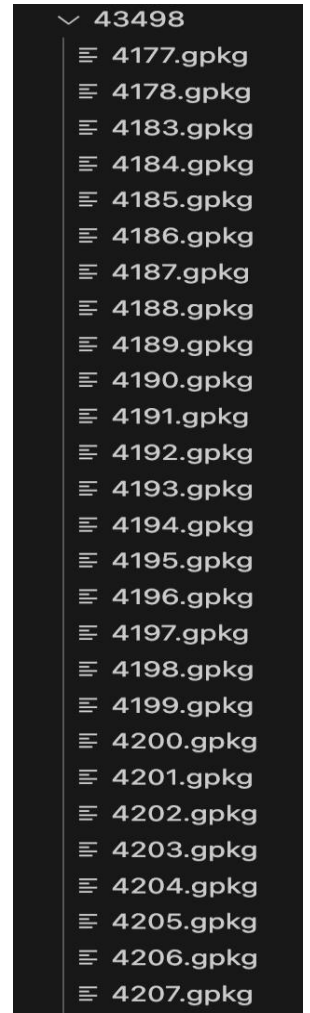
Big statistics data product

Tiling system

Our IWP data is organized and store under Tiling system, so we can search for IWPs from desired region based on their Tiling information.

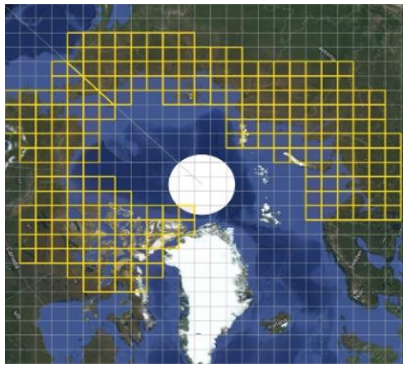


References:

- Microsoft Bing Tiling System
- Open Geospatial Consortium

Workflow of the big statistics data mapping

Step 1: Create 230 grids to cover the entire Pan-Arctic region.

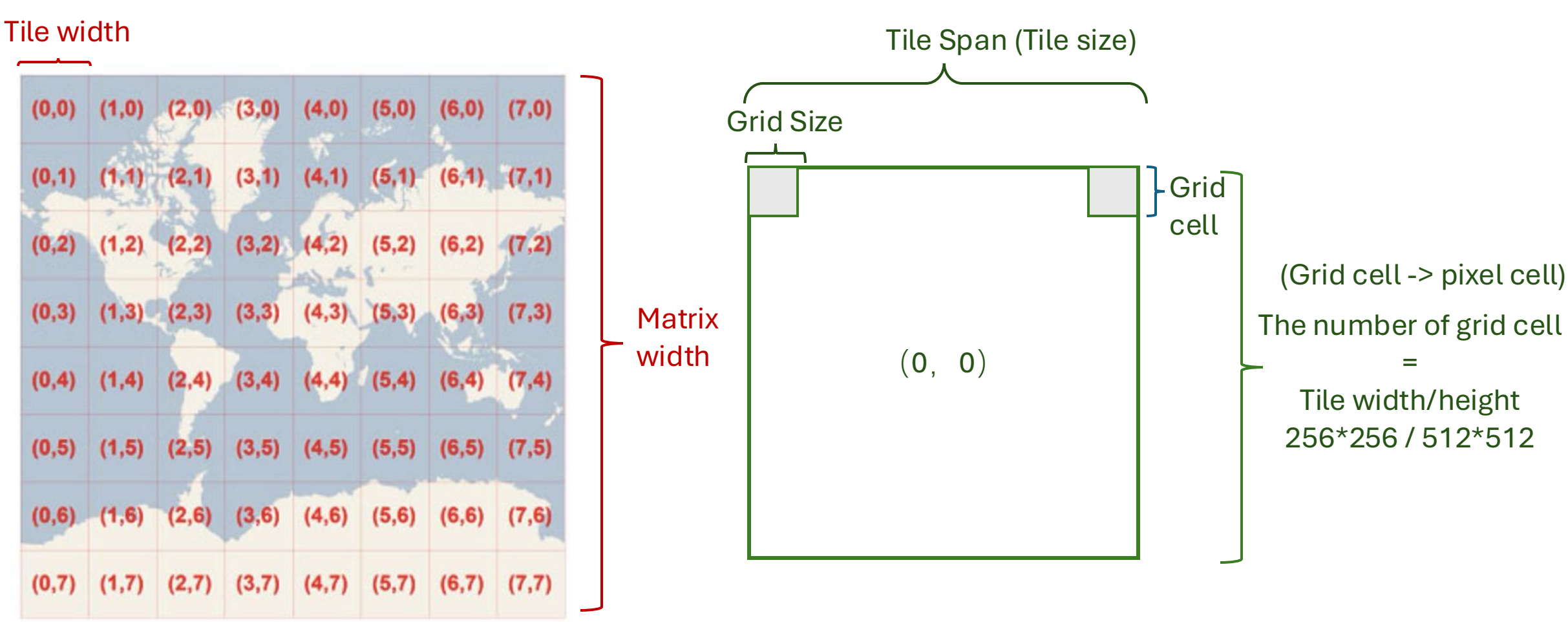


Step 2: Set the pixel of each grid to 1km.

Step 3: (Download geopackages first if needed) Get BBOX of each pixel given X, Y, Z, and search for IWP under
(refer to Juliet's code)
https://github.com/PermafrostDiscoveryGateway/viz-info/blob/main/helpful-code/prepocessing/bounding_box_tiles.ipynb

Step 4: Calculate all IWPs within a pixel

Step 5: Map all pixels within a grid, and slice all grids together



Given a pair of pixel XY coordinates, you can easily determine the tile XY coordinates of the tile containing that pixel:

```
tileX = floor(pixelX / 256)
```

```
tileY = floor(pixelY / 256)
```

Reference: https://docs.ogc.org/is/17-083r4/17-083r4.html#toc15

Tile Size /Matrix Size Calculation

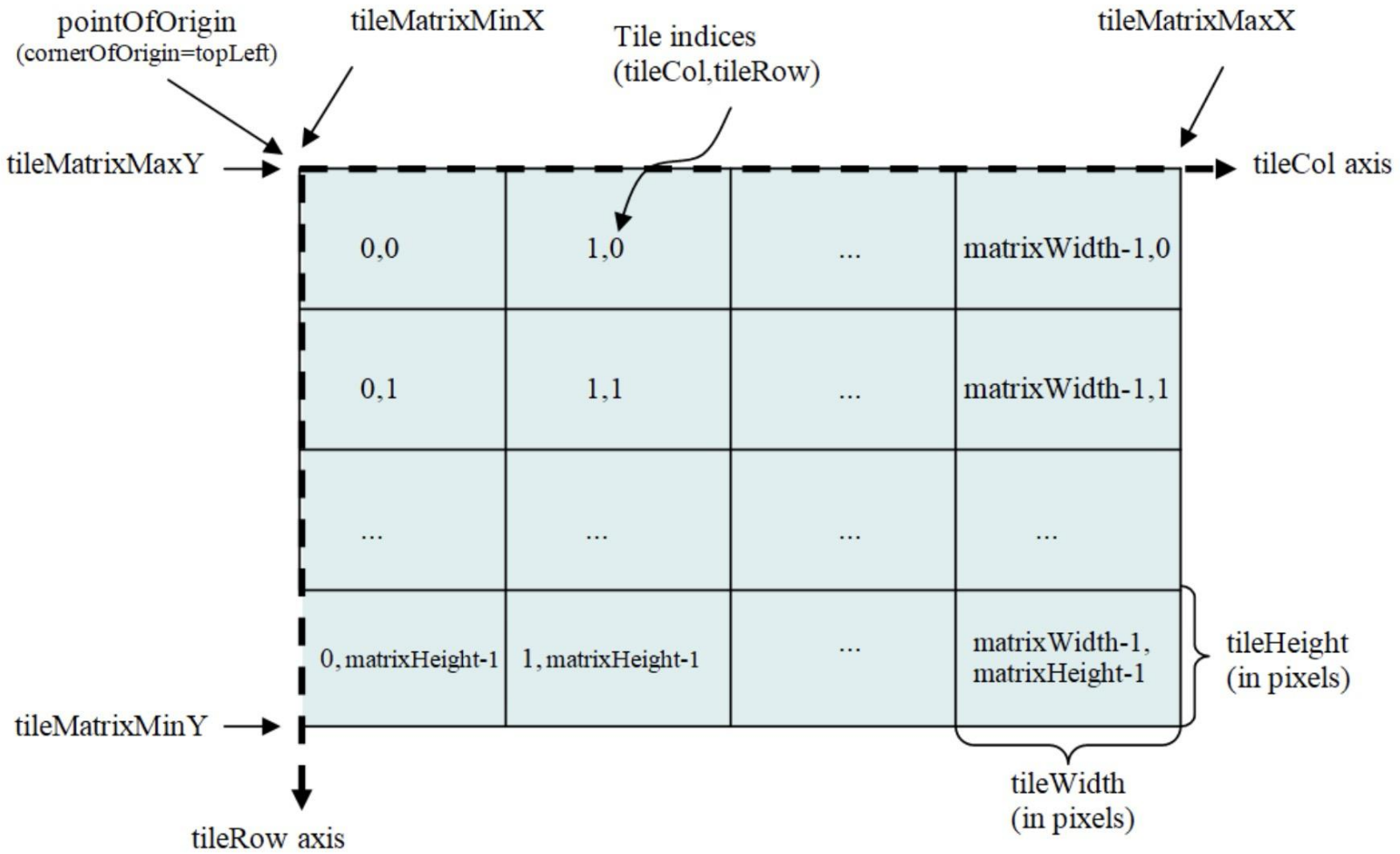
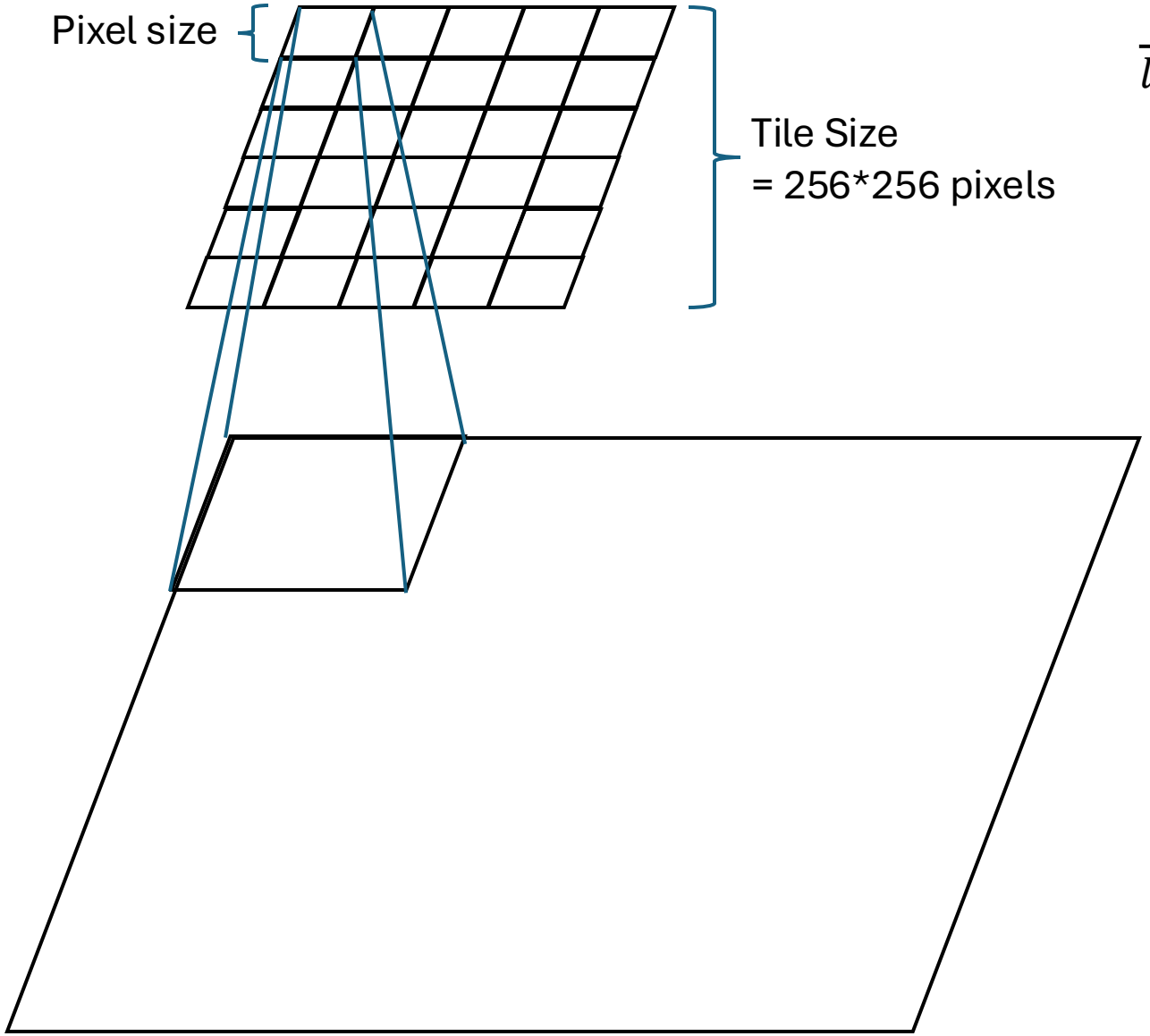


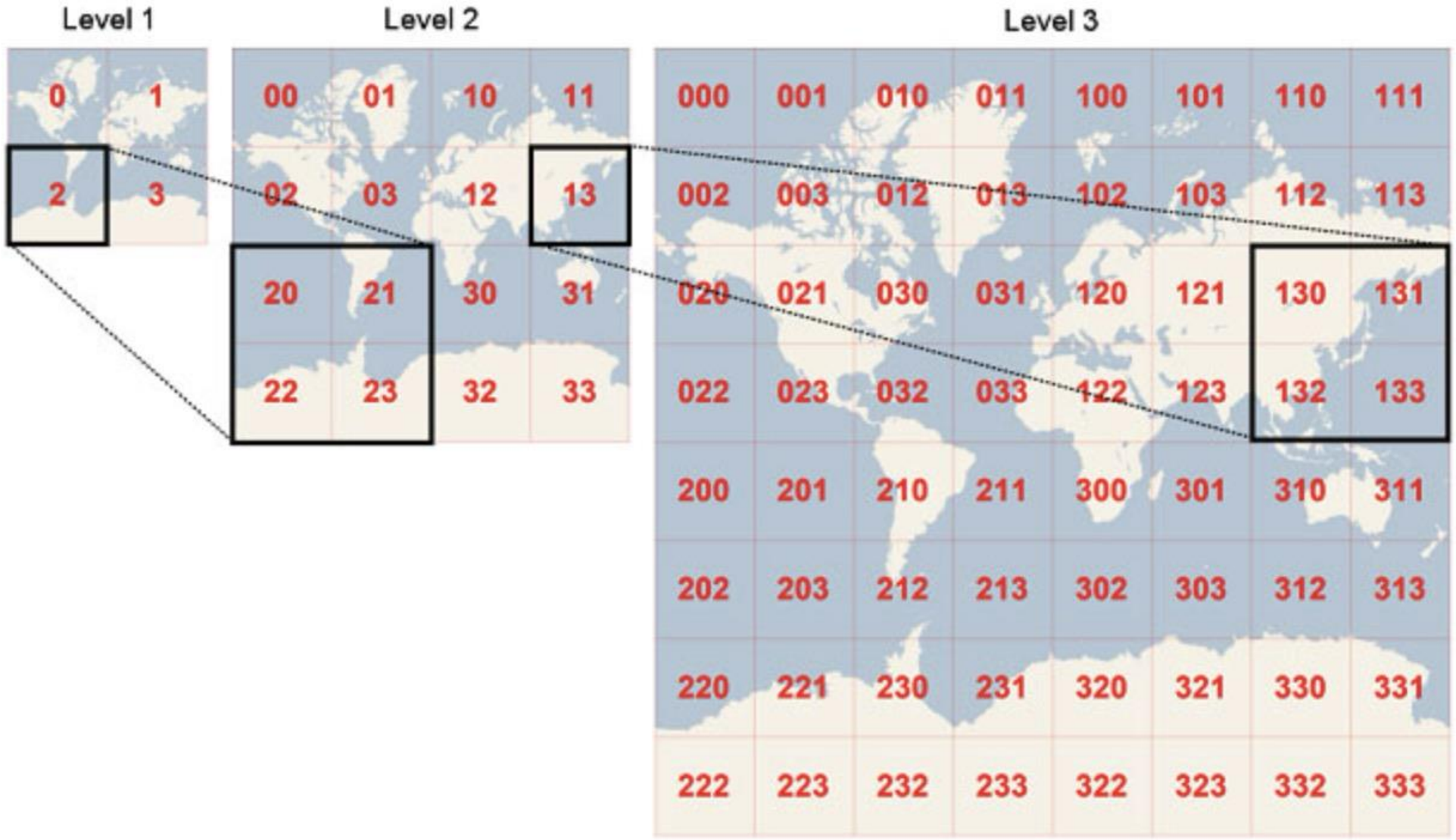
Figure 5 – Tile Space (the corner of origin is topLeft)

Tile Size /Matrix Size Calculation



$$\frac{each\ pixel\ size\ on\ map}{length\ on\ the\ real\ world\ per\ pixel} = \frac{1}{Scale\ Denominator}$$

Tile Matrix Set



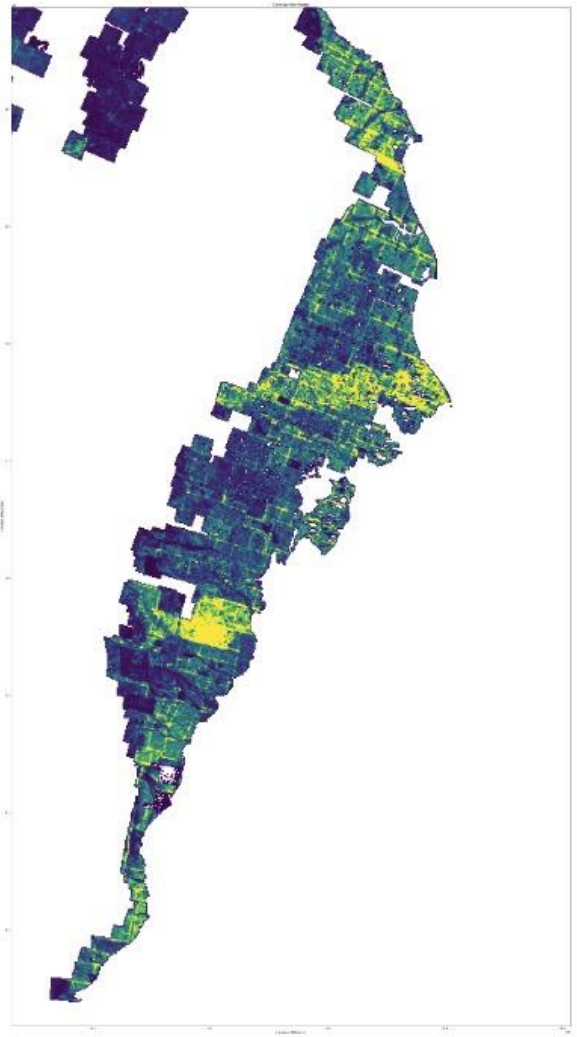
Create a new geom column with centroid points as the geometry

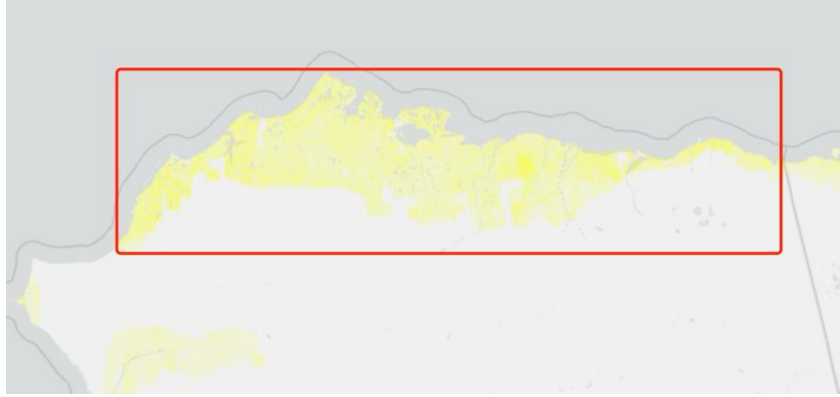
Create the column

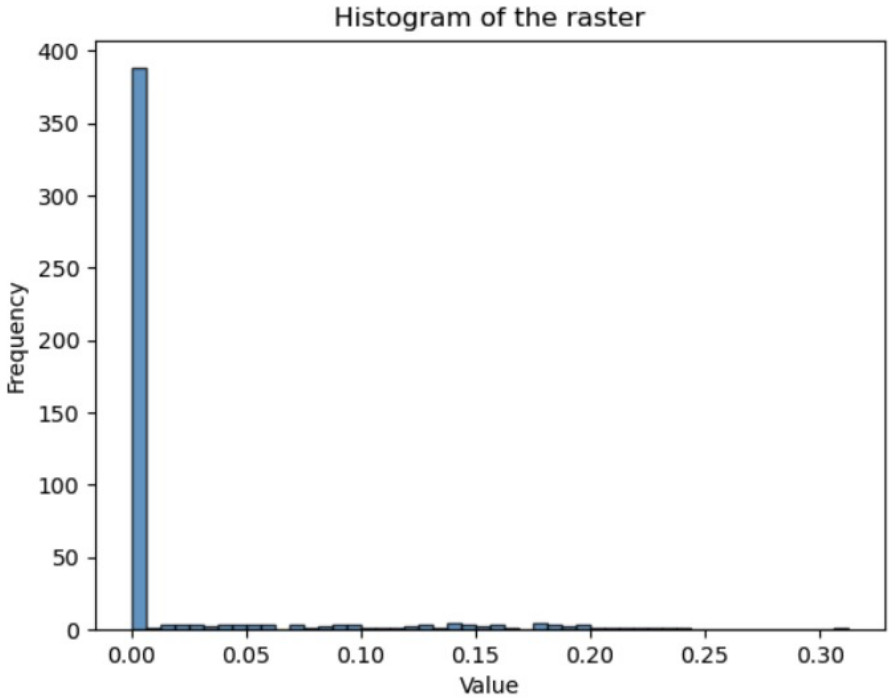
```
SELECT AddGeometryColumn('public', 'alaska_all_3413','geom_centroid',3413,'POINT',2) ;
```

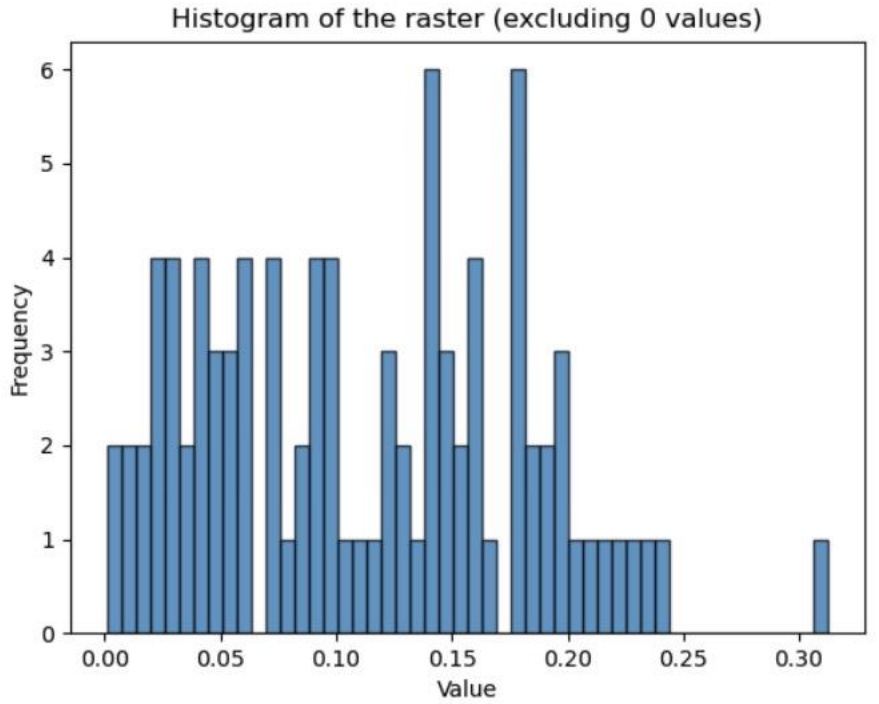
Update the values in the column

```
update public.Alaska_all_3413
set geom = ST_SetSRID(
ST_MakePoint(
"Longitude"::double precision,
"Latitude"::double precision
), 4326)
```



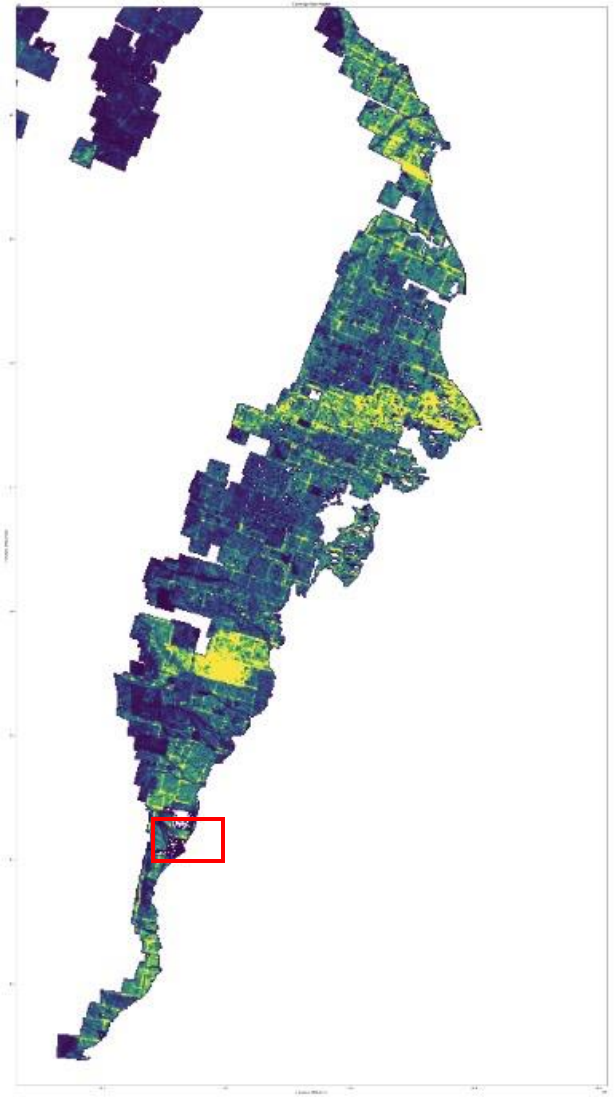




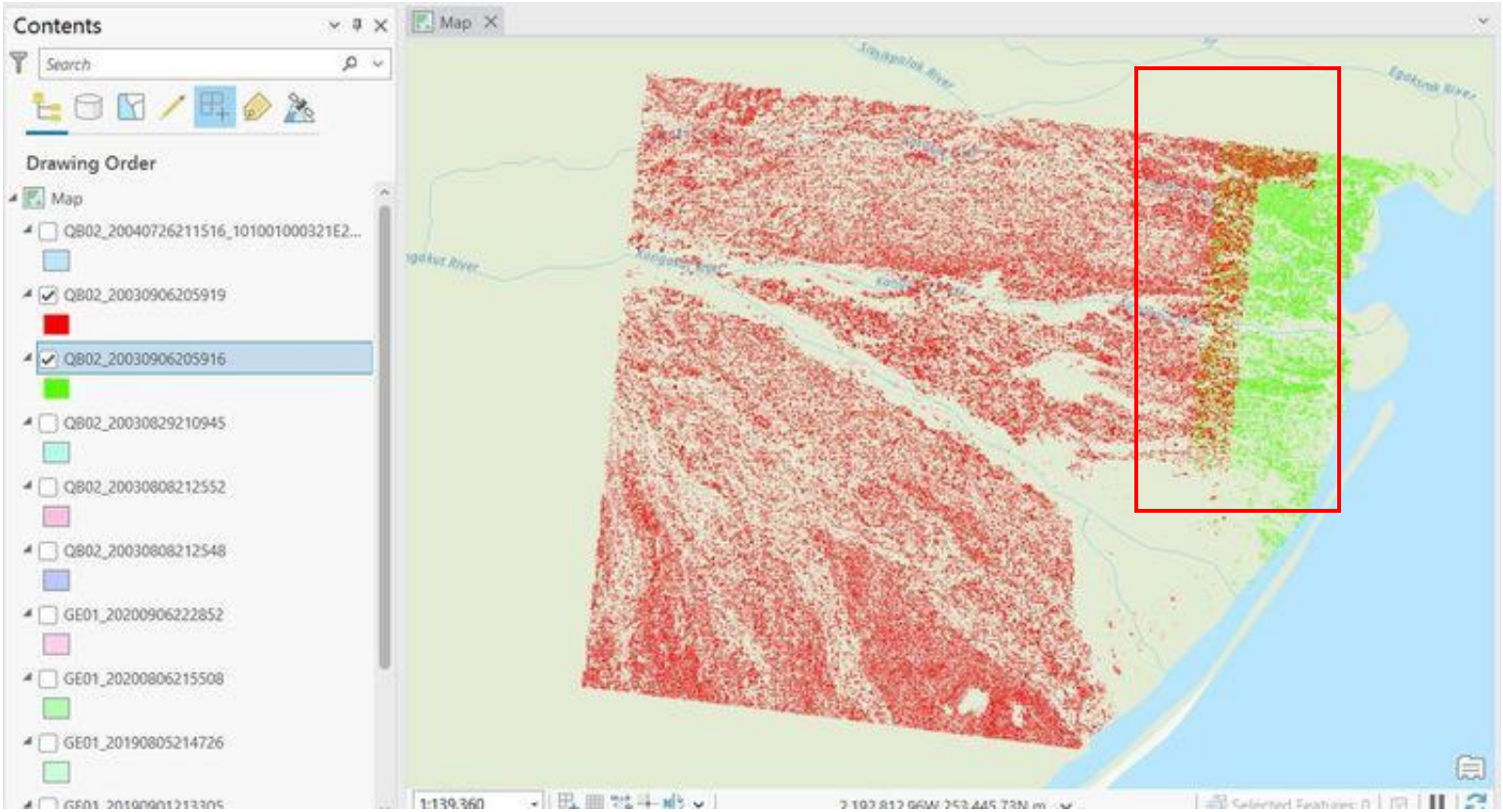


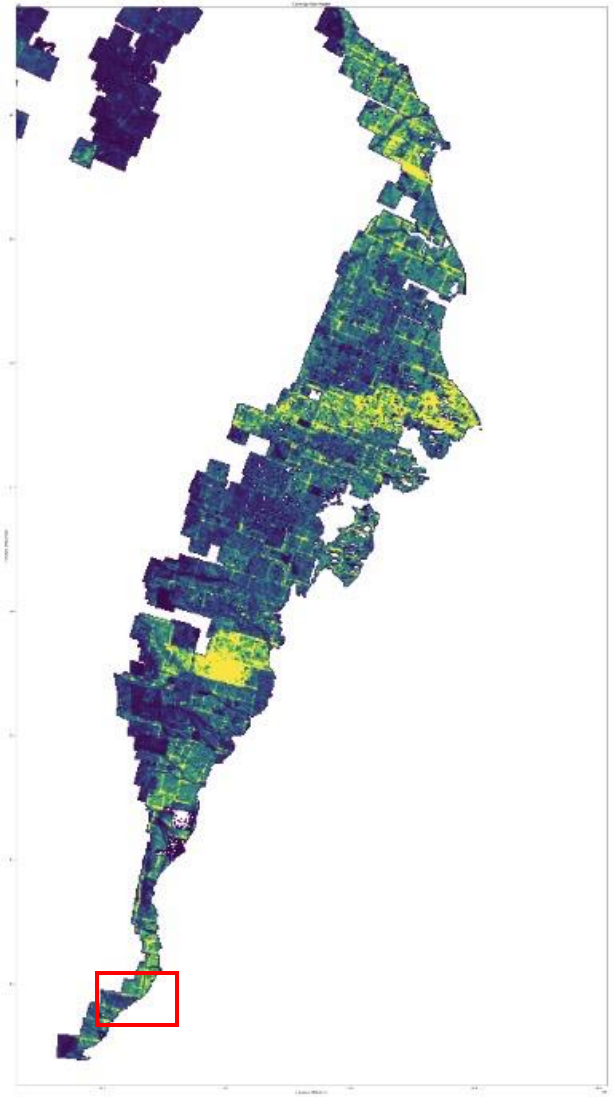

```
Overall Min: 0.0
Overall Max: 0.31268838
Overall Mean: 0.02122589
Overall Standard Deviation: 0.052890483
```

Overlapped area

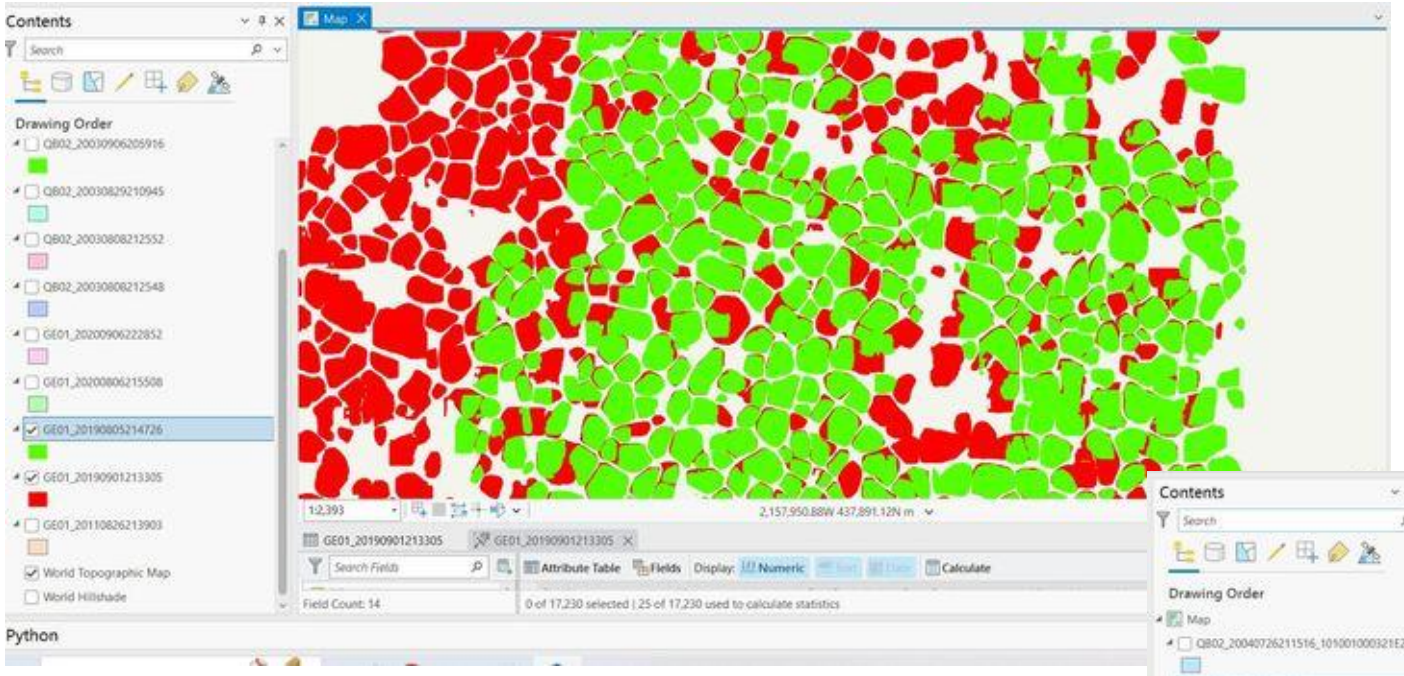


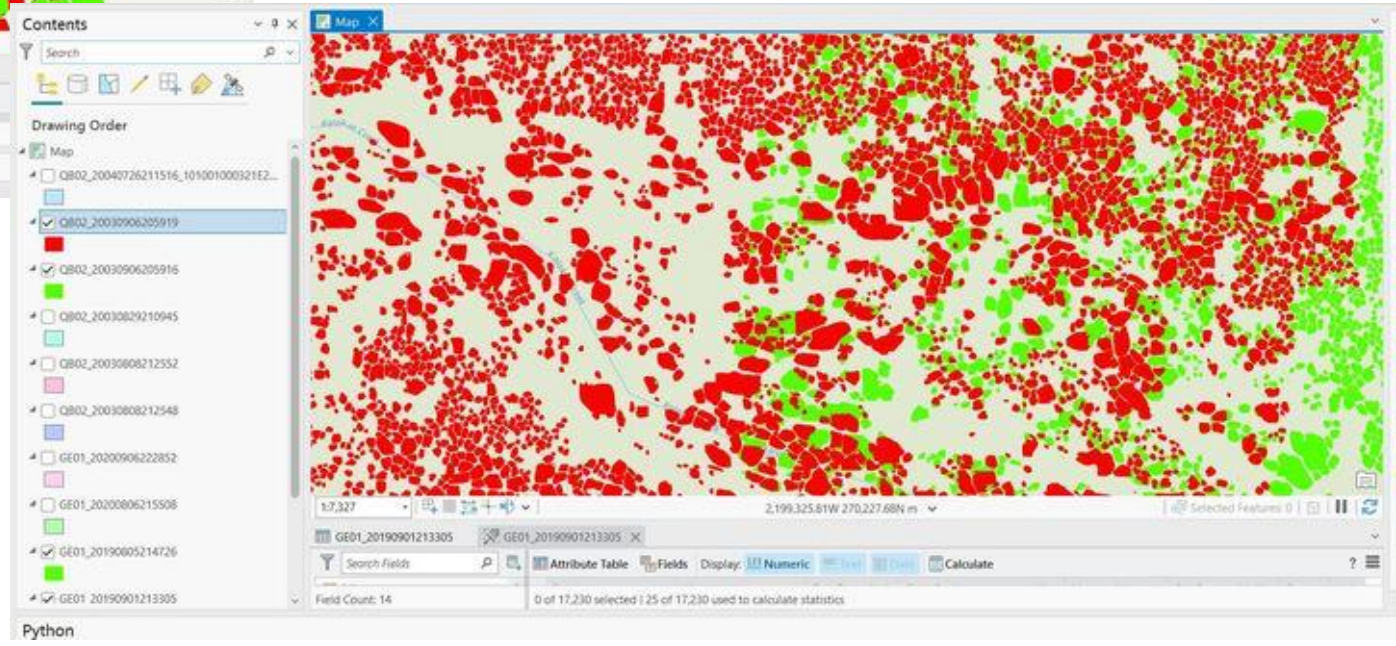
Overlapped area



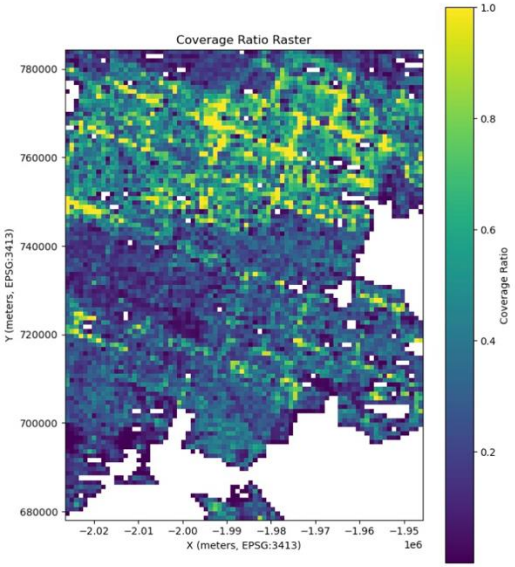


Overlapped area

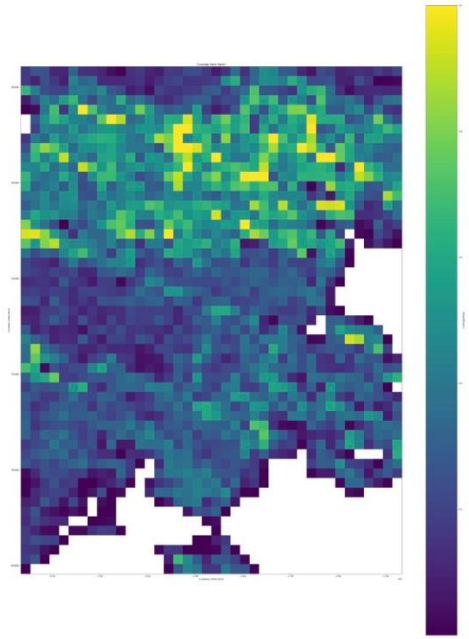




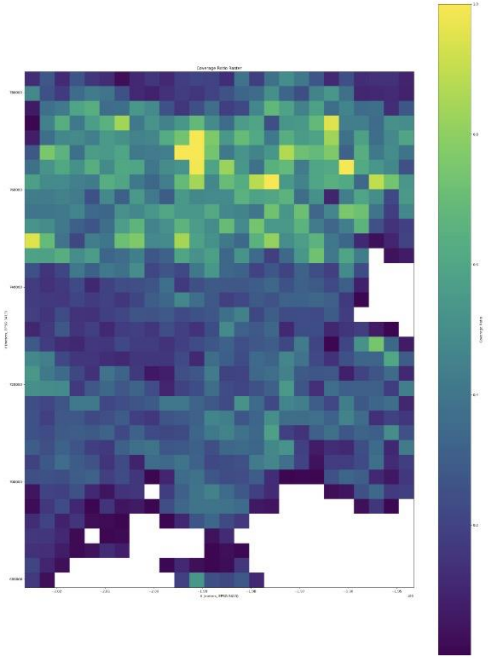
Aggregate from 1*1km



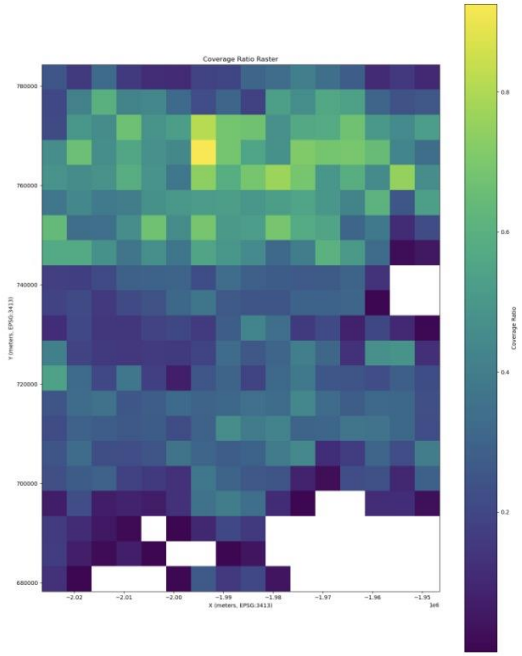
1*1km



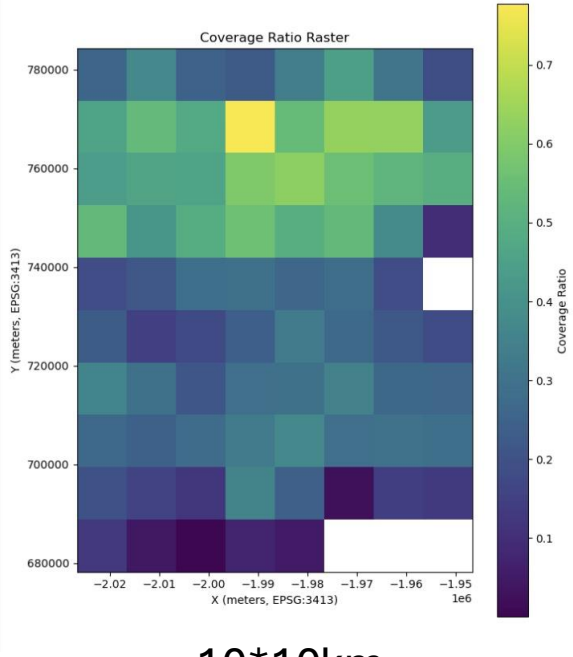
2*2km



3*3km



5*5km



10*10km

Geopackage downloading

```
(base) xchen@cici2labasuedu:~/data/geopackages/iwp_geopackage_high/WGS1984Quad/15$ ls
100     11599   13200   14786   16434   18029   19727   2408    3994    43410   44996   4658    4950    6536    8375
1000    116     13201   14787   16435   1803    19728   2409    3995    43411   44997   46580   4951    6537    8376
10000   11600   13202   14788   16436   18030   19729   241     3996    43412   44998   46581   4952    6538    8377
10001   11601   13203   14789   16437   18031   19730   2410    3997    43413   44999   46582   4953    6539    8378
10002   11602   13204   1479    16438   18032   19731   2411    3998    43414   45      46583   4954    654     8379
10003   11603   13205   14790   16439   18033   19732   2412    3999    43415   450     46584   4955    6540    8380
10004   11604   13206   14791   1644    18034   19733   2413    40      43416   4500    46585   4956    6541    8381
10005   11605   13207   14792   16440   18035   19734   2414    400     43417   45000   46586   4957    6542    8382
10006   11606   13208   14793   16441   18036   19735   2415    4000    43418   45001   46587   4958    6543    8383
10007   11607   13209   14794   16442   18037   19736   2416    4001    43419   45002   46588   4959    6544    8384
10008   11608   1321    14795   16443   18038   19737   2417    4002    4342    45003   46589   496     6545    8385
10009   11609   13210   14796   16444   18039   19738   2418    4003    43420   45004   4659    4960    6546    8386
1001    11610   13211   14797   16445   1804    19739   2419    4004    43421   45005   46590   4961    6547    8387
10010   11611   13212   14798   16446   18040   19740   242     4005    43422   45006   46591   4962    6548    8388
10011   11612   13213   14799   16447   18041   19741   2420    4006    43423   45007   46592   4963    6549    8389
10012   11613   13214   148     16448   18042   19742   2421    4007    43424   45008   46593   4964    655     8390
10013   11614   13215   1480    16449   18043   19743   2422    4008    43425   45009   46594   4965    6550    8391
10014   11615   13216   14800   1645    18044   19744   2423    4009    43426   4501    46595   4966    6551    8392
10015   11616   13217   14801   16450   18045   19745   2424    401     43427   45010   46596   4967    6552    8393
10016   11617   13218   14802   16451   18046   19746   2425    4010    43428   45011   46597   4968    6553    8394
10017   11618   13219   14803   16452   18047   19747   2426    4011    43429   45012   46598   4969    6554    8395
10018   11619   1322    14804   16453   18048   19748   2427    4012    4343    45013   46599   497     6555    8396
10019   11620   13220   14805   16454   18049   19749   2428    4013    43430   45014   466     4970    6556    8397
1002    11621   13221   14806   16455   1805    19750   2429    4014    43431   45015   4660    4971    6557    8398
10020   11622   13222   14807   16456   18050   19751   243     4015    43432   45016   46600   4972    6558    8399
10021   11623   13223   14808   16457   18051   19752   2430    4016    43433   45017   46601   4973    6559    84
10022   11624   13224   14809   16458   18052   19753   2431    4017    43434   45018   46602   4974    656     8400
10023   11625   13225   1481    16459   18053   19754   2432    4018    43435   45019   46603   4975    6560    8401
10024   11626   13226   14810   1646    18054   19755   2433    4019    43436   4502    46604   4976    6561    8402
10025   11627   13227   14811   16460   18055   19756   2434    402     43437   45020   46605   4977    6562    8403
10026   11628   13228   14812   16461   18056   19757   2435    4020    43438   45021   46606   4978    6563    8404
10027   11629   13229   14813   16462   18057   19758   2436    4021    43439   45022   46607   4979    6564    8405
10028   11630   1323    14814   16463   18058   19759   2437    4022    4344    45023   46608   498     6565    8406
10029   11631   13230   14815   16464   18059   19760   2438    4023    43440   45024   46609   4980    6566    8407
1003    11632   13231   14816   16465   1806    19761   2439    4024    43441   45025   4661    4981    6567    8408
10030   11633   13232   14817   16466   18060   19762   244     4025    43442   45026   46610   4982    6568    8409
```

Directories: >40,000

Z-x-y

After we process the input data, the output tiles are organized by z-y-x indices for the tileset, and we no longer use the original file hierarchy. There are two main reasons for this

You could check which tiles intersect with Alaska, e.g. in python you could use morecantile and and an overlay operation in geopandas.

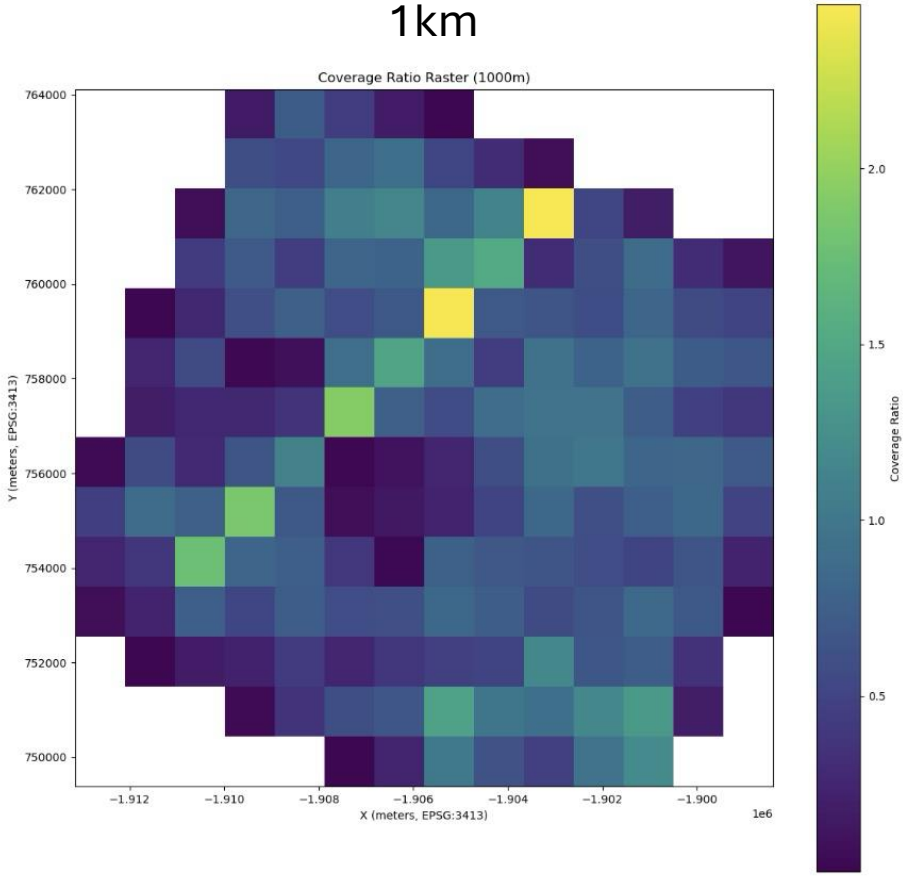
~ 90 geopackages

```
(alaska_shpMapping2) (base) xchen@cici2labasuedu:~/data/geopackages/iwp_geopackage_high/WGS1984Quad/15/100$ ls
3358.gpkg   3368.gpkg   3378.gpkg   3388.gpkg   3398.gpkg   3408.gpkg   3418.gpkg   3428.gpkg   3452.gpkg
3359.gpkg   3369.gpkg   3379.gpkg   3389.gpkg   3399.gpkg   3409.gpkg   3419.gpkg   3429.gpkg   3455.gpkg
3360.gpkg   3370.gpkg   3380.gpkg   3390.gpkg   3400.gpkg   3410.gpkg   3420.gpkg   3430.gpkg   3465.gpkg
3361.gpkg   3371.gpkg   3381.gpkg   3391.gpkg   3401.gpkg   3411.gpkg   3421.gpkg   3431.gpkg   3470.gpkg
3362.gpkg   3372.gpkg   3382.gpkg   3392.gpkg   3402.gpkg   3412.gpkg   3422.gpkg   3432.gpkg   3471.gpkg
3363.gpkg   3373.gpkg   3383.gpkg   3393.gpkg   3403.gpkg   3413.gpkg   3423.gpkg   3433.gpkg   index.html
3364.gpkg   3374.gpkg   3384.gpkg   3394.gpkg   3404.gpkg   3414.gpkg   3424.gpkg   3434.gpkg   'index.html?C=D;O=A'
3365.gpkg   3375.gpkg   3385.gpkg   3395.gpkg   3405.gpkg   3415.gpkg   3425.gpkg   3435.gpkg   'index.html?C=M;O=A'
3366.gpkg   3376.gpkg   3386.gpkg   3396.gpkg   3406.gpkg   3416.gpkg   3426.gpkg   3436.gpkg   'index.html?C=N;O=D'
3367.gpkg   3377.gpkg   3387.gpkg   3397.gpkg   3407.gpkg   3417.gpkg   3427.gpkg   3440.gpkg   'index.html?C=S;O=A'
```

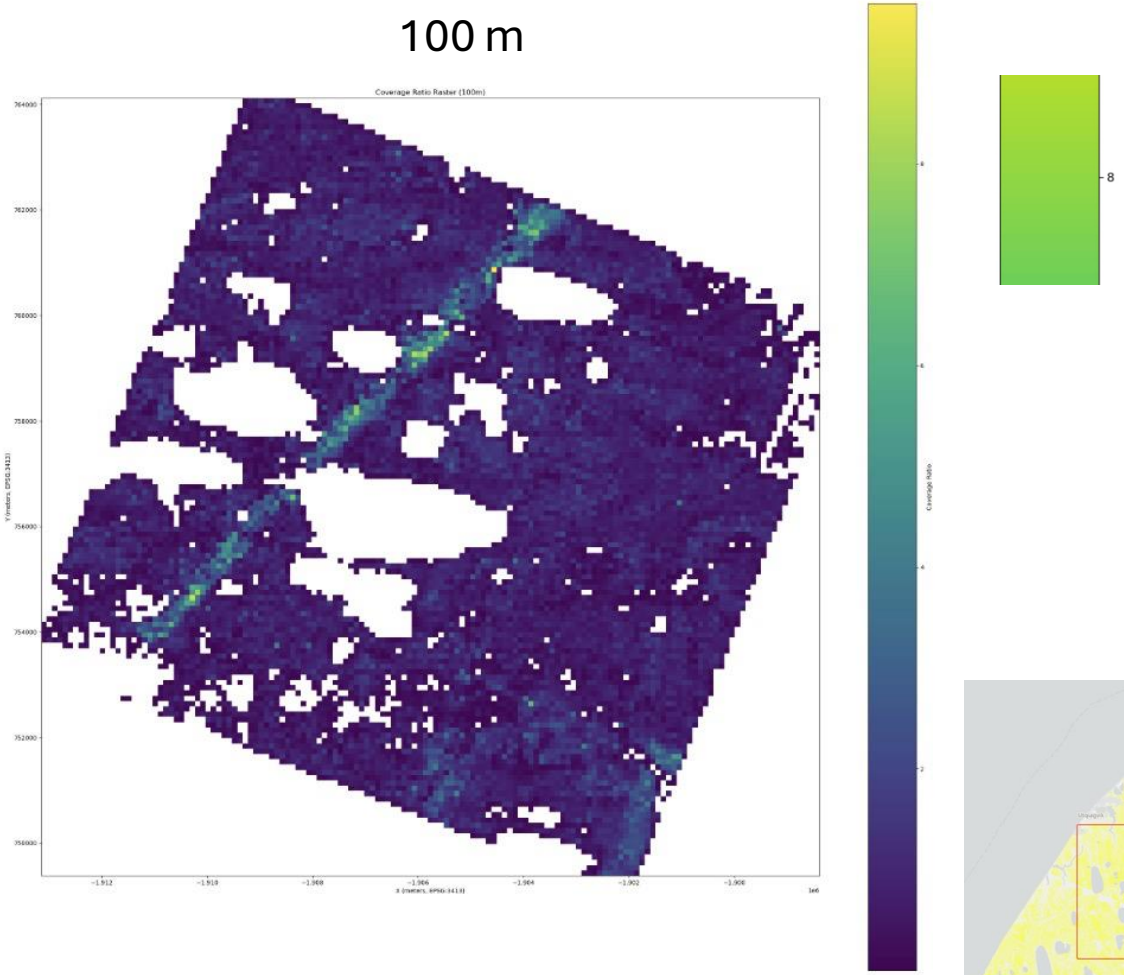
Coverage ratio raster at small scale

Before Deduplicated

1km



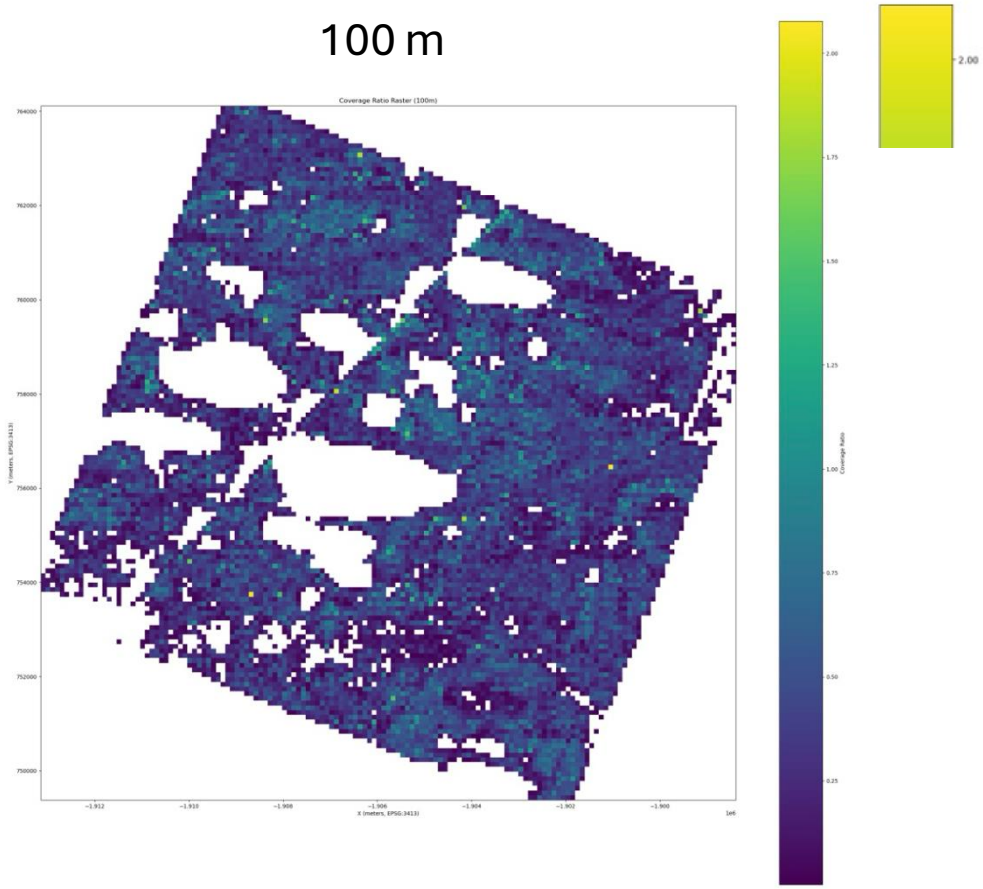
100 m



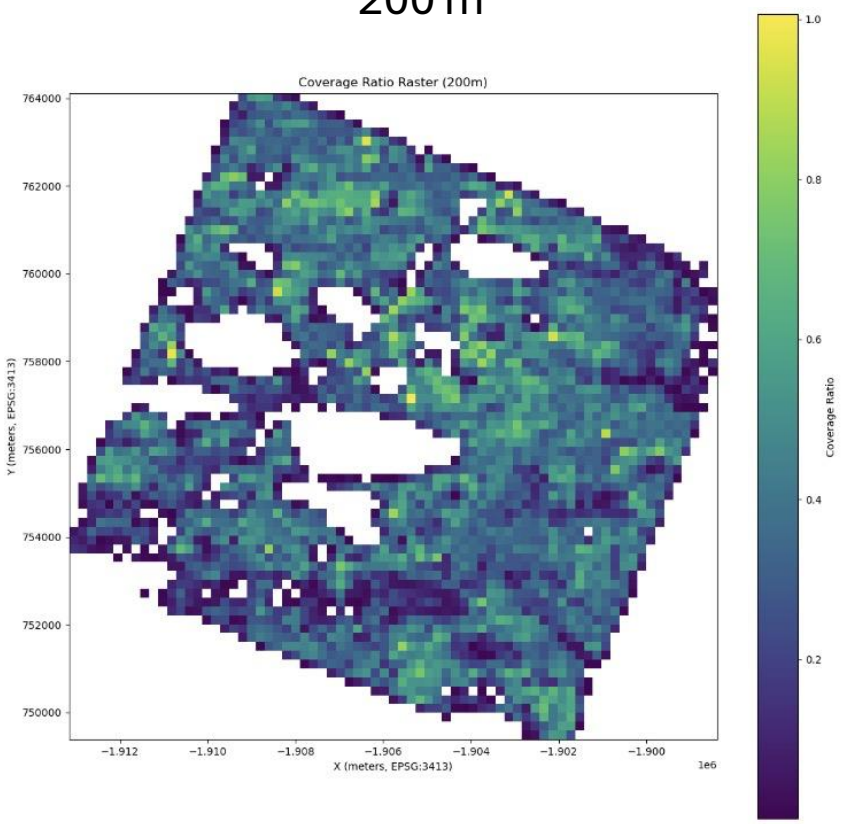
Coverage ratio raster at median scale

After Deduplicated

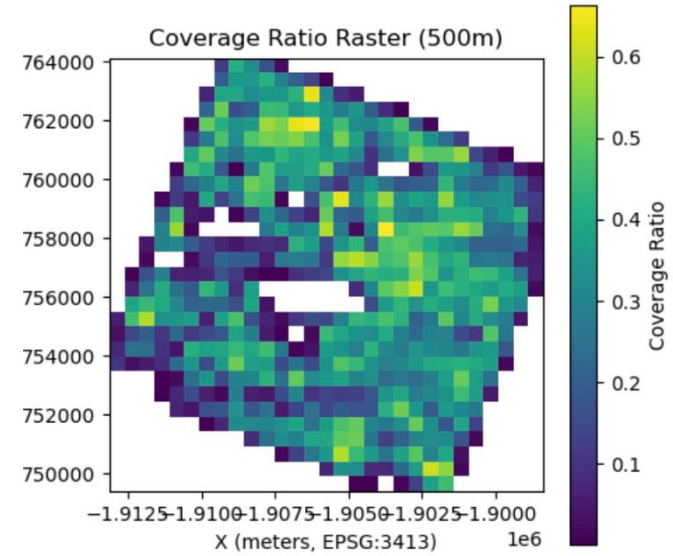
100 m



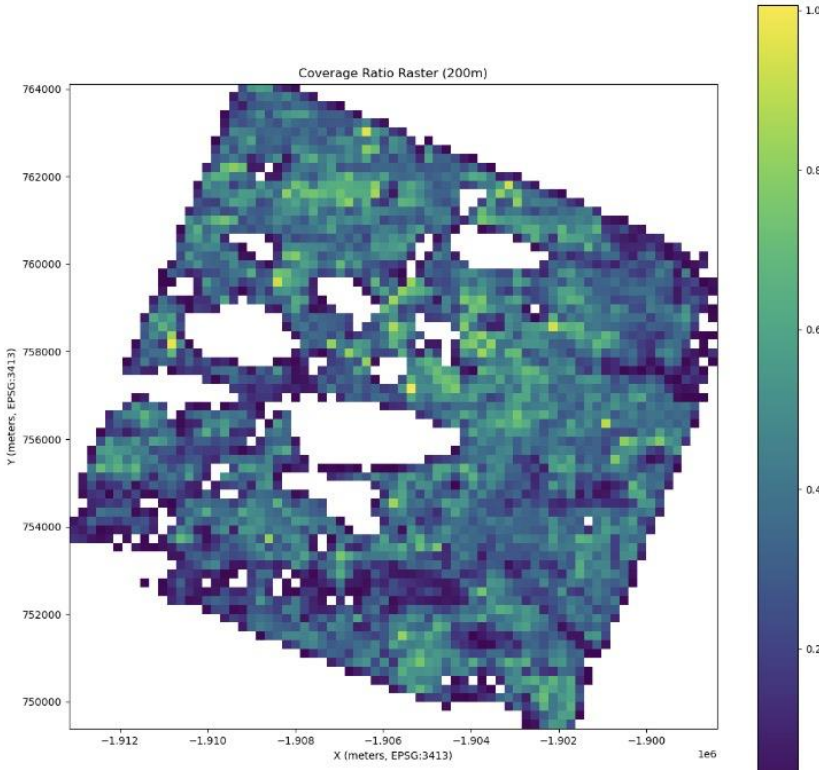
200 m



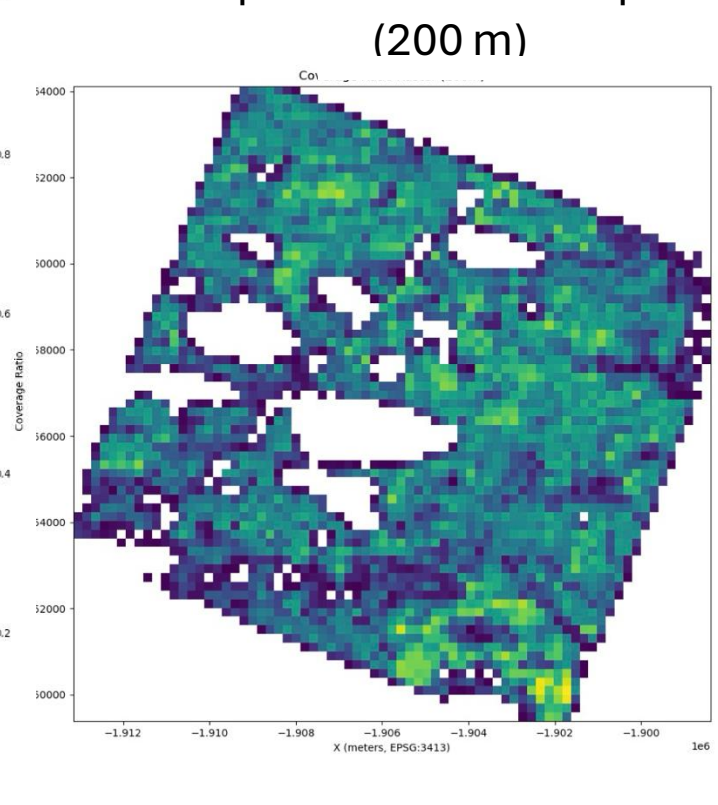
500 m



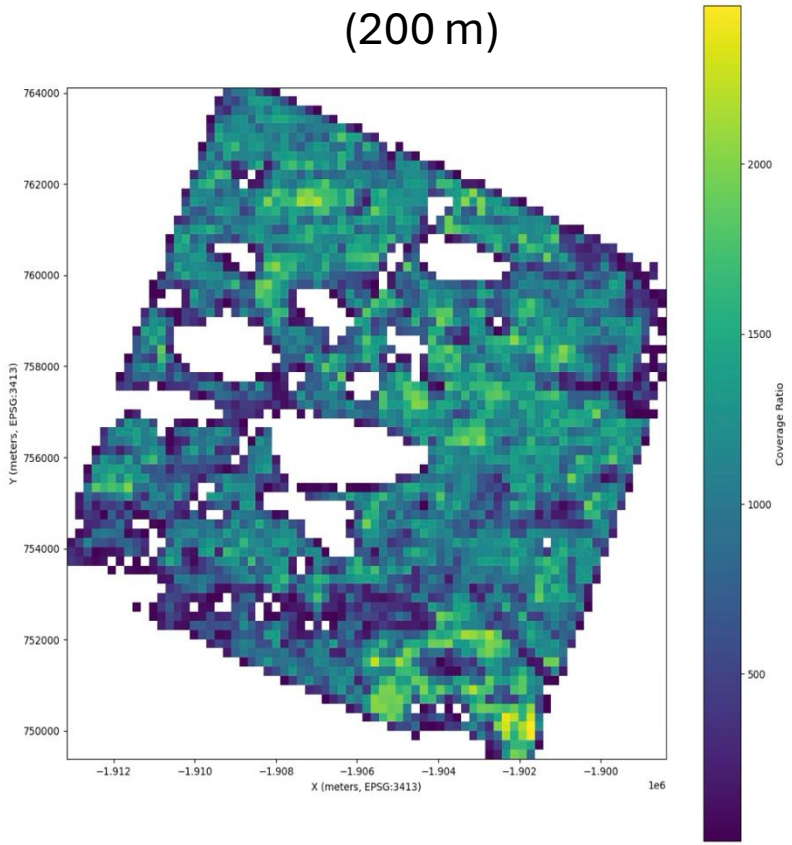
Different maps



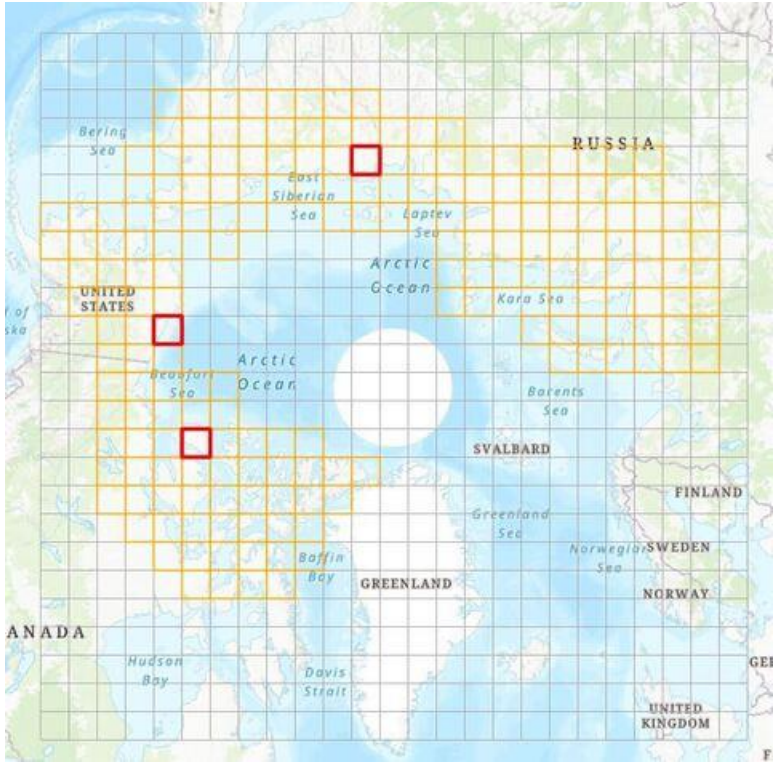
perimeter sum map (200 m)



Length sum map (200 m)



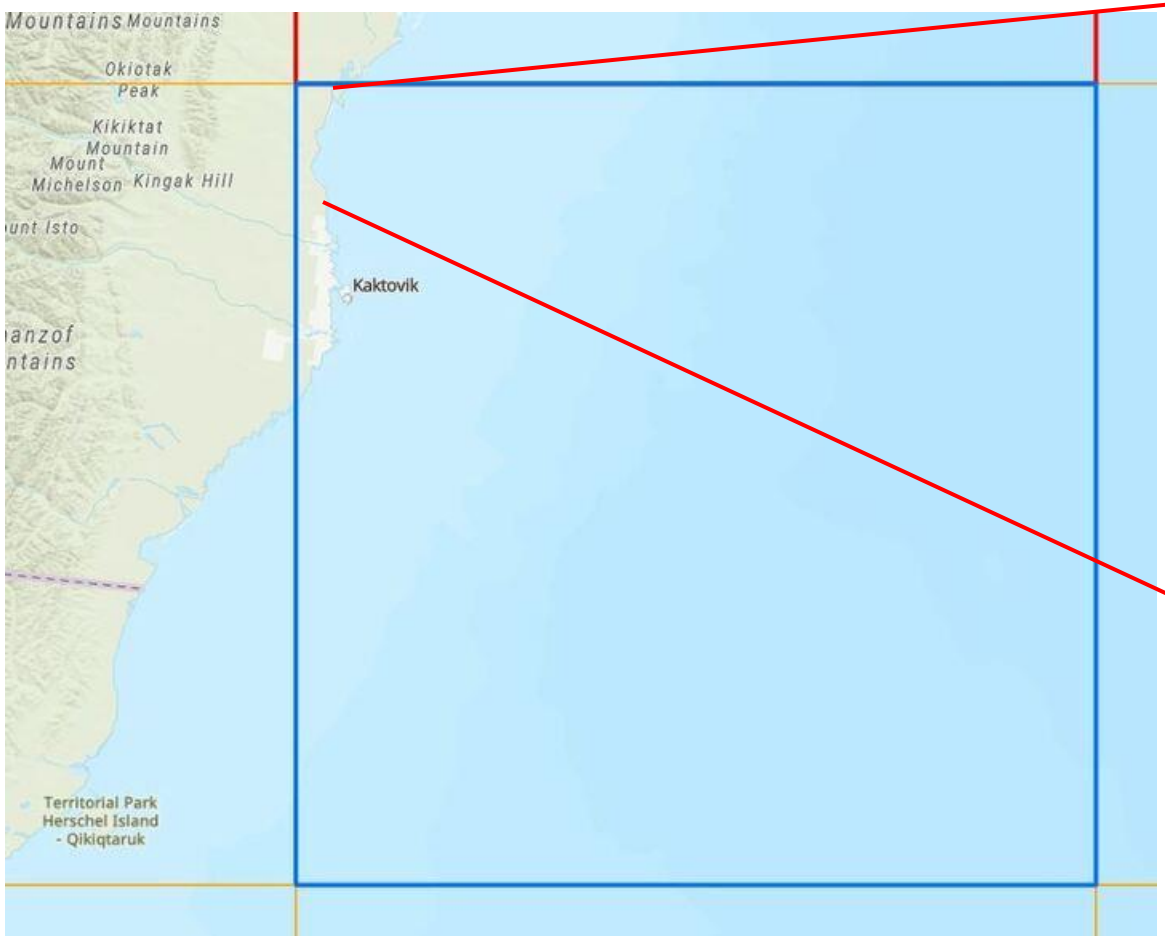
Create Grids (256km*256km)

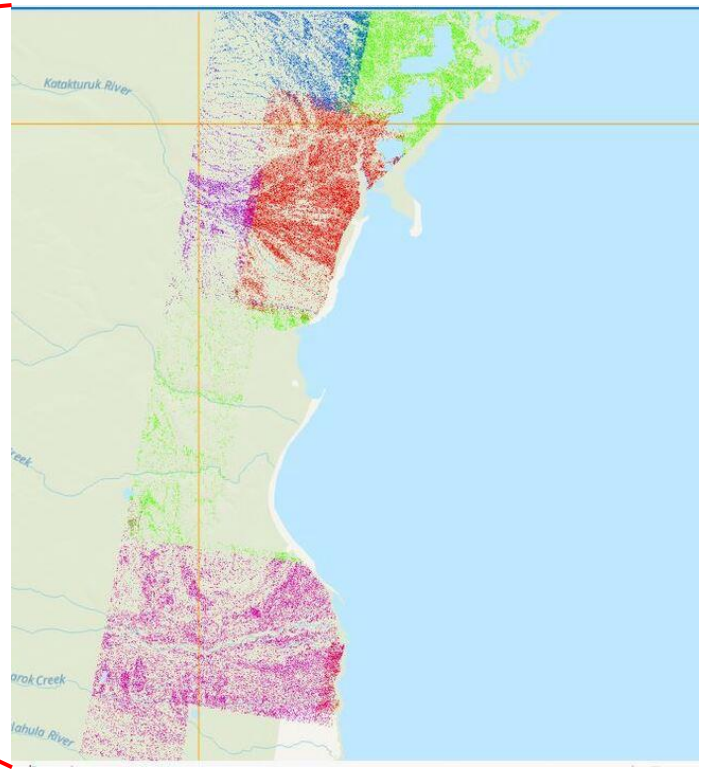


EPSG: 3413

EPSG: 4326

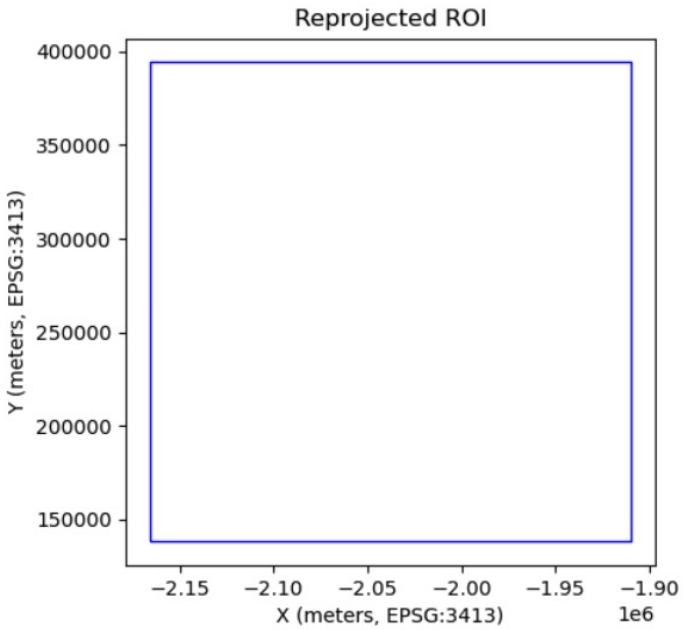
Select a grid to test



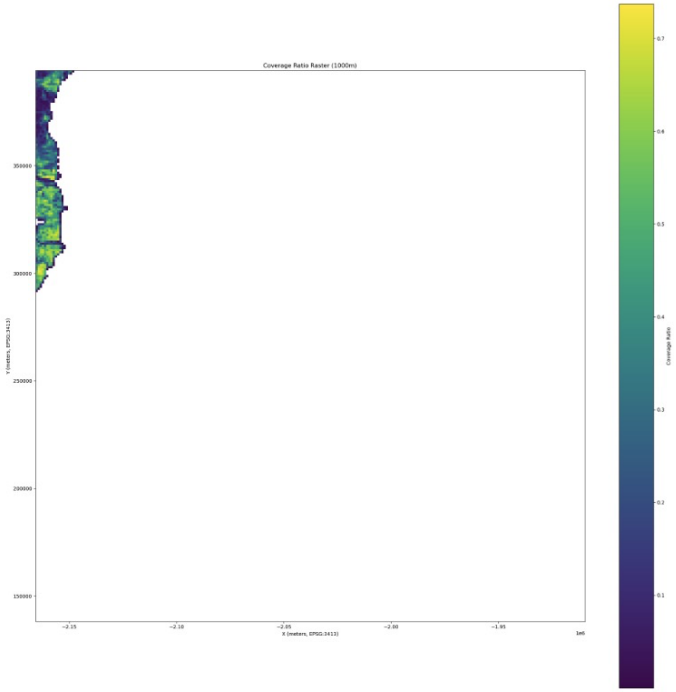


Validation: projection

EPSG: 3413



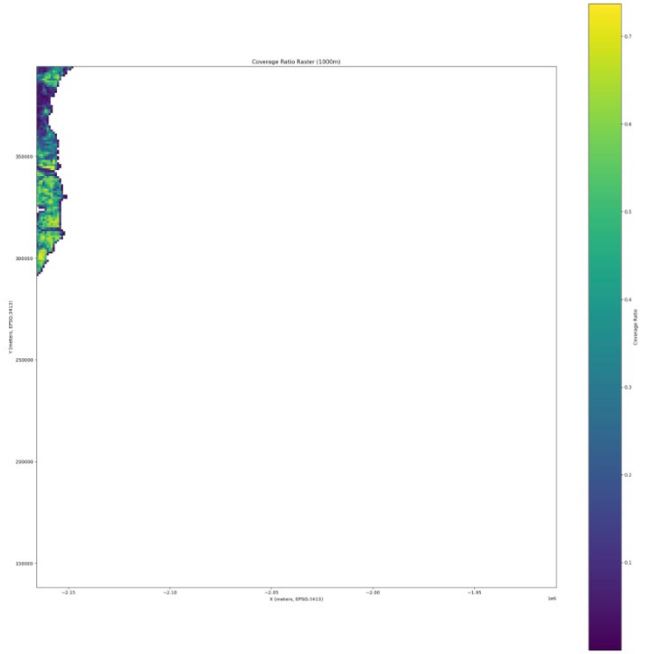
EPSG: 3413



EPSG: 3413

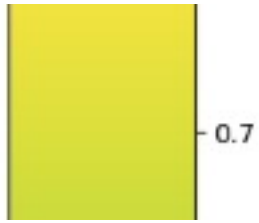
256*256 pixels

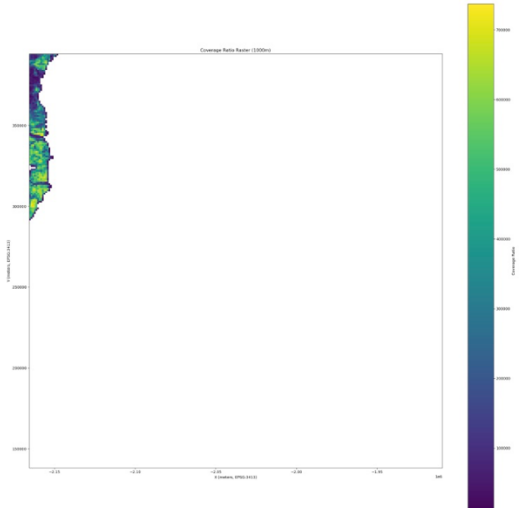
Validation: value



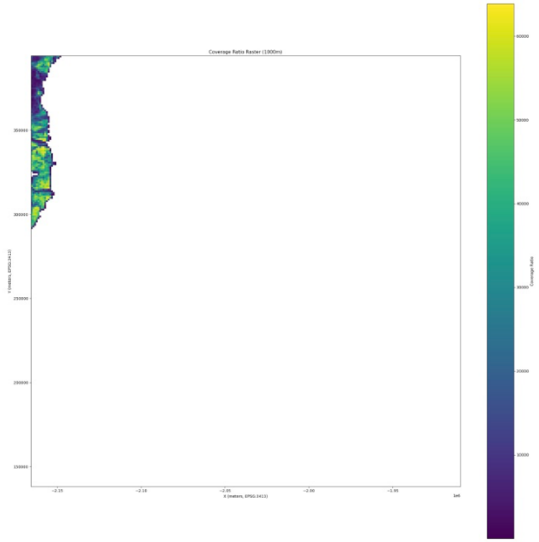
Coverage Ratio

Max





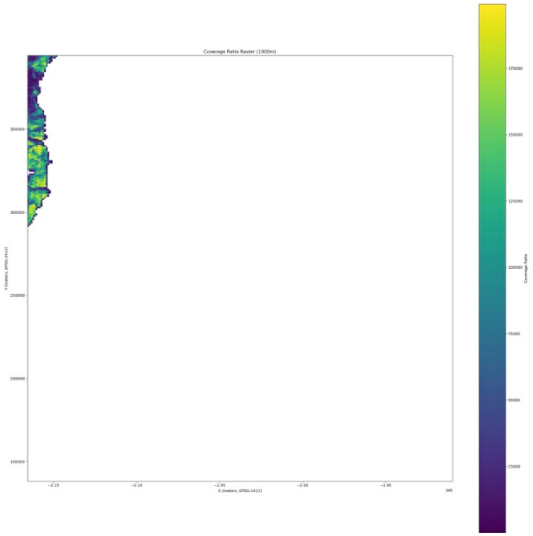
Coverage sum
Max: 700,000



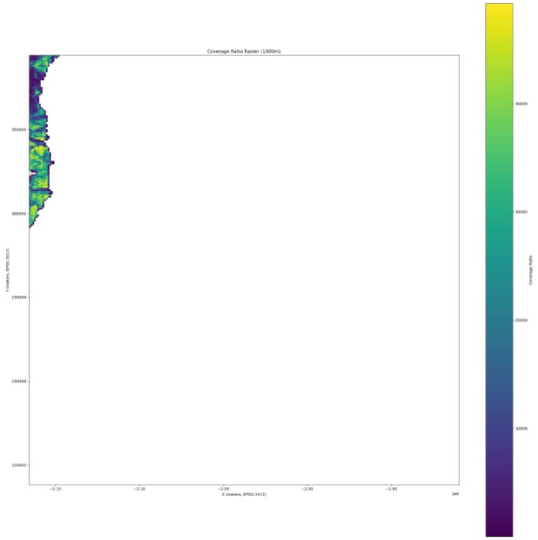
Length sum
Max: 60,000

Plot size: 25*25

IWP Count

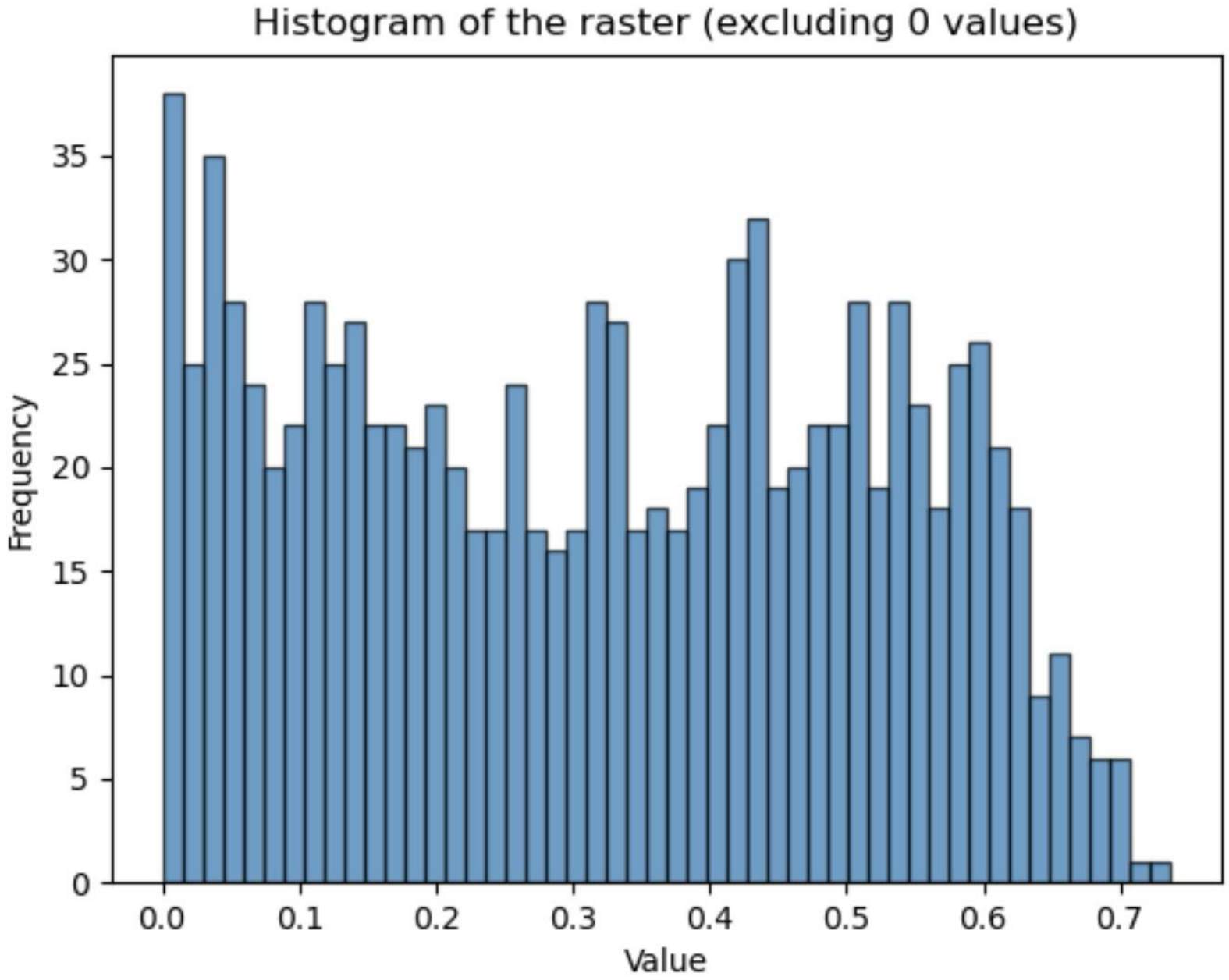


Perimeter sum
Max: 175,000



Width sum
Max: 40,000

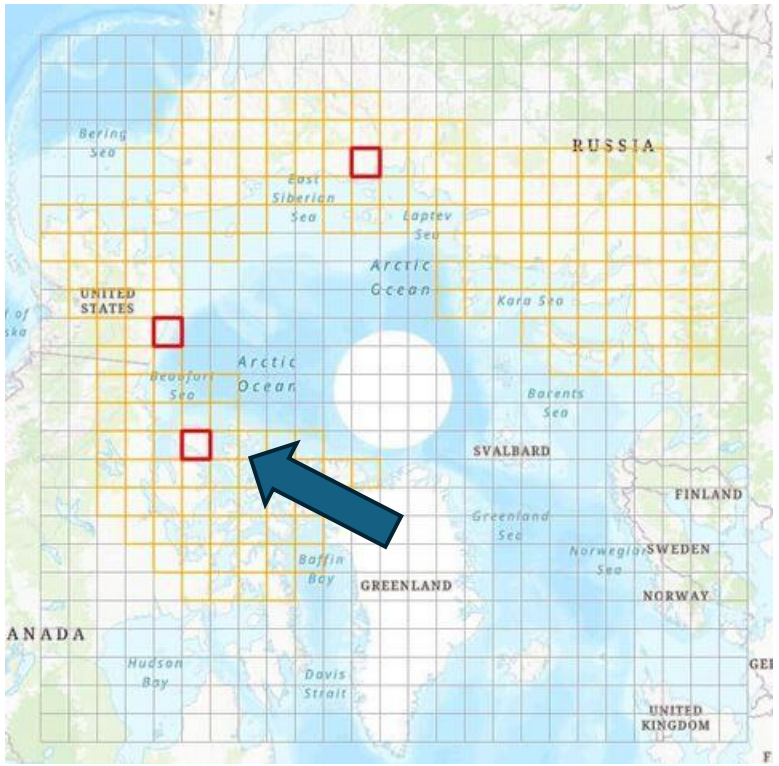
Validation: value



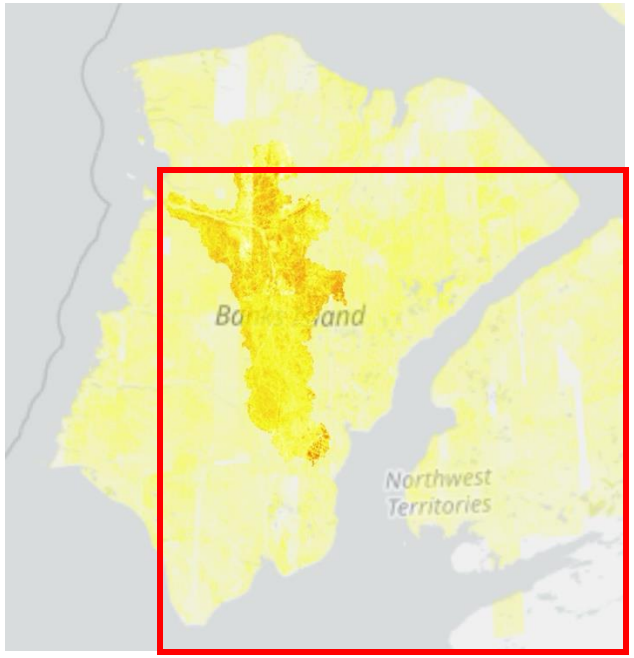
Remaining issues

~600,000 gpkg
~130~200G gpkg

1. Data storage for dense area



EPSG: 3413



EPSG: 3413

EPSG: 4326

EPSG: 4326

Process validation

From 256*256 1km pixels, we will select the two pixels

In our workflow

Convert the 256km*256km grid into a 256*256 raster (with 1km as a pixel size)

256*256 1km pixels
Choose two of pixels

Query 1

Two pixels generated via ArcGIS tools

Code: to convert to query

Query 2

Compare query

get area sum in database

Overlapped GPKG in a pixel

get area sum in database

Compare data

Get area sum in ArcGIS

~708,460 m^2

BBOX coordinates check

Query 1: generated from our raster

(1, 13)

SELECT area, perimeter, length, width, geom_centroid FROM grid_gpkg_3413 polygons WHERE ST_within(polygons.geom_centroid, ST_GeomFromText('POLYGON((-2153757.915805478 393242.08426331764, -2153757.915805478 394242.08426331764, -2152757.915805478 394242.08426331764, -2152757.915805478 393242.08426331764, -2153757.915805478 393242.08426331764))', 3413)) AND staging_duplicated = false

→389180.4

(12, 4)

SELECT area, perimeter, length, width, geom_centroid FROM grid_gpkg_3413 polygons WHERE ST_within(polygons.geom_centroid, ST_GeomFromText('POLYGON((-2162757.9158000015 382242.0841999985, -2161757.9158000015 382242.0841999985, -2161757.9158000015 383242.0841999985, -2162757.9158000015 383242.0841999985, -2162757.9158000015 382242.0841999985))', 3413)) AND staging_duplicated = false

→319280.18

BBOX coordinates check

Query 2: generated from ArcGIS and converted via code

(1, 13)

SELECT area, perimeter, length, width, geom_centroid FROM grid_gpkg_3413 polygons WHERE ST_within(polygons.geom_centroid, ST_GeomFromText('POLYGON((-2153757.915805478 393242.08426331764, -2153757.915805478 394242.08426331764, -2152757.915805478 394242.08426331764, -2152757.915805478 393242.08426331764, -2153757.915805478 393242.08426331764))', 3413)) AND staging_duplicated = false

Number: 2785

→389180.408

(12, 4)

SELECT area, perimeter, length, width, geom_centroid FROM grid_gpkg_3413 polygons WHERE ST_within(polygons.geom_centroid, ST_GeomFromText('POLYGON((-2162757.915805478 382242.08426331764, -2162757.915805478 383242.08426331764, -2161757.915805478 383242.08426331764, -2161757.915805478 382242.08426331764, -2162757.915805478 382242.08426331764))', 3413)) AND staging_duplicated = false

Number: 1190

→319280.195

Total area sum:
389180.408+319280.195 = 708460.603

Data in database vs. geopackages data

Summarize: Center in

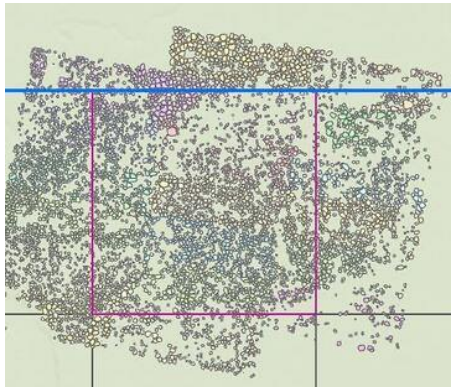
	G	H	I	J	K	L
1	Area	CentroidX	CentroidY	Perimeter	Length	Width
3955	214.572	-2161845.664	382414.794	65.013	19.589	14.234
3956	426.427	-2161900.34	382433.877	80.24	23.578	23.318
3957	139.862	-2161919.297	382436.698	46.198	15.687	11.582
3958	50.771	-2161867.973	382477.609	27.843	9.1	6.783
3959	103.209	-2161878.4	382481.406	41.173	13.143	9.514
3960	199.666	-2161874	382443.722	56.545	17.383	15.305
3961	319.623	-2161813.553	382441.414	67.553	21.407	18.992
3962	299.104	-2161785.53	382446.53	67.86	20.03	19.518
3963	200.192	-2161795.245	382431.044	55.033	16.133	15.263
3964	233.864	-2161792.776	382462.358	73.459	20.35	16.96
3965	212.293	-2161827.955	382427.084	57.983	19.057	14.508
3966	98.649	-2161798.697	382411.144	41.968	14.22	9.304
3967	156.698	-2161810.109	382417.886	51.354	16.96	12.295
3968	73.044	-2161824.257	382451.039	35.217	11.466	8.551
3969	78.831	-2161844.082	382459.405	34.676	10.599	9.514
3970	420.815	-2161842.919	382442.734	82.947	26.888	18.655
3971	87.25	-2161785.975	382405.293	36.73	10.341	10.175
3972	366.623	-2161873.119	382376.372	79.688	26.859	18.15
3973	481.845	-2161852.511	382368.733	86.491	29.98	20.651
3974	91.459	-2161828.946	382384.356	37.137	11.717	9.621
3975	205.102	-2161826.282	382369.619	55.378	16.54	15.324
3976	67.607	-2161842.797	382380.366	31.234	9.751	8.687
3977	708460.603			204080.118	64540.57	50296.361
3978						

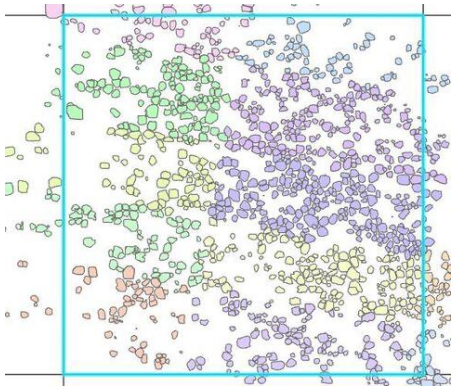
The previous results:
708460.603

127953.55
+76126.57
=204080.12

40786.008
+ 23754.56
=64540.568

31559.664
+18736.695
=50296.359





Data installation

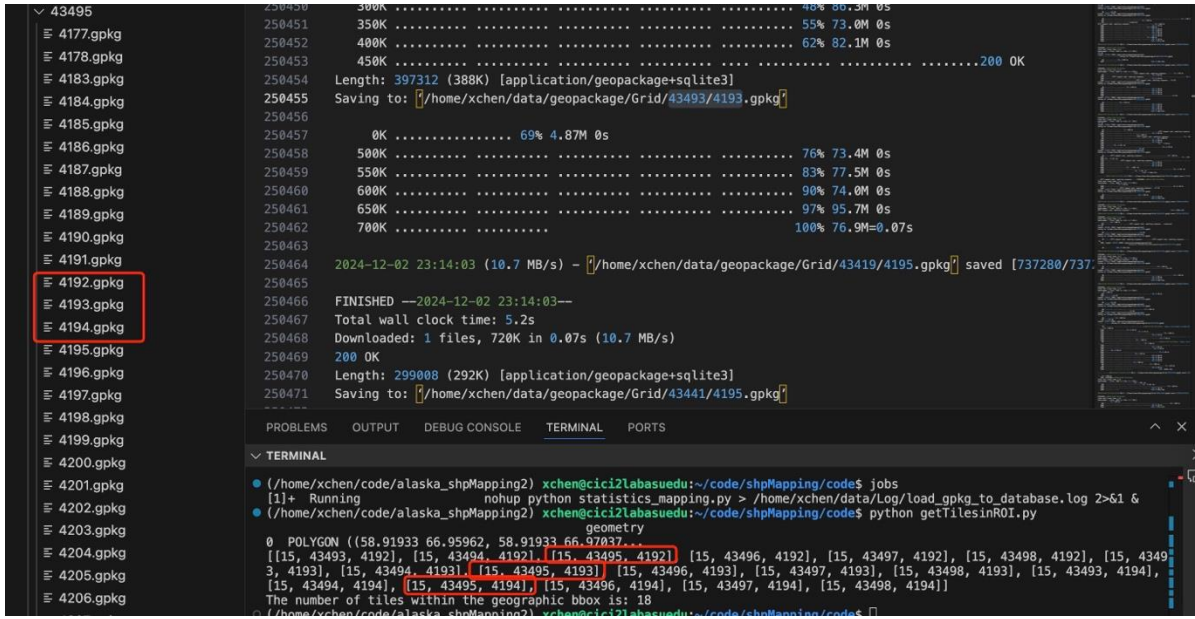
- Uninterrupted
  - Retry up to 30 times
  - Timeout: 120 seconds

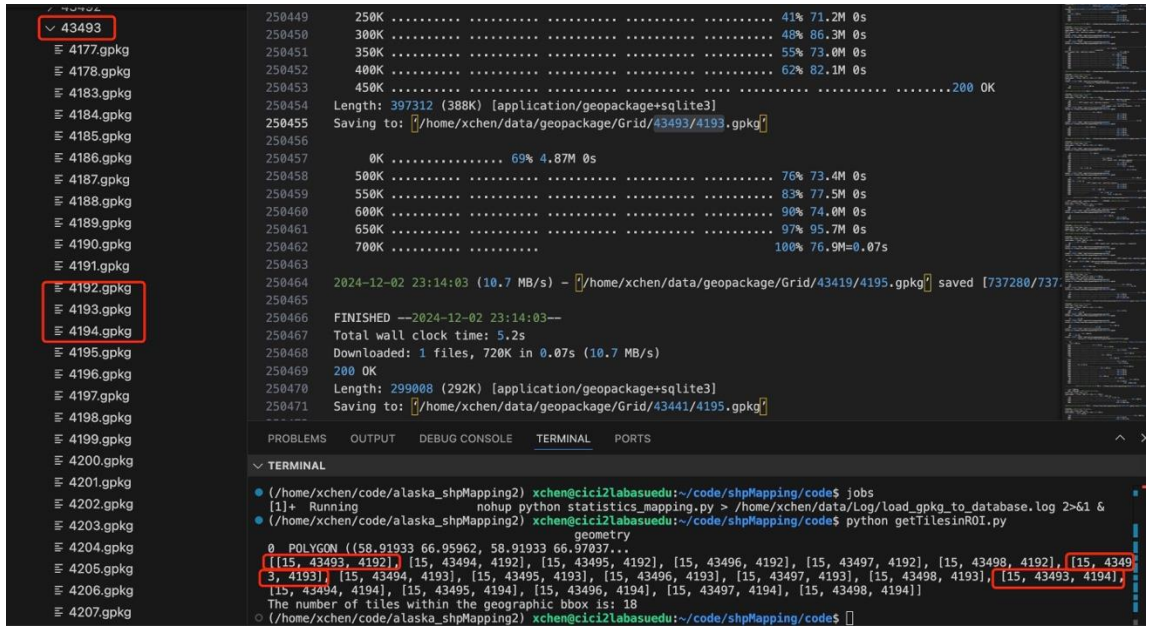
➢No fetch failures due to various network connection issues (Connection, timeout, SSL, etc.)

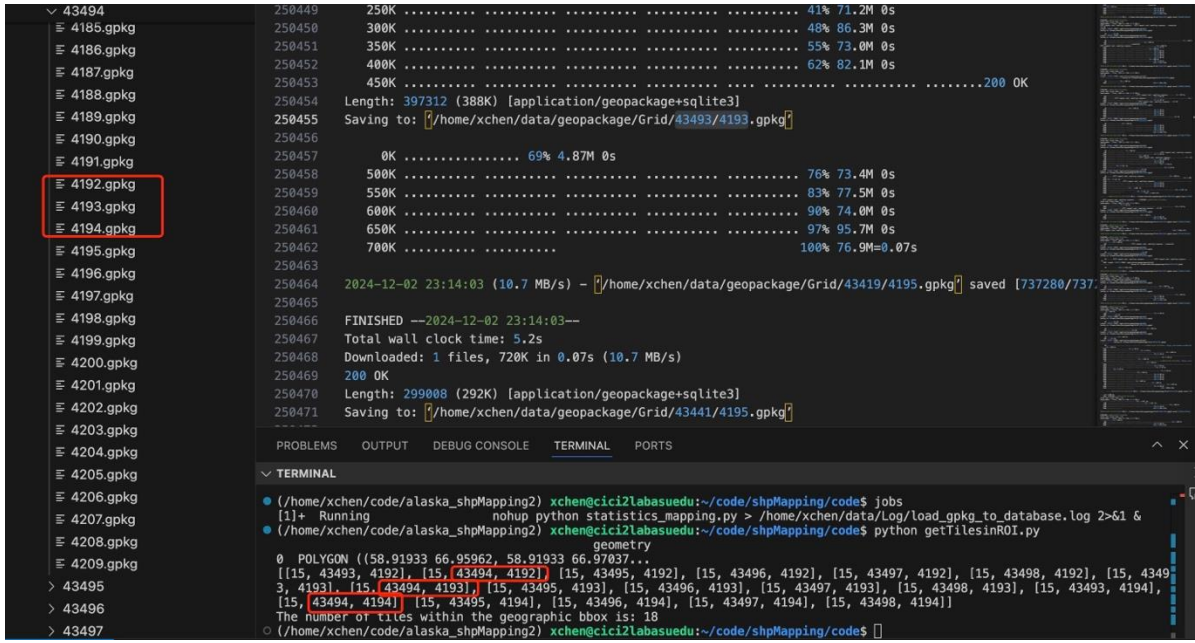
➢Total runtime for this grid: 5719 seconds (1.5h)

Data installation

- Validation
  - Randomly select a pixel: Compare the tiles within those pixels with installed GPKG



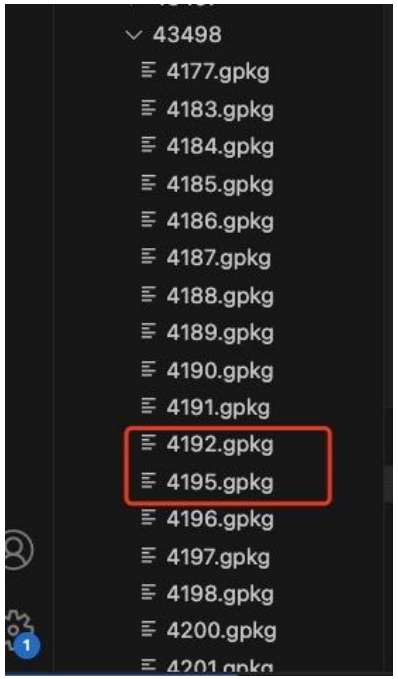




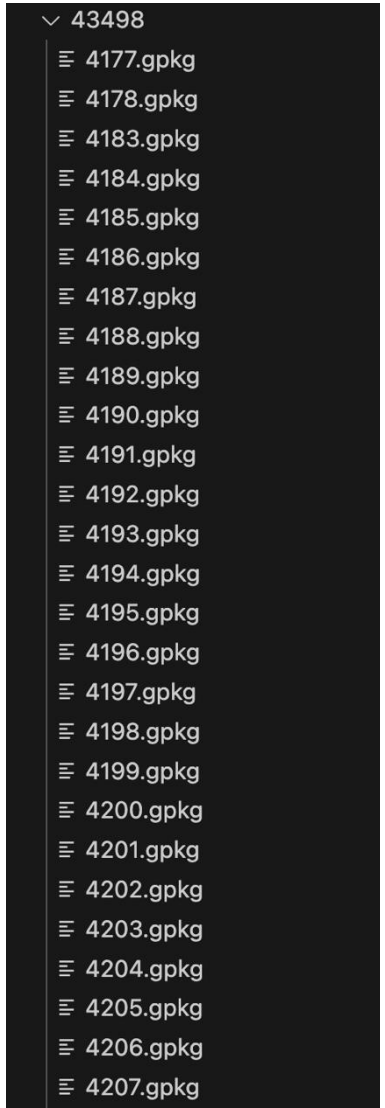
Data installation

- Validation
  - Randomly selected a Tile X: Check continuous tiles

Previous



Current



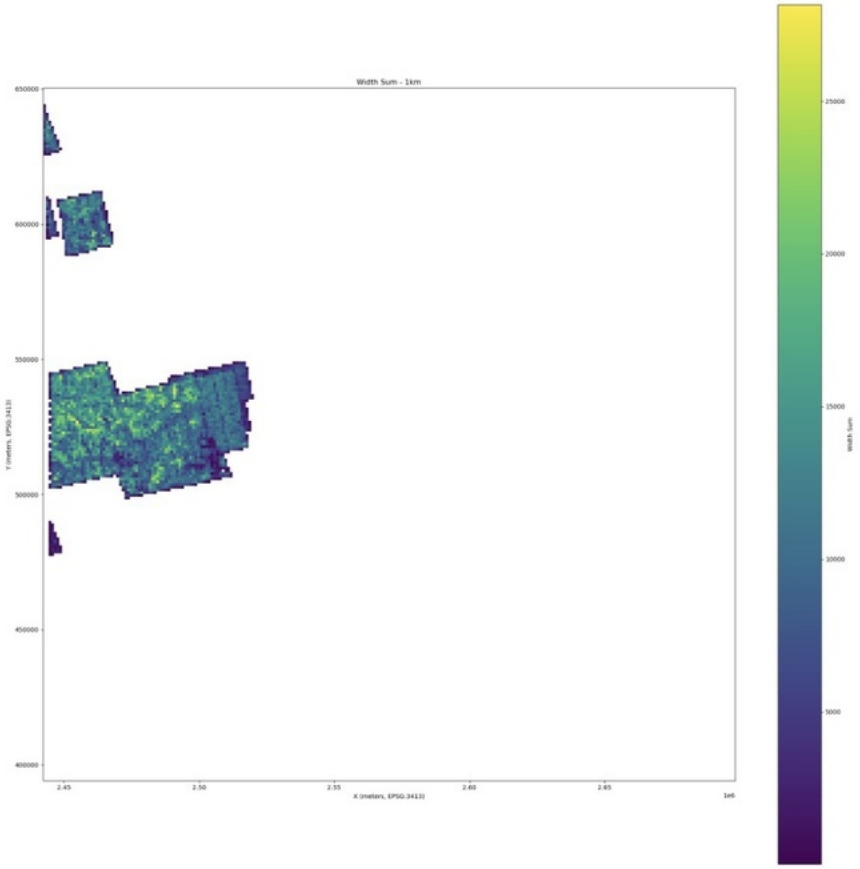
Data installation

- Validation
  - Randomly selected a Tile X: Check continuous tiles

Previous		Current
18,826 GPKG		22,157 GPKG

Total time in processing 1 grid

- Data installation: 1.5h
- Data loading into database: 1.5h
- Ploting data: 0.5~1h



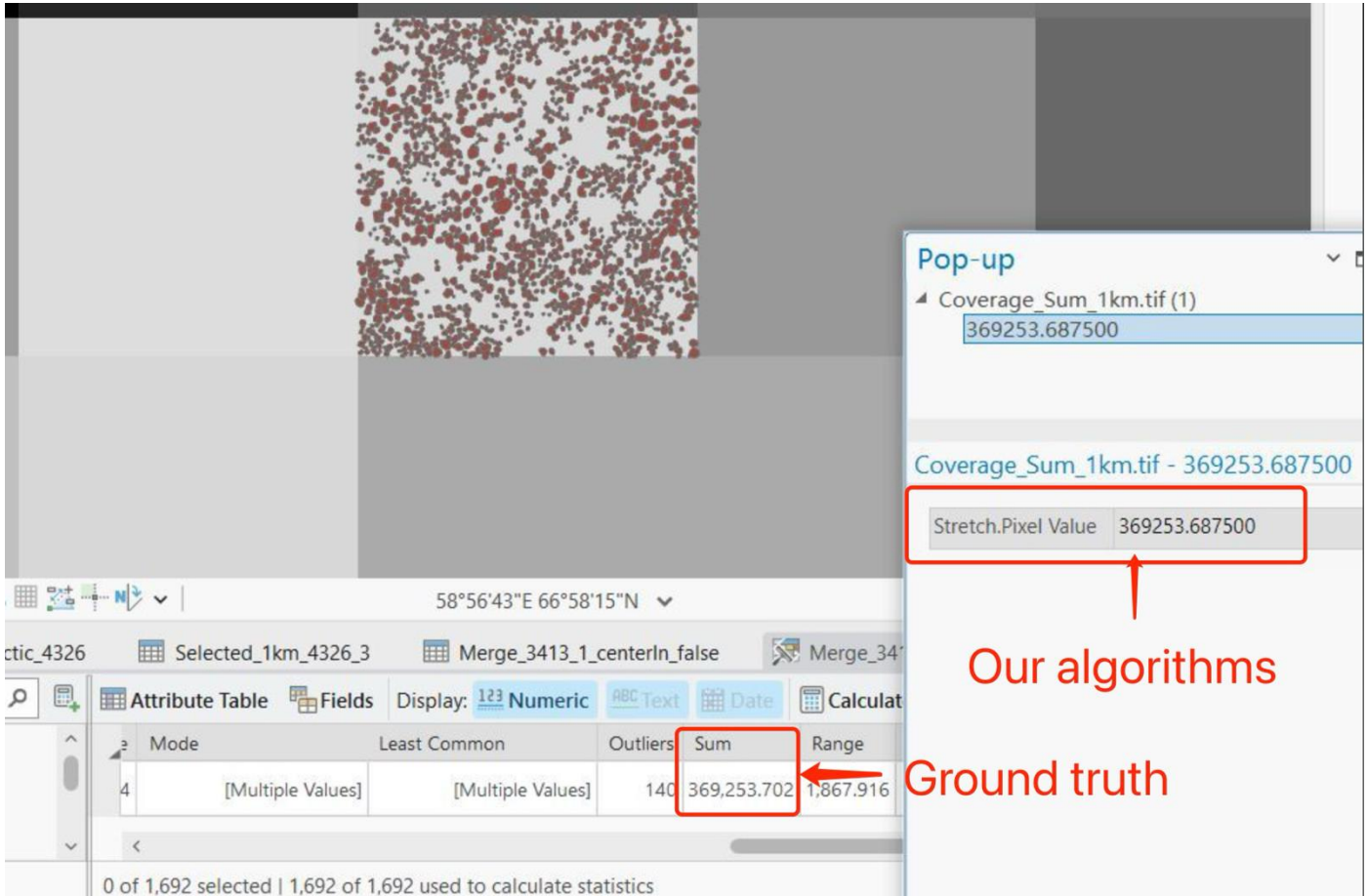
Meeting 12/10/2024

- Validation
  - Values within a pixel
  - Automatic and complete program

Validate

```
TERMINAL
(/home/xchen/code/alaska_shpMapping2) xchen@cici2labasuedu:~/code/shpMapping/code$ python getTilesinROI.py
                                        geometry
0  POLYGON ((58.91933 66.95962, 58.91933 66.97037...
[[15, 43493, 4192], [15, 43494, 4192], [15, 43495, 4192], [15, 43496, 4192], [15, 43497, 4192], [15, 43498, 4192], [15, 43493, 4193], [15, 43494, 4193], [15, 43495, 4193], [15, 43496, 4193], [15, 43497, 4193], [15, 43498, 4193], [15, 43493, 4194], [15, 43494, 4194], [15, 434
95, 4194], [15, 43496, 4194], [15, 43497, 4194], [15, 43498, 4194]]
The number of tiles within the geographic bbox is: 18
(/home/xchen/code/alaska_shpMapping2) xchen@cici2labasuedu:~/code/shpMapping/code$
```

1. Select one pixel
2. Get ground truth (coverage sum in that pixel)
3. Compare our values from the algorithms <-> Ground truth



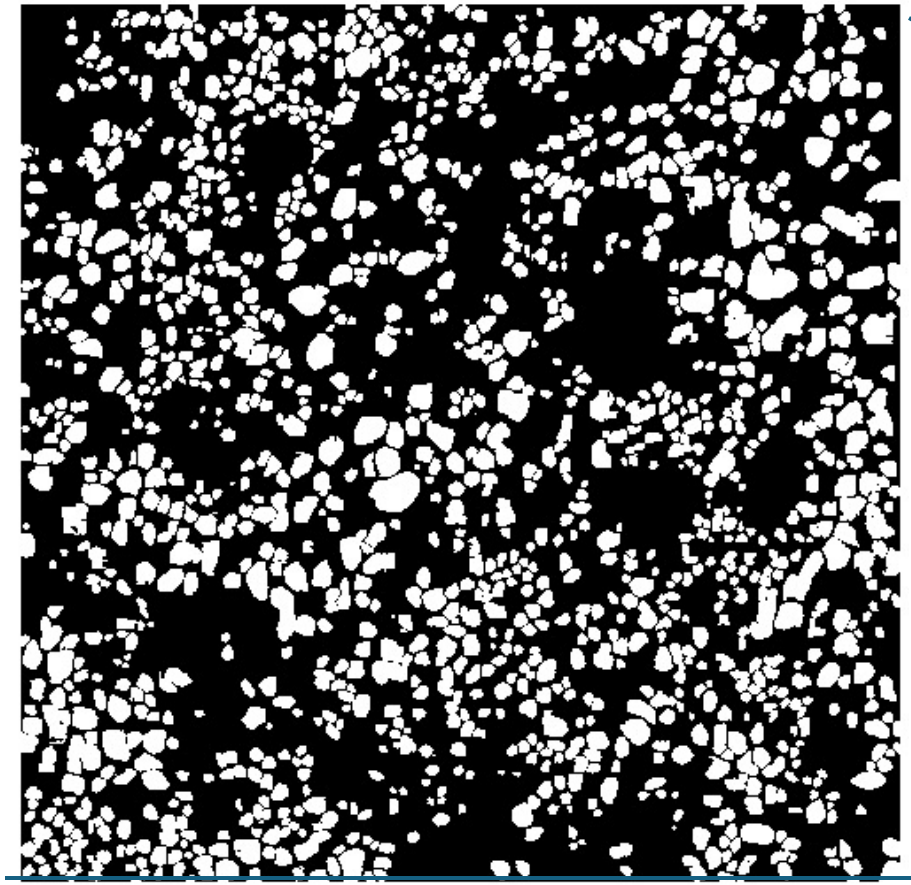
Trough length

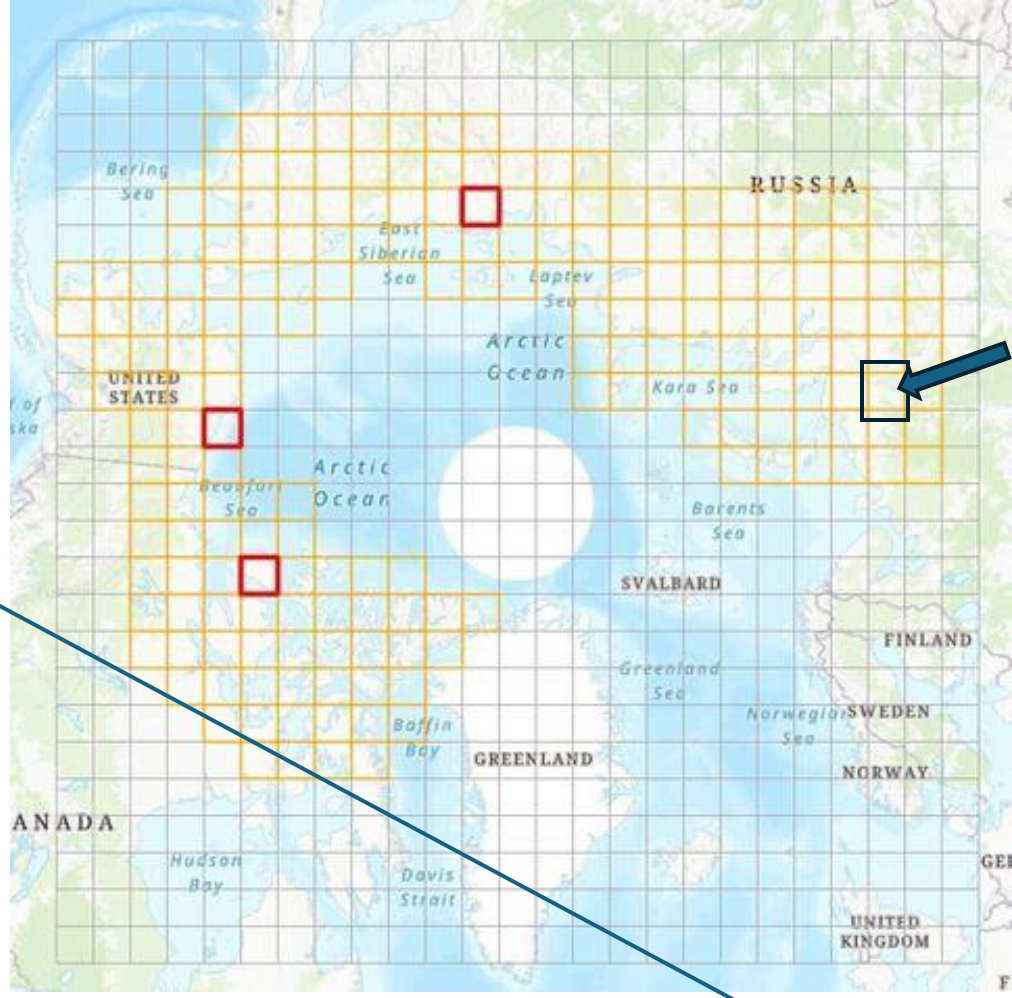
Test on 1 pixel within a grid

Width > IWP gap

Multiple Subset >5-20

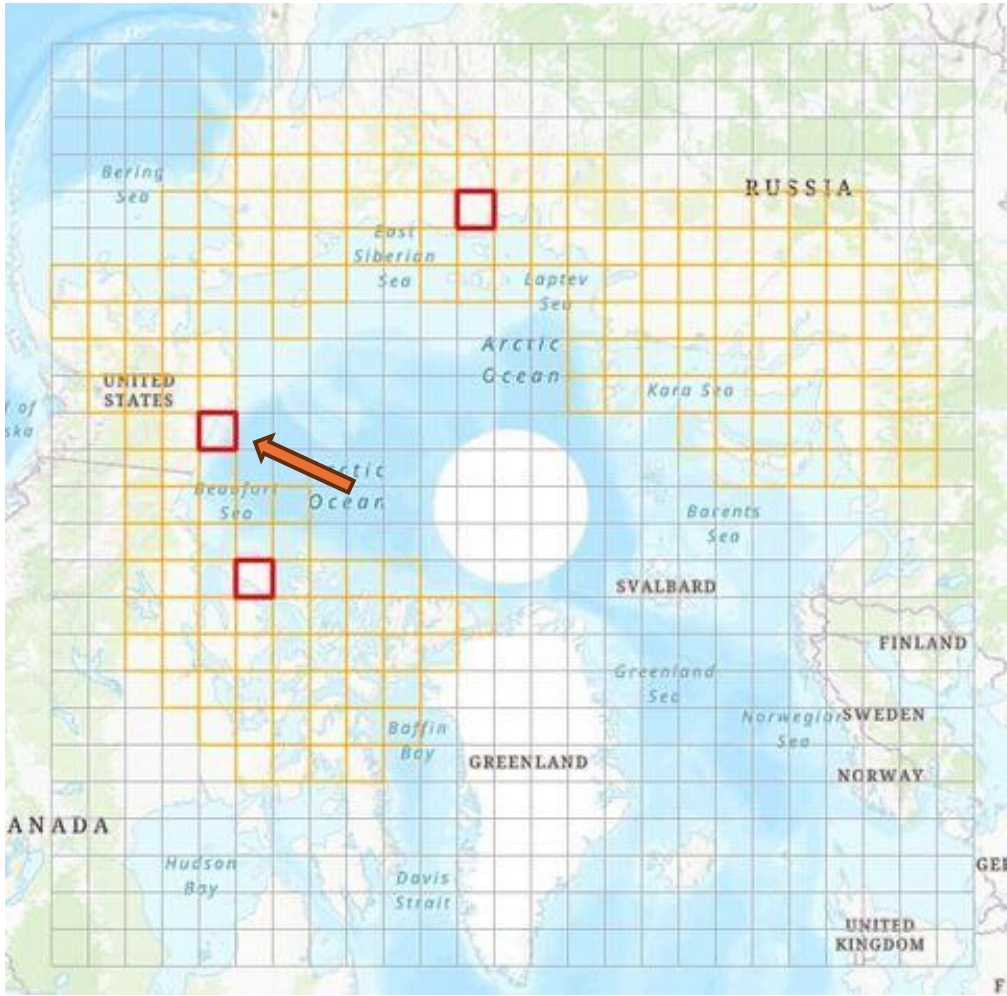
Centroid distance -> Intersect with polygons



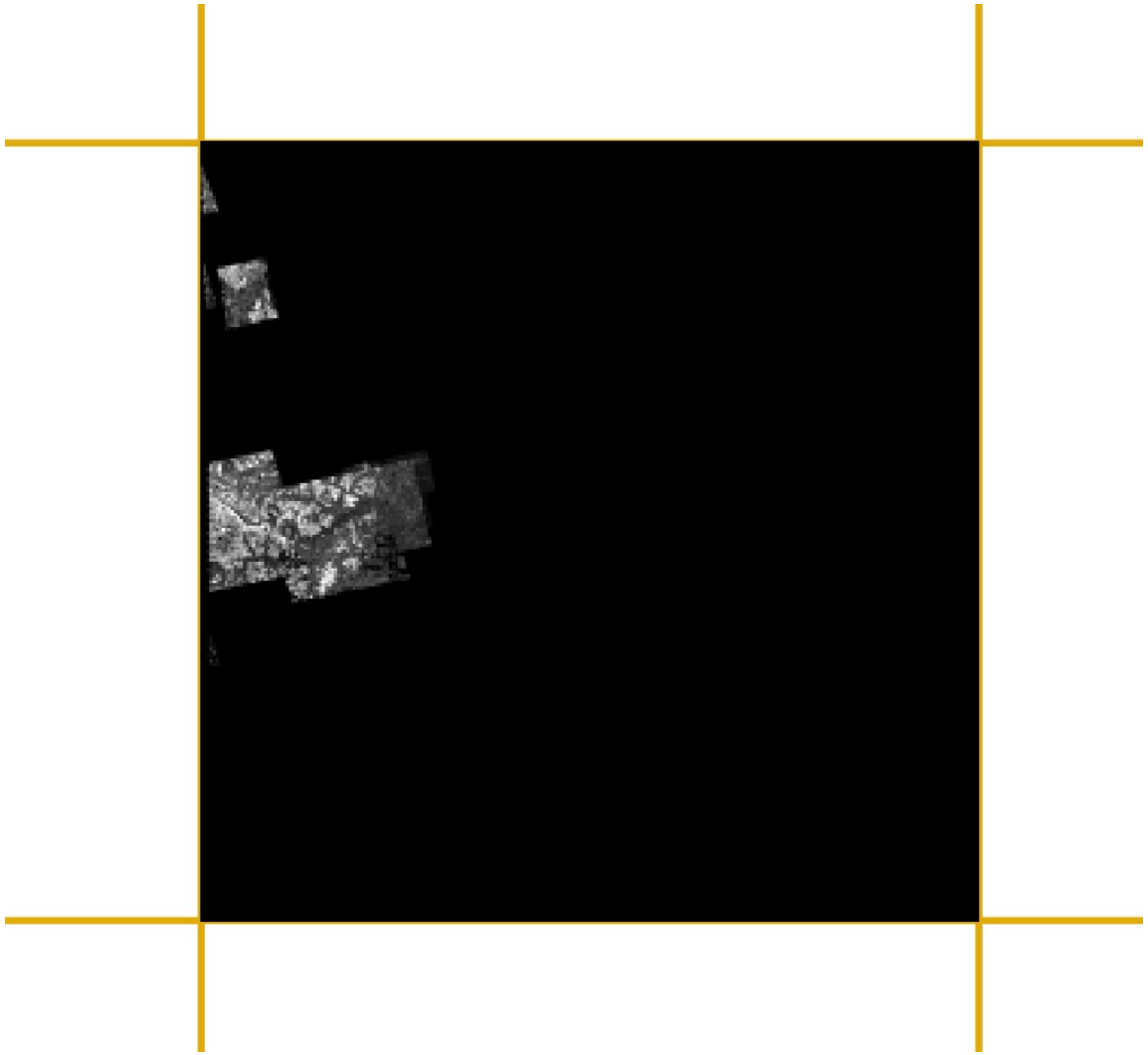


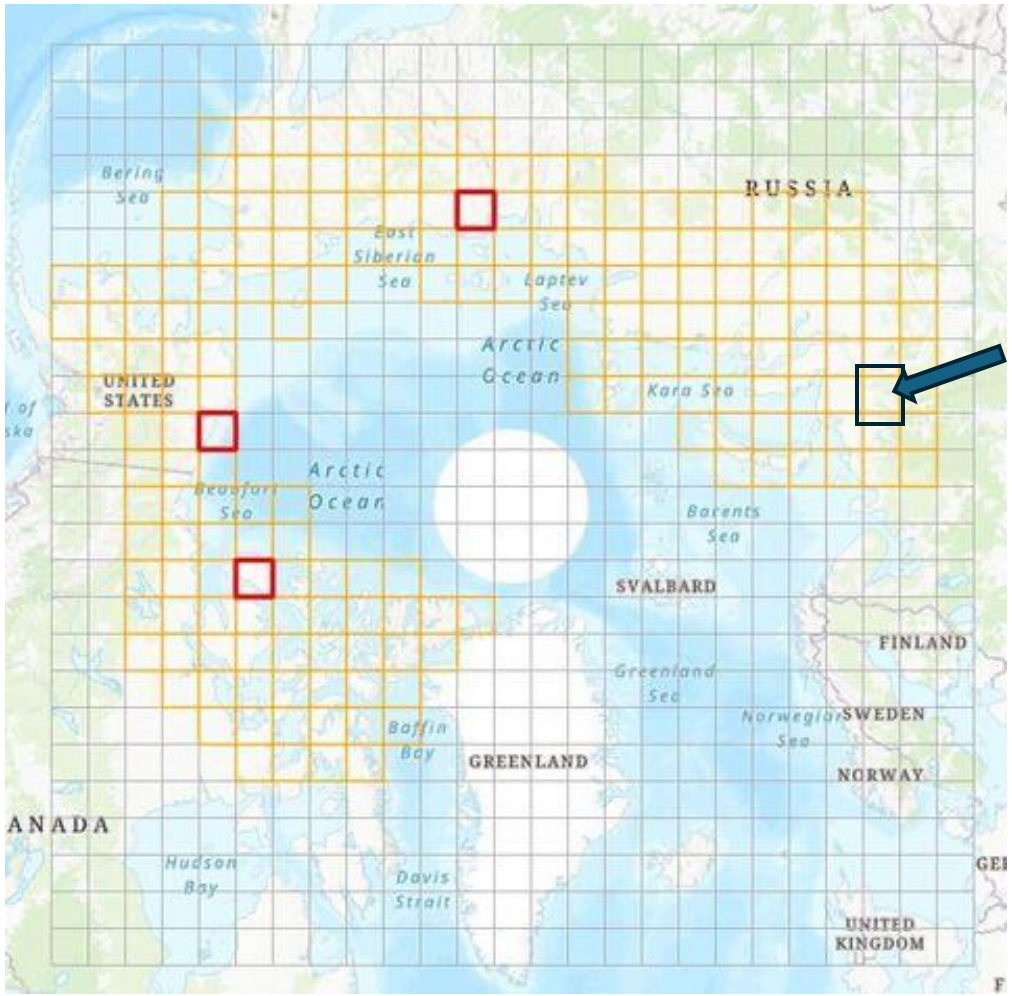
North Slope, Alaska

- 85,656,062 IWPs



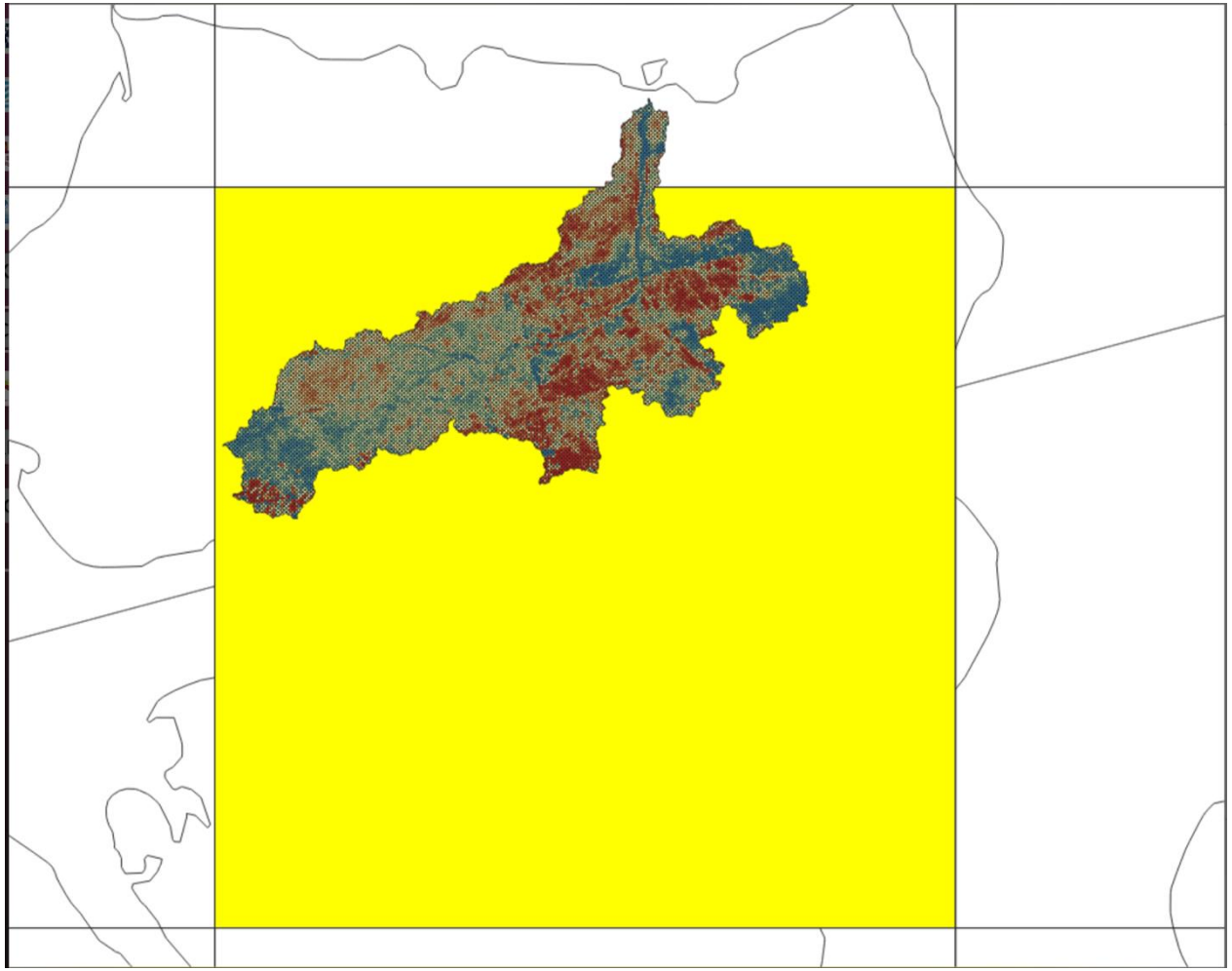
Processed grid





Meeting on 03/04/2025

- Cross validation on percentage coverage calculation with Elias' results
  - Yellow is one 256km by 256km at 1km resolution tile in our processing
  - The other is the watershed in Bank island that Elias processed.



Our results vs. Elias result on % coverage

- Computational efficiency improvement:
  - If we don’t consider data downloading time, each 256km^2 will take 20-30min on average.
  - So far, we are no longer using database
  - Processing the 240 tiles (a week)

Meeting 3/4/2025

- Results Evaluation
  - Accuracy evaluation: The degree of precision and correctness in predicting pipeline conditions or identifying anomalies.
  - Reliability evaluation: The consistency and stability of the results produced by the analytics techniques over multiple evaluations.
  - Scalability evaluation: The ability of the technique to handle large volumes of pipeline data and scale with increasing data sizes.
  - Computational efficiency: The speed and efficiency of the technique in processing and analyzing data within acceptable time frames.
  - Interpretability: The ease of understanding and interpreting the results generated by the analytics technique.

Accuracy evaluation

- IW network
  - Validate with external validation set (Elias work), which is generated outside of our algorithm-building process.

- Other statistics
  - Validate with Elias “coverage ratio” dataset

Reliability evaluation

- IW network:
  - Compared with hand-labelled IW network (randomly choose 1 pixels in Alaska, Canada, or Russia): relative error

- Other statistics
  - data downloading & querying: randomly choose 3× pixels in Alaska, Canada, and Russia. Compare results from our workflow and the manually downloaded gpkg files and calculated in ArcGIS.

Scalability & Computational Efficiency

- Compared parallel processing with non-parallel processing in data loading process in 1 grid
  - Speed up
  - # of concurrent processing
  - Focus on Alaska

Scientific validation

Their IWP dataset starts from 2001 to 2021, if our IWP dataset get trained on the images (2005-2008) within one of their 5 sites, then it should be fine. I'm not sure anything about their field sites now, so couldn't know if the datasets are consistent (edited)

Diameters of IWP (max) and total area of IWP. The smaller the IWP, the more ice wedges we have on ground.

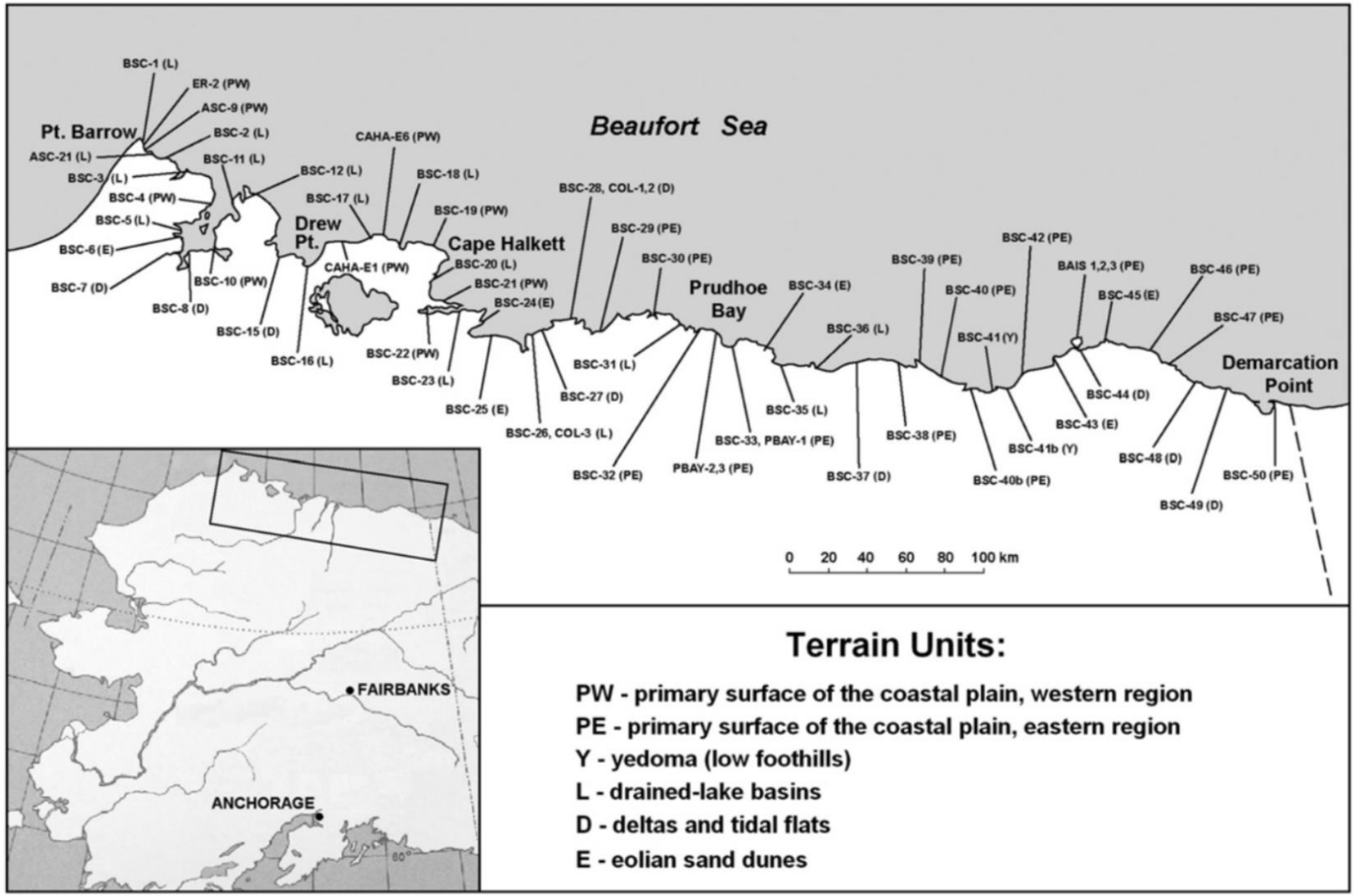


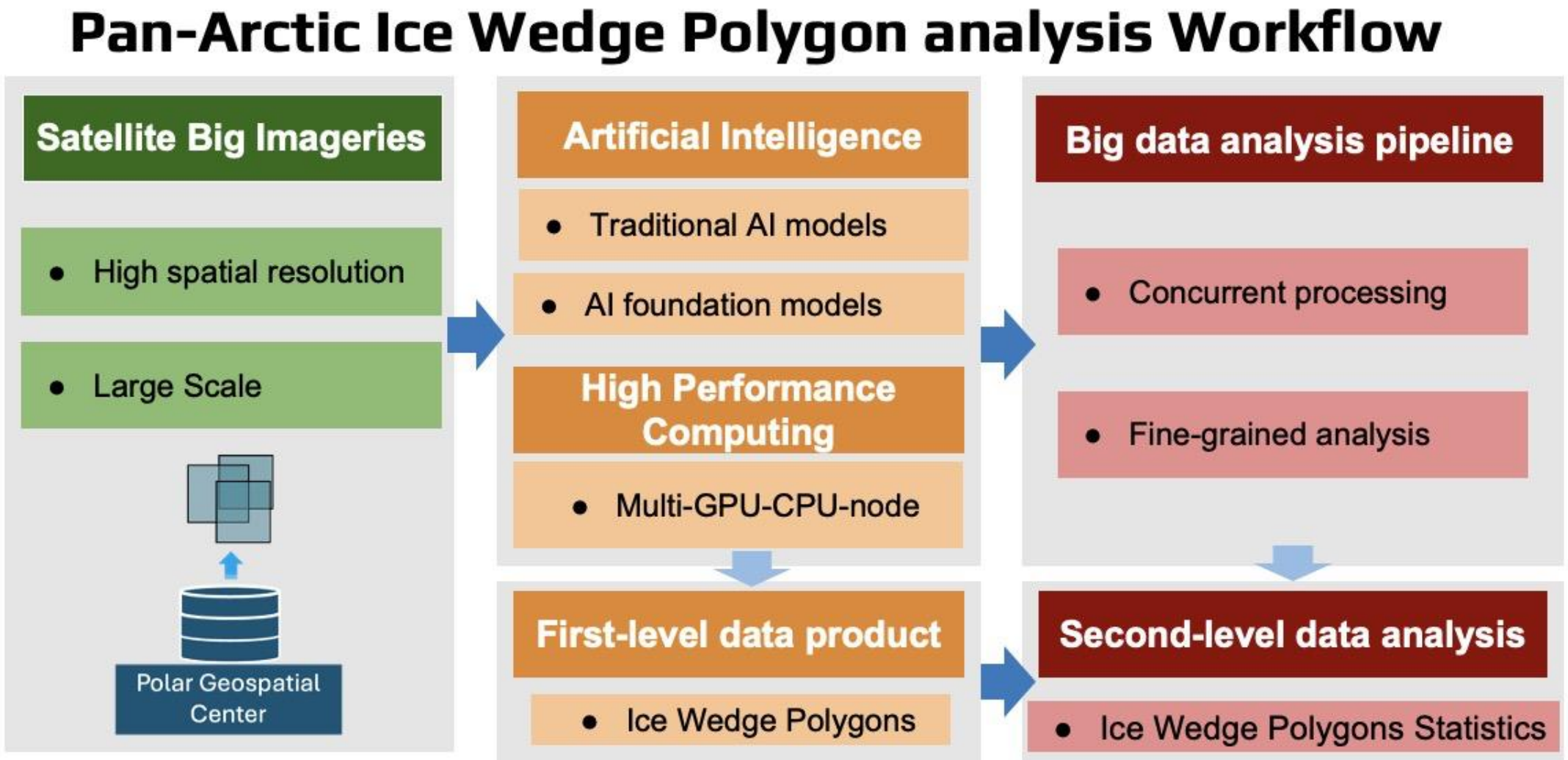
Fig. 1. Location of field sites studied in 2005–2008 along the Beaufort Sea coast from Point Barrow to the Canadian border.

https://www.sciencedirect.com/science/article/abs/pii/S0165232X12001644

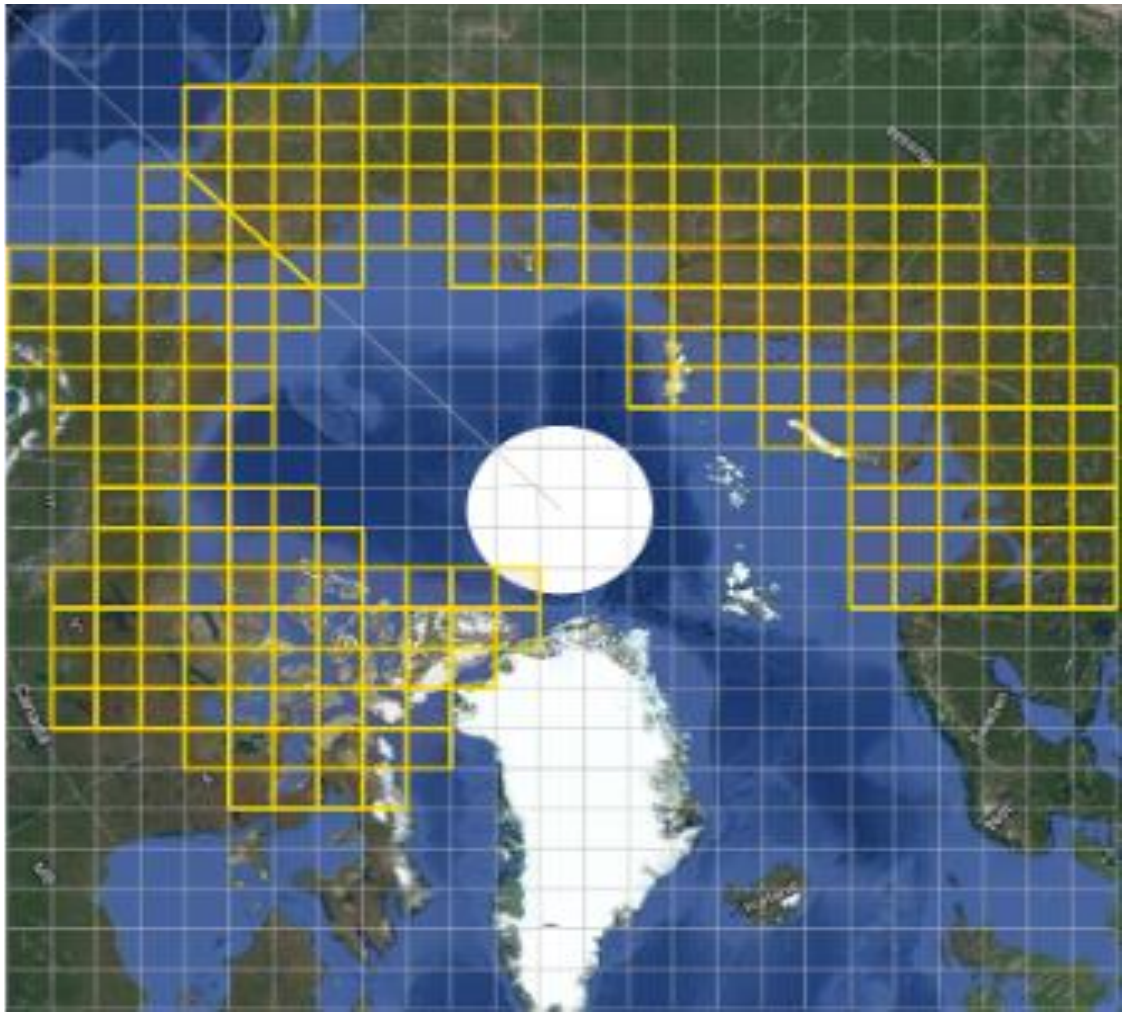
3/28/2025

- Mapping in Alaska

Workflow


Pan-Arctic Ice Wedge Polygon analysis Workflow
Satellite Big Imageries
High spatial resolution
Large Scale
Polar Geospatial Center
Artificial Intelligence
Traditional AI models
AI foundation models
High Performance Computing
Multi-GPU-CPU-node
First-level data product
Ice Wedge Polygons
Big data analysis pipeline
Concurrent processing
Fine-grained analysis
Second-level data analysis
Ice Wedge Polygons Statistics

Grids



Statistics mapping in Alaska