Data challenge 2018-2019

Prediction of pollutant concentrations

- Schedule and milestones
- Rules
- Evaluation
- Presentation of data
- Presentation of scores
- Format of final code

Schedule

- November 7. Composition of teams
- December 13. Development of some basic solution (baseline): prediction of **mean concentrations (“part 1”)**.
- February 6-8. Intensive session, partly supervised: prediction of **conditional distributions (“part 2”)**

https://msiam.imag.fr/collab:data_challenge

Composition of teams

- Deadline: November 7 (after: random assignment)

- Rules:
  - **4 members (exception: 1 team)**
  - **at least one SIGMA student**
  - **no more than 1 MSIAM student, 2 SIGMA, 2 MoSIG students per team.**

- Register on CodaLab (see secret link)

https://competitions.codalab.org/competitions/20430?secret_key=9c39a737-0965-4120-89a4-d8b6cc08eddb

Baseline and oral presentation

- Short (12 min.) oral presentation of some baseline solution, including performance evaluation and perspectives for 2nd part of challenge.
- Compare evaluation metrics with dummy solution (mean of Chimere and last observed value).
- You may express some particular needs in terms of computational power and libraries for the intensive session in February (with a brief justification).

Intensive session

- From February 6 (9:00 AM) to 8 (6:00 PM)
- Supervision by tutors (not full time)
- February 8, 9 AM: delivery of the prediction data set and production of a prediction (rankings).
- February 8, 2 PM: defense of the whole project and final proposal with questions from teachers and other team's members.
- Report of 4 to 10 pages with experience feedback (description and comparison of the different approaches that were tried), metrics and execution times.

Assessment

- 1/3 score (ranking / performance)
- 1/3 report
- 1/3 oral presentation

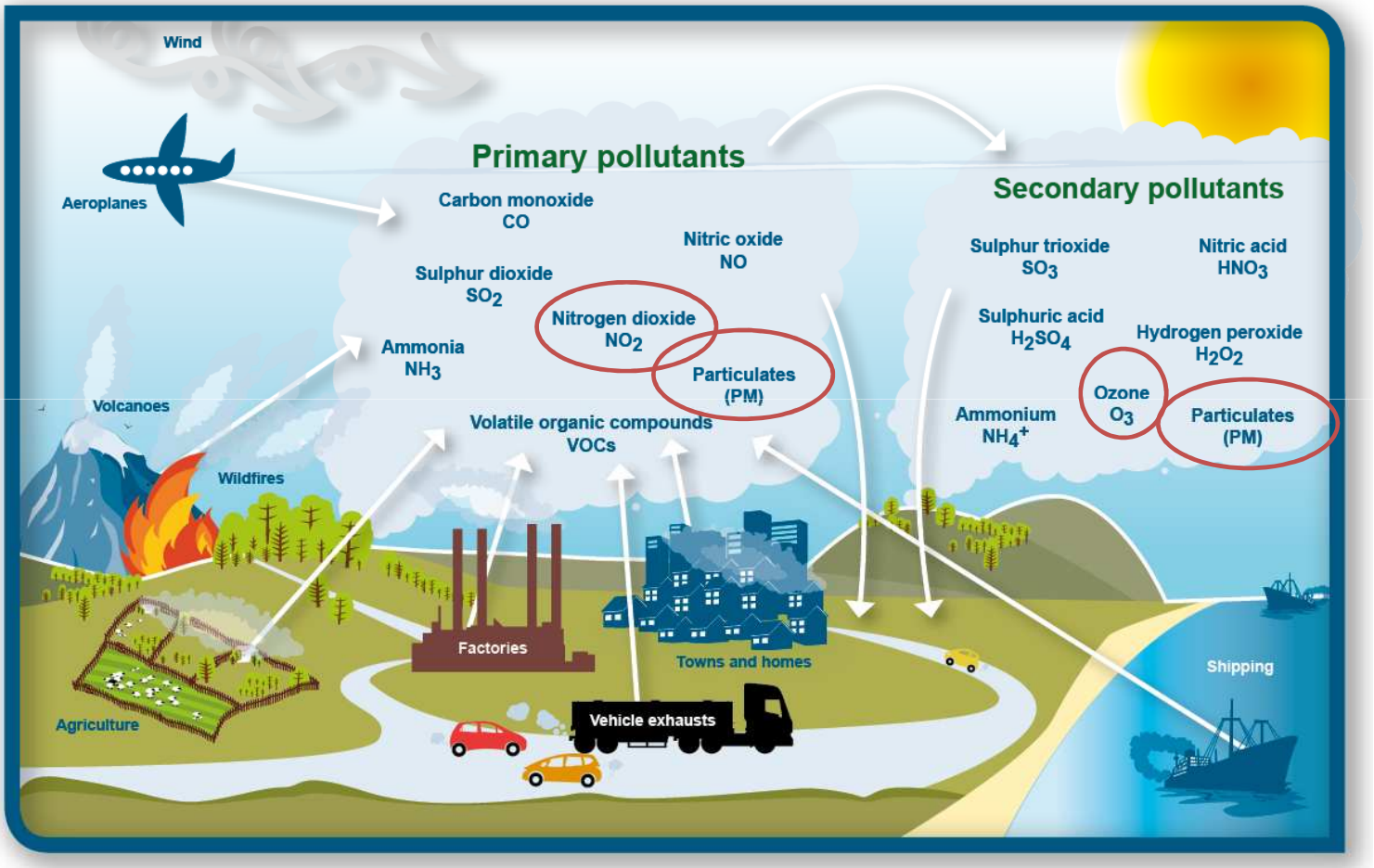
Training data set

- Measurement sites' description (coordinates, typology)

For each station (107), every hour 2012 → 2016:

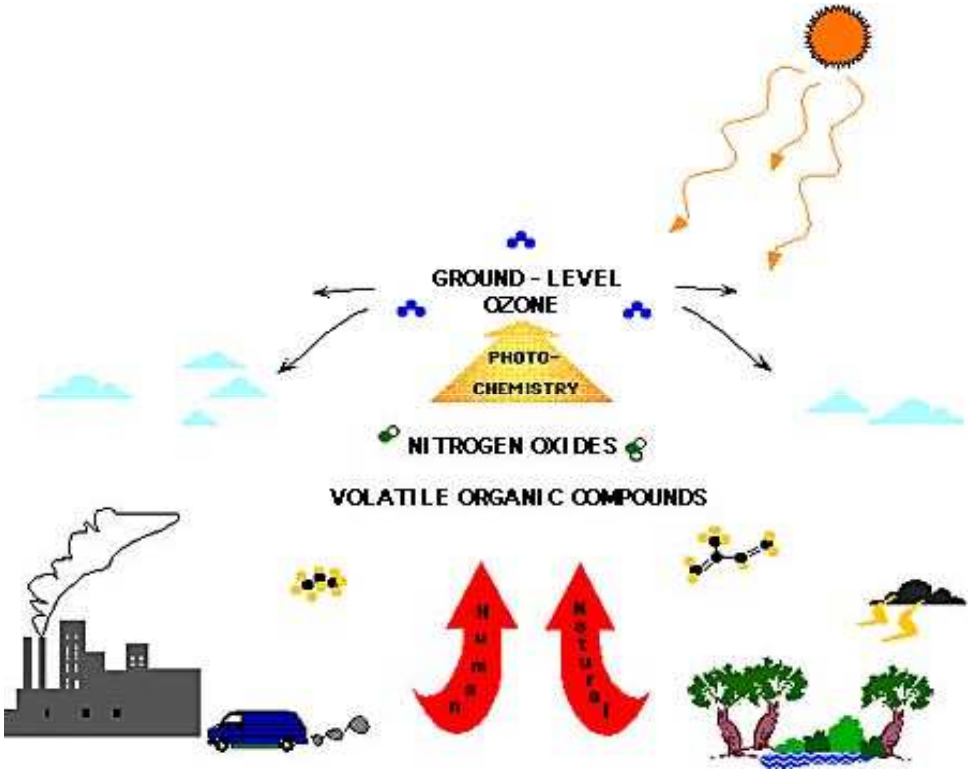
- Measurements (when available)
- CHIMERE + WRF data

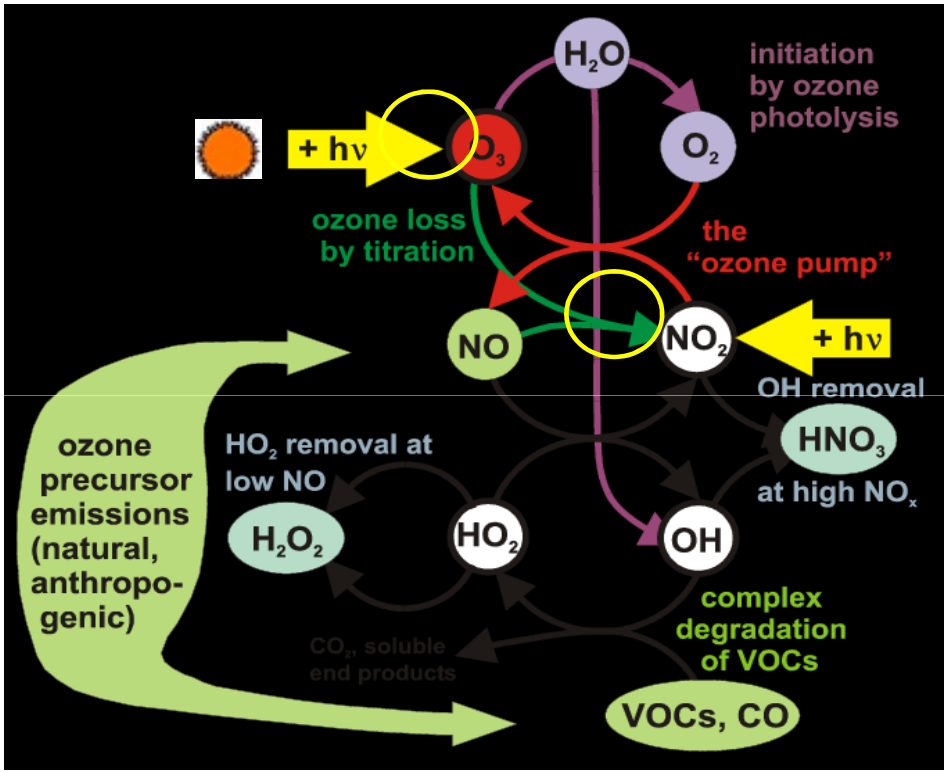
Pollutants and sources



Atmo votre parten'air AUVERGNE-RHONE-ALPES

Pollutants and interactions in the atmosphere



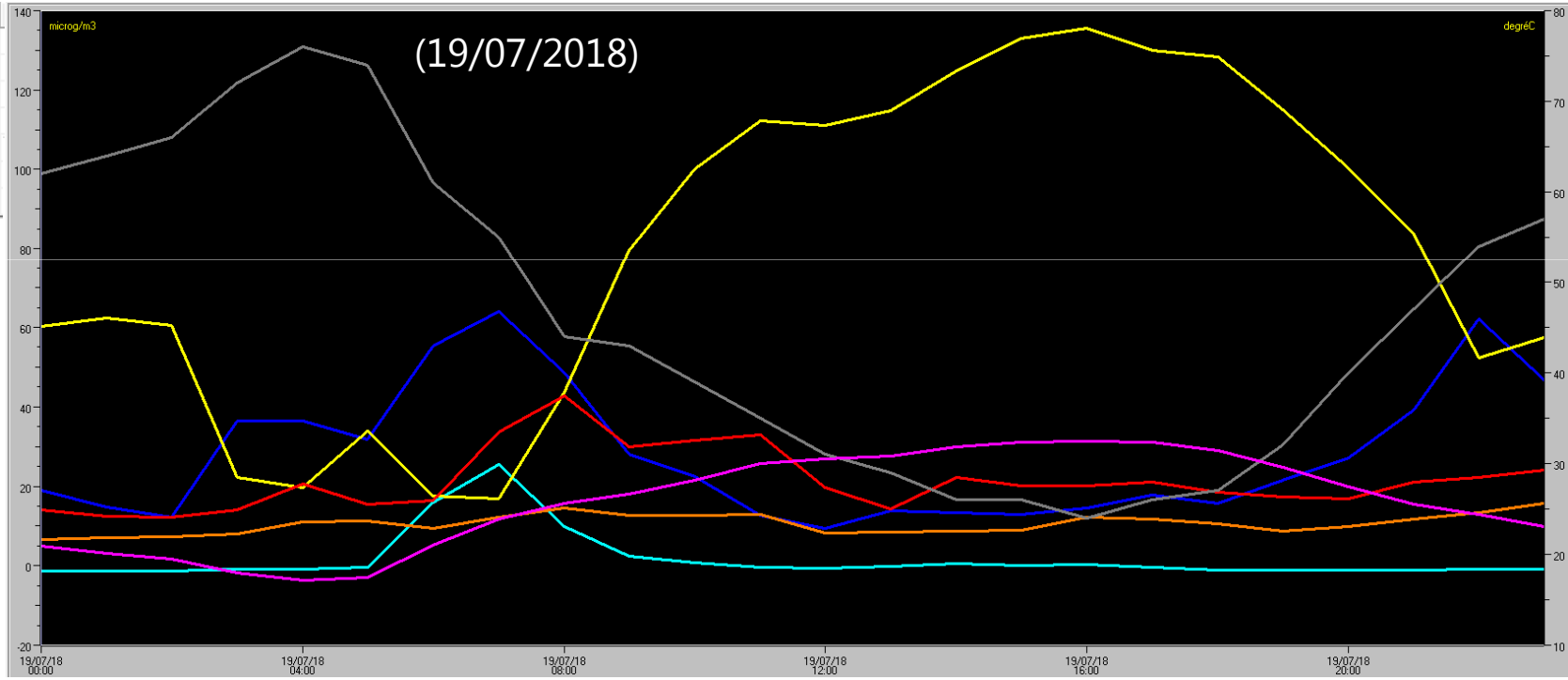


Atmo votre parten'air AUVERGNE-RHONE-ALPES

Pollutants and interactions in the atmosphere

Example during summer

Station	Mesure	Type	Courbe
LYON Centre	*Monoxyde d'azote*	Horaire	
LYON Centre	*Dioxyde d'azote*	Horaire	
LYON Centre	*Ozone*	Horaire	
LYON Centre	*Particule PM10 FDMS*	Horaire	
LYON Centre	*Particule PM2.5 FDMS*	Horaire	
MF LYON BRON	*Humidité*	Horaire	
MF LYON BRON	*Température*	Horaire	

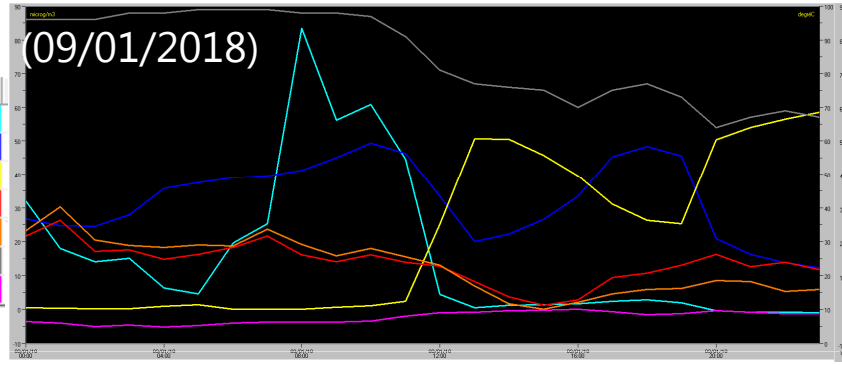


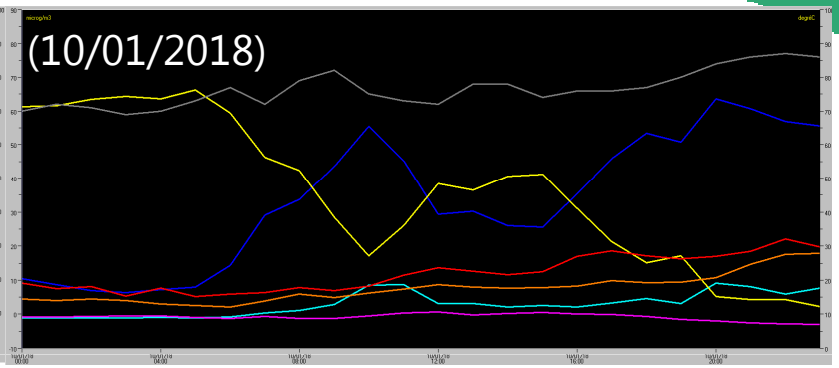
Atmo votre parten'air AUVERGNE-RHONE-ALPES

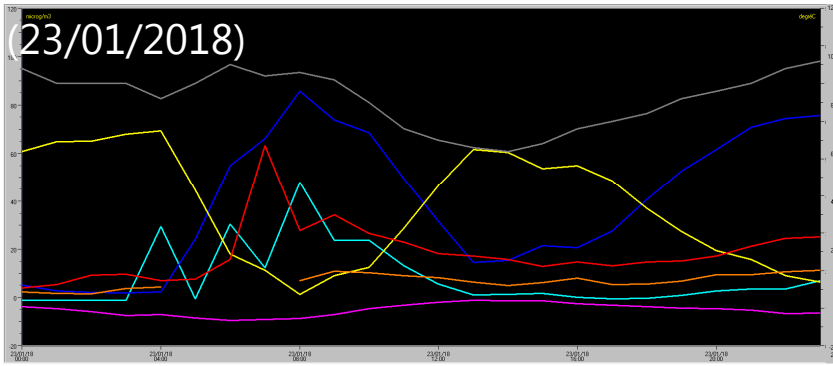
Pollutants and interactions in the atmosphere

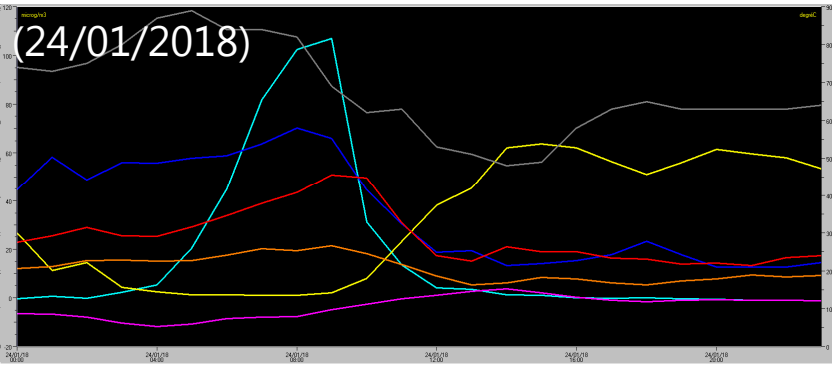
Example during winter

Station	Mesure	Type	Courbe
LYON Centre	*Monoxyde d'azote*	Horaire	
LYON Centre	*Dioxyde d'azote*	Horaire	
LYON Centre	*Ozone*	Horaire	
LYON Centre	*Particule PM10 FDMS*	Horaire	
LYON Centre	*Particule PM2.5 FDMS*	Horaire	
MF LYON BRON	*Humidité*	Horaire	
MF LYON BRON	*Température*	Horaire	





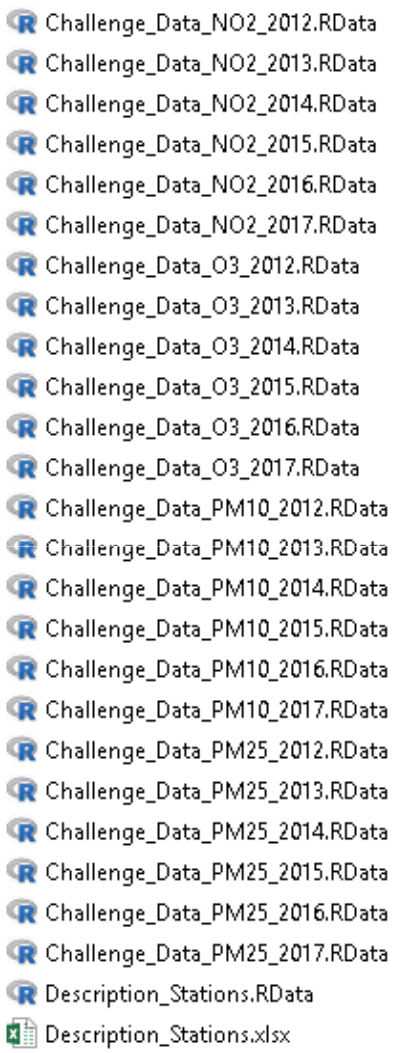




Atmo votre parten'air AUVERGNE-RHONE-ALPES

Measurements Data files

24 files > hourly concentrations for :
4 polluants : NO2, O3, PM10, PM2.5
6 years for each pollutants
+ 2 files > Description of the stations
(1 Excel & 1 RData)



Atmo votre parten'air AUVERGNE-RHONE-ALPES

Measurements sites description

Description_Stations.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	idPolair	nom_station	coord_x_l9	coord_y_l9	X_lamb2	Y_lamb2	LON	LAT	Département	Zone_EPCI	typologie	NO2_influenc	NO2_2012	NO2_2013	NO2_2014	NO2_2015
2	15013	Champ_sur_Drac	914668.44	6445968.5	867090.82	2014539.48	.7286593518272	45.0795945160278	Isère	CC Sud Grenoblois	Peri-Urbaine	Fond	1	1	1	1
3	15017	Fontaine_les_Balmes	910970.88	6458165.5	863285.49	2026715.96	.6869696547199	45.1904939464065	Isère	CA Grenoble	Urbaine	Fond	1	1	1	1
4	15018	Voiron_Urbain	902699	6476726	854847	2045221	.5894433313129	45.3600137204882	Isère	CA Pays Voironnais	Urbaine	Fond	1	1	1	1
5	15031	Ecrins	973315.81	6439213	925852.02	2008282.66	.4695006287800	44.9980956246051	Isère	CC du Briançonnais	Rurale	NA				
6	15038	Saint_Martin_Heres	916202	6457516	868526.98	2026110.36	.7532526468808	45.1830269553378	Isère	CA Grenoble	Urbaine	Fond	1	1	1	1
7	15039	Grenoble_Rocade_Sud	912411.38	6454650.5	864757.05	2023210.11	.7037695670100	45.1584201584421	Isère	CA Grenoble	Peri-Urbaine	Trafic	1	1	1	1
8	15043	Grenoble_les_Frenes	914916	6455102	867260	2023683	.7358243832881	45.1617064773352	Isère	CA Grenoble	Urbaine	Fond	1	1	1	1
9	15045	Vif	910678.63	6443436.5	863119.88	2011969.78	.6768932856663	45.0580393471876	Isère	CA Grenoble	Peri-Urbaine	Fond	1	1	1	1
10	15046	Grenoble_Boulevards	913650	6457175	865975	2025746	.7206301771691	45.1807546902755	Isère	CA Grenoble	Urbaine	Trafic	1	1	1	1
11	15048	Gresivaudan	925945.81	6468636.5	878184.26	2037325.38	.8823959874747	45.2799648220649	Isère	CC du Pays du Grésivaudan	Peri-Urbaine	Fond	1	1	1	1

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
1	NO2_2015	NO2_2016	NO2_2017	O3_influen	O3_20	O3_20	O3_20	O3_20	O3_20	O3_20	PM10_influen	PM10_2(	PM10_2(	PM10_2(	PM10_2(	PM10_2(	PM10_2(	PM25_influen	M25_2	M25_2	M25_2	M25_2	M25_2	M25_2
2	1	1	1	Fond	1	1	1	1	1	1	NA							NA						
3	1	1	1	Fond	1	1	1	1	1	1	Fond	1	1	1	1	1	1	NA						
4	1	1	1	Fond	1	1	1	1	1	1	Fond	1	1	1	1	1		NA						
5				Fond	1	1	1	1	1	1	NA							NA						
6	1	1	1	Fond	1	1	1	1	1	1	Fond	1	1	1	1	1	1	NA						
7	1	1	1	NA							Trafic	1	1	1	1	1	1	Trafic	1	1	1	1	1	1
8	1	1	1	Fond	1	1	1	1	1	1	Fond	1	1	1	1	1	1	Fond	1	1	1	1	1	1
9	1	1	1	Fond	1	1	1	1	1	1	Fond	1	1	1	1	1	1	NA						
10	1	1	1	NA							Trafic	1	1	1	1	1	1	NA						

Stations | Explanations

Atmo votre parten'air AUVERGNE-RHONE-ALPES

Measurements sites description

Description_Stations.xlsx

	A	B
1	**Column_Name**	**Explanation**
2	idPolair	ID of the station (XXYYY > XX = Organisme ID ; YYY = Station ID)
3	nom_station	Name of the station
4	coord_x_l93	Coord in Lambert 93
5	coord_y_l93	Coord in Lambert 93
6	X_lamb2	Coord in Lambert 2
7	Y_lamb2	Coord in Lambert 2
8	LON	Coord Longitude in Decimal Degree
9	LAT	Coord Latitude in Decimal Degree
10	Département	Zone Level Department
11	Zone_EPCI	Zone Level Cross-Town
12	typologie	Typology of the Station : Urban, Periurban, Rural
13	NO2_influence	Influence of the measure NO2 : background (fond), traffic (trafic), industrial (industriel)
14	NO2_2012	NO2 measurement representative this year ? (1=Yes ; ""=No)
15	NO2_2013	NO2 measurement representative this year ? (1=Yes ; ""=No)
16	NO2_2014	NO2 measurement representative this year ? (1=Yes ; ""=No)
17	NO2_2015	NO2 measurement representative this year ? (1=Yes ; ""=No)
18	NO2_2016	NO2 measurement representative this year ? (1=Yes ; ""=No)
19	NO2_2017	NO2 measurement representative this year ? (1=Yes ; ""=No)
20	O3_influence	Influence of the measure Ozone : background (fond) (No Ozone measurment in trafic or industrial influence)
21	O3_2012	Ozone measurement representative this year ? (1=Yes ; ""=No)
22	O3_2013	Ozone measurement representative this year ? (1=Yes ; ""=No)
23	O3_2014	Ozone measurement representative this year ? (1=Yes ; ""=No)
24	O3_2015	Ozone measurement representative this year ? (1=Yes ; ""=No)
25	O3_2016	Ozone measurement representative this year ? (1=Yes ; ""=No)
26	O3_2017	Ozone measurement representative this year ? (1=Yes ; ""=No)
27	PM10_influence	Influence of the measure PM10 : background (fond), traffic (trafic), industrial (industriel)

Stations | Explanations

« Data_pollutant » description

Challenge_Data_NO2_2012.RData

```
Organisme;Station;Mesure;Date;Valeur
15;013;004;01/01/2012 00:00;46
15;013;004;01/01/2012 01:00;42
15;013;004;01/01/2012 02:00;38
15;013;004;01/01/2012 03:00;37
15;013;004;01/01/2012 04:00;36
15;013;004;01/01/2012 05:00;31
15;013;004;01/01/2012 06:00;33
15;013;004;01/01/2012 07:00;33
15;013;004;01/01/2012 08:00;33
15;013;004;01/01/2012 09:00;28
15;013;004;01/01/2012 10:00;22
15;013;004;01/01/2012 11:00;17
15;013;004;01/01/2012 12:00;24
15;013;004;01/01/2012 13:00;23
15;013;004;01/01/2012 14:00;12
15;013;004;01/01/2012 15:00;6
15;013;004;01/01/2012 16:00;8
15;013;004;01/01/2012 17:00;32
15;013;004;01/01/2012 18:00;33
15;013;004;01/01/2012 19:00;41
15;013;004;01/01/2012 20:00;42
15;013;004;01/01/2012 21:00;39
15;013;004;01/01/2012 22:00;32
15;013;004;01/01/2012 23:00;31
15;013;004;02/01/2012 00:00;25
15;013;004;02/01/2012 01:00;14
15;013;004;02/01/2012 02:00;14
15;013;004;02/01/2012 03:00;28
15;013;004;02/01/2012 04:00;26
15;013;004;02/01/2012 05:00;23
15;013;004;02/01/2012 06:00;24
15;013;004;02/01/2012 07:00;26
15;013;004;02/01/2012 08:00;30
```

Description_Stations.xlsx

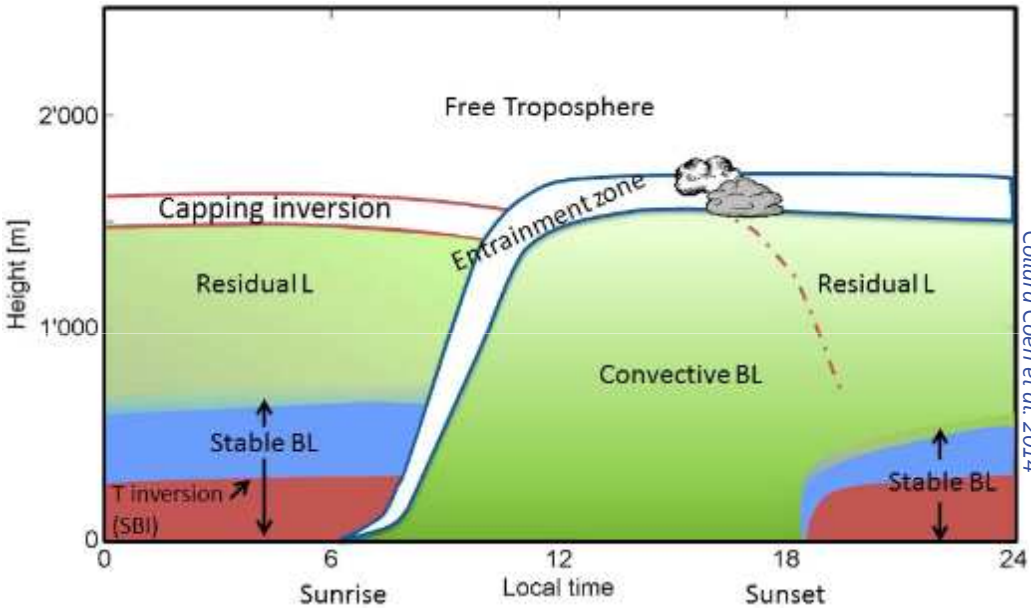
	A	B	C
1	idPolair	nom_station	coord_x_l9
2	15013	Champ_sur_Drac	914668.44
3	15017	Fontaine_les_Balmes	910970.88
4	15018	Voiron_Urbain	902699

	A	B
1	Column_Name	Explanation
2	idPolair	ID of the station (XXYYY > XX = Organisme ID ; YYY = Station ID)

« Meteo data »

At each measurement site, each hour 2012 to 2016:

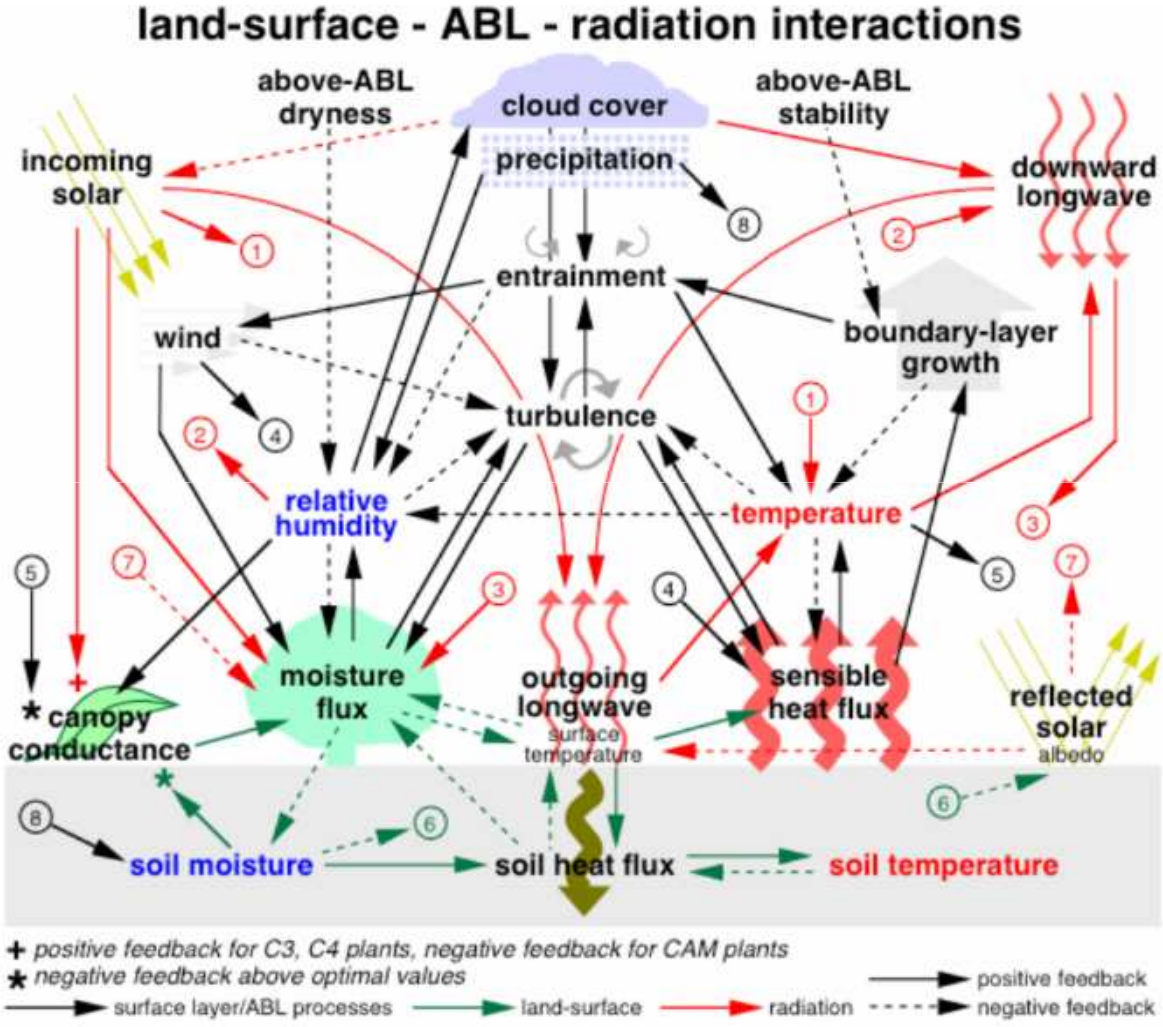
- **T2 : 2mtemperature** , °C
- **Q2 : 2m specific humidity,** kg/kg
- **RH2 : 2m relative humidity**, %
- **U10, V10 : 10m wind components U & V**, m/s
- **VV10, DV10 : 10m wind speed & direction**, m/s , deg
- **PSFC : surface pressure**, Pa
- **PRECIP : precipitation**, mm,
- **PBLH : PBL height : mixing height**, m
- **HFX : sensible heat flux**, $W.m^{-2}$
- **LH : latent heat flux(surface evaporation)**, $W.m^{-2}$
- **ALBEDO**
- **SNOWC : flag indicating snow coverage (1 for snow cover)**
- **Geop500hPa et geop750hPa : geopotential altitudes at 500 et 850hPa**, m



https://dtcenter.org/events/wrf-nmm_tutorial06_summer/Presentations/Ek_WRFNMM_tutorial_Aug2006_land-radiation.pdf

Rechercher

Zoom automatique



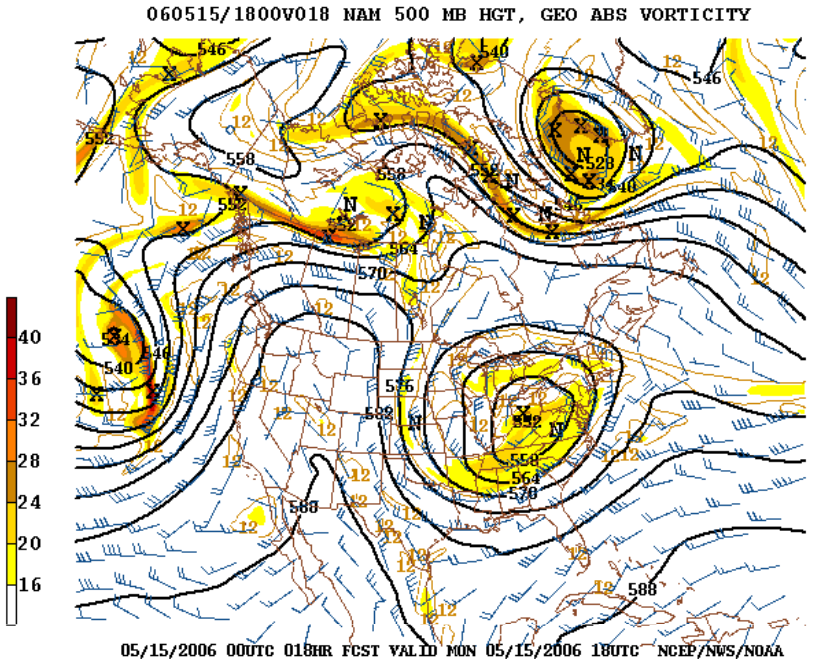
votre parten'air AUVERGNE-RHONE-ALPES

Geopotential 500 / 700 hPa

- https://en.wikipedia.org/wiki/Geopotential_height :

- Geopotential height is a vertical coordinate referenced to Earth's mean sea level

- geopotential height of a certain pressure level = the geopotential height at which that pressure occurs



At an elevation of h, the **geopotential** is defined as:

$$\Phi(h) = \int_0^h g(\phi, z)\, dz\,,$$

where $g(\phi, z)$ is the acceleration due to gravity, ϕ is latitude, and z is the geometric elevation. Thus geopotential is the gravitational potential energy per unit mass at that elevation h.[1]

The **geopotential height** is:

$$Z_g(h) = \frac{\Phi(h)}{g_0}\,,$$

which normalizes the geopotential to g_0, the standard gravity at mean sea level.[citation needed]

➔ Global & synoptic patterns

« Meteo data » : WRF output

WRF_*YEAR*.RData

YEAR = 2012 → 2016

dataframe « wrfdata » : 1 row = 1 hour : date column

idPolair = id station ex 15017, 15203…

```
> load('meteoWRF_2013.RData')
> ls()
[1] "wrfData"
> head(wrfData)
                 date idPolair  T2     Q2 U10 V10  PSFC PBLH   LH   HFX ALBEDO
1 2013-01-01 00:00:00     7001 3.1 0.0044 0.7 5.0 96993    0  0.0   0.0    0.1
2 2013-01-01 01:00:00     7001 3.3 0.0045 1.7 7.5 96947  530 17.9 -65.1    0.1
3 2013-01-01 02:00:00     7001 3.6 0.0046 2.3 8.5 96965  563 18.4 -48.7    0.1
4 2013-01-01 03:00:00     7001 3.9 0.0046 2.3 7.6 96973  557 24.4 -47.1    0.1
5 2013-01-01 04:00:00     7001 4.3 0.0047 1.8 7.2 96959  582 25.8 -56.8    0.1
6 2013-01-01 05:00:00     7001 4.8 0.0048 1.6 7.1 96945  552 23.9 -52.1    0.1
  SNOWC HR2 VV10 DV10 PRECIP
1     0  90  5.1  188   0.00
2     0  90  7.7  193   0.01
3     0  90  8.8  195   0.17
4     0  89  8.0  197   0.41
5     0  88  7.4  194   0.36
6     0  87  7.3  192   0.26
```

« Meteo data » : Geop500hPa et geop750hPa

Geop.*idstation*.*YEAR*.**d02.RData**

YEAR = 2012 → 2016

Idstation : id station ex : 15017, 15203…

d02 = *domain_geop* = France

→ dataframe:

date id_polair geo_p_500hPa geo_p_850hPa

```
> load('Geop.15017.2012.d02.RData')
> ls()
[1] "out"
> str(out)
'data.frame':   8785 obs. of  4 variables:
 $ date         : int  2012010100 2012010101 2012010102
 $ id_polair    : num  15017 15017 15017 15017 15017 ..
 $ geop_p_500hPa: num  5550 5553 5556 5560 5564 ...
 $ geop_p_850hPa: num  1436 1434 1434 1437 1437 ...
> head(out)
        date id_polair geop_p_500hPa geop_p_850hPa
1 2012010100     15017      5550.076      1436.409
2 2012010101     15017      5553.286      1434.187
3 2012010102     15017      5555.764      1434.181
4 2012010103     15017      5560.354      1436.774
5 2012010104     15017      5563.615      1437.479
6 2012010105     15017      5567.679      1439.133
>
```

« Data pollution » : CHIMERE ouput

CHIMERE_YEAR.**RData**

YEAR = 2012 → 2016

→ dataframe polChimere

val = concentration, µg.m^{-3}

Param = O3, NO2, PM10, PM25

```
> load('CHIMERE_2013.RData')
> head(polChimere)
                  date  val idPolair param
1 2013-01-01 00:00:00 42.8    15017    O3
2 2013-01-01 00:00:00 42.9    15018    O3
3 2013-01-01 00:00:00 43.3    15038    O3
4 2013-01-01 00:00:00 43.2    15039    O3
5 2013-01-01 00:00:00 43.3    15043    O3
6 2013-01-01 00:00:00 42.4    15045    O3
> str(polChimere)
'data.frame':   271787276 obs. of  4 variables:
 $ date    : chr  "2013-01-01 00:00:00" "2013-01-01 00:00:
 $ val     : num  42.8 42.9 43.3 43.2 43.3 42.4 43.3 42.4
 $ idPolair: num  15017 15018 15038 15039 15043 ...
 $ param   : Factor w/ 4 levels "O3","NO2","PM10",..: 1 1
```

Scoring

- Part 1 (October → December). Prediction of mean concentrations
  - **Station *s*, time *t*, pollutant *p*, horizon *h* (7 AM to J+2 11 PM), ground truth** $m_{t+h,s,p}$
  - **Constraints: measurements known up to 6:00 AM at J0.**
  - **Score:**

$$\sum_{t=t_0}^{t_{\max}} \sum_{s=1}^{135} \sum_{h=0}^{2} \sum_{p=1}^{4} (3-h)\, p \left(\frac{m_{t+h,s,p} - \hat{m}_{t+h,s,p}}{\hat{s}_p} \right)^2 \left| \log \left(\frac{\hat{m}_{t+h,s,p}}{m_{t+h,s,p}} \right) \right|$$

  with *p*=4 for PM10, *p*=3 for PM25, *p*=2 for NO_2, *p*=1 for O_3 and $\hat{s}_p$ **: standard deviation on training set.**
  - **Training: 2012-2016. Evaluation: 2017 (available on CodaLab only).**
- Part 2 (January - February). Prediction of conditional distributions.
  - **Score:**

$$\sum_{t=t_0}^{t_{\max}} \log \hat{\wp}\left((m_{t+h,s,p})_{h=0,1,2;s=1,\ldots,135,\, p=1,2,3,4} \,\middle|\, (m_{r,s,p})_{r<t;s=1,\ldots,135,\, p=1,2,3,4}, (x_r)_{r<t} \right)$$

  - x_r **various other predictors**
  - **Constraints: execution time < 10 min in prediction on 80 cores**
  - $\int \hat{\wp}(\boldsymbol{m})\, d\boldsymbol{m} = 1$

Code

- Participants are asked to use Gitlab @gricad-gitlab.univ-grenoble-alpes.fr
- You will have a project per team to collaborate on your code.

- The code will have to be written in Python or R and follow some specific input/output formatting.
- In particular, you will have to provide how many days of history you want to use as regressors and you will need to handle missing data.