SUPERVISED REGRESSION LEARNING FOR PREDICTIONS OF 3-1000 NM AEROSOL PARTICLE SIZE DISTRIBUTIONS FROM PM2.5, TOTAL PARTICLE NUMBER, TRACE GASES AND METEOROLOGICAL PARAMETERS AT HYYTIÄLÄ SMEAR II STATION

Yusheng Wu[1], R. Cai[1], M. Zaidan[1], J. Kuula[2], H. Timonen[2], P. Aalto[1], M. Kulmala[1], K. Lehtipalo[1], T. Petäjä[1], J. Kangasluoma[1]

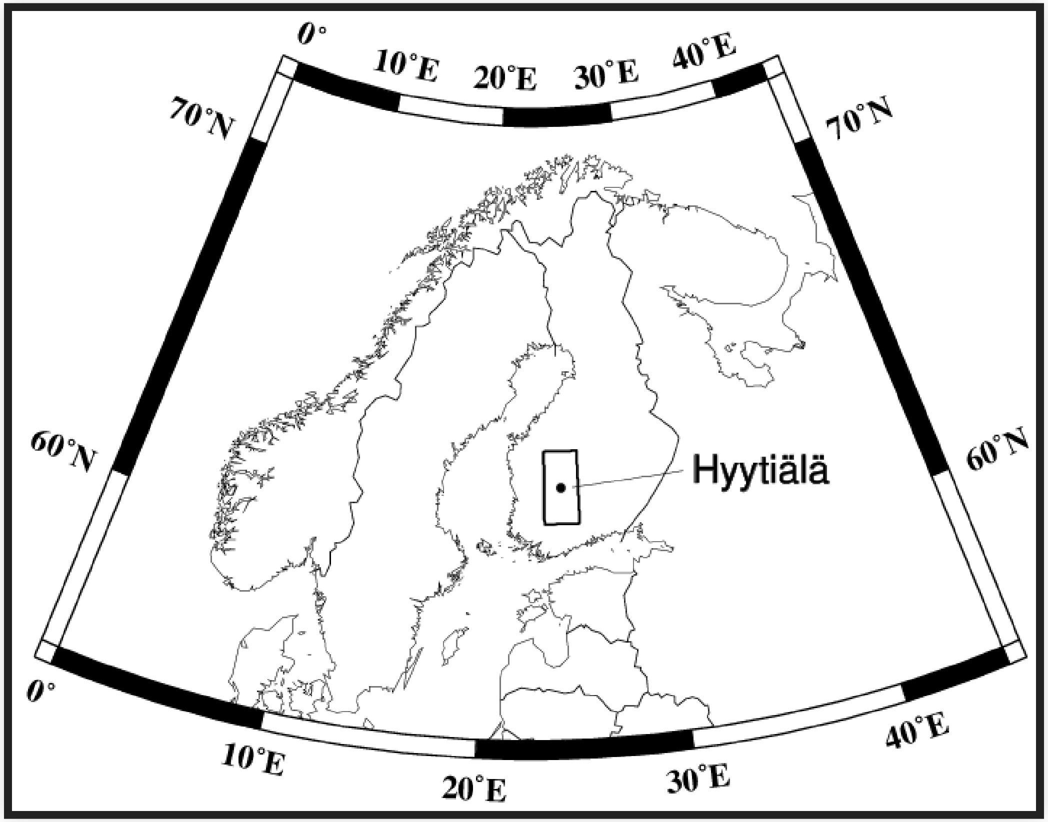
[1] Institute for Atmospheric and Earth System Research / Physics, Faculty of Science, University of Helsinki, Finland
[2] Finnish Meteorological Institute, Erik Palménin aukio 1, 00560 Helsinki, Finland

MOTIVATION

- Health effect of sub-micron particles, especially the ultrafine particles (<100 nm)
- Epidemiological studies are lacking suitable data
- Air quality stations routine measurement: $PM_{2.5}$, meteorological and trace gases data
- The temptation of using machine learning to fill this gap

FIELD MEASUREMENTS

- Hyytiälä Forestry Field Station (SMEAR II)
  - A very comprehensive regional boreal forest environment
  - Famous for new particle formation studies
  - Southen Finland (61°50.845' N, 24°17.686' E, 180 m a.s.l.)



MEASURED INPUT DATA

Catogories	Features	Time resolution
Meteorology	wind, T, RH, P, Rad	1 min
Trace gases	CO, SO_2, NO_x, O_3	5 min
Particle	$PM_{2.5}$, N_{tot}	10 min

MEASURED OUTPUT DATA

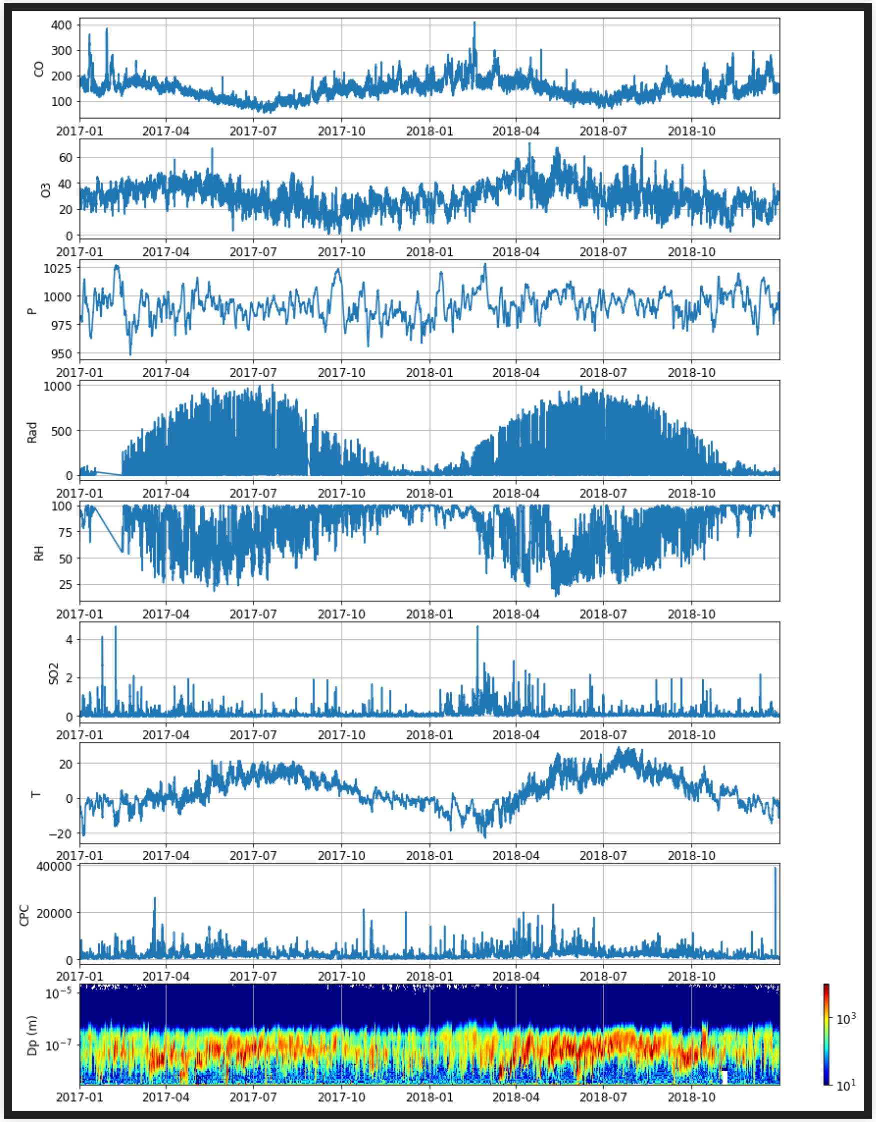
Instruments	Features	Time resolution
DMPS	3 nm ~ 1 μm	10 min
APS	500 nm ~ 20 μm	10 min

OUTPUT FORMAT

Separation point: 650 nm

Combined to 94 bins of normalized number concentration (dN/dlogDp)

data time-series



FEATURE ENGINEERING

- Abstract date-time information: weekend, season
- Encoding: wind direction, hour of day
- Date before the predicting day: memory/delay effect

PERFORMANCE METRICS - R^2_SCORES

- predicted vs observed(test)
- R^2, pronounced "R squared"
  - coefficient of determination
  - Nash-Sutcliffe model efficiency coefficient
- vary from -∞ to 1 for a non-linear regression or a linear regression without including an intercept

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}\left(Q_m^t - Q_o^t\right)^2}{\sum_{t=1}^{T}\left(Q_o^t - \bar{Q}_o\right)^2}$$

PERFORMANCE METRICS - NMAE

- NRMSE: normalized mean-absolute-error
- vary from 0 to ∞

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$\mathrm{NMAE} = \frac{\mathrm{MAE}}{\bar{y}}$$

PERFORMANCE METRICS - NRMSE

- NRMSE: normalized root-mean-square-error
- vary from 0 to ∞

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\bar{y}}$$

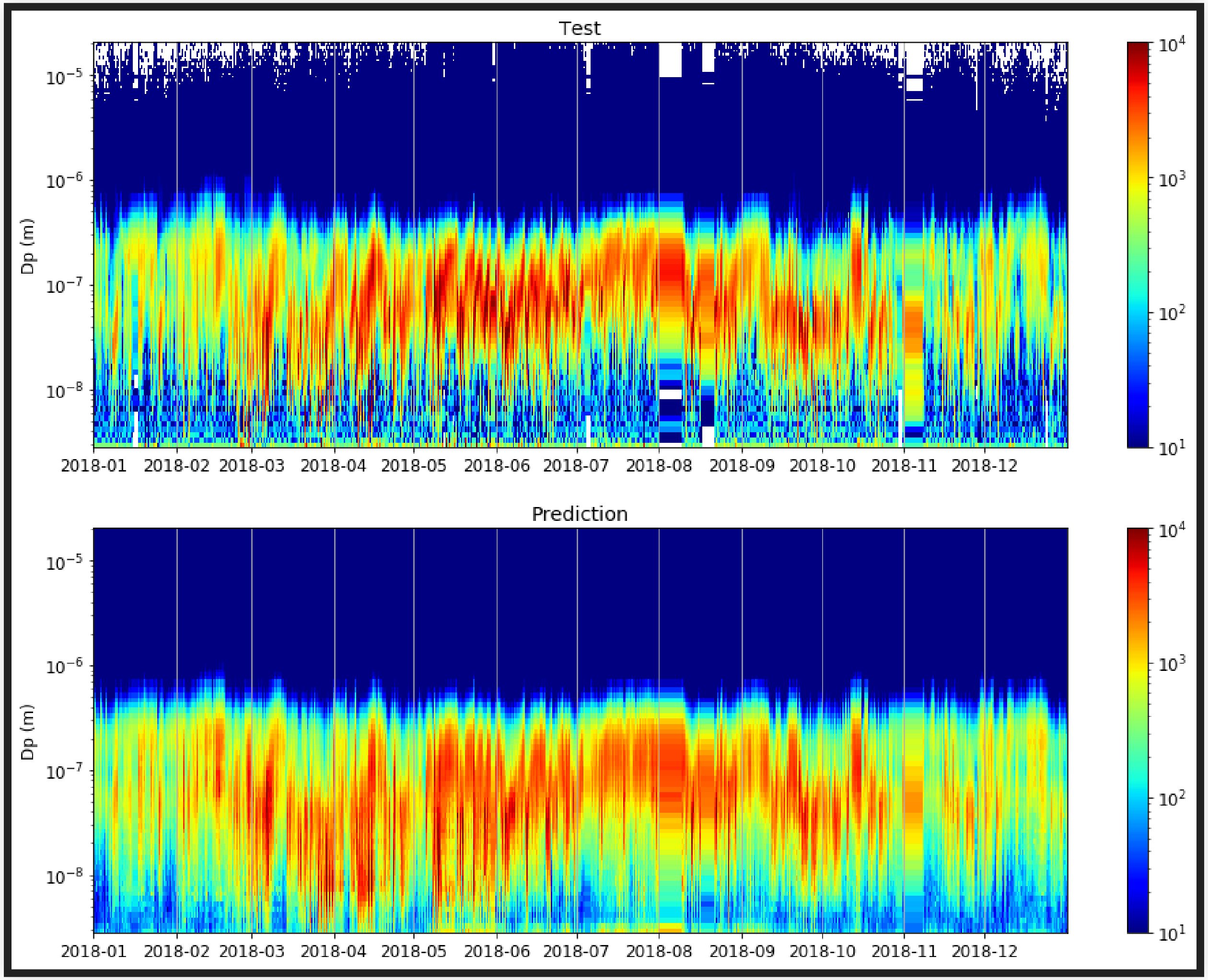
RANDOM FOREST

Hyperparameter tuning: random search

```
param_distributions =
{
        'n_estimators': [10, 30, 100, 300],
        'max_features': ['auto', 'sqrt'],
        'max_depth': [10, 30, 100, None],
        'min_samples_split': [2, 5, 10],
        'min_samples_leaf': [1, 2, 4],
        'bootstrap': [True, False]
}
```

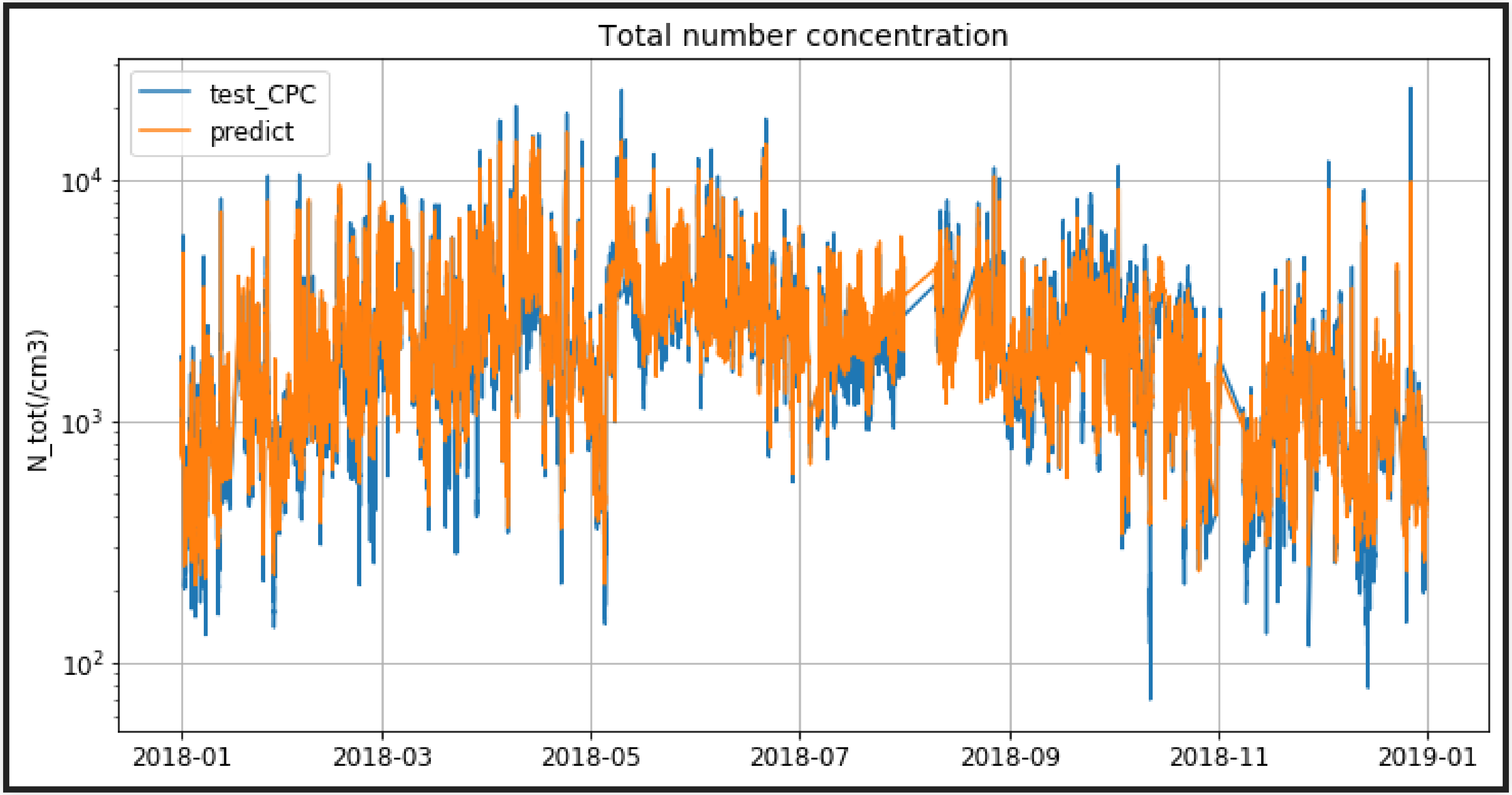
RANDOM FOREST PREDICT TIME SERIES

- Feature number: 9 (raw)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 3, n_iter = 100



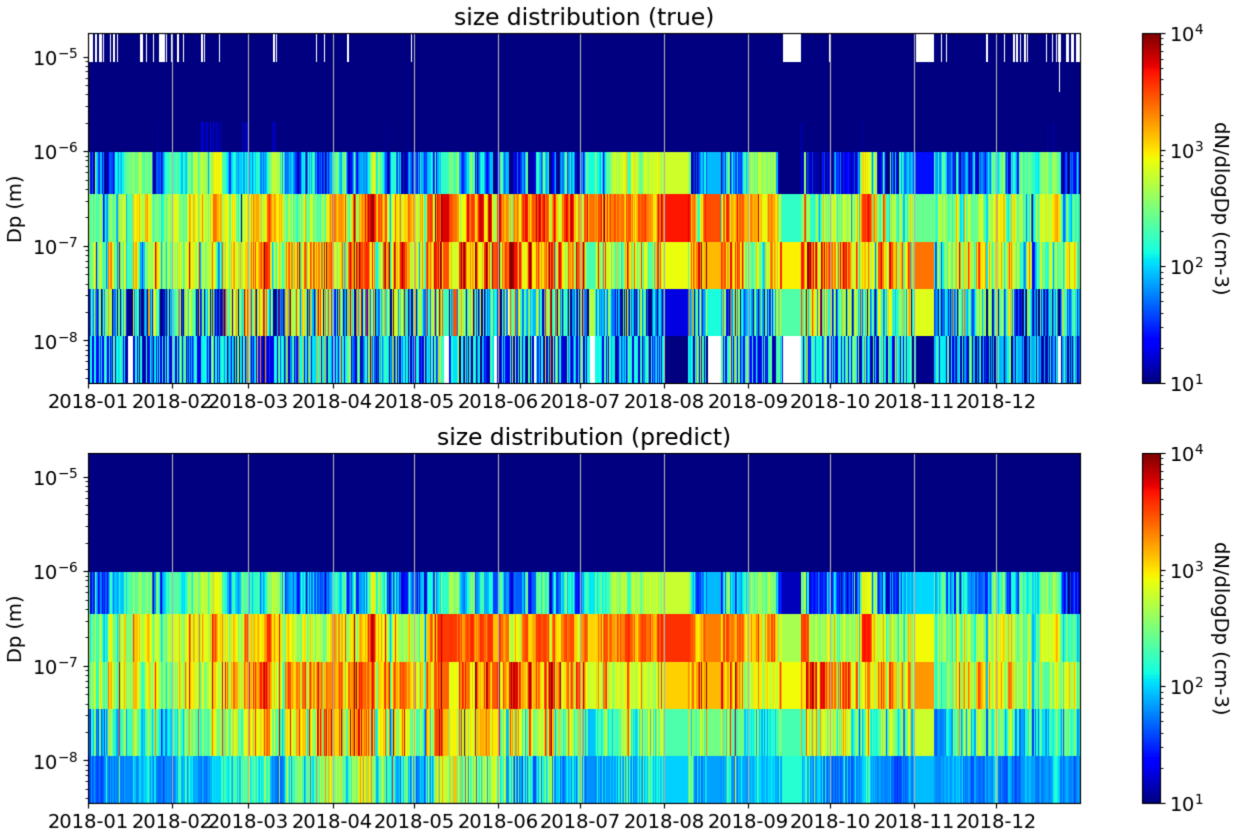
RANDOM FOREST PREDICT N_{TOT}

- Feature number: 9 (raw)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 5, n_iter = 100



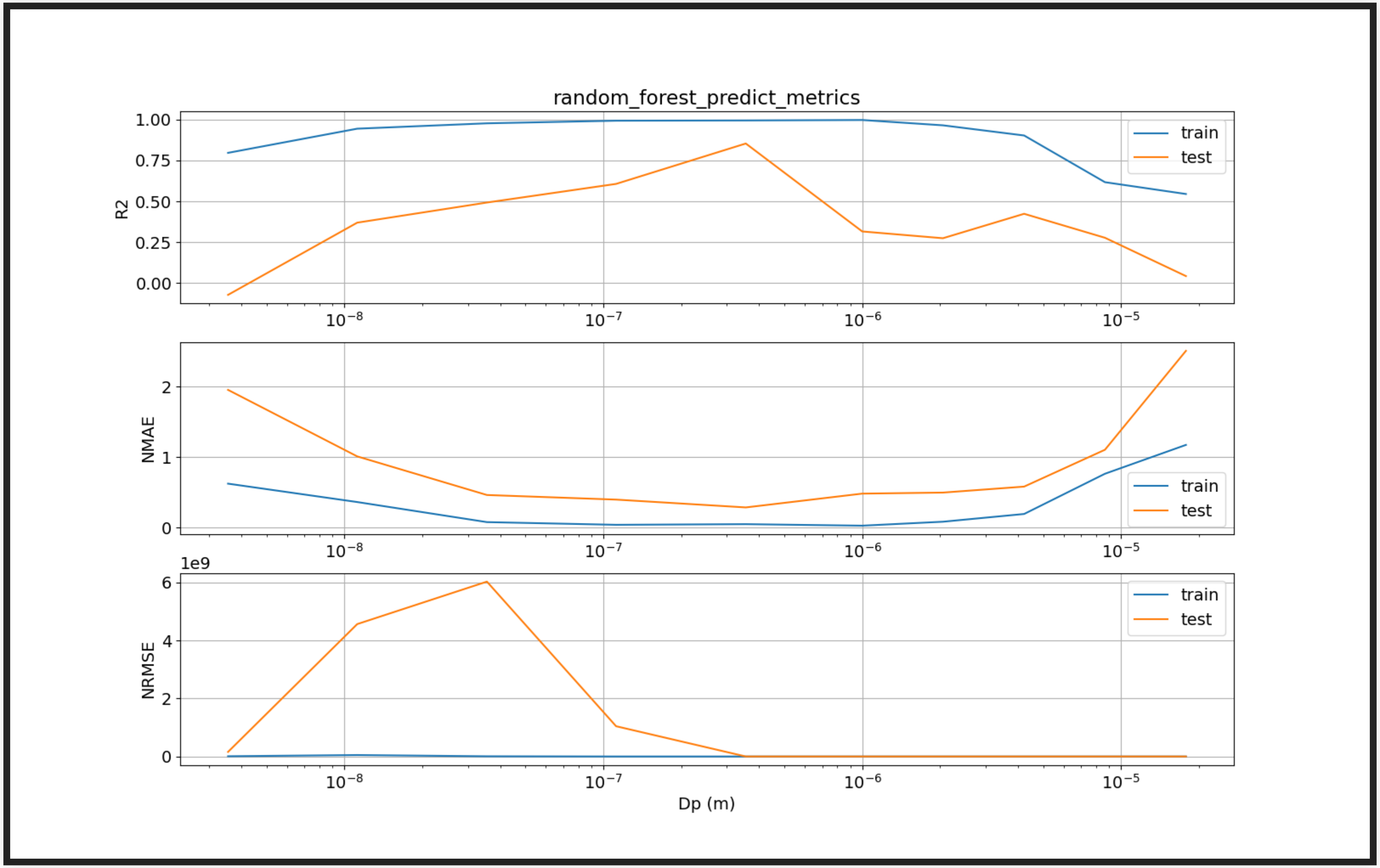
RANDOM FOREST PREDICT TIME SERIES

- Feature number: 12 (raw) -> 46 (derived)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 5, n_iter = 40



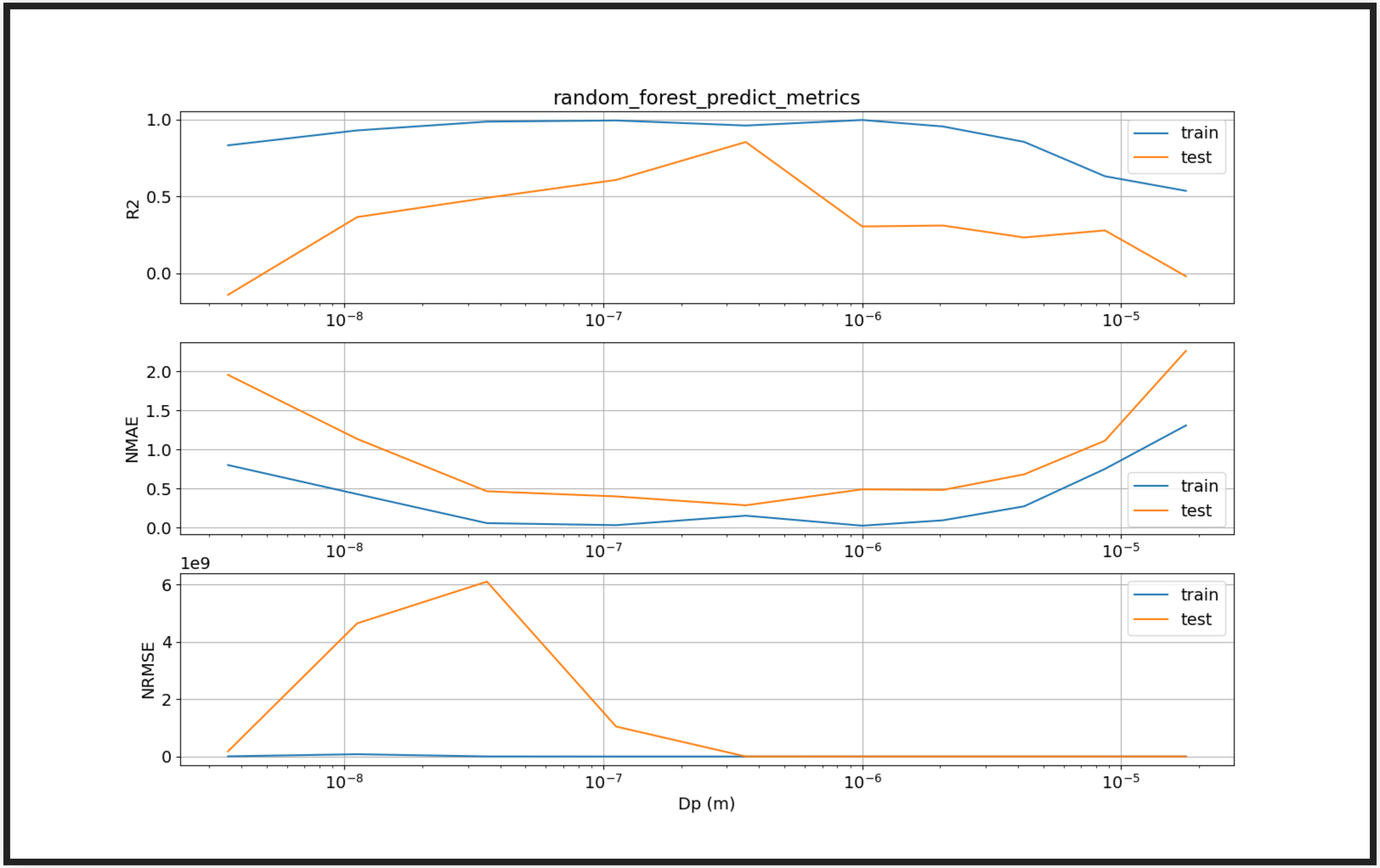
RANDOM FOREST PREDICT METRICS

- Feature number: 12 (raw) -> 46 (derived)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 5, n_iter = 40



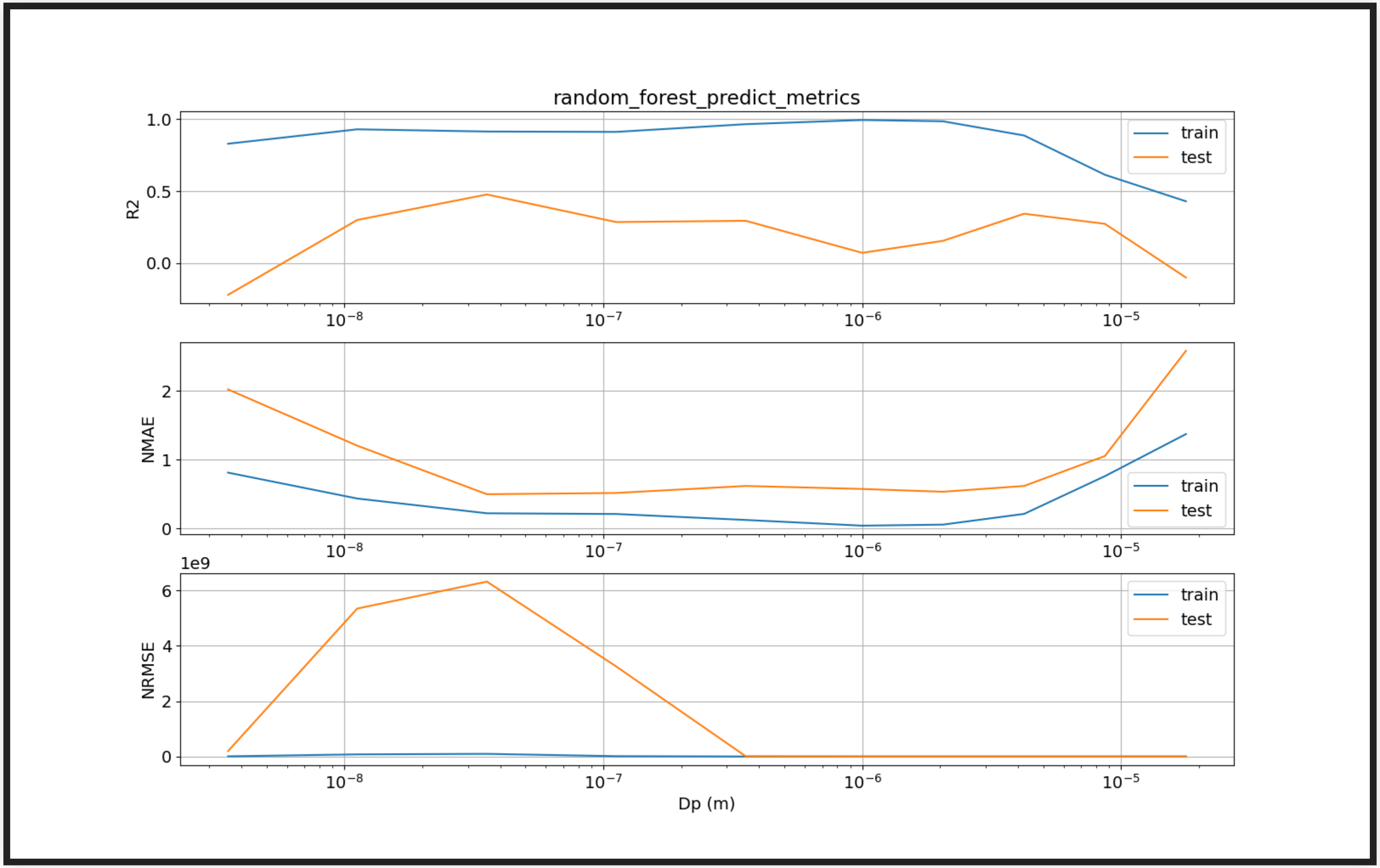
RANDOM FOREST PREDICT METRICS

- Feature number: 12 (raw) -> 46 (derived)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 3, n_iter = 20



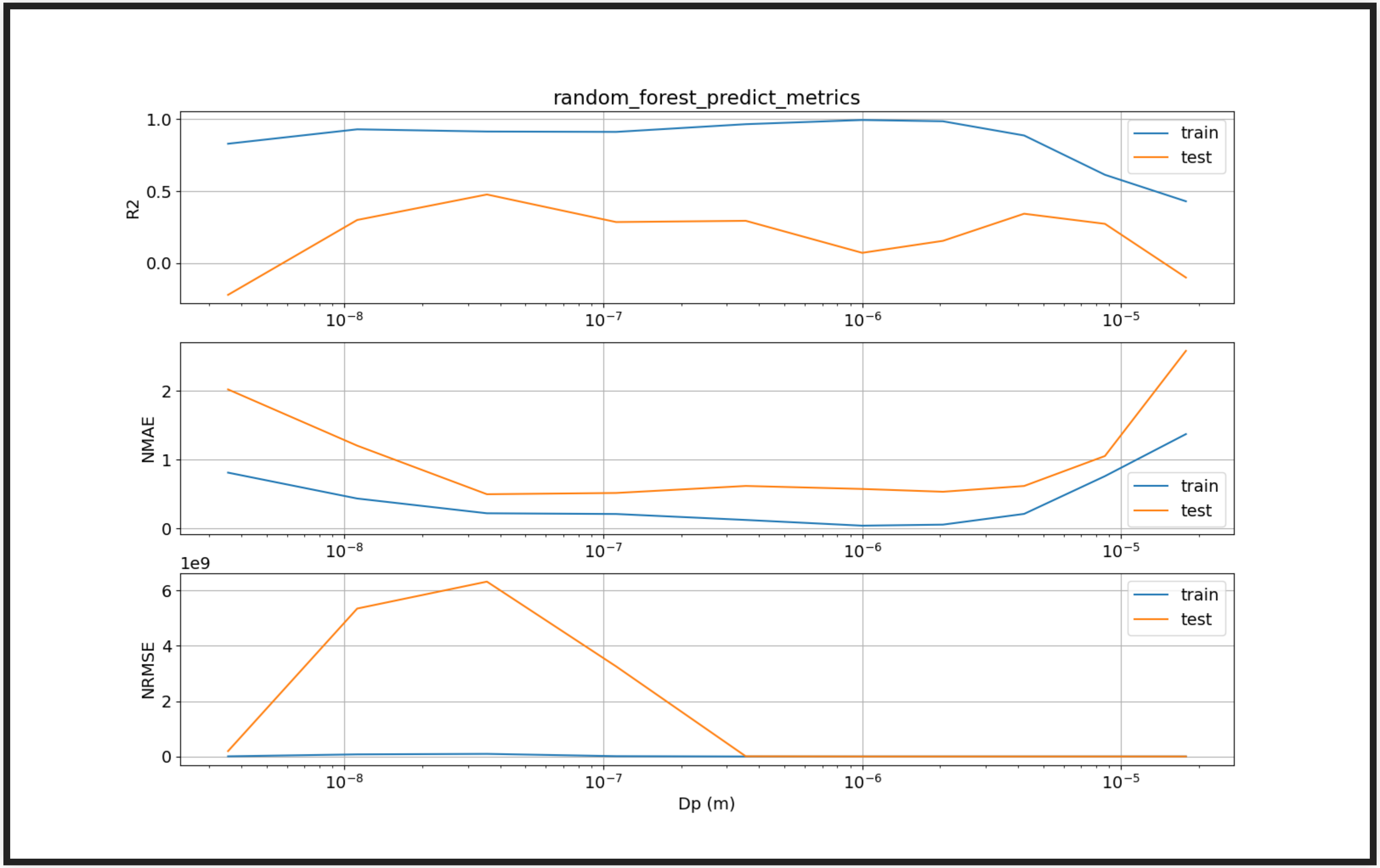
RANDOM FOREST PREDICT METRICS

- Without PM2.5
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 3, n_iter = 20



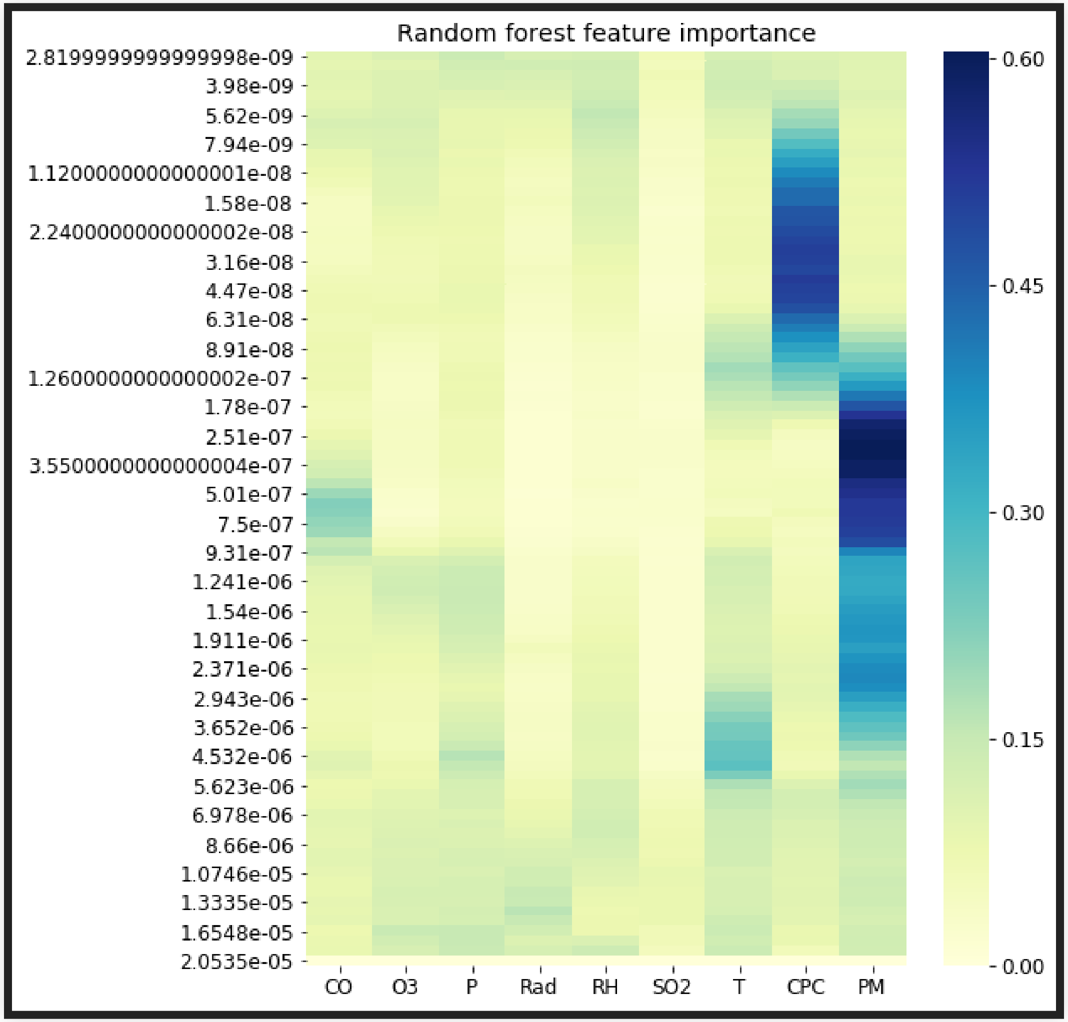
RANDOM FOREST PREDICT METRICS

- Without PM2.5, Ntot
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 3, n_iter = 20



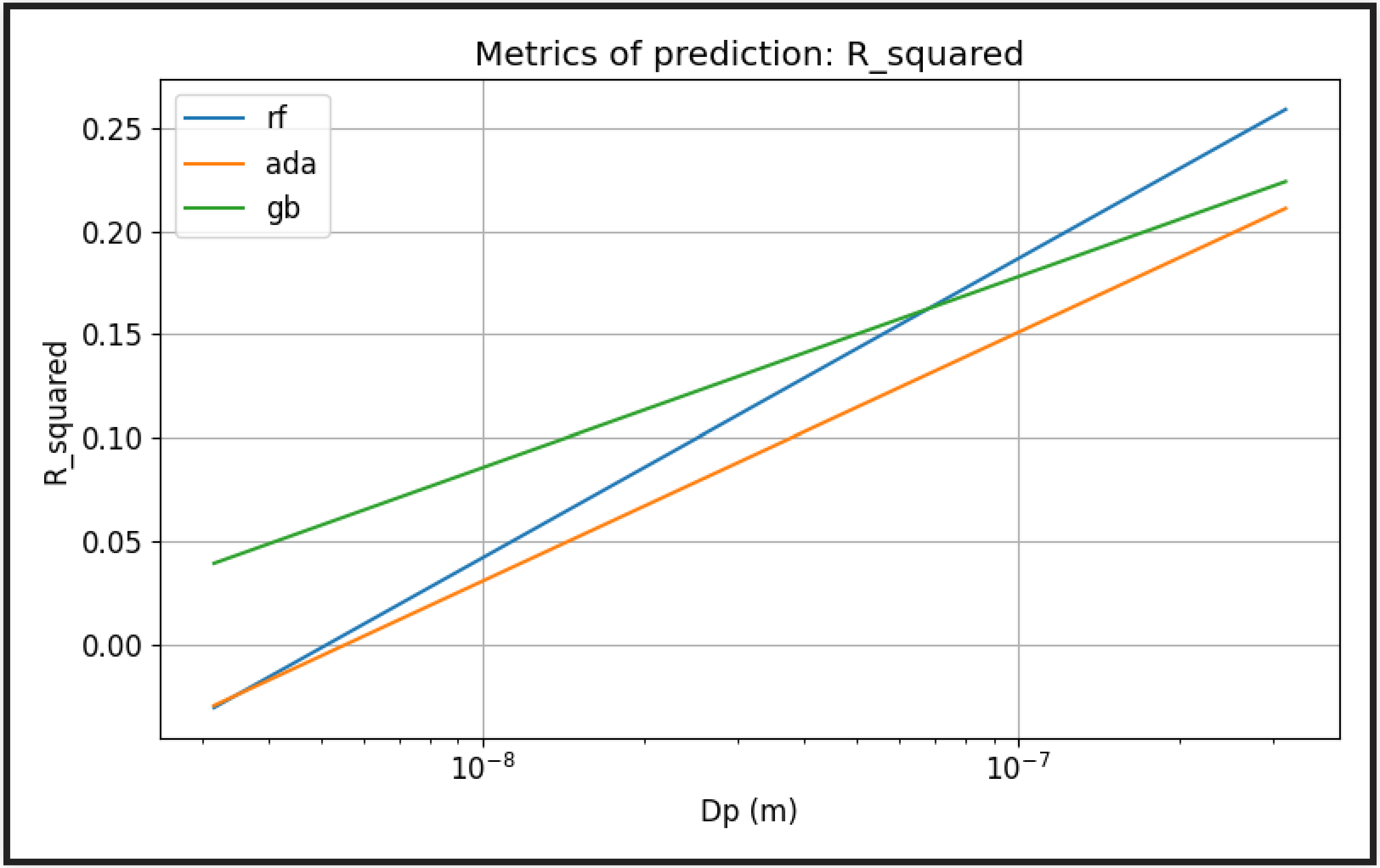
RANDOM FOREST FEATURE IMPORTANCE

- Feature number: 9 (raw)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 3, n_iter = 100



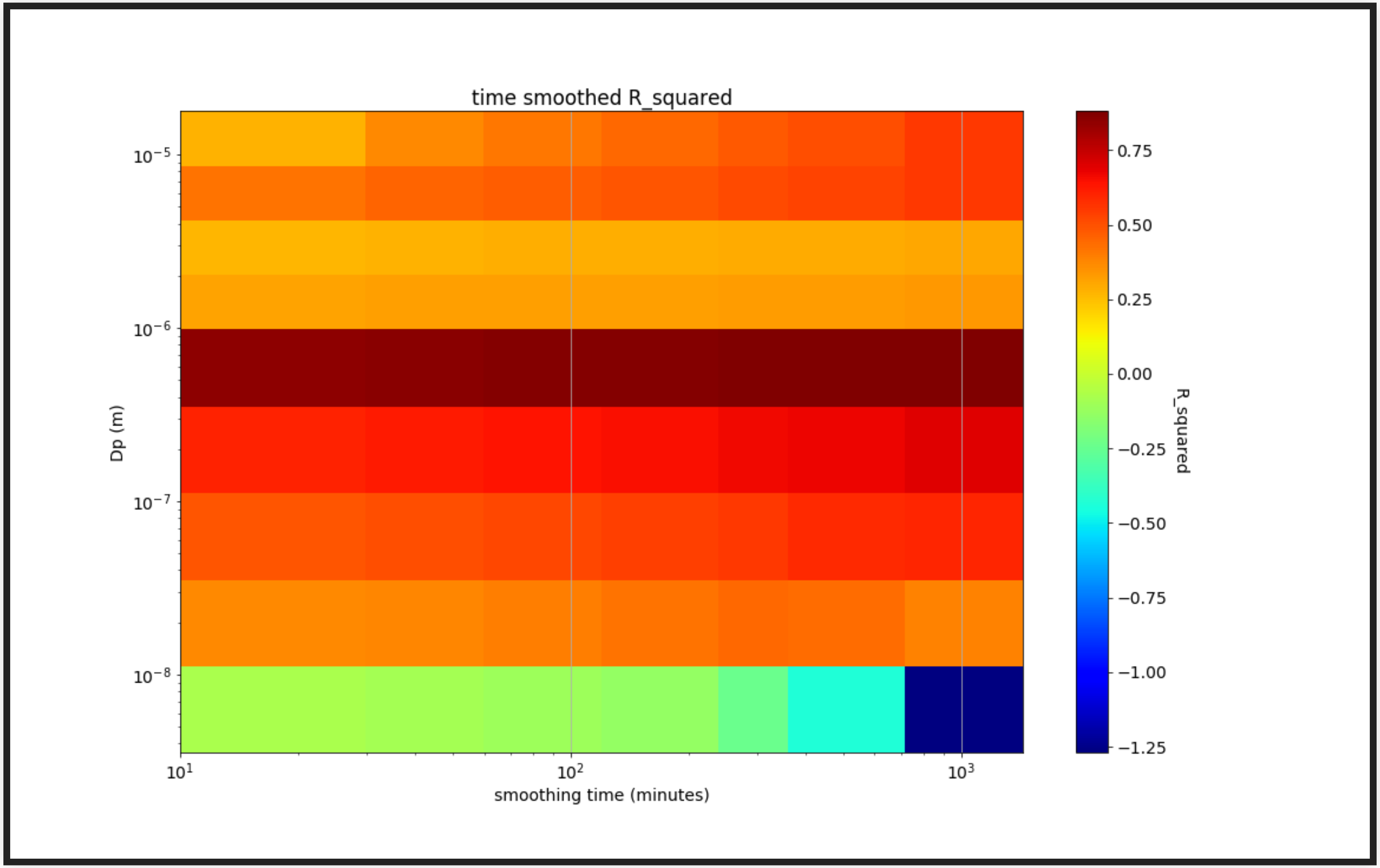
RANDOM FOREST, ADABOOST, GRADIANT BOOST

- Feature number: 9 (raw)
- 1 year training data, 10-min time resolutio
- Parameters: n-fold cv = 3, n_iter = 100



PREDICTION TIME SMOOTHING

- Feature number: 12 (raw) -> 46 (derived)
- 1 year training data, time resolutio: 10-min -> 1-day
- Parameters: n-fold cv = 3, n_iter = 100



SUMMARY

- Feasible to predict partical size distribution from routine data
- The prediction of middle size range (100-1000 nm) is relative easier
- PM2.5 and Ntot are relative important features for the predcition

NEXT TO DO

- Other learning algorithms
- Trade-off: accuracy, explainability, computational consumption
- Generalizing models or methodologies for different places

THANK YOU