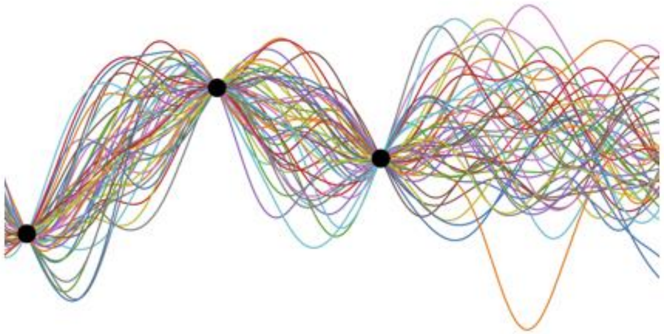
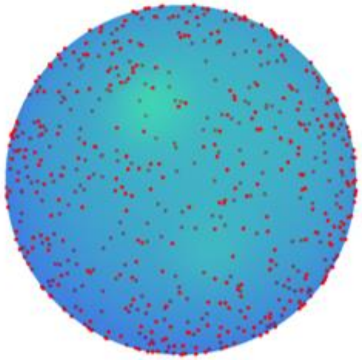
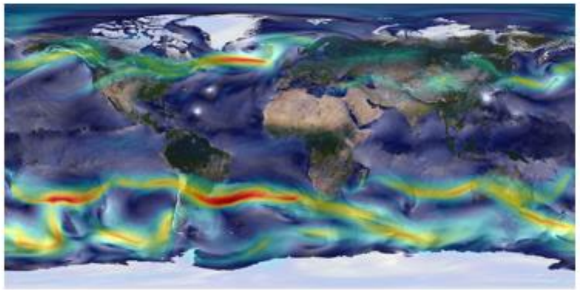
Emission Sector Contributions to Air Quality and Public Health in China from 2010–2050 using Emulators

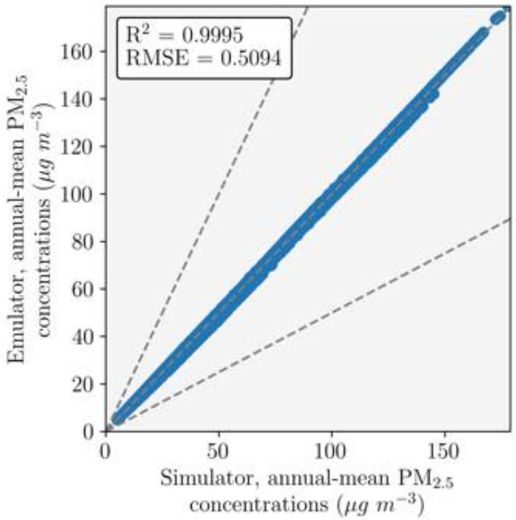
UNIVERSITY OF LEEDS

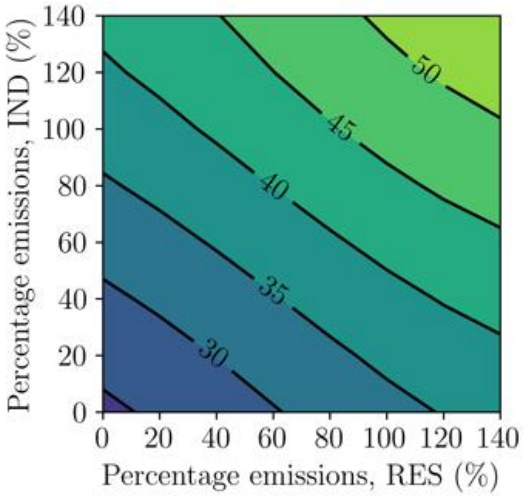
Simulator → Emulator → Sensitivity analysis











Images 1, 2, and 3

Motivation

UNIVERSITY OF LEEDS

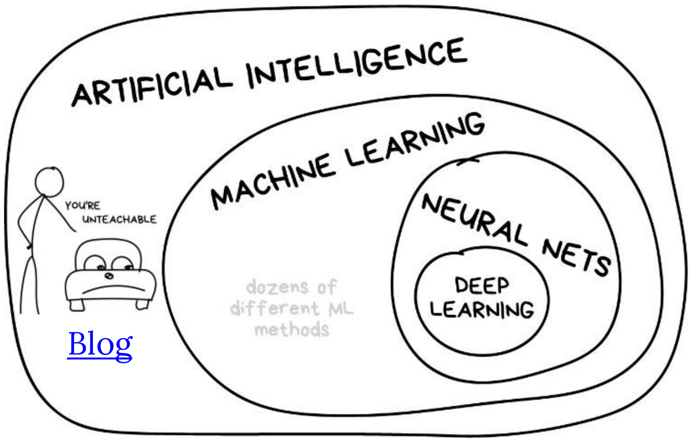
Problem

- Atmospheric models are useful, but slow and expensive.
    - Compromises:
        - Reduce the model accuracy (e.g., reduced complexity, coarser resolution, increase parameterisation).
        - Reduce the model precision (e.g., reduced precision chips).
        - Reduce the number of experiments.
        - Use a bigger computer.

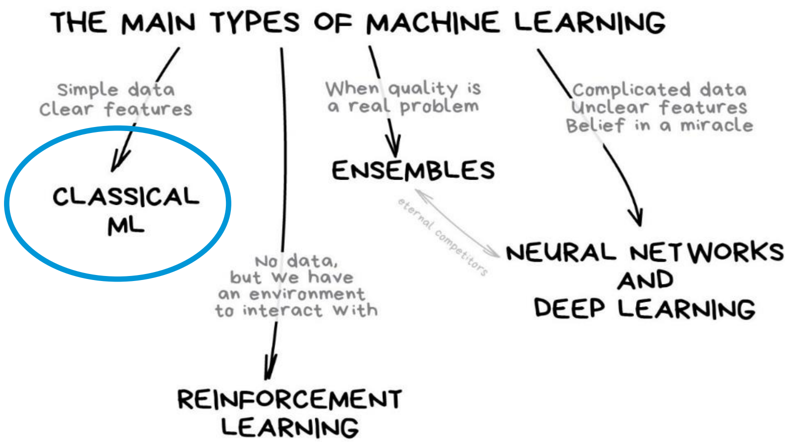
Alternative solution

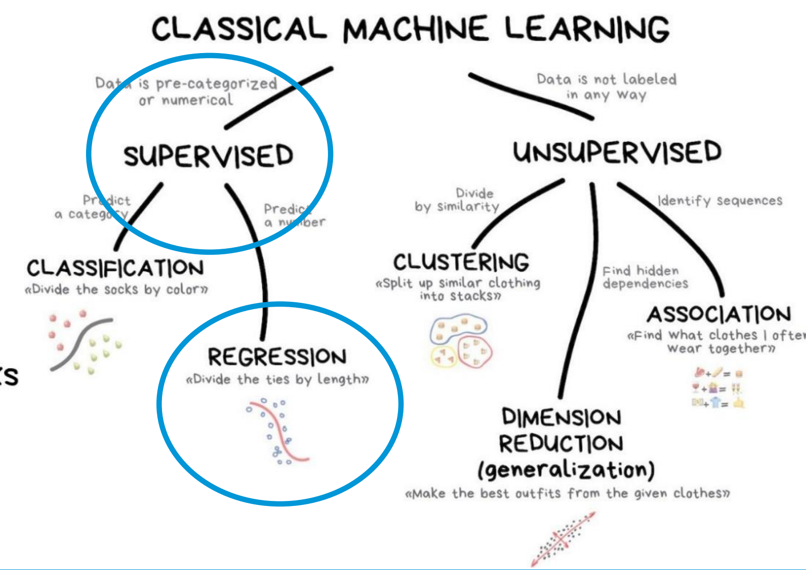
- Emulate these atmospheric models with machine learning models.
    - Trained on simulation data to learn statistical associations between inputs and outputs.
    - Cheaper and faster to run, enabling many more experiments.
- Previous studies have used emulators to:
    - Predict air quality, weather, and climate.
    - Represent processes, such as convection and chemistry.
    - Explore uncertainties and sensitivities.

- Associational (not explanatory) knowledge.
- Useful *tool* for:
  - Prediction problems (patterns recognition, induction).
  - Problems cannot program (e.g., image recognition).
  - Faster approximations to problems that can program (e.g., spam classification).
- Classical.
- Supervised.
- Regression.



[Blog]





Methods

UNIVERSITY OF LEEDS

Problem

- *Identify.*
- *Input(s).*
- *Output(s).*

Simulations

- *Design.*
- *Run.*
- *Evaluate.*

Emulators

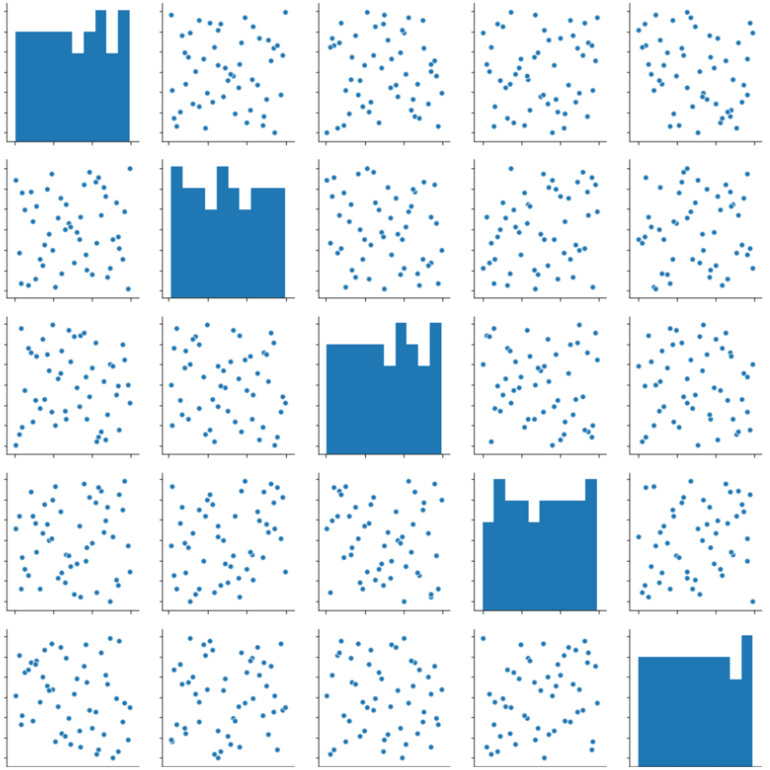
- *Design.*
- *Optimise.*
- *Evaluate.*
- *Predict.*

Problem

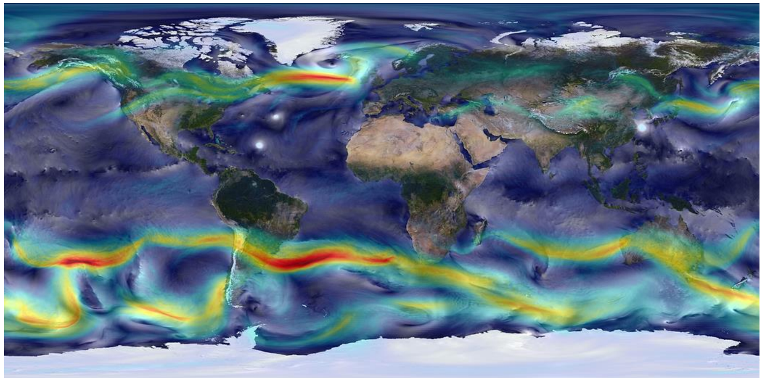
- *Identify*.
  - Predict air quality from emission changes in China.
    - There have been recent reductions in emissions and fine particulate matter ($PM_{2.5}$) concentrations in China.
    - However:
      - $PM_{2.5}$ exposure remains high.
      - Ozone (O_3) exposure is increasing.
      - The associated disease burden is substantial.
        - >10% healthy life lost to disease in 2019 (GBD 2019, 2020).
    - Goal: Explore how emission scenarios can improve health.
- *Input(s)*.
  - 5 key anthropogenic emission sectors.
    - Residential (RES), industrial (IND), land transport (TRA), agriculture (AGR), and power generation (ENE).
- *Output(s)*.
  - $PM_{2.5}$ concentrations.
  - O_3 concentrations.

Simulations

- *Design.*
  - Maxi-min Latin hypercube space-filling designs to select the scalings to apply to the inputs.
    - Near-random sample of parameter values (0–150%) from all input combinations.
    - Independently for both the training and test data.

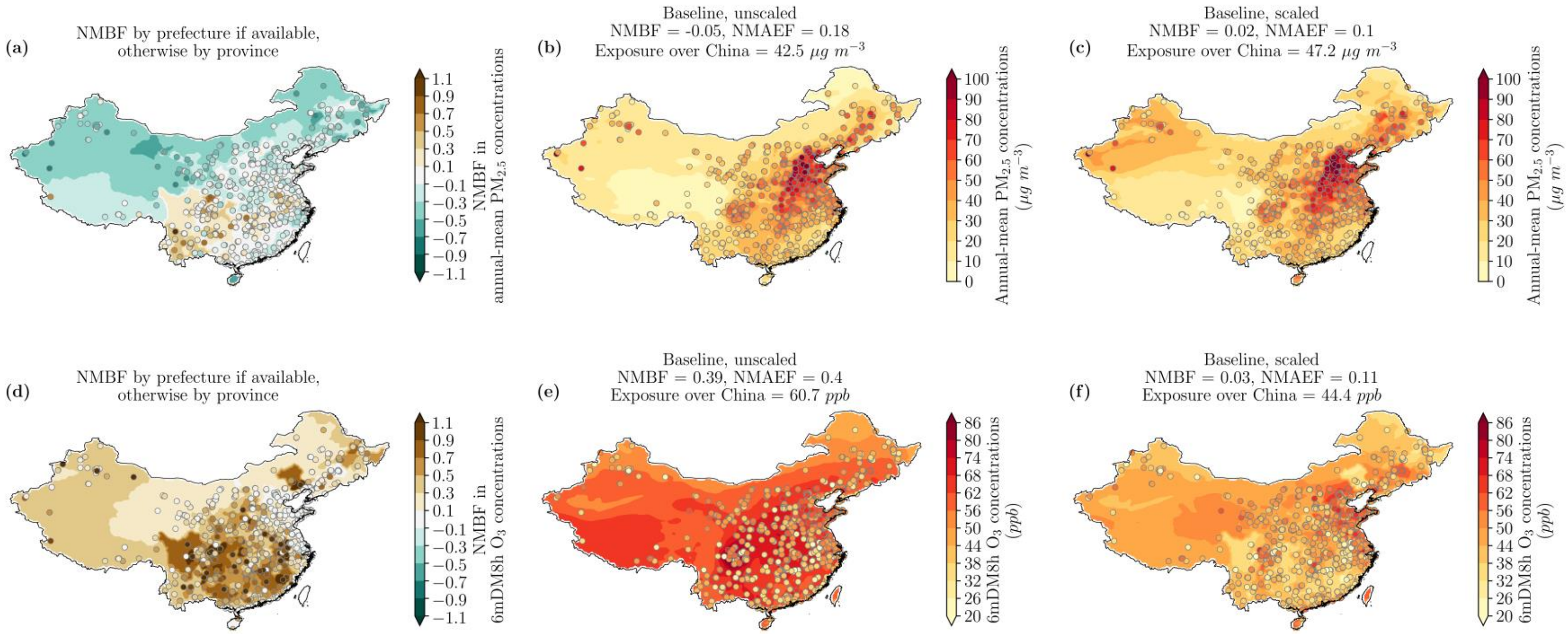


- *Run.*
  - Complex air quality model (WRFChem).
  - 1 year for the control (*evaluate*).
  - 50 years of training simulations.
  - 5 years of testing simulations.



Simulations

- *Evaluate.*
  - Evaluate the control run against measurements to ensure it accurately predicts outputs.
    - Normalised mean bias factor (NMBF): −0.05 ($PM_{2.5}$) and 0.39 (O_3).
  - Tuned to measurements to improve accuracy.
    - NMBF: 0.02 ($PM_{2.5}$) and 0.03 (O_3).

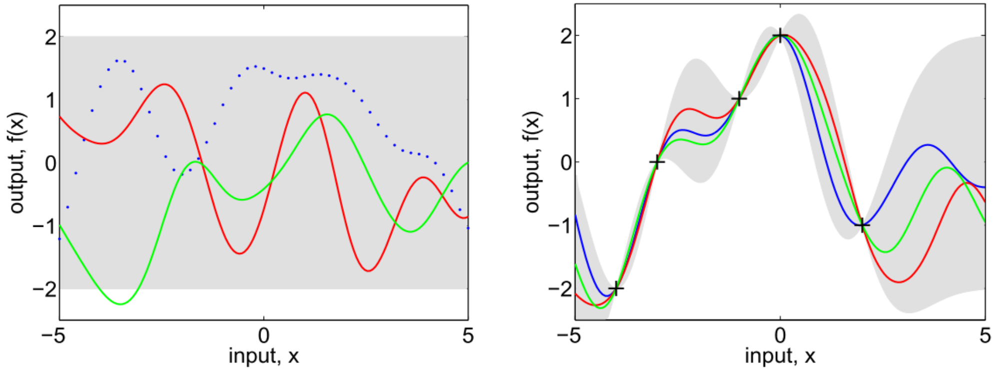


Emulators

- *Design.*
    - Gaussian process regressor.
        - Notice trends well (when similar inputs have similar outputs).
        - Flexible and accurate with smaller data sets.
        - Prior: Output (Gaussian) = function(input)
        - Posterior: Output (Gaussian) = function(input and observations)

Gaussian probability distribution
Bayesian inference

Rasmussen, C. E., & Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*



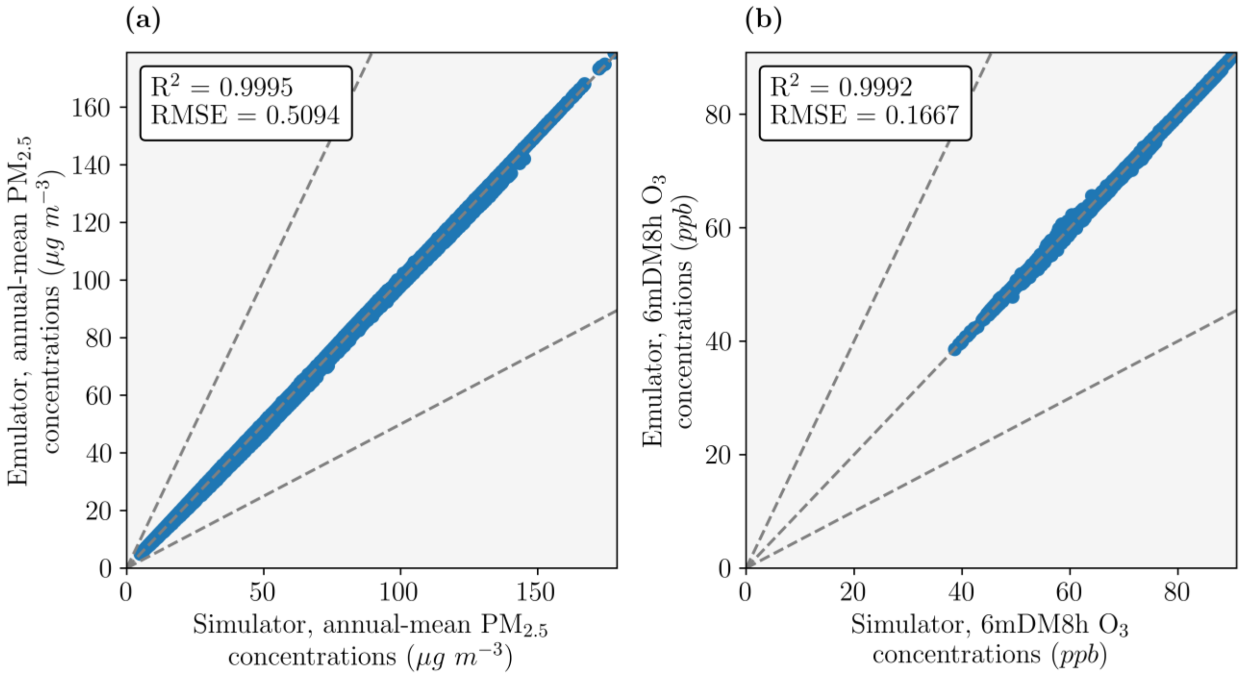
(a), prior (b), posterior

- 1 emulator per grid cell from the simulator (30,556 in total).
- Trained on the 50 years of simulator data.

Emulators

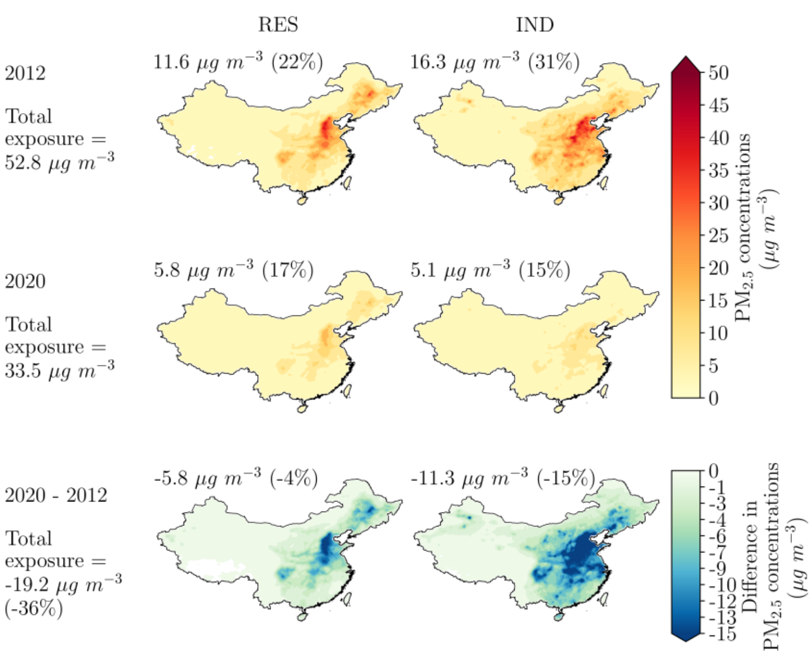
UNIVERSITY OF LEEDS

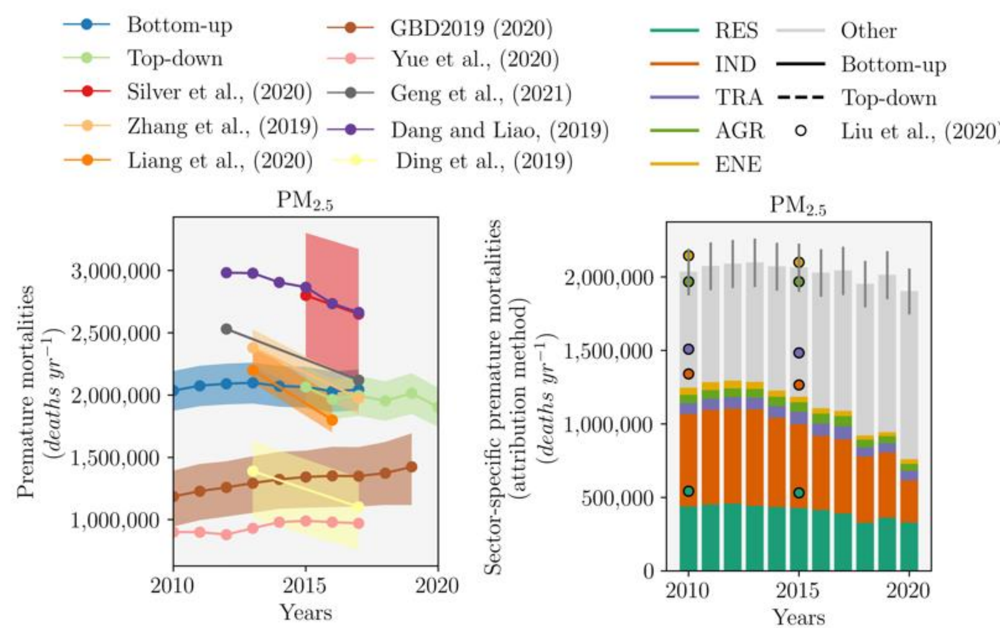
- *Optimise.*
  - Pre–process inputs to make more Gaussian–like.
  - Optimised hyperparameters (for accuracy) using genetic programming (via automated machine learning tool).
- *Evaluate.*
  - Predict unseen test data and evaluate against simulated values.
    - Coefficient of determination (R^2) = 0.999 for both outputs.



Emulators

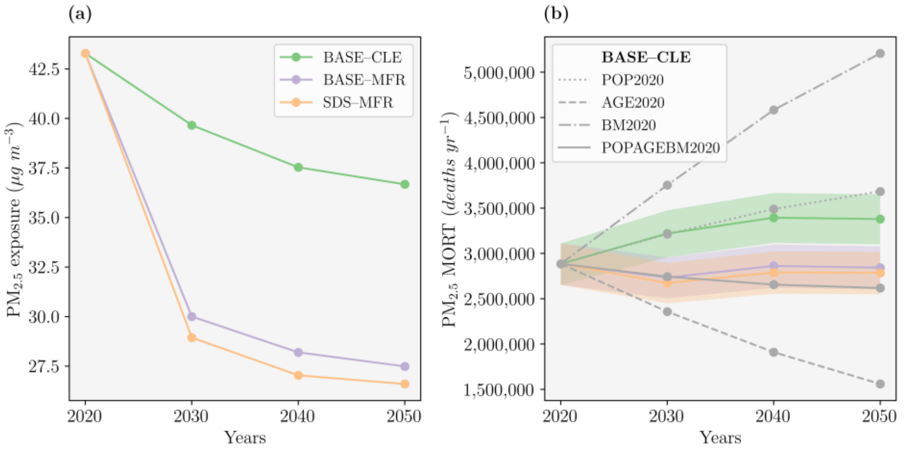
- *Predict.*
  - Results: 2010–2020
    - $PM_{2.5}$ exposure
      - Peaks in 2012, at 52.8 μg m^{-3}.
      - Reduces by 36% in 2020, to 33.5 μg m^{-3}, reaching NAQT.
      - 187,800 (95UI: 179,900–194,200) avoided deaths per year.
      - 58% from industry and 29% from residential emissions.

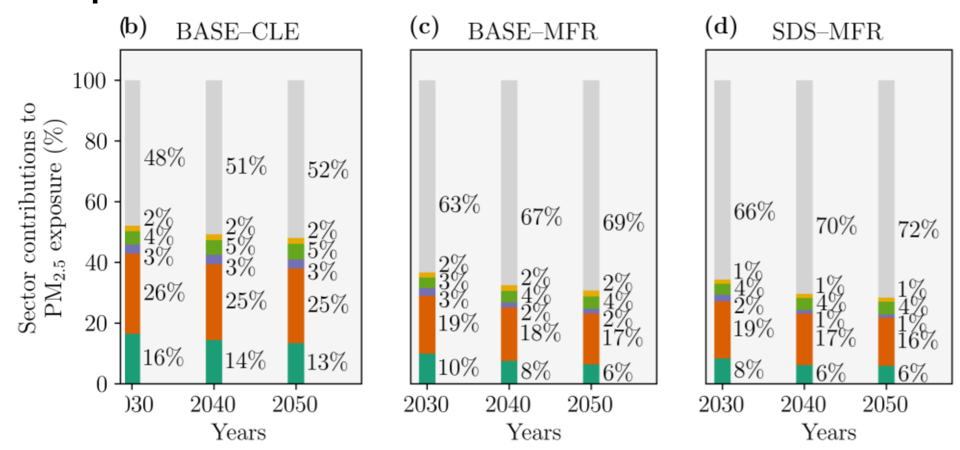




Emulators

- Results: 2020–2050 (Emission scenarios from ECLIPSEv6b)
  - BASE–CLE: baseline climate with current air quality legislation.
  - BASE–MFR: baseline climate with best air pollution technologies.
  - SDS–MFR: sustainable development climate with best AP tech.
  - $PM_{2.5}$ exposure:
    - −15% in BASE–CLE, −36% in BASE–MFR, and −39% in SDS–MFR.
  - $PM_{2.5}$ disease burden:
    - +17% in BASE–CLE, −1% in BASE–MFR, and −3% in SDS–MFR.
    - Population ageing increasing disease susceptibility.
  - Most of benefits by 2030, mainly from reductions in IND and RES.
  - After 2030, other sources more important.





Summary

UNIVERSITY OF LEEDS

Problem

- Atmospheric models have large computational demands, which limit the number of experiments.

One solution

- Emulate these atmospheric models using machine learning models.
  - Predict outputs from specific statistical associations with inputs.
  - Accurate, flexible, fast, and cheap (once setup).

Application

- Predict air quality from emission changes in China.
- Emulators predicted 99.9% of the variance in air quality.
- Large public health benefits by 2030 possible by best air pollution technologies in industry and residential sectors.
- After 2030, other sectors increase in importance.

Further information

- Short–term air quality prediction: Code and Paper.
- Long–term air quality prediction: Code and Papers (in prep.).