London School of Hygiene & Tropical Medicine

Improving Health Worldwide

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Satellite-based machine learning models to estimate high-resolution environmental exposures across the UK

Rochelle Schneider
Research Fellow in Geospatial Data Science
PhD Geospatial Analytics | MSc GIS | MRes Remote Sensing

Visiting Researcher at ECMWF
Centre on Climate Change and Planetary Health, LSHTM

Antonio Gasparrini
Professor of Biostatistics and Epidemiology
PhD Medical Statistics | Postgraduate Medical Statistics and Biometry | MSc Biostatistics

Centre for Statistical Methodology, LSHTM
Centre on Climate Change and Planetary Health, LSHTM

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

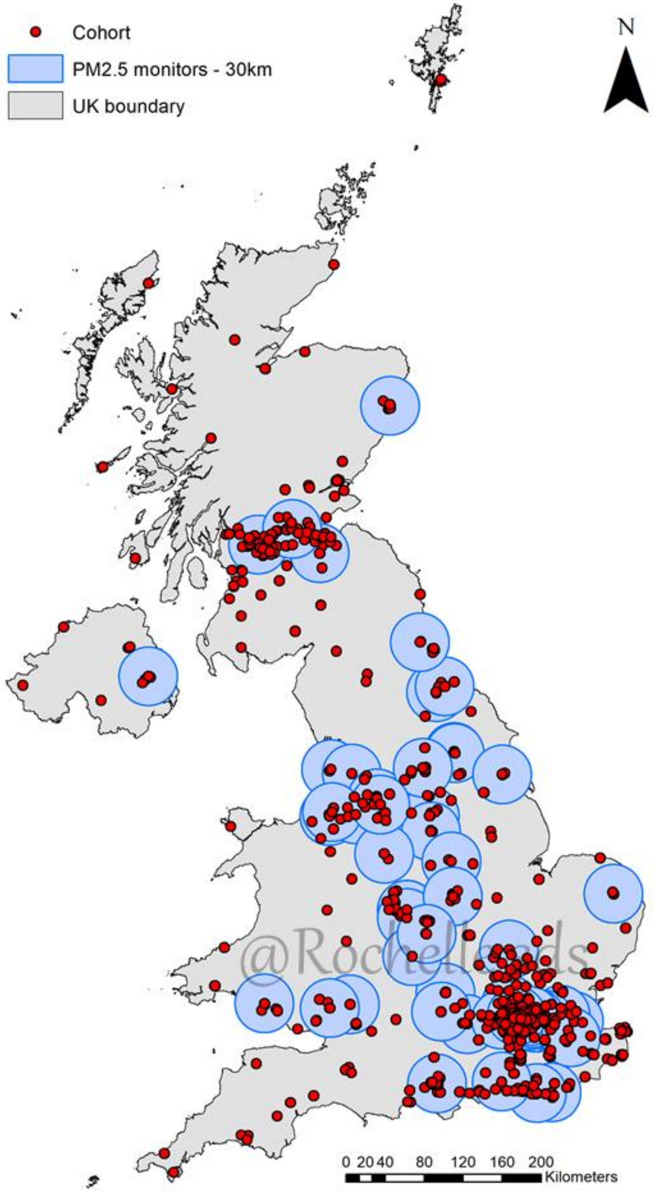
Current issues in environmental epidemiology

Specific aspects:

- Widespread exposure to environmental factors - *e.g.* air pollution - often affecting the whole population;
- Often small risks - *e.g.*, a RR of 1.0051 (95%CI: 1.0007-1.0093) for an increase of $10\mu gr/m^3$ of PM_{10} (Samet NEJM 2000);
- Need to perform epidemiological analyses on large populations.

Traditional Limitations

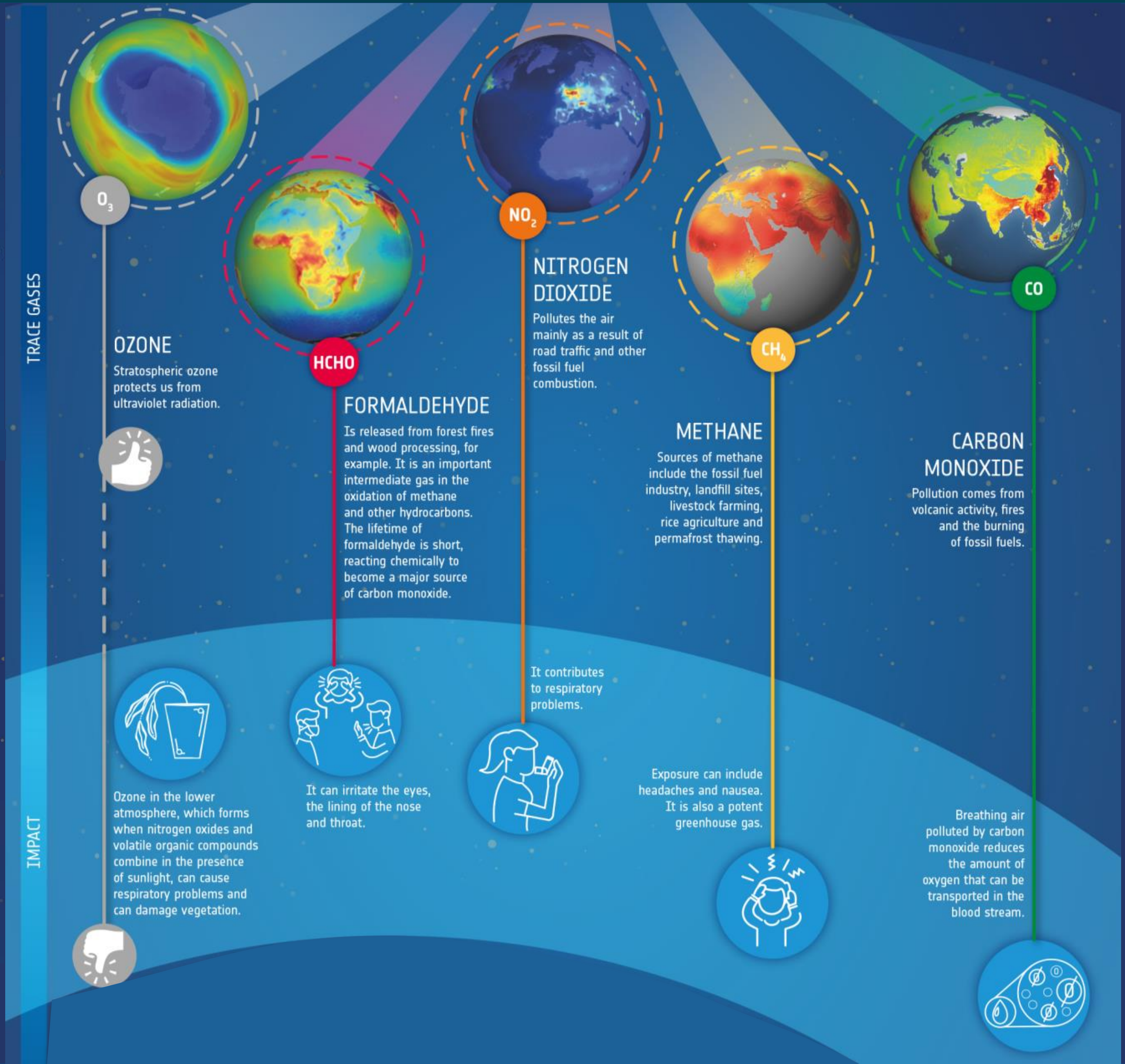
- Outcomes/exposures with low temporal and/or spatial resolution, roughly aggregated, and lack of individual / small-area information;
- Partial coverage, especially in rural areas and in low/middle income countries, posing limitations in country-wide or global analyses.
- Linkage between various temporal and spatial scales



Novel data sources for environmental epidemiological studies

→ THE AIR WE BREATHE

Air pollution is a major environmental health problem that affects millions of people around the world. Satellite data and computer models can show how pollution accumulates and how it is carried in the air. Mapping the global atmosphere every day, Sentinel-5P provides high-resolution data on a multitude of trace gases and information on aerosols that affect air quality and the climate. Offering advances in coverage and resolution, Sentinel-5P is set to take air-quality monitoring to a new level.



Wildfires

Pollen

aerosol cci PM / Dust

land surface temperature cci

Project ST- UK

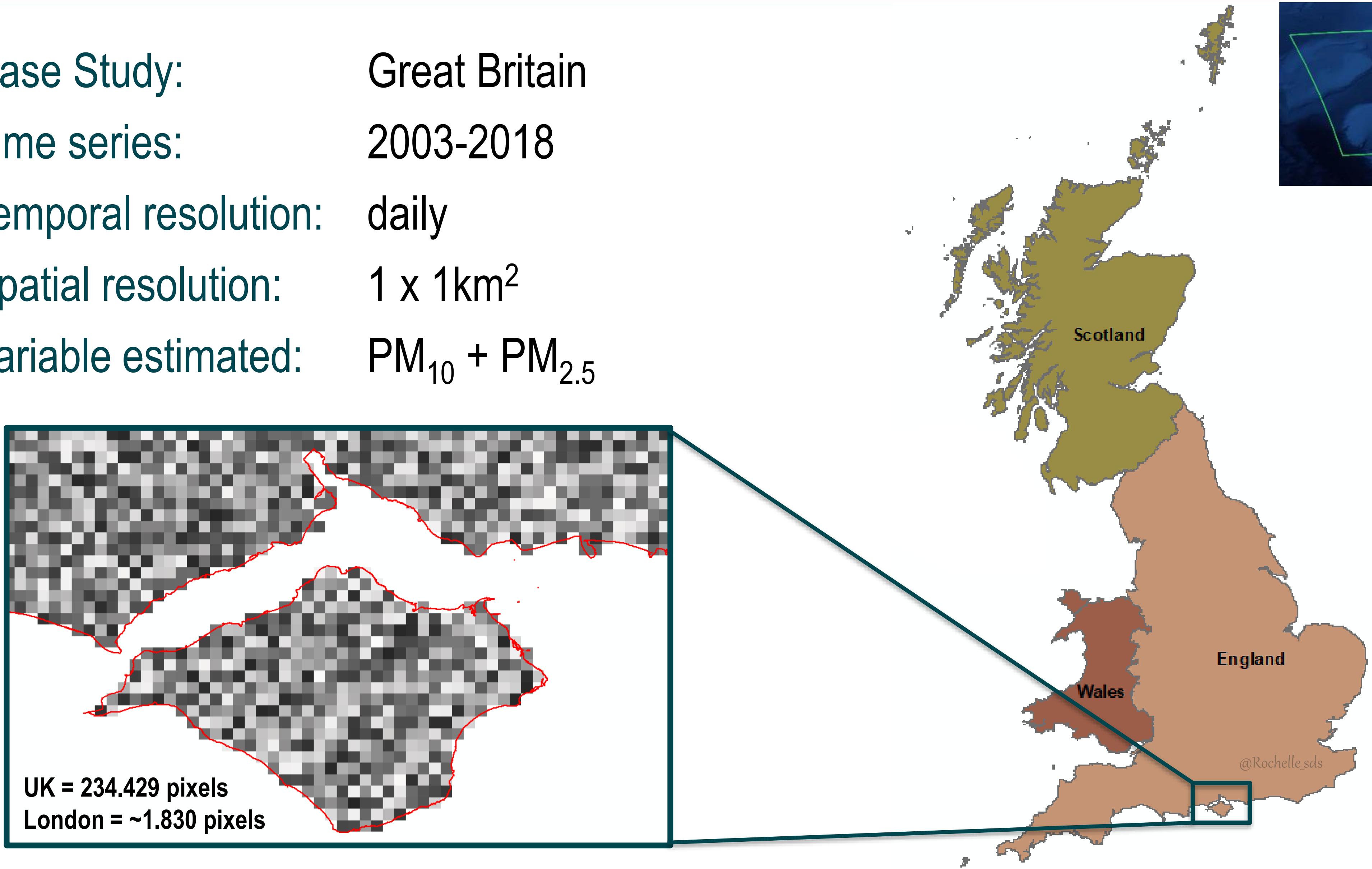
(Spatio-temporal UK exposure modelling)

Estimation of daily PM_{10} and $PM_{2.5}$ concentrations using satellite-based machine learning models

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

CASE STUDY

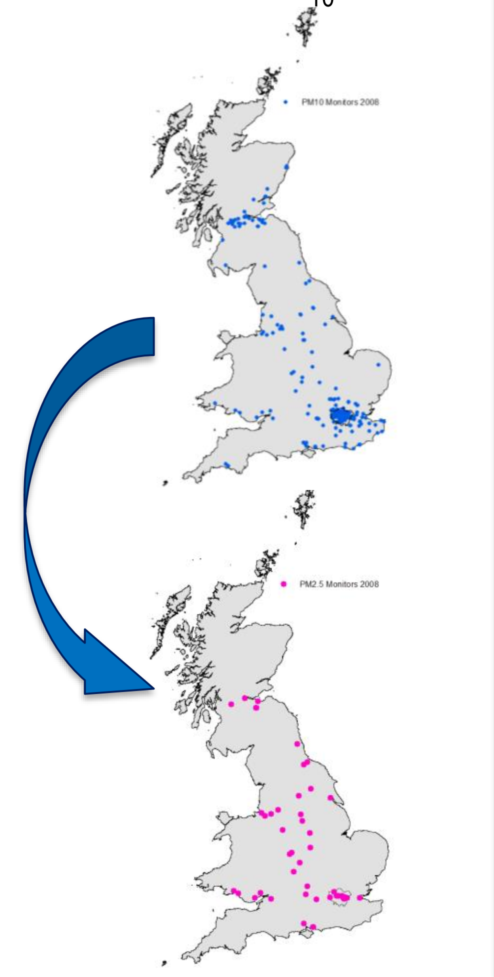
Case Study:	Great Britain
Time series:	2003-2018
Temporal resolution:	daily
Spatial resolution:	1 x 1km^2
Variable estimated:	PM_{10} + $PM_{2.5}$

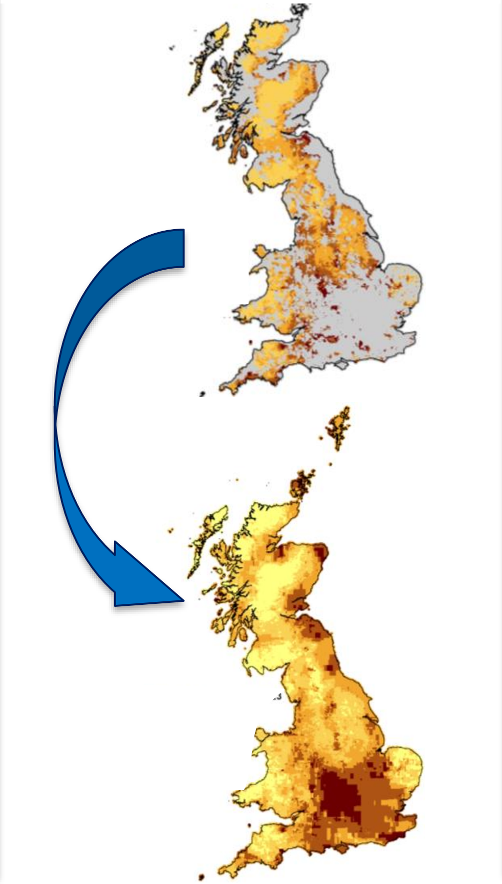


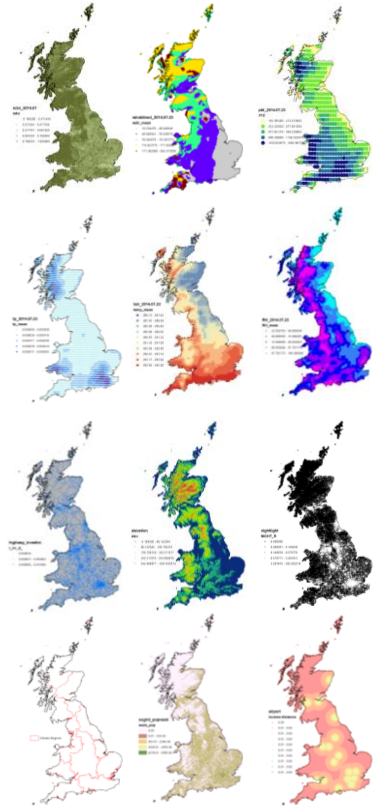
Multi-stage machine learning spatio-temporal model for $PM_{2.5}$

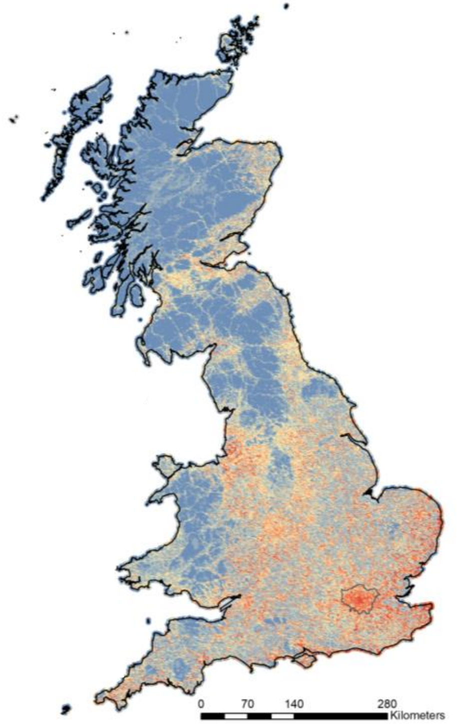
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

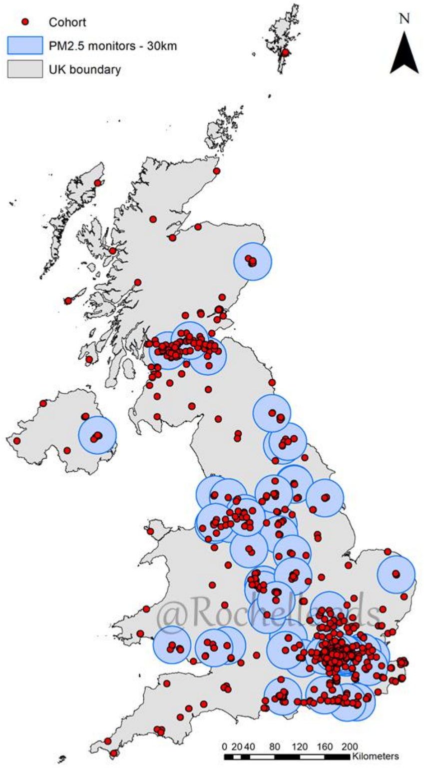
Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Predict $PM_{2.5}$ concentrations from co-located PM_{10} stations.	Predict satellite-aerosol from Copernicus atmospheric reanalysis aerosol.	Test different satellite-based machine learning models to predict $PM_{2.5}$ concentrations from the predictors.	Predict daily $PM_{2.5}$ concentrations at 1 km^2 using the parsimonious satellite-based machine learning model.	Predict $PM_{2.5}$ concentrations at the monitor level using small-scale predictors





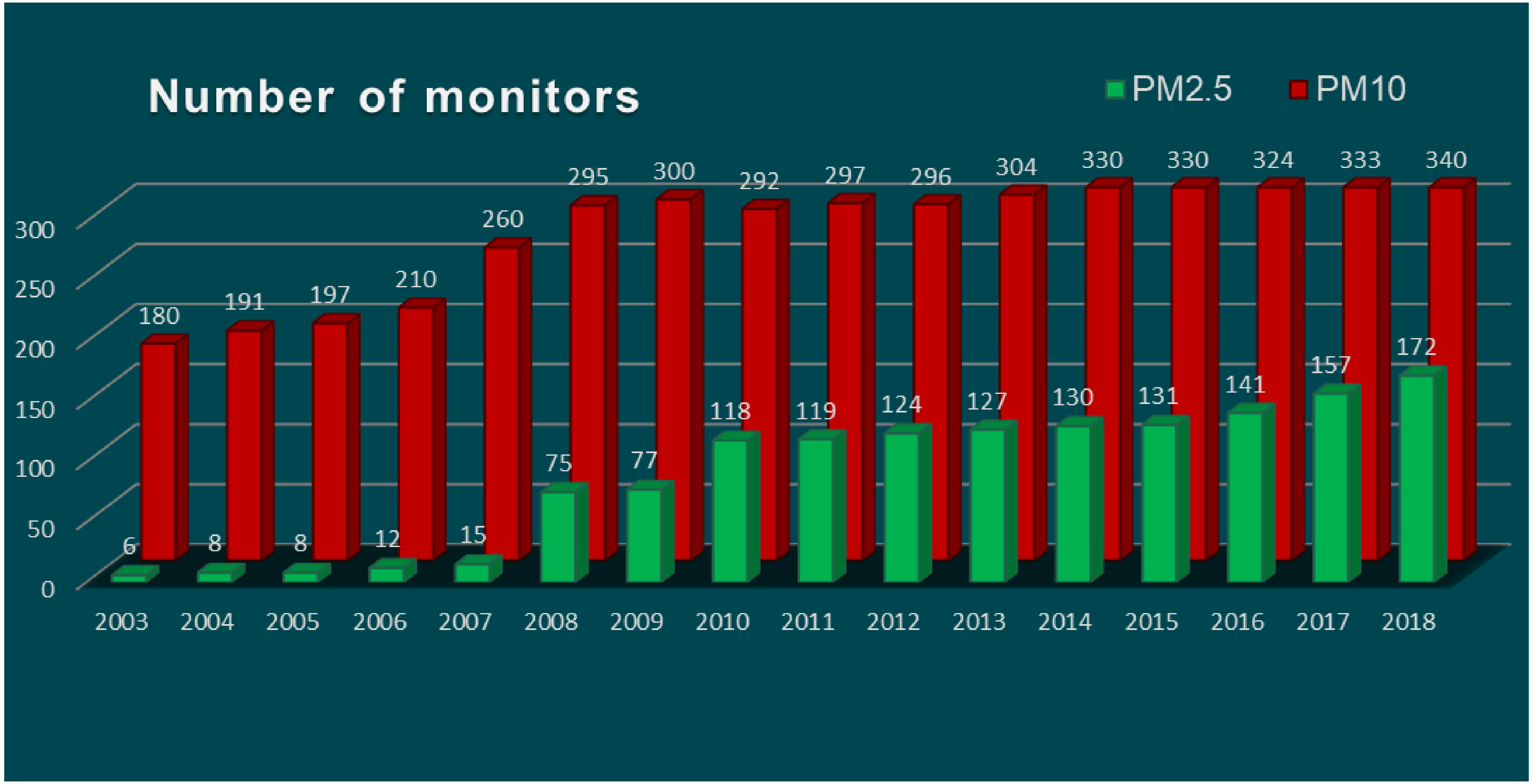




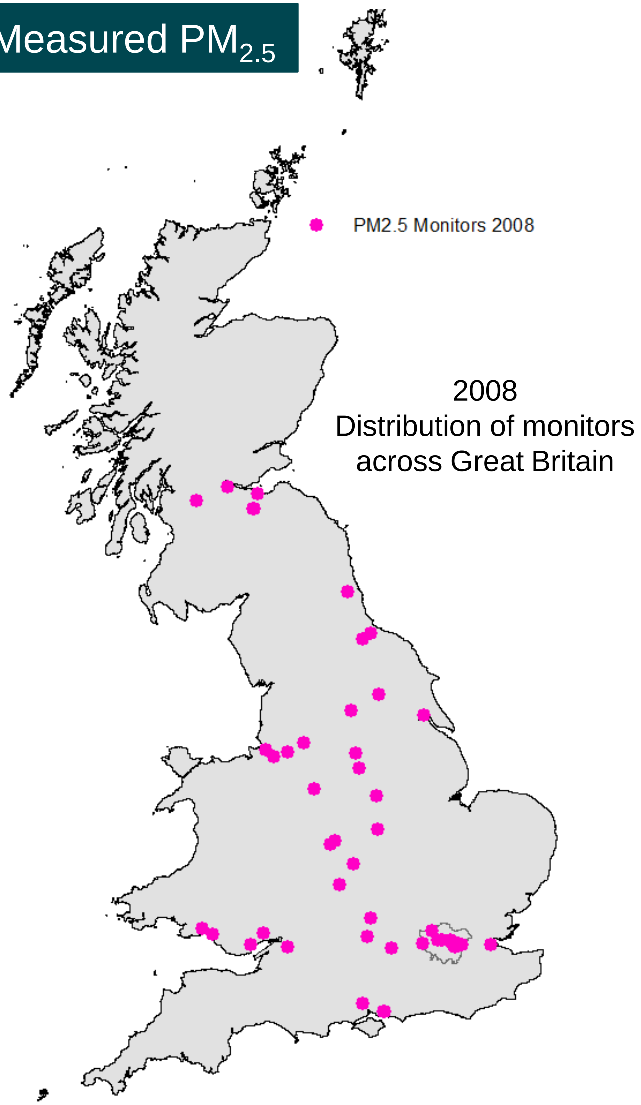


STAGE 1: Gap-filling Model

Predict daily $PM_{2.5}$ concentrations from co-located PM_{10} stations



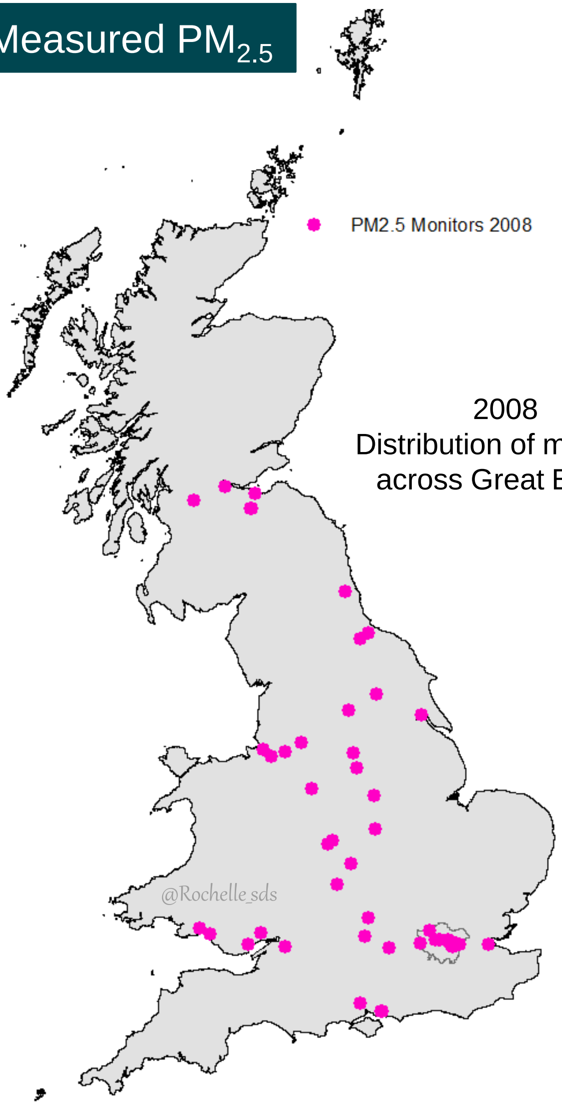
Measured $PM_{2.5}$



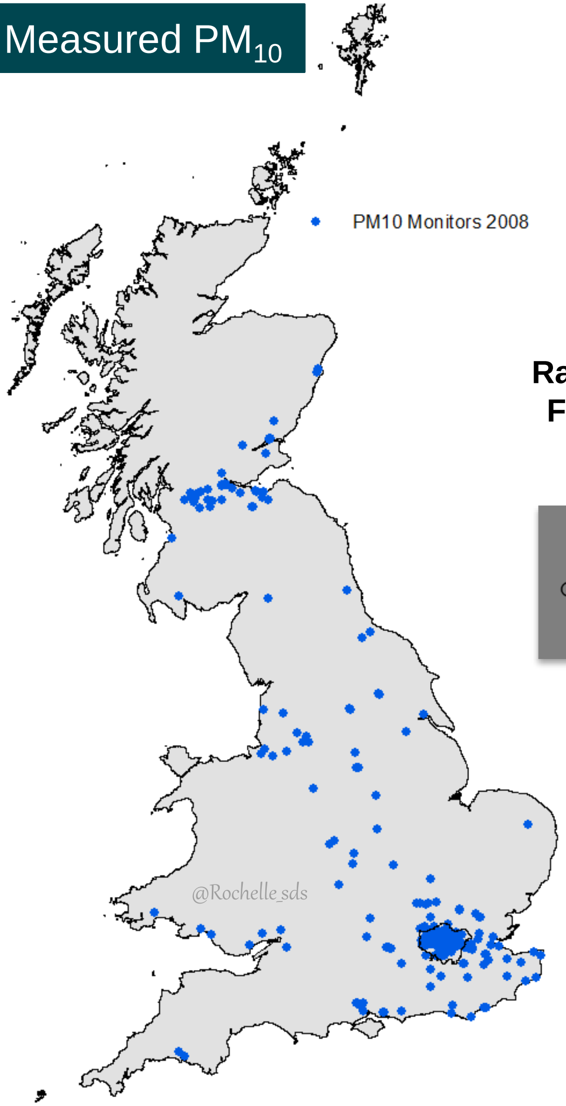
STAGE 1: Gap-filling Model

Predict daily $PM_{2.5}$ concentrations from co-located PM_{10} stations

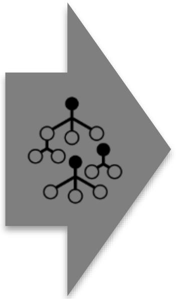
Measured $PM_{2.5}$



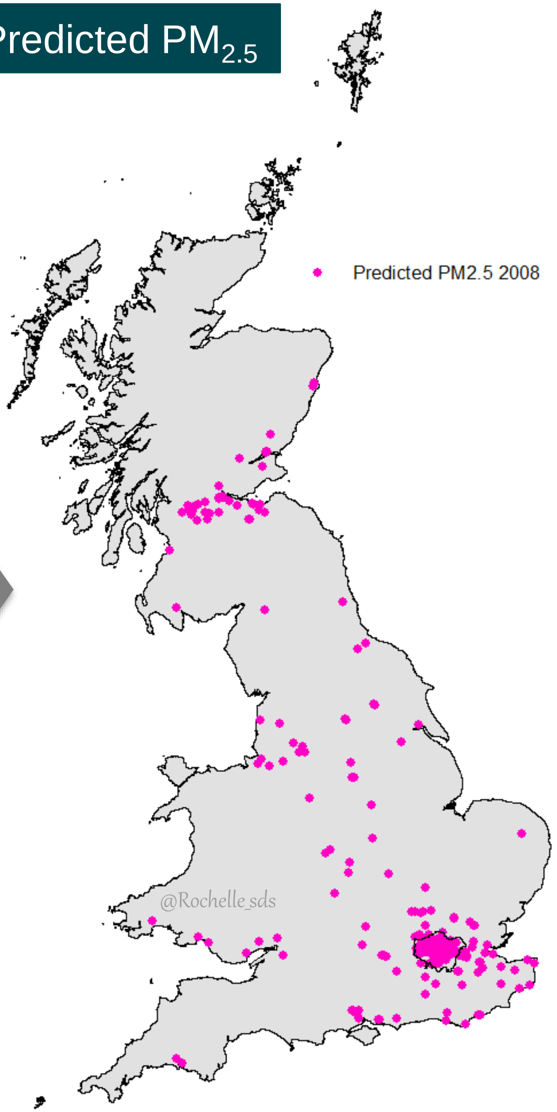
Measured PM_{10}

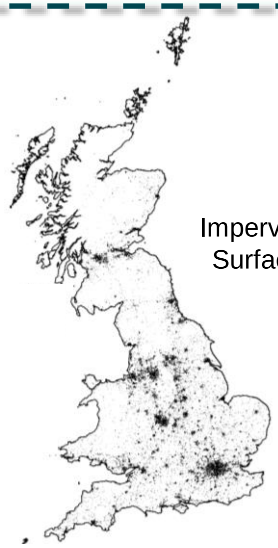


Random Forest

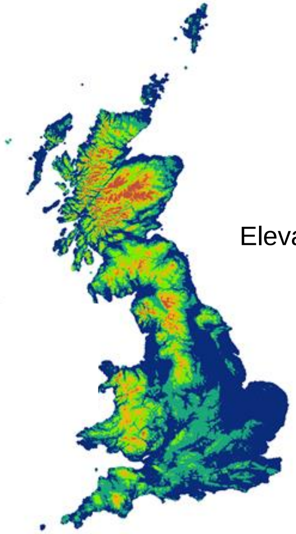


Predicted $PM_{2.5}$





Impervious Surfaces



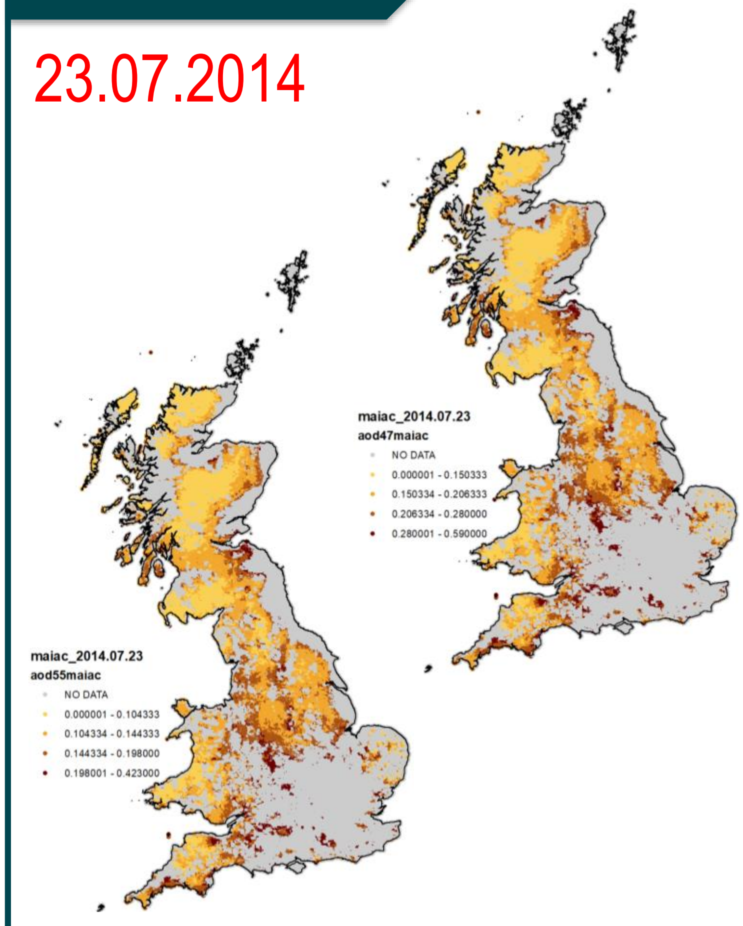
Elevation

STAGE 2: AOD GAP-FILLING MODEL

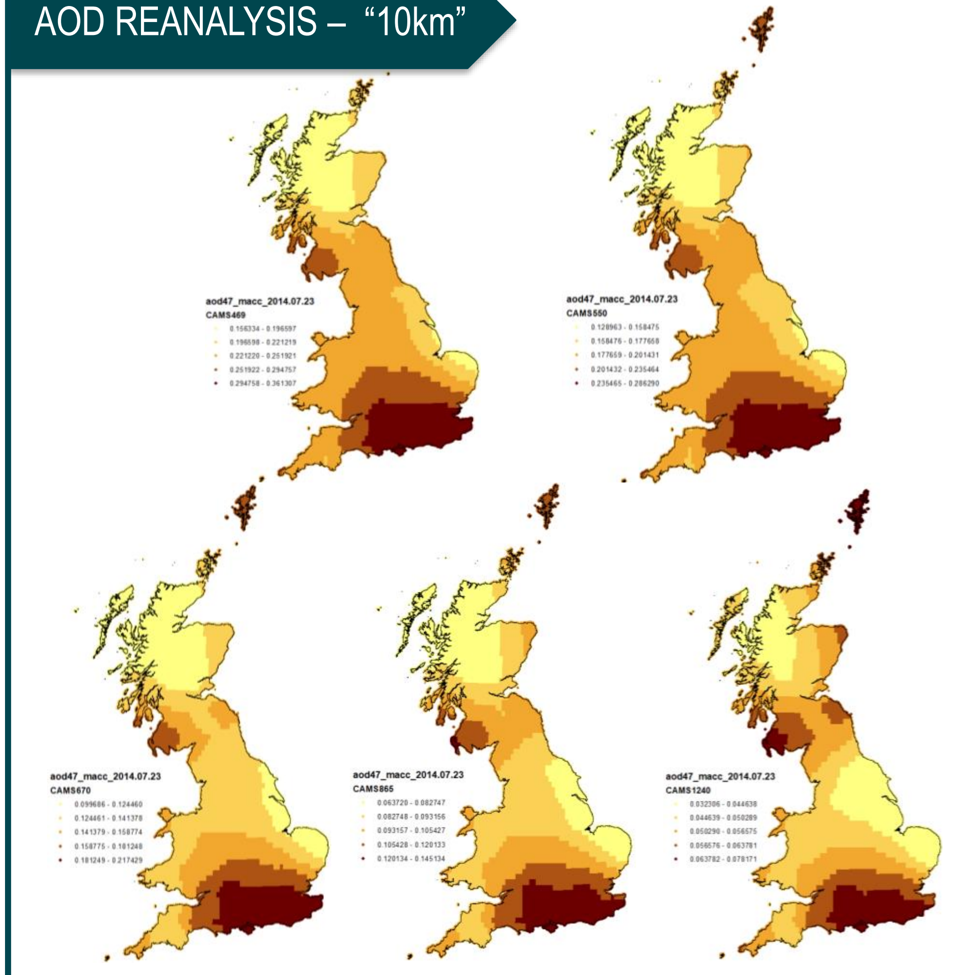
Predict AOD (2 λ / day) from MAIAC-MODIS using AOD (5λ, 7t) from CAMS

AOD MAIAC – 1km

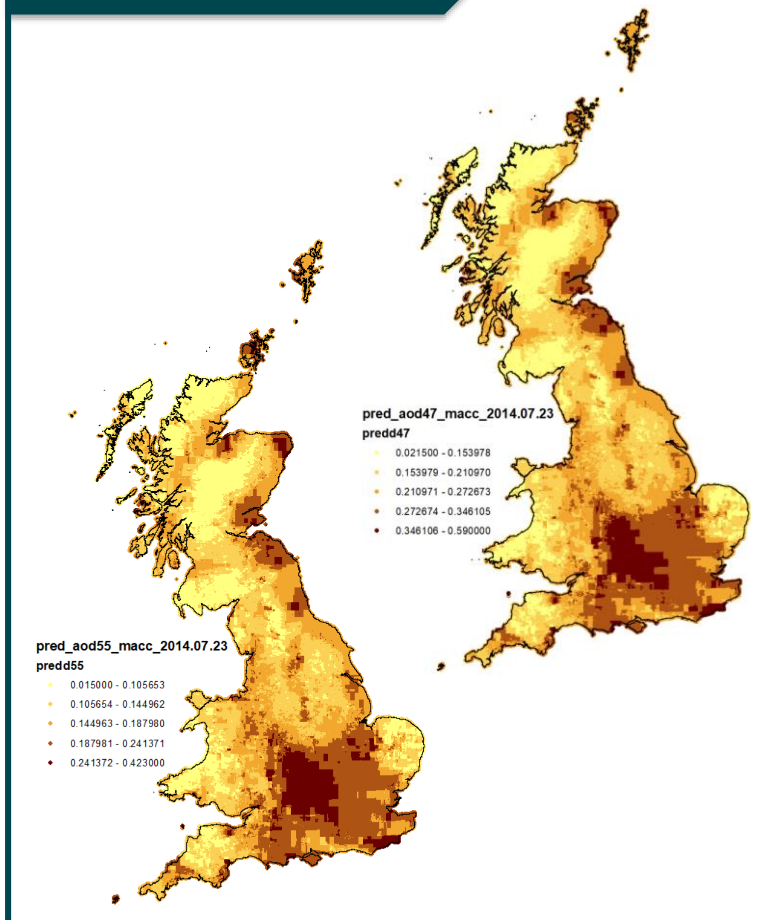
23.07.2014



AOD REANALYSIS – "10km"

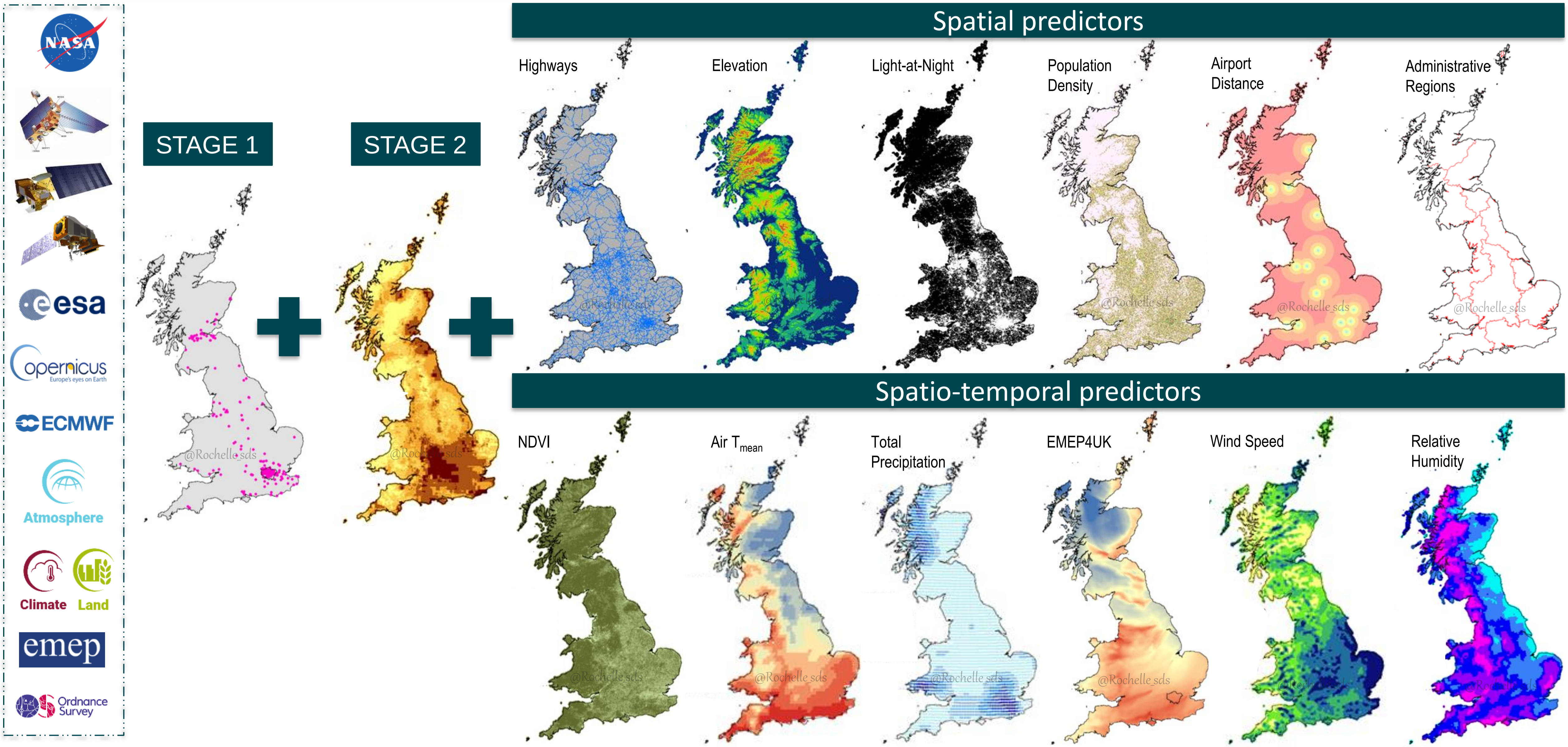


PREDICTED AOD – 1km



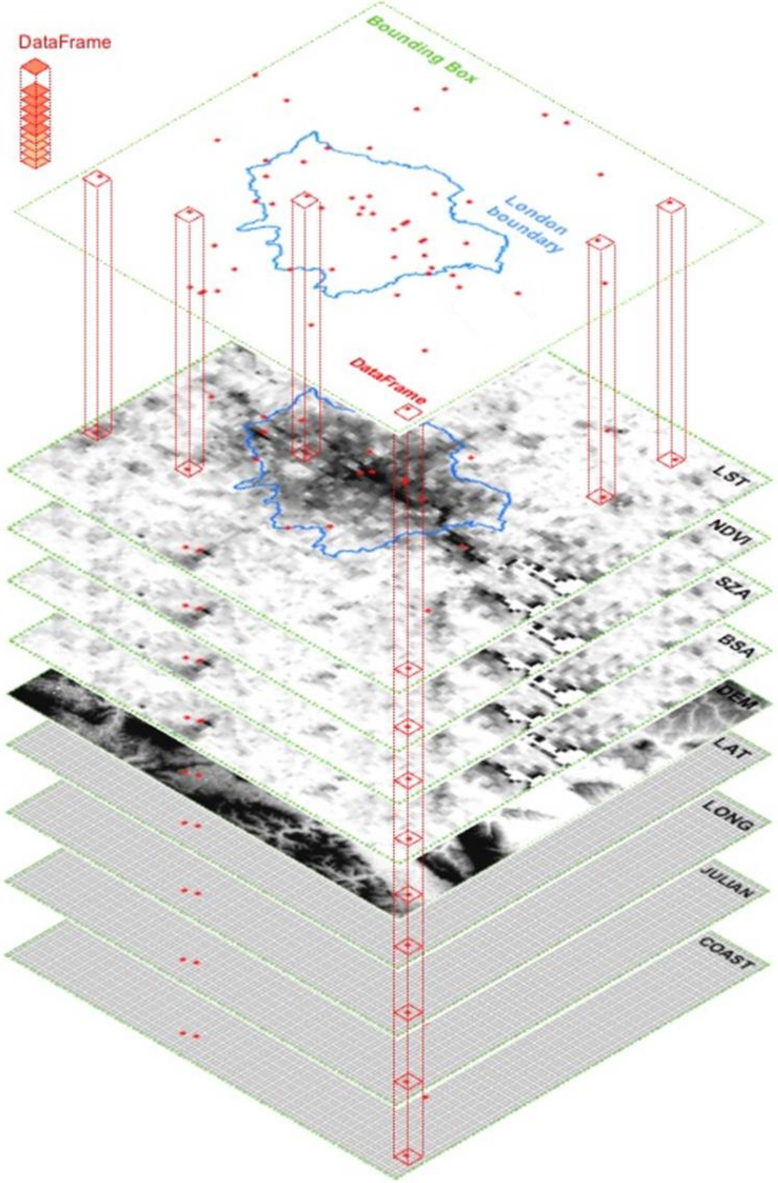
STAGE 3: Test ≠ Satellite-based ML Models

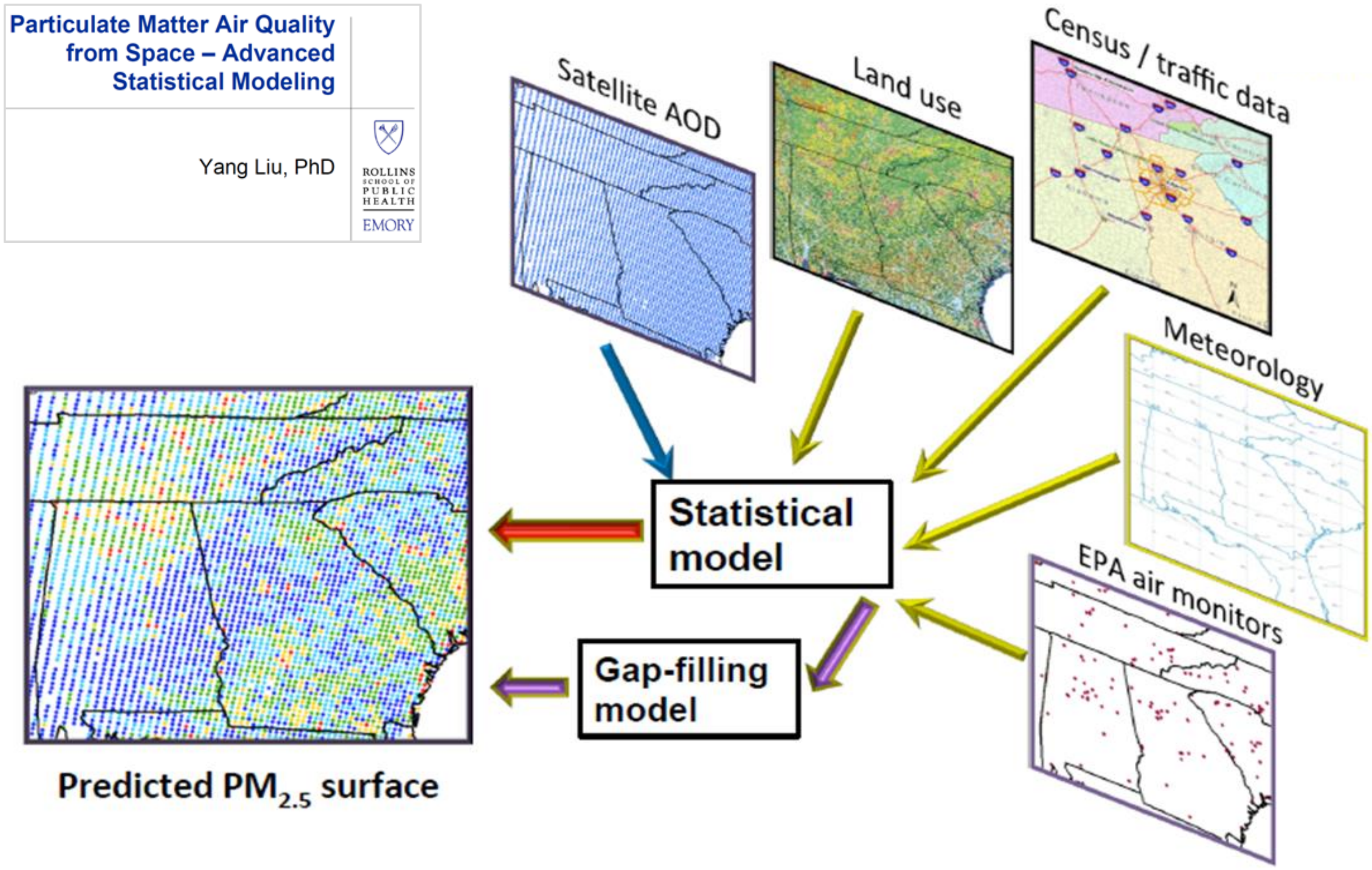
Predict $PM_{2.5}$ at 1 km^2 using the parsimonious satellite-based ML model



Data Synchronisation + Gap-filling Model + Advanced Statistical Modelling

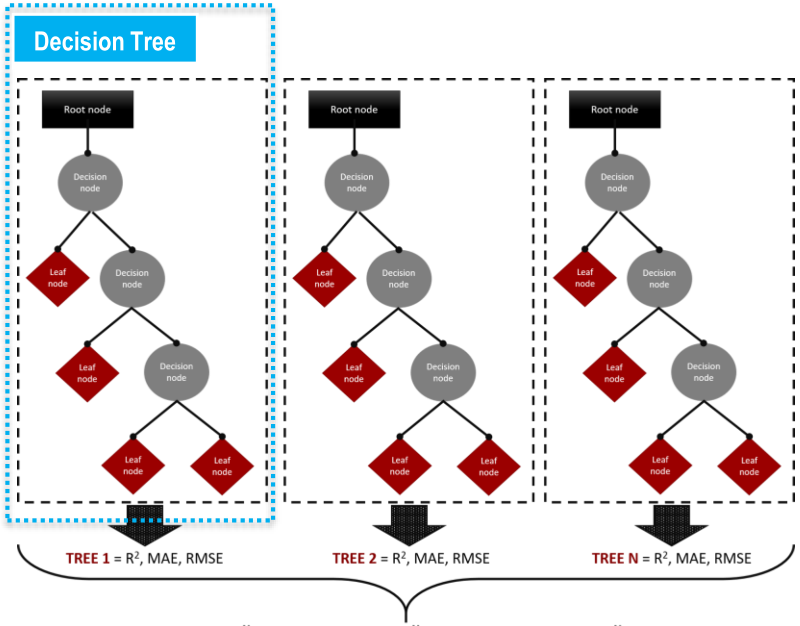
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE





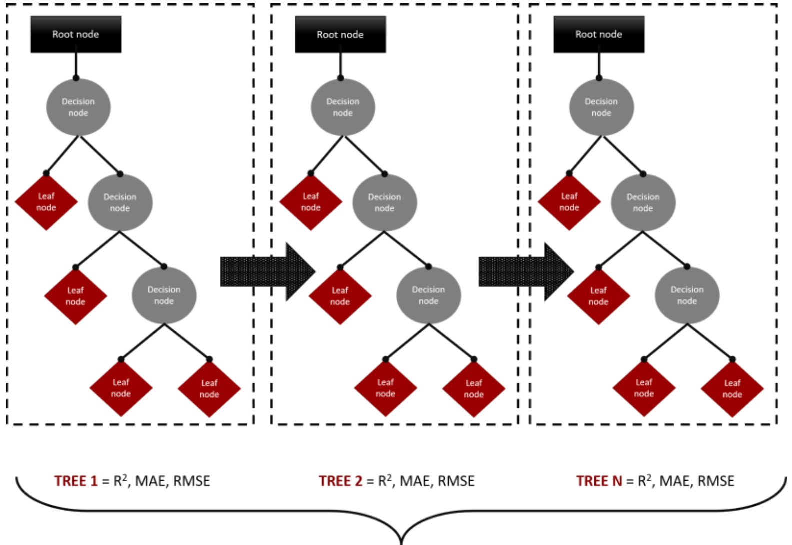
Methods: Machine Learning Algorithms

Random Forest



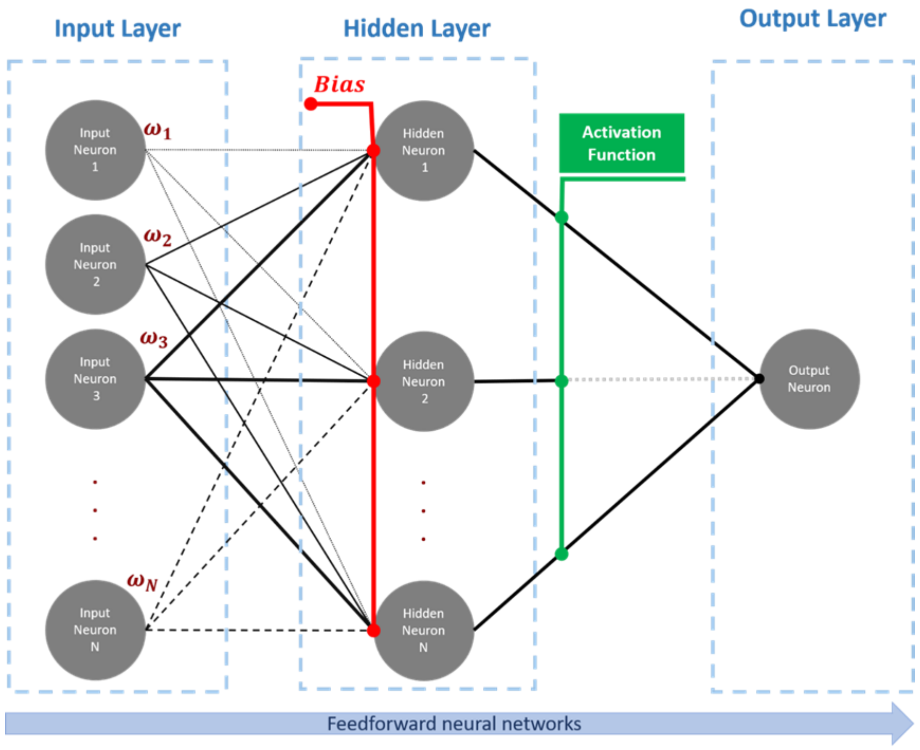
$$R^2_{mean} = \frac{1}{N}\sum_{i=1}^{N} R_i^2 \qquad MAE_{mean} = \frac{1}{N}\sum_{i=1}^{N} MAE_i \qquad RMSE_{mean} = \frac{1}{N}\sum_{i=1}^{N} RMSE_i$$

Gradient Boosting



$$R^2_{mean} = \frac{1}{N}\sum_{i=1}^{N} wR_i^2 \qquad MAE_{mean} = \frac{1}{N}\sum_{i=1}^{N} wMAE_i \qquad RMSE_{mean} = \frac{1}{N}\sum_{i=1}^{N} wRMSE_i$$

Neural Network



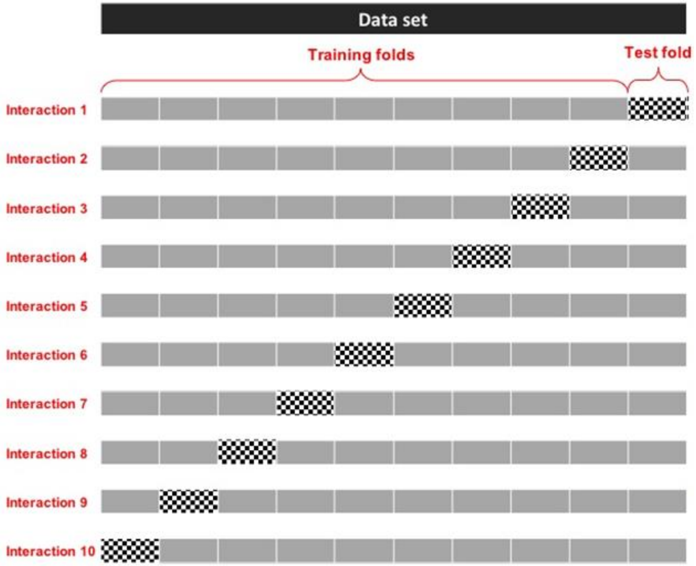
Results: STAGE 1 - Gap-filling Model

```r
################################################################
### OPTIMISATION USING SPATIO-TEMPORAL CROSS-VALIDATION ###
################################################################

for (ntree in c(100, 500, 1000)){
  for (mtry in c(3,5,7)){                    ## DEFAULT: sqrt(num.predictors)

    cat("\n", "ntree=", ntree, "| mtry=", mtry, "\n")

    final = foreach(i=1:10, .packages=c("data.table","ranger"))%dopar% {

      mod       <- ranger(pm25 ~ pm10           + julian    +
                                 latitude       + longitude +
                                 SiteTypeCode + year      +
                                 month          + dow,
                         data=subset(stg1, split!=i),
                         num.trees=ntree,
                         mtry=mtry,
                         respect.unordered.factors=TRUE,
                         verbose=TRUE )

      test          <- subset(stg1, split==i)
      test$pred.rf  <- predict(mod, test)$predictions
      test$itercv   <- i
      test[, list(date, code.source, pm25, pred.rf, itercv)]}

  mod.cv = do.call(rbind, final)

###########################################
### REGRESS PREDICTED AGAINST OBSERVED  ###
###########################################
linear.pm25   <- lm(pm25~pred.rf, data=mod.cv)
r2.all        <- summary(linear.pm25)$r.squared
```

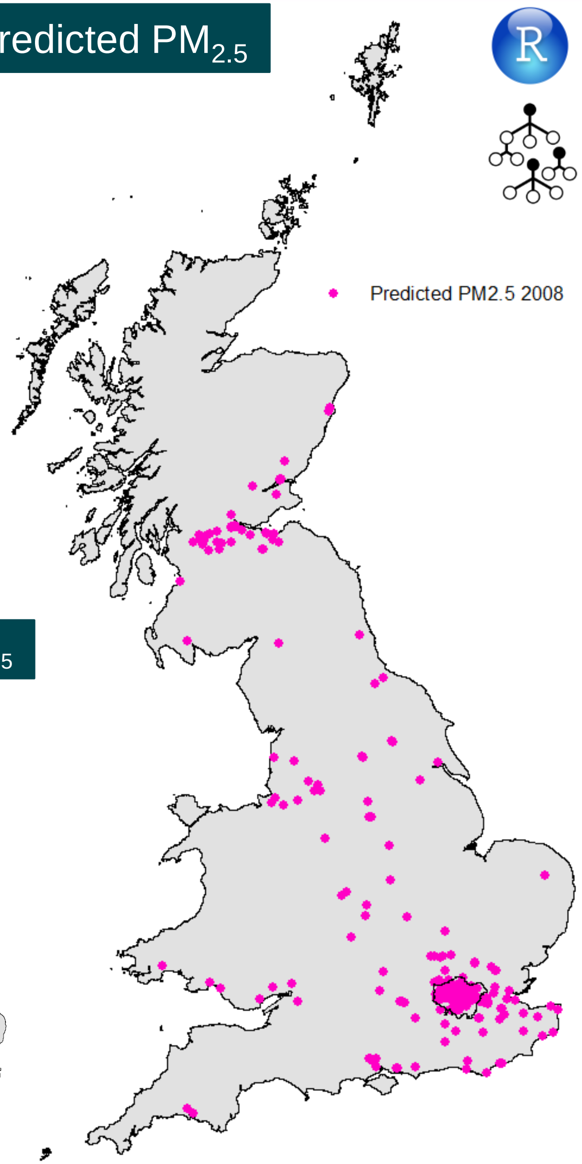
```r
###############
### RESULTS ###
###############
```

variable.importance	
pm10	15351814.3
longitude	1158180.8
latitude	1066338.3
julian	875359.1
year	868112.3
month	505182.1
SiteTypeCode	292421.0
dow	266770.8

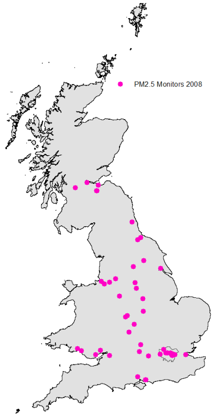
R^2 *all*	ntree=100	ntree=500	ntree=1000
mtry=3	0.860	0.861	0.861
mtry=5	0.849	0.850	0.850
mtry=7	0.842	0.843	0.843

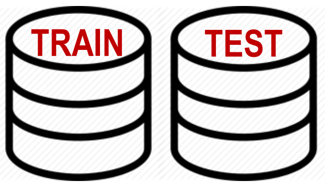


Predicted $PM_{2.5}$



Measured $PM_{2.5}$





Predictors: 8 variables

Train/Test data: Air Quality monitors

Total Size: ~ 260.000

Model Design: 1 model for 16 years

Results: STAGE 2 – AOD Gap-filling Model

```r
##################################
# \\\\\\ CROSS-VALIDATION \\\\\ #
##################################

ranger.model <- caret::train (formula,
                              data      = stage2.train,
                              method    = "ranger",
                              metric    = "RMSE",
                              num.trees = 50,
                              tuneGrid  = expand.grid(.mtry = 20),
                              trControl = trainControl(
                                          method="cv",
                                          number=10,
                                          allowParallel=T,
                                          verboseIter=T)
                              )

##############################################
#### REGRESS THE PREDICTED AGAINST OBSERVED ####
##############################################

stage2.valid$pred  <- predict(ranger.model, stage2.valid)
linear.model       <- lm(formula.linear, data=stage2.valid)
r2.stage2          <- summary(linear.model)$r.squared
```

- Optimised RF : ntree=50 | mtry=20

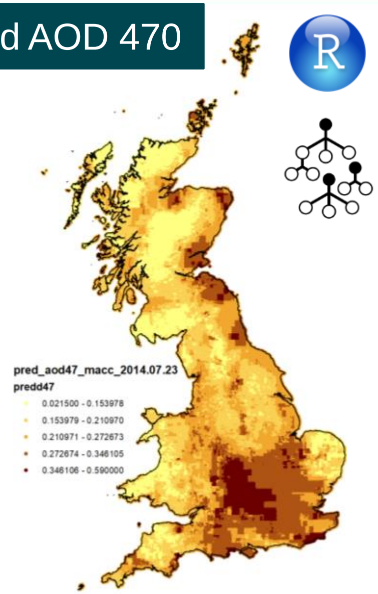
```
###############
### RESULTS ###
###############
```

variable.importance

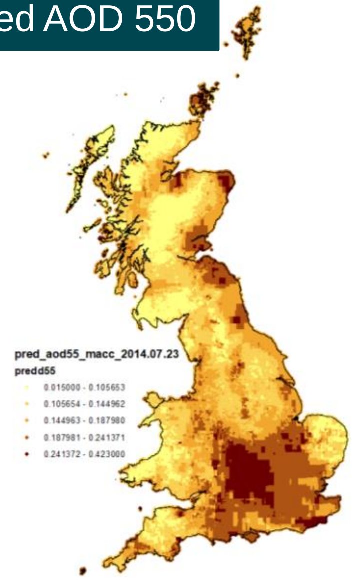
Julian	9926.3528
CAMS_469_12	4722.1932
CAMS_550_12	4340.4057
y	2176.8599
x	1818.0451
CAMS_AOD670_12	874.2431
CAMS_AOD865_12	572.2148
CAMS_AOD469_15	531.0689
CAMS_AOD1240_12	513.8520
CAMS_AOD1240_18	479.6304

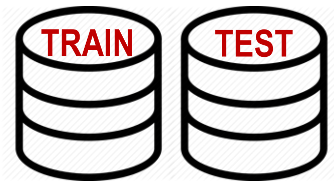
R^2 .stage 2	AOD 470	AOD 550
2003	0.960	0.960
2004	0.926	0.926
2005	0.935	0.934
2006	0.955	0.955
2007	0.960	0.961
2008	0.932	0.931
2009	0.935	0.935
2010	0.919	0.918
2011	0.957	0.970
2012	0.939	0.939
2013	0.942	0.942
2014	0.921	0.920
2015	0.914	0.914
2016	0.923	0.923
2017	0.911	0.910
2018	0.921	0.921

Predicted AOD 470



Predicted AOD 550





Predictors: 43 variables

Train/Test data: Satellite 1km^2 grid

Total Size: ~ 7 Million by year and AOD type (112 Million for 16 years)

Model Design: 1 model by year and AOD type (i.e. 16 models for AOD 470 and 16 for AOD550)

Results: STAGE 3 – Test ≠ Satellite-based ML Models

```r
###########################################################
### OPTIMISATION USING SPATIO-TEMPORAL CROSS-VALIDATION ###
###########################################################

for (ntree in c(100, 500,1000)){
  for (mtry in c(7,70,20)){                    ## DEFAULT: sqrt(42) = 7

    cat("\n", "ntree=", ntree, "| mtry=", mtry, "\n")

    final=foreach (i=1:10, .packages=c("data.table","ranger")) %dopar% {
      print(i)

      mod                  <- ranger(formula.pm25,
                                     data=subset(stg3, split!=i),
                                     num.trees=ntree,
                                     mtry=mtry,
                                     respect.unordered.factors=TRUE,
                                     verbose=TRUE)
      test                 <- subset(stg3, split==i)
      test$logpred.rf      <- predict(mod, test)$predictions
      test$exppred.rf      <- exp(test$logpred.rf)
      test$itercv <- i
      test[, list(day, osgrid.id, code.source, pm25pred.meas,
                  exppred.rf, logpm25, logpred.rf, itercv)]
    }

    stage3.cv = do.call(rbind, final)

    ################################################
    #### REGRESS THE PREDICTED AGAINST OBSERVED ####
    ################################################
    linear.all   <- lm(pm25pred.meas~exppred.rf, data=stage3.cv)
    r2.all       <- summary(linear.all)$r.squared
}}
```

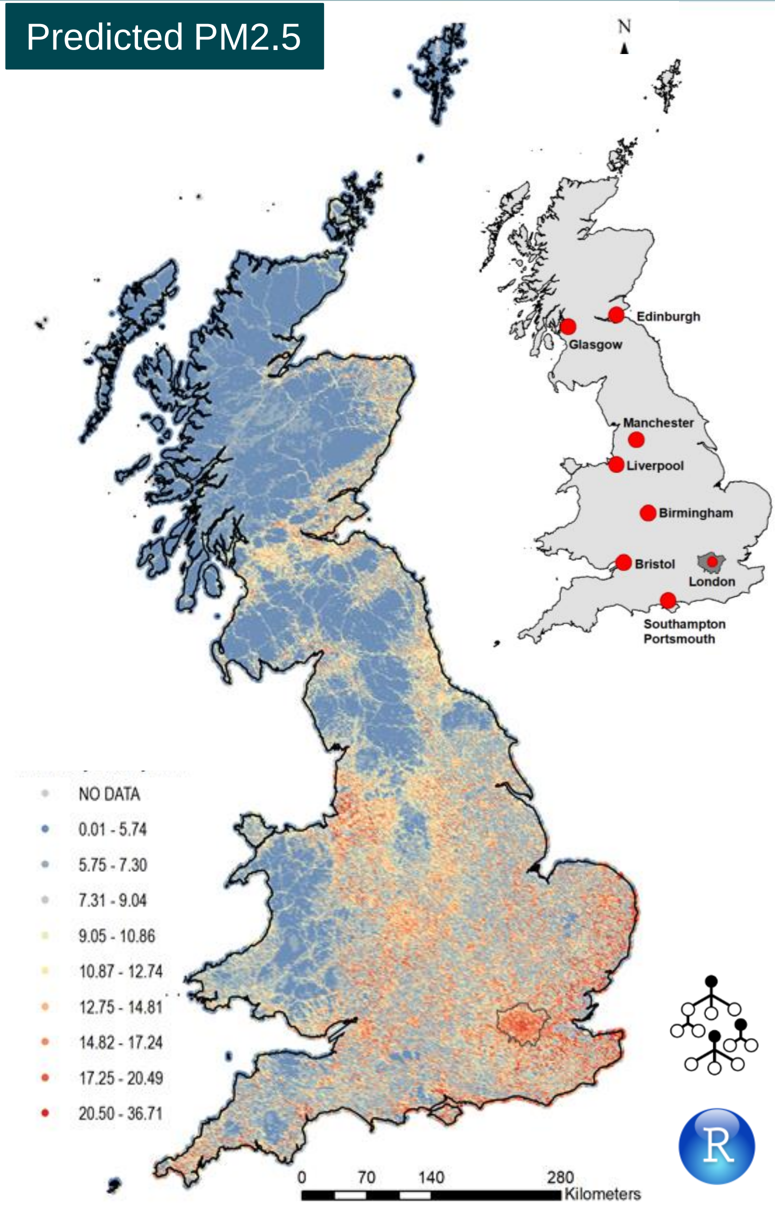
```
###############
### RESULTS ###
###############
```

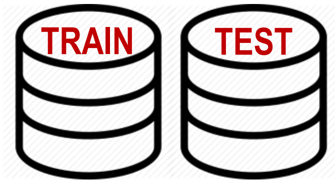
Variable Importance

Wind Direction	3774.673120
Distance from Sea	2921.701911
Wind Speed	1803.228131
Day of year	1780.716491
Mean Precipitation	1725.783224
Month(factor)	1700.169657
Planetary Boundary Layer_12h	1471.014341
Mean Temperature	1408.845640
Planetary Boundary Layer_0h	1291.013732
Mean Sea Pressure	1103.636722
Elevation	873.211102
Stage2-AOD550	833.542341
Stage2-AOD470	822.952578

R^2.all	ntree=100	ntree=500	ntree=1000
mtry=7	0.670	0.675	0.675
mtry=10	0.687	0.691	0.691
mtry=20	0.700	0.702	0.702

Predicted PM2.5





Predictors: 42 variables

Train/Test data: Air Quality monitors

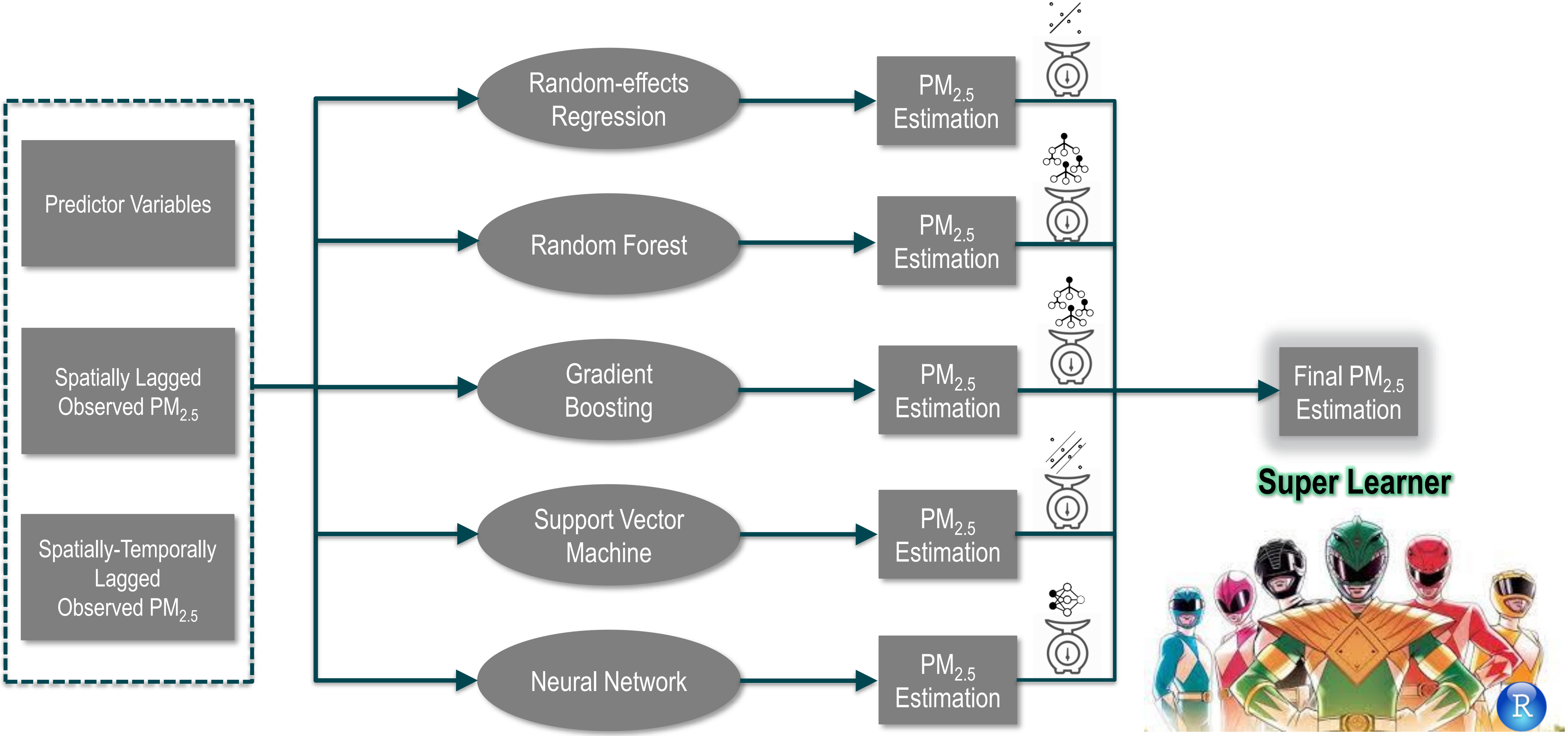
Total Size: ~ 100.000 by year

Model Design: 1 model by year

Next Steps

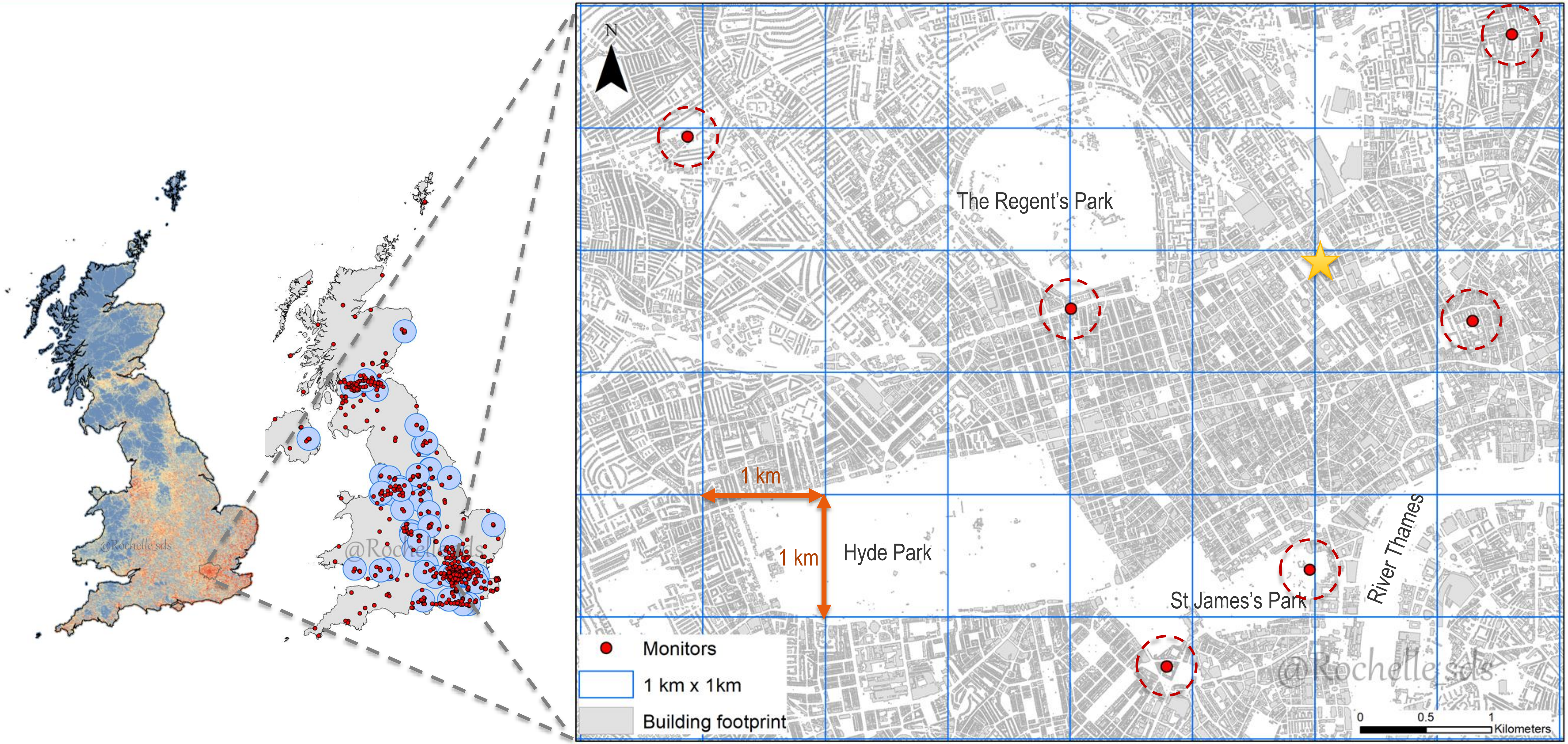
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Explore the Ensemble ML Models



STAGE 5: Downscaling the $PM_{2.5}$ Estimations

Predict $PM_{2.5}$ concentrations at the monitor level using small-scale predictors



Connect with Health Data

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Copernicus
Europe's eyes on Earth

ECMWF

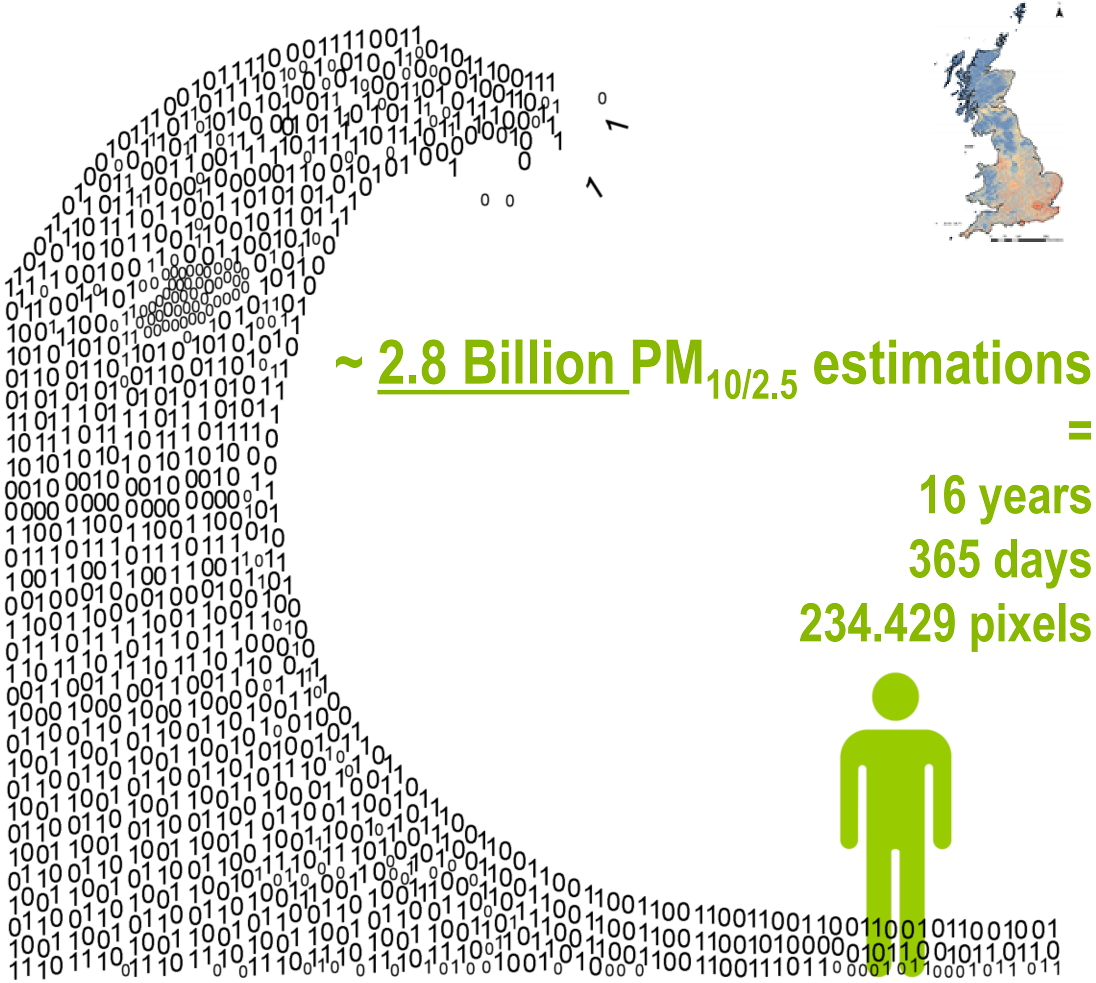
esa

Atmosphere

Climate Land

emep

Ordnance Survey



~ 2.8 Billion $PM_{10/2.5}$ estimations

=

16 years

365 days

234.429 pixels

+

biobank uk
Improving the health of future generations

NHS Digital

Hospital Episode Statistics (HES)

ELSA English Longitudinal Study of Ageing

NEXT STEPS
LEARNING FROM YOUR GENERATION

Millennium Cohort Study
CHILD OF THE NEW CENTURY

Longitudinal studies (cohorts)

Connect with Worldwide Health Data

MCC Collaborative Research Network
An international research program on the associations between environmental stressors, climate, and health

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

United Kingdom

biobank uk
Improving the health of future generations

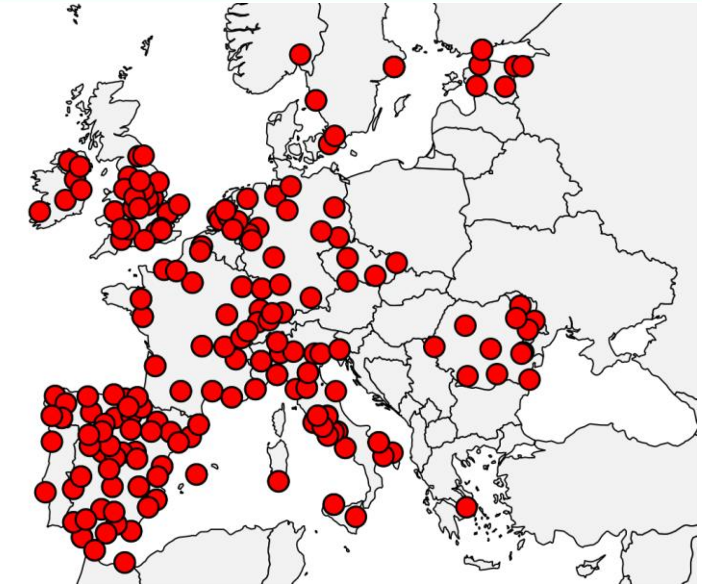
NHS Digital

BCS70
1970 British Cohort Study

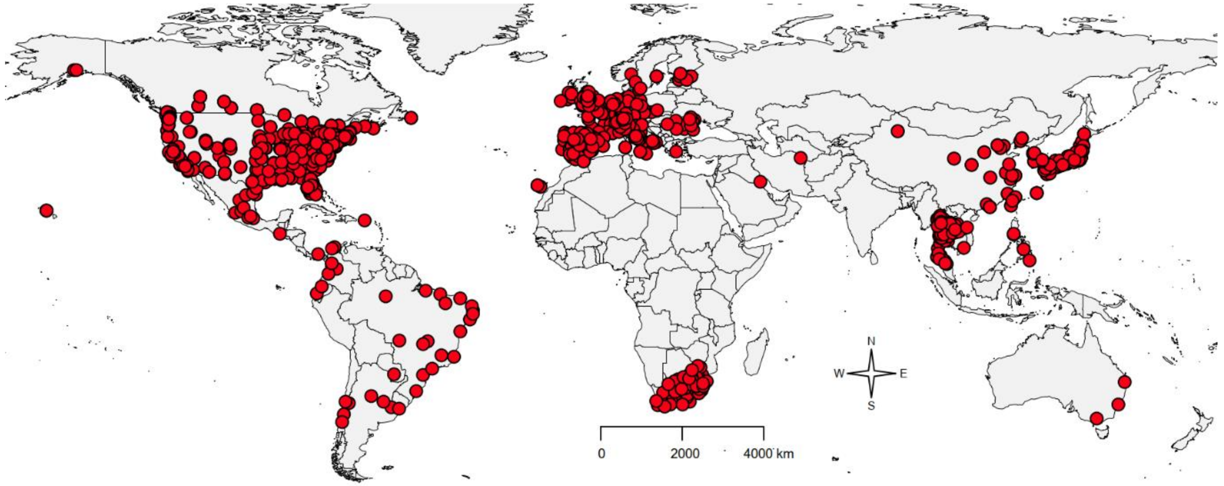
ELSA English Longitudinal Study of Ageing

Office for National Statistics

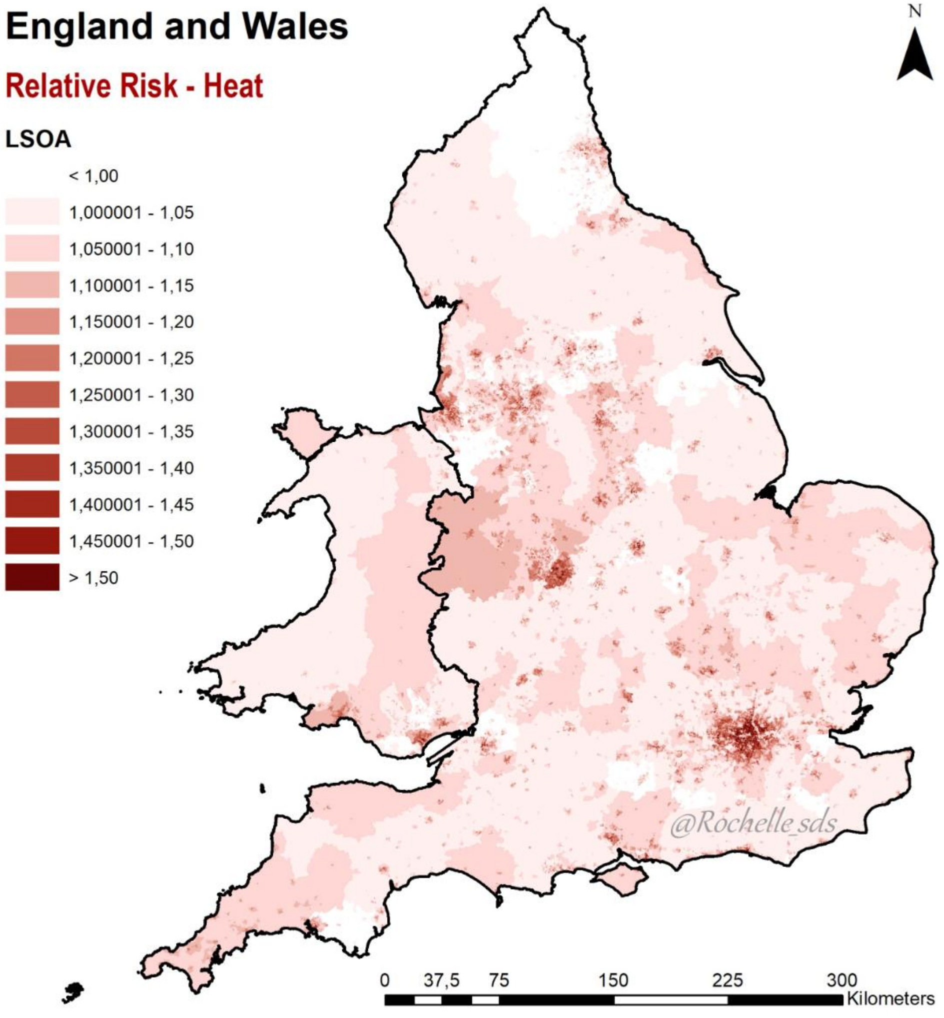
European MCC Locations

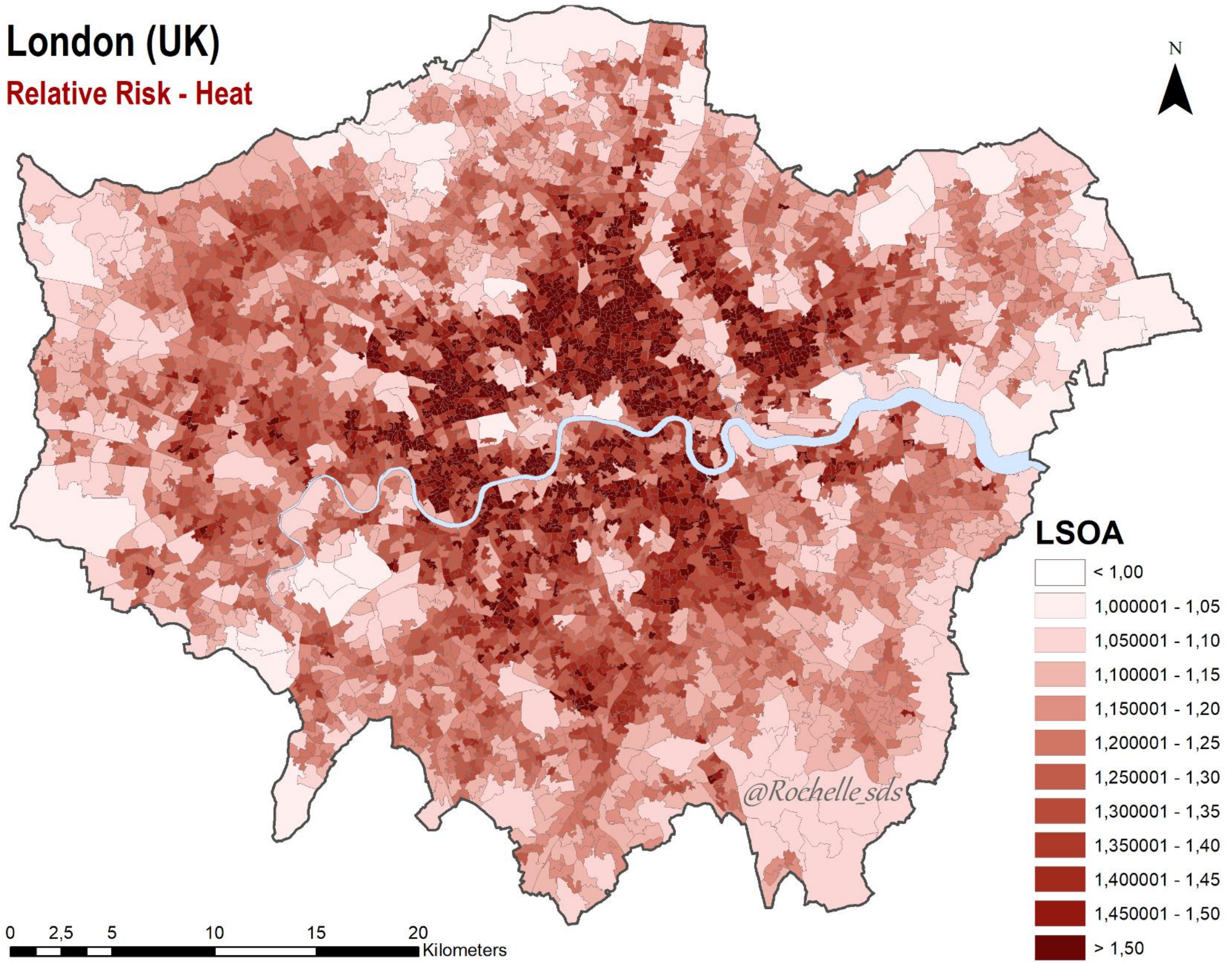


Global MCC Locations

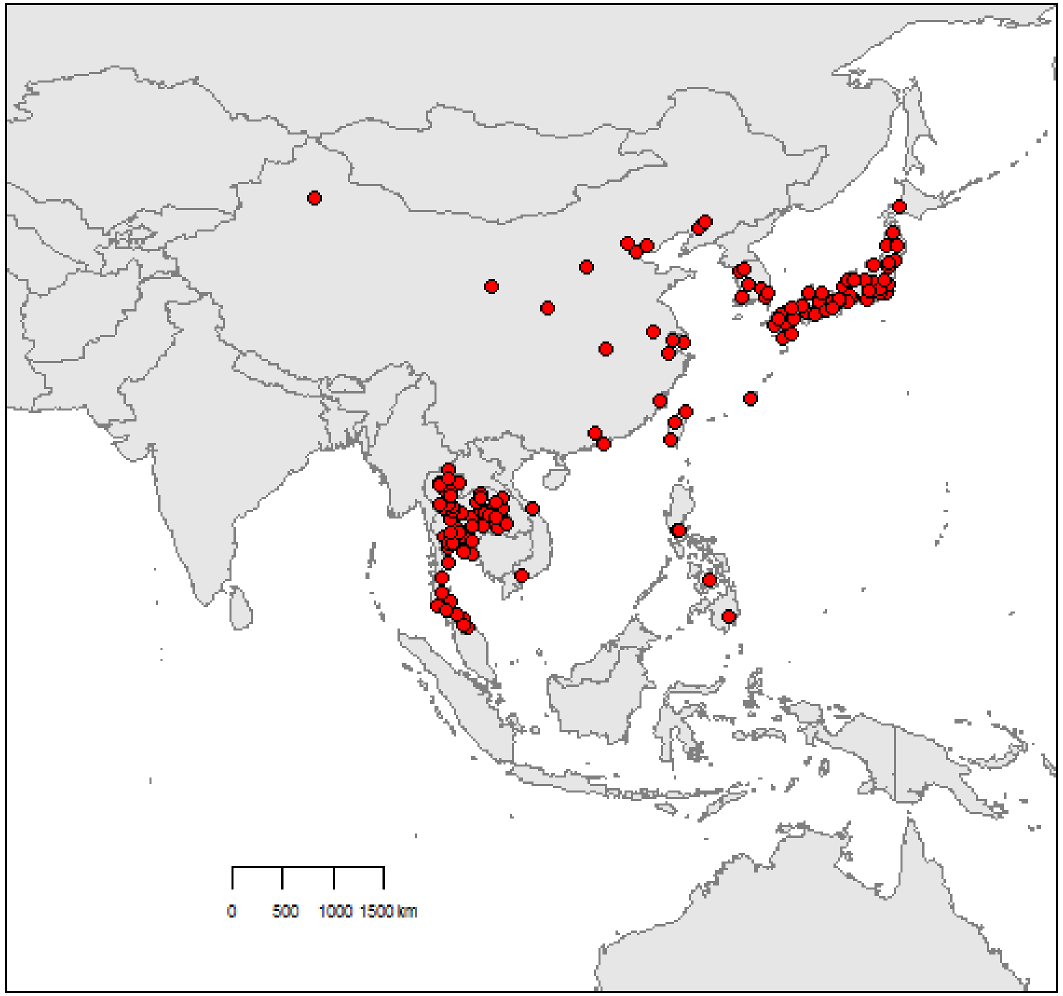


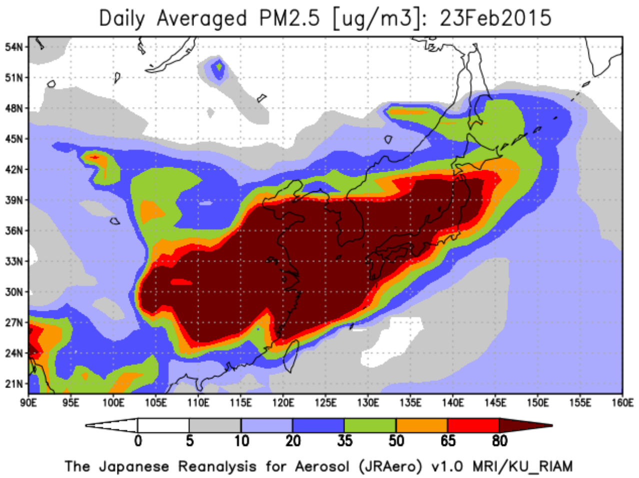
Modelling Area–specific Environmental Risks

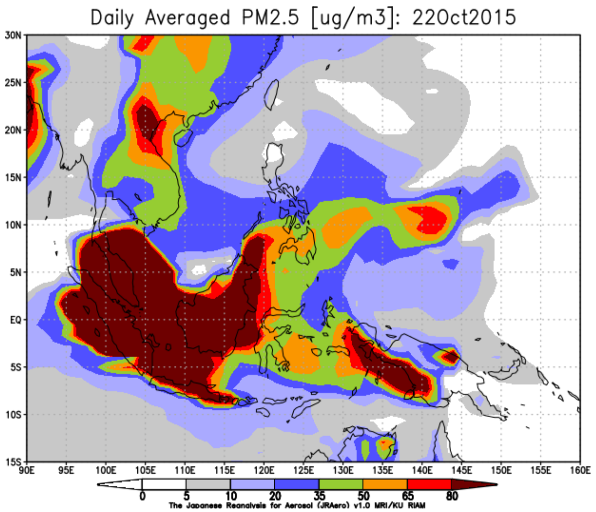




Analysis of Health Effects of Transboundary Pollution







Thank you

Rochelle Schneider
Senior Research Fellow in Geospatial Data Science

Antonio Gasparrini
Professor of Biostatistics and Epidemiology

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE