# Data assimilation (DA) for atmospheric composition (AC)

M.Sofiev,

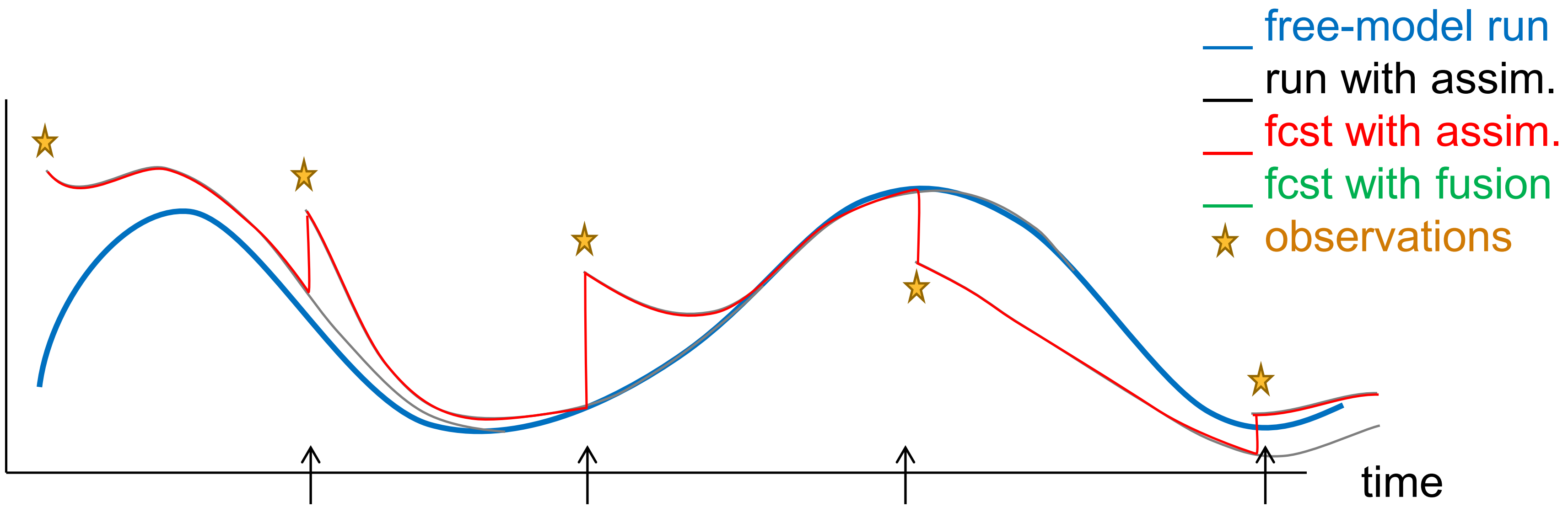
with material from J.Vira, A.Uppstu

# Outline

- Introduction: DA from the bird's view
- Atmospheric composition model as a part of Earth-System model
- Challenges from the DA viewpoint
- Approaches to perform DA for the system violating basic assumptions
- Illustrations
  - Control variable selection
  - Technique selection
- Summary

## **Data assimilation in a classical form:** a bird's view

- The approach designed for meteorological forecasting
  - Corrects the model state, i.e. the predicted variables (T,q,U,V,p,...)
  - Works there



# AC problem is bulky...

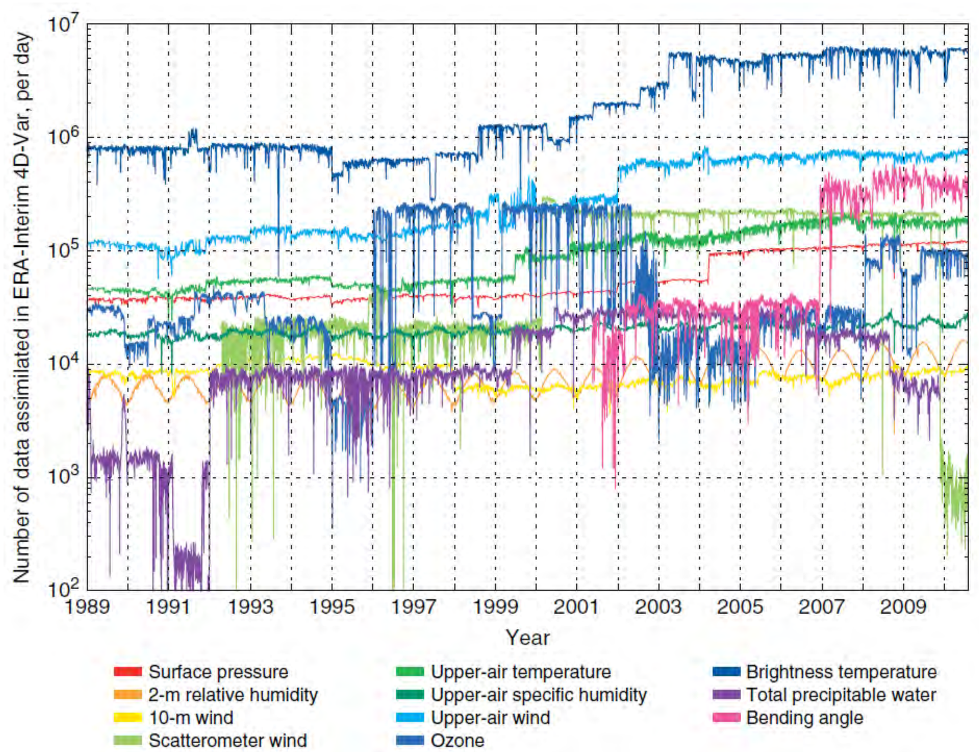
*“He inhaled a breath of humid morning breeze and let in nitrogen, oxygen, argon, xenon & radon, steam, carbon monoxide, nitrogen dioxide, tetra-ethyl lead, benzene, some mould spores, a bacteria fleet, anonymous body hair, a pigeon ectoparasite, anemophilous pollen, a drop of sulphur dioxide flown from a distant factory, and a particle of dust carried by the night sirocco.*

*In other words he breathed air of the city”*

(Stefano Benni "*Achille piè veloce*", Mondadori, Italy, 2003)
Courtesy of G.D’Amato

# ...and much worse observed

- Che … s of nur
  - ➢ I … c … ( … heme
  - ➢ / … ole

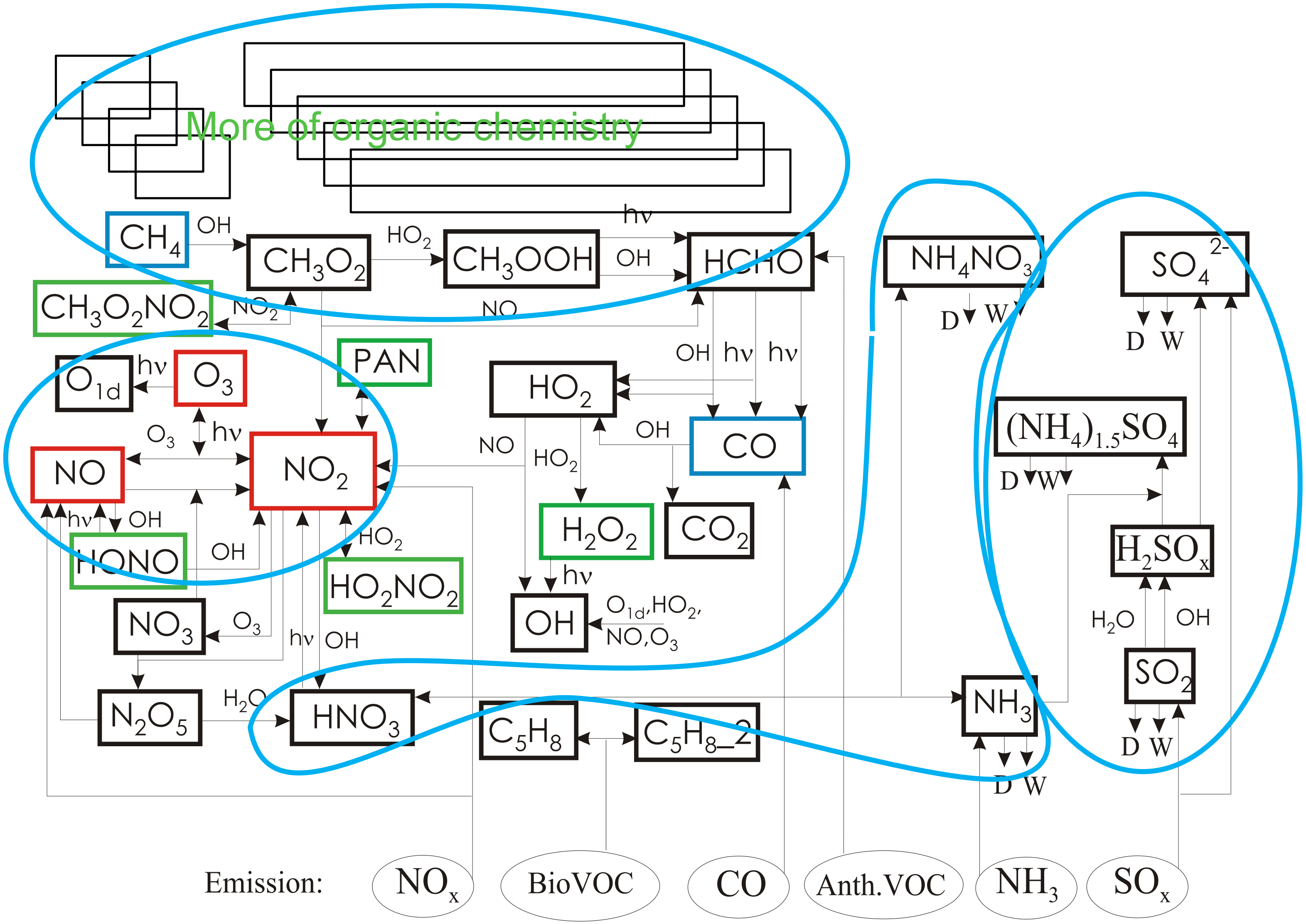



Daily count of observations in ERA-Interim

| | o3 | no2 | pm25 | pm10 | so2 |
|---|---|---|---|---|---|
| 20161101 | 9839 | 11424 | 3746 | 8628 | 5826 |

Daily count of observations in CAMS: example of 1.11.2016

# Chemistry scheme for SOx/NOx/NHx



# Model variables and parameters

Model variables

model state concentrations ~100

$c_1$
…
$c_{nobs}$
…
$c_n$

observed (~10)

non-observed (~100)

emission fluxes ~10

$e_1$
…
$e_m$

emission (10-20)

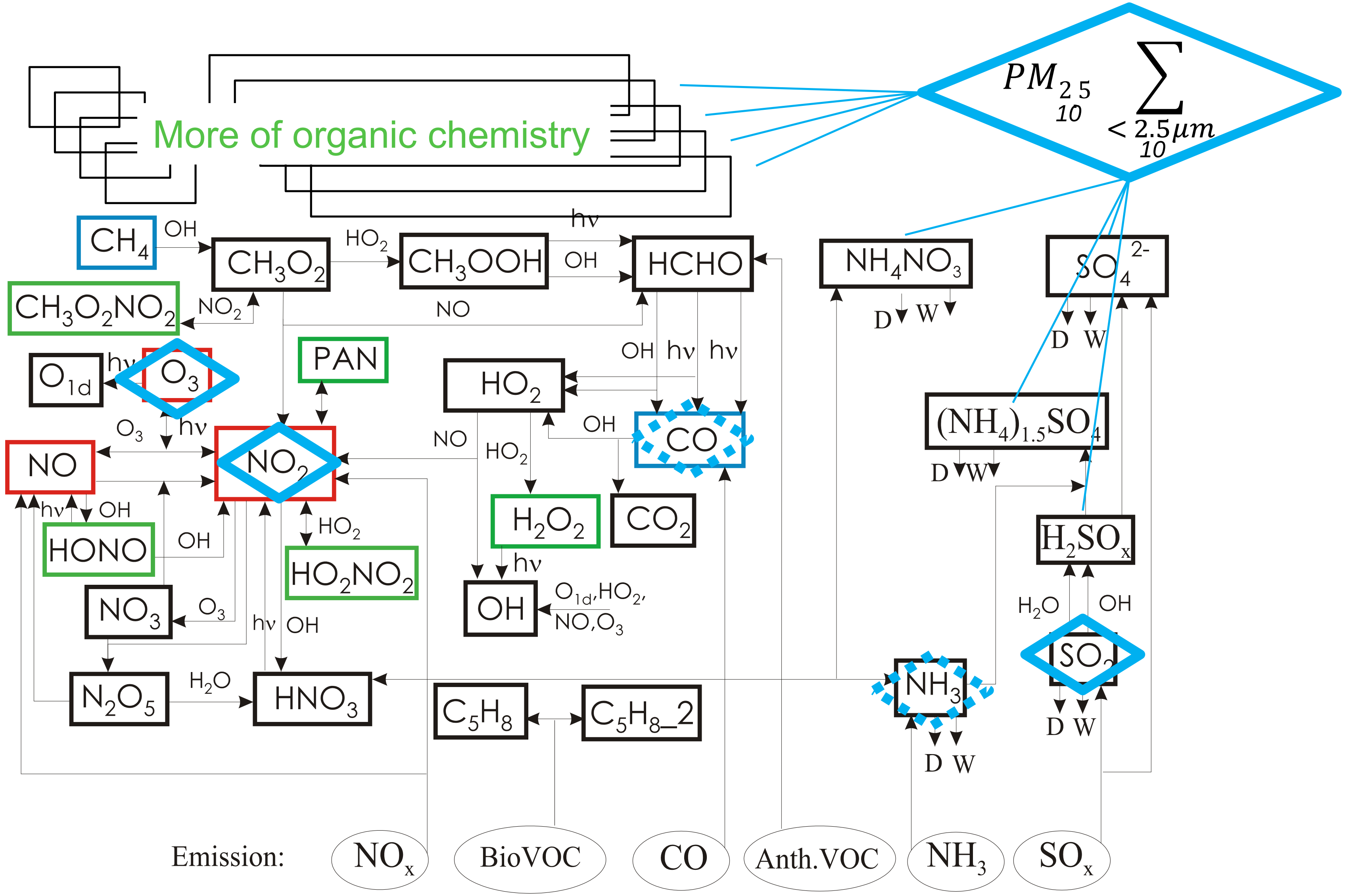
model parameters
meteorology ~10
parameterizations ??

$p_1$
…
$p_k$

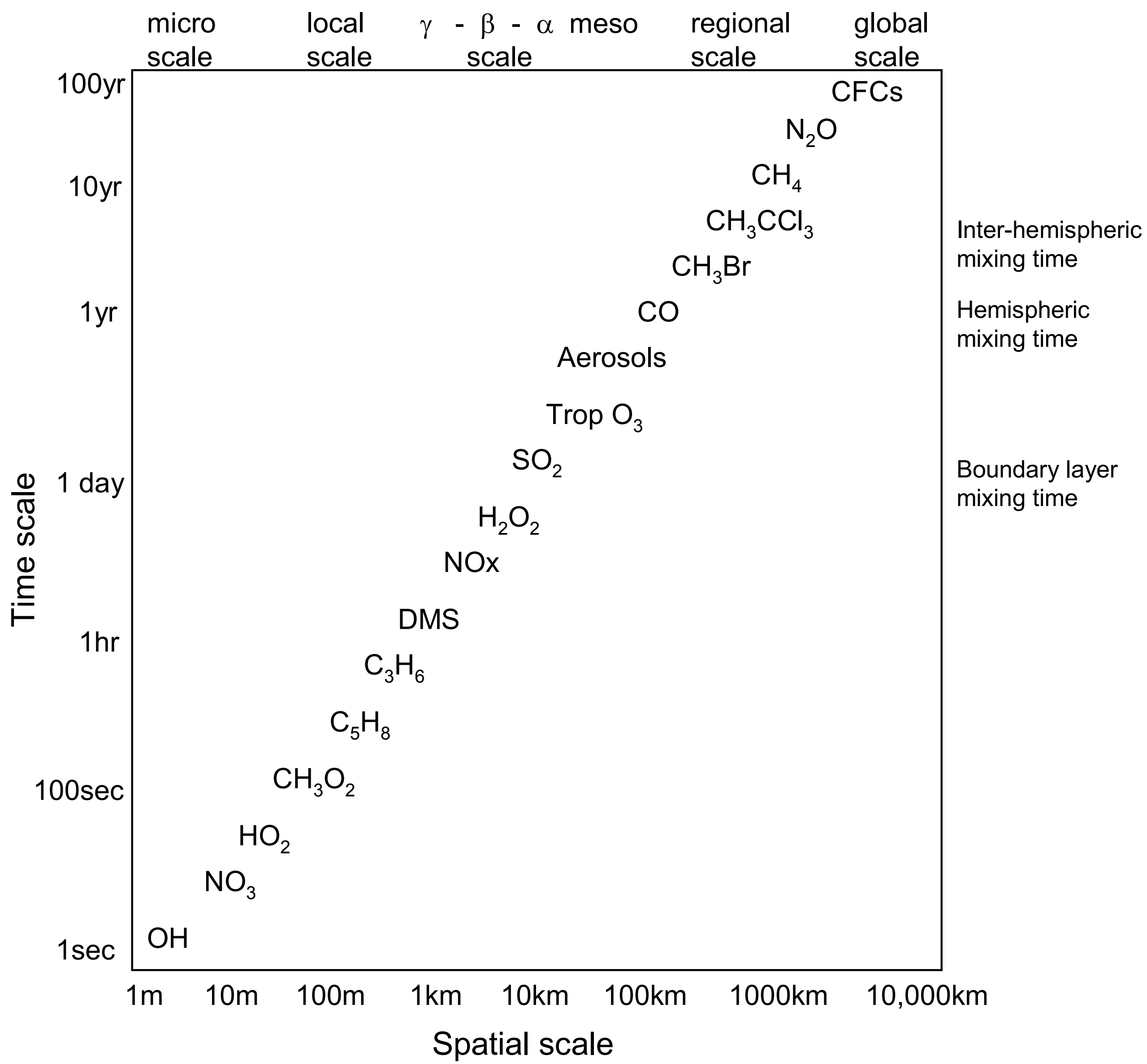
physical state (meteorology, ~10)

Each variable is a map of $10^6$ - $10^8$ grid cells

# What do we observe?



# Spatial and temporal scales in AC

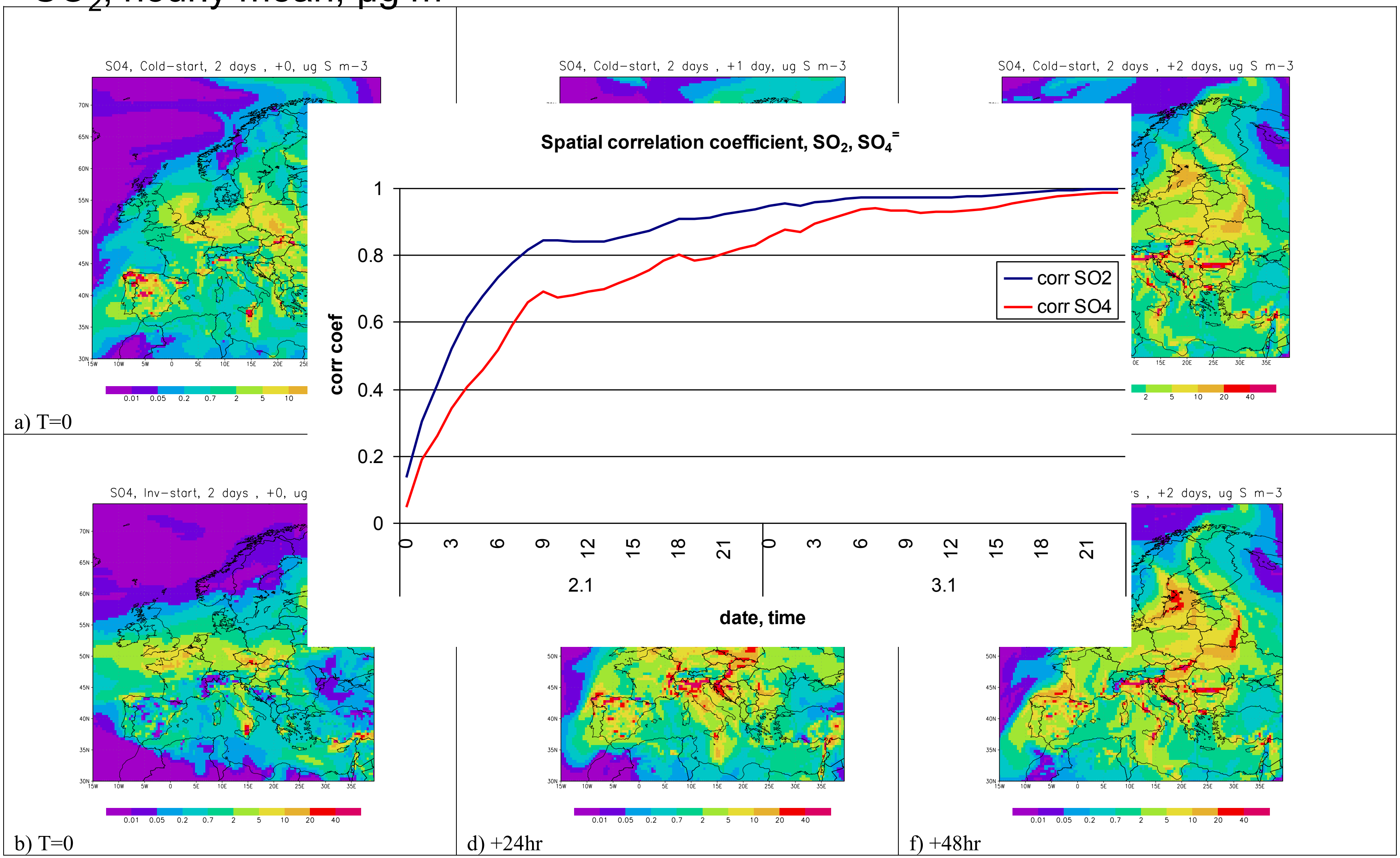


# ...and much worse observed

- Chemical-system state vector contains concentrations of numerous species…
- … and constraining this vector is not enough:
  - forced motion of this non-autonomous system may be (and often is) the most significant
  - the own system relaxation is often fast and quickly eliminate the effect of DA

# Memory of the troposphere

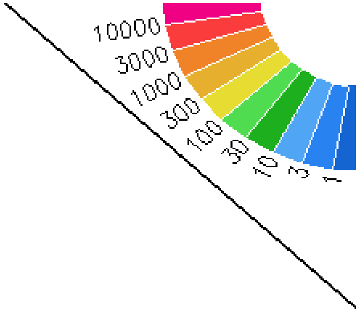
$SO_2$, hourly mean, µg $m^{-3}$

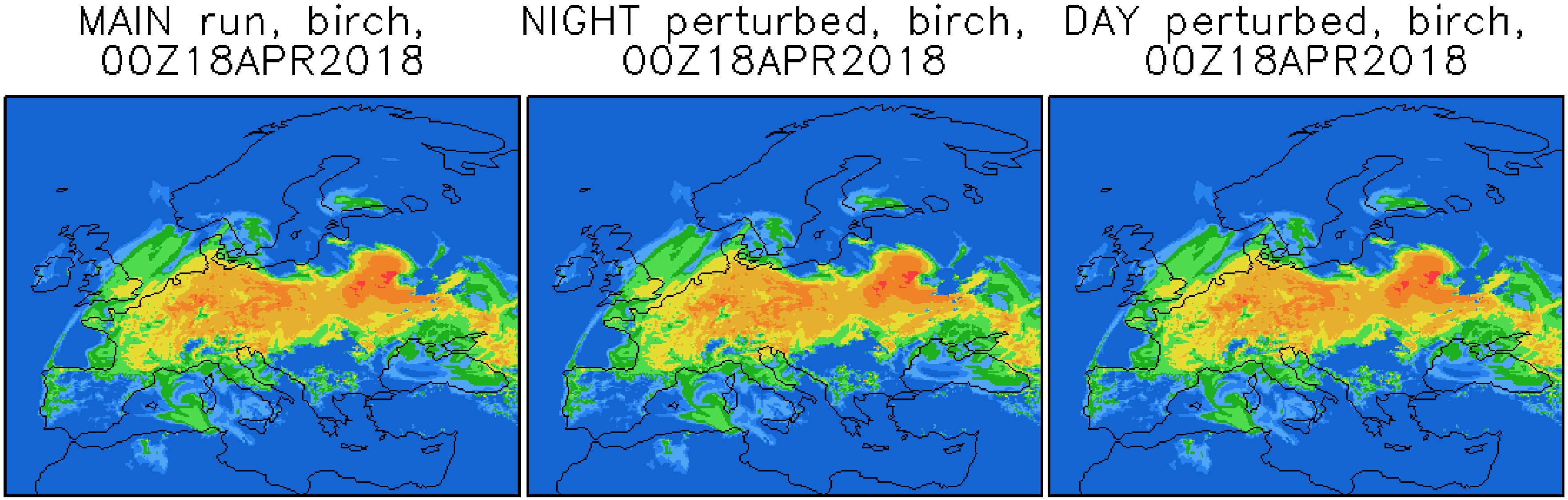


# Pollen test case (no chemistry)

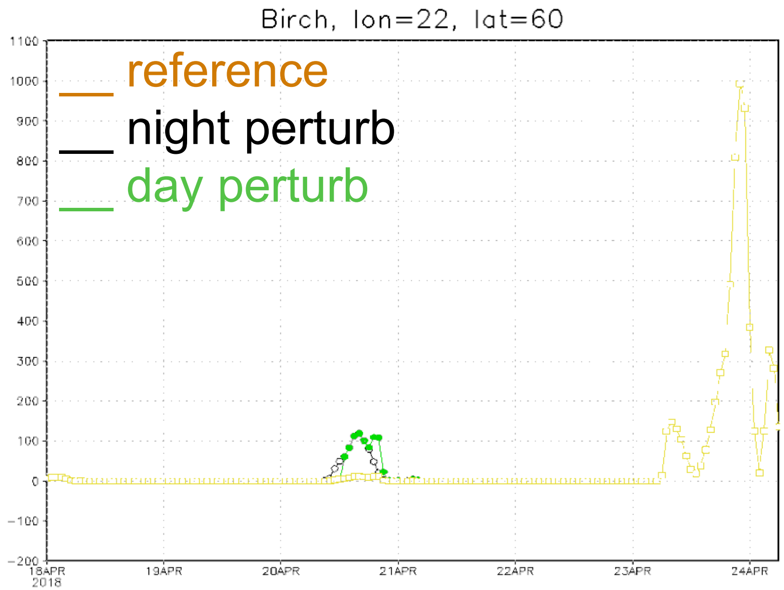
- Season 2018
- Birch
- Europe
- Resolution ~20 km $\times$ 1 hr
- Runs
  - standard
  - midnight perturbation: **00:00** 20.04.2018 all concentrations * **10**
  - mid-day perturbation: **12:00** 20.04.2018 all concentrations * **10**
- Remark: extreme case, usual data assimilation is much more conservative

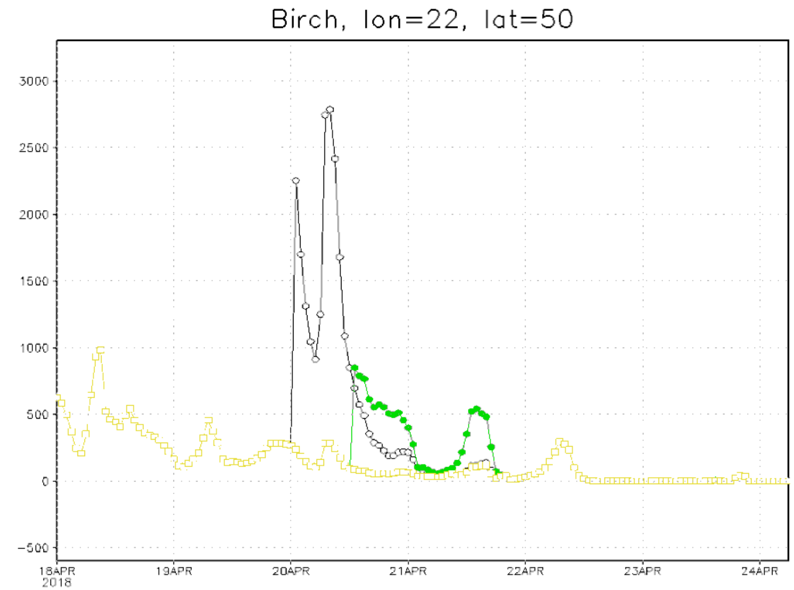
# Birch hindcast with SILAM, 2018

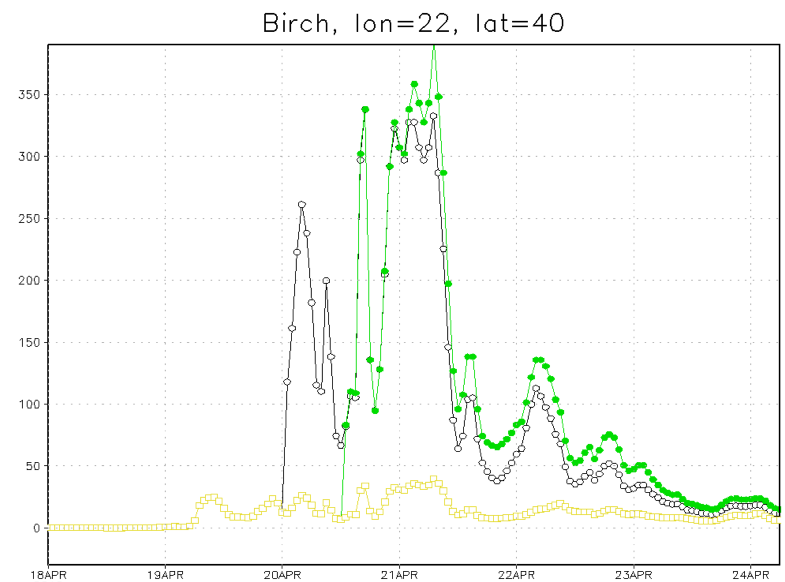


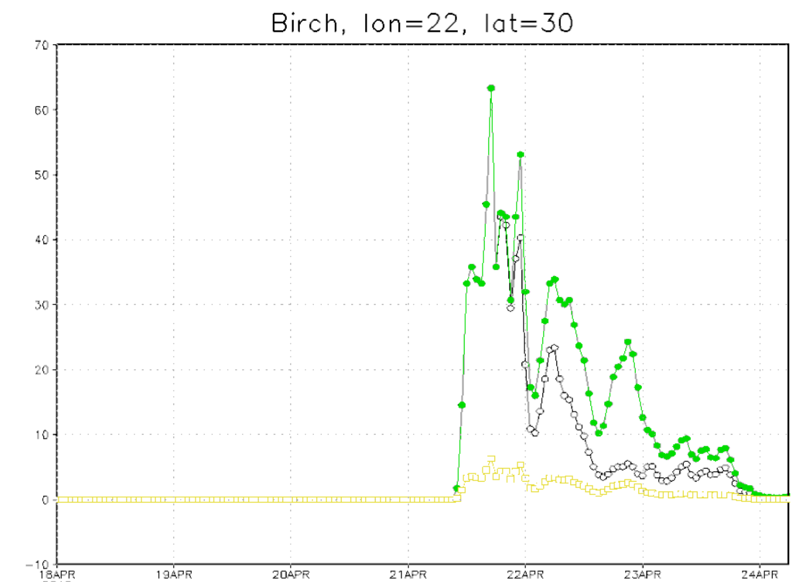



# Time series


Birch, lon=22, lat=60
reference
night perturb
day perturb
18APR 2018
19APR
20APR
21APR
22APR
23APR
24APR


Birch, lon=22, lat=50
18APR 2018
19APR
20APR
21APR
22APR
23APR
24APR


Birch, lon=22, lat=40
18APR 2018
19APR
20APR
21APR
22APR
23APR
24APR

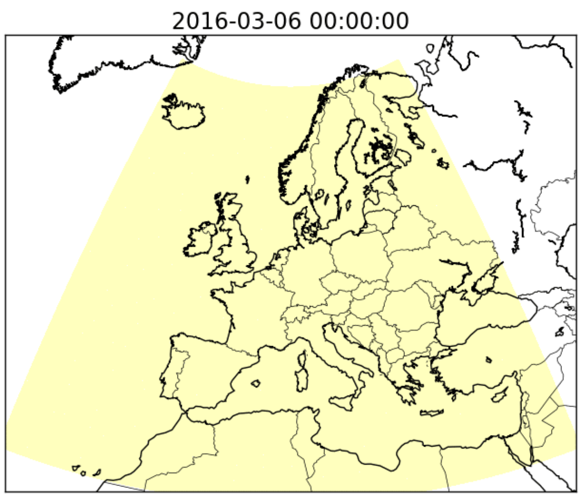

Birch, lon=22, lat=30
18APR 2018
19APR
20APR
21APR
22APR
23APR
24APR

# Summary 1

- Classical assimilation of concentrations makes little sense: the model forgets the impact much too fast
- Reason: mathematically, the system has short relaxation time, thus being driven by external forcing rather than by initial conditions

# Emergency (and not only) applications

- Sharp edges of the plumes, high-frequency variability
- Small uncertainty in wind fields generates incompatible predictions
- Example: hypothetical Etna eruption, plume predicted with two meteorological datasets: ECMWF ERA-Interim and ECMWF operational IFS




- Area with above-threshold 200 µg $m^{-3}$ ash concentrations
- The light blue areas are computed with ERA-Interim
- The light red areas are computed with IFS
- The dark purple colour indicates the areas where the threshold is exceeded using both datasets

# Summary 2.

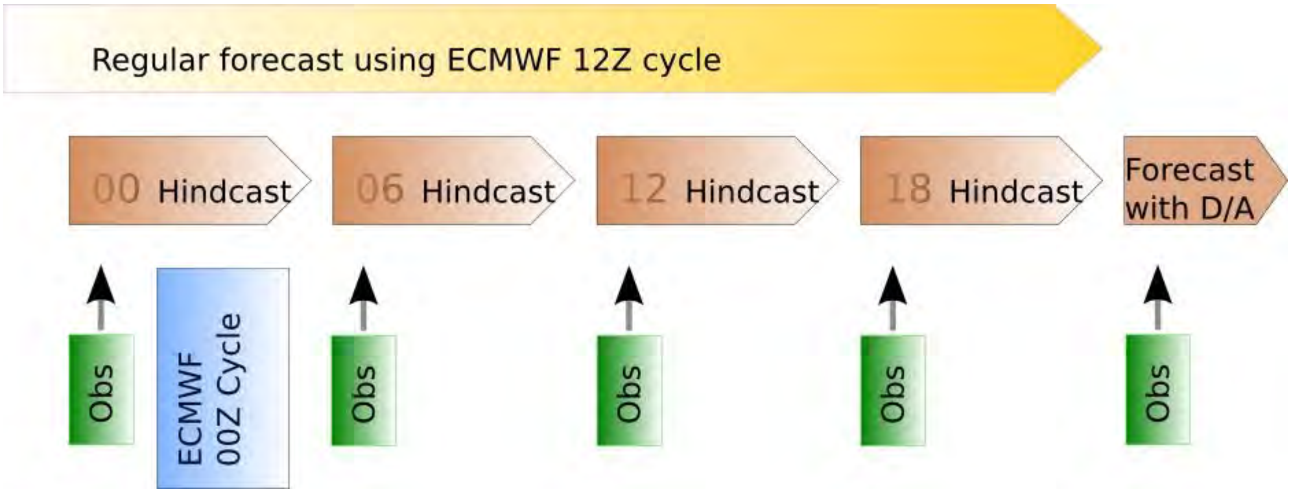
- The atmospheric composition problem is:
  - non-autonomous
  - non-linear
  - has nearly full frequency spectrum of processes from 1/sec to 1/year closely related to spatial resolution
  - has relaxation towards the forced motion of a few hours in the troposphere, a year in the stratosphere
- Observed <10% of species, strong ties with non-observed ones
  - several reservoir species, in-essence, none observed
  - observations primarily near-surface (in-situ) or column-integrated (nadir-looking satellites)
- Depending on the problem, distribution function can be strongly non-Gaussian, e.g. bi-modal in emergency applications

# How to handle such system?

- Ignore the difficulties and apply known techniques with available observations. State estimation with
  - OI / 3D-VAR
- Account for the system constraints and chemical links. State estimation with
  - 4D-VAR / EnKF
- Expand the control variable
  - include emission fluxes
  - include meteorology
- Consider non-classical forms of “DA-looking” techniques
  - data fusion
  - optimised ensemble

# Business as usual

- How bad are the problems?
- Task: the state estimation with in-situ or satellite-column data for available species
- Example:
  - ➢ SILAM chemistry transport model
  - ➢ European domain
  - ➢ Technique: 3D-VAR



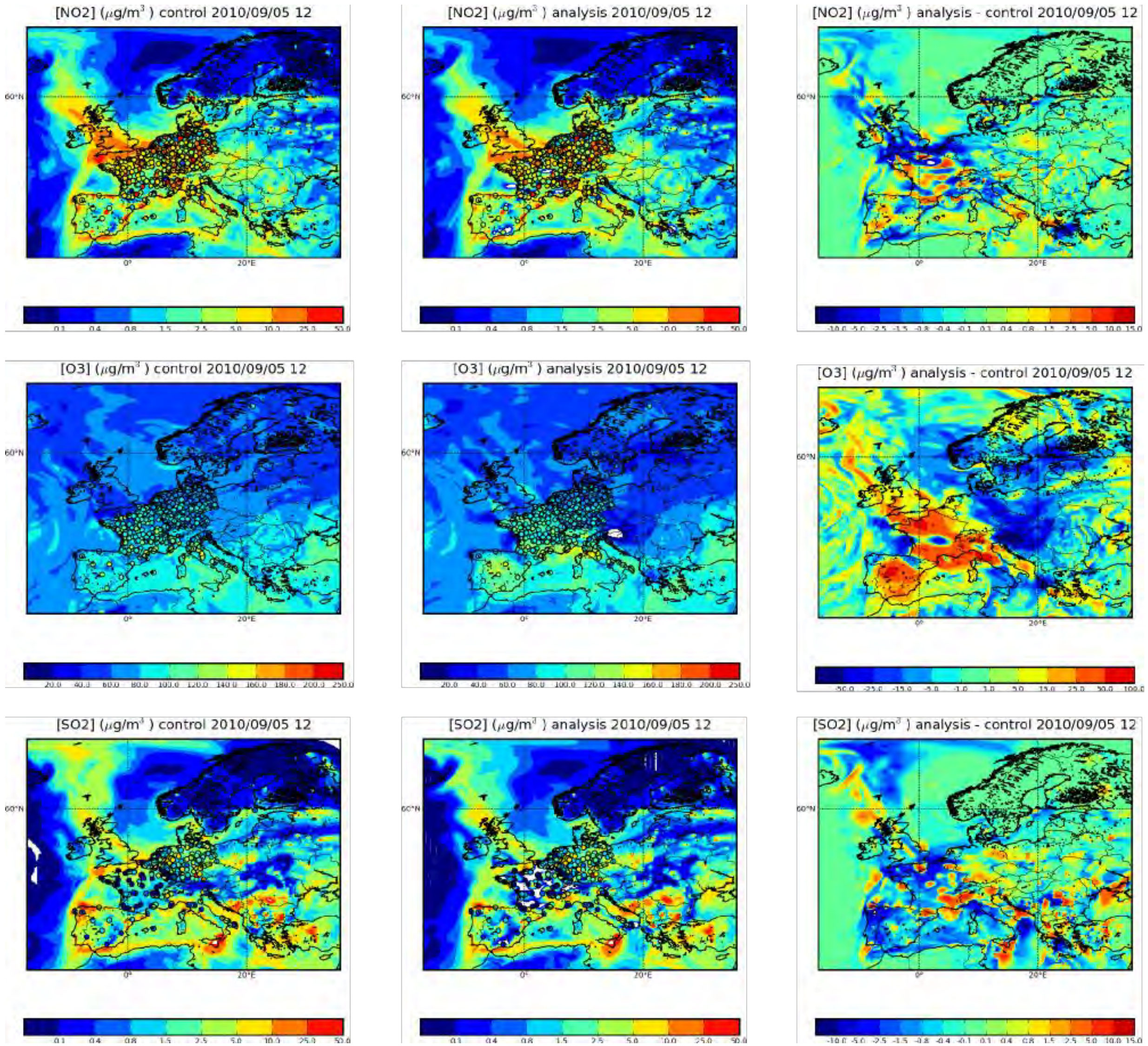
# 3D-VAR outcome for $NO_2$, $O_3$, $SO_2$

control | analysis | difference

$NO_2$

$O_3$

$SO_2$

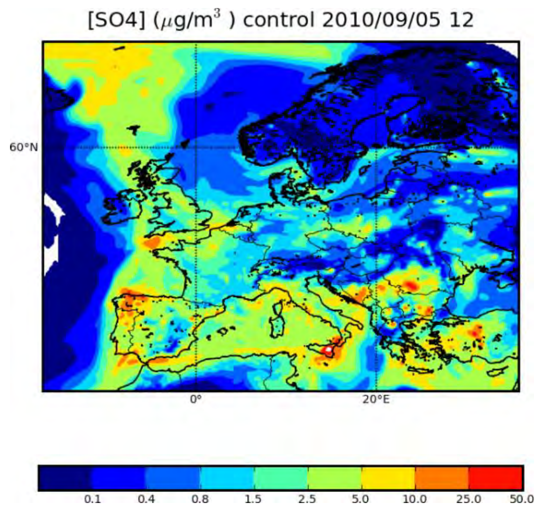


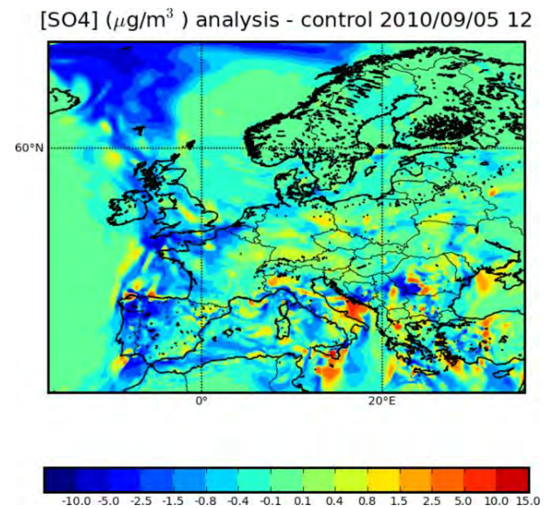
# Non-observed species? Scores?

control

difference

$SO_4$




| RMSE (μg/m³) | Reference Forecast, no DA | Analysis | Forecast with DA |
|---|---|---|---|
| O3 | 29.1 | 22.2 | 26.6 |
| NO2 | 19.3 | 17.5 | 18.5 |
| SO2 | 5.88 | 5.64 | 5.99 |
| PM2.5 | 10.1 | 9.21 | 9.33 |

# Real-life AQ case: China

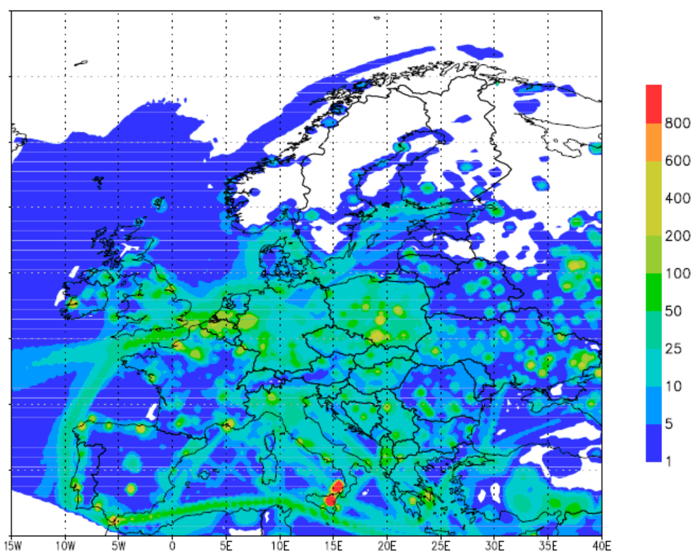
- Model inter-comparison, $PM_{2.5}$
  - China, 800 stations
  - scores over 1-14.03.2016

4D-VAR, column AOD from low-orbit satellite too infrequent, system relaxes to forced motion

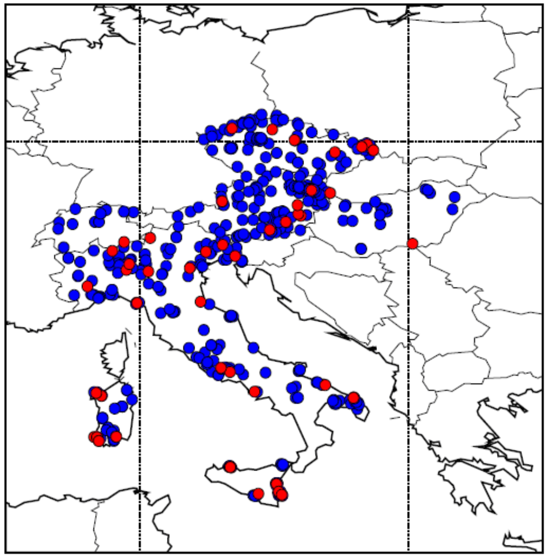
# In-depth with SOx issue

- SILAM experiment 8-22.02.2006
- 3D-VAR, 4D-VAR
- state estimation problem

$SO_2$ emission




$SO_2$ observations

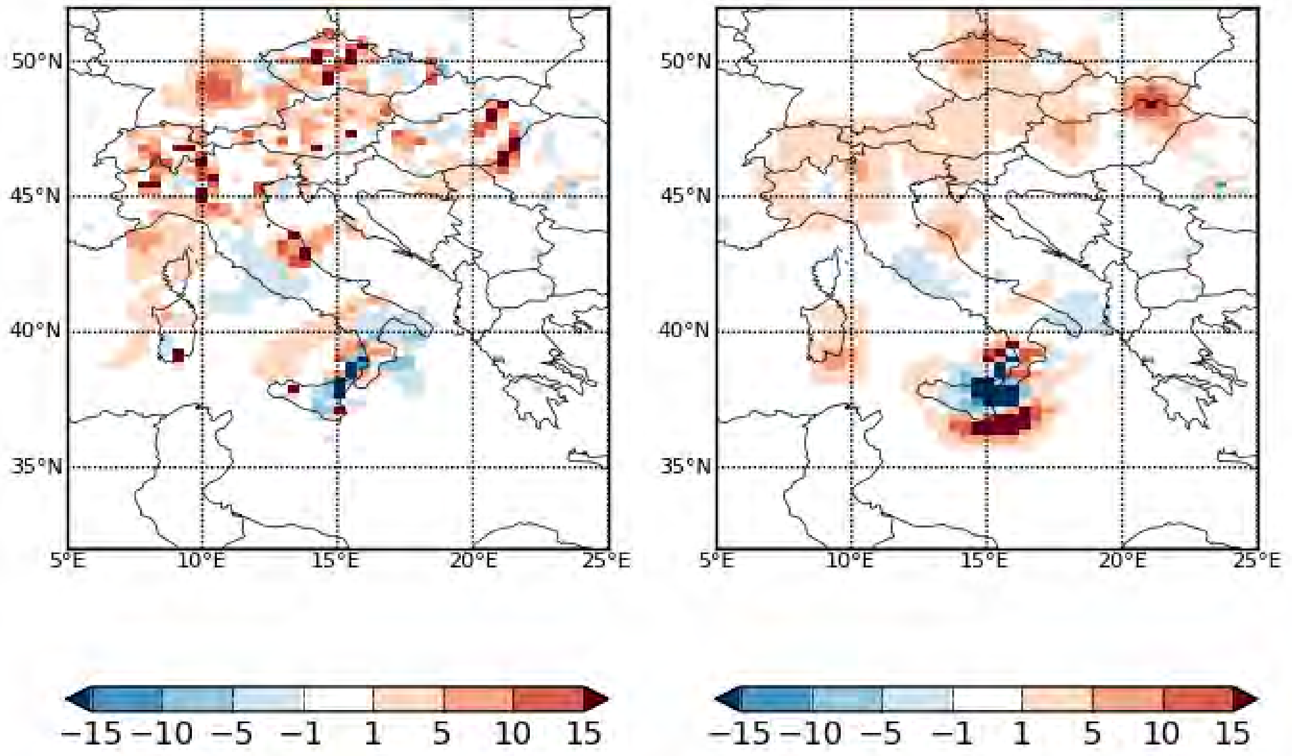


# Effect of complexity: 3D-VAR vs 4D-VAR

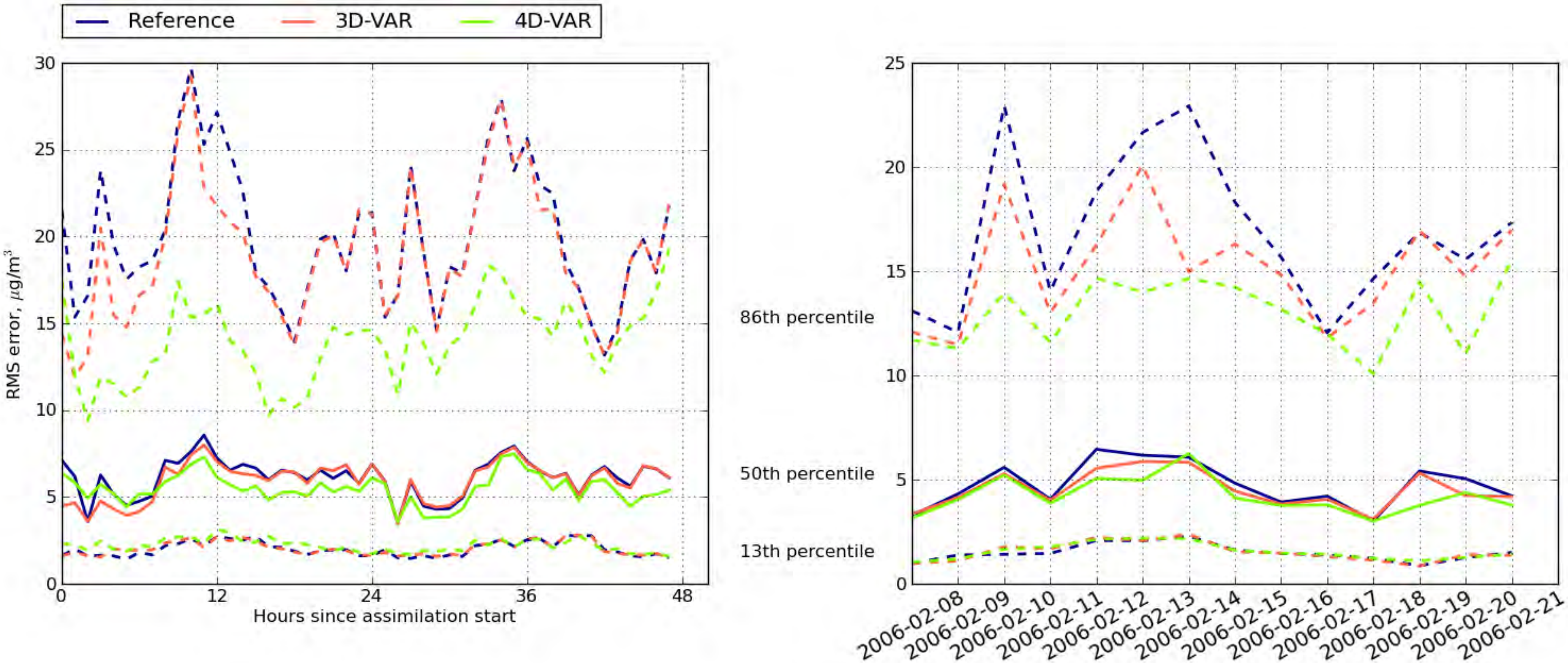
- $SO_2$ near-surface concentration, changes due to DA

4D-VAR
diagonal error covariance

3D-VAR
non-diagonal error covariance



# Effect on scores



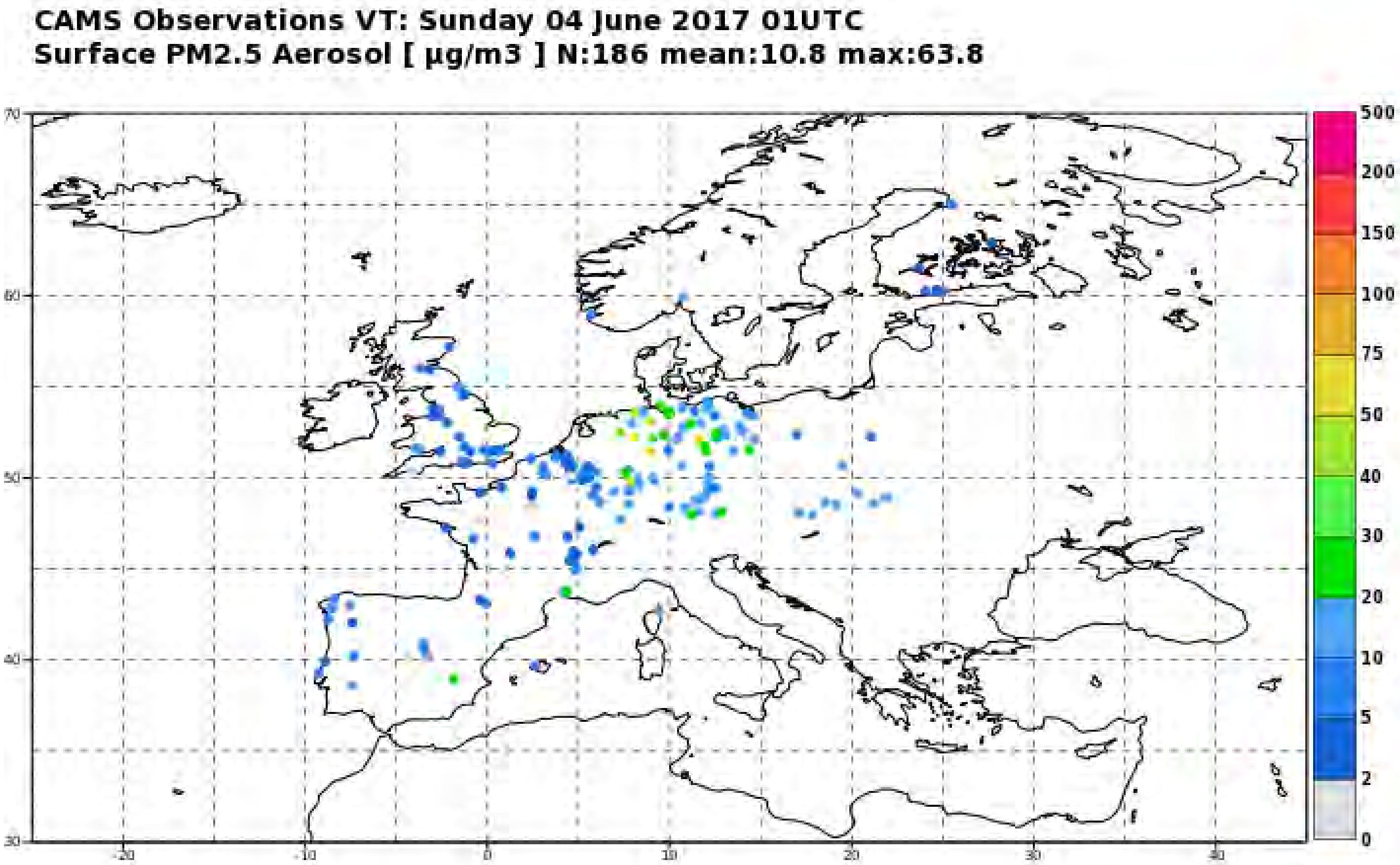
# Can we assimilate PM?

- PM is a sum of several species, i.e. not a system predicted variable or parameter, cannot be a control variable
- Let’s create a assimilation-PM, which can have positive and negative concentrations
  - that one can serve as a control variable.
  - we cannot propagate the correction to the model state
  - … but we can advect and deposit this aPM
- Examples:
  - operational SILAM analysis within Copernicus Atmospheric Monitoring Service
  - MarcoPolo model intercomparison for China

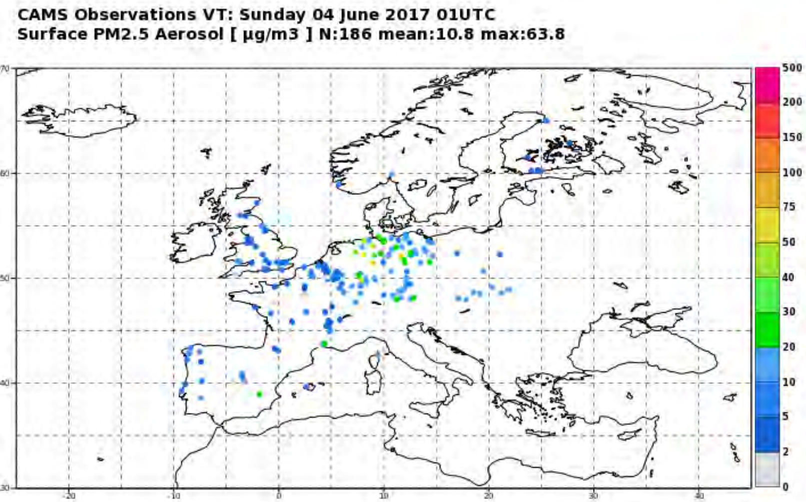
# SILAM setup in CAMS analysis

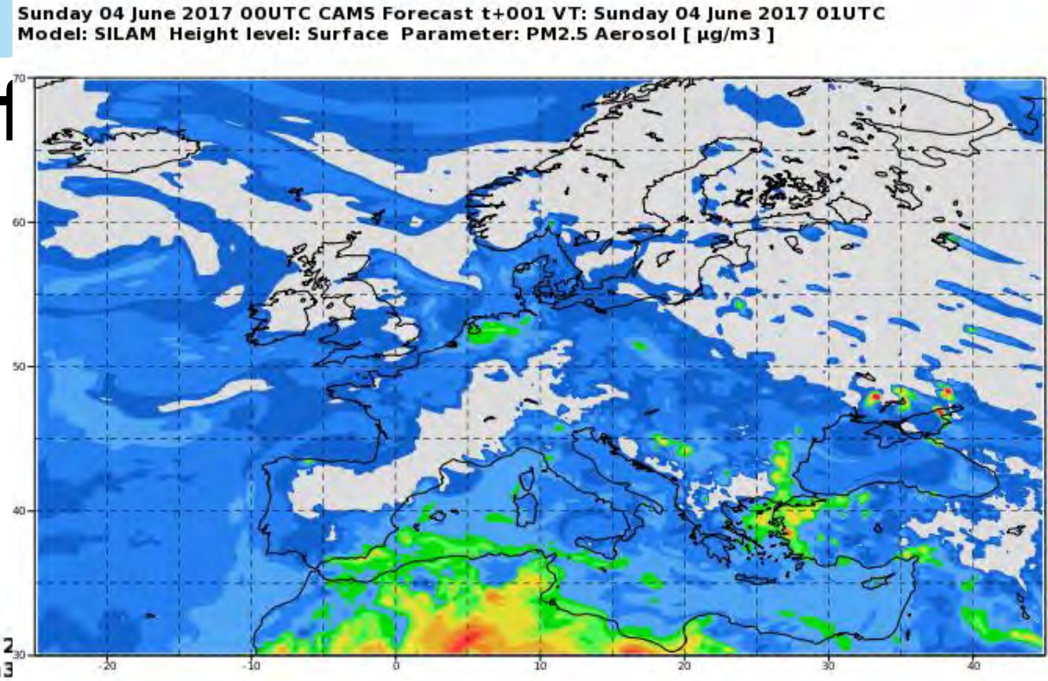
- European domain
- 10km resolution
- daily analysis of the last-day data
- in-situ observations of $NO_2$, $SO_2$ $O_3$, $PM_{2.5}$, $PM_{10}$.
- 3D-VAR
- error covariance as before: non-diagonal in all spatial dimensions
- hourly update of the model state with no chemistry at the state update step
- full chemistry during the model time integration between the assimilation steps

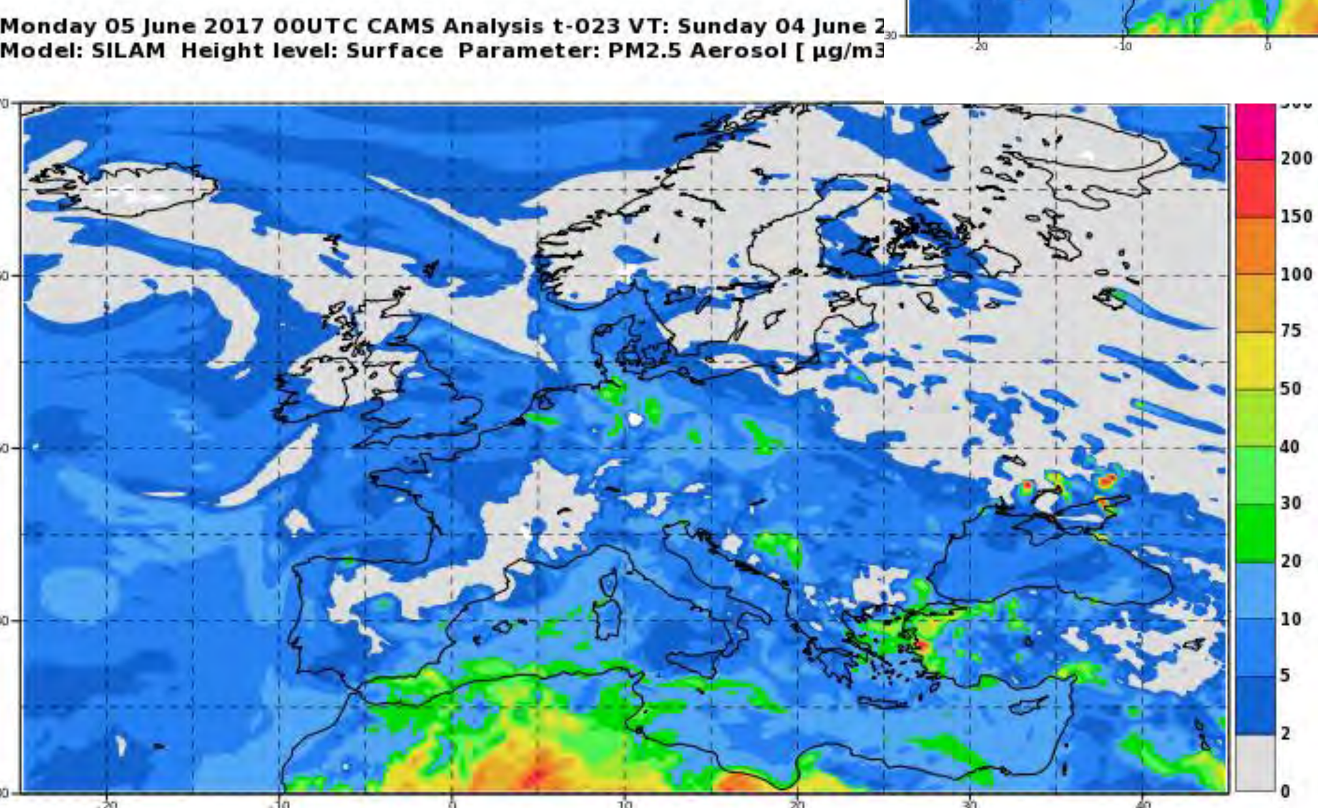
# $PM_{2.5}$ last Sunday: observations



of



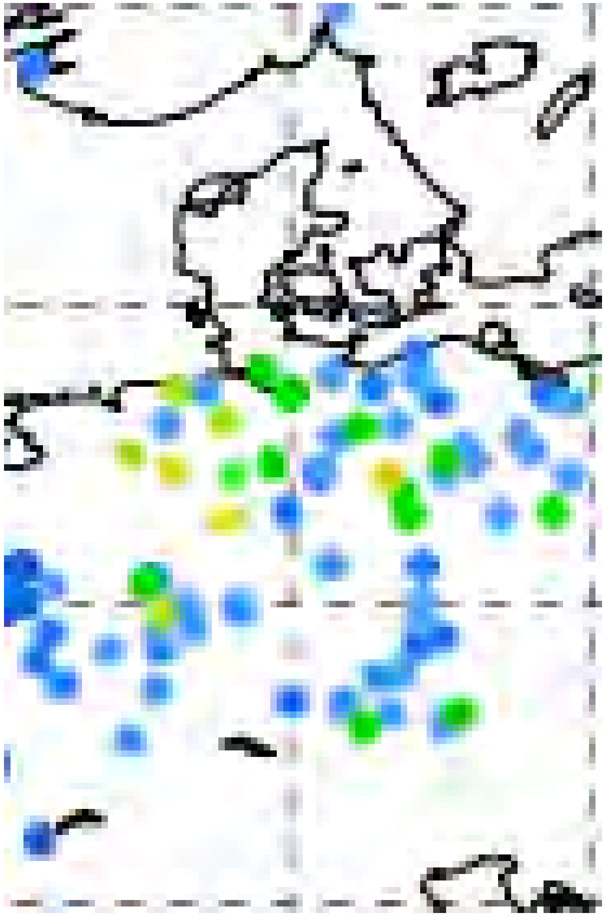




↑ Forecast

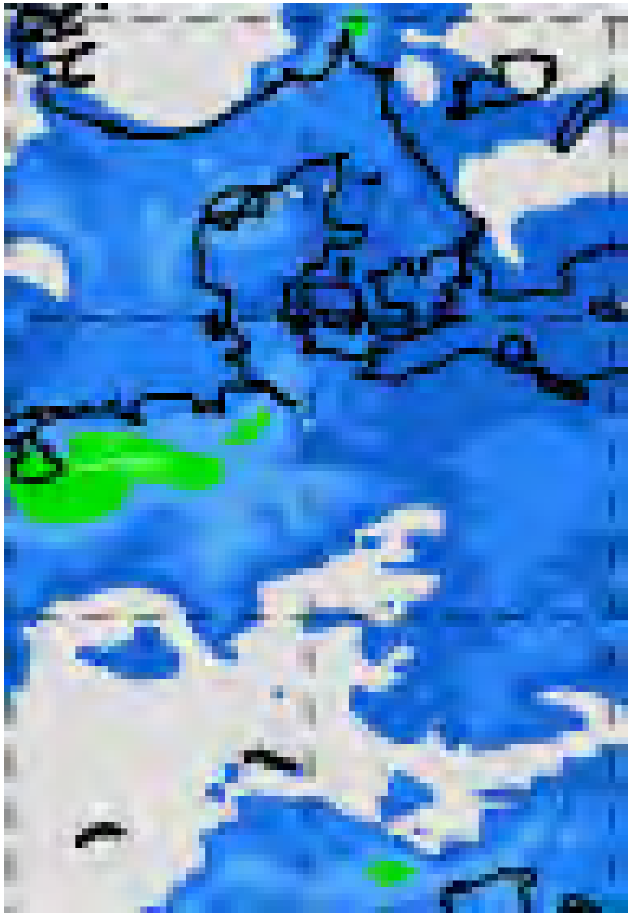
← Analysis

# Zoom towards high-variability area

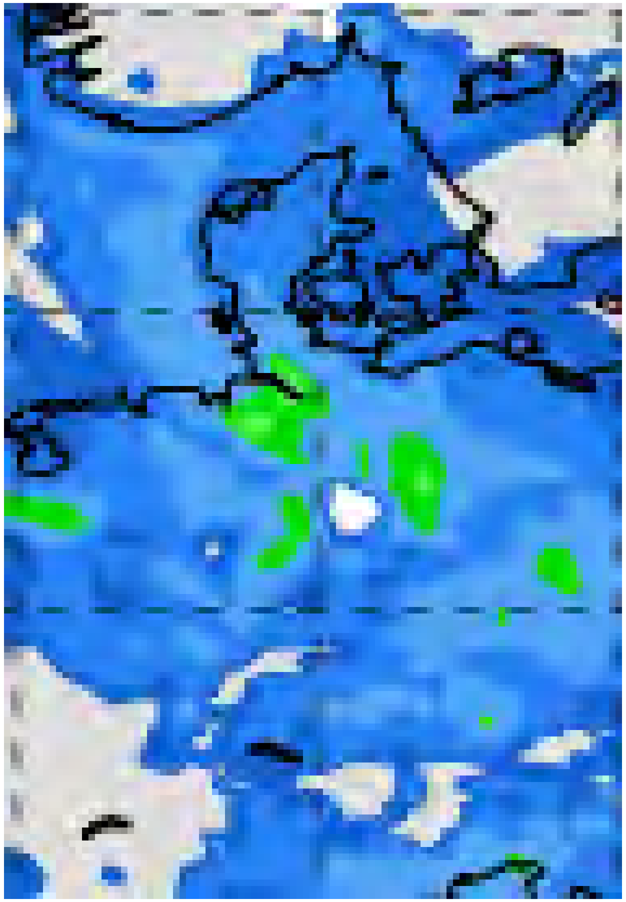
Observations



Forecast



analysis 3D-VAR



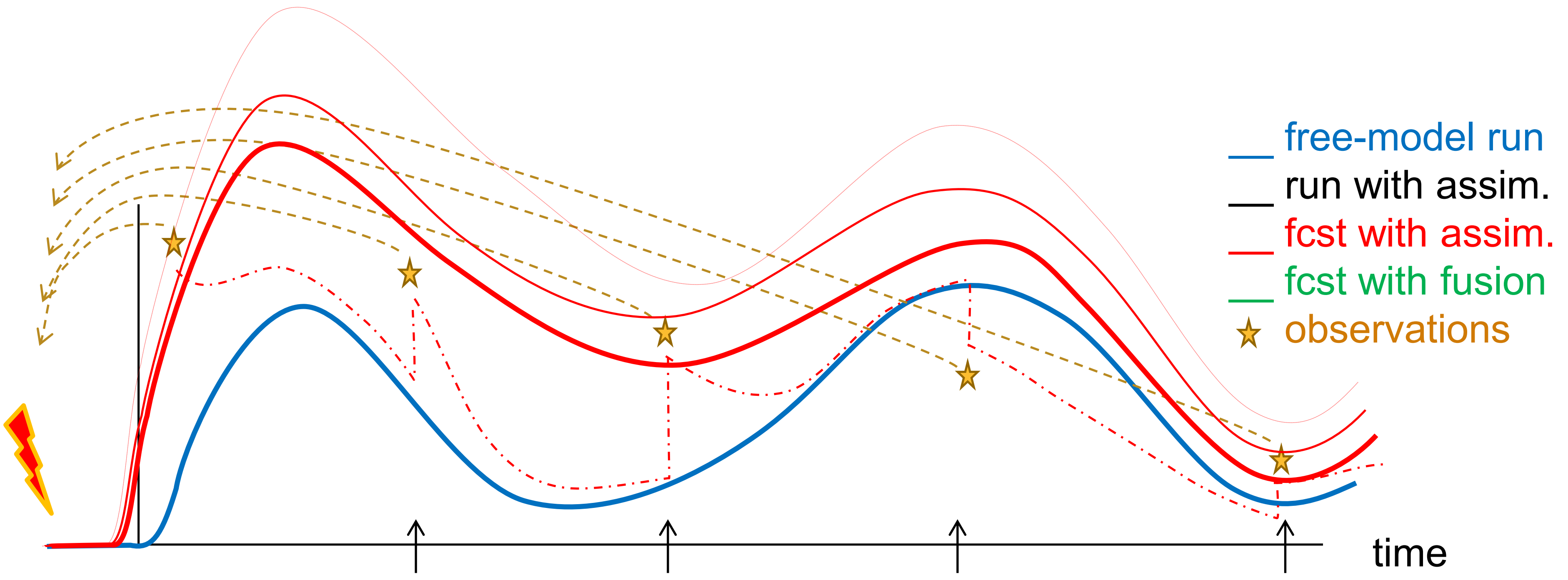
# How to handle such system?

- Ignore the difficulties and apply known techniques with available observations. State estimation with
  - OI / 3D-VAR
- Account for the system constraints and chemical links. State estimation with
  - 4D-VAR / EnKF
- Expand the control variable: find what has longer impact
  - include emission fluxes
  - include meteorology
- Consider non-classical forms of “DA-looking” techniques
  - data fusion
  - optimised ensemble

# Expand the control variable

- Reminder:
  - predicted variables are concentrations of many species
  - their assimilation does not make much sense due to short model memory
- Can we find something that does have a longer impact?
- Controlling parameters:
  - emission fluxes
  - meteorological data
  - model internal parameters and coefficients

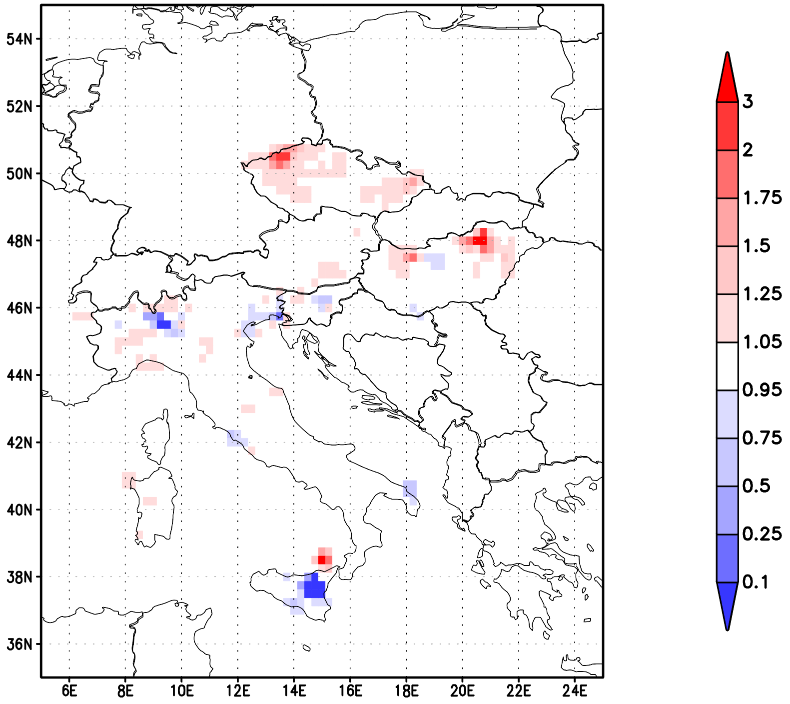
# Source term inversion



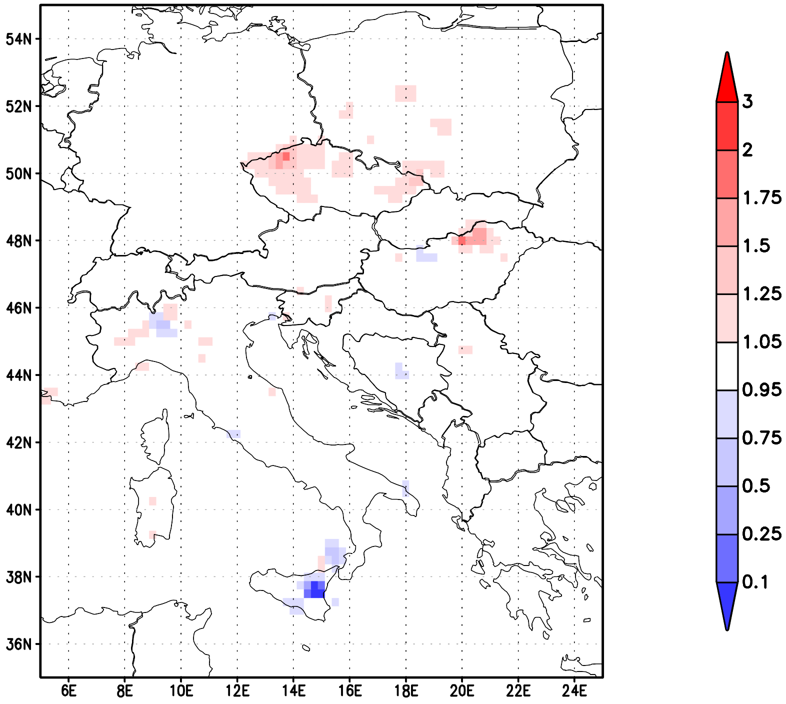
# Emission correction factor

- Same SOx experiment, now with 4D-VAR towards emission

Day 1 correction



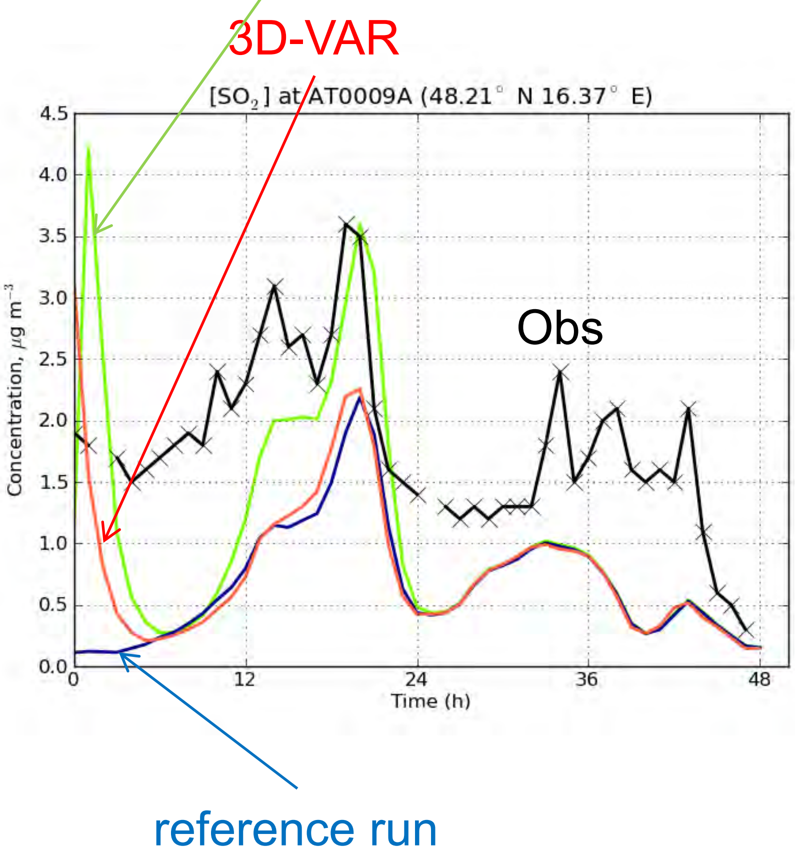

Weeks 1-2 mean correction



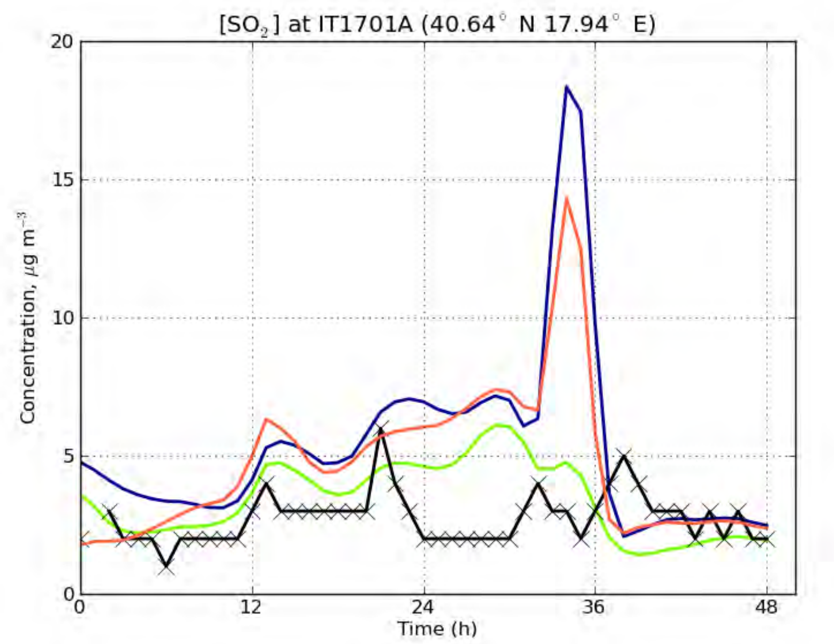
# Comparison of the approaches

4D-VAR state+emissoin

Site: AT 9, Austria



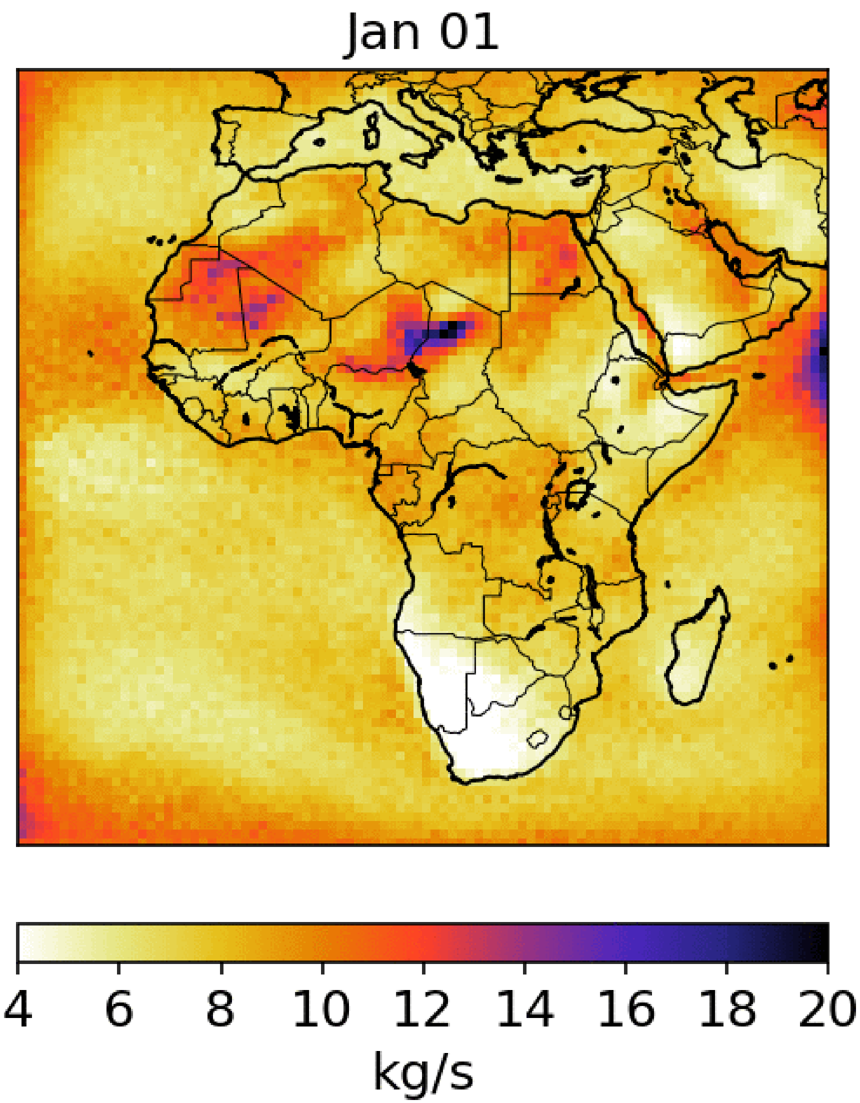

Site: IT 17, Italy



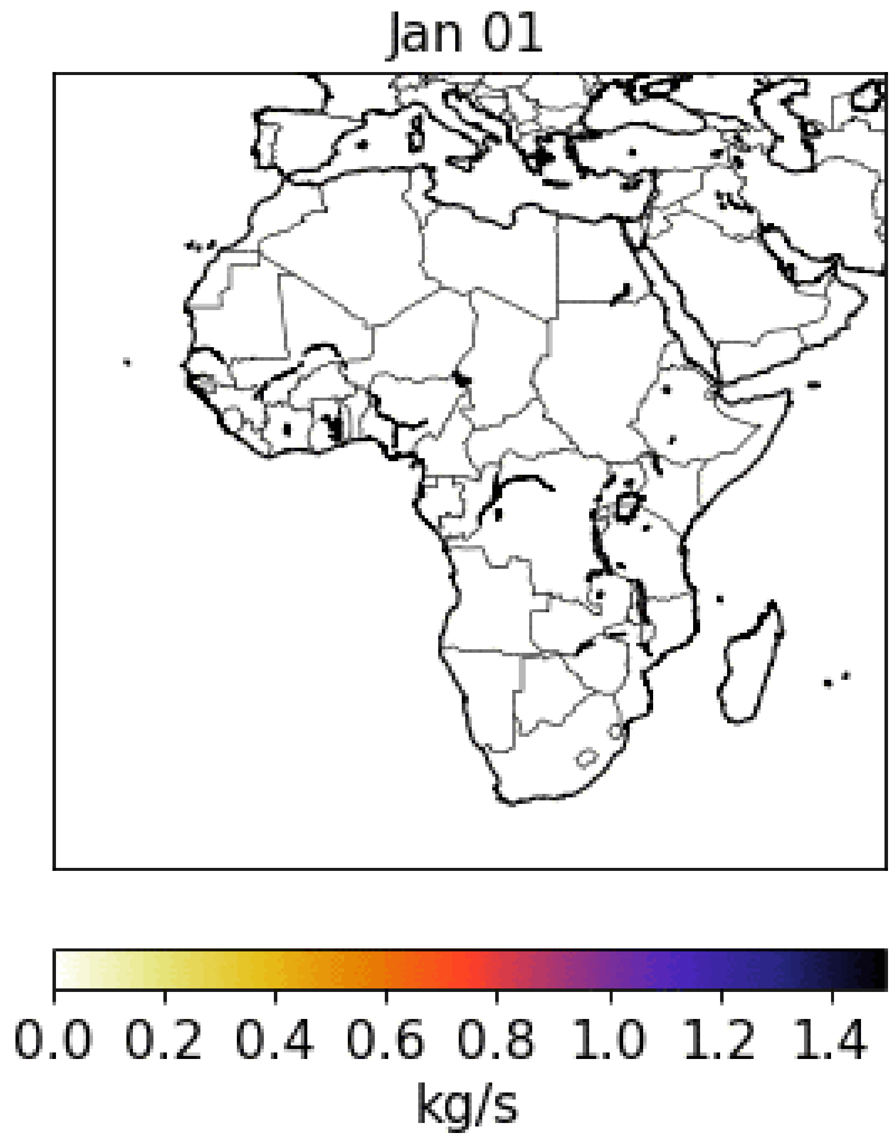
# African emission experiment

- Experiment concept: construct African emission from scratch, without any prior knowledge
- Input: MODIS AOD, full 2016
- Starting point: constant homogenous emission all over domain
- Method: SILAM EnKF assimilation of emission correction factor
- Evaluation: Aeronet for full 2018

**PM emission**




**AOD**



**RMSE**

| | SILAM old | SILAM test | SILAM 2018 | SDS-WAS median |
|---|---|---|---|---|
| Sahel/Sahara | 0.39 | 0.34 | 0.30 | 0.31 |
| Middle East | 0.42 | 0.27 | 0.26 | 0.33 |
| Mediterranean | 0.20 | 0.15 | 0.18 | 0.15 |
| All regions | 0.35 | 0.29 | 0.26 | 0.28 |

**Correlation coefficient**

| | SILAM old | SILAM test | SILAM 2018 | SDS-WAS median |
|---|---|---|---|---|
| Sahel/Sahara | 0.39 | 0.42 | 0.47 | 0.75 |
| Middle East | 0.13 | 0.56 | 0.56 | 0.52 |
| Mediterranean | 0.60 | 0.57 | 0.56 | 0.72 |
| All regions | 0.44 | 0.47 | 0.52 | 0.74 |

**Bias**

| | SILAM old | SILAM test | SILAM 2018 | SDS-WAS median |
|---|---|---|---|---|
| Sahel/Sahara | -0.22 | -0.15 | -0.03 | -0.18 |
| Middle East | -0.21 | -0.10 | -0.02 | -0.18 |
| Mediterranean | -0.14 | 0.00 | 0.06 | -0.10 |
| All regions | -0.20 | -0.10 | 0.00 | -0.16 |

**Fractional gross error**

| | SILAM old | SILAM test | SILAM 2018 | SDS-WAS median |
|---|---|---|---|---|
| Sahel/Sahara | 0.88 | 0.59 | 0.49 | 0.56 |
| Middle East | 0.88 | 0.43 | 0.37 | 0.56 |
| Mediterranean | 1.16 | 0.43 | 0.44 | 0.85 |
| All regions | 0.96 | 0.53 | 0.46 | 0.64 |

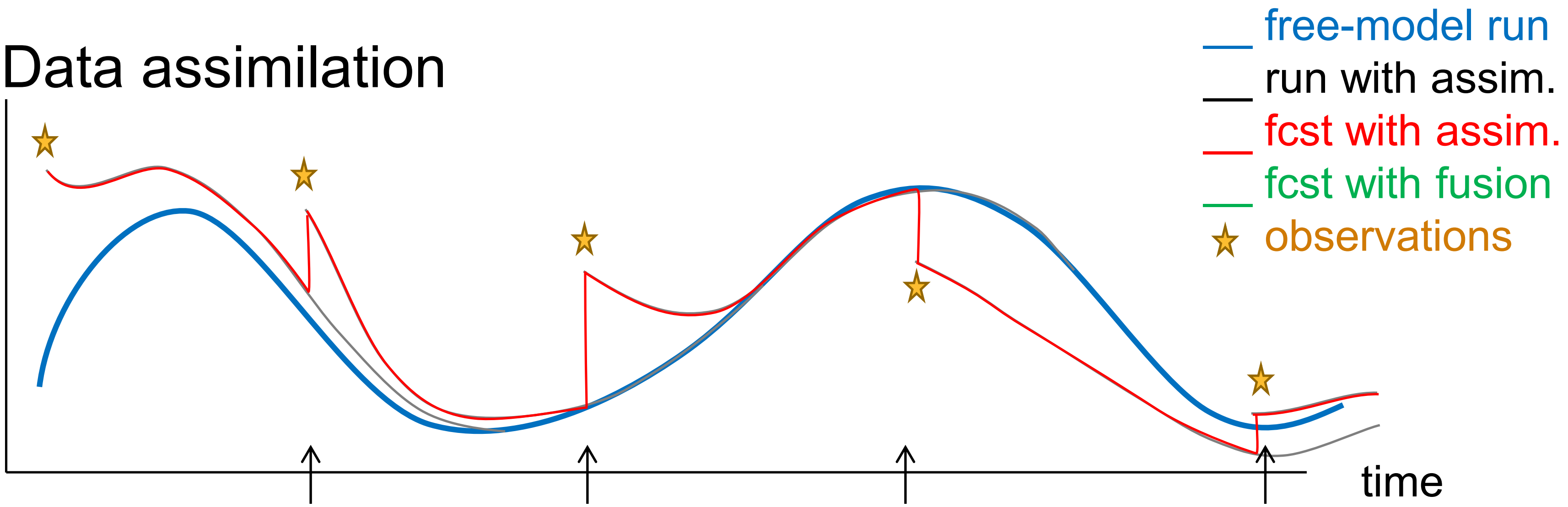
Evaluation model run outperforms not only SILAM operational skills but also SDS-WAS ensemble (over 10 models!)

# How to handle such system?

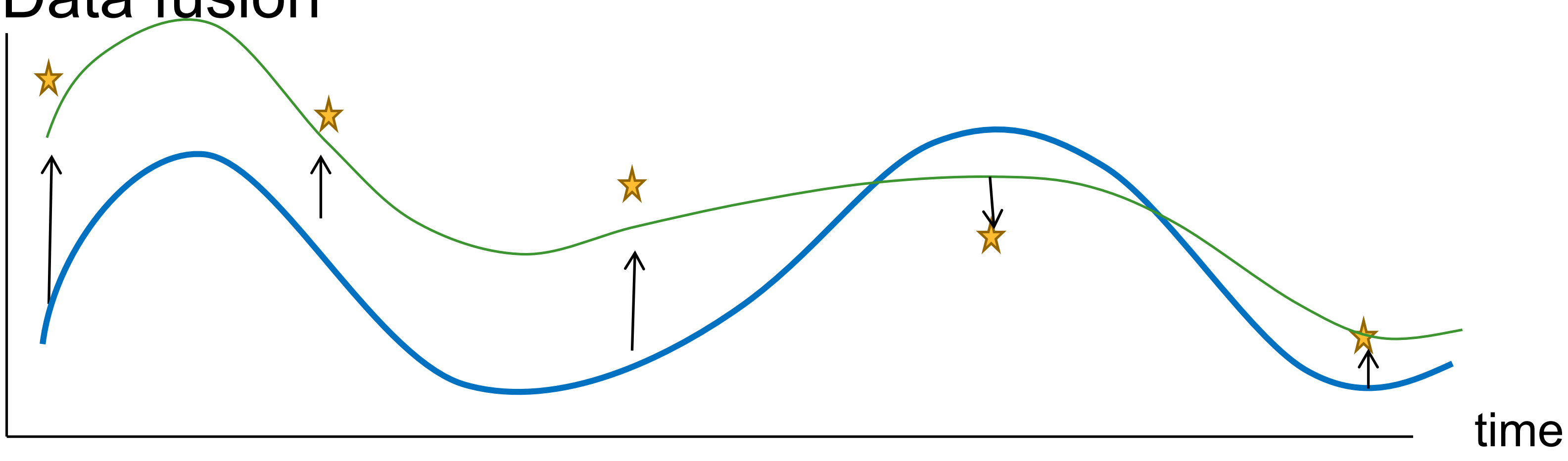
- Ignore the difficulties and apply known techniques with available observations. State estimation with
  - OI / 3D-VAR
- Account for the system constraints and chemical links. State estimation with
  - 4D-VAR / EnKF
- Expand the control variable: find what has longer impact
  - include emission fluxes
  - include meteorology
- Consider non-classical forms of “DA-looking” techniques
  - data fusion
  - optimised ensemble

# Ways of involving observations

- Data assimilation



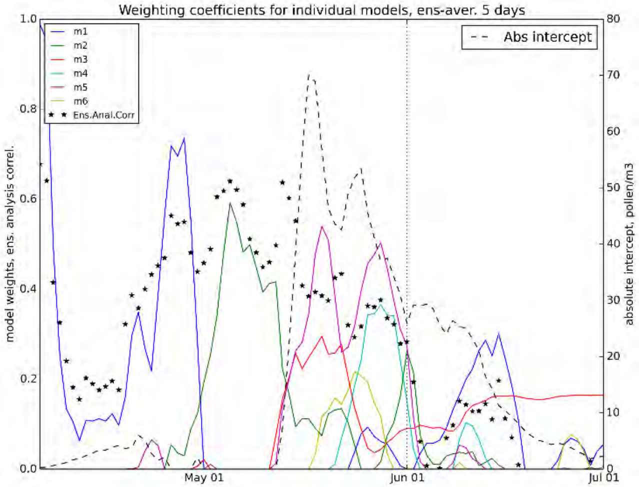

- Data fusion



# Data fusion vs data assimilation

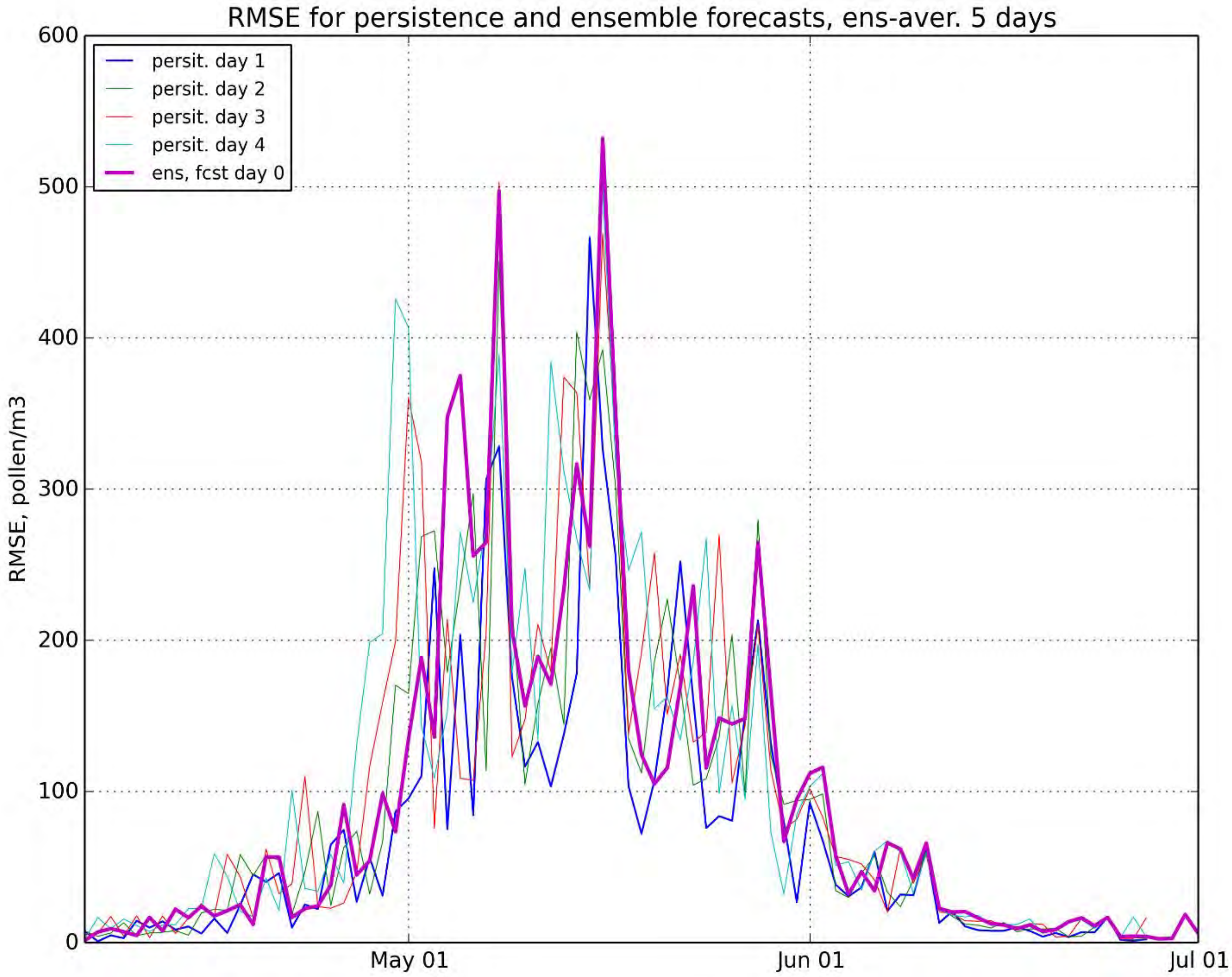
- DA: data are used to adjust model internal variables, parameters or forcing
  - model is “informed” about deviation from the observations and asked to behave better
- DF: data are used to adjust model output after the simulations are finished
  - model has no clue about its errors, it runs without feedback from observations
  - all corrections are applied as post-processing of the model predictions
- A simple example: bias correction
- Promising: error of model predictions (e.g., bias) can be less varying than the predictions themselves

# Ensemble-based data fusion: works!




Weights of individual models

Error of individual models, simple and optimised ensemble



# Summary

- Atmospheric composition is tough for data assimilation: violates almost all assumptions behind DA methods
  - non-linear, non-autonomous, non-Gaussian, correlated errors, very small fraction of observed phase space
- Classic methods give <20% of improvement for the analysis, next to nothing for the follow-up forecast
  - Still, useful in some (few) applications
- Expansion of control variable is among the most-promising approaches
  - Has longer forecasting horizon and wider correlation distance
  - Own complexity: adjoint and ensemble generation
- Data fusion technology shows very promising first results
  - can be applied together with data assimilation: fully independent approach