MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN
Proyecto PCI2024-155075-2 financiado por MICIU/AEI /10.13039/501100011033 y Cofinanciado por la Unión Europea

ITM2026

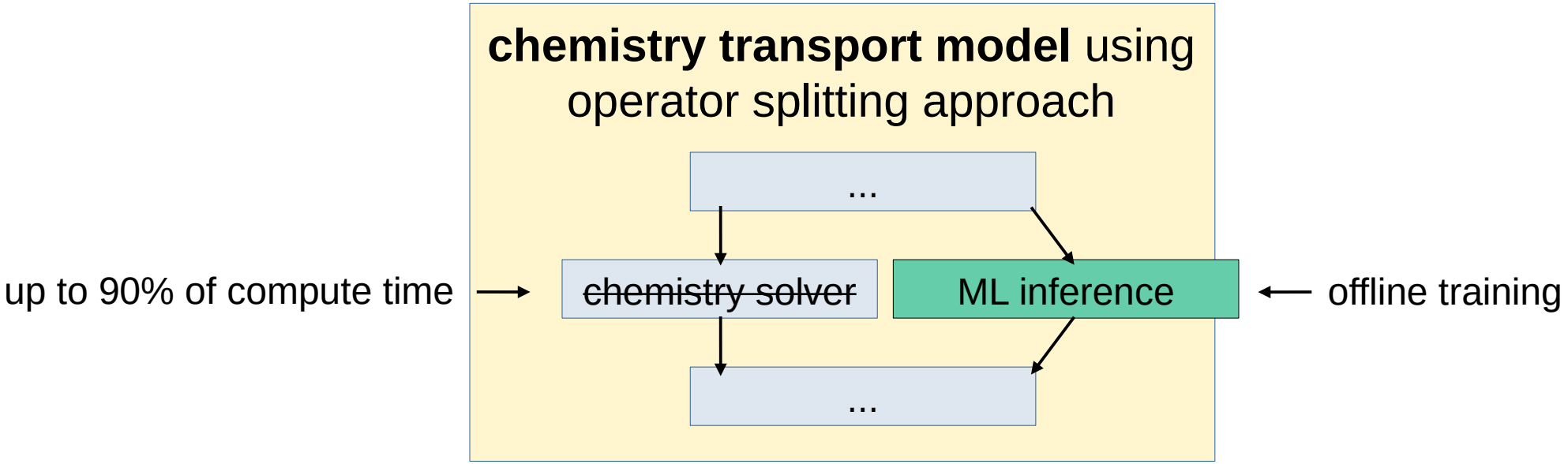
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Emulating tropospheric chemistry mechanisms with deep neural networks

Alessio Melli, Camille Mouchel-Vallon, Isidre Mas Magre, Hervé Petetin, and Oriol Jorba Casellas
Barcelona Supercomputing Center, Barcelona (Spain)

Tuesday, 21 April, 15h20

Framework



Goal of the study: speed up CTM using ML inference

Preliminary findings: model performance on two datasets of the same mechanism

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Case study

POLLU mechanism **20 species, 25 reactions** [1]

Simplified tropospheric ozone formation mechanism widespread in the literature on stiff chemical kinetics

$$\mathrm{NO} + \mathrm{O_3} \xrightarrow{k_{02}} \mathrm{NO_2} \qquad k_{02} = 1.8 \times 10^{-14}\,\mathrm{m^3 molec^{-1} s^{-1}}$$

$$\vdots$$

$$\mathrm{HCHO} + \mathrm{OH} \xrightarrow{k_{06}} \mathrm{HO_2} + \mathrm{CO} \qquad k_{06} = 1.0 \times 10^{-11}\,\mathrm{m^3 molec^{-1} s^{-1}}$$

stiff system
does not depend on *T, p*
no photolysis

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

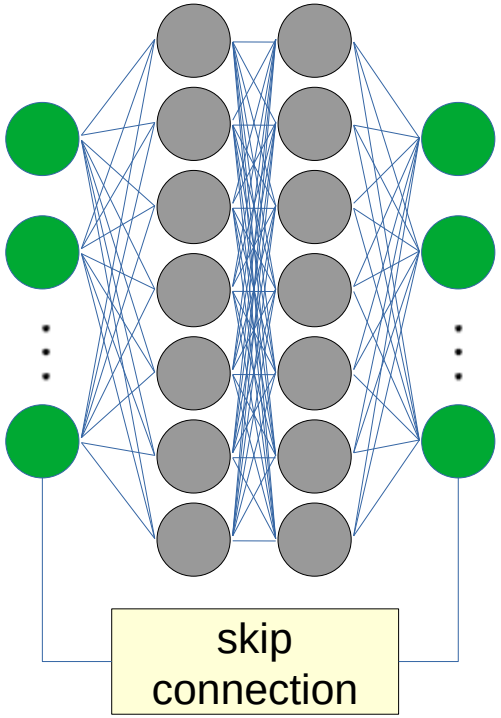
[1] Verwer, Jan G. *SIAM Journal on Scientific Computing* 15.5 (1994): 1243-1250

Baseline architecture

Simple neural network:

fully connected multilayer perceptron (fc-MLP) with skip connection

20 input features: conc at *t*



20 target features: conc at *t*+1

The skip connection:

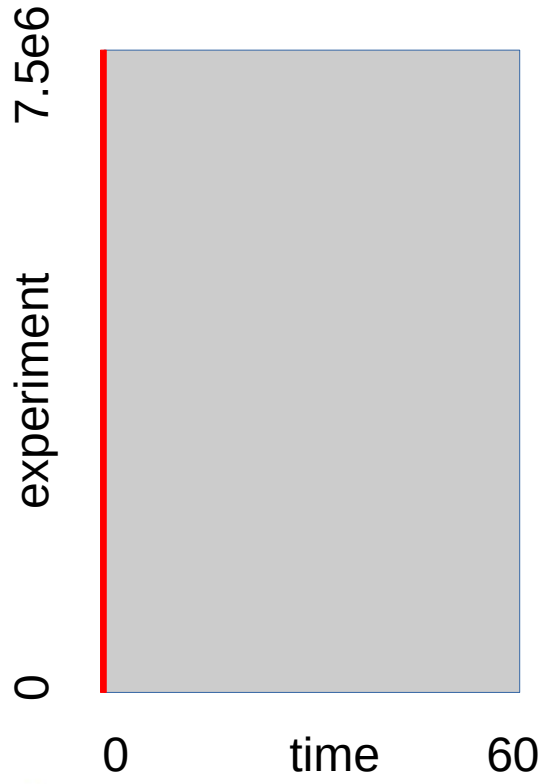
- Sums the inputs to the model output before determining loss
- Removes the design choice of predicting concentration or concentration change
- Improves gradients

hyperparameter tuning on limited search space:

learning rate
neurons per layer (width)
hidden layers (depth)

Data generation

From 7.5 millions uniformly sampled initial conditions we generated 60-timestamp simulations using BSC's CAMP boxmodel. From this, we built two datasets (same # data points):

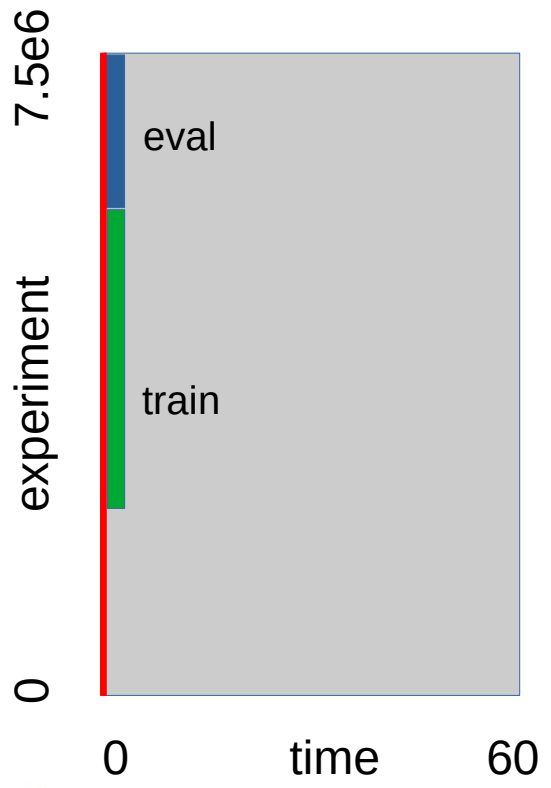


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

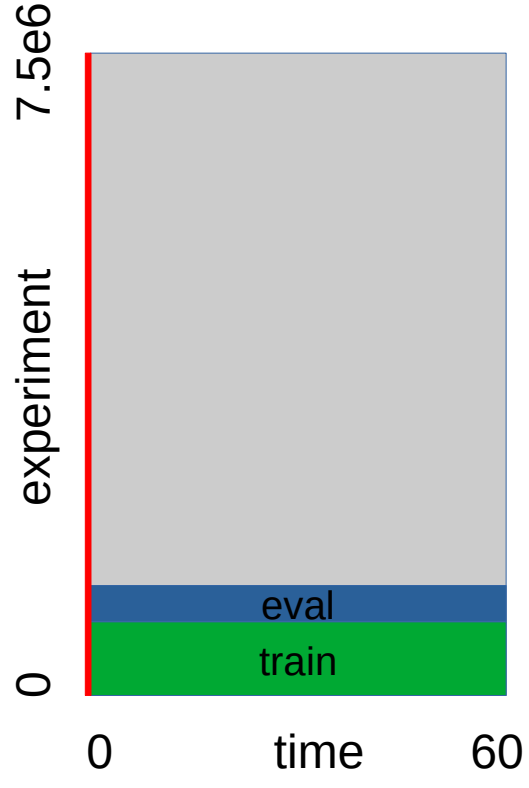
Data generation

From 7.5 millions uniformly sampled initial conditions we generated 60-timestamp simulations using BSC's CAMP boxmodel. From this, we built two datasets (same # data points):

1. First timestamp (1st-ts)



2. Full trajectories (60-traj)

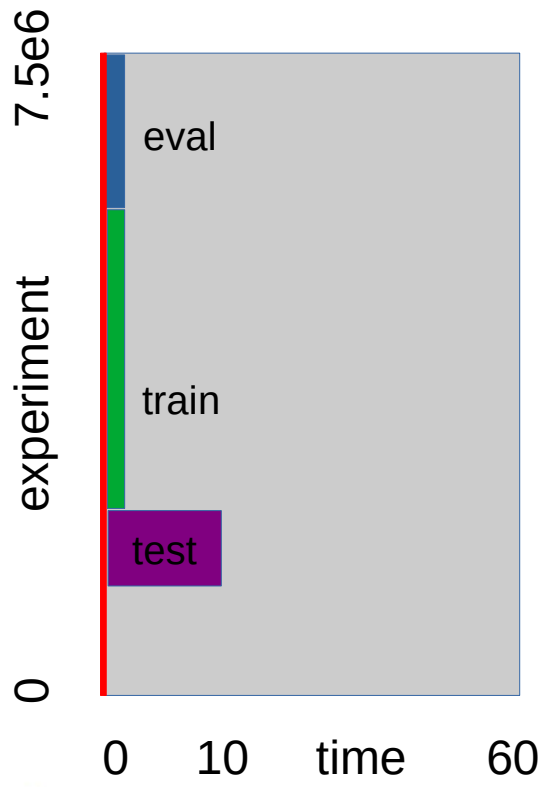


Each area of the same color has the same surface

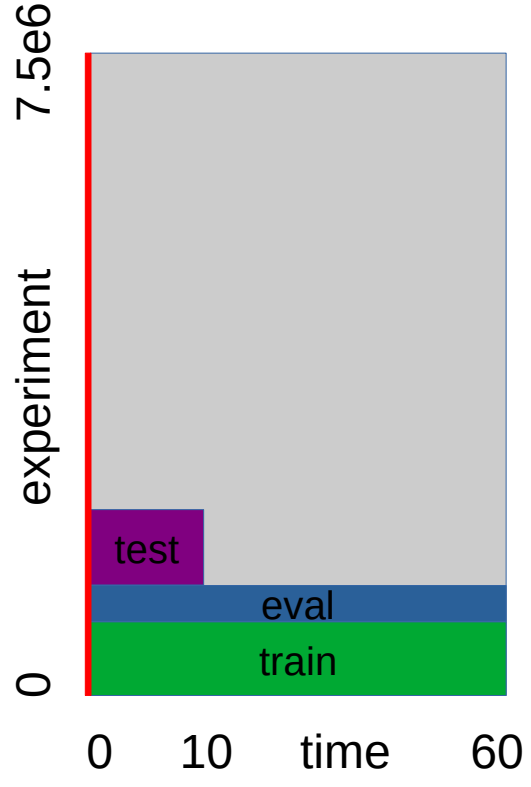
Data generation

From 7.5 millions uniformly sampled initial conditions we generated 60-timestamp simulations using BSC's CAMP boxmodel. From this, we built two datasets (same # data points):

1. First timestamp (1st-ts)



2. Full trajectories (60-traj)



data points

train: 3.6 millions
eval: 2.4 millions
traineval: 6 millions

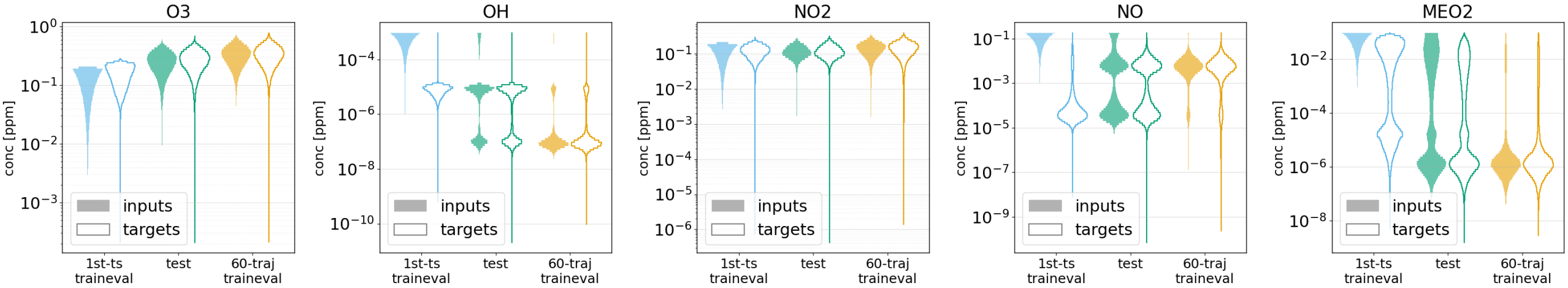
+ additional test (never seen in training): 13.7 millions

Each area of the same color has the same surface

Data generation

Initial ranges for uniform sampling are selected based on typical atmospheric values, with some margin.

For many species, the inputs and targets to the ML model can assume significantly different distributions.



The differences in features' distribution will affect the **performance**.

n.b.:
1st-ts refers to the best* fc-MLP model trained on 1st-ts traineval subset
60-traj refers to the best* fc-MLP model trained on 60-traj traineval subset

*after hyperparameter tuning

Before comparing performances...

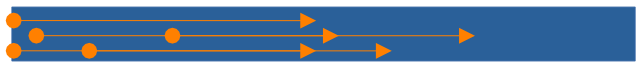
has the model learned?

Each model is evaluated on their **eval** subset: they have been used in training and are different for each dataset.

definition: **autoregressive inference**

From **all points** on both eval subsets, we can let the run the model in *autoregressive mode* (*i.e.*, the model uses its own output as following input) for a duration (AR horizon) up to 10 timestamps.

***e.g.*, Full trajectories (60-traj)**



eval

AR horizon ranges from 1 to 10

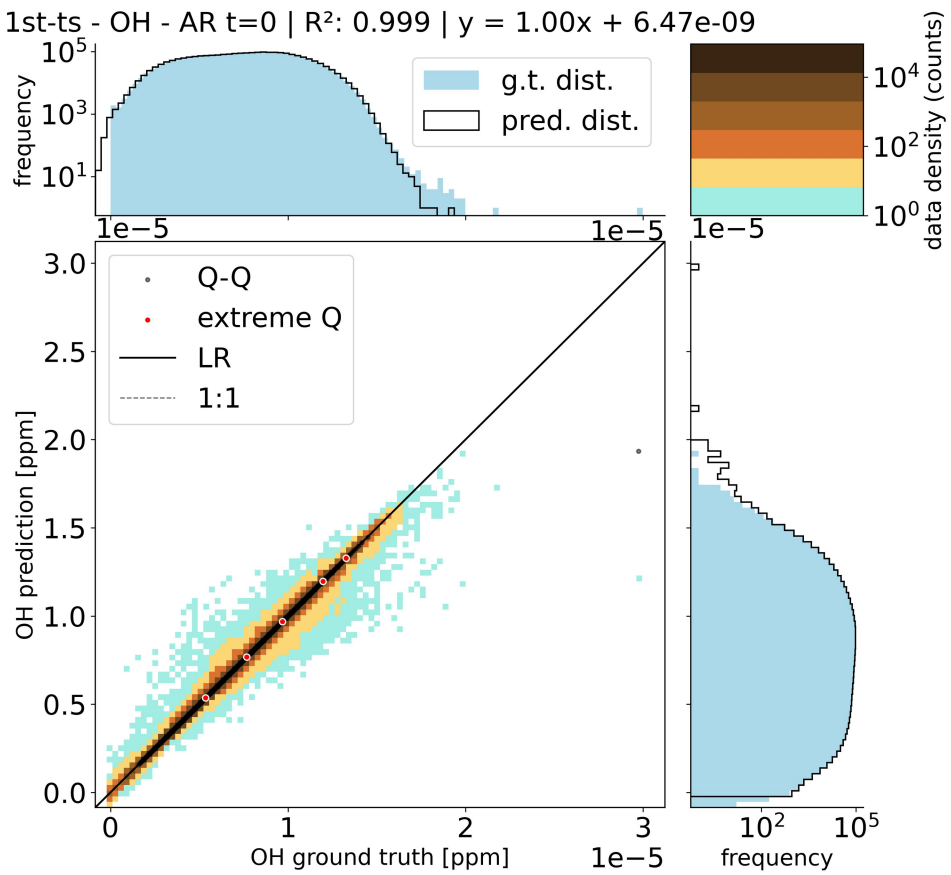
When AR horizon is equal to 1, this is the one-step inference.

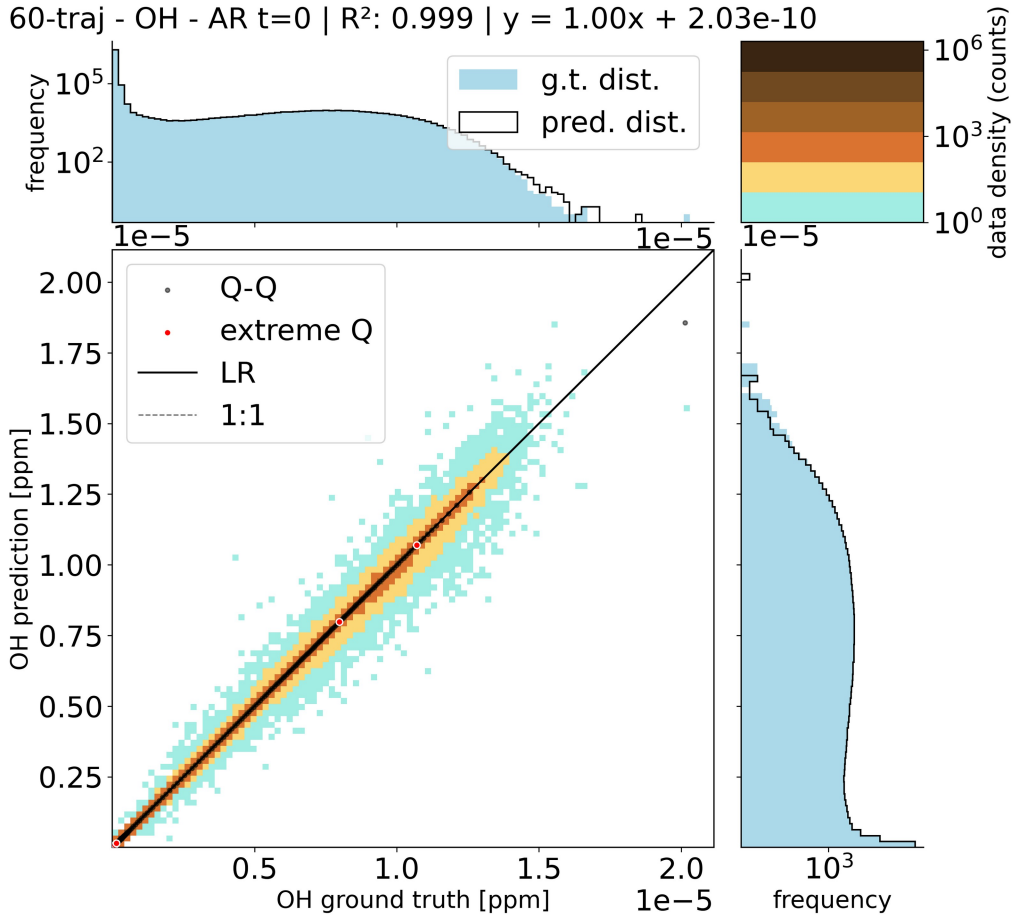
one-step performance on eval subset

AR horizon = 1

performance on eval subsets: OH

The models show they have learned: 19 species show R^2 above 0.9985, PCC above 0.999, and a LR between ground truth and predictions with a slope close to 1. Worst performance: C2O3 (next slide)



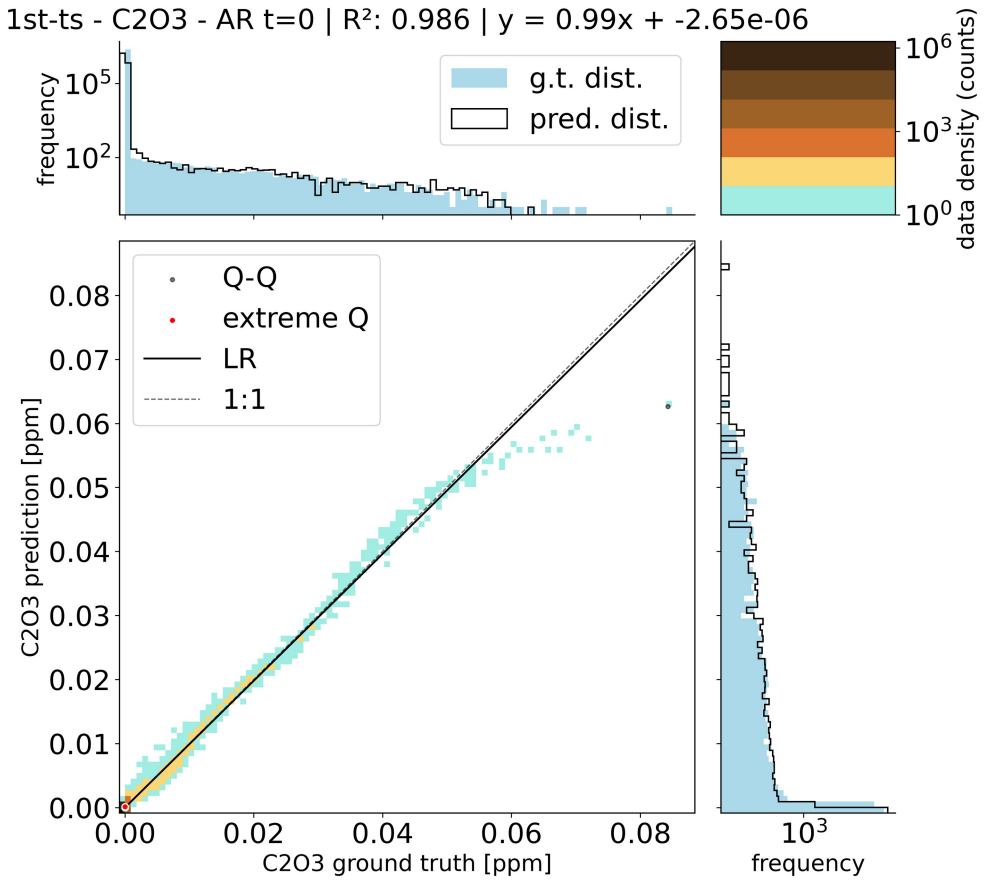


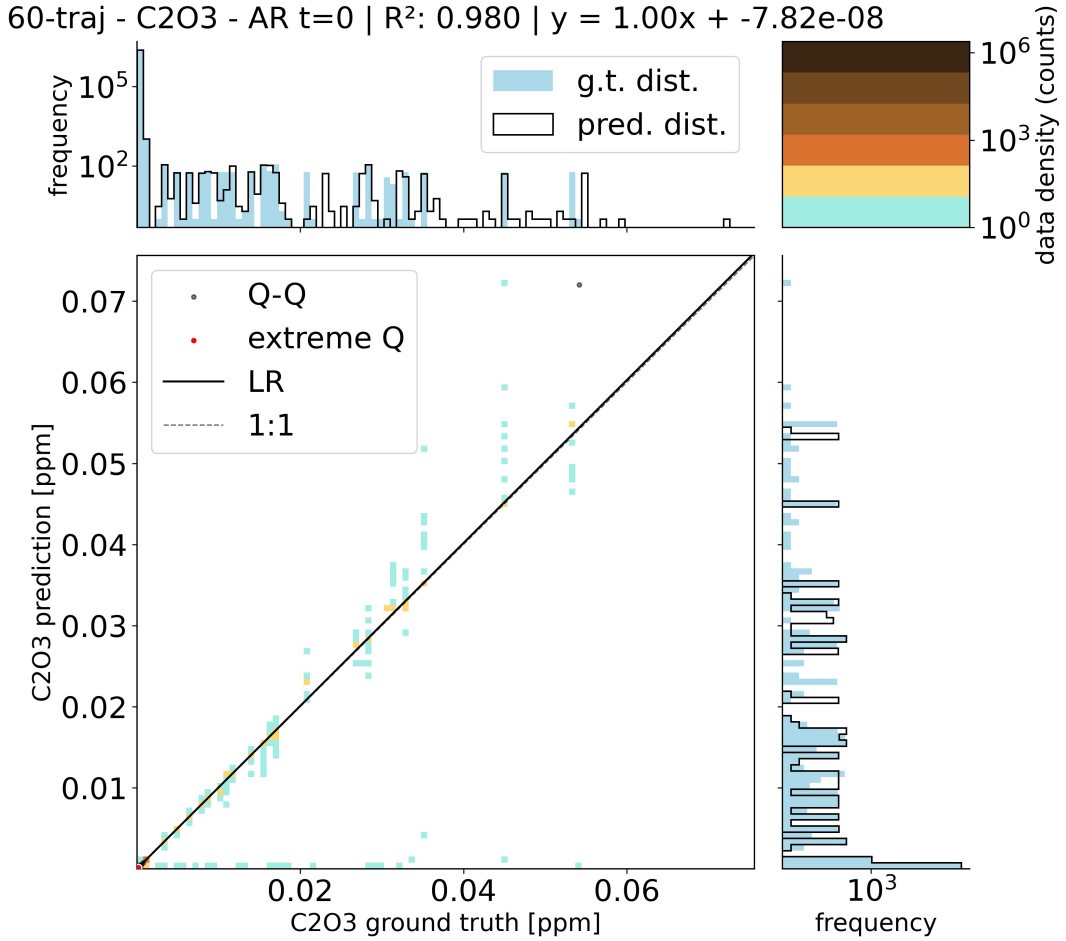
1st-ts eval ≠ **60-traj eval**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

performance on eval subsets: C2O3

Worst performance is still acceptable, given the highly skewed distribution. The model learns mainly to predict null values. This is especially true for **60-traj** dataset, where most points are concentrated towards 0.





1st-ts eval ≠ **60-traj eval**

AR performance on eval subset

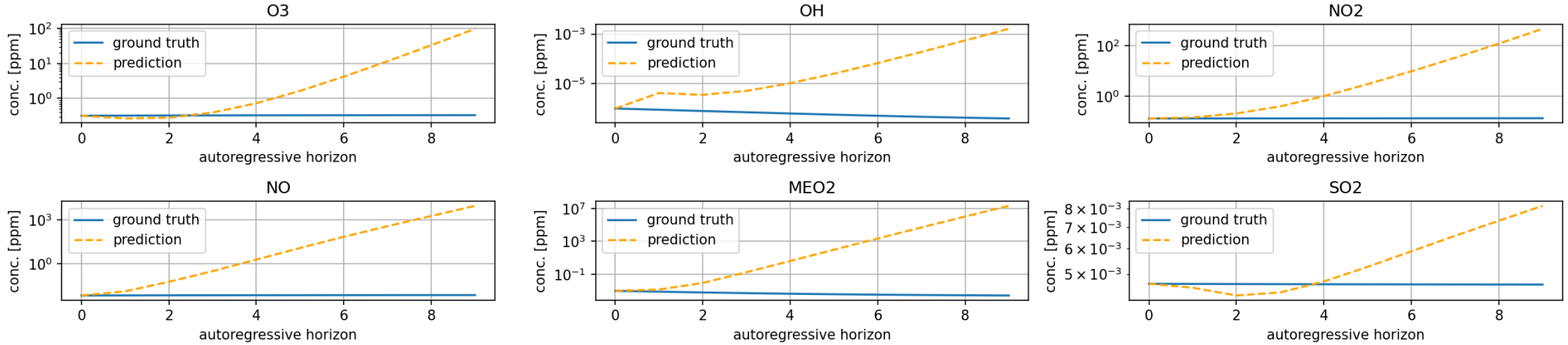
AR horizon = 1...10: performance is expected to worsen due to error accumulation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

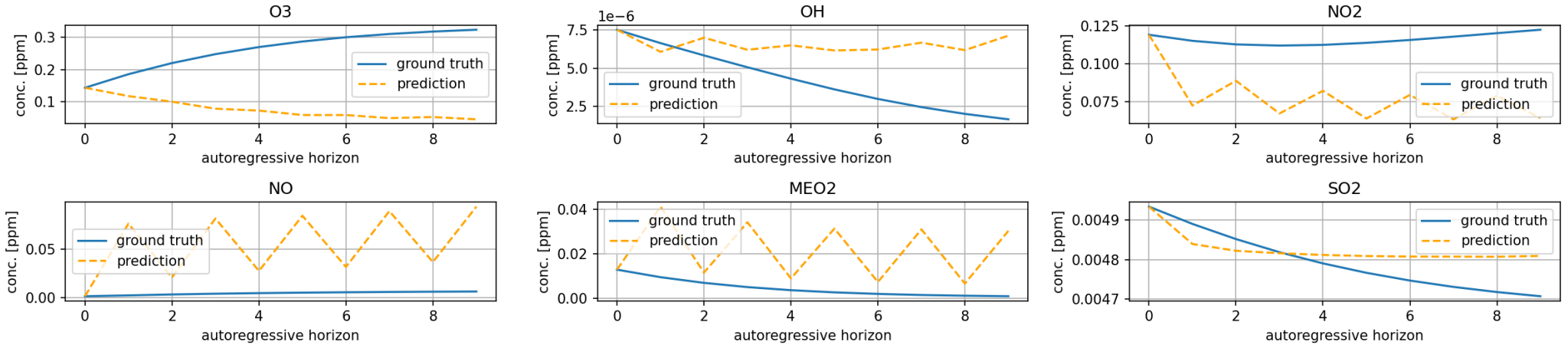
AR performance on the eval subset

The error accumulation acts differently between models. It's counterituitive: The model that did *not* train on full trajectories seems to behave better in AR.

mean 60-traj (log-scale)



mean 1st-ts



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

comparing performances

each model is evaluated on the same test subset which has not been seen in training guaranteeing a meaningful comparison

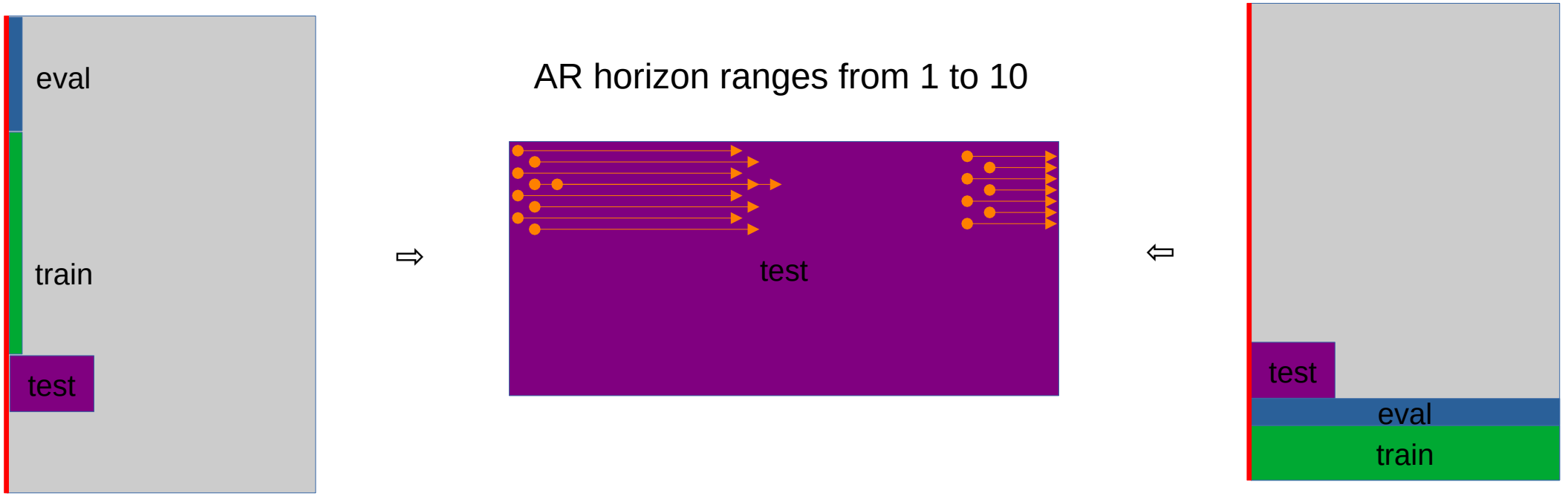
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

AR performance on the test subset

At first, let's look at the test subset: the performance is directly comparable as the test subset has not been seen in training by either model.

1. First timestamp (1st-ts)

2. Full trajectories (60-traj)



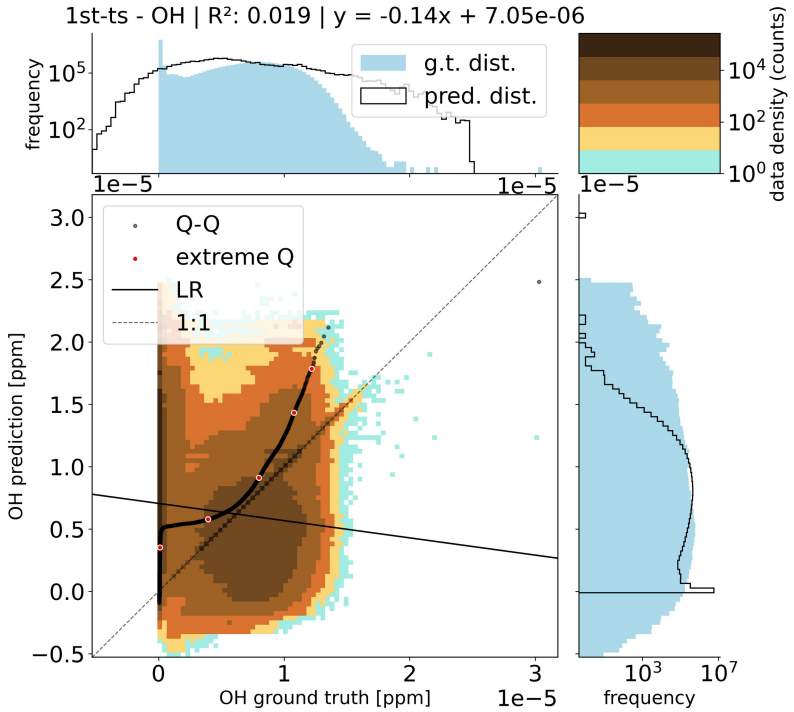
one-step performance

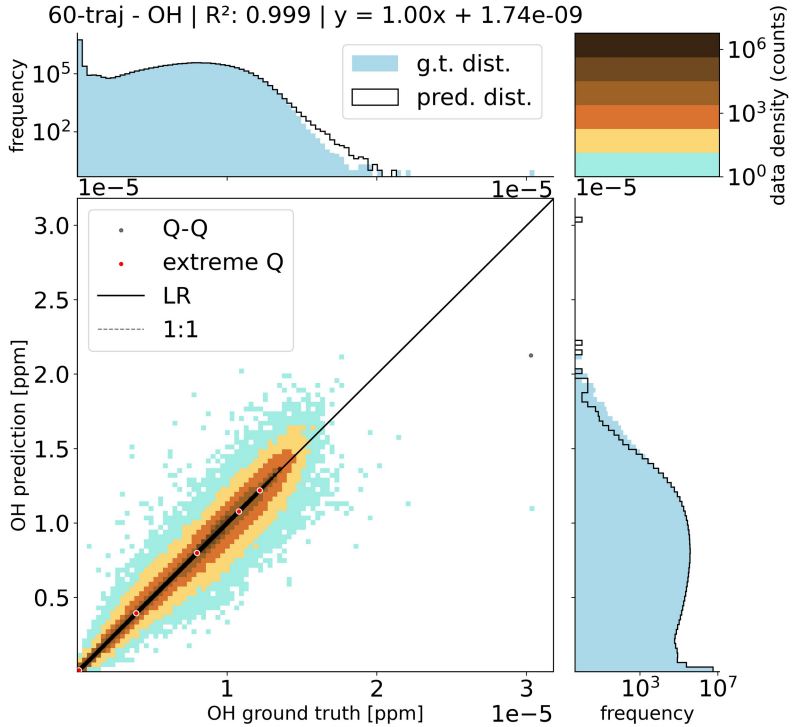
AR horizon = 1, test subset

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

performance on the test subset: OH

The worsening of the performance is immediate for 1st-ts. Best performance: SO2 with R^2 of 0.995 and slope 1.00. Performance is good on 19 species (R^2 > 9.998) for 60-traj. Worse performance again on C2O3.

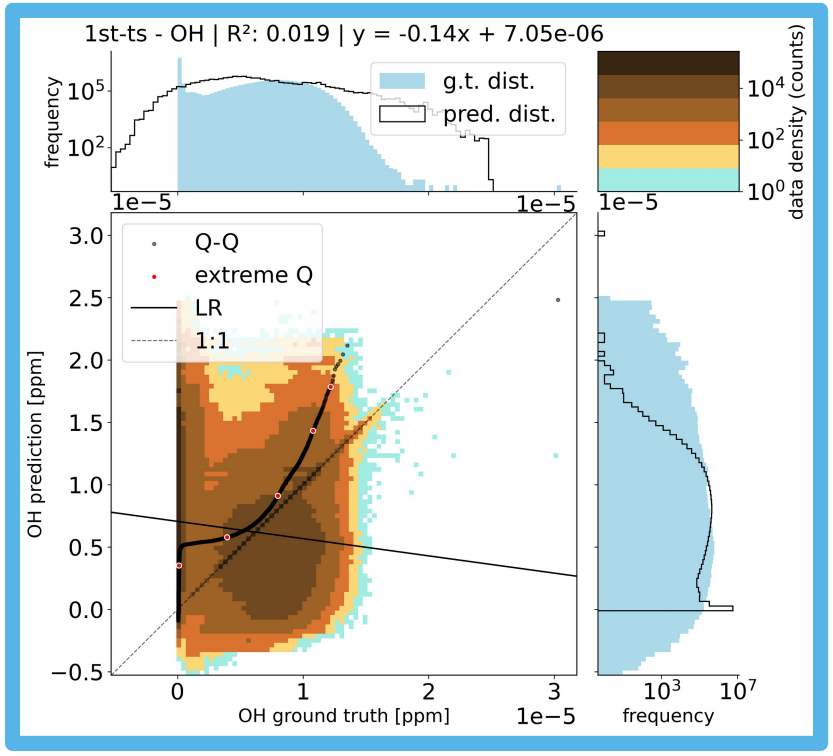


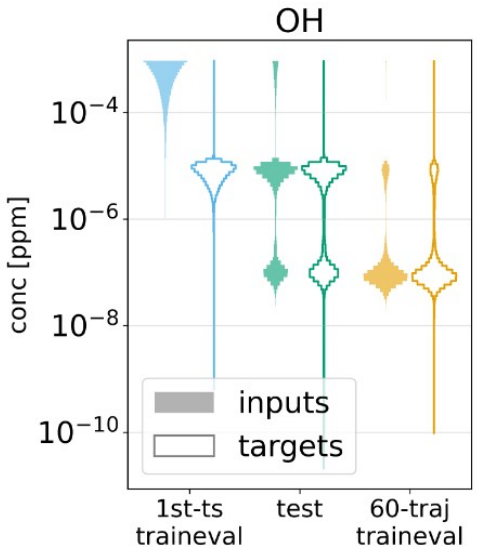


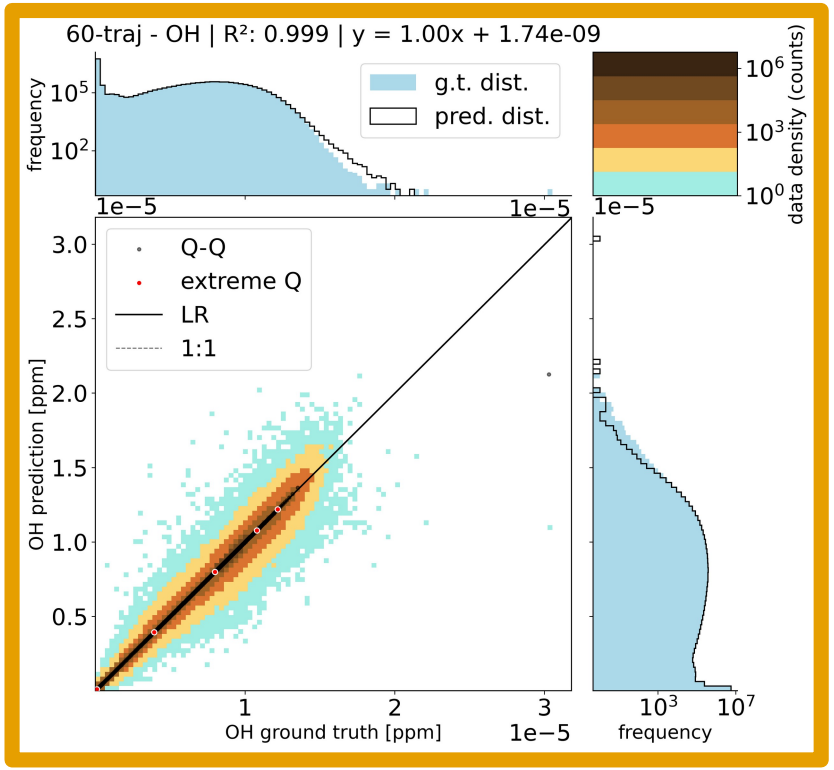
1st-ts test = 60-traj test

performance on the test subset: OH

The worsening of the performance is immediate for 1st-ts. Best performance: SO2 with R^2 of 0.995 and slope 1.00. Performance is good on 19 species (R^2 > 9.998) for 60-traj. Worse performance again on C2O3.





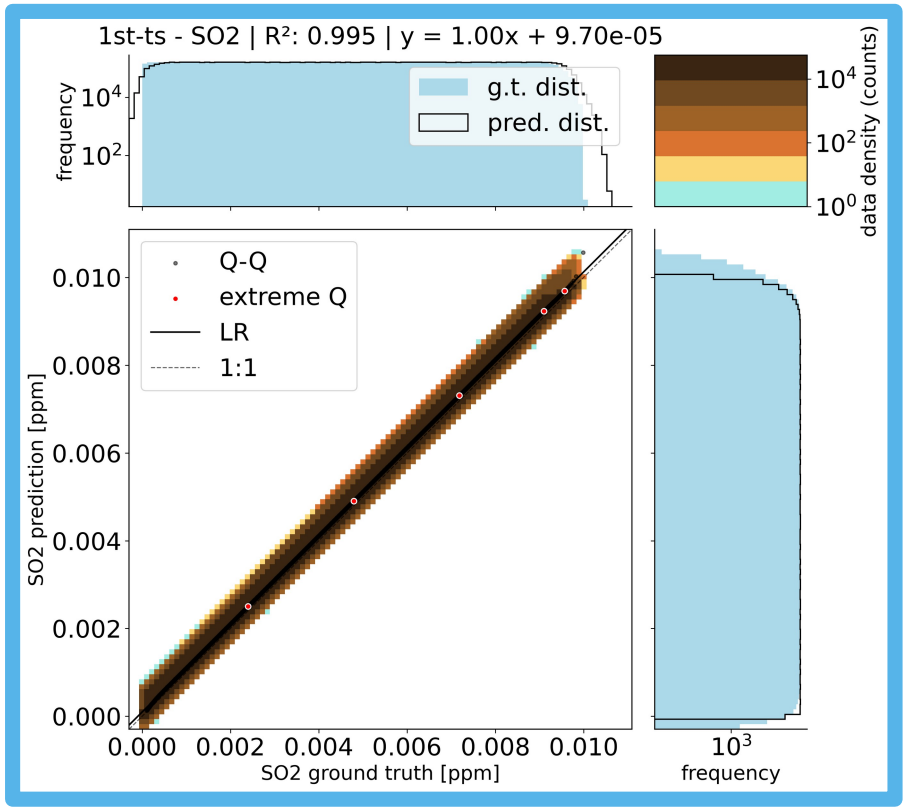


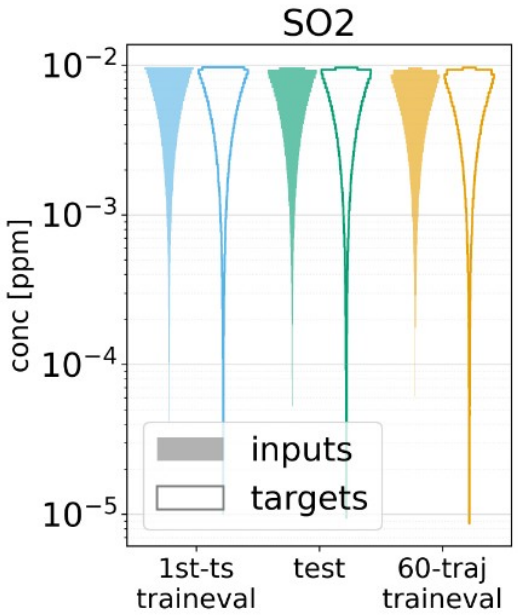
All species distribution agreement is better between 60-traj and test.

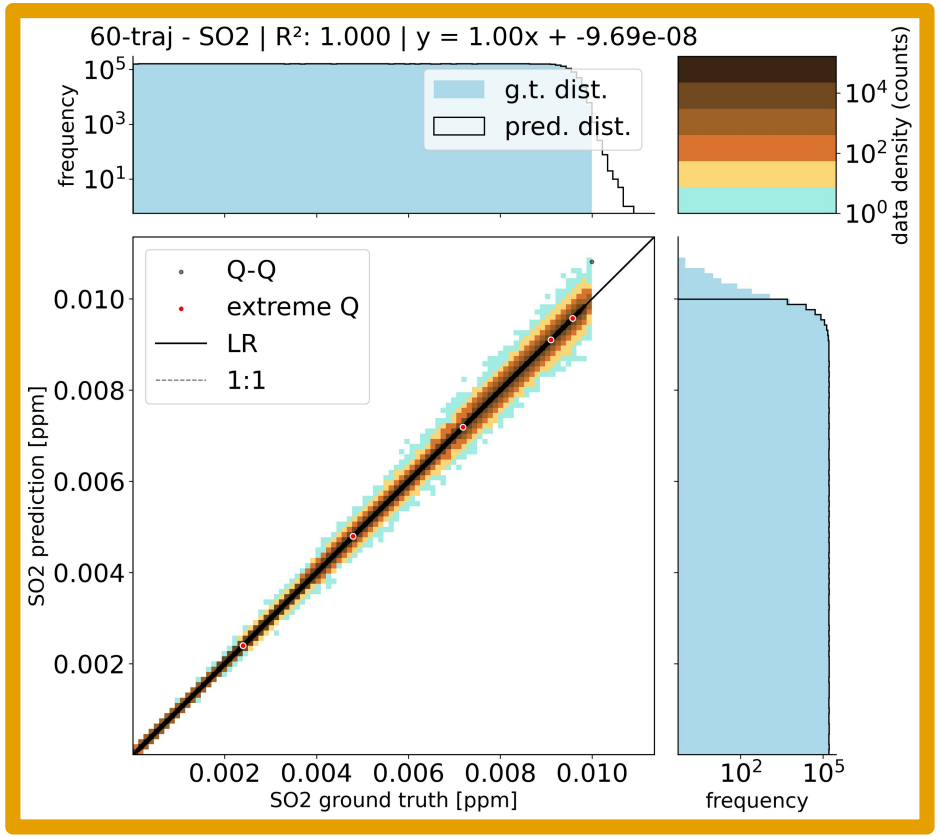
Similar shape in 1st-ts: NO, O3P, HO2, MEO2, C2O3, PAN, CH3O, O1D, NO3

performance on the test subset: SO2

Similar distributions give better results: For SO2, also 1st-ts has good performance.







SO2 distributions are all in agreement:
it is slowly decaying with time.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

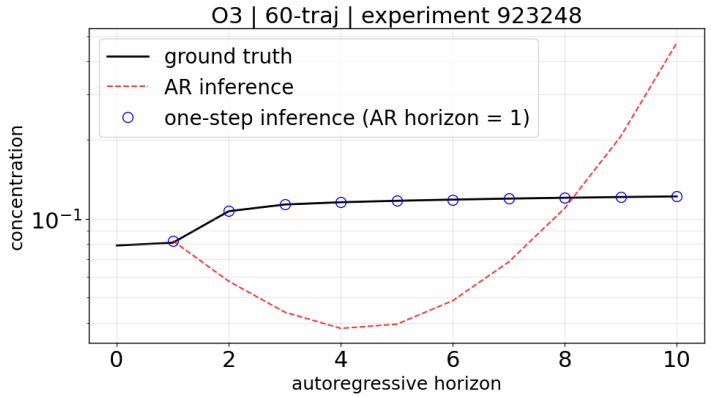
AR performance

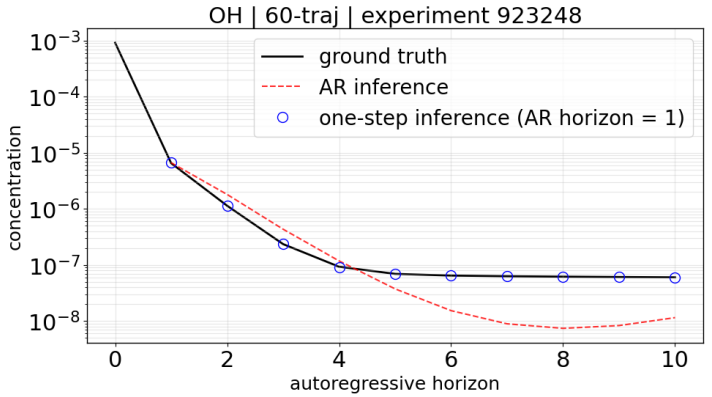
AR horizon = 1...10, test subset: performance is expected to worsen due to error accumulation

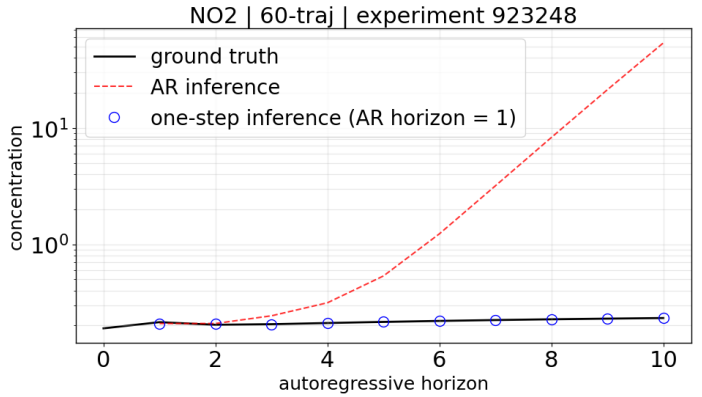
AR performance on the test subset

The error accumulation acts differently between models. It's counterituitive: Again, the model that did *not* train on full trajectories seems to behave better in AR.

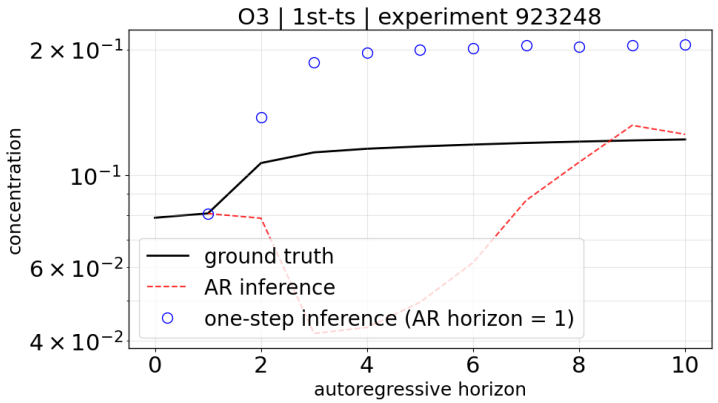
60-traj (log-scale)

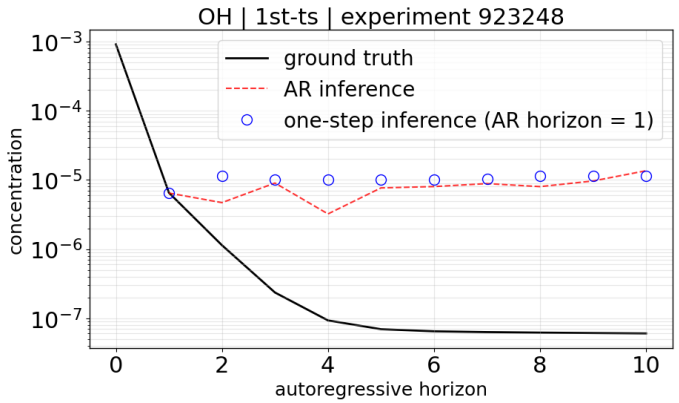


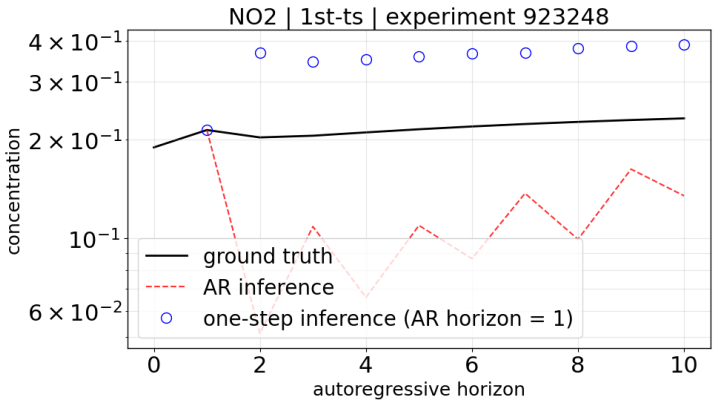




1st-ts (log-scale)







Barcelona Supercomputing Center
Centro Nacional de Supercomputación

open questions

- AR performance for 1st-ts shows bounded behavior ⇨ how to exploit it?
- Some species behave globally worse than others (C2O3, NO2, MEO2) ⇨ can a tailored preprocessing/different dataset design help?

next steps

- Improve preprocessing and/or dataset design on short-lived species
- Include model constraints based on physics
- CTM target: CB05 mechanism

Thanks for your kind attention