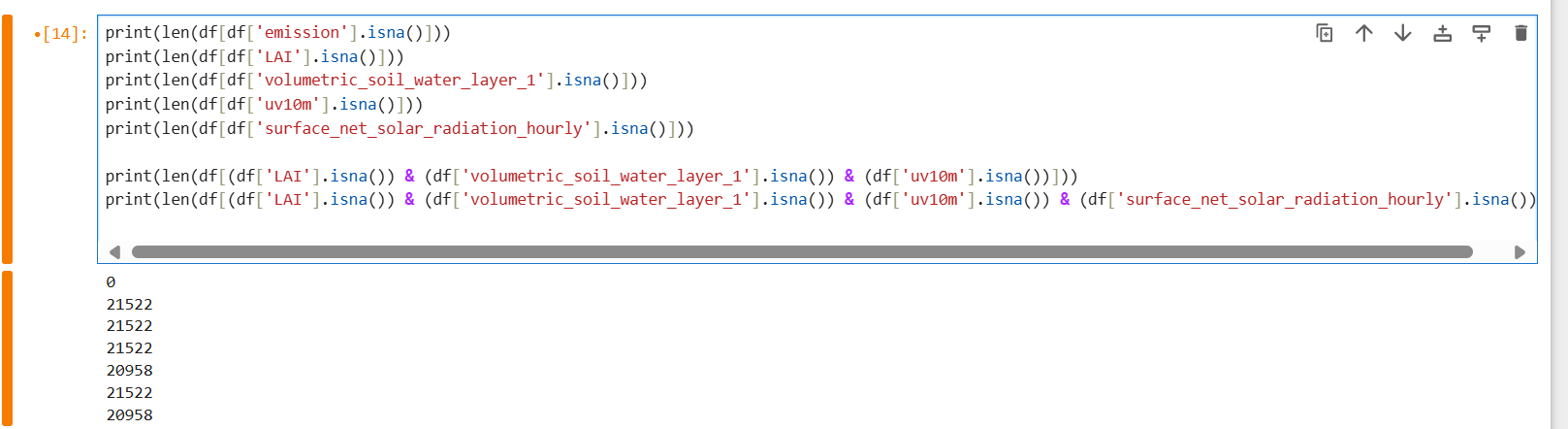
关于NE & ERA5数据集空值和0值的问题

Chenya

latest update: 20241024

[ERA5-Land: data documentation - Copernicus Knowledge Base - ECMWF Confluence Wiki](https://confluence.ecmwf.int/display/CKB/ERA5-Land:+data+documentation)

1. **空值**

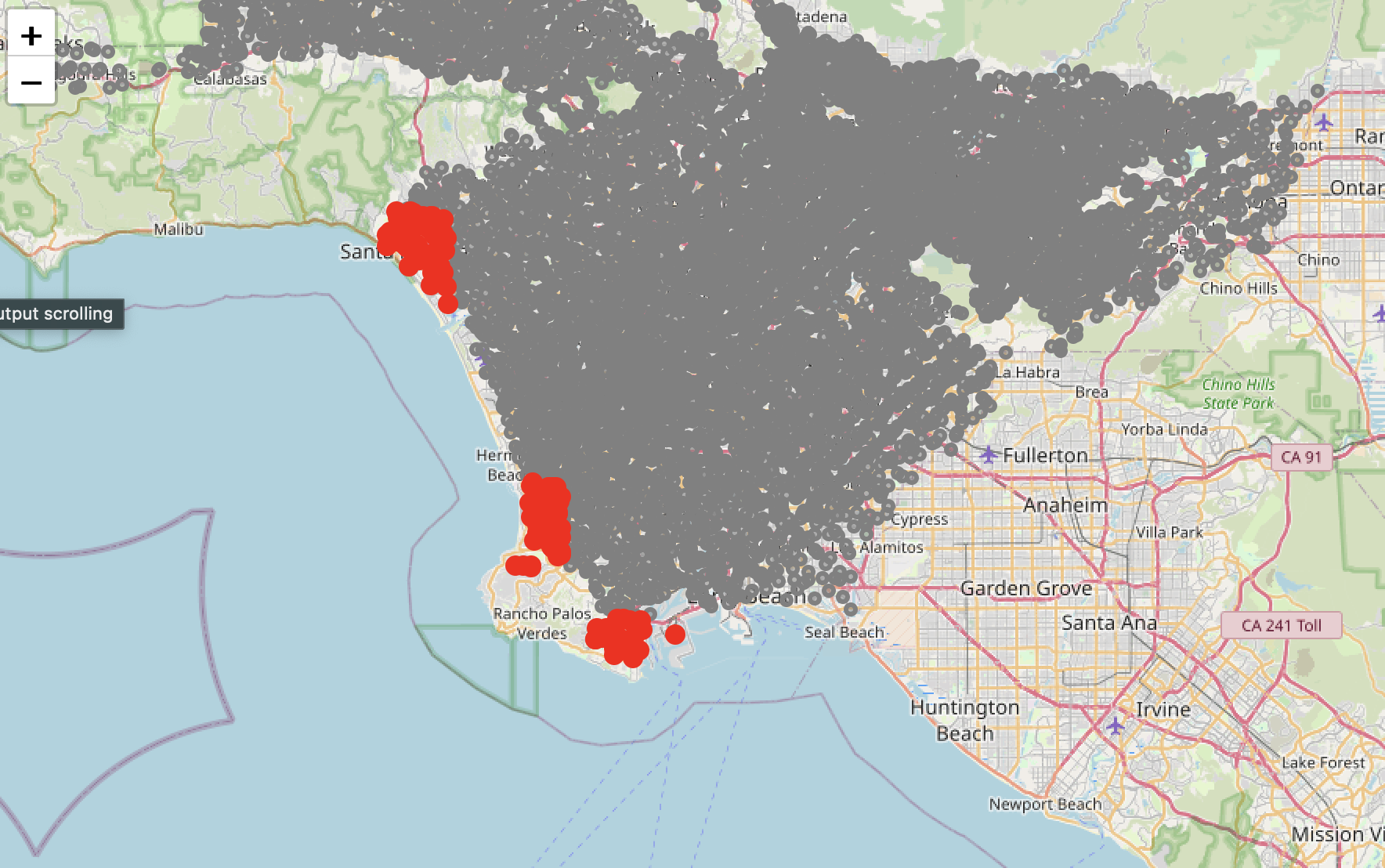


这里首先需要强调的是emission来自GRACED数据集，原理是根据其他AE数据集和部门排放数据等建模再gridded；其他气象数据来自era5 hourly land data reanalysis，来自于已有的卫星数据、实地测定数据建模再对齐分辨率。所以再分析几个feature的空值和0值时应该回归这个数据集的测定、计算方法和规定。

* emission：这里没有空值，跳过
* LAI/soil water/uv10m这里有21522条数据为空，即几个数据都是空值，其中20958条hourly radiation为空。已重新join过，数据无误。

以下documentation也讲了具体LAI的计算过程等，modis数据集-model-对齐时间分辨率：[IFS Documentation CY45R1 - Part IV : Physical processes | ECMWF](https://www.ecmwf.int/en/elibrary/80895-ifs-documentation-cy45r1-part-iv-physical-processes)

通过观察可知，era5的空值主要分布在沿海地区，但是这并不意味着沿海的数据都是nan，下图以LA为例，红色代表空值，灰色为其他正常数据：



era5的空间分辨率为0.1°，如以上格子示意（比例可能不对），我们的数据集关注land数据即陆地的性质，所以一个格子内涉及海、陆两种下垫面时可能会对卫星测量、陆地的模型精准度都造成影响，特别是海洋面积远大于陆地面积时，这些数据被assign为nan。所以，我认为这些era5为空值的数据应该删去，共2w条左右（140w数据总数），是我们运用era5数据集扩大了NE数据分辨率的牺牲，避免这些数据对我们model的影响。

1. **0值**

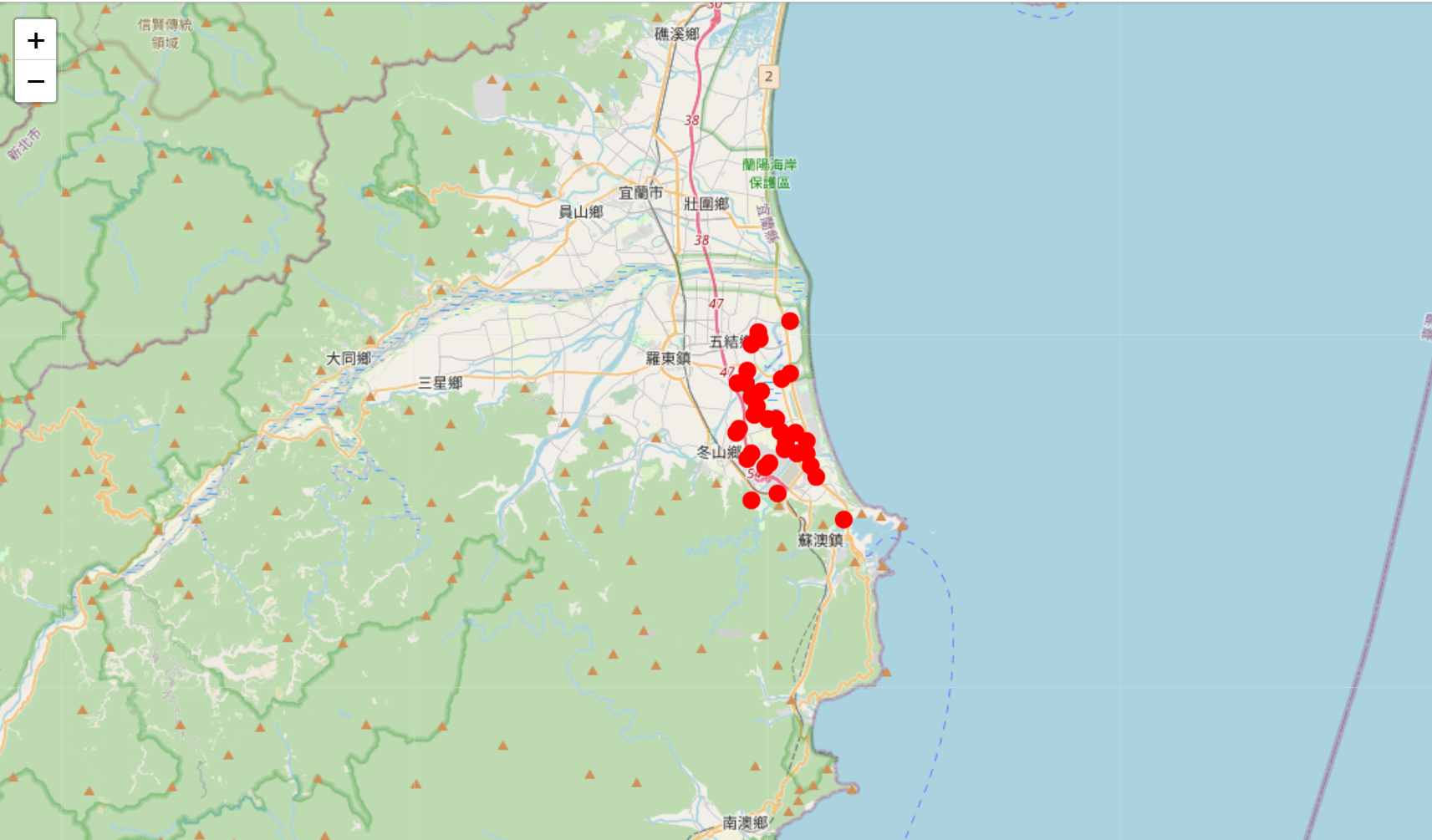


0值目前出现在ae、lai、radiation三个feature：

* AE：原本的代码错误已解决，目前共2k条数据ae为0。随机出现，无规则形状，可能零星出现在城市、河边、沙漠内builtup等，大部分城市只是个别数据为0，部分城市0值点较多（以台湾省屏东县为例，三个日期22-08-31等出现大量0值，COVID??），可能由原本的数据缺失（某段时间、或某个区域数据上报的问题）造成。由此，我认为应该删去这2k条数据。
* 沿海

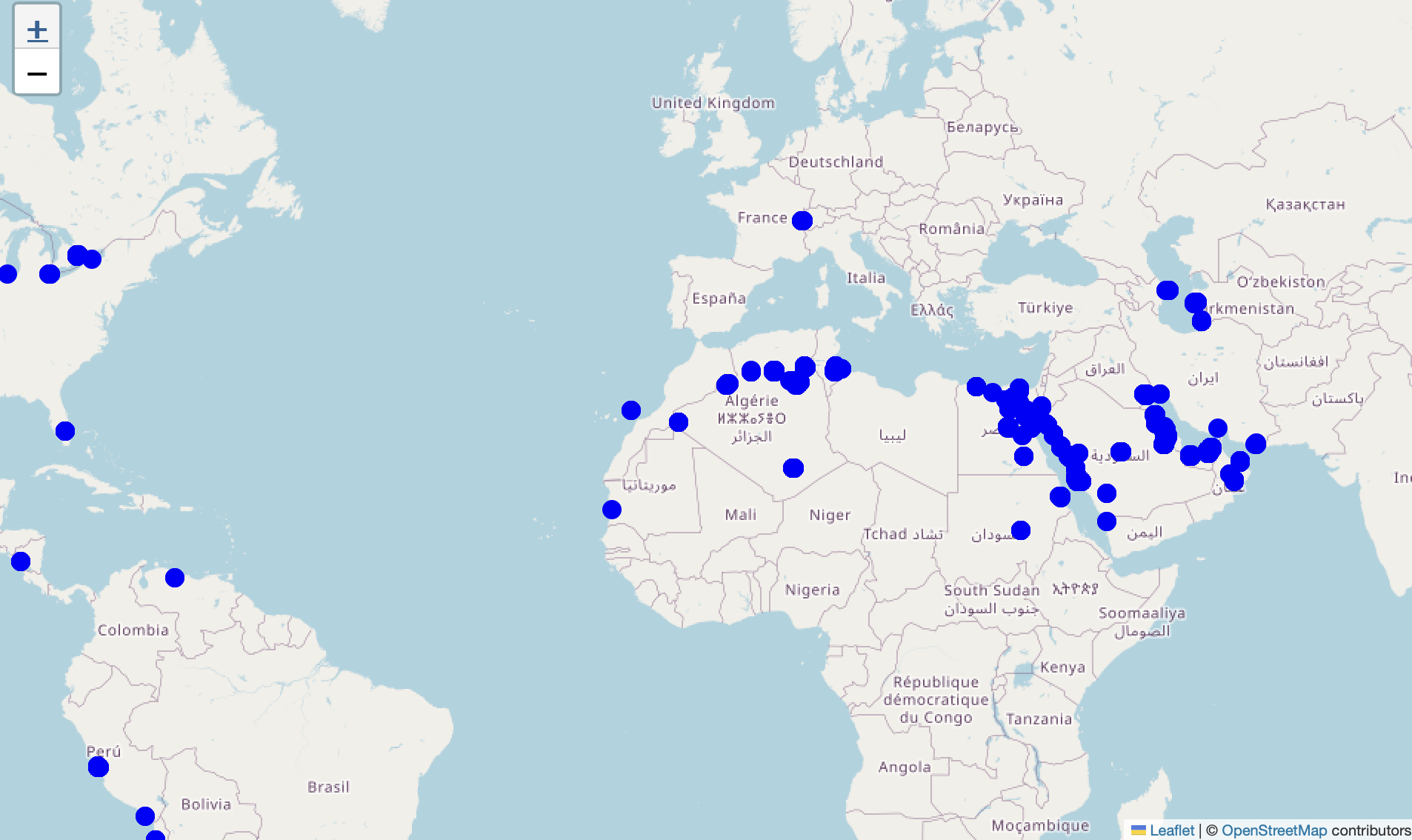
（已联系graced数据集团体请教这个问题）

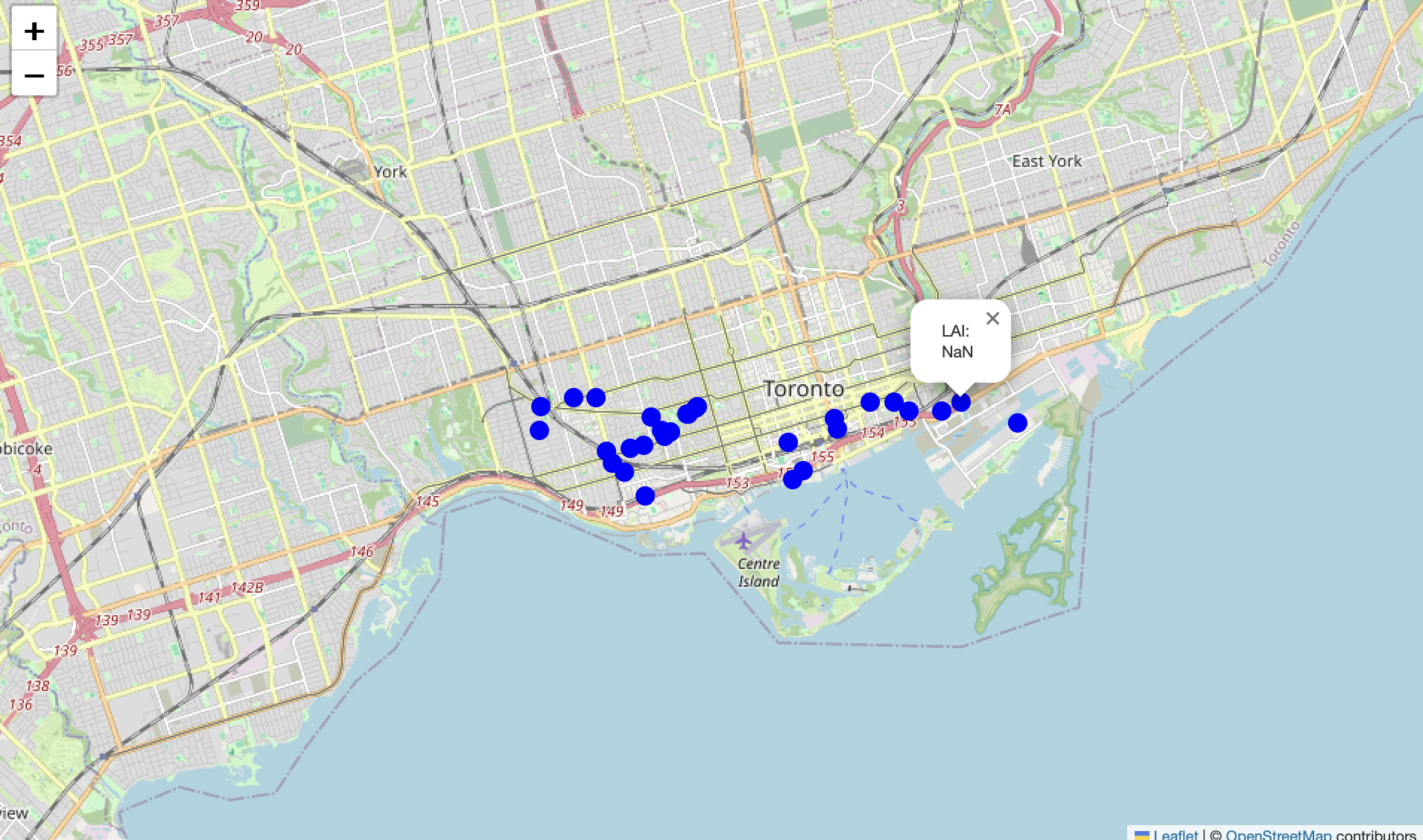




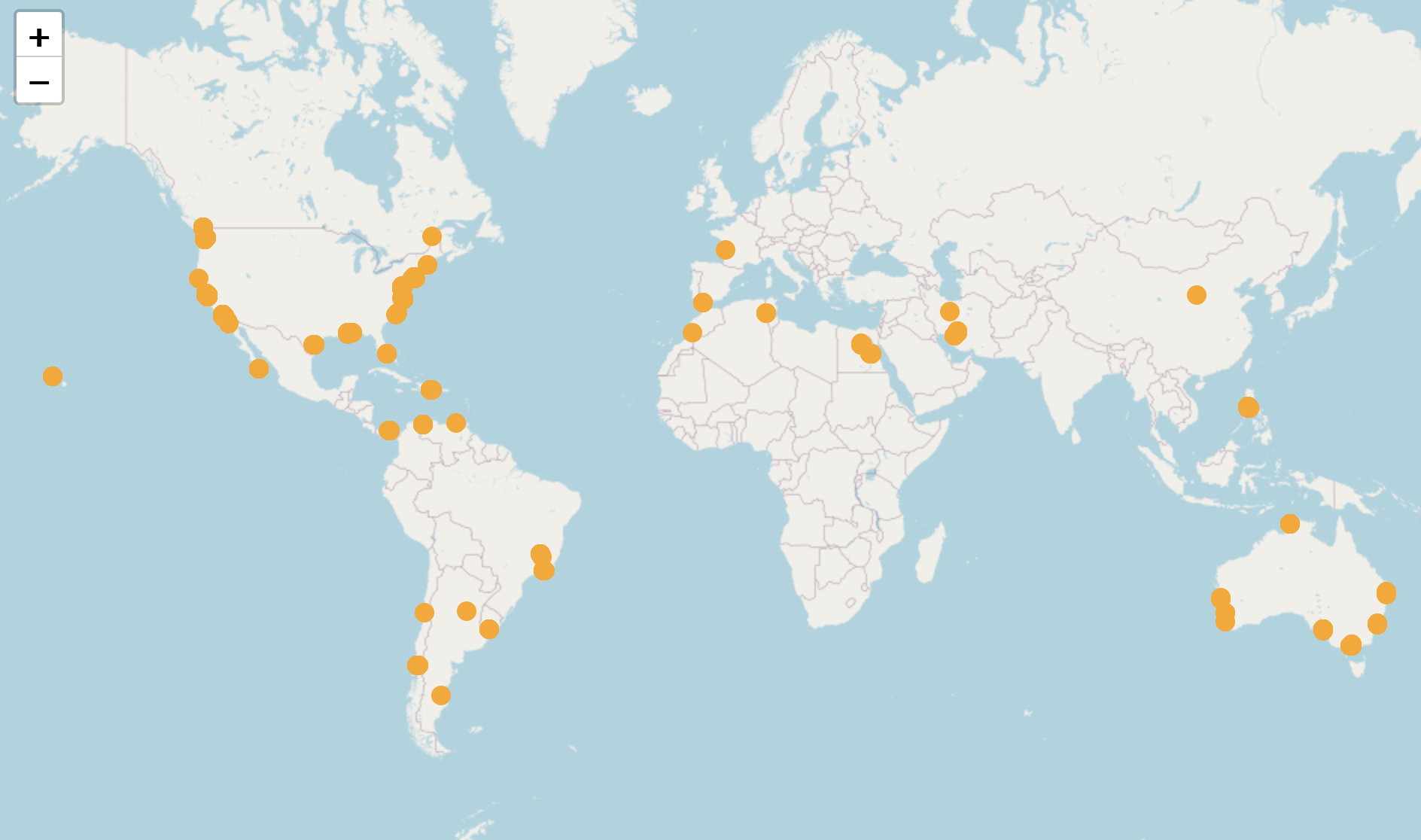
* LAI：0值出现在北非沙漠、其他城市地区等，如下图所示，在官方documentation中也介绍了lai为0值时指没有植被/裸露土地等，雨水可以没有阻挡的到达地面。所以我认为应该保留这几条数据。 <https://codes.ecmwf.int/grib/param-db/67>

*「This parameter has a value of 0 over bare ground or where there are no leaves. 」*





* Radiation：600条数据为0，其中564条数据的lai为空，和lai空值分布一致。官方文档中没有对它的特殊值的解释，但其计算的原理是短波辐射减去地面反射，所以除了反射率极高、极夜现象（不可能，我们研究的是中纬度以下地区），这个值不可能为绝对的0。剩余36条主要分布在北非沙漠，可以保留或不保留。



*「This parameter is the amount of solar radiation (also known as shortwave radiation) that reaches a horizontal plane at the surface of the Earth (both direct and diffuse) minus the amount reflected by the Earth's surface (which is governed by the albedo). 」*

1. **总结！**

* 根据以上分析，空值和零值应当分别处理
* 确认feature后可以重新spatial join一下，减少不同轮次的代码、计算、版本差异，避免这部分误差 已完成
* era5数据以秒为单位先合并，era5空值可以直接填充，除了lai外其他数据的0值不参与计算。

1. **以秒为单位合并**

era5的空值不参与计算，0值参与计算

结果：

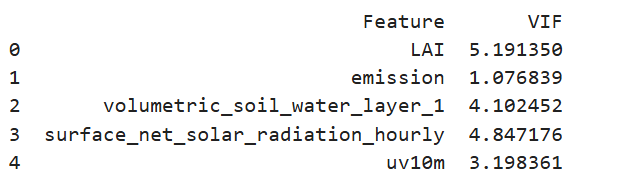
aggregate出来共237621 rows （空值和0值怎么处理？？）

空值约3k rows，即这个group都是nan，主要分布沿海地区

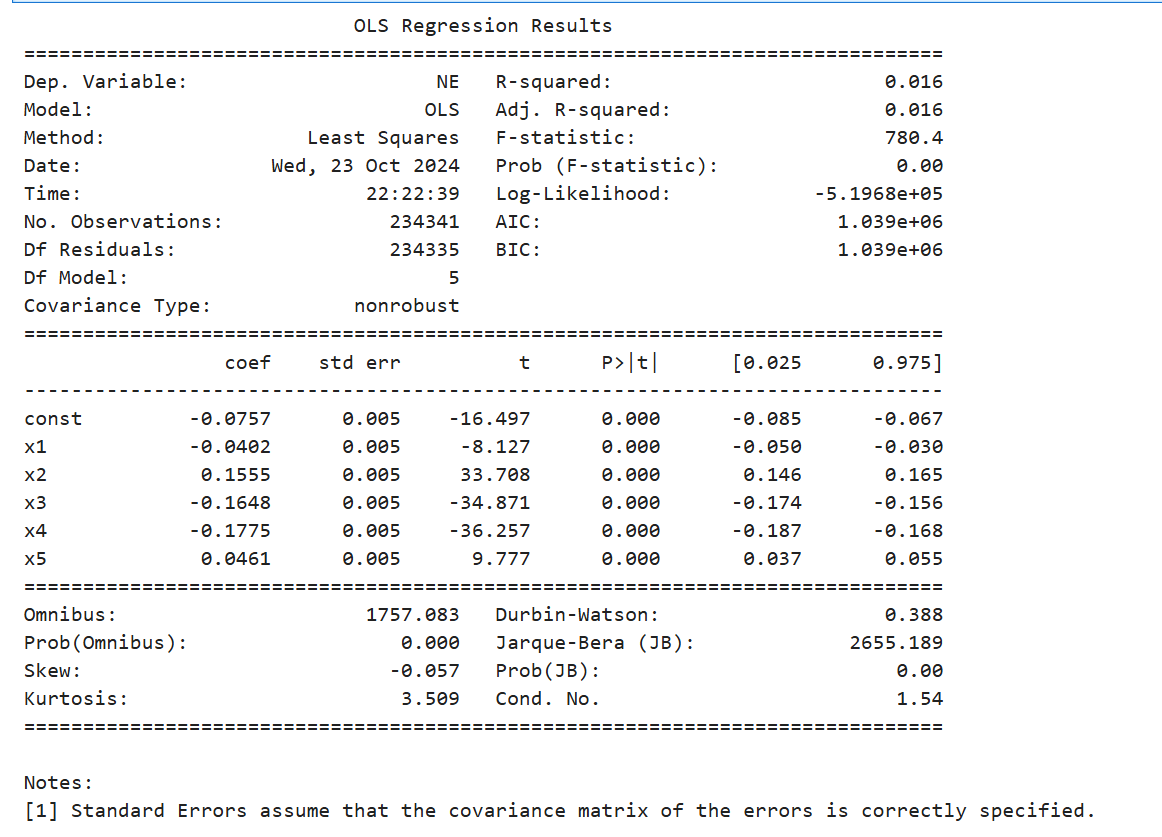
****

****

去掉空值后，还剩234341 rows，通过了共线性VIF检验 （0.1度 23.04w）



OLS R2 = 0.016 (0.001⬆)

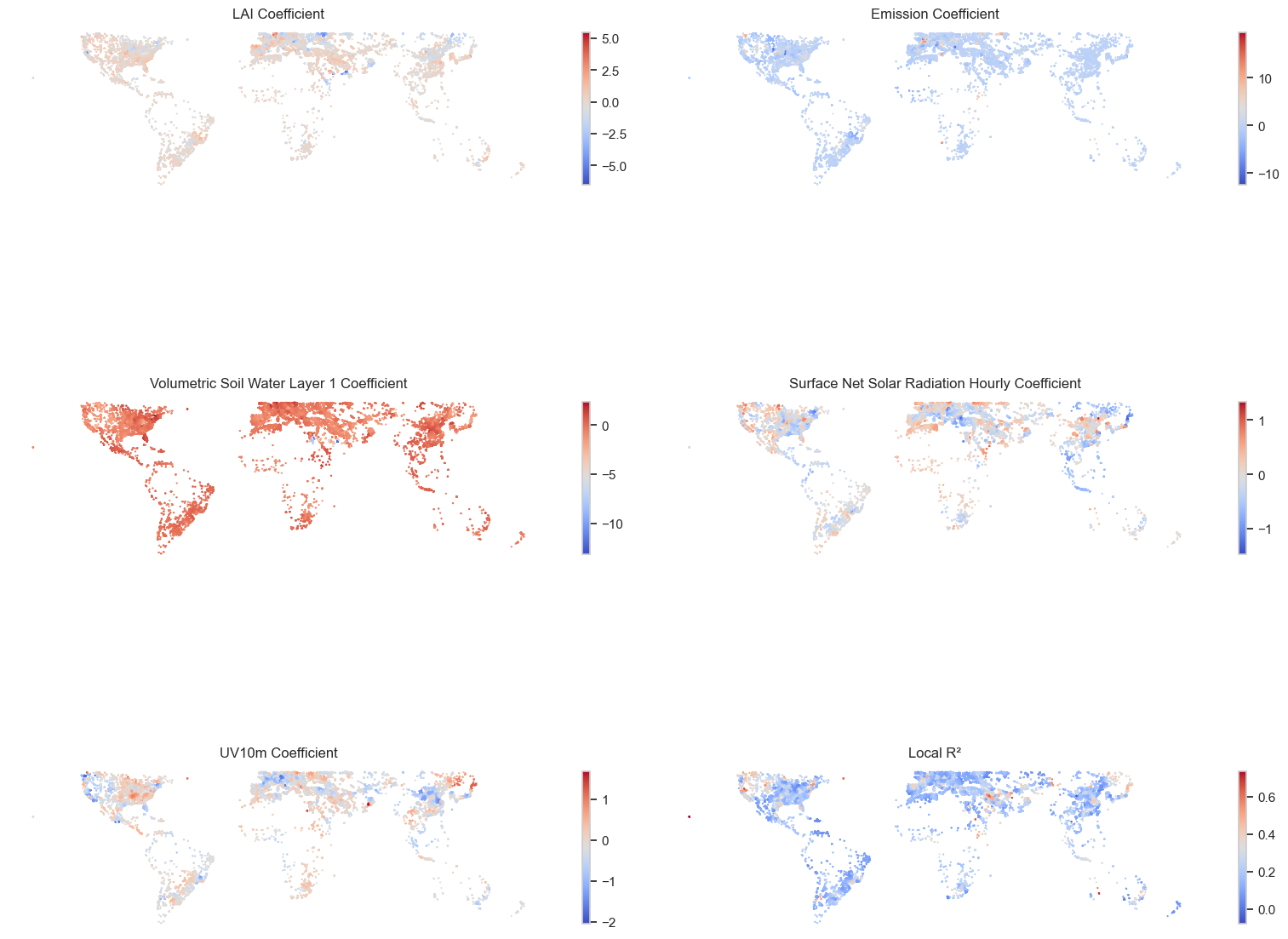
****

GWR R2 = ~0.4

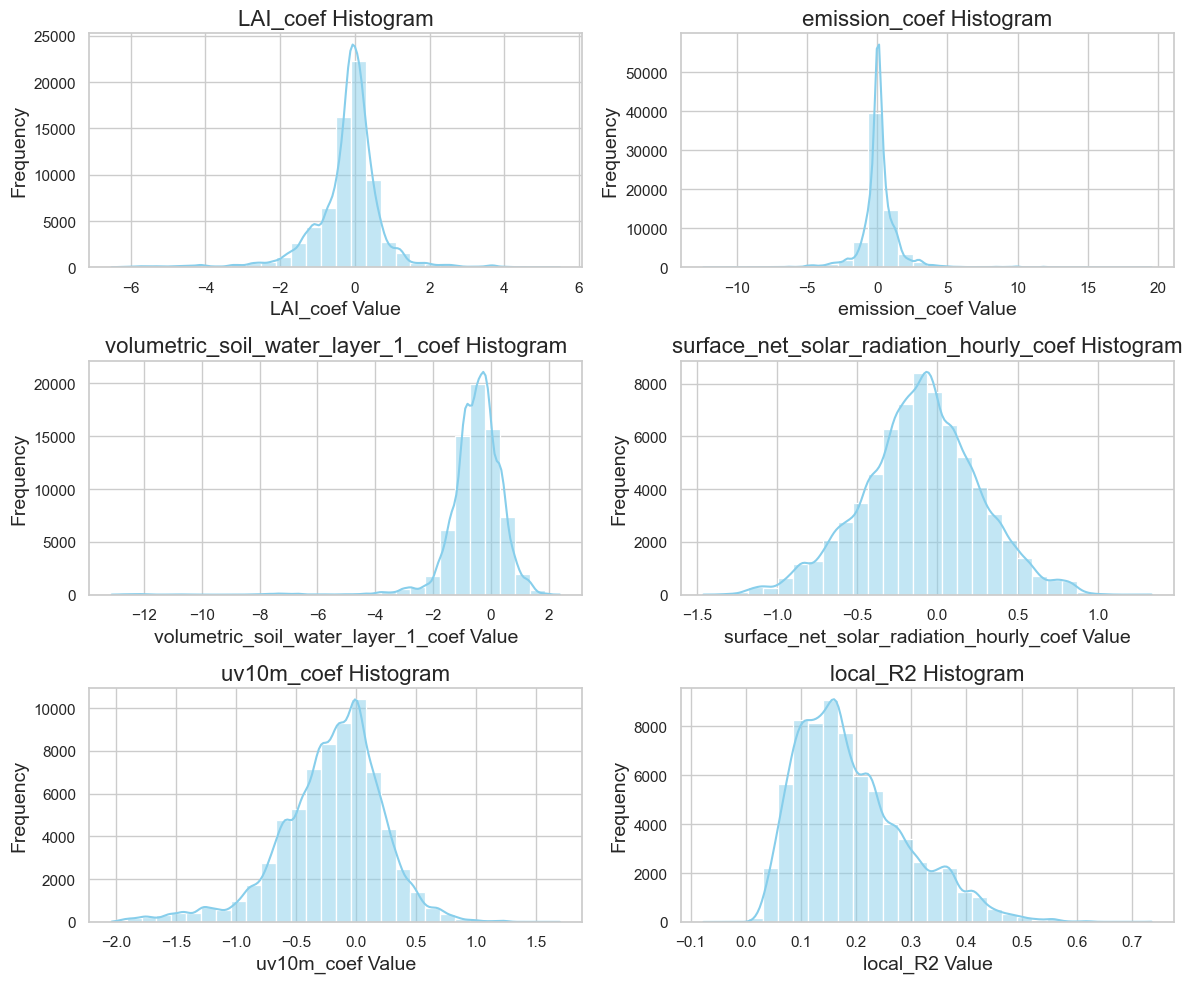
bw=289

global-continental level

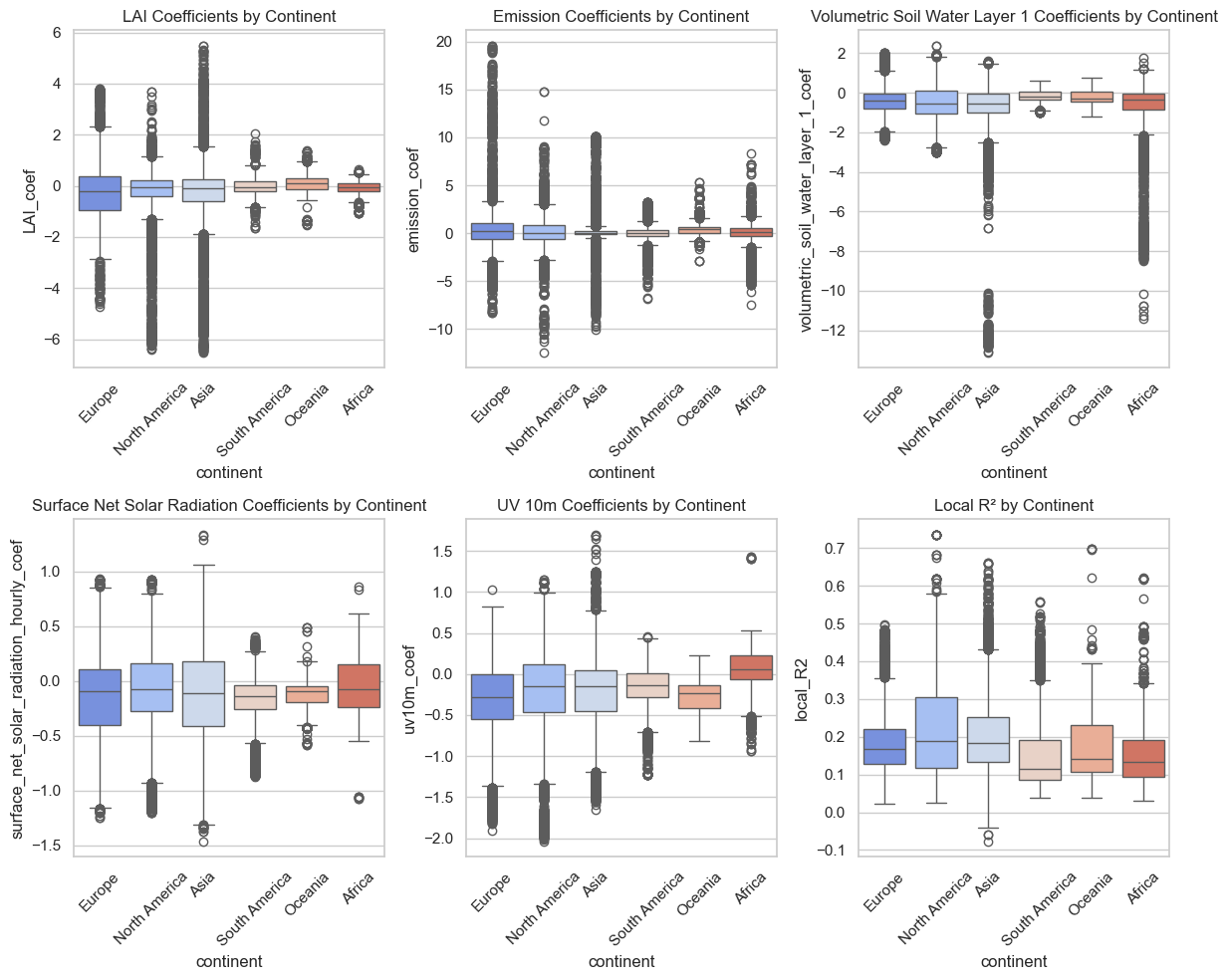
global coefs and local r2 map:



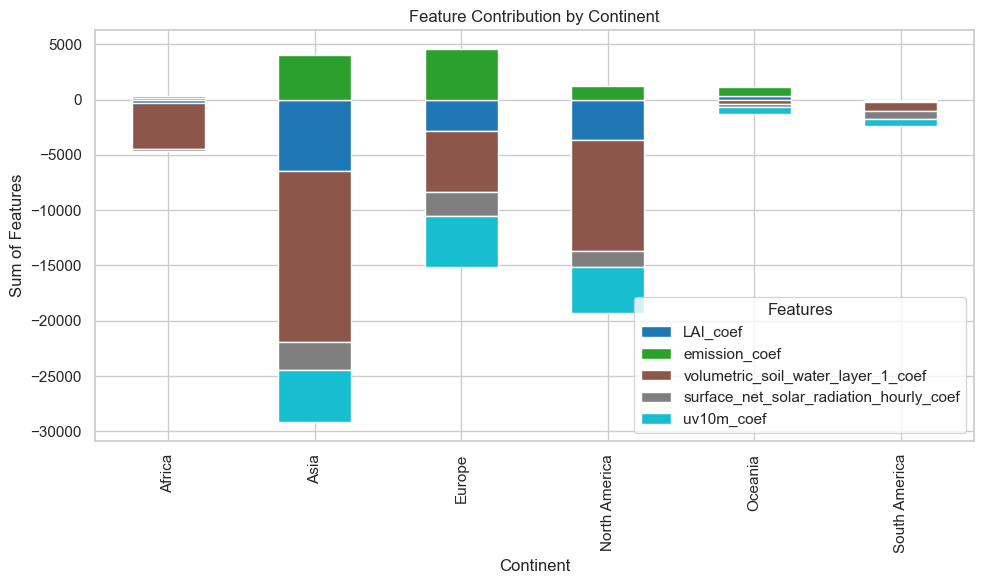
coefs + local r2 histogram:



boxplot of coefs across continents:



feature contribution (abs) percentage:



(not very accurate⬆)

GTWR: running