in stage 77.0 (TID 6513)
(10.0.1.22 executor 20): java.io.IOException: incorrect header check
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.inflateBytesDir
ect(Native Method)
        at org.apache.hadoop.io.compress.zlib.ZlibDecompressor.decompress(Zlib
Decompressor.java:225)
        at org.apache.hadoop.io.compress.DecompressorStream.decompress(Decompr
essorStream.java:111)
        at org.apache.hadoop.io.compress.DecompressorStream.read(DecompressorS
tream.java:105)
        at java.base/java.io.InputStream.read(InputStream.java:205)
        at org.apache.hadoop.util.LineReader.fillBuffer(LineReader.java:191)
        at org.apache.hadoop.util.LineReader.readDefaultLine(LineReader.java:2
27)
        at org.apache.hadoop.util.LineReader.readLine(LineReader.java:185)
        at org.apache.hadoop.mapreduce.lib.input.LineRecordReader.skipUtfByteO
rderMark(LineRecordReader.java:158)
        at org.apache.hadoop.mapreduce.lib.input.LineRecordReader.nextKeyValue
(LineRecordReader.java:198)
        at org.apache.spark.sql.execution.datasources.RecordReaderIterator.has
Next(RecordReaderIterator.scala:39)
        at org.apache.spark.sql.execution.datasources.HadoopFileLinesReader.ha
sNext(HadoopFileLinesReader.scala:69)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator$$anon$17.hasNext(Iterator.scala:814)
        at org.apache.spark.sql.catalyst.csv.CSVExprUtils$.extractHeader(CSVEx
prUtils.scala:54)
        at org.apache.spark.sql.catalyst.csv.CSVHeaderChecker.checkHeaderColum
nNames(CSVHeaderChecker.scala:126)
        at org.apache.spark.sql.catalyst.csv.UnivocityParser$.parseIterator(Un
ivocityParser.scala:410)
        at org.apache.spark.sql.execution.datasources.csv.TextInputCSVDataSour
ce$.readFile(CSVDataSource.scala:104)
        at org.apache.spark.sql.execution.datasources.csv.CSVFileFormat.$anonf
un$buildReader$2(CSVFileFormat.scala:137)
        at org.apache.spark.sql.execution.datasources.FileFormat$$anon$1.apply
(FileFormat.scala:154)
        at org.apache.spark.sql.execution.datasources.FileFormat$$anon$1.apply
(FileFormat.scala:139)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.org
$apache$spark$sql$execution$datasources$FileScanRDD$$anon$$readCurrentFile(Fil
eScanRDD.scala:209)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.next
Iterator(FileScanRDD.scala:270)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.hasN
ext(FileScanRDD.scala:116)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at org.apache.spark.sql.catalyst.expressions.GeneratedClass$GeneratedI
teratorForCodegenStage1.processNext(Unknown Source)
        at org.apache.spark.sql.execution.BufferedRowIterator.hasNext(Buffered
RowIterator.java:43)
        at org.apache.spark.sql.execution.WholeStageCodegenExec$$anon$1.hasNex
t(WholeStageCodegenExec.scala:760)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at org.apache.spark.ContextAwareIterator.hasNext(ContextAwareIterator.
scala:39)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator$GroupedIterator.fill(Iterator.scala:1211)
        at scala.collection.Iterator$GroupedIterator.hasNext(Iterator.scala:12
17)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at scala.collection.Iterator.foreach(Iterator.scala:943)
        at scala.collection.Iterator.foreach$(Iterator.scala:943)
        at scala.collection.AbstractIterator.foreach(Iterator.scala:1431)
        at org.apache.spark.api.python.PythonRDD$.writeIteratorToStream(Python
RDD.scala:307)
        at org.apache.spark.sql.execution.python.PythonUDFRunner$$anon$1.write
IteratorToStream(PythonUDFRunner.scala:53)
        at org.apache.spark.api.python.BasePythonRunner$WriterThread.$anonfun
$run$1(PythonRunner.scala:431)
        at org.apache.spark.util.Utils$.logUncaughtExceptions(Utils.scala:206
6)
        at org.apache.spark.api.python.BasePythonRunner$WriterThread.run(Pytho
nRunner.scala:265)
```

```
[Stage 77:=====================================================> (85 + 2) / 8
7]
```

```
1.0
-0.32876037279202674
0.09299648062050242
-0.06331018563061122
0.3312033526046111
0.3006229306878314
0.6753268242721471
-0.4755451378532127
0.12583315702655154
0.12086842825139503
0.034331141735920535
-0.4267477826478016
-0.2178706092637409
-0.32876037279202674
1.0
-0.11909568602952875
0.36200632337062716
-0.5020376784371261
-0.0014540612997474694
-0.350961547761044
0.6735140271750317
-0.2741576276988989
0.06321237683956434
0.10439125607078491
0.4316703744634368
0.07661478387815497
0.09299648062050242
-0.11909568602952875
1.0
0.006777504004157043
0.2153222401224591
0.16975822987134329
0.08709698251812974
-0.1462826215517284
0.24788545577616988
0.11625903713403446
0.026746802442033174
-0.08290270347322395
-0.0888182592086532 2
-0.06331018563061122
0.36200632337062716
0.006777504004157043
1.0
0.003639619876053058
-0.1164933176576662
-0.2888605873040968
0.17371136029774578
0.023391035076909903
-0.07036735416628287
-0.07396206189818443
-0.08477744442787437
0.1489064797441997
0.3312033526046111
-0.5020376784371261
0.2153222401224591
0.003639619876053058
1.0
0.35536936090462834
0.3114912473701309
-0.5463279117708657
```

```
0.801024295475604
0.28817118096764355
0.1064060310397576
-0.38914078791341483
-0.24740710707577296
0.3006229306878314
-0.0014540612997474694
0.16975822987134329
-0.1164933176576662
0.35536936090462834
1.0
0.2667299518452327
-0.36163980203732354
0.29870296299116816
0.4640976919640432
0.2577164929200181
-0.21507930004145864
-0.38817699077108214
0.6753268242721471
-0.350961547761044
0.08709698251812974
-0.2888605873040968
0.3114912473701309
0.2667299518452327
1.0
-0.45411622088804576
0.1207082760114408
0.17260189427670672
0.07914025689505096
-0.30014876534004603
-0.3108042855012886
-0.4755451378532127
0.6735140271750317
-0.1462826215517284
0.17371136029774578
-0.5463279117708657
-0.36163980203732354
-0.45411622088804576
1.0
-0.28685371296960016
-0.18097483525456154
-0.04178416037259742
0.5099277025521041
0.22598903199968487
0.12583315702655154
-0.2741576276988989
0.24788545577616988
0.023391035076909903
0.801024295475604
0.29870296299116816
0.1207082760114408
-0.28685371296960016
1.0
0.2517467981698815
0.12149827701647714
-0.14900896652226722
-0.201892265808317
0.12086842825139503
0.06321237683956434
0.11625903713403446
```

```
-0.07036735416628287
0.28817118096764355
0.4640976919640432
0.17260189427670672
-0.18097483525456154
0.2517467981698815
1.0
0.3069888844477944
-0.1174956716673092
-0.4865817385500211
0.034331141735920535
0.10439125607078491
0.026746802442033174
-0.07396206189818443
0.1064060310397576
0.257716492920018l
0.07914025689505096
-0.04178416037259742
0.12149827701647714
0.3069888844477944
1.0
0.034895435278292856
-0.45873285910546235
-0.4267477826478016
0.4316703744634368
-0.08290270347322395
-0.08477744442787437
-0.38914078791341483
-0.21507930004145864
-0.30014876534004603
0.5099277025521041
-0.14900896652226722
-0.1174956716673092
0.034895435278292856
1.0
0.11270724127115021
-0.2178706092637409
0.07661478387815497
-0.08881825920865322
0.1489064797441997
-0.24740710707577296
-0.38817699077108214
-0.3108042855012886
0.22598903199968487
-0.201892265808317
-0.4865817385500211
-0.45873285910546235
0.11270724127115021
1.0
[(1.0, 'pressure_maximum_wind', 'pressure_maximum_wind'), (1.0, 'pressure_surf
ace', 'pressure_surface'), (1.0, 'pressure_tropopause', 'pressure_tropopaus
e'), (1.0, 'relative_humidity_zerodegc_isotherm', 'relative_humidity_zerodegc_
isotherm'), (1.0, 'snow_depth_surface', 'snow_depth_surface'), (1.0, 'temperat
ure_surface', 'temperature_surface'), (1.0, 'temperature_tropopause', 'tempera
ture_tropopause'), (1.0, 'total_cloud_cover_entire_atmosphere_single_layer',
'total_cloud_cover_entire_atmosphere_single_layer'), (1.0, 'total_precipitatio
n_surface_3_hour_accumulation', 'total_precipitation_surface_3_hour_accumulati
on'), (1.0, 'vegetation_surface', 'vegetation_surface'), (1.0, 'visibility_sur
face', 'visibility_surface'), (1.0, 'wilting_point_surface', 'wilting_point_su
rface'), (1.0, 'wind_speed_gust_surface', 'wind_speed_gust_surface'), (0.80102
```

```
4295475604, 'snow_depth_surface', 'total_precipitation_surface_3_hour_accumula
tion'), (0.801024295475604, 'total_precipitation_surface_3_hour_accumulation',
'snow_depth_surface'), (0.6753268242721471, 'pressure_maximum_wind', 'temperat
ure_tropopause'), (0.6753268242721471, 'temperature_tropopause', 'pressure_max
imum_wind'), (0.6735140271750317, 'pressure_surface', 'total_cloud_cover_entir
e_atmosphere_single_layer'), (0.6735140271750317, 'total_cloud_cover_entire_at
mosphere_single_layer', 'pressure_surface'), (-0.5463279117708657, 'snow_depth
_surface', 'total_cloud_cover_entire_atmosphere_single_layer'), (-0.5463279117
708657, 'total_cloud_cover_entire_atmosphere_single_layer', 'snow_depth_surfac
e'), (0.5099277025521041, 'total_cloud_cover_entire_atmosphere_single_layer',
'wilting_point_surface'), (0.5099277025521041, 'wilting_point_surface', 'total
_cloud_cover_entire_atmosphere_single_layer'), (-0.5020376784371261, 'pressure
_surface', 'snow_depth_surface'), (-0.5020376784371261, 'snow_depth_surface',
'pressure_surface'), (-0.4865817385500211, 'vegetation_surface', 'wind_speed_g
ust_surface'), (-0.4865817385500211, 'wind_speed_gust_surface', 'vegetation_su
rface'), (-0.4755451378532127, 'pressure_maximum_wind', 'total_cloud_cover_ent
ire_atmosphere_single_layer'), (-0.4755451378532127, 'total_cloud_cover_entire
_atmosphere_single_layer', 'pressure_maximum_wind'), (0.4640976919640432, 'tem
perature_surface', 'vegetation_surface'), (0.4640976919640432, 'vegetation_sur
face', 'temperature_surface'), (-0.45873285910546235, 'visibility_surface', 'w
ind_speed_gust_surface'), (-0.45873285910546235, 'wind_speed_gust_surface', 'v
isibility_surface'), (-0.4541622088804576, 'temperature_tropopause', 'total_c
loud_cover_entire_atmosphere_single_layer'), (-0.4541622088804576, 'total_clo
ud_cover_entire_atmosphere_single_layer', 'temperature_tropopause'), (0.431670
3744634368, 'pressure_surface', 'wilting_point_surface'), (0.4316703744634368,
'wilting_point_surface', 'pressure_surface'), (-0.4267477826478016, 'pressure_
maximum_wind', 'wilting_point_surface'), (-0.4267477826478016, 'wilting_point_
surface', 'pressure_maximum_wind'), (-0.38914078791341483, 'snow_depth_surfac
e', 'wilting_point_surface'), (-0.38914078791341483, 'wilting_point_surface',
'snow_depth_surface'), (-0.38817699077108214, 'temperature_surface', 'wind_spe
ed_gust_surface'), (-0.38817699077108214, 'wind_speed_gust_surface', 'temperat
ure_surface'), (0.36200632337062716, 'pressure_surface', 'relative_humidity_ze
rodegc_isotherm'), (0.36200632337062716, 'relative_humidity_zerodegc_isother
m', 'pressure_surface'), (-0.36163980203732354, 'temperature_surface', 'total_
cloud_cover_entire_atmosphere_single_layer'), (-0.36163980203732354, 'total_cl
oud_cover_entire_atmosphere_single_layer', 'temperature_surface'), (0.35536936
090462834, 'snow_depth_surface', 'temperature_surface'), (0.35536936090462834,
'temperature_surface', 'snow_depth_surface'), (-0.350961547761044, 'pressure_s
urface', 'temperature_tropopause'), (-0.350961547761044, 'temperature_tropopau
se', 'pressure_surface'), (0.3312033526046111, 'pressure_maximum_wind', 'snow_
depth_surface'), (0.3312033526046111, 'snow_depth_surface', 'pressure_maximum_
wind'), (-0.32876037279202674, 'pressure_maximum_wind', 'pressure_surface'),
(-0.32876037279202674, 'pressure_surface', 'pressure_maximum_wind'), (0.311491
2473701309, 'snow_depth_surface', 'temperature_tropopause'), (0.31149124737013
09, 'temperature_tropopause', 'snow_depth_surface'), (-0.3108042855012886, 'te
mperature_tropopause', 'wind_speed_gust_surface'), (-0.3108042855012886, 'wind
_speed_gust_surface', 'temperature_tropopause'), (0.3069888844477944, 'vegetat
ion_surface', 'visibility_surface'), (0.3069888844477944, 'visibility_surfac
e', 'vegetation_surface'), (0.3006229306878314, 'pressure_maximum_wind', 'temp
erature_surface'), (0.3006229306878314, 'temperature_surface', 'pressure_maxim
um_wind'), (-0.30014876534004603, 'temperature_tropopause', 'wilting_point_sur
face'), (-0.30014876534004603, 'wilting_point_surface', 'temperature_tropopaus
e'), (0.29870296299116816, 'temperature_surface', 'total_precipitation_surface
_3_hour_accumulation'), (0.29870296299116816, 'total_precipitation_surface_3_h
our_accumulation', 'temperature_surface'), (-0.2888605873040968, 'relative_hum
idity_zerodegc_isotherm', 'temperature_tropopause'), (-0.2888605873040968, 'te
mperature_tropopause', 'relative_humidity_zerodegc_isotherm'), (0.288171180967
64355, 'snow_depth_surface', 'vegetation_surface'), (0.28817118096764355, 'veg
etation_surface', 'snow_depth_surface'), (-0.28685371296960016, 'total_cloud_c
over_entire_atmosphere_single_layer', 'total_precipitation_surface_3_hour_accu
```

```
mulation'), (-0.28685371296960016, 'total_precipitation_surface_3_hour_accumul
ation', 'total_cloud_cover_entire_atmosphere_single_layer'), (-0.2741576276988
989, 'pressure_surface', 'total_precipitation_surface_3_hour_accumulation'),
(-0.2741576276988989, 'total_precipitation_surface_3_hour_accumulation', 'pres
sure_surface'), (0.2667299518452327, 'temperature_surface', 'temperature_tropo
pause'), (0.2667299518452327, 'temperature_tropopause', 'temperature_surfac
e'), (0.2577164929200181, 'temperature_surface', 'visibility_surface'), (0.257
7164929200181, 'visibility_surface', 'temperature_surface'), (0.25174679816988
15, 'total_precipitation_surface_3_hour_accumulation', 'vegetation_surface'),
(0.2517467981698815, 'vegetation_surface', 'total_precipitation_surface_3_hour
_accumulation'), (0.24788545577616988, 'pressure_tropopause', 'total_precipita
tion_surface_3_hour_accumulation'), (0.24788545577616988, 'total_precipitation
_surface_3_hour_accumulation', 'pressure_tropopause'), (-0.24740710707577296,
'snow_depth_surface', 'wind_speed_gust_surface'), (-0.24740710707577296, 'wind
_speed_gust_surface', 'snow_depth_surface'), (0.22598903199968487, 'total_clou
d_cover_entire_atmosphere_single_layer', 'wind_speed_gust_surface'), (0.225989
03199968487, 'wind_speed_gust_surface', 'total_cloud_cover_entire_atmosphere_s
ingle_layer'), (-0.2178706092637409, 'pressure_maximum_wind', 'wind_speed_gust
_surface'), (-0.2178706092637409, 'wind_speed_gust_surface', 'pressure_maximum
_wind'), (0.2153222401224591, 'pressure_tropopause', 'snow_depth_surface'),
(0.2153222401224591, 'snow_depth_surface', 'pressure_tropopause'), (-0.2150793
0004145864, 'temperature_surface', 'wilting_point_surface'), (-0.2150793000414
5864, 'wilting_point_surface', 'temperature_surface'), (-0.201892265808317, 't
otal_precipitation_surface_3_hour_accumulation', 'wind_speed_gust_surface'),
(-0.201892265808317, 'wind_speed_gust_surface', 'total_precipitation_surface_3
_hour_accumulation'), (-0.18097483525456154, 'total_cloud_cover_entire_atmosph
ere_single_layer', 'vegetation_surface'), (-0.18097483525456154, 'vegetation_s
urface', 'total_cloud_cover_entire_atmosphere_single_layer'), (0.1737113602977
4578, 'relative_humidity_zerodegc_isotherm', 'total_cloud_cover_entire_atmosph
ere_single_layer'), (0.17371136029774578, 'total_cloud_cover_entire_atmosphere
_single_layer', 'relative_humidity_zerodegc_isotherm'), (0.17260189427670672,
'temperature_tropopause', 'vegetation_surface'), (0.17260189427670672, 'vegeta
tion_surface', 'temperature_tropopause'), (0.16975822987134329, 'pressure_trop
opause', 'temperature_surface'), (0.16975822987134329, 'temperature_surface',
'pressure_tropopause'), (-0.14900896652226722, 'total_precipitation_surface_3_
hour_accumulation', 'wilting_point_surface'), (-0.14900896652226722, 'wilting_
point_surface', 'total_precipitation_surface_3_hour_accumulation'), (0.1489064
797441997, 'relative_humidity_zerodegc_isotherm', 'wind_speed_gust_surface'),
(0.1489064797441997, 'wind_speed_gust_surface', 'relative_humidity_zerodegc_is
otherm'), (-0.1462826215517284, 'pressure_tropopause', 'total_cloud_cover_enti
re_atmosphere_single_layer'), (-0.1462826215517284, 'total_cloud_cover_entire_
atmosphere_single_layer', 'pressure_tropopause'), (0.12583315702655154, 'press
ure_maximum_wind', 'total_precipitation_surface_3_hour_accumulation'), (0.1258
3315702655154, 'total_precipitation_surface_3_hour_accumulation', 'pressure_ma
ximum_wind'), (0.12149827701647714, 'total_precipitation_surface_3_hour_accumu
lation', 'visibility_surface'), (0.12149827701647714, 'visibility_surface', 't
otal_precipitation_surface_3_hour_accumulation'), (0.12086842825139503, 'press
ure_maximum_wind', 'vegetation_surface'), (0.12086842825139503, 'vegetation_su
rface', 'pressure_maximum_wind'), (0.1207082760114408, 'temperature_tropopaus
e', 'total_precipitation_surface_3_hour_accumulation'), (0.1207082760114408,
'total_precipitation_surface_3_hour_accumulation', 'temperature_tropopause'),
(-0.11909568602952875, 'pressure_surface', 'pressure_tropopause'), (-0.1190956
8602952875, 'pressure_tropopause', 'pressure_surface'), (-0.1174956716673092,
'vegetation_surface', 'wilting_point_surface'), (-0.1174956716673092, 'wilting
_point_surface', 'vegetation_surface'), (-0.1164933176576662, 'relative_humidi
ty_zerodegc_isotherm', 'temperature_surface'), (-0.1164933176576662, 'temperat
ure_surface', 'relative_humidity_zerodegc_isotherm'), (0.11625903713403446, 'p
ressure_tropopause', 'vegetation_surface'), (0.11625903713403446, 'vegetation_
surface', 'pressure_tropopause'), (0.11270724127115021, 'wilting_point_surfac
e', 'wind_speed_gust_surface'), (0.11270724127115021, 'wind_speed_gust_surfac
```

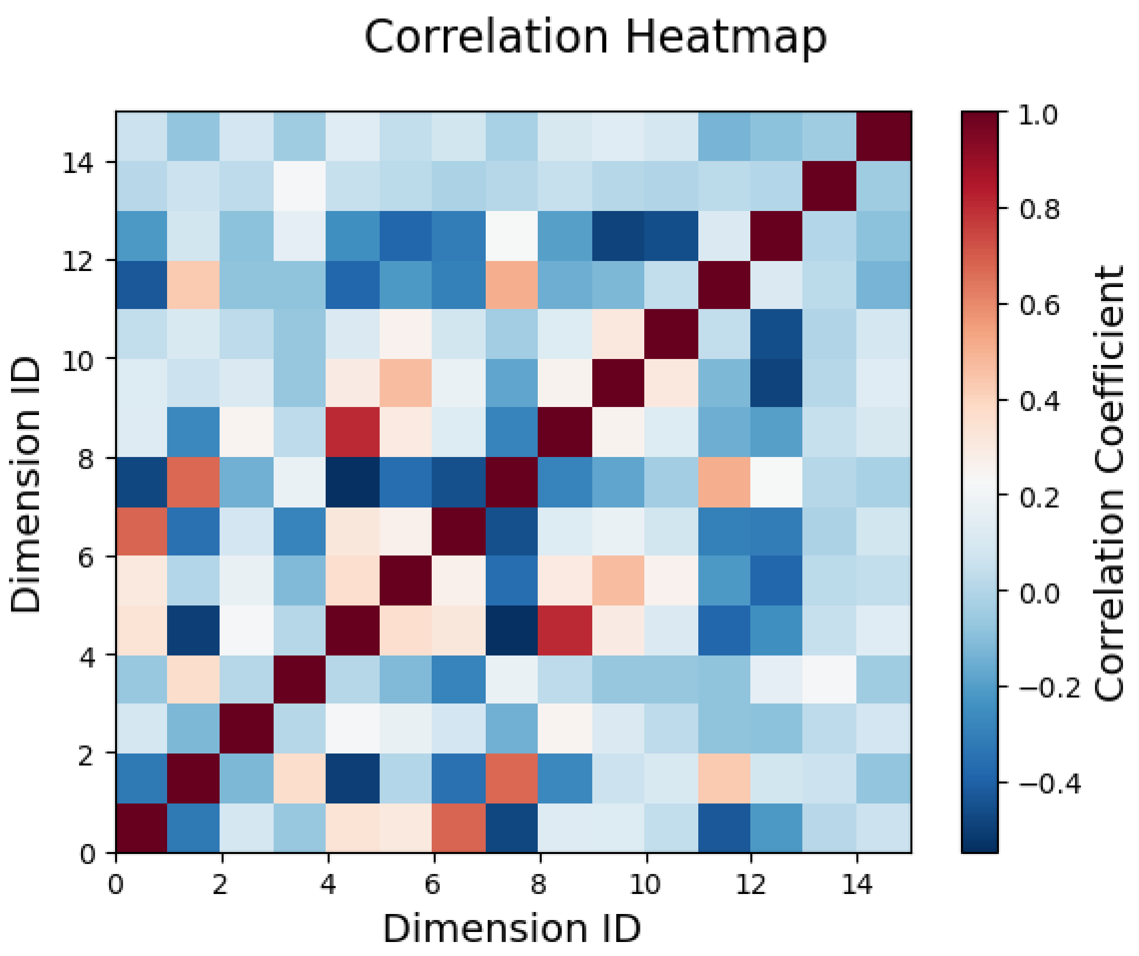
```
e', 'wilting_point_surface'), (0.1064060310397576, 'snow_depth_surface', 'visi
bility_surface'), (0.1064060310397576, 'visibility_surface', 'snow_depth_surfa
ce'), (0.1043912560707849l, 'pressure_surface', 'visibility_surface'), (0.1043
9125607078491, 'visibility_surface', 'pressure_surface'), (0.0929964806205024
2, 'pressure_maximum_wind', 'pressure_tropopause'), (0.09299648062050242, 'pre
ssure_tropopause', 'pressure_maximum_wind'), (-0.08881825920865322, 'pressure_
tropopause', 'wind_speed_gust_surface'), (-0.08881825920865322, 'wind_speed_gu
st_surface', 'pressure_tropopause'), (0.08709698251812974, 'pressure_tropopaus
e', 'temperature_tropopause'), (0.08709698251812974, 'temperature_tropopause',
'pressure_tropopause'), (-0.08477744442787437, 'relative_humidity_zerodegc_iso
therm', 'wilting_point_surface'), (-0.08477744442787437, 'wilting_point_surfac
e', 'relative_humidity_zerodegc_isotherm'), (-0.08290270347322395, 'pressure_t
ropopause', 'wilting_point_surface'), (-0.08290270347322395, 'wilting_point_su
rface', 'pressure_tropopause'), (0.07914025689505096, 'temperature_tropopaus
e', 'visibility_surface'), (0.07914025689505096, 'visibility_surface', 'temper
ature_tropopause'), (0.07661478387815497, 'pressure_surface', 'wind_speed_gust
_surface'), (0.07661478387815497, 'wind_speed_gust_surface', 'pressure_surfac
e'), (-0.07396206189818443, 'relative_humidity_zerodegc_isotherm', 'visibility
_surface'), (-0.07396206189818443, 'visibility_surface', 'relative_humidity_ze
rodegc_isotherm'), (-0.07036735416628287, 'relative_humidity_zerodegc_isother
m', 'vegetation_surface'), (-0.07036735416628287, 'vegetation_surface', 'relat
ive_humidity_zerodegc_isotherm'), (-0.06331018563061122, 'pressure_maximum_win
d', 'relative_humidity_zerodegc_isotherm'), (-0.06331018563061122, 'relative_h
umidity_zerodegc_isotherm', 'pressure_maximum_wind'), (0.06321237683956434, 'p
ressure_surface', 'vegetation_surface'), (0.06321237683956434, 'vegetation_sur
face', 'pressure_surface'), (-0.04178416037259742, 'total_cloud_cover_entire_a
tmosphere_single_layer', 'visibility_surface'), (-0.04178416037259742, 'visibi
lity_surface', 'total_cloud_cover_entire_atmosphere_single_layer'), (0.0348954
35278292856, 'visibility_surface', 'wilting_point_surface'), (0.03489543527829
2856, 'wilting_point_surface', 'visibility_surface'), (0.034331141735920535,
'pressure_maximum_wind', 'visibility_surface'), (0.034331141735920535, 'visibi
lity_surface', 'pressure_maximum_wind'), (0.026746802442033174, 'pressure_trop
opause', 'visibility_surface'), (0.026746802442033174, 'visibility_surface',
'pressure_tropopause'), (0.023391035076909903, 'relative_humidity_zerodegc_iso
therm', 'total_precipitation_surface_3_hour_accumulation'), (0.023391035076909
903, 'total_precipitation_surface_3_hour_accumulation', 'relative_humidity_zer
odegc_isotherm'), (0.006777504004157043, 'pressure_tropopause', 'relative_humi
dity_zerodegc_isotherm'), (0.006777504004157043, 'relative_humidity_zerodegc_i
sotherm', 'pressure_tropopause'), (0.003639619876053058, 'relative_humidity_ze
rodegc_isotherm', 'snow_depth_surface'), (0.003639619876053058, 'snow_depth_su
rface', 'relative_humidity_zerodegc_isotherm'), (-0.0014540612997474694, 'pres
sure_surface', 'temperature_surface'), (-0.0014540612997474694, 'temperature_s
urface', 'pressure_surface')]
```

In [68]:

```python
import sys
import numpy as np
import matplotlib.pyplot as plt

plt.suptitle('Correlation Heatmap', fontsize=16)
plt.xlabel('Dimension ID', fontsize=14)
plt.ylabel('Dimension ID', fontsize=14)

plt.pcolor(corr_mat, cmap='RdBu_r')
cb = plt.colorbar()
cb.set_label('Correlation Coefficient', fontsize=14)
plt.plot()
```

Out[68]:

```
[]
```



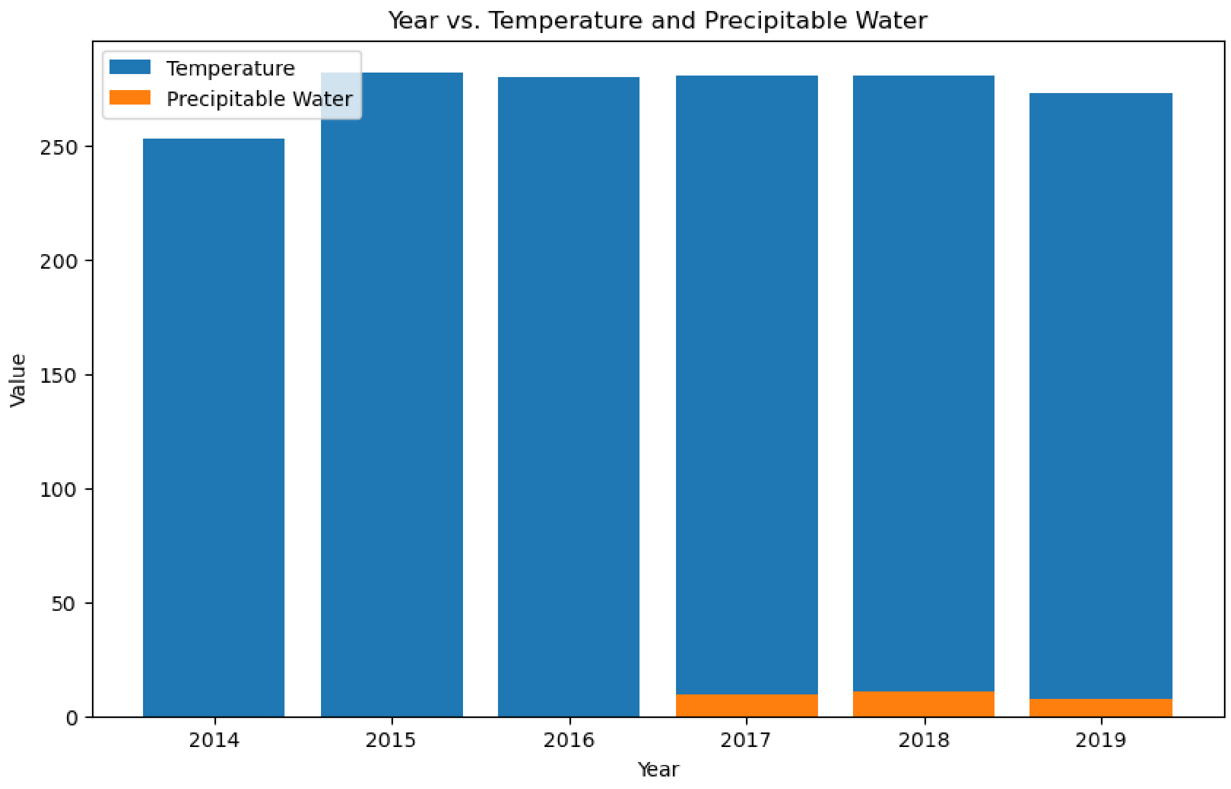
In [69]:
```
# snow_depth_surface and albedo are strongly co-related.
# precipitable_water_entire_atmosphere_single_layer and temperature_surface are
# pressure_tropopause and temperature_tropopause are strongly co-related.
temp_surface = spark.sql("""
                        Select YEAR(FROM_UNIXTIME(time/1000)),
                         AVG(CASE WHEN temperature_surface == 'null' THEN 0.0 E
                        AVG(CASE WHEN precipitable_water_entire_atmosphere_sing
                        from df_temp
                         where geohash LIKE '9x%'
                         GROUP BY YEAR(FROM_UNIXTIME(time/1000)) """).collect()
# print(temp_surface)
# temp_surface.show(3)
```

In [70]:
```
print(temp_surface[:5])
```

```
[Row(year(from_unixtime((time / 1000), yyyy-MM-dd HH:mm:ss))=2015, avg(CASE WH
EN (temperature_surface = null) THEN 0.0 ELSE temperature_surface END)=282.149
6568564598, avg(CASE WHEN (precipitable_water_entire_atmosphere_single_layer =
null) THEN 0.0 ELSE precipitable_water_entire_atmosphere_single_layer END)=0.
0), Row(year(from_unixtime((time / 1000), yyyy-MM-dd HH:mm:ss))=2016, avg(CASE
WHEN (temperature_surface = null) THEN 0.0 ELSE temperature_surface END)=280.3
302151301624, avg(CASE WHEN (precipitable_water_entire_atmosphere_single_layer
= null) THEN 0.0 ELSE precipitable_water_entire_atmosphere_single_layer END)=
0.0), Row(year(from_unixtime((time / 1000), yyyy-MM-dd HH:mm:ss))=2018, avg(CA
SE WHEN (temperature_surface = null) THEN 0.0 ELSE temperature_surface END)=28
0.6623647452187, avg(CASE WHEN (precipitable_water_entire_atmosphere_single_la
yer = null) THEN 0.0 ELSE precipitable_water_entire_atmosphere_single_layer EN
D)=10.883690347968521), Row(year(from_unixtime((time / 1000), yyyy-MM-dd HH:m
m:ss))=2019, avg(CASE WHEN (temperature_surface = null) THEN 0.0 ELSE temperat
ure_surface END)=272.9755704882924, avg(CASE WHEN (precipitable_water_entire_a
tmosphere_single_layer = null) THEN 0.0 ELSE precipitable_water_entire_atmosph
ere_single_layer END)=7.682247617823717), Row(year(from_unixtime((time / 100
0), yyyy-MM-dd HH:mm:ss))=2014, avg(CASE WHEN (temperature_surface = null) THE
N 0.0 ELSE temperature_surface END)=253.19901869158733, avg(CASE WHEN (precipi
table_water_entire_atmosphere_single_layer = null) THEN 0.0 ELSE precipitable_
water_entire_atmosphere_single_layer END)=0.0)]
```

In [71]:

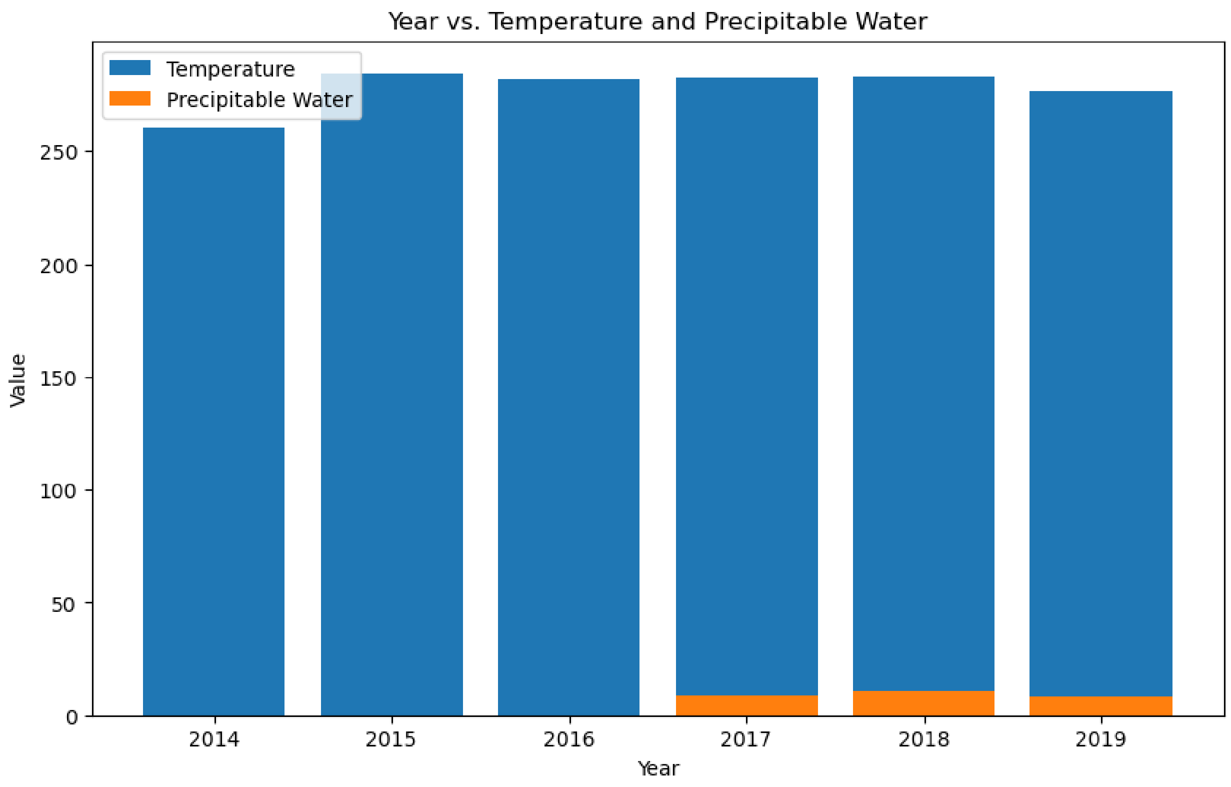
```python
import matplotlib.pyplot as plt


years = [row[0] for row in temp_surface]
temperature = [row[1] for row in temp_surface]
precipitable_water = [row[2] for row in temp_surface]

plt.figure(figsize=(10, 6))

plt.bar(years, temperature, label='Temperature')
plt.bar(years, precipitable_water, label='Precipitable Water')

plt.xlabel('Year')
plt.ylabel('Value')
plt.title('Year vs. Temperature and Precipitable Water')

plt.legend()
plt.show()
```



In [ ]:
```
# In 2017 2018 2019, we can observe increase in precipitable water and ultimate
```

In [72]:
```
# snow_depth_surface and albedo are strongly co-related.
# precipitable_water_entire_atmosphere_single_layer and temperature_surface are
# pressure_tropopause and temperature_tropopause are strongly co-related.
temp_surface = spark.sql("""
                        Select YEAR(FROM_UNIXTIME(time/1000)),
                         AVG(CASE WHEN temperature_surface == 'null' THEN 0.0 E
                        AVG(CASE WHEN precipitable_water_entire_atmosphere_sinc
                        from df_temp
                         where geohash LIKE '9r%'
                         GROUP BY YEAR(FROM_UNIXTIME(time/1000)) """).collect()
# print(temp_surface)
# temp_surface.show(3)
```

In [73]:
```
import matplotlib.pyplot as plt


years = [row[0] for row in temp_surface]
temperature = [row[1] for row in temp_surface]
precipitable_water = [row[2] for row in temp_surface]

plt.figure(figsize=(10, 6))

plt.bar(years, temperature, label='Temperature')
plt.bar(years, precipitable_water, label='Precipitable Water')

plt.xlabel('Year')
plt.ylabel('Value')
plt.title('Year vs. Temperature and Precipitable Water')
plt.legend()
plt.show()
```



In [ ]:
```
# Co-relation are not region based because we see similar trend for wyoming anc
```