#PlanDeRecuperación

Plan de Recuperación, Transformación y Resiliencia

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA TERCERA DEL GOBIERNO
MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO

AEMet Agencia Estatal de Meteorología

Financiado por la Unión Europea
NextGenerationEU

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Bringing the complexity of organic chemistry to climate models with machine learning techniques

Camille Mouchel-Vallon, Alma Hodzic, John Schreck, Charlie Becker, Keeley Lawrence, Siyuan Wang, Jinkyul Choi, Daven Henze, David John Gagne

NSF | NCAR Operated by UCAR

PASC24, Zürich, 2024/06/05

Context

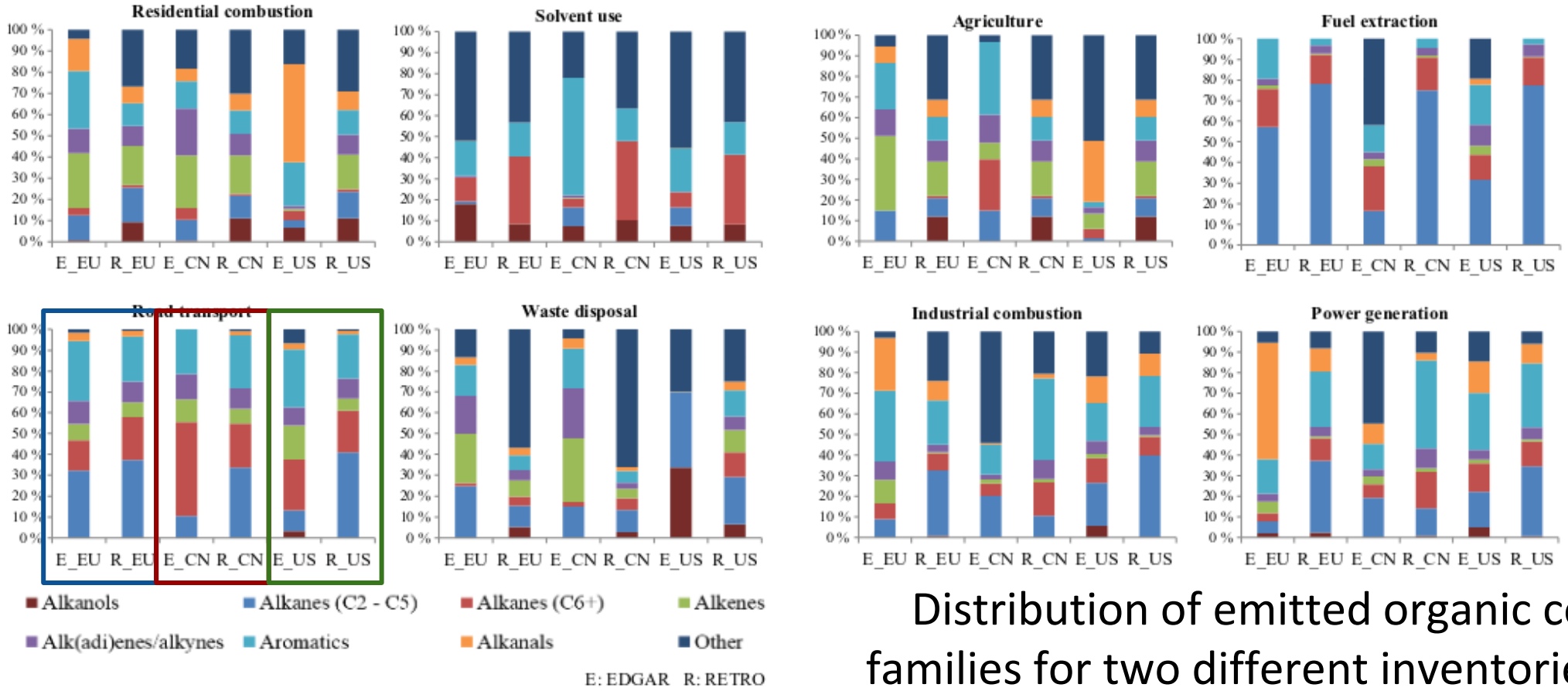
Plan de Recuperación, Transformación y Resiliencia

1st objective: A high resolution spatial (up to 2.5x2.5 km) and temporal (hourly) emissions system to compute emissions of **primary atmospheric pollutants** for Spain to be used in the national **air quality prediction** system.

Gaseous primary atmospheric pollutants
NO_x, CO, SO_x, NH_3, **NMVOCs**

Non-Methane Volatile Organic Compounds are emitted by human and natural activities and their chemical evolution in the atmosphere has an impact on air quality, health and climate

Primary organic compounds: many sources, high diversity, high uncertainty



Distribution of emitted organic compounds families for two different inventories in **Europe**, **China** and **USA**

(Huang et al., 2017)

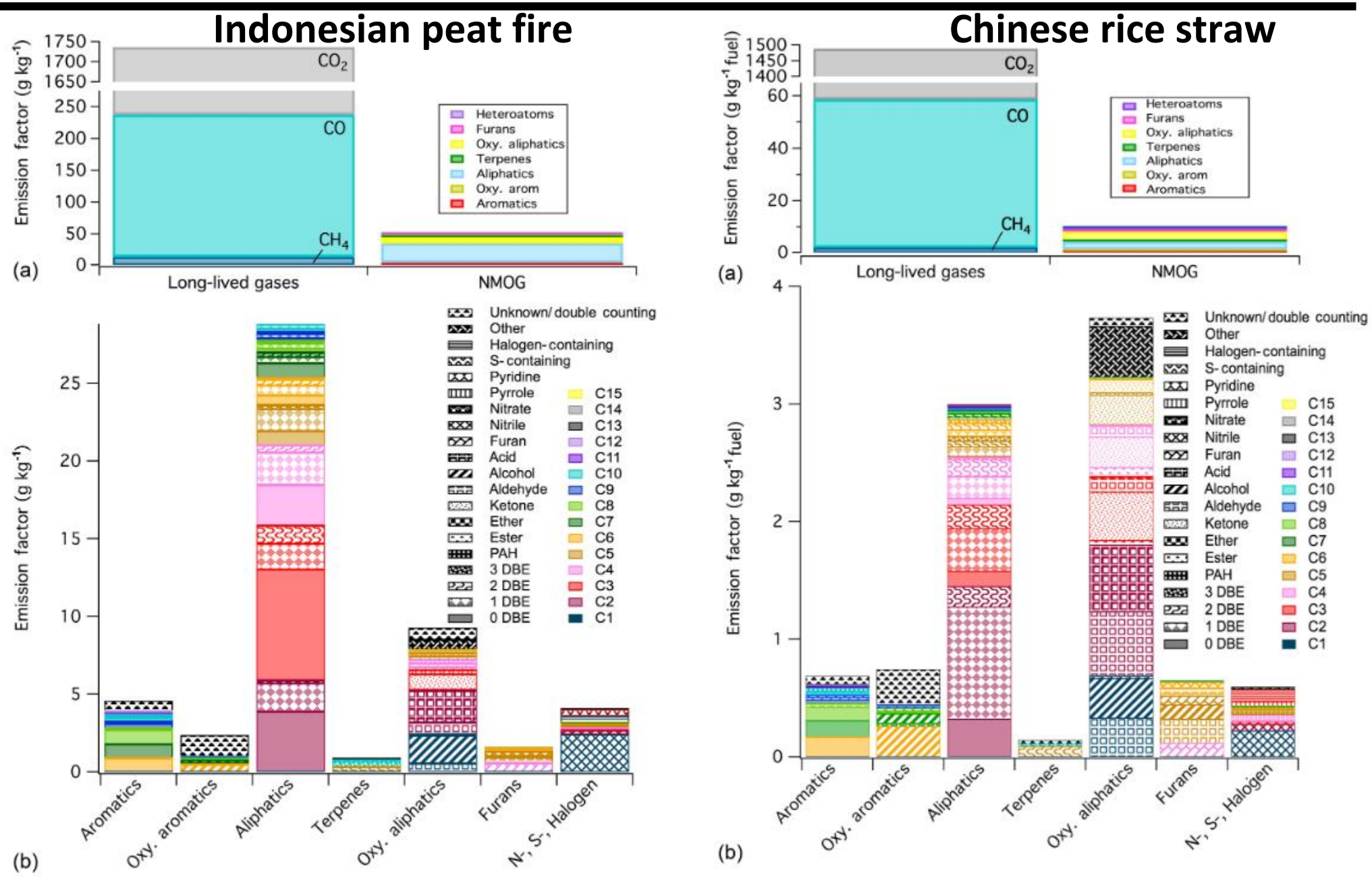
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Financiado por la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

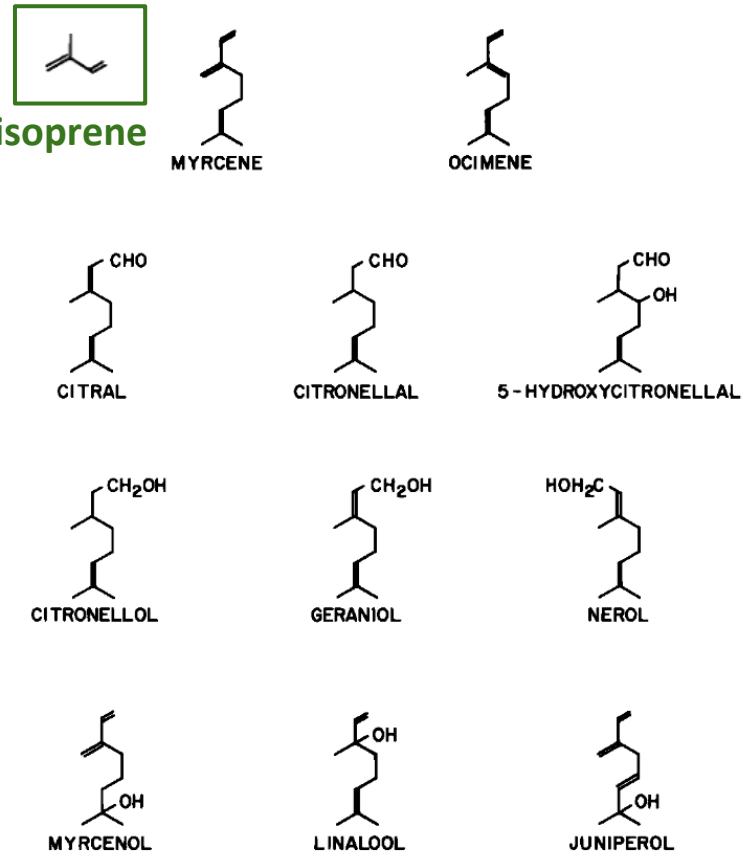
Primary organic compounds: many sources, high diversity, high uncertainty

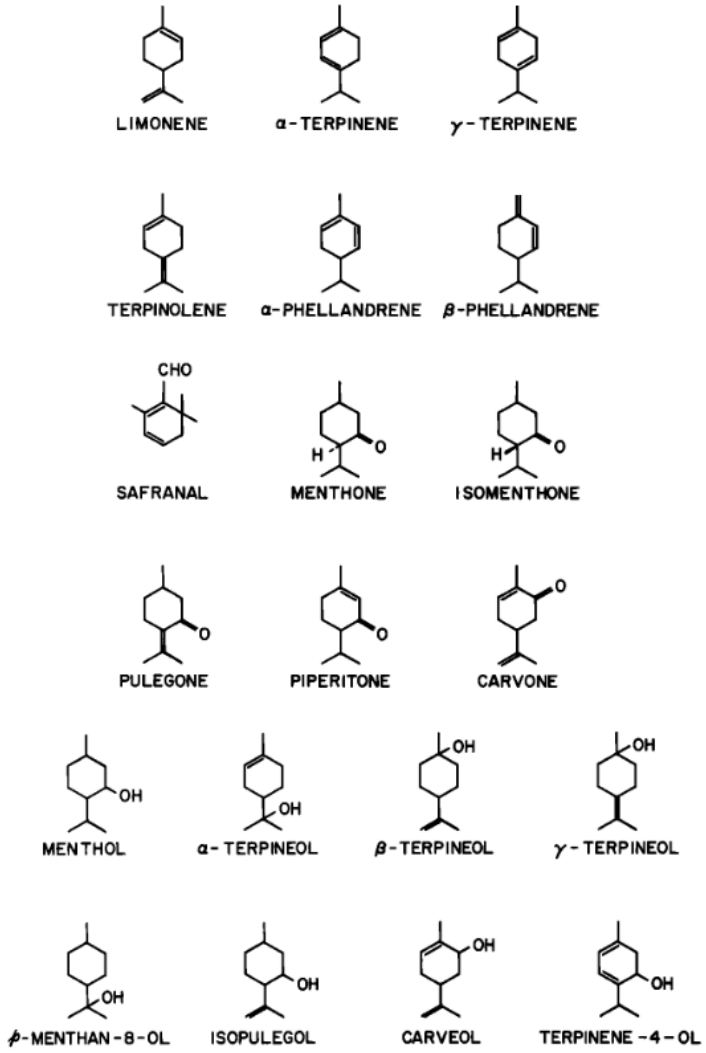
Measured fire emissions composition (Hatch et al., 2017)

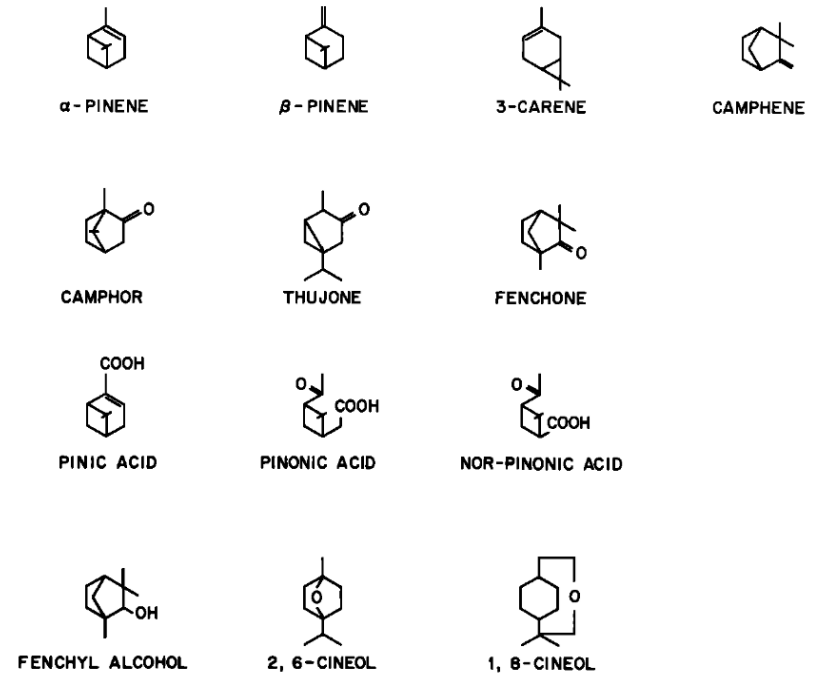


Barcelona Supercomputing Center
Centro Nacional de Supercom

Primary organic compounds: many sources, high diversity, high uncertainty



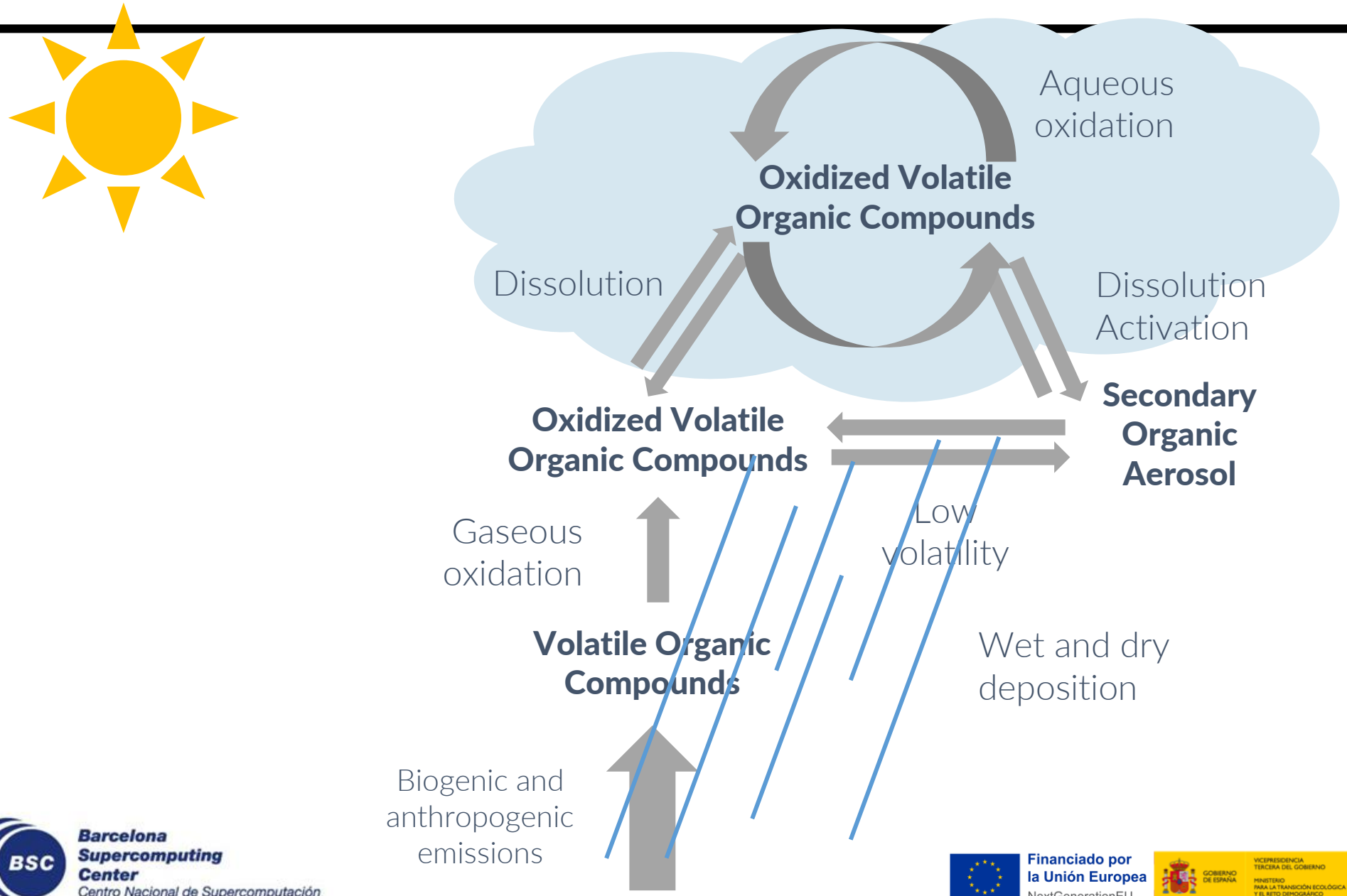




Inventory of terpenoids emitted by plants
(Graedel, 1979)

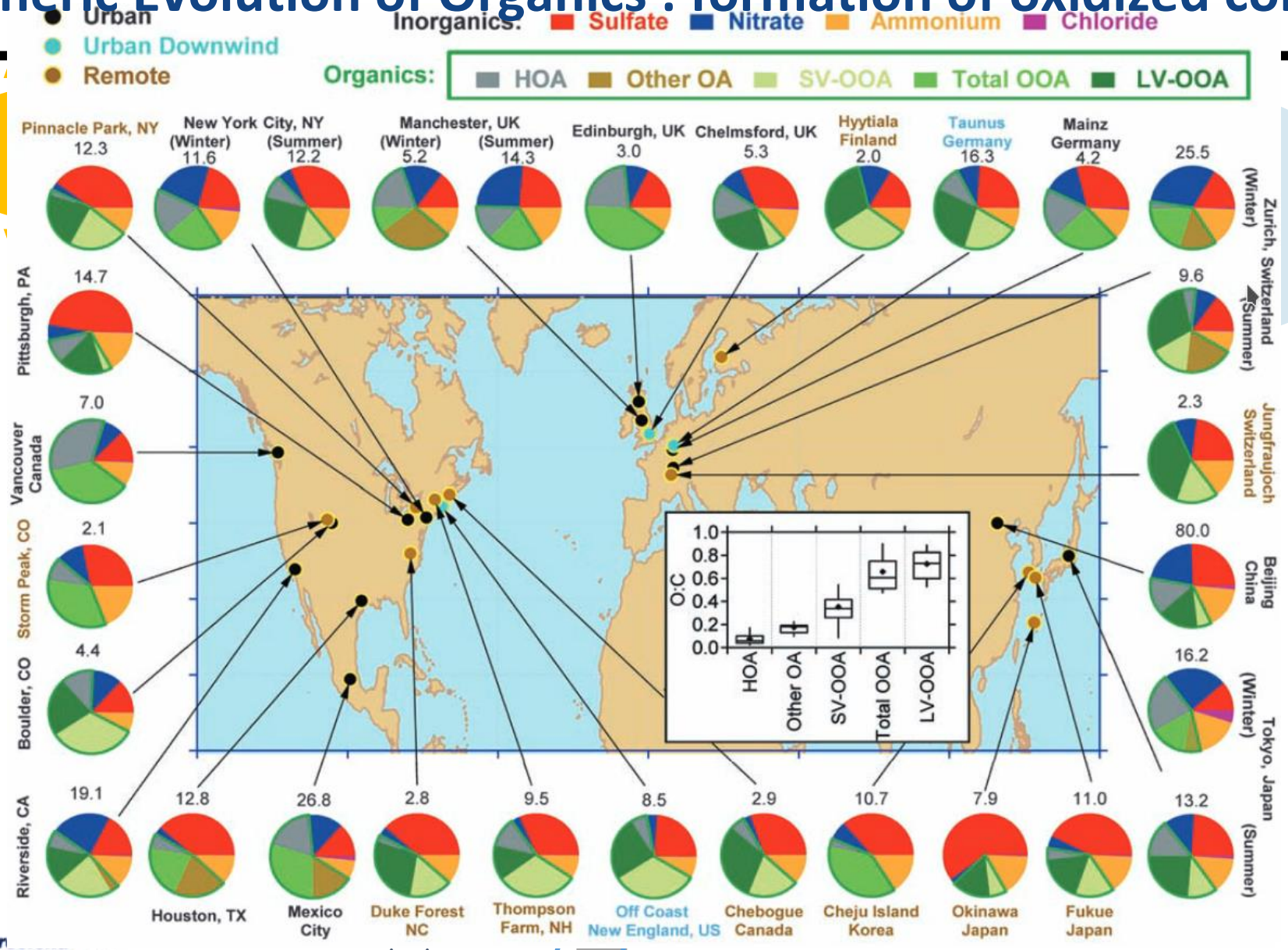
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea
NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Atmospheric Evolution of Organics : formation of oxidized compounds



Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

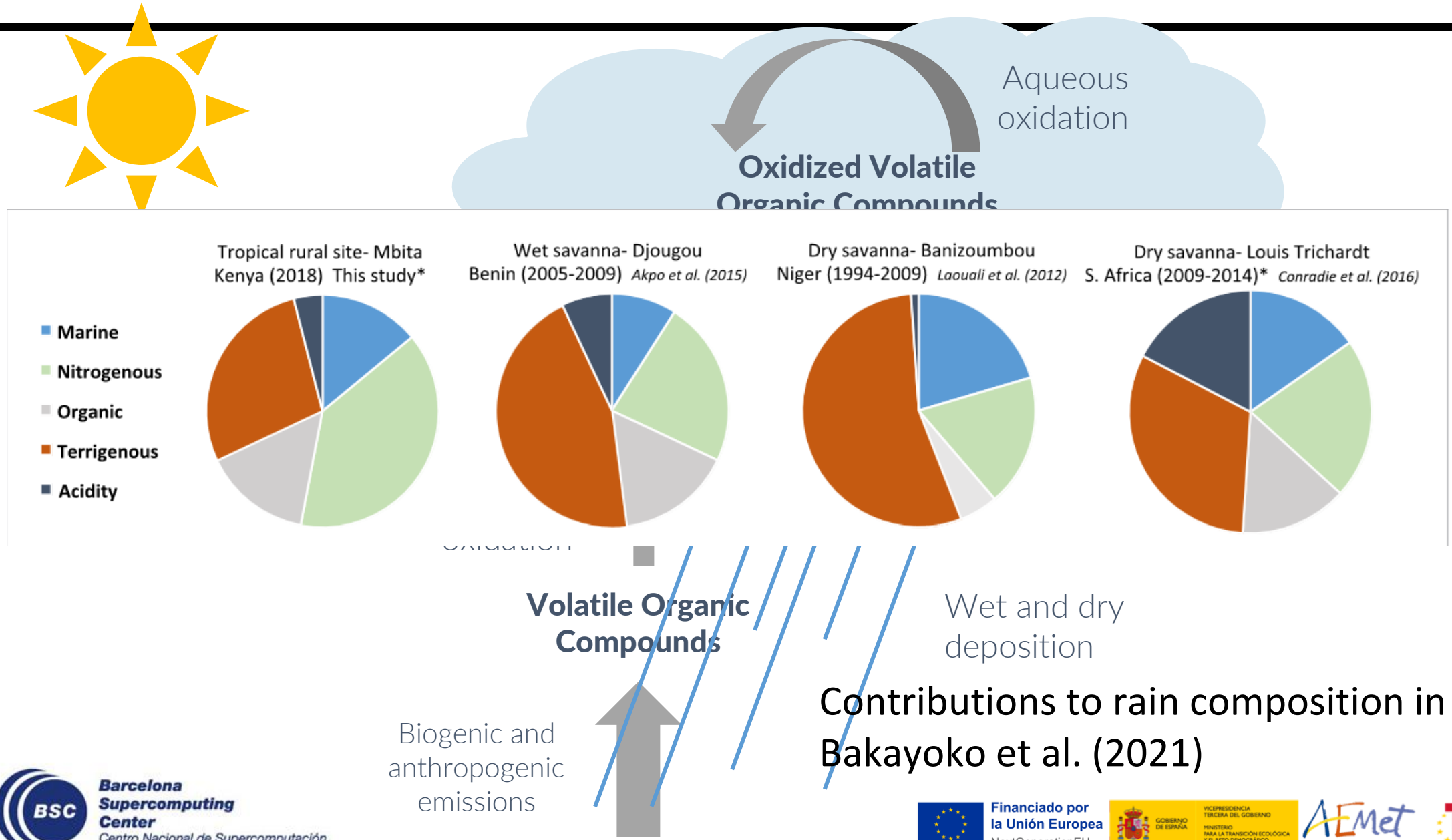
Atmospheric Evolution of Organics : formation of oxidized compounds



Measured PM1 composition Jimenez et al. (2009)

emissions

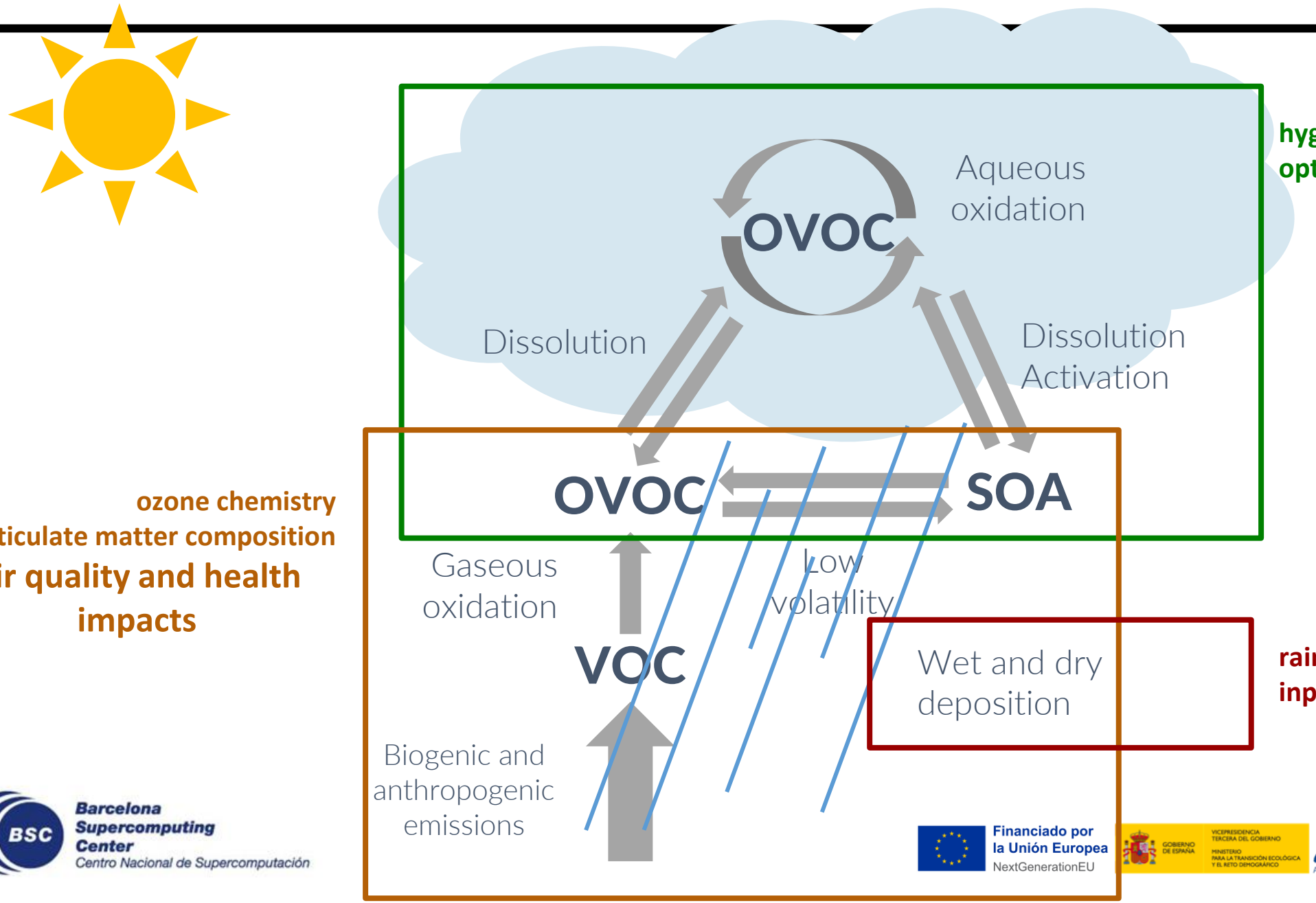
Atmospheric Evolution of Organics : formation of oxidized compounds



Contributions to rain composition in the tropics
Bakayoko et al. (2021)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Impacts



hygroscopicity
optical properties
Climate impacts

ozone chemistry
particulate matter composition
Air quality and health impacts

rain composition
inputs to soils
Ecosystems impacts

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

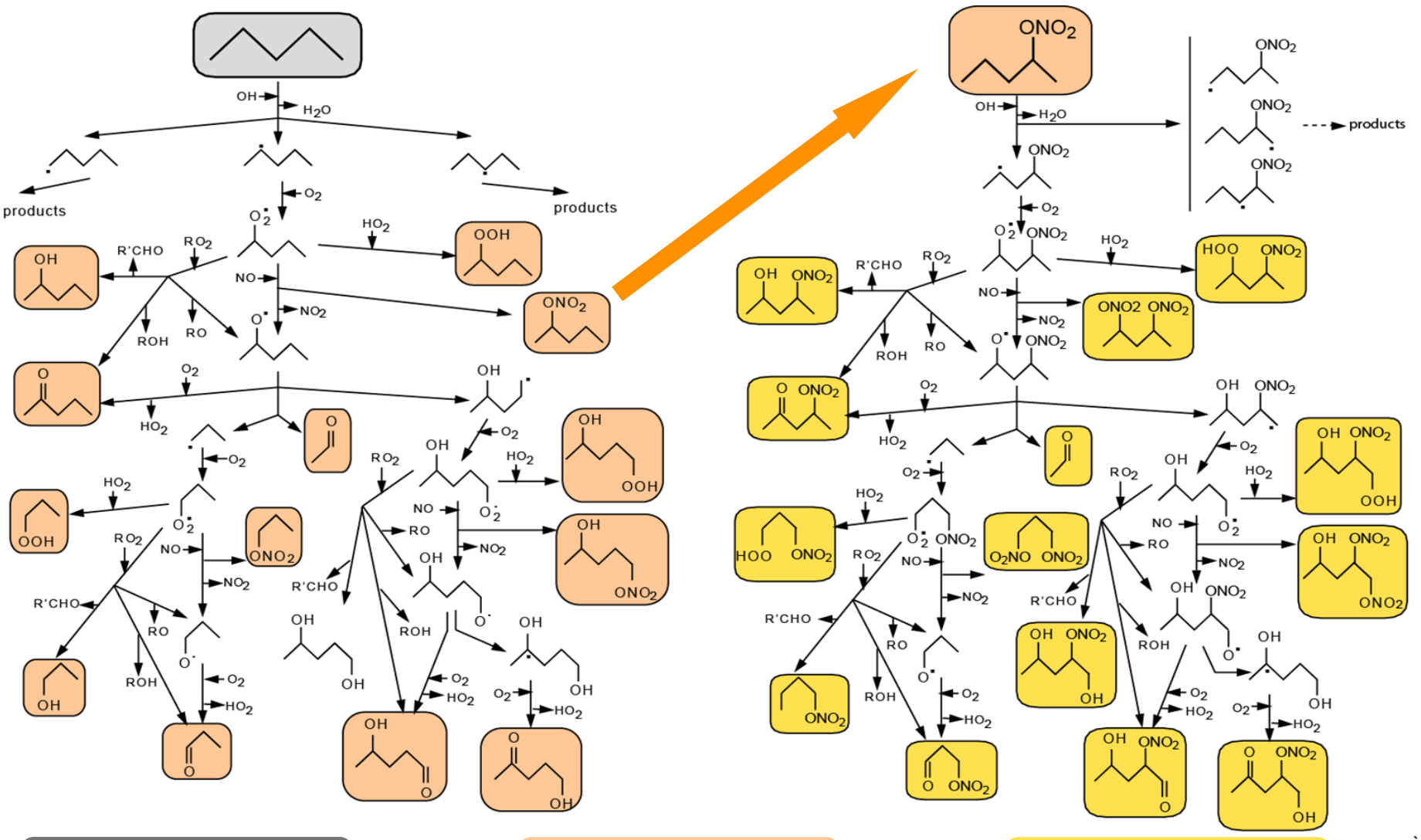
Financiado por la Unión Europea
NextGenerationEU

Vicepresidencia Tercera del Gobierno
Ministerio para la Transición Ecológica y el Reto Demográfico

AEMet
Agencia Estatal de Meteorología

Plan de Recuperación, Transformación y Resiliencia

Progressive and complex oxidation of organics



Oxidized Volatile Organic Compounds

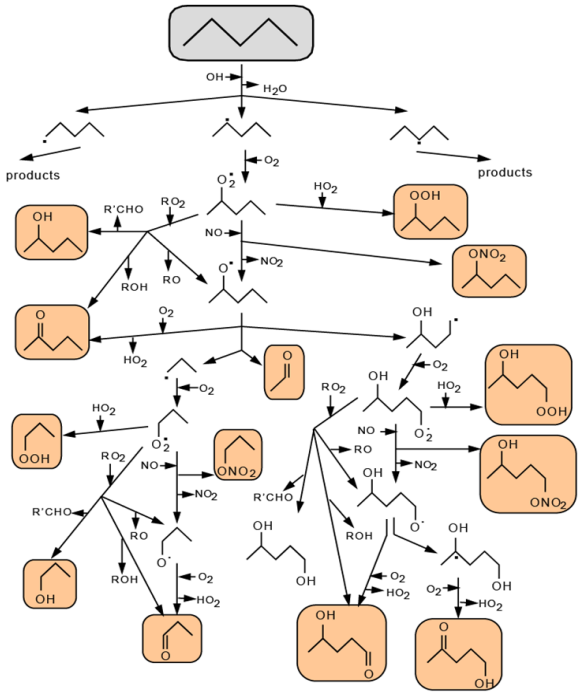
Gaseous oxidation

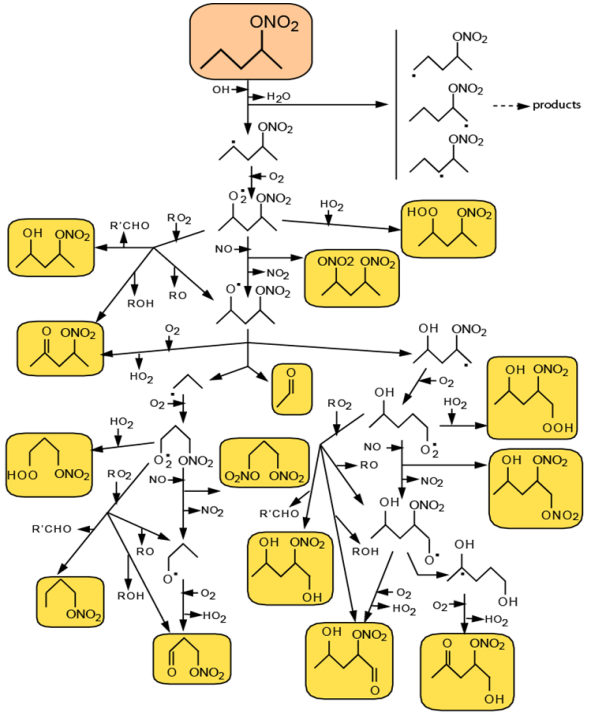
Volatile Organic Compounds

How to bring this complexity to air quality models?

Parent Hydrocarbon → **1st Generation Species** → **2nd Generation Species** → CO_2

How I Learned to Stop Worrying and Love the Complexity





Master Chemical Mechanism (Jenkin et al., 1997; 15yrs 2-3people ?)

- 143 precursors from C_1 to C_{12}
- 17000 reactions of 6700 species
- Simplified!

CLEPS (Mouchel-Vallon et al., 2017; 2yrs 2people)

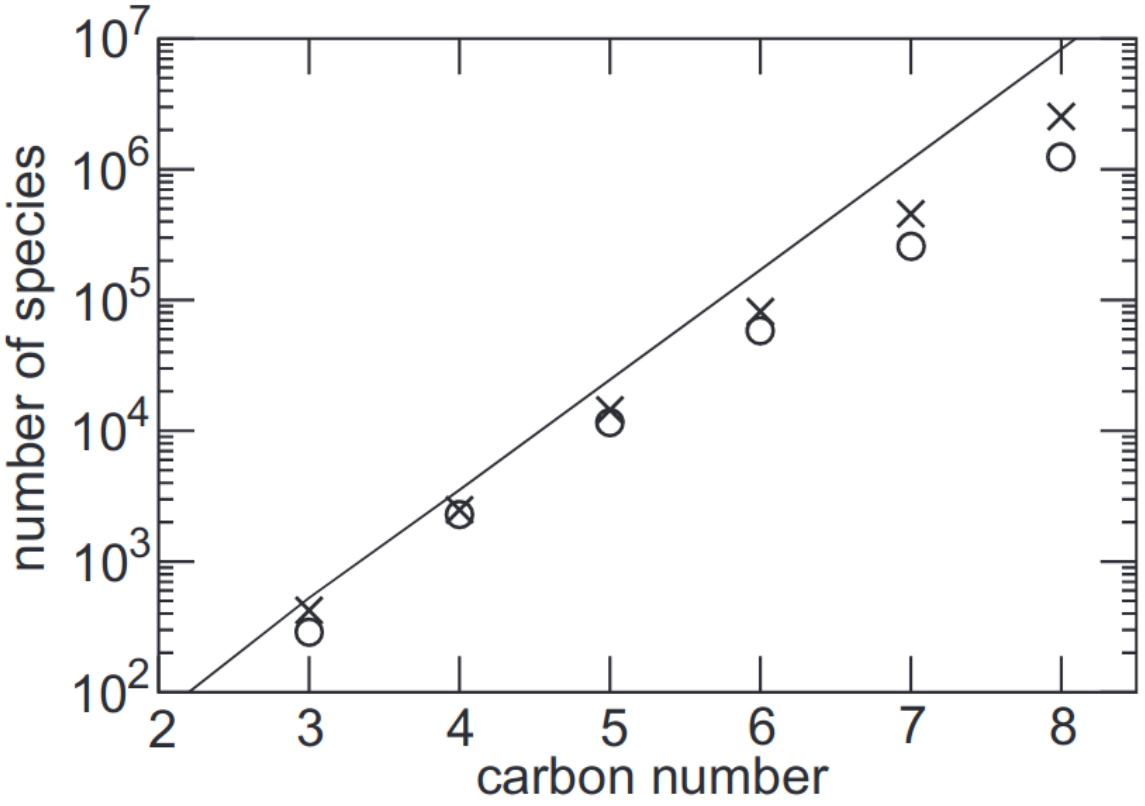
- Cloud oxidation of C_1-C_4 products from isoprene oxidation
- 1315 reactions involving 717 chemical species
- Simplified!

Handwritten detailed mechanisms: high potential for errors and time consuming (creation and maintenance)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
AEMet Agencia Estatal de Meteorología

How I Learned to Stop Worrying and Love the Complexity



Potential number of distinct species in a full oxidation mechanism as a function of the size of the precursor

Aumont et al. (2005)

Master Chemical Mechanism (Jenkin et al., 1997; 15yrs 2-3people ?)

- 143 precursors from C_1 to C_{12}
- 17000 reactions of 6700 species
- Simplified!

CLEPS (Mouchel-Vallon et al., 2017; 2yrs 2people)

- Cloud oxidation of C_1-C_4 products from isoprene oxidation
- 1315 reactions involving 717 chemical species
- Simplified!

Handwritten detailed mechanisms: high potential for errors and time consuming (creation and maintenance)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

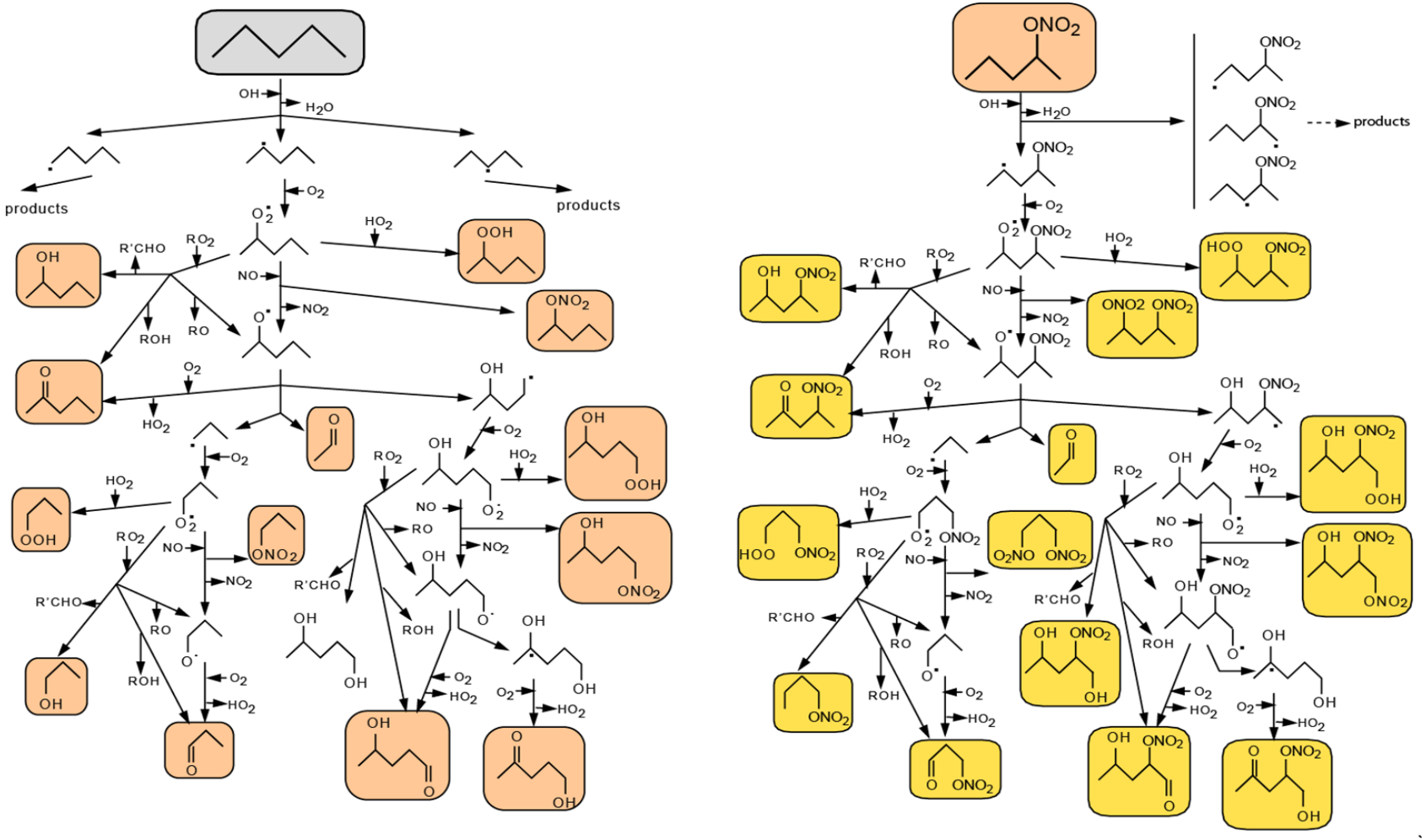
We need to go faster

Writing the chemical mechanism

Solving the ODEs in a (3D) model

Financiado por la Unión Europea
NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Automating boring things



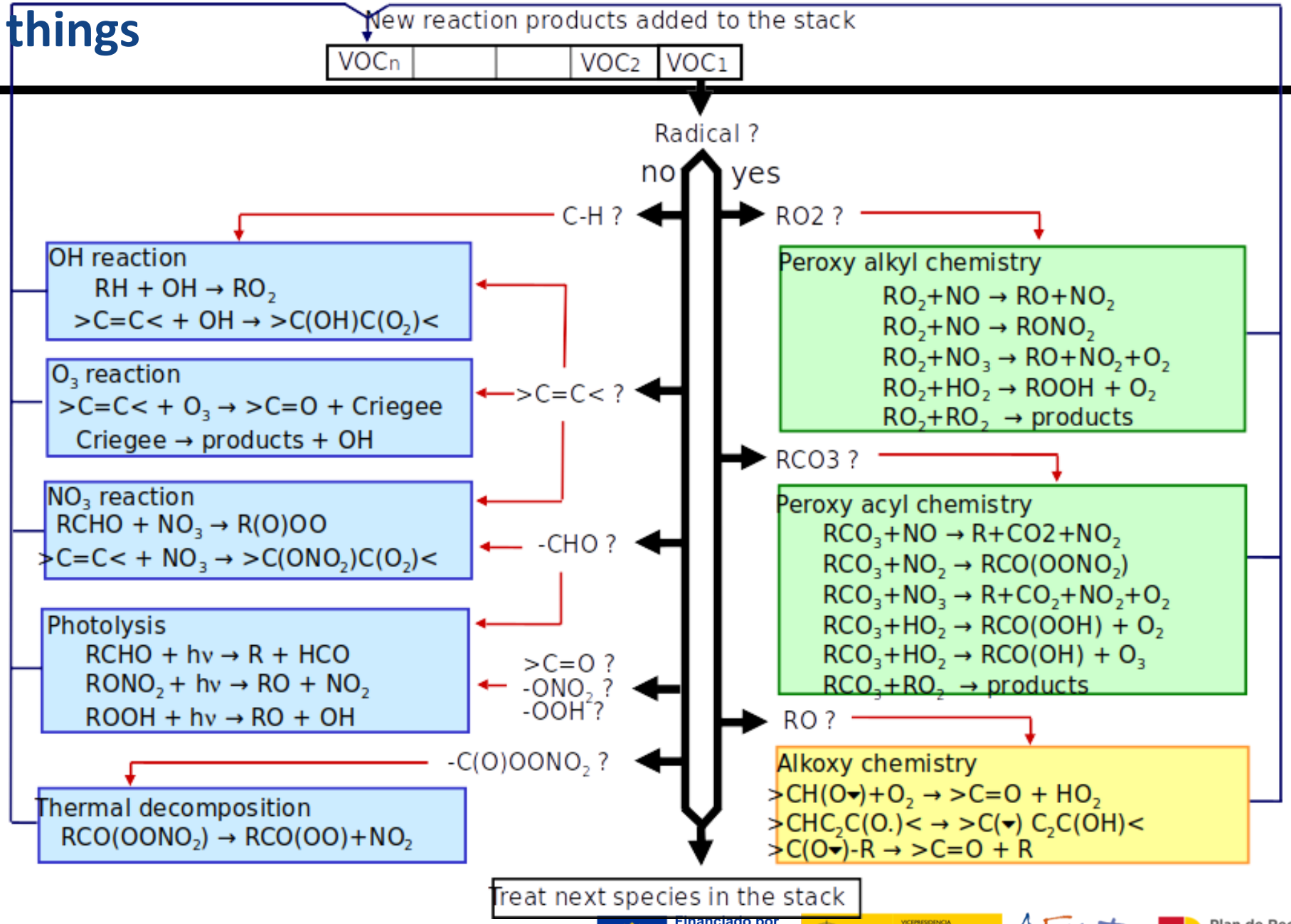
Systematic and repeated oxidation steps

Generator for Explicit Chemistry and Kinetics of Organics in the Atmosphere (GECKO-A, Aumont et al., 2005)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea
NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Automating boring things

Aumont et al., 2005



Accelerating things

Writing the chemical mechanism

GECKO-A

12 biogenic, 53 anthropogenic precursors

23 million reactions involving 4.4 million species

(Mouchel-Vallon et al., 2020)

Solving the ODEs in a (3D) model

Simulating 2 days in 2 grid cells

≈ 36 hours on 16 cores

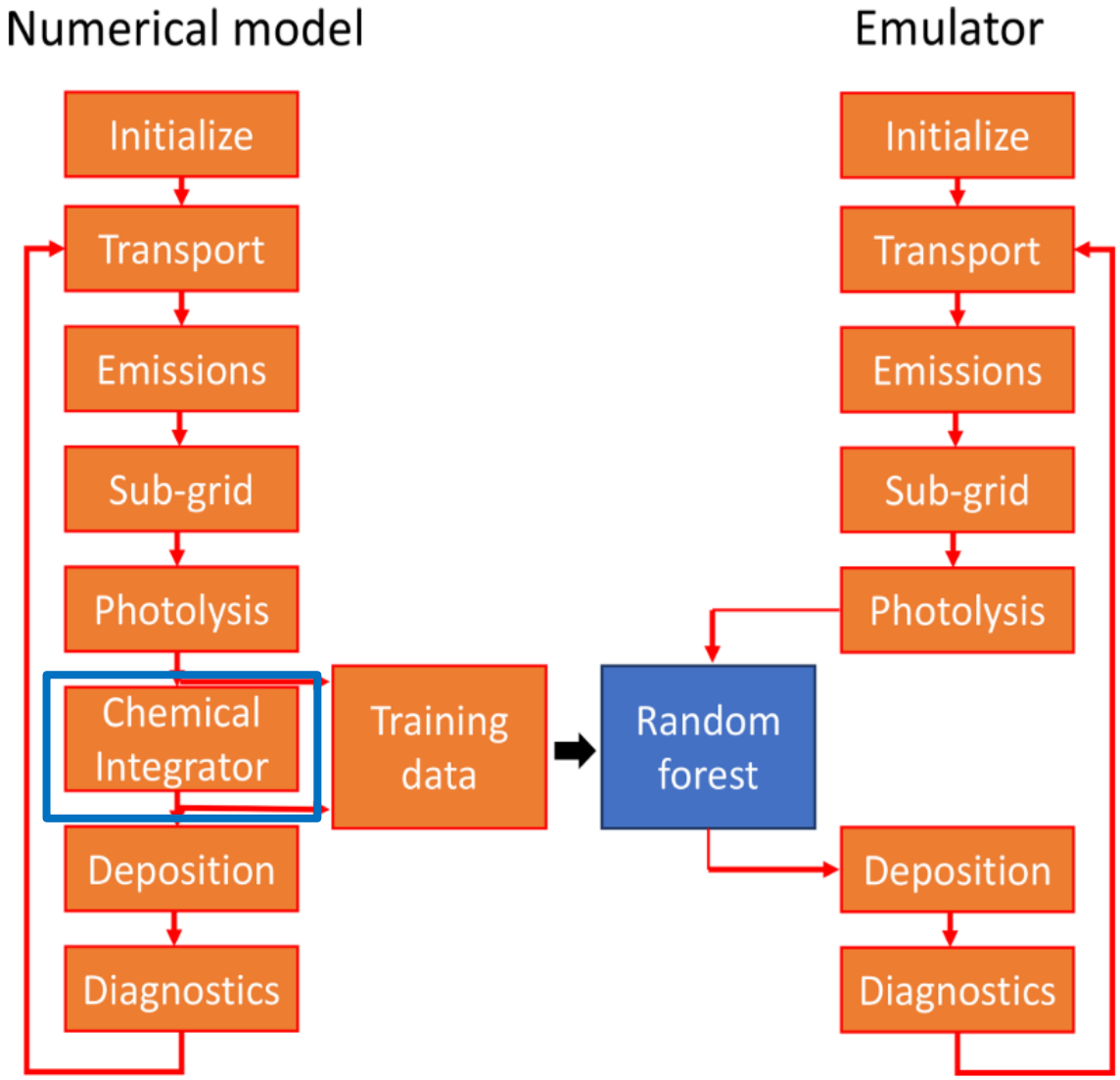
Machine learning?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Financiado por la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

Emulating atmospheric chemistry



Computationally expensive

Use GECKO-A 0D explicit simulations as training dataset

Emulate the behavior of the detailed model with machine learning

- **Neural Networks (NN)**
- **Random Forest (RF)**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

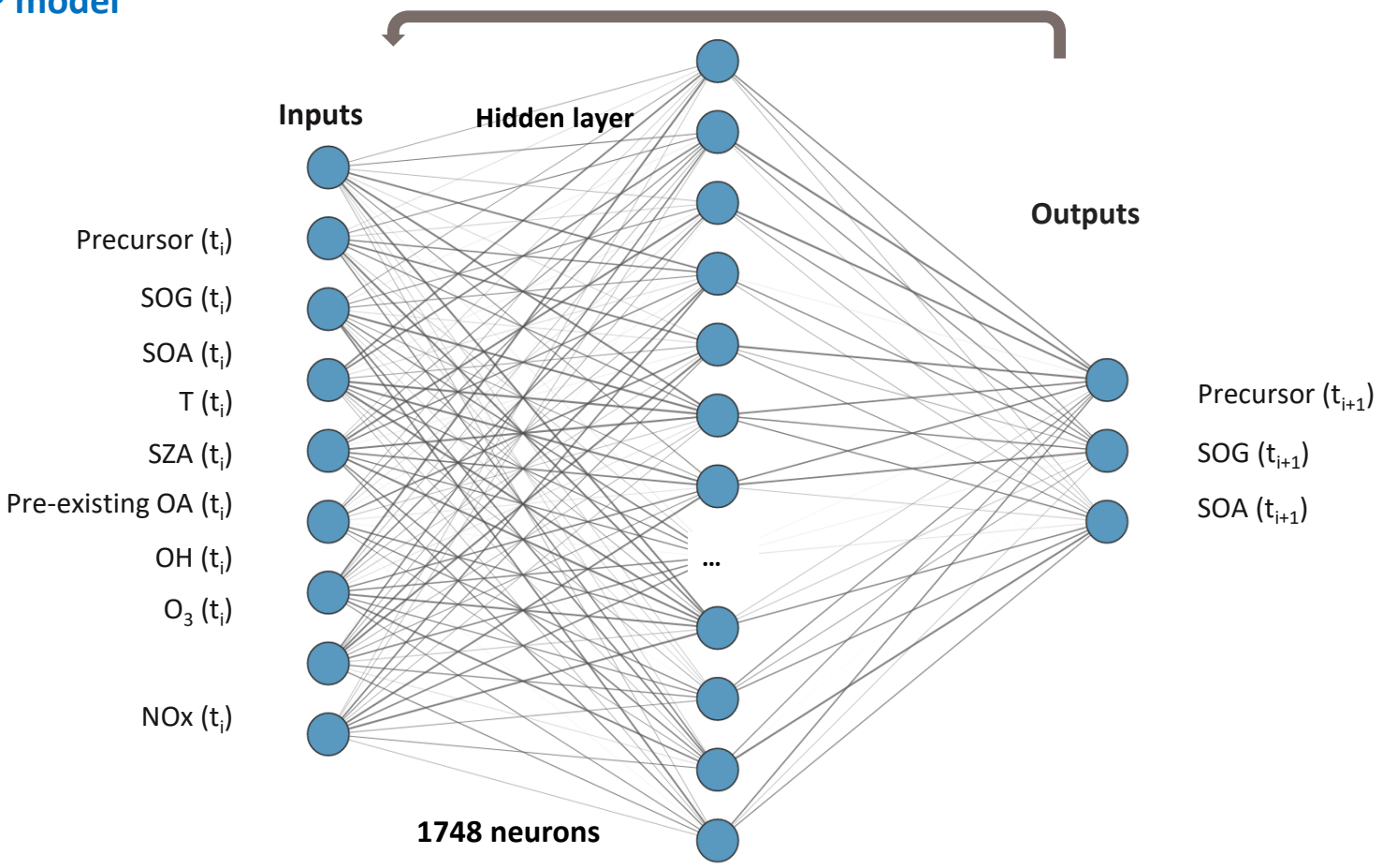
Financiado por la Unión Europea
NextGenerationEU

Agencia Estatal de Meteorología

Plan de Recuperación, Transformación y Resiliencia

NN approach for predicting time-series of concentrations

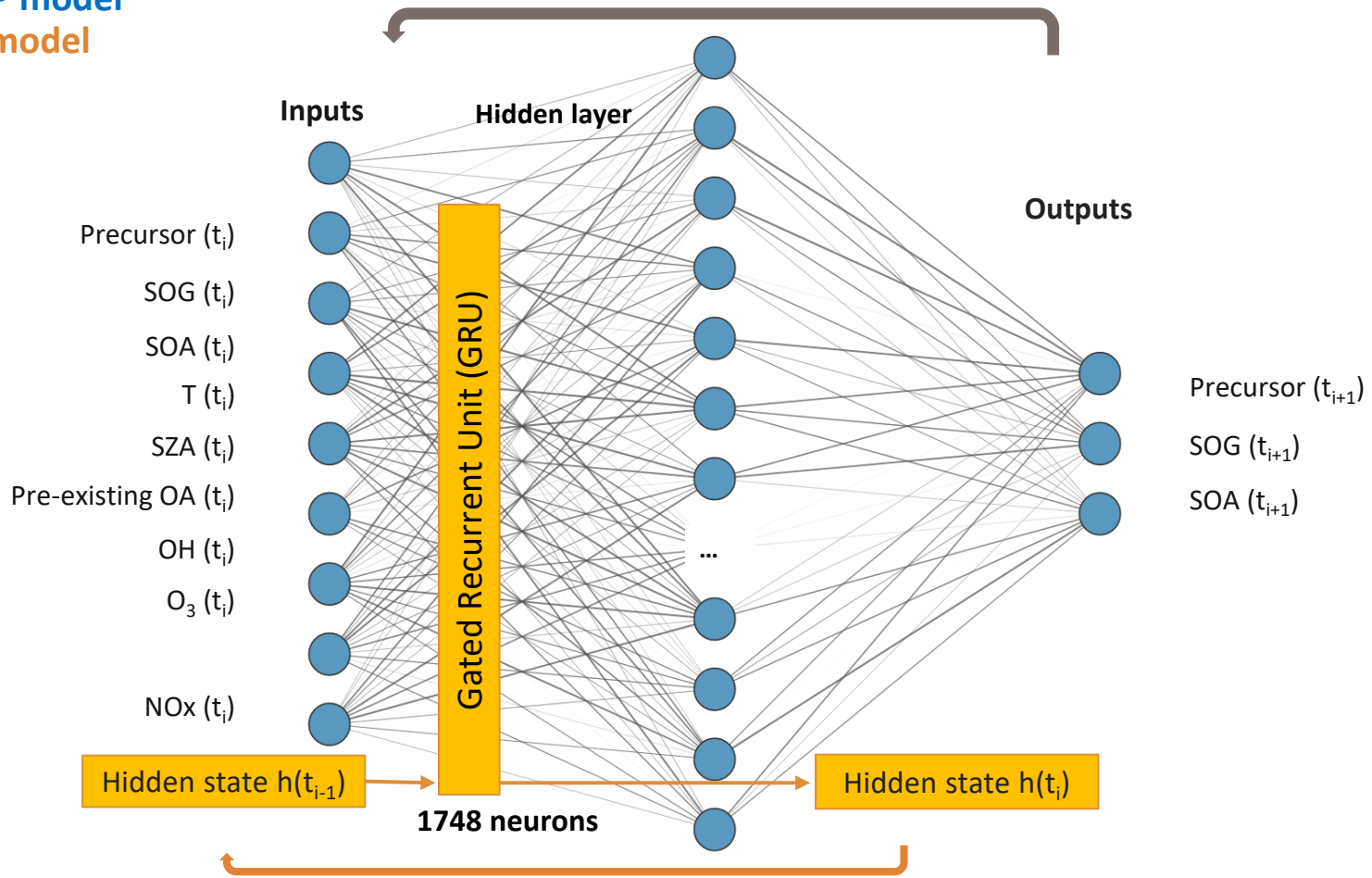
1. Multi-layer perceptron MLP model



Schreck et al. JAMES 2022: Neural network emulation of the formation of OA based on the explicit GECKO-A chemistry model

NN approach for predicting time-series of concentrations

1. Multi-layer perceptron MLP model
2. Gated-recurrent unit GRU model



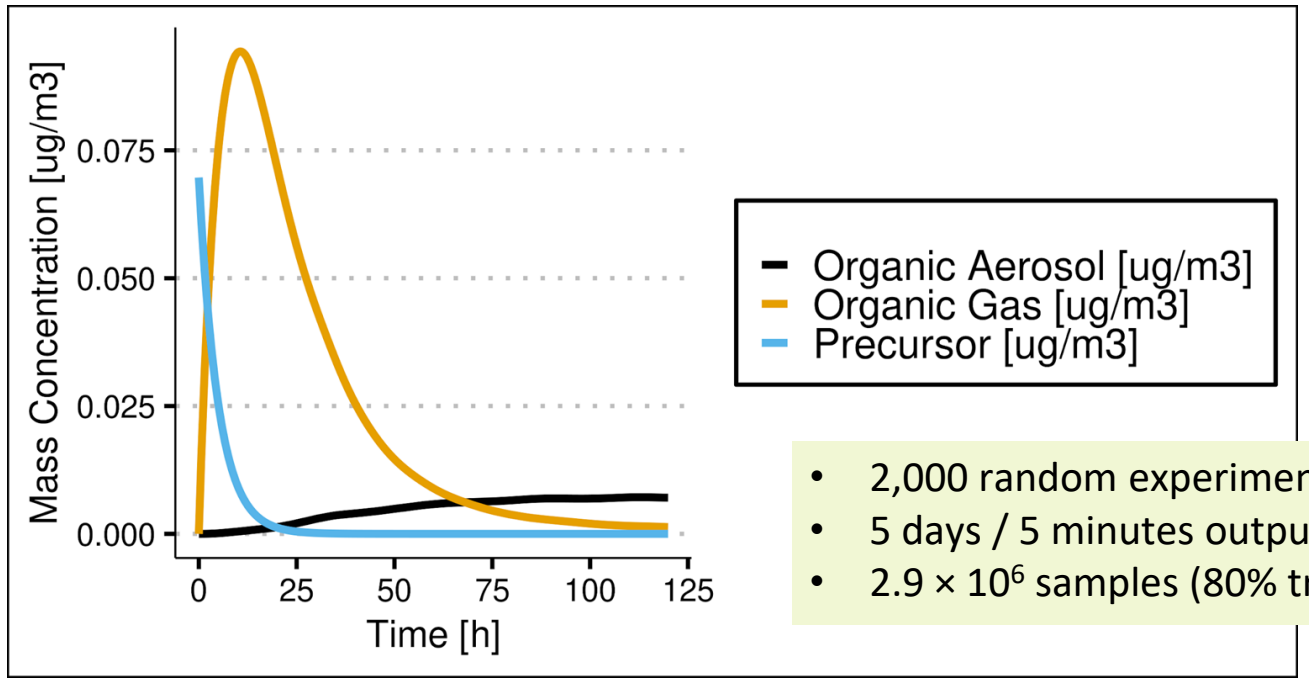
Schreck et al. JAMES 2022: Neural network emulation of the formation of OA based on the explicit GECKO-A chemistry model

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Generate Neural Network training dataset with GECKO-A

- 3 precursors: **toluene**, **dodecane**, **a-pinene**
- Random environmental conditions
- No diurnal variations
- Initial precursor amounts of 10 ppt, 0.1 and 1 ppb

Temperature	240 - 320 K	Uniform
Solar zenith angle (SZA)	0-90 degrees	Uniform
Pre-existing aerosols	0.01-10 $\mu g/m^3$	Logarithmic
Ozone	1 - 150 ppb	Uniform
NOx	0.01-10 ppb	Logarithmic
OH	10^1–10^6 molecules/cm^3	Uniform

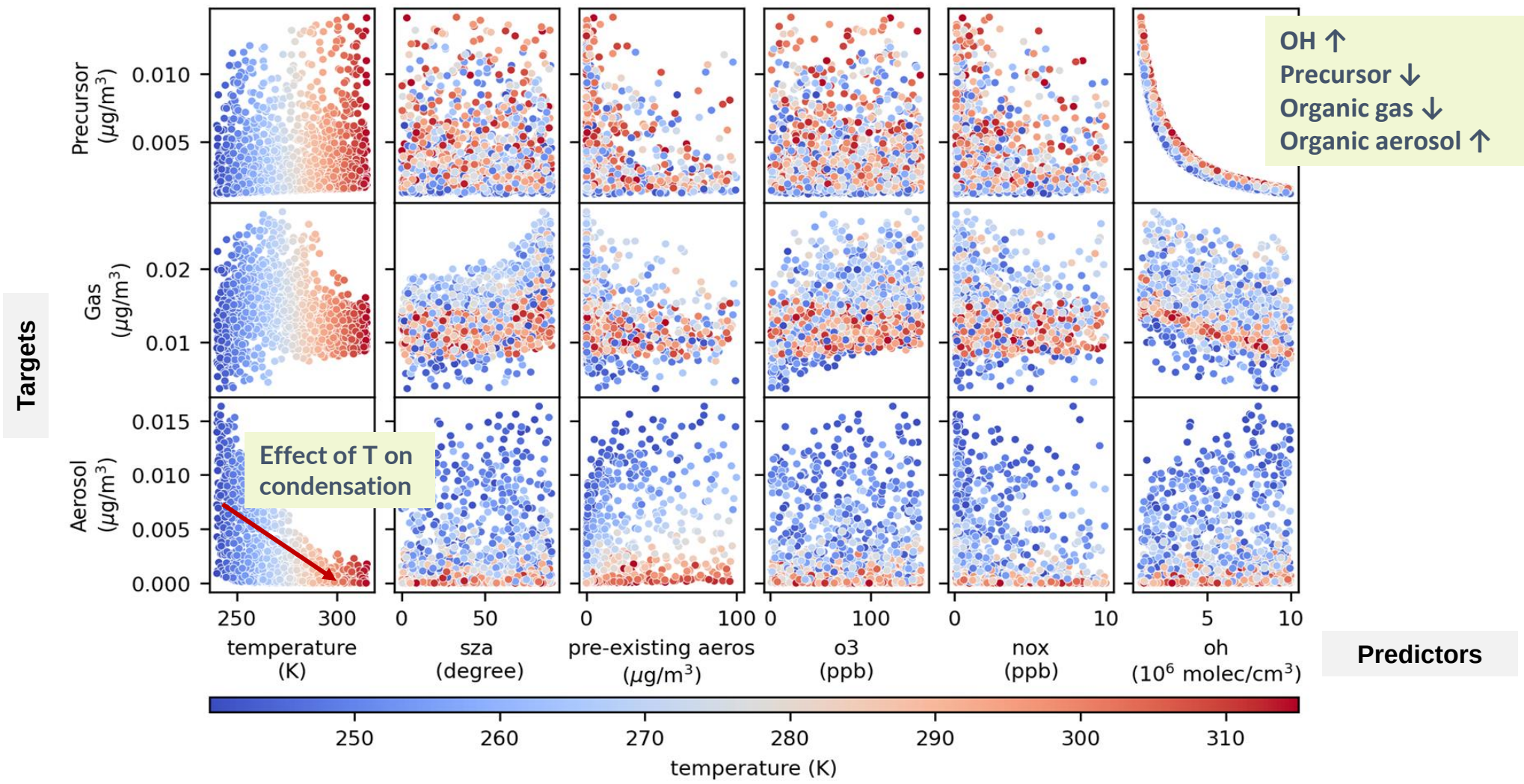


- 2,000 random experiments for each precursor
- 5 days / 5 minutes output
- 2.9×10^6 samples (80% training, 20% validation)

Schreck et al., JAMES, 2022

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Neural Network training dataset



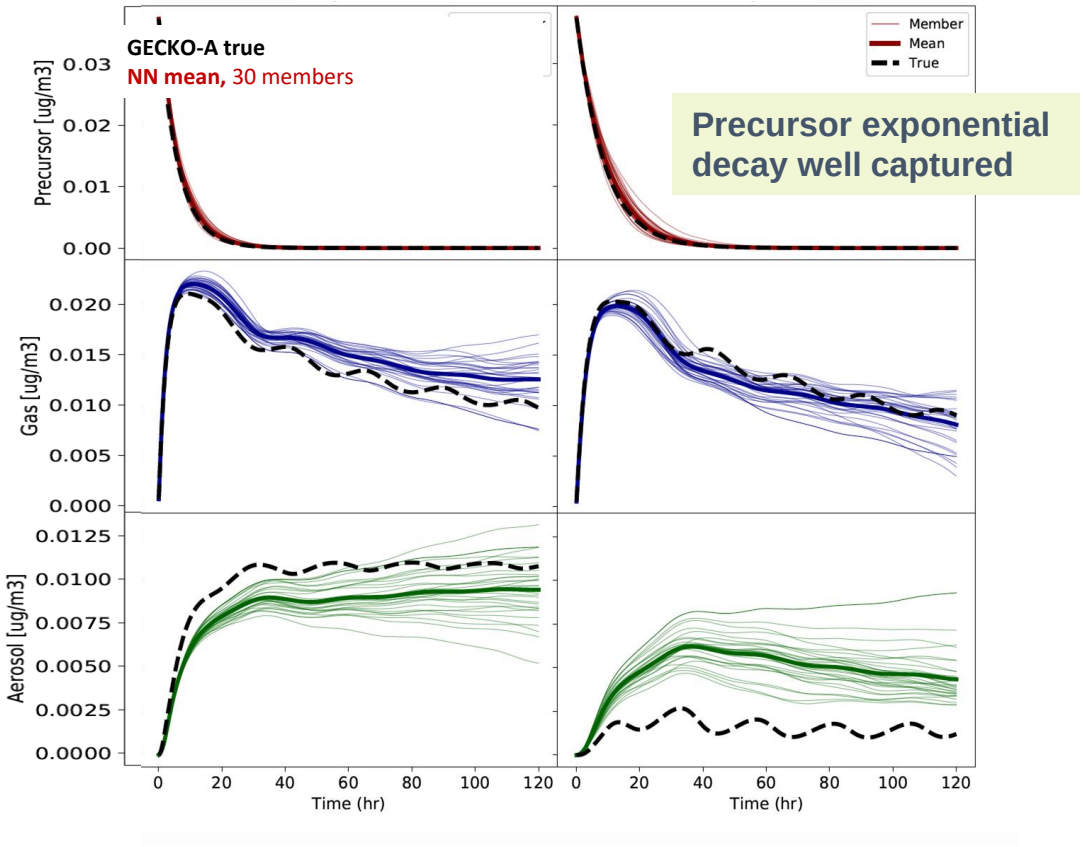
Schreck et al., JAMES, 2022

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
AEMet Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Multi-Layer Perceptron vs. Gated-Recurrent Unit network

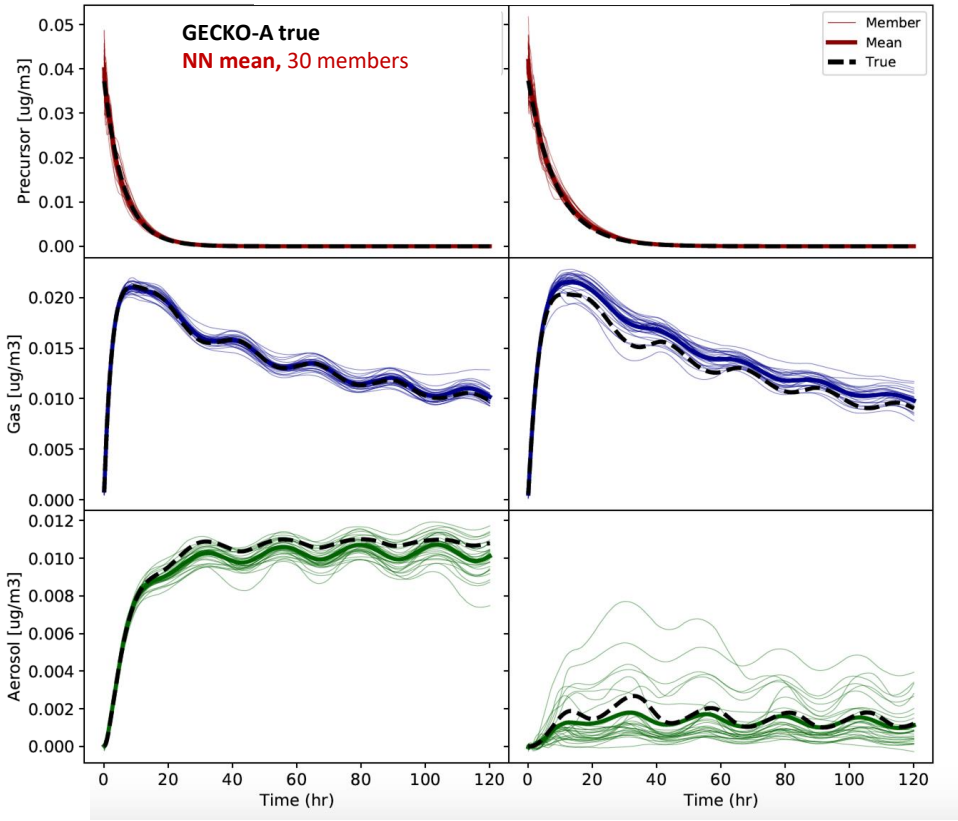
30-ensemble member predictions for Toluene

Multi-layer perceptron



Precursor exponential decay well captured

Gated-recurrent unit

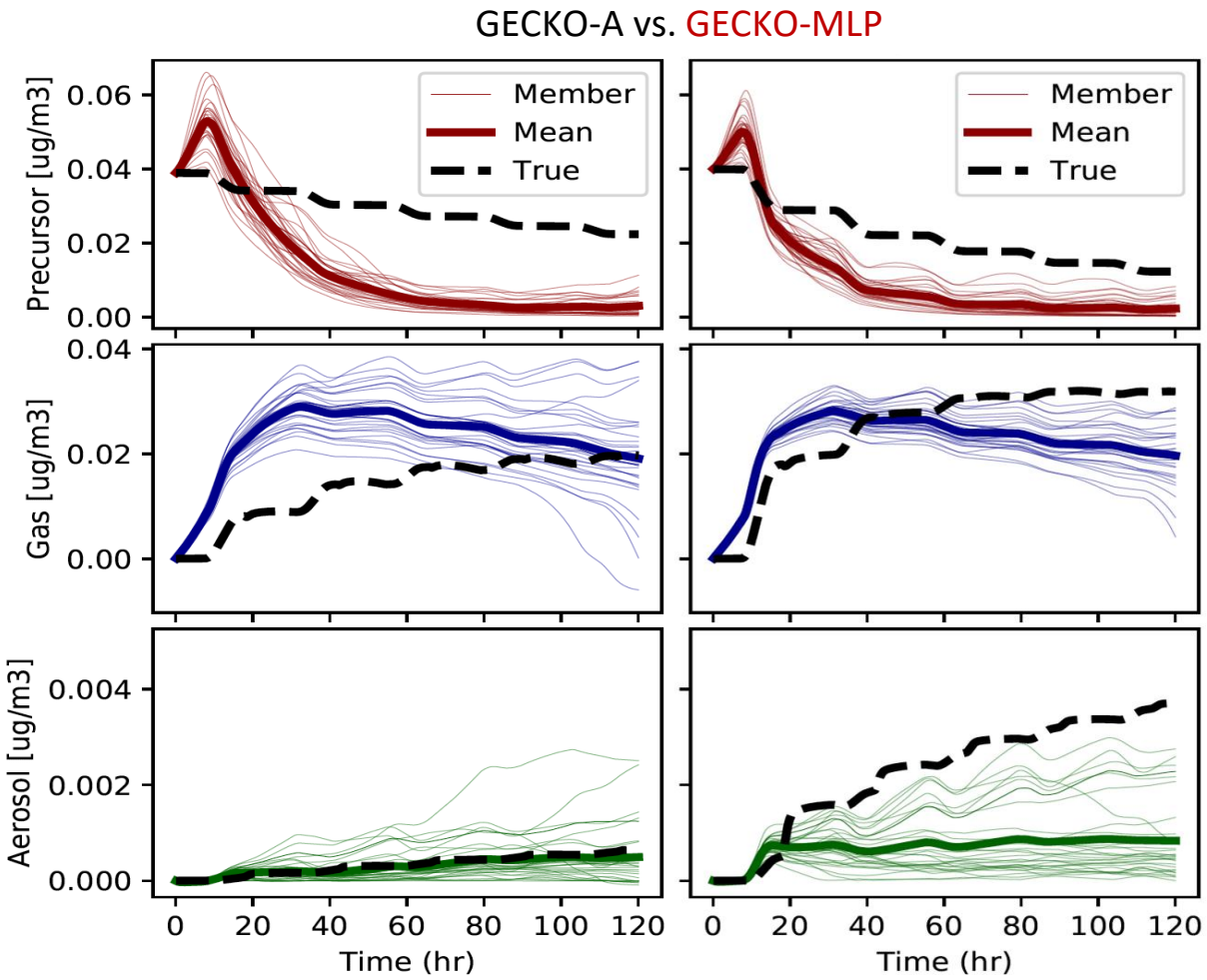


⇒GRU performs better but is challenging to implement in a 3D model

Schreck et al., JAMES, 2022

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NSF NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Vicepresidencia Tercera del Gobierno · Ministerio para la Transición Ecológica y el Reto Demográfico
AEMet Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Application in a box model for diurnally varying conditions



⇒ Neural Networks trained on datasets built under constant conditions cannot reproduce realistic diurnal cycles

⇒ Raises the question of the representativeness of the training data sets

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Computational gain GECKO-A vs. GECKO-NN

- For Toluene, GECKO-NN is 4×10^2 times faster than GECKO-A on CPU, and 10^4 times faster on GPU

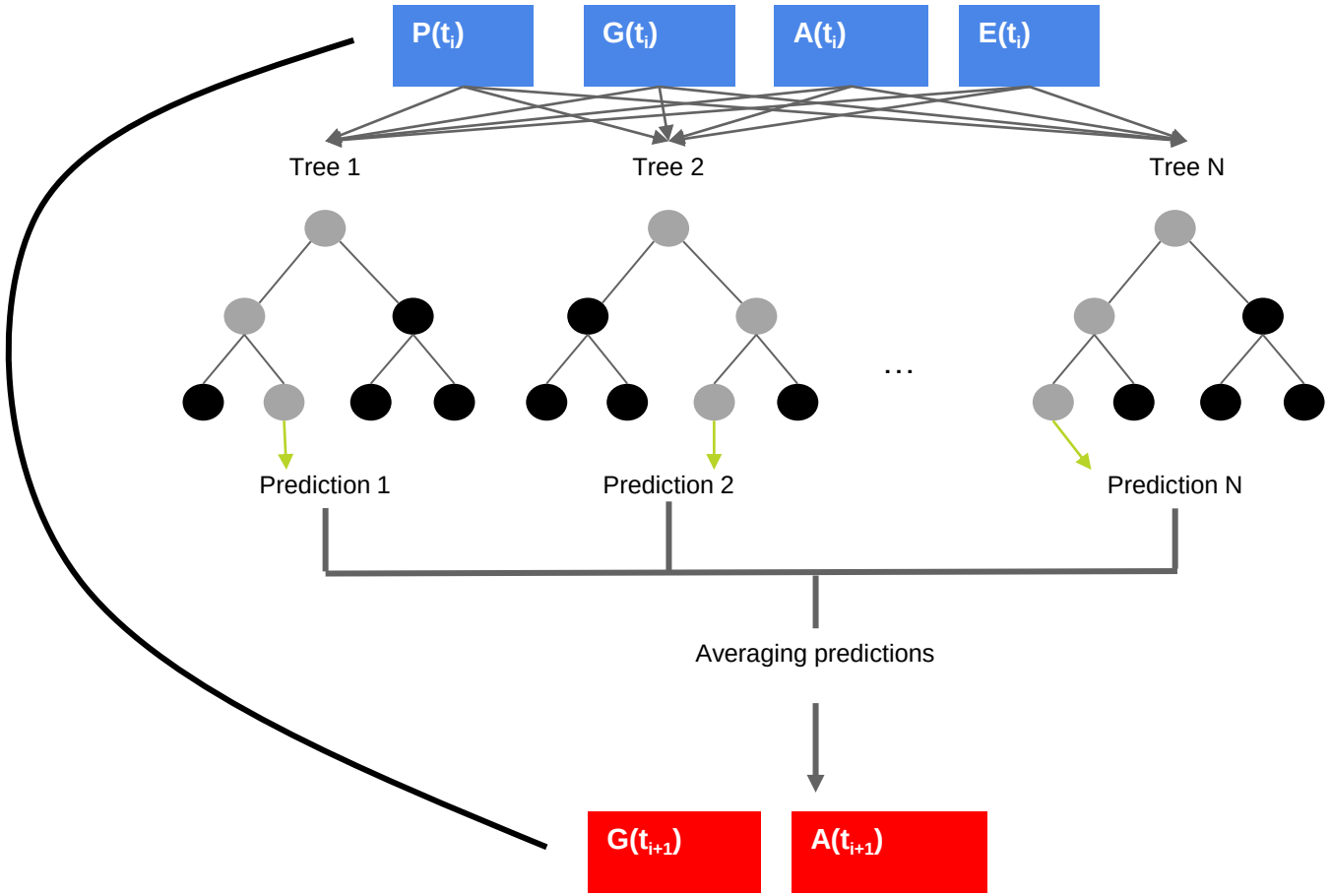
Model	Toluene		Dodecane		α-pinene	
GECKO-A	0.9 s	1	71 s	1	220 s	1
MLP CPU	2.1 µs	430	0.8 µs	8.88×10^4	1.6 µs	1.38×10^5
MLP GPU	0.08 µs	11250	0.07 µs	1.01×10^6	0.08 µs	2.75×10^6
GRU CPU	3.1 µs	290	3.2 µs	2.22×10^4	3.3 µs	6.67×10^4
GRU GPU	0.38 µs	2368	0.38 µs	1.87×10^5	0.38 µs	5.79×10^5

Ratio of GECKO-A / GECKO-NN

Timing per 5min timestep

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest approach



Inputs :

- Precursor concentration
- Gas organic concentration
- Aerosol organic concentration
- Temperature
- Solar zenith angle
- Pre-existing aerosol
- O_3
- NOx
- OH

$P(t_i)$ $G(t_i)$ $A(t_i)$ $E(t_i)$

Output :

- Gas organic concentration
- Aerosol organic concentration

$G(t_{i+1})$ $A(t_{i+1})$

Mouchel-Vallon & Hodzic, JGR, 2023:
Toward emulating an explicit organic chemistry mechanism with random forest models.

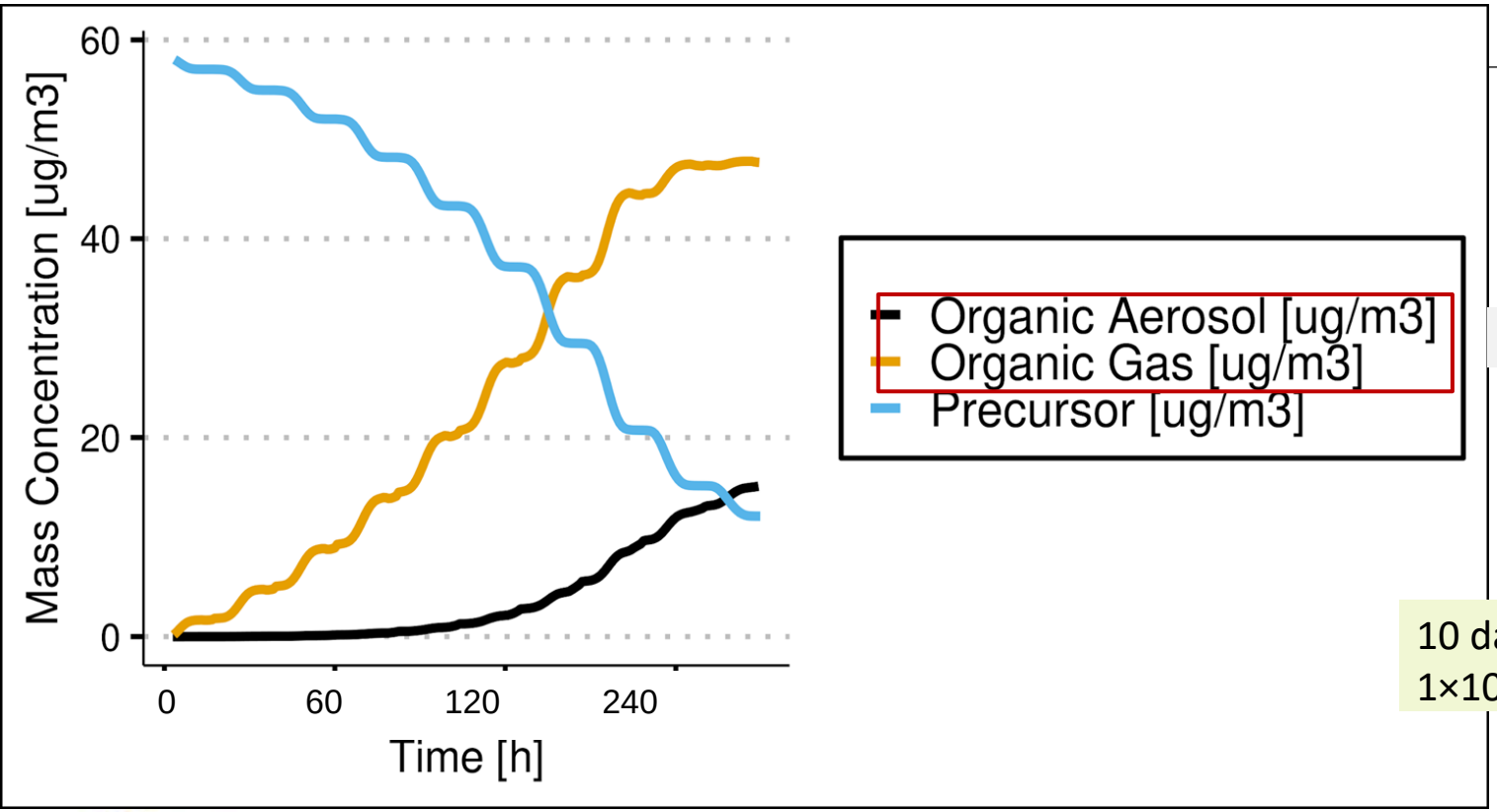
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest: Realistic training dataset

- 2 precursors: **toluene**, **dodecane**
- Random environmental conditions
- With diurnal variations
- Random initial precursor concentration

Parameter	Range	Parameter	Range
Latitude [°]	80S–80N	Relative humidity [%]	3–102
Temperature [K]	216–313	Atmospheric pressure [atm]	0.5–1.02
Preexisting aerosol seed [μg m^{-3}]	0.03–340	Initial NO_x [ppb]	10^{-4}–42
Initial precursor [ppb]	0–16	Initial CO	33–1012
Initial O_3 [ppb]	1–100	NO Emission [molec cm^{-2} s^{-1}]	10^{7}–10^{9}

From 3D model tropospheric ranges

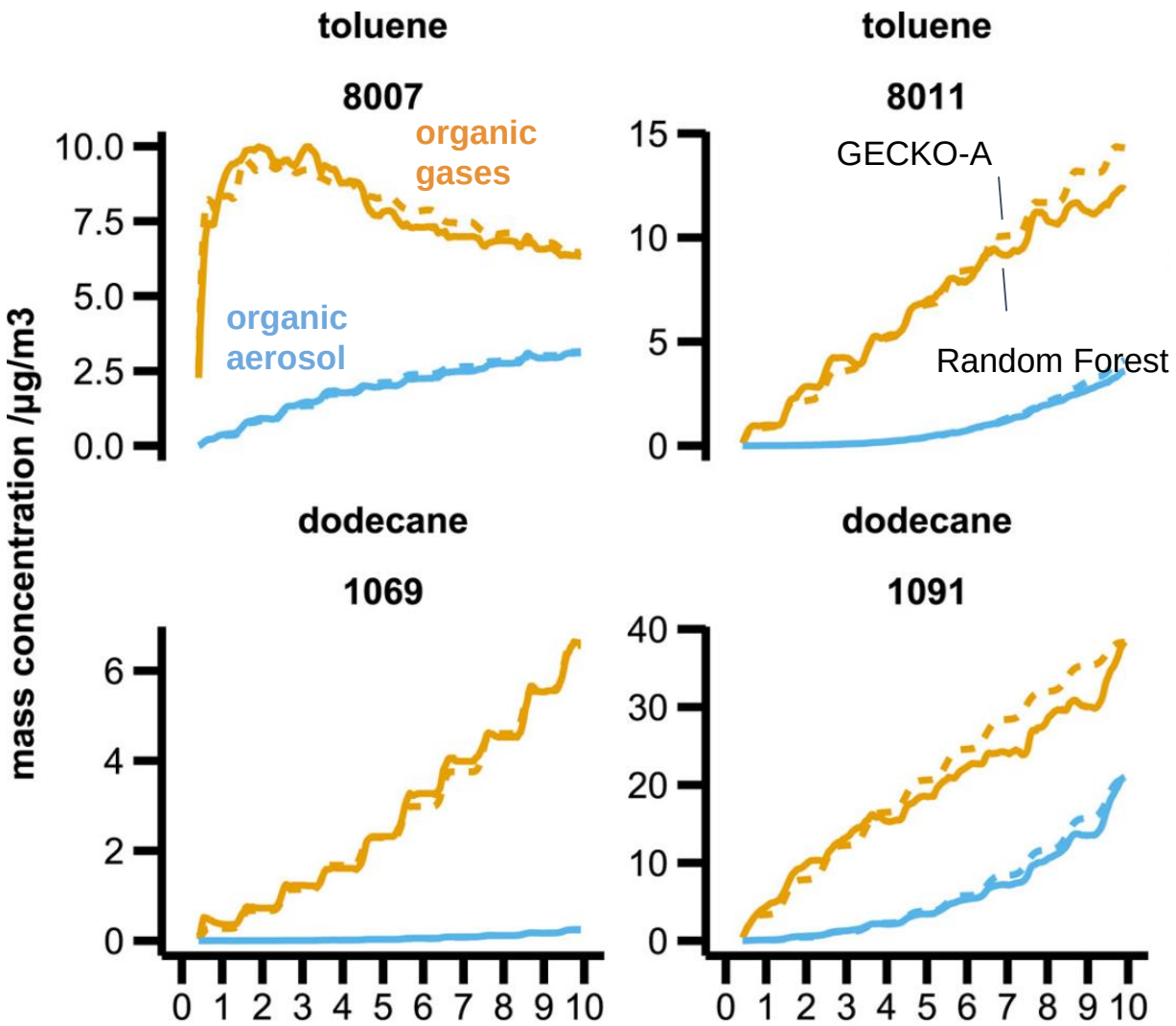


Only 2 targets

10 days / 5 min output
1×10^6 samples/precursor

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NSF NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

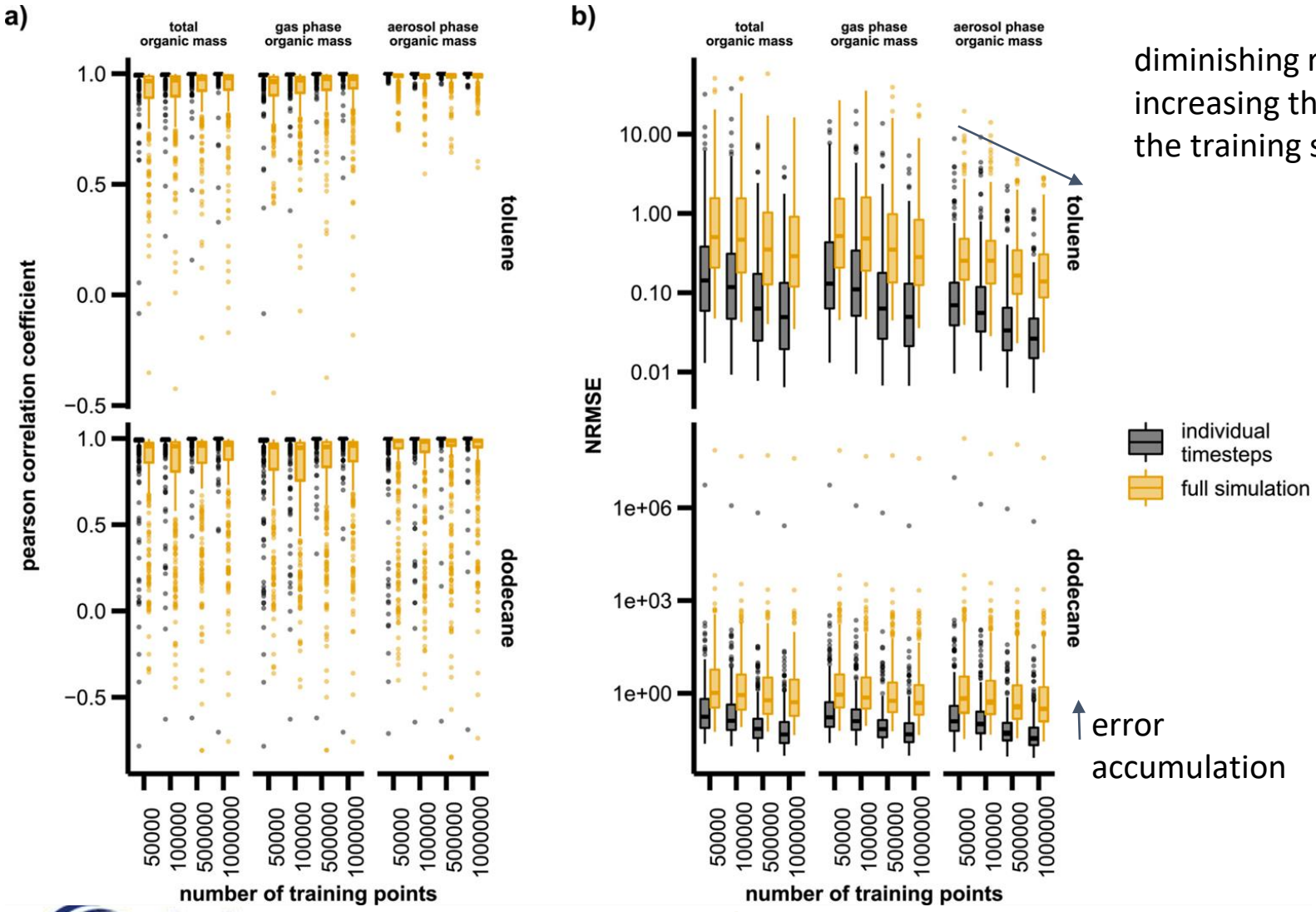
Random Forest results: examples



RF perform similarly to NN
The training set allows reproducing the diurnal cycles

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
AEMet Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest results: performance



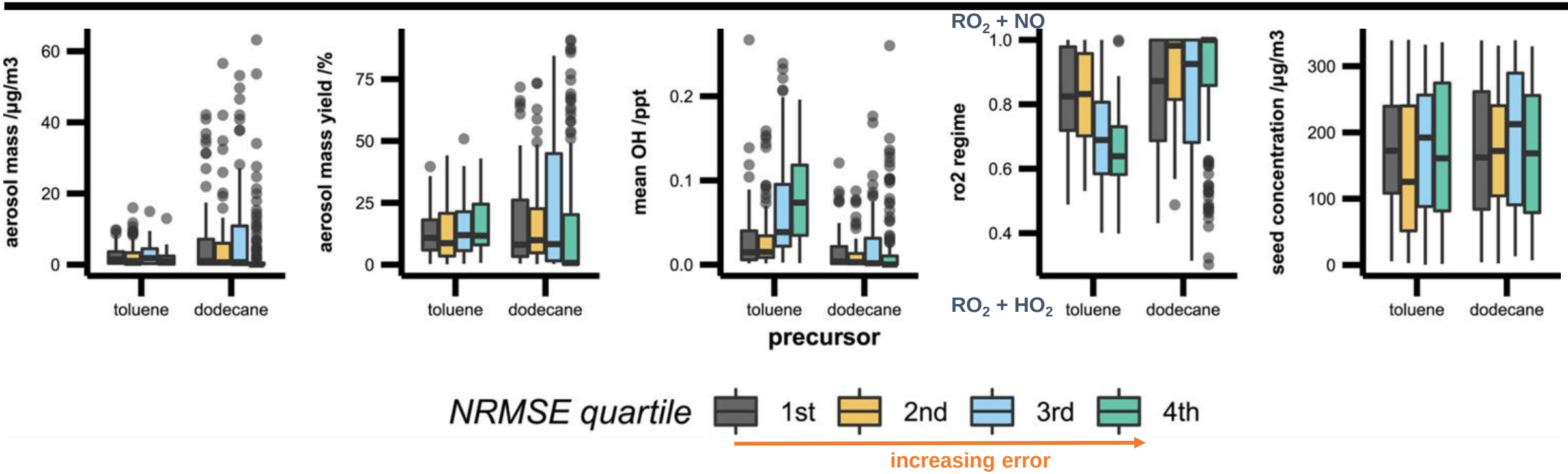
diminishing returns of increasing the size of the training sample

error accumulation

Increasing the number of training samples improves performance

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest results: errors distribution

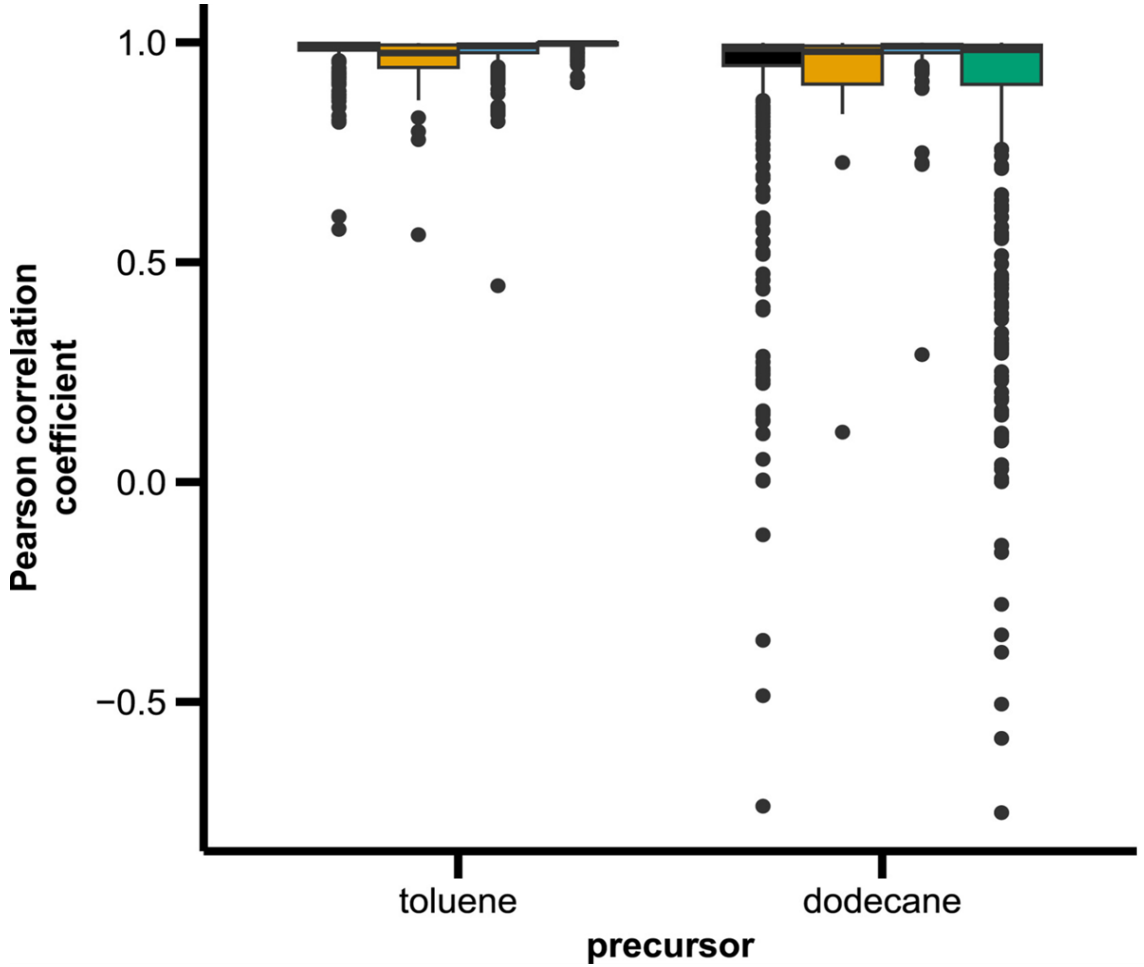


Little sensitivity of errors to the aerosol quantity to predict (seed, yield, mass)
Errors sensitive to NOx regime and OH mixing ratios: underrepresented regimes in training set

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest results: specializing random forests

Do performances improve if the range of chemical regimes to predict is reduced?



random forest

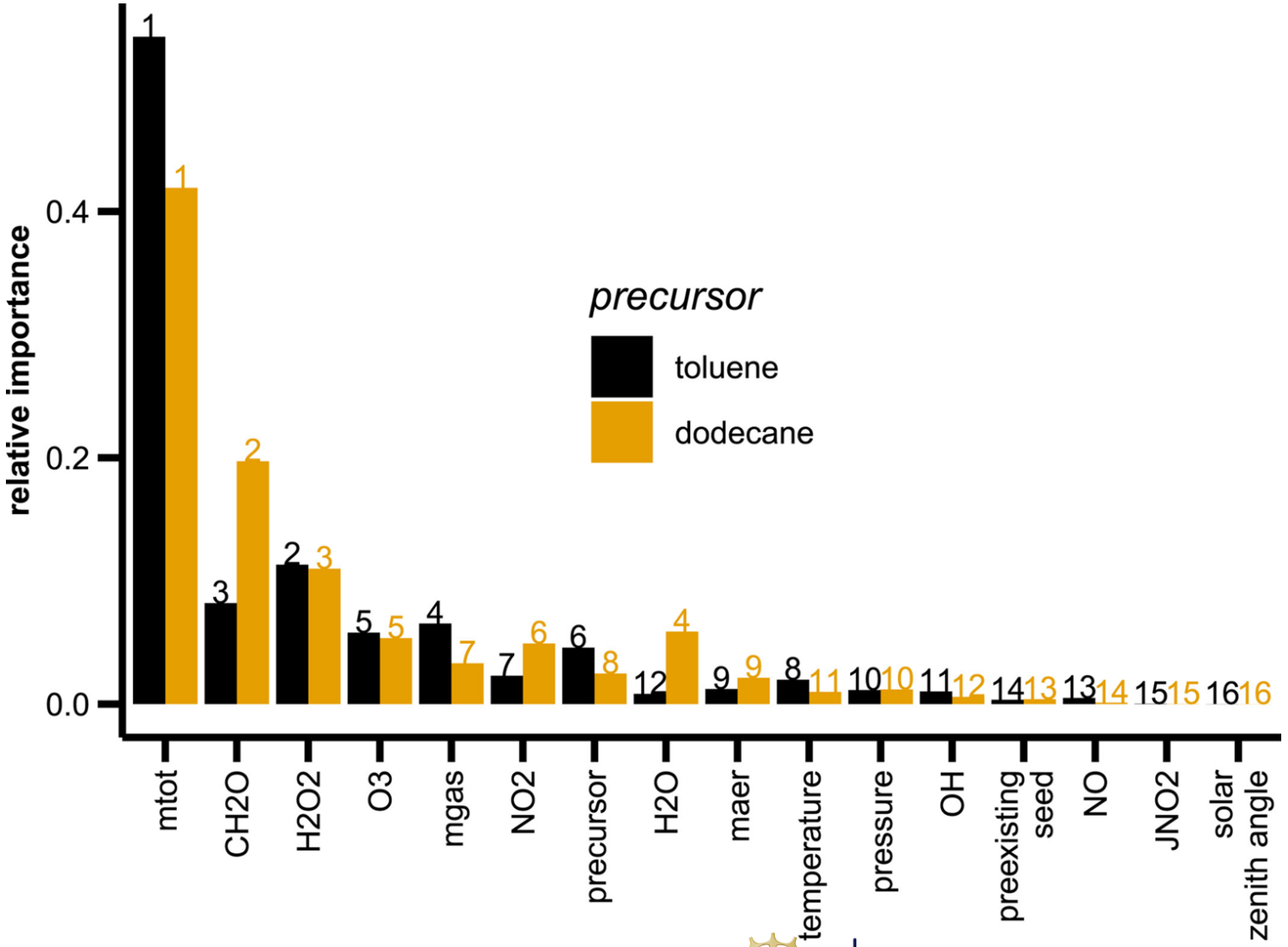
- original
- low nox
- mid nox
- high nox

# training points	
toluene	dodecane
1×10^5	7×10^4
6×10^5	3×10^5
4×10^5	7×10^5

Specialization requires **more** or **better constructed** training data

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest results: predictors relative importance



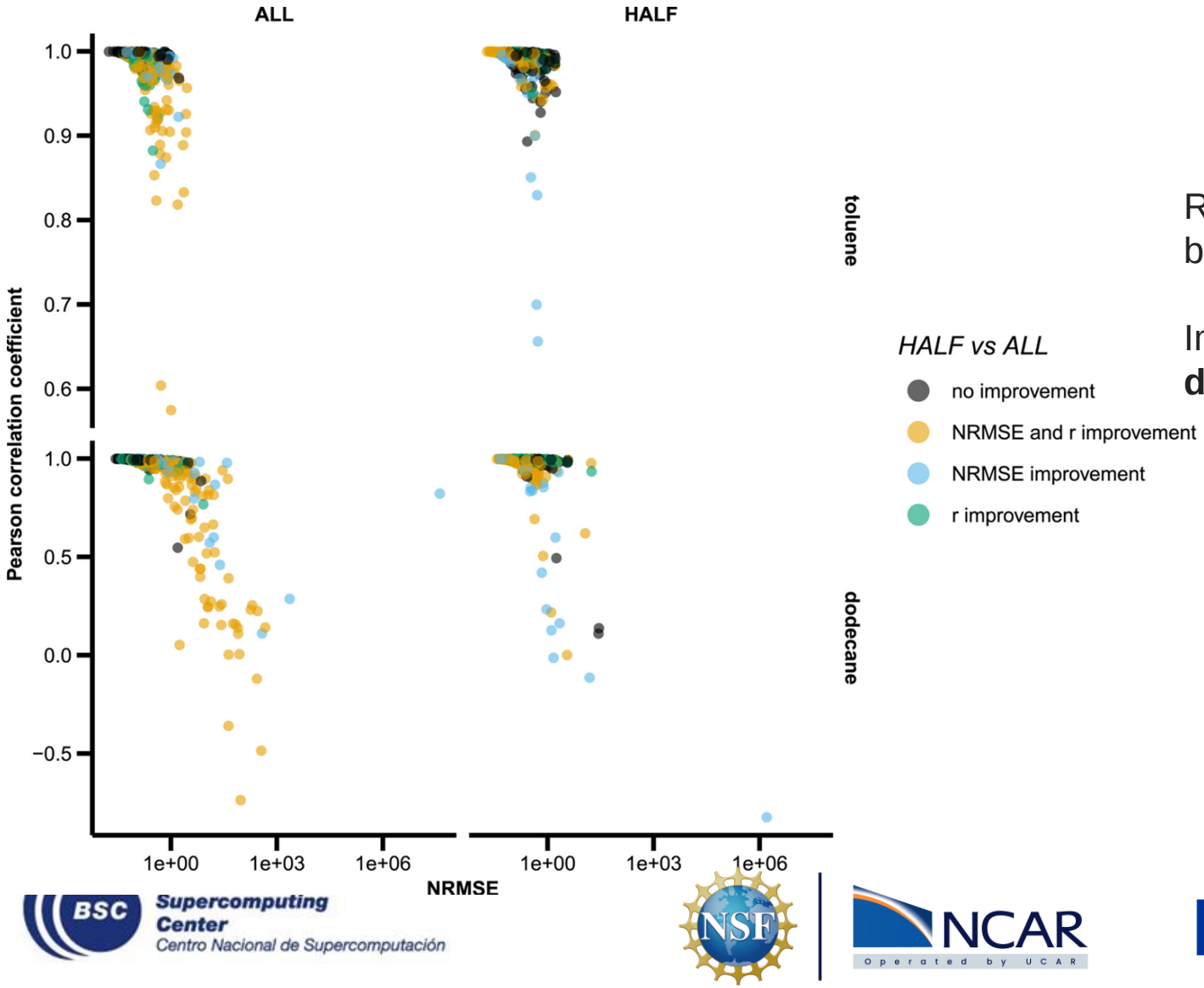
The importance of predictors (mean decrease impurity) is similar for **both** species

Distribution of importance and predictive ability among **correlated variables**

Center
Centro Nacional de Supercomputación
NSF
NCAR
Operated by UCAR
Financiado por la Unión Europea
NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

Random Forest results: reducing the number of predictors

Do performances improve if the number of predictor is reduced?



Reducing the number of (**correlated?**) predictors is beneficial for the worst performing simulations

Importance of **selecting predictors** and **dimensionality reduction**

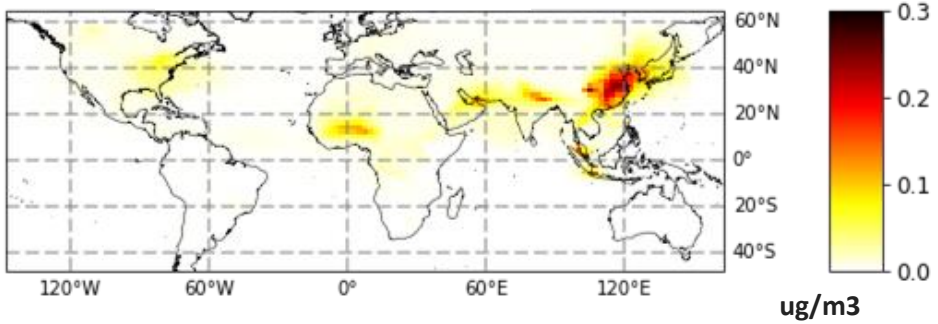
BSC Supercomputing Center
Centro Nacional de Supercomputación
NSF
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Gobierno de España
Vicepresidencia Tercera del Gobierno
Ministerio para la Transición Ecológica y el Reto Demográfico
AEMet Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

First tests in a global model: VBS vs. GECKO-NN

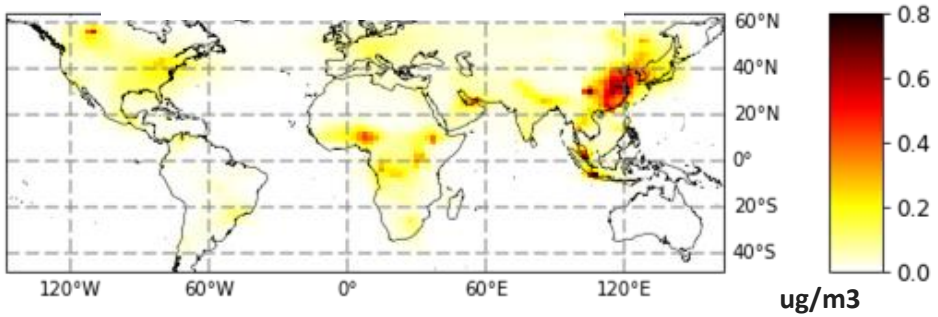
Implementing GECKO-NN in Geos-Chem

GEOS-Chem Monthly average Toluene SOA (May 2016)

GECKO-NN-MLP



VBS Hodzic et al., 2016



GECKO-NN simulations are

- stable over several months
- within a factor of 2 of the VBS parametrization for Toluene-SOA.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
NCAR Operated by UCAR
Financiado por la Unión Europea NextGenerationEU
Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia

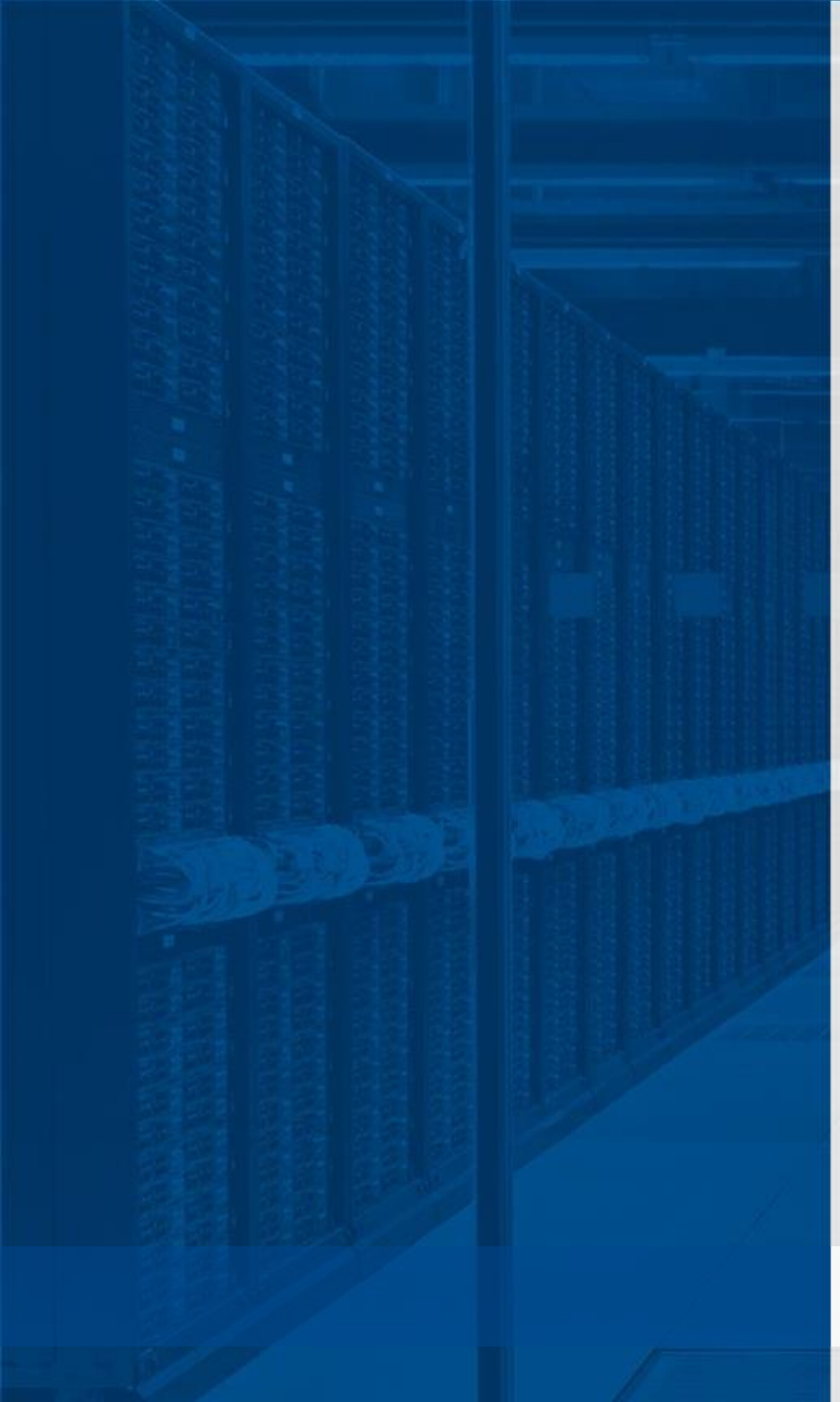
Conclusions and Outlook

- It is possible to emulate the behavior of detailed atmospheric chemistry models with machine learning
- Long term stability can be achieved for recurrent neural networks with GRU
- The training dataset must be carefully constructed to cover all environmental conditions
- Random forests can perform similarly to NN+GRU
- Predictors selection is crucial

Current and future works at BSC

- Bring the complexity of organic chemistry to air quality models, built on the development of detailed Spanish emissions
- Explore ML use, with lessons learned from this work: start again from the basics and systematic exploration by (i) progressively increasing chemical complexity (from toy mechanisms to GECKO-A complexity) and (ii) testing multiple families of ML techniques (RF, NN, GraphNet …)
- Implementation in the MONARCH air quality model

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you for your attention

camille.mouchel@bsc.es

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Vicepresidencia Tercera del Gobierno · Ministerio para la Transición Ecológica y el Reto Demográfico
AEMet Agencia Estatal de Meteorología
Plan de Recuperación, Transformación y Resiliencia