Introduction to methods in environmental epidemiology 1

Robbie M Parks, PhD

21st July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

A little bit about me…

An audio diary from the COP28 climate conference, Part 1

Sights, sounds and scenes from the largest climate gathering on the planet.

By EHN Editors • Dec 13, 2023 • 1 min read

Print | PDF | Email

SPARK LAB NYC

Environmental Health News

Donate

Newsletters ▾ Topics ▾ Special Projects ▾ About Us ▾ Guides ▾

Home > Originals > Justice

A protest inside COP27 demanding Loss and Damage funding. (Credit: Robbie Parks)

Mar 29, 2023

Op-ed: Mourning family and climate change in the age of loss and damage

Grief is a consequence of the natural cycle of life and death, but it can be exacerbated by negligence and unjust approaches to climate change.

Robbie Parks

SPARK Lab NYC, Environmental Health Sciences, Columbia University, New York

Climate Change 🤝 Health 🤝 Justice

SPARK Lab NYC

Goal: To improve health and wellbeing with research and action on climate-related exposures.

By: Uncovering hidden burdens of climate-related exposures on health, including:

- How health is impacted by climate-related exposures
- Who is impacted most
- What needs to be done for positive change

Work relevant to:

- Climate and health
- Mental and behavioral health
- Environmental justice
- Urban health

SPARK Lab NYC

The Team
Publications
Press Coverage
GitHub
Lab Wiki
Reddit R/Heat_prep
G-TROPIC
HEATE
Lab Activities
What Does It Mean?

Robbie M. Parks, PhD

About Me
CV
Columbia University Faculty Website
P8301: Atmospheric And Climate Science For Public Health
SHARP Bayesian Modeling For Environmental Health Workshop
GitHub
Bluesky
LinkedIn
Google Scholar
ORCID
Global Heat Health Information Network (GHHIN)
E-Mail

SPARK Lab NYC

The Team

01. Current Lab Associates
02. Previous Lab Associates
03. Future Lab Associates

SPARK LAB NYC

Current Lab Associates

Felipe L. Antequera (Post-doctoral Fellow)
Wil Lieberman-Cribbin (Post-doctoral Fellow)
Victoria D. Lynch (Post-doctoral Fellow)
Ufuoma Ovienmhada (Post-doctoral Fellow)

Theo Rashid (Research Software Engineer)

Heather McBrien (PhD Candidate)
Raenita (Rae) Spriggs (PhD Candidate)
Allison L. Stewart Ruano (PhD Candidate)

Fintan Mooney (PhD Student)

Lincole Jiang (Research Assistant)
Jenna Nicole Swartz (Research Assistant)

Hiroaki Murayama (Medical Student)
Lisa Yamasaki (Medical Doctor)

Maren Hale (Research Co-ordinator)

Jiayi Ge (Master's Student)
Yuanyu Lu (Master's Student)
Naomi Simon-Kumar (Master's Student)
Iris Yang (Master's Student)

Gabriela Imbert (Barnard Senior)
Hannah Sanghvi (Barnard Senior)

Jackson Chen (High school intern)

Acknowledgements

- Thank you to:
  - Dr. Marianthi-Anna Kioumourtzoglou (Brown University, USA) for course guidance and help for days 1 and 2.
  - Dr. Garyfallos Konstantinoudis (Imperial College London, UK) for input for day 3.
  - Dr. Cascade Tuholske (Montana State University, USA) for heat exposure lectures on day 3.

Goals for the workshop

- This workshop is a three-day intensive course of:
  - lectures; and
  - hands-on labs with code.

  - To provide an approachable and practical overview of concepts, techniques, and data analysis methods used in environmental epidemiology with applications in environmental health.

Goals for the workshop

- By the end of the workshop, participants should be familiar with the following topics:
  - Theories and assumptions.
  - Data structuring.
  - Model likelihood structures (Normal, Poisson, Bernoulli, Binomial).
  - Quantile regression.
  - Non-linear exposure-response curves.
  - Case crossover design.
  - Time series design.
  - Distributed lag non-linear models (DLNM).
  - Case crossover with hybrid DLNM.
  - Treed distributed non-linear models (TDLNMs).
  - Brief history of Bayesian methods.
  - Overview of Bayesian methods.
  - Bayesian regression.
  - Hierarchical modelling.

This course is for those who

- Are interested or who have heard about some of the methods in environmental epidemiology.
- Work in environmental health (or adjacent fields).
- Have worked in coding but not necessarily in these topics.
- Would like some tools and know-how to get started in an approachable and friendly setting.
- People who would like to have some fun while learning!
- Some equations will be there but mainly for reference.
- Lots of code though throughout the three days…

R

- R is an interactive environment developed by statisticians for data analysis.
- A more detailed Introduction to R can be found at https://www.r-project.org.
- R is the environment we will use throughout the workshop.
- But this isn't a course to learn R…
- There will lots of code, though we will restrict the equations as much as possible.

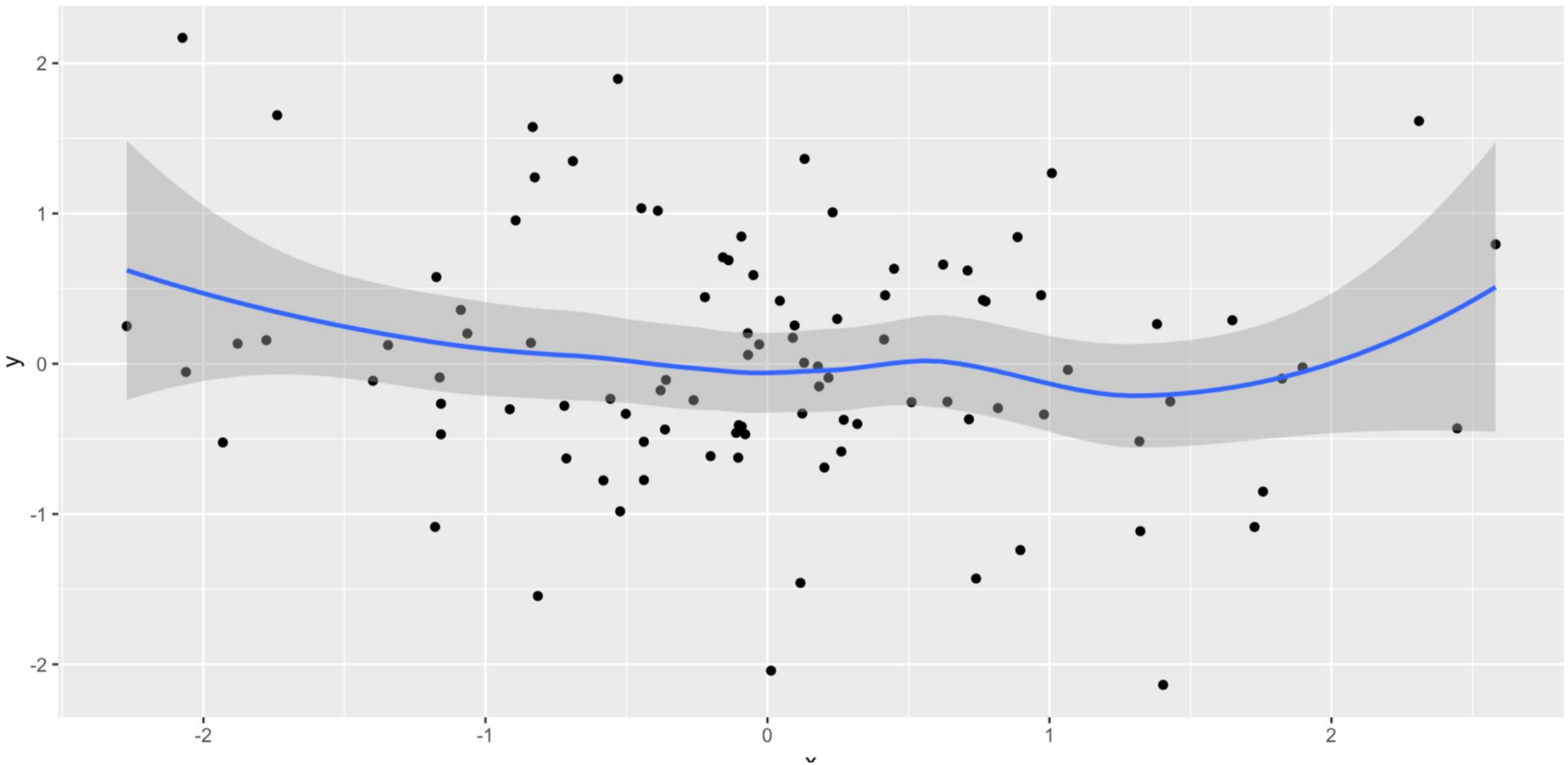
R

- (Sample) R code:

```r
# Load packages
library(ggplot2)

# Create a dataset
set.seed(100)
data = data.frame(x=rnorm(100),y=rnorm(100))

# Plot and rough fit
p <- ggplot(data, aes(x, y)) +
  geom_point() +
  geom_smooth(method = "loess")

plot(p)
```

R

- Output of code:



Rstudio by Posit

- RStudio is an integrated development environment (IDE) for R.
- RStudio provides a convenient graphical interface to R, making it more user-friendly, and providing many useful features.
- Such features include direct code execution, tools for plotting, history, debugging and workspace management.
- A more detailed background on to RStudio can be found at https://posit.co/about/.

Rstudio by Posit

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

Flights - RStudio

Go to file/function | Addins | Flights

flights-example.R

Source on Save | Run | Source

```r
library(nycflights13)     ## package containing flights dataset
library(lubridate)
library(dplyr)
library(ggplot2)

head(flights, n = 3)
daily <- flights %>%
  mutate(date = make_date(year, month, day)) %>%
  count(date) %>%
  mutate(wday = wday(date, label = TRUE))
head(daily, n = 3)
ggplot(daily, aes(wday, n)) +
  geom_boxplot(outlier.colour = "hotpink") +
  labs(x = "Weekday", y = "Flights",
       subtitle = "Number of 2013 New York Flights Each Weekday")
```

1:1 (Top Level) R Script

Environment | History | Connections | Tutorial

Import Dataset | 346 MiB | List

R | Global Environment

Data

daily 365 obs. of 3 variables

```
$ date: Date[1:365], format: "2013-01-01" "2013-01-02" …
$ n   : int [1:365] 842 943 914 915 720 832 933 899 902…
$ wday: Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 3 …
```

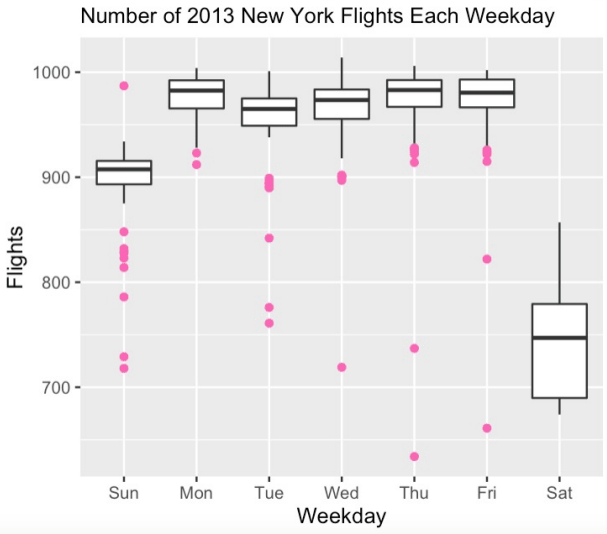
Console | Terminal | Jobs

~/Documents/Flights/

```
# A tibble: 3 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>
1  2013     1     1      517            515         2      830            819        11 UA
2  2013     1     1      533            529         4      850            830        20 UA
3  2013     1     1      542            540         2      923            850        33 AA
# … with 9 more variables: flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
> daily <- flights %>%
+   mutate(date = make_date(year, month, day)) %>%
+   count(date) %>%
+   mutate(wday = wday(date, label = TRUE))
> head(daily, n = 3)
# A tibble: 3 x 3
  date           n wday
  <date>     <int> <ord>
1 2013-01-01   842 Tue
2 2013-01-02   943 Wed
3 2013-01-03   914 Thu
> ggplot(daily, aes(wday, n)) +
+   geom_boxplot(outlier.colour = "hotpink") +
+   labs(x = "Weekday", y = "Flights",
+        subtitle = "Number of 2013 New York Flights Each Weekday")
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export



Posit (RStudio) Cloud

- RStudio Cloud is a cloud-based RStudio which runs projects and code in the cloud.
- RStudio Cloud allows convenient scaling and sharing of code, including for training programs such as SHARP.
- Registration for free is available via https://posit.cloud/.
- We will go through how to navigate Rstudio Cloud together.
- More details are available via https://docs.posit.co/cloud/.

Posit (RStudio) Cloud

- Time to register on
  - https://posit.cloud/.

- Then we will sync GitHub project to your computer

Outline

- What is environmental epidemiology?
- Historical milestones
- Major environmental exposures
- Study designs in environmental epidemiology
- Linear regression
- Theories and assumptions
- Key concepts
- Challenges and limitations
- Data structuring

What is environmental epidemiology?

- Study of environmental exposures and health outcomes.
- Identifies preventable environmental risks.
- Informs regulation and policy.

Historical milestones

- John Snow & cholera: Waterborne disease.
- Minamata disease: Mercury poisoning.
- London Smog (1952): Air pollution mortality.
- Lead and IQ: Neurodevelopment in children.

Major environmental exposures

- Air pollution ($PM_{2.5}$, NO_x, O_3).
- Water pollutants (arsenic, lead, microbes).
- Soil & food (pesticides, heavy metals).
- Chemical exposures (BPA, PFAS, endocrine disruptors).
- Radiation (UV, radon).

Smogs

- London smog occurred when particles swell in size under high relative humidity.
- Some particles serve as nuclei on which fog droplets form.
- SO_2 dissolves in fog droplets where it is oxidized to form sulfuric acid.
- Many people died in aftermath of severe London fog episode (the Great Smog) in December 1952, due to a cold air inversion which trapped the smog.
- After the Great Smog laws passed to stop use of coal on open fires for domestic heating and emissions of black smoke.

Smogs

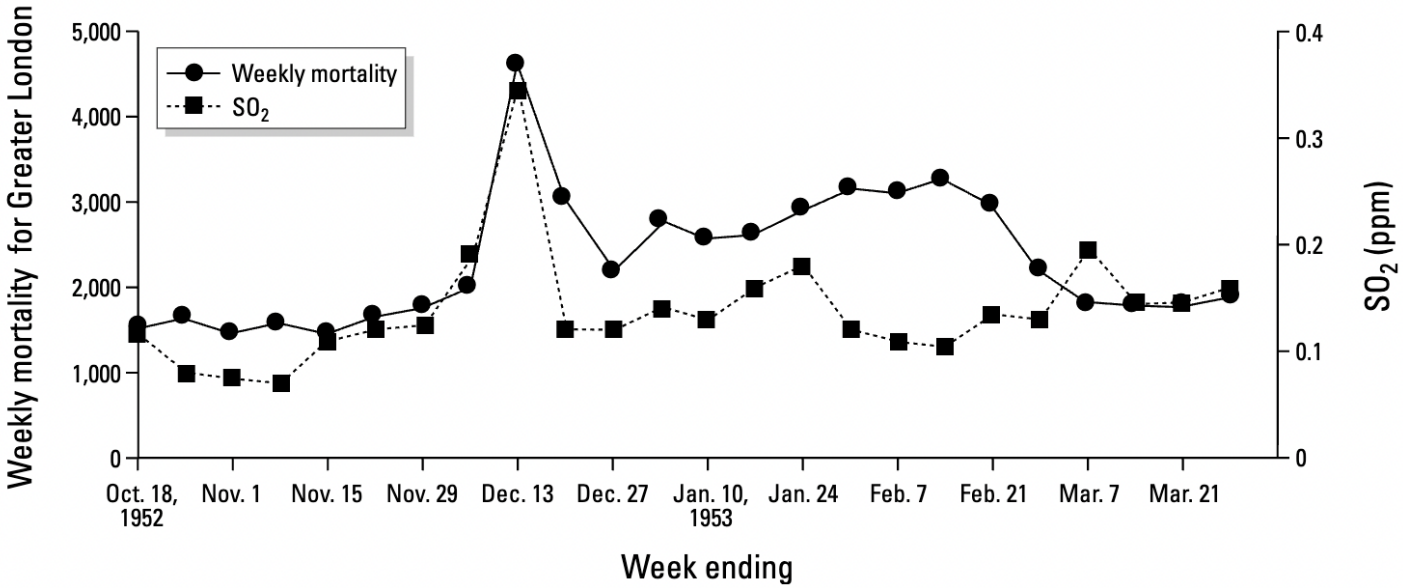


Figure 1. Approximate weekly mortality and SO_2 concentrations for Greater London, 1952–1953.

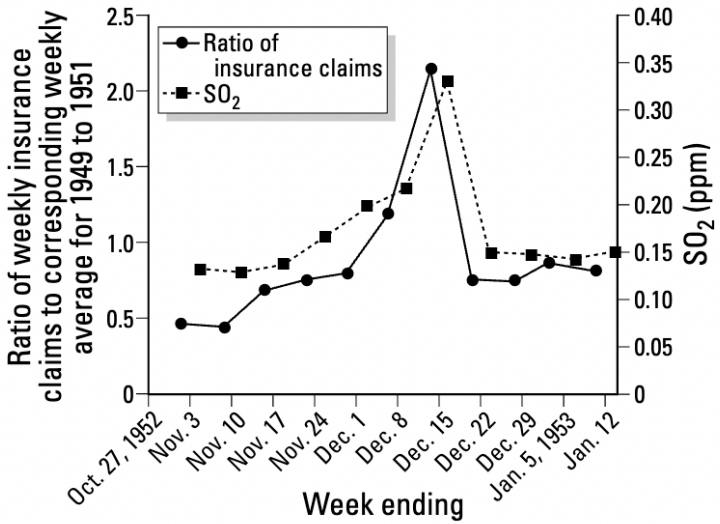


Figure 2. Concentrations and ratios of insurance claims for new sickness in London and Middlesex to corresponding weekly averages for 1949–1951.

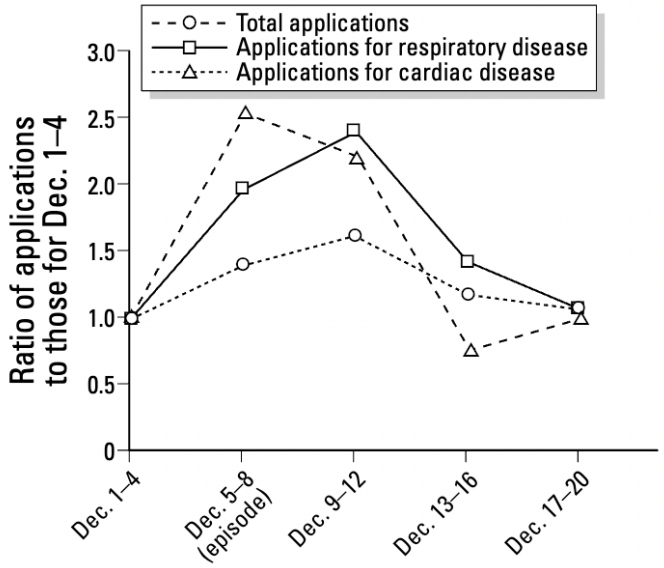


Figure 3. Ratio of applications for emergency bed service in Greater London to those from 1–4 December 1952.

Bell et al., EHP

Smogs

Getty Images

Smogs

Getty Images

Smogs

- Dominance of combustion-based cars produces high concentrations of NO (TRAP), which can be converted to NO_2 via:

$$2NO + O_2 \rightarrow 2NO_2 \qquad (5.36)$$

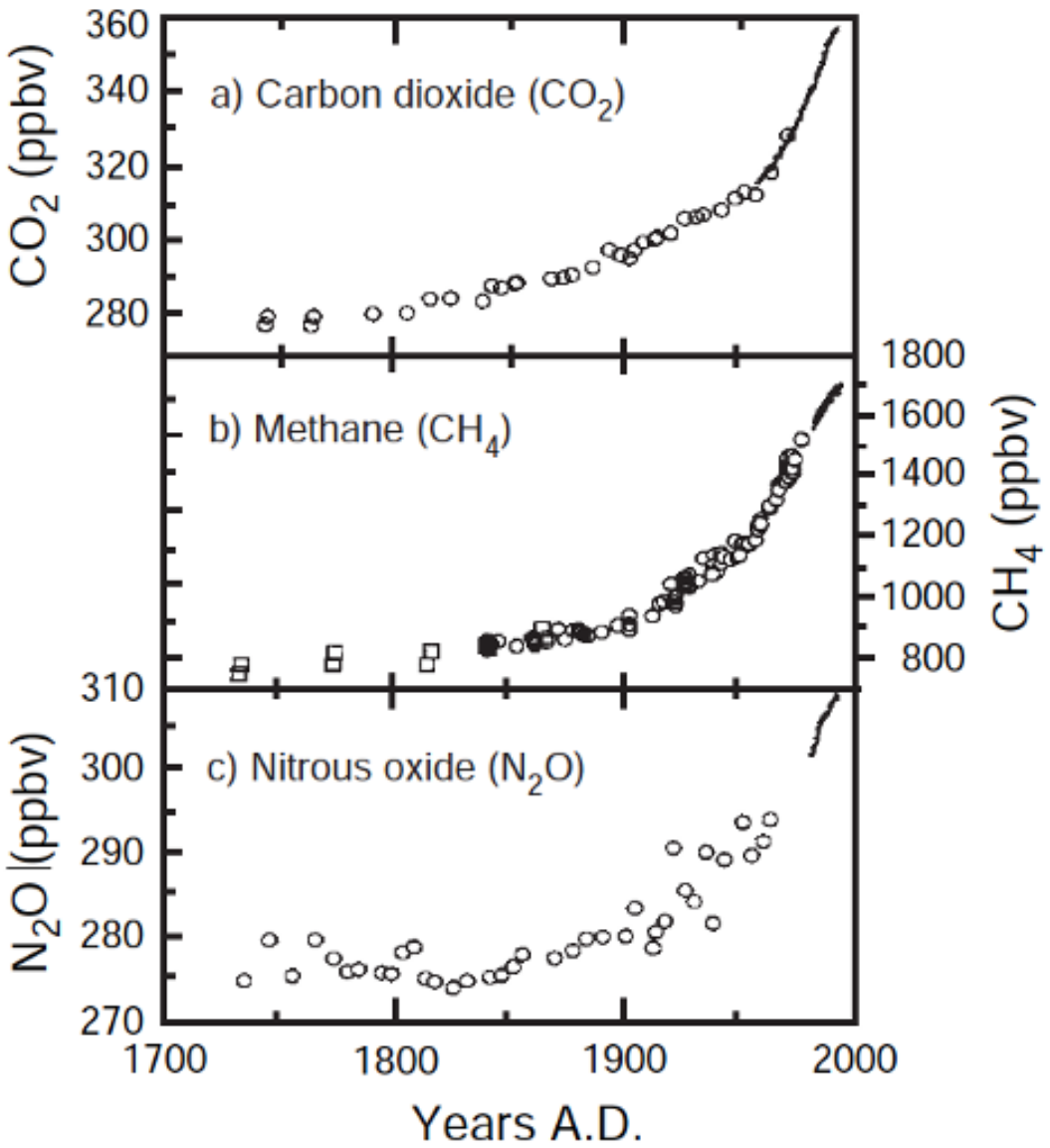
- Rate of production of NO_2 higher in colder temperatures.
- Large problem in cities with cold winters.
- NO_2 can then be involved in production of O_3.

Smogs

- Los Angeles smog originates from car emissions in combination with sunlight and stagnant meteorological conditions.
- Characterized by high concentrations of a large variety of pollutants such as nitrogen oxides, O_3, CO, hydrocarbons, aldehydes, sulfuric acid.
- Results in combination of reactions summarized by, .e.g,:

$$CH_4 + 2O_2 + 2NO \rightarrow H_2O + 2NO_2 + HCHO \tag{5.41}$$

Regional and global pollution



Wallace and Hobbs

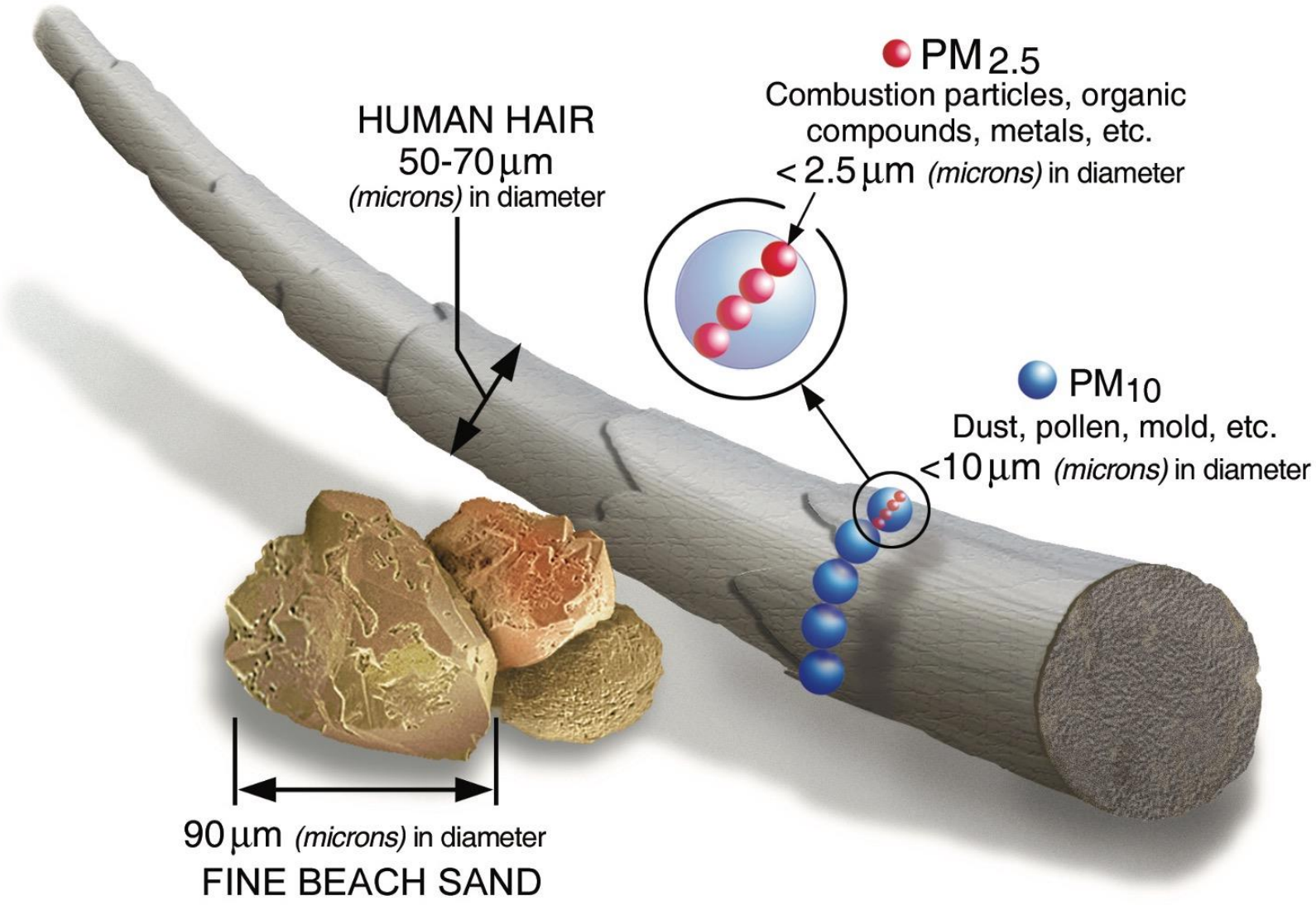
Regional and global pollution

- Another way to describe air pollution is via its diameter.
- Very common measure is particulate matter of < 2.5 μg, $PM_{2.5}$.
- Many studies of this on health outcomes.

Regional and global pollution

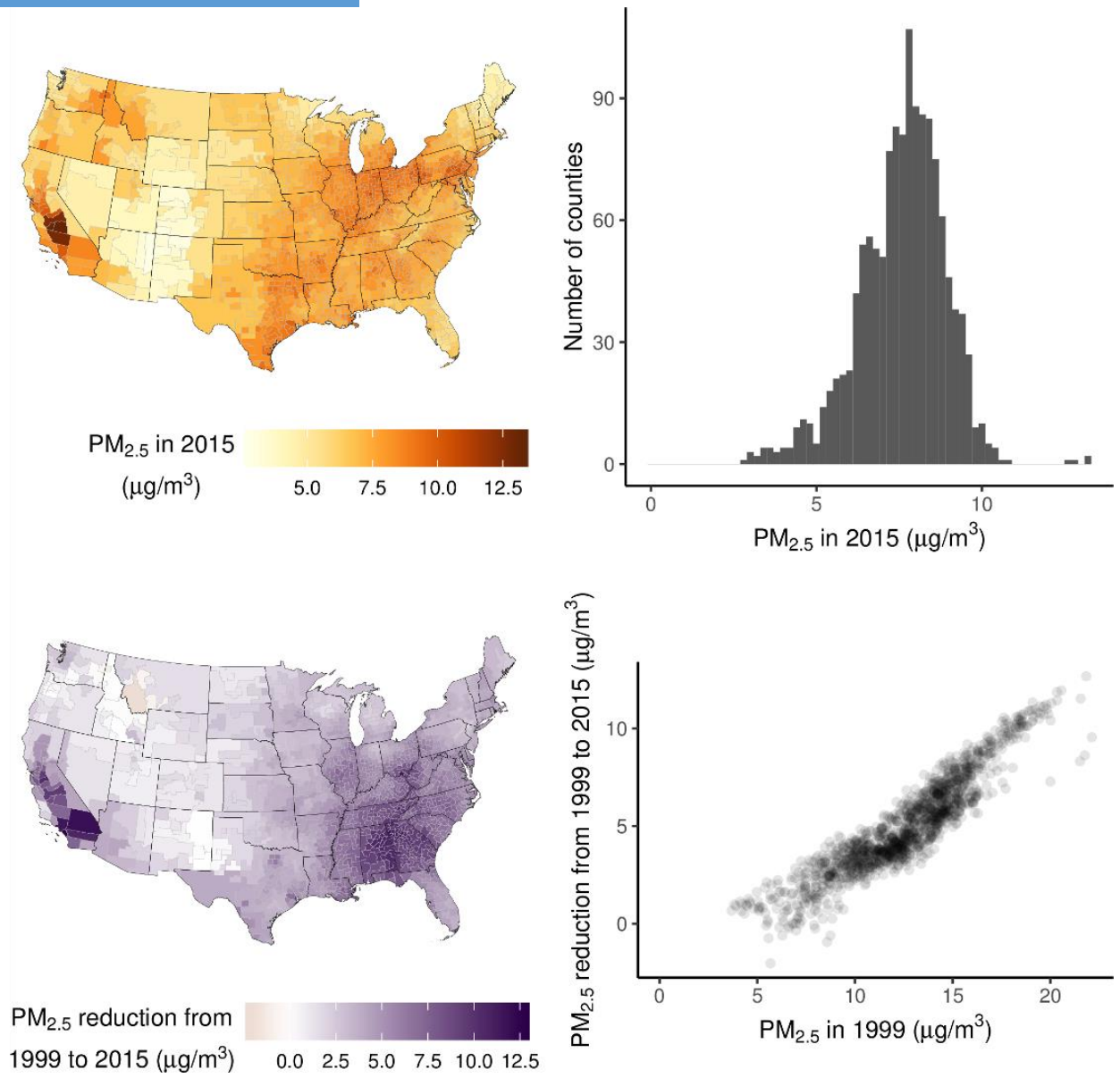
- Air pollution has been consistently associated with both acute and chronic adverse health:
  - All-cause mortality.
  - Cardiovascular mortality and morbidity.
  - Respiratory mortality and morbidity.
  - Cerebrovascular events.
  - Lung cancer.
  - Neurodevelopmental disorders.
  - Neurodegeneration.
  - Pregnancy and reproductive outcomes.
  - Etc…

Regional and global pollution



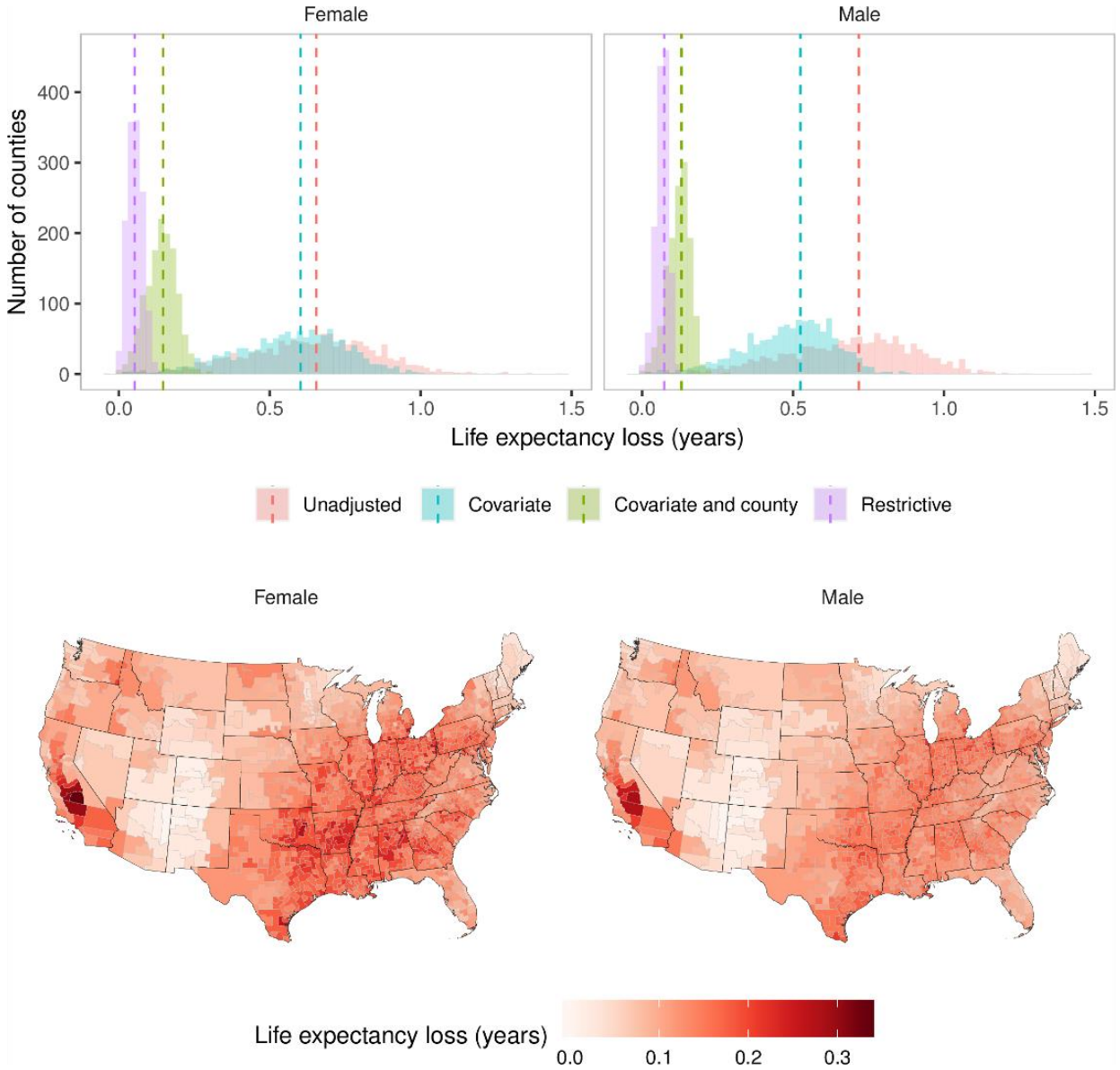
EPA

Regional and global pollution



Bennett et al., PLOS Medicine

Regional and global pollution



Bennett et al., PLOS Medicine

Climate-related Exposures and Health

Missing or understudied connections?

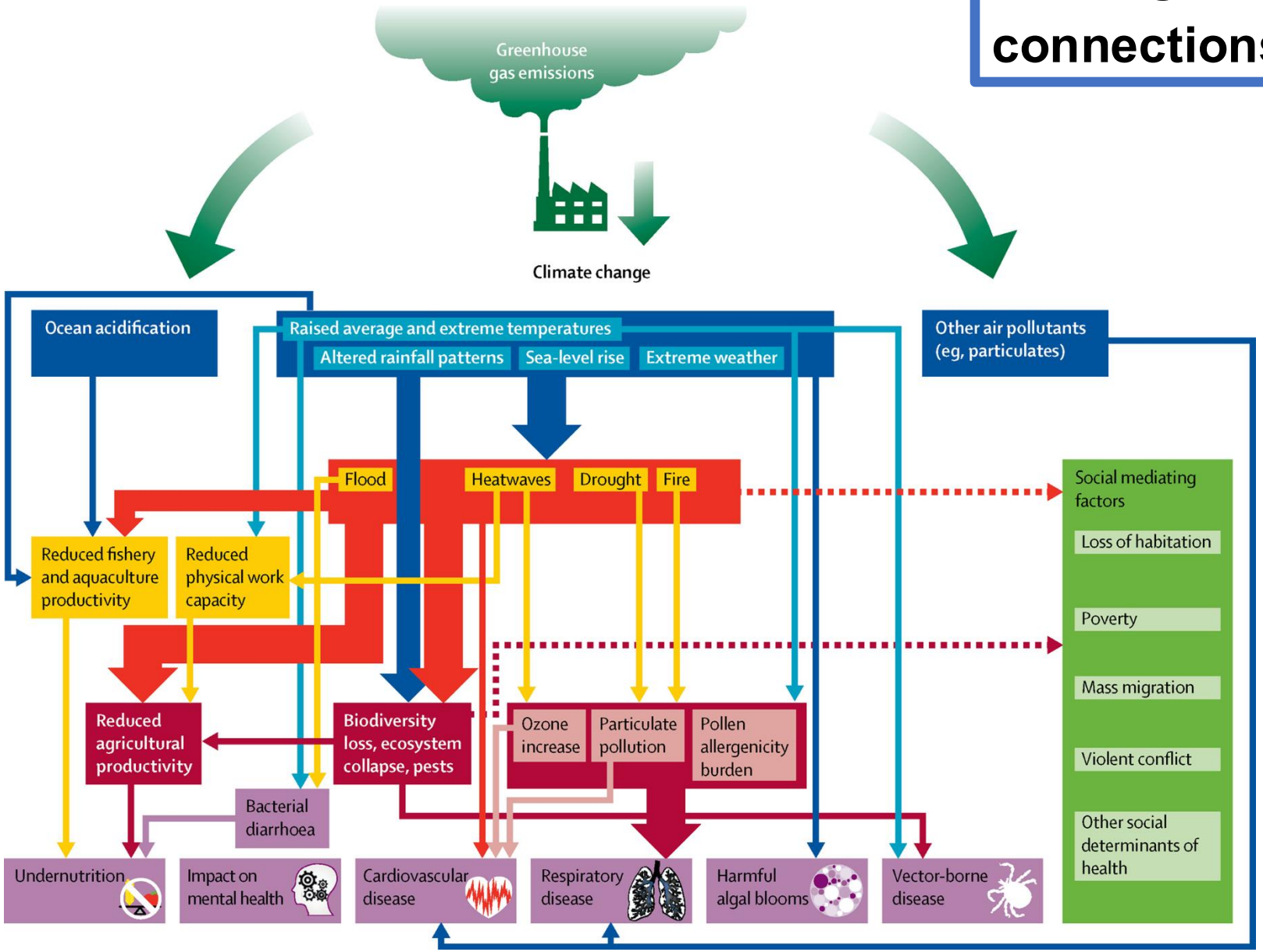
Emissions and pollutants

Climate-related phenomena

Climate-related exposures

Societally-relevant phenomena

Health outcomes



Mediating and modifying factors

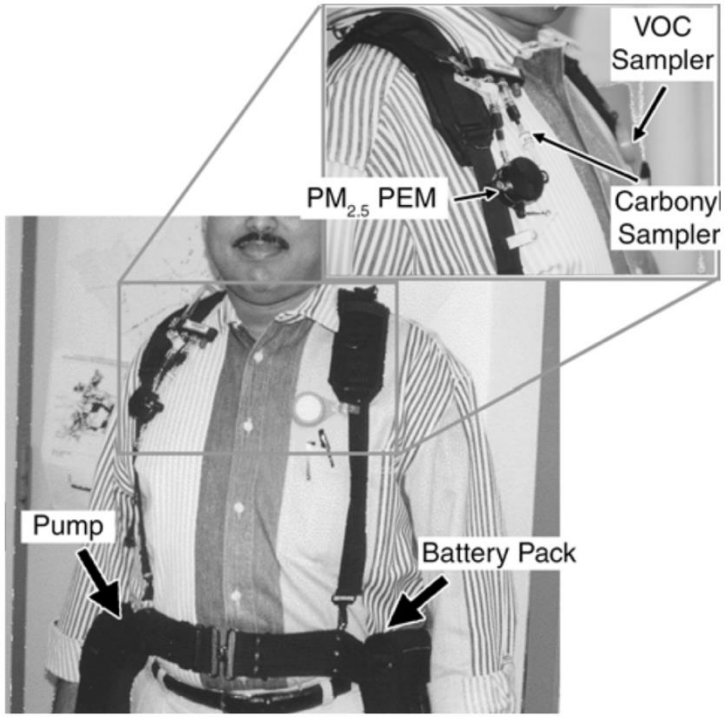
Lancet Countdown 2018

Quantitative study designs in environmental epidemiology

- Cross-sectional.
- Case-control.
- Cohort (prospective/retrospective).
- Time-series and case-crossover.

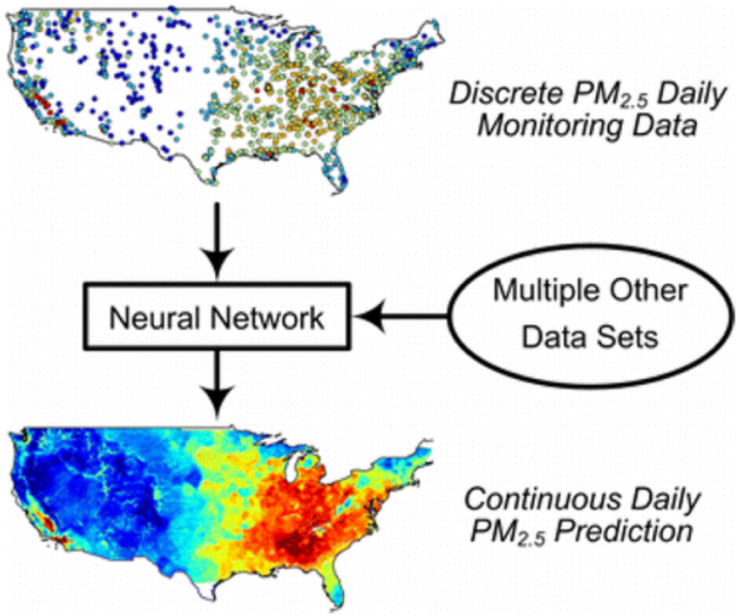
Air pollution exposure assessment: Ideally…

- Air pollution associated with numerous adverse health outcomes
- Ideally: Personal exposures
  - Expensive and intrusive
  - Not feasible for long time periods with many subjects



Air pollution exposure assessment: More recently…

- "Fancier" spatio-temporal prediction models
- Fuse remote sensing data, predictions from chemical transport models, met and GIS variables
- Using more robust methods that yield higher predictive accuracy, e.g. random forests and neural networks
- Very high spatial and temporal resolution (i.e., daily predictions at 1×1 km^2 grids)
- Increasing importance of computational scalability



Exposure challenges

- Exposure misclassification.
- Indoor vs. outdoor environments.
- Time-activity patterns.
- Co-exposures and mixtures.
- Lack of historical data.

Linear regression (a very quick overview)

- Looks at the change in one variable (outcome of interest; Y) that corresponds to a given change in the other variable (exposure of interest; X).
- Regression line:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

  - Where i here is the unit of observation (e.g. study participant)
- β_0 is the intercept - How do you interpret this?
  - $E[Y|X = 0] = \beta_0$
  - i.e., expected value of Y when $X = 0$.

Linear regression (a very quick overview)

- Looks at the change in one variable (outcome of interest; Y) that corresponds to a given change in the other variable (exposure of interest; X).
- Regression line:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

  - Where i here is the unit of observation (e.g. study participant)
- β_1 is the slope - How do you interpret this?
  - The change in Y, on average, that corresponds to a one unit increase in X.

Linear regression (a very quick overview)

- Looks at the change in one variable (outcome of interest; Y) that corresponds to a given change in the other variable (exposure of interest; X).
- Regression line:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \widehat{Y}_i + \varepsilon_i$$

  - Where i here is the unit of observation (e.g. study participant)
- ε_i are the errors (aka residuals): $Y_i - \widehat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i)$
  - Get the β_0 and β_1 that minimize $\sum_{i=1}^{N} \widehat{\varepsilon_i}^2$

Linear regression (a very quick overview)

- R Code:

5: Fit Regression Model

```
mod <- lm(AveBMI ~ AvePM + PerBlack + PerLatinx + PerAsianAm + MedHInc +
            MedHVal + LTHS + FemaleUnemp + MaleUnemp + ClimateRegion,
          data = dta, na.action = na.omit)
```

Linear regression (a very quick overview)

- R Code:

6: Review Model Results

```r
summary(mod)
```

```
Call:
lm(formula = AveBMI ~ AvePM + PerBlack + PerLatinx + PerAsianAm +
    MedHInc + MedHVal + LTHS + FemaleUnemp + MaleUnemp + ClimateRegion,
    data = dta, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9089 -0.6638 -0.1131  0.5264 12.2713

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.595e+01  7.663e-01  33.864  < 2e-16 ***
AvePM                      1.303e-01  4.992e-02   2.611  0.00936 **
```

Theories and assumptions

1. <u>Independence</u>: Y_i's are independent given X_i's.
2. <u>Linearity</u>: mean value of Y is a linear function of X.
3. <u>Normality</u>: $\varepsilon \sim N(0, \sigma^2)$.
4. <u>Homoscedasticity</u>: constant variance of the errors, i.e., σ^2 does not depend on X.

Confounding and effect modification

- Confounding: Third variable distorting association.
- Example: age, smoking, SES.
- Effect modification: Exposure effect differs by subgroup.
- Use stratified analysis or interaction terms.

Data structuring

What is the research question?	What kind of data do we have?	Any other considerations?
• Exposure? • Health outcome? • Basis for studying their links? • Plausibility? • Hypothesis?	• Type of records? • Time resolution? • Spatial resolution? • Similarity between neighbouring units? • Coverage?	• Covariates and potential confounders? • Effect modification? • Linearity/non-linearity? • Computing resources?

Collaboration is (of course) key!

Outline

- What is environmental epidemiology?
- Historical milestones
- Major environmental exposures
- Study designs in environmental epidemiology
- Linear regression
- Theories and assumptions
- Key concepts
- Challenges and limitations
- Data structuring

Getting ready for the lab

- This lab will involve taking some models and concepts from the **Introduction to methods in environmental epidemiology 1** lecture and introduce you to the way linear regression works with an example from environmental epidemiology.

Application

- How can you imagine applying this learning to your data and your research questions?

Questions

- Questions?

Introduction to methods in environmental epidemiology 1

Robbie M Parks, PhD

21st July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH