NILU Norsk institutt for luftforskning
Norwegian Institute for Air Research

Toward Assimilation of Crowdsourcing Data using the EnKF

William Lahoz and Philipp Schneider
NILU: wal@nilu.no

Thanks to Sam-Erik Walker

EnKF Workshop 2014, Steinsland, Os, Norway
24 June, 2014

Outline

- Need for information
  - Examples
  - Data assimilation
- Crowdsourcing – a novel information source
  - What is it?
  - Mobile phone use
  - The EU Citizens' Observatory -> what the citizen needs
- Data assimilation and crowdsourcing – NILU effort
  - The roadmap: observations, model and DA
  - The challenges: spatio-temporal scales
  - What is being done – early results
- Outlook for data assimilation and crowdsourcing
  - Dealing with the challenges

Need for information

Need for information:

Main challenges to society require information for an intelligent response, including making choices on future action examples:

- Climate change
- Impact of extreme weather
- Environmental degradation:

Loss of natural habitat, impact on biodiversity, impacts of pollution (water, air)

We can take action according to information obtained:

- Future behaviour of system of interest, future events – prediction
- Test understanding of system & its dynamic response & adjust understanding – hypothesis testing
- Assess the Earth Climate System (e.g. climate change) – monitoring

Data assimilation: combine observations + models + errors

What is crowdsourcing?

Citizen Science:

A novel & recent development for observing the Earth System provided by activities from citizens involved in Science – people accumulating knowledge to learn about & respond to environmental threats & as public participation in scientific research.

Crowdsourcing:

Associated with Citizen Science

«The act of taking a job traditionally performed by a designated agent (usually an employee) & outsourcing to an undefined, generally large of people in the form of an open call» *Howe (2010)*

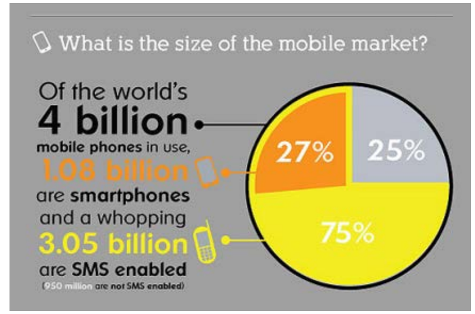
Examples:

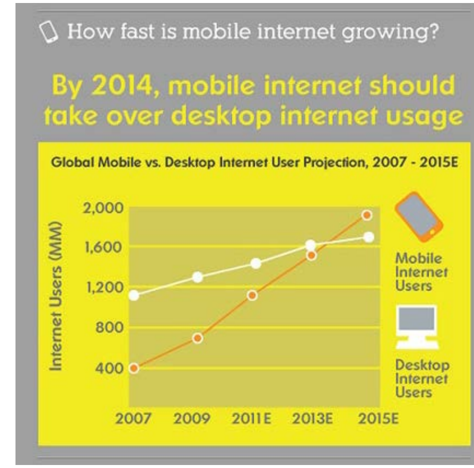
Observations by amateurs of birds & butterflies - monitoring the environment

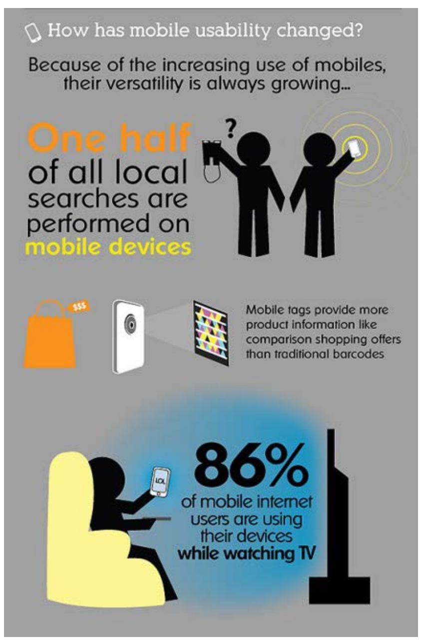
Lahoz and Schneider 2014, Front. Env. Sci.

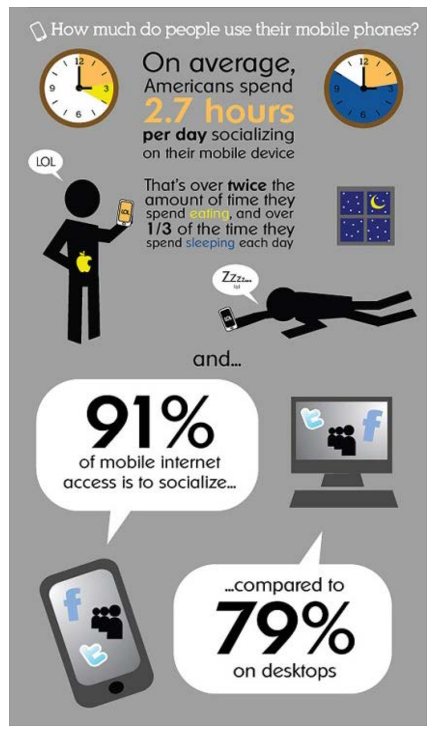
Citizens' Observatory

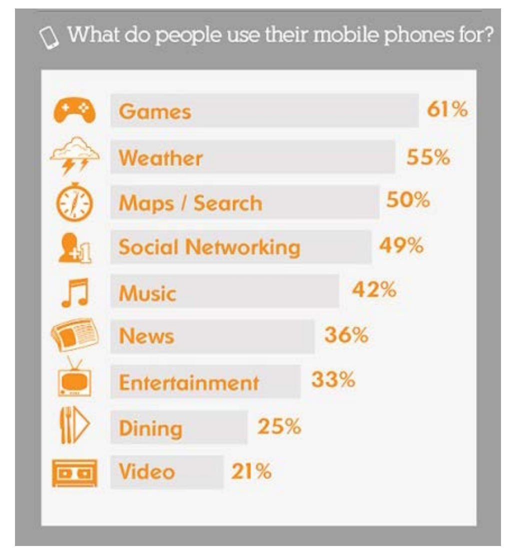
- ***Growth*** in mobile use
- ***Change*** in mobile usage
- ***Increasing*** range of features







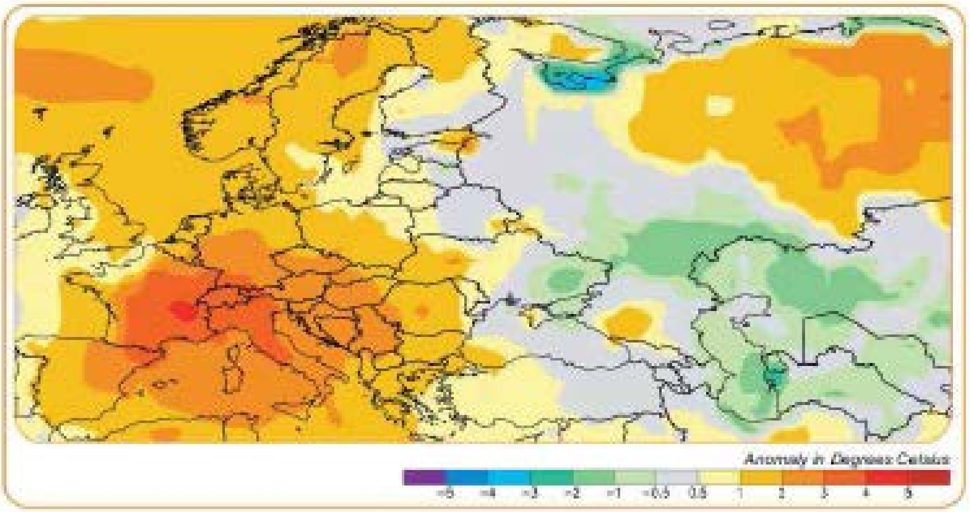




Source: http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics

Societal concern: health and economic cost (Billions of Euros)

European Summer of 2003



Temperature anomaly (°C)

June-Aug 2003 (Europe)

Climatological base period 1998-2003

Red +ve anomalies; blue –ve anomalies

(Courtesy UNEP)

Estimated European heat wave of 2003 caused loss of 14802 lives (mainly elderly) in France (http://www.grid.unep-ch/product/publication/download/ew_heat_wave.en.pdf)

High temperatures increase tropospheric O_3 amounts, & anticyclonic conditions ensured their persistence (*Vautard et al., Atmos Env., 2005*)

Potential application of crowdsourcing

Data assimilation & crowdsourcing

Crowdsourscing: New work at NILU – CITI-SENSE project

The roadmap:

- Observations: microsensors (static/mobile platforms); citizens
- Model: EPISODE air quality model for Oslo
- Data Assimilation: EnKF, SQRT variant from *Sakov and Oke 2008*

The challenges – technical, implementation:

- Spatio-temporal scales – «street level»: what citizen wants
- Characterization of errors
- Providing user-friendly information

What is being done at NILU – early results

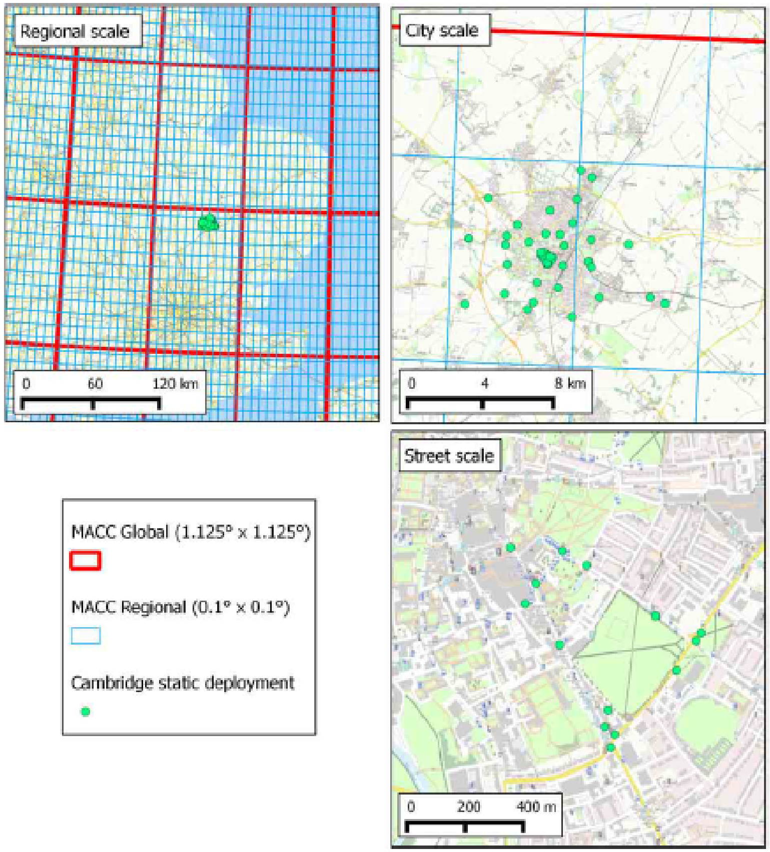


FIGURE 9 | Illustration of significant differences in spatial scale between operational atmospheric modeling and typical data assimilation applications (case 1); and urban air quality applications (case 2). Spatial scales associated with case 1 are exemplified by the global and regional grids used by the MACC-II project as a precursor of the Copernicus Atmospheric Monitoring Service—top left-hand panel (labeled regional scale). Spatial scales associated with case 2 are exemplified by the observations of gases relevant for urban air quality (CO, NO, and NO_2) collected by low-cost, high-density monitoring networks by the University of Cambridge—top right-hand panel (labeled city scale), and bottom right-hand panel (labeled street scale). Spatial resolutions of the global and regional scale MACC models identified in the top two panels are, respectively, 1.125° × 1.125° and 0.1° × 0.1°. The University of Cambridge data are described in Mead et al. (2013).

The challenges:

- Significantly different spatial scales vs NWP (street level vs c. 10 km)
- Model development (smaller spatial scales)
- Noisy information from users/microsensors
- User-friendly representation of uncertainty

- Merging of data from traditional sources (satellite, in situ) with Citizen Science data

- Quality of data from low-cost sensors
- Data security & privacy

Challenges addressed in EU-funded CITI-SENSE project

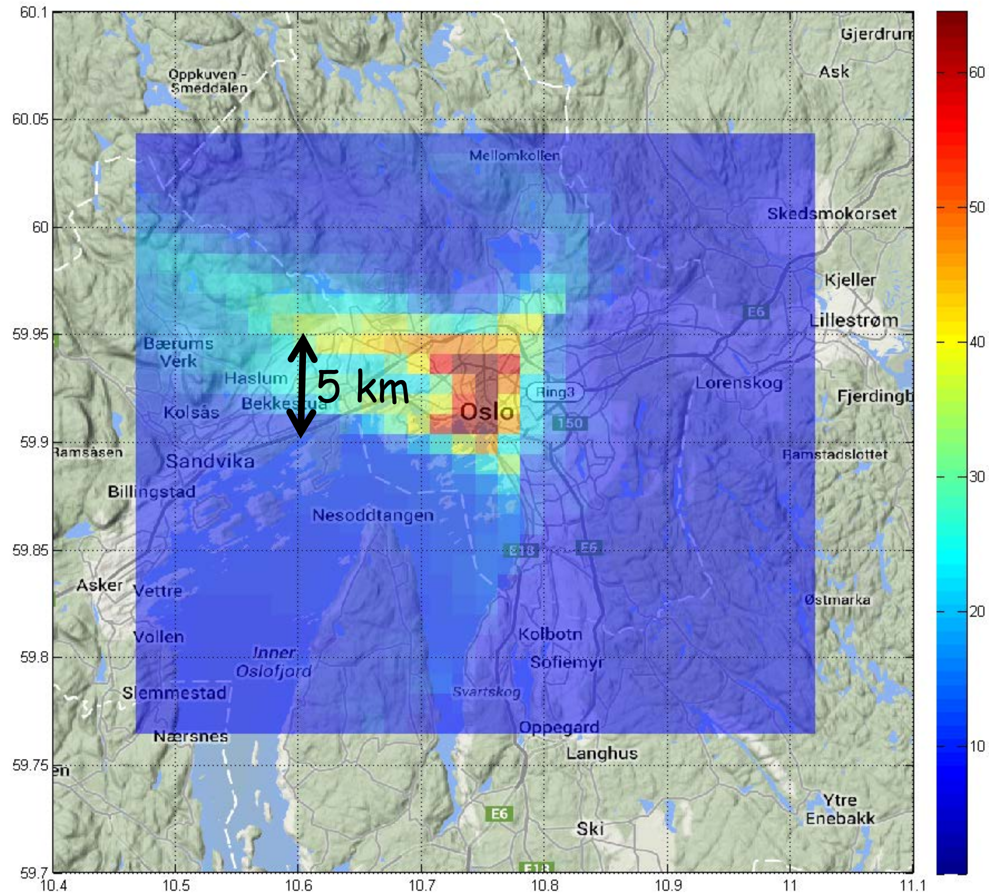
Also: NWP going to smaller spatial scales
 - e.g. for convection

WOW project at UK Met Office
http://wow.metoffice.gov.uk

Model

The EPISODE model

- Developed by *Slørdal et al. (2008)*
- 3-D combined Eulerian / Lagrangian air pollution dispersion model, developed at NILU
- Main focus on urban & local-to-regional scale applications
- Provides gridded fields of ground-level hourly average concentrations
- Spatial resolution down to 100m
- Time step between 10 s and 300 s
- Schemes for advection, turbulence, deposition, and chemistry

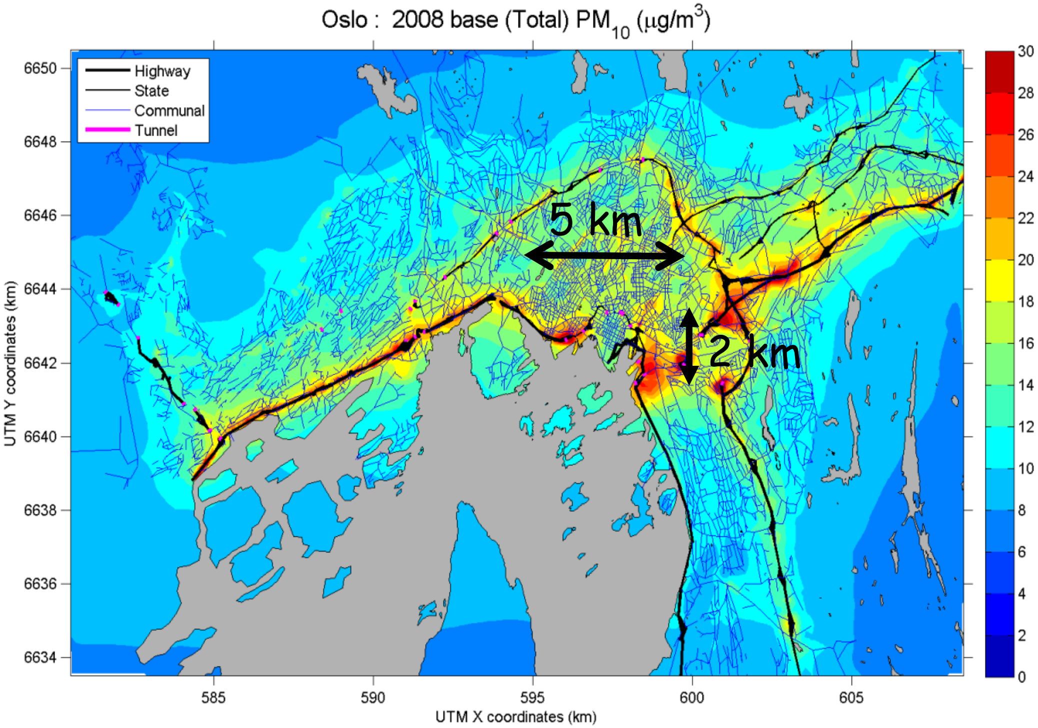


Example output for NO_2 from the EPISODE model over Oslo, here at 1 km spatial resolution.

Data fusion: test concepts toward challenging DA approach
Application of Land User Regression – LUR

- Any spatially exhaustive dataset related to observation
- In LUR this is generally land use, traffic etc.
- Output from high-resolution dispersion model
- Or all of the above…

- LUR provides input dataset for geostatistical data fusion by residual kriging, conceptually simple way to simulate & test the combination model/obs



High-resolution map of PM_{10} in Oslo from the EPISODE dispersion model. These maps are ideally suited as a spatially distributed auxiliary dataset.

Data assimilation

Two methods from *Sakov & Oke*:

- EnSRKF - Ensemble Transform Kalman filter (ETKF) using a symmetric Ensemble Transform Matrix (ETM) – MWR 2008
- DEnKF- Deterministic Ensemble Kalman Filter (DEnKF) using a linear approximation to the Ensemble Square Root Filter (ESRF) update matrix – Tellus 2008

Code implementation:

- Windows 7 and Visual Studio 2012
- Intel Visual Fortran Composer XE 2013
- Intel Math Kernel Library 11.1
    - Basic Linear Algebra Subprograms (BLAS)
    - Linear algebra package (LAPACK)
- Ensemble Kalman Filter Fortran module
    - Common ensemble methods routines
    - ETKF with symmetric ETM subroutine
    - DEnKF subroutine

Data assimilation for the Oslo AQ forecast system (Bedre Byluft)

- The system calculates 2-day forecasts of NO_2, PM_{10} and $PM_{2.5}$ hourly conc. in a grid (29 x 18 x 35) (1 km) and at individual receptor points (AQ stations);
- Data assimilation is introduced to improve the initial conc. fields in the dispersion model (EPISODE) for each 2-day forecast using available AQ obs. at the stations;
- For this purpose we use the mean preserving ETM ensemble square root Kalman Filter from Sakov & Oke (2008);
- We are in the early stages of development of this system and run tests for the period 2 Dec – 8 Dec 2013 (Mon-Sun) using 8 ensemble members (1 control + 7 perturbed).

AQ stations proxy for crowdsourcing information

- Episode model run on an hourly basis, using hourly emissions, meteorology & background conc.

- Internal time step in Episode for numerical solution of advection-diffusion equations varies with meteorology (most notably with wind speed), but is typically between 30 and 120 seconds, c. 60 timesteps per hour of simulation

- Every day at midnight (24h) we assimilate AQ obs. from one or more stations in Oslo from the same hour (24h) - i.e., current time window for assimilation is 1 hr

- This updates the initial conc. fields for Episode each day, i.e., for the next 48h forecast

EnSRKF (ETKF with symmetric ETM) – N ensemble members

$$\mathbf{X}^f = \left[\mathbf{X}_1^f, \ldots, \mathbf{X}_N^f\right]; \quad \mathbf{x}^f = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i^f$$

Forecast

$$\mathbf{A}^f = \left[\mathbf{A}_1^f, \ldots, \mathbf{A}_N^f\right] = \left[\mathbf{X}_1^f - \mathbf{x}^f, \ldots, \mathbf{X}_N^f - \mathbf{x}^f\right]$$

Forecast anomaly

$$\mathbf{P}^f = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{X}_i^f - \mathbf{x}^f)(\mathbf{X}_i^f - \mathbf{x}^f)^T = \frac{1}{N-1}\mathbf{A}^f\mathbf{A}^{fT}$$

Background/forecast errors

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^f) \quad \mathbf{K} = \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}$$

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^f$$

Analysis and analysis errors

$$\mathbf{A}^a = \mathbf{A}^f \mathbf{T}$$

Update ensemble anomalies via ETM **T**
Match eqn for $\mathbf{P}^a$
Analysed anomalies remain zero-centred

$$\mathbf{T} = \left[\mathbf{I} + \frac{1}{N-1} \left(\mathbf{HA}^f \right)^T \mathbf{R}^{-1} \left(\mathbf{HA}^f \right) \right]^{-1/2} \quad ; \quad \mathbf{S} = \mathbf{HA}^f$$

$$\mathbf{I} + \frac{1}{N-1} \mathbf{S}^T \mathbf{R}^{-1} \mathbf{S} = \mathbf{WEW}^T$$

$$\mathbf{T} = \mathbf{WE}^{-1/2} \mathbf{W}^T$$

Singular value decomposition with **W** orthonormal and **E** diagonal with +ve e.values

Sakov & Oke follow the ETKF formalism of Bishop et al. (2001)

Sakov & Oke 2008a – NILU subroutine

```
call ensrkf(ndim, nens, nobs, Xf_ens, xf,
            Yf_ens, yf, y, R, Xa_ens, xa))
```

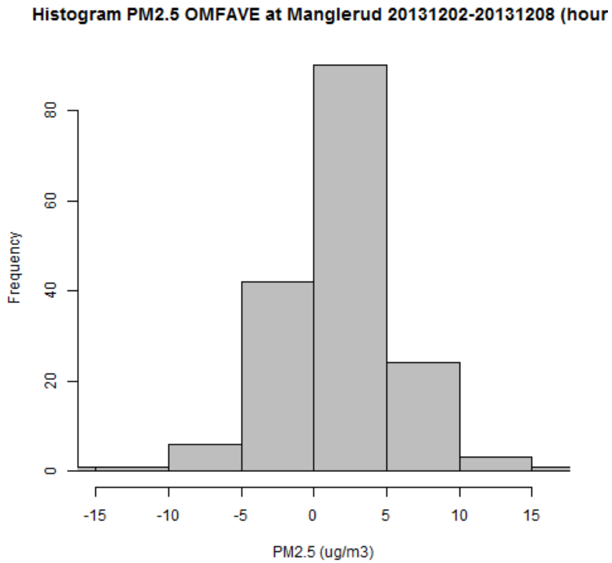
```
ndim = Number of state variables
nens = Number of ensemble members
nobs = Number of observations
Xf_ens = Forecasted ensemble ndim x nens
xf = Mean of forecasted ensemble ndim
Yf_ens = Forecasted (simulated) observations nobs x nens
yf = Mean of forecasted (simulated) observations nobs
y = Real observations nobs
R = Diagonal of R matrix (observation errors) nobs
Xa_ens = Analysed ensemble ndim x nens
xa = Mean of analysed ensemble (the analysed state) ndim
```

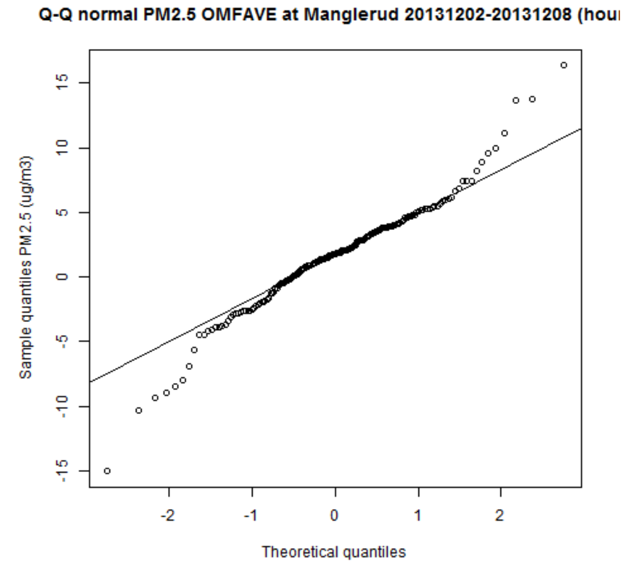
Ensemble set up

- Ensembles are created by perturbing emission data (domestic heating and traffic) and background conc. from MACC (MACC ensemble mean) using 5% relative error standard deviation (SD) – mean of perturbed ensemble is zero;
- Met. data from HARMONIE model (Met Norway) is currently not perturbed (same for all ensemble members);
- Model state is the ground level values in the 3-D initial conc. grid in the EPISODE dispersion model;
- In the EnKF we currently use:
  - 2.5% relative error SD @ 100 μg/m^3 for observations
  - 50%, 50% and 40% relative error SD @ 100 μg/m^3 for NO_2, PM_{10} and $PM_{2.5}$ model error resp. (repr. + subgrid scale (traffic) model error)
  - Diagonal **R**
- DA system tests
  - OmF & OmA
  - Errors tested using chi-square approach for each AQ station
  - Later: vs independent data

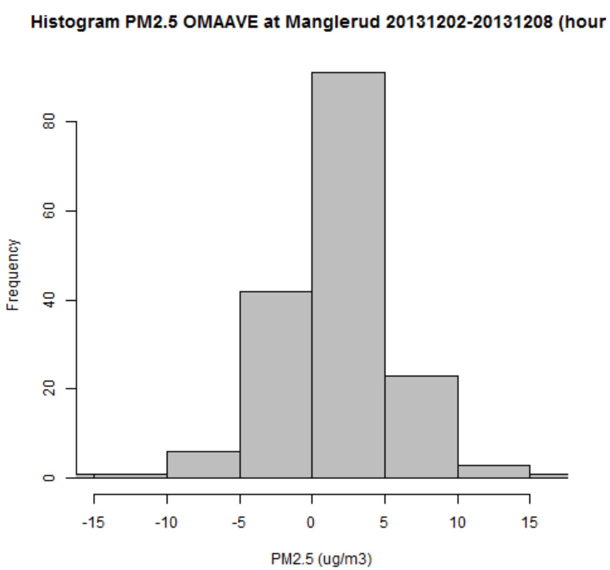
Tests

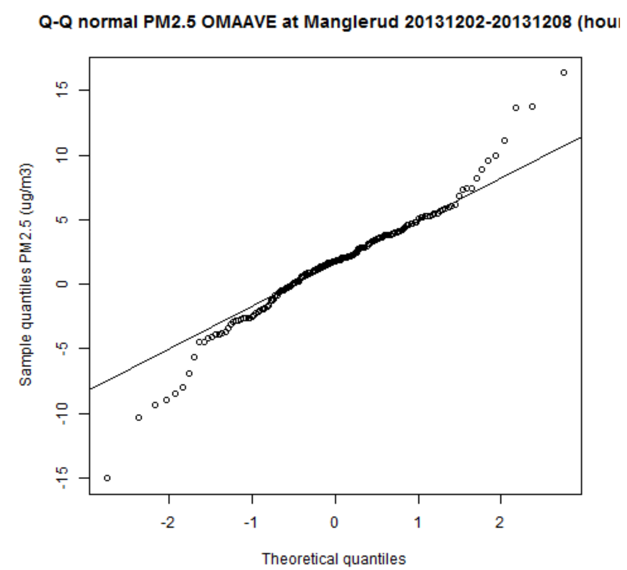
Manglerud AQ station





OmF

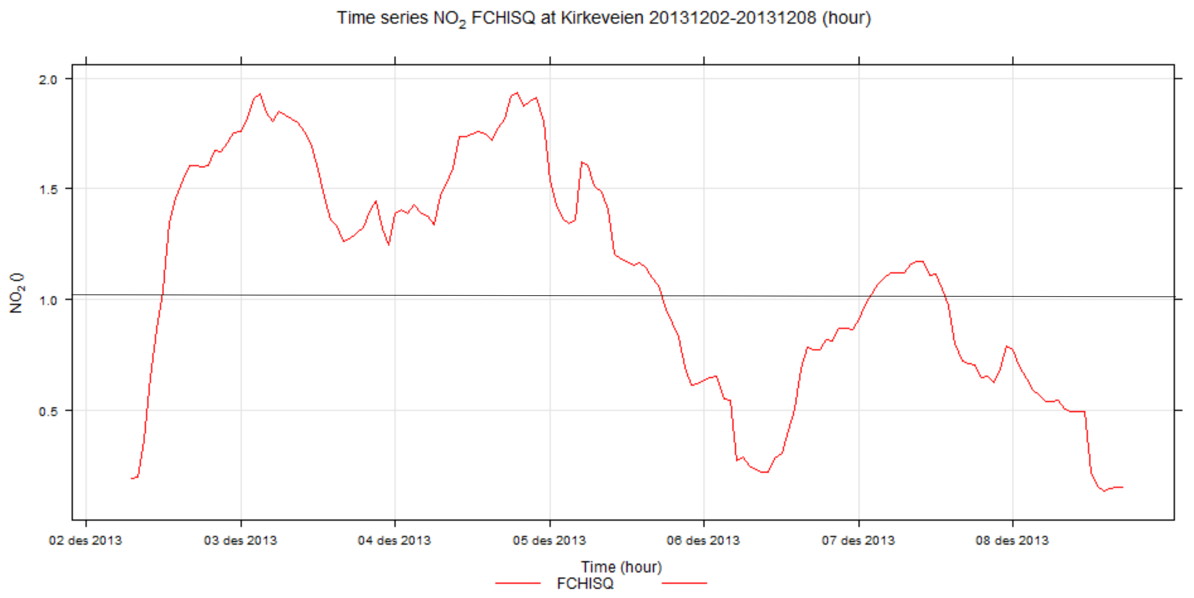




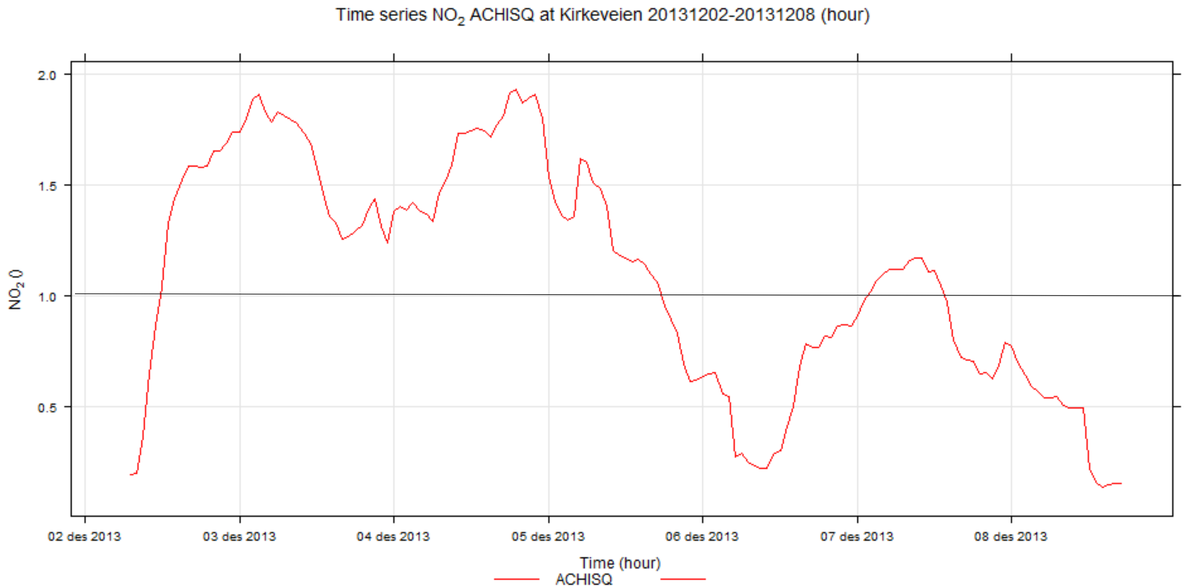
OmA

Chi-square: test of observational errors – Kirkeveien AQ station

OmF



OmA



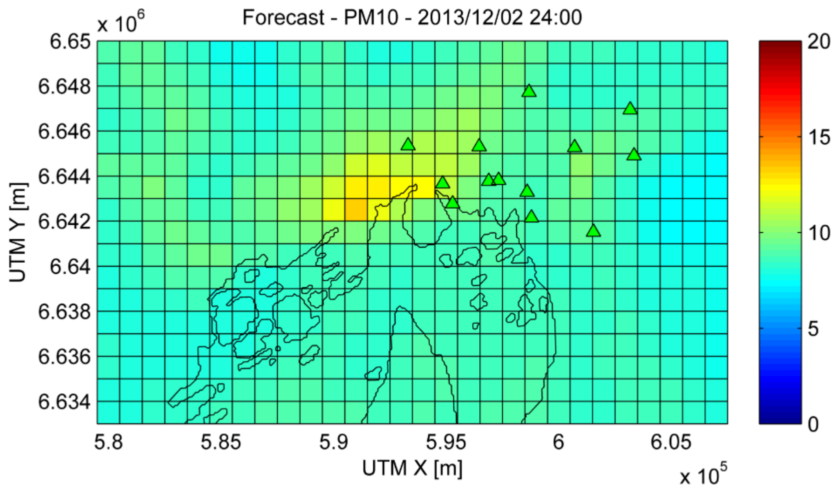
Chi-square test results for AQ stations

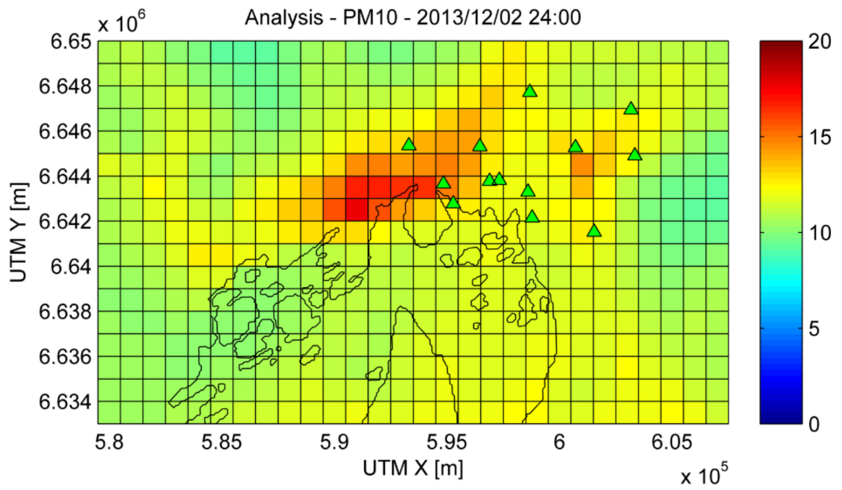
	NO2 % RELATIVE ERROR SD AT 100 ug/m3	PM2.5 % RELATIVE ERROR SD AT 100 ug/m3	PM10 % RELATIVE ERROR SD AT 100 ug/m3
Alnabru	65	47	59
Bygdoy Alle	85	42	63
Hjortnes	74	31	108
Kirkeveien	52	28	63
Manglerud	57	28	59
Rv4 Aker Sykehus	42	22	52
Skoyen	NA	NA	NA
Smestad	82	31	74
Sofienbergparken	NA	36	59
Akebergveien	69	33	50
Gronland	76	NA	NA

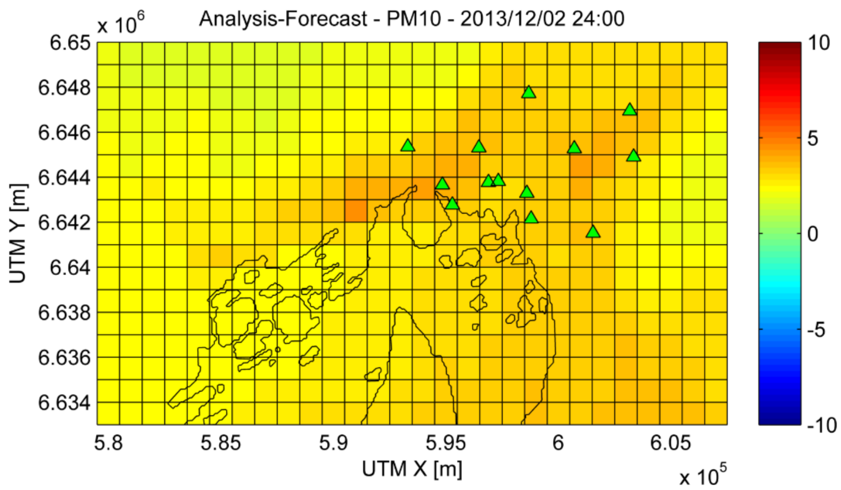
Relative model error SD in % at each station necessary to make the weekly average of the chi-square statistic approximately equal to 1 (for each compound)
The relative observation error SD is 2.5% for all stations

Analyses

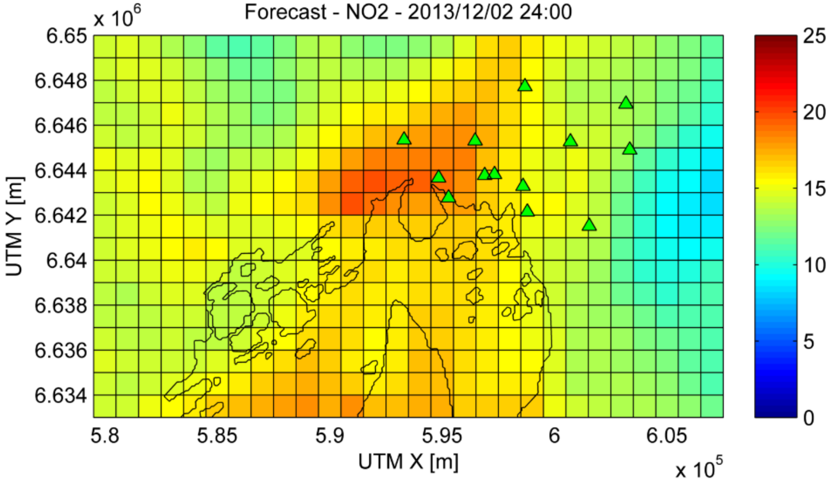
PM10 : Fields at 2400 2-Dec-2012

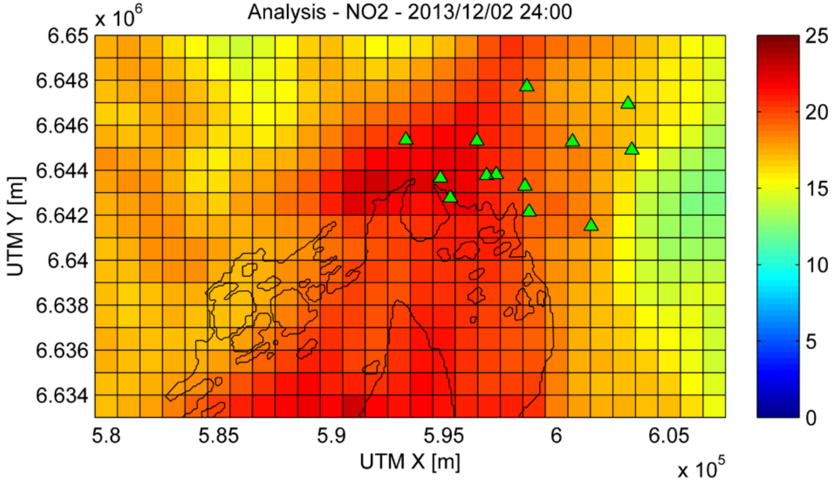


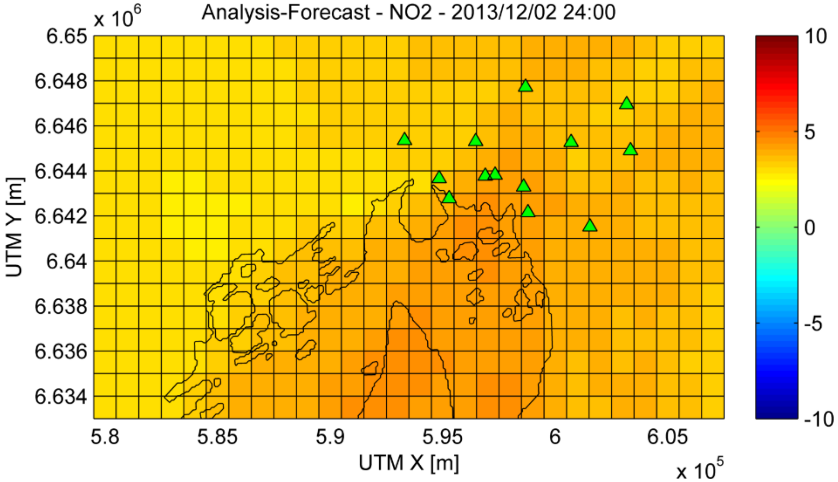




NO_2 : Fields at 2400 2-Dec-2012







Conclusions

- EnKF DA system set up for AQ forecast/analysis for Oslo
- High spatial resolution (1 km – aiming to go lower); high temporal resolution

  Proxy for crowdsourcing development
- Early results – promising, but much work to be done (technical issues)

  Model error; localization; perturbation of ensemble elements; …
- Discussion welcome!

Outlook for data assimilation

Focus is on mainly on three areas (*Lahoz and Schneider, 2014*):

- Improved representation of observational & model errors, including development of hybrid variational/ensemble methods;
- Extension to include & couple various elements of Earth System;
- Reduction in spatial scales being simulated & forecast: getting closer to needs of users—e.g. for weather centers -> representation of convective scales.

Fully coupled, higher-resolution & more accurate reanalyses of Earth System expected to lead to better understanding of climate variability & predictability of weather events.

All apply to "crowdsourcing":

- Citizens' Observatory concept - use of mobile phone platforms:

  EU CITI-SENSE: http://citi-sense.nilu.no; http://greenweek2013.eu/

-A lot of challenges:

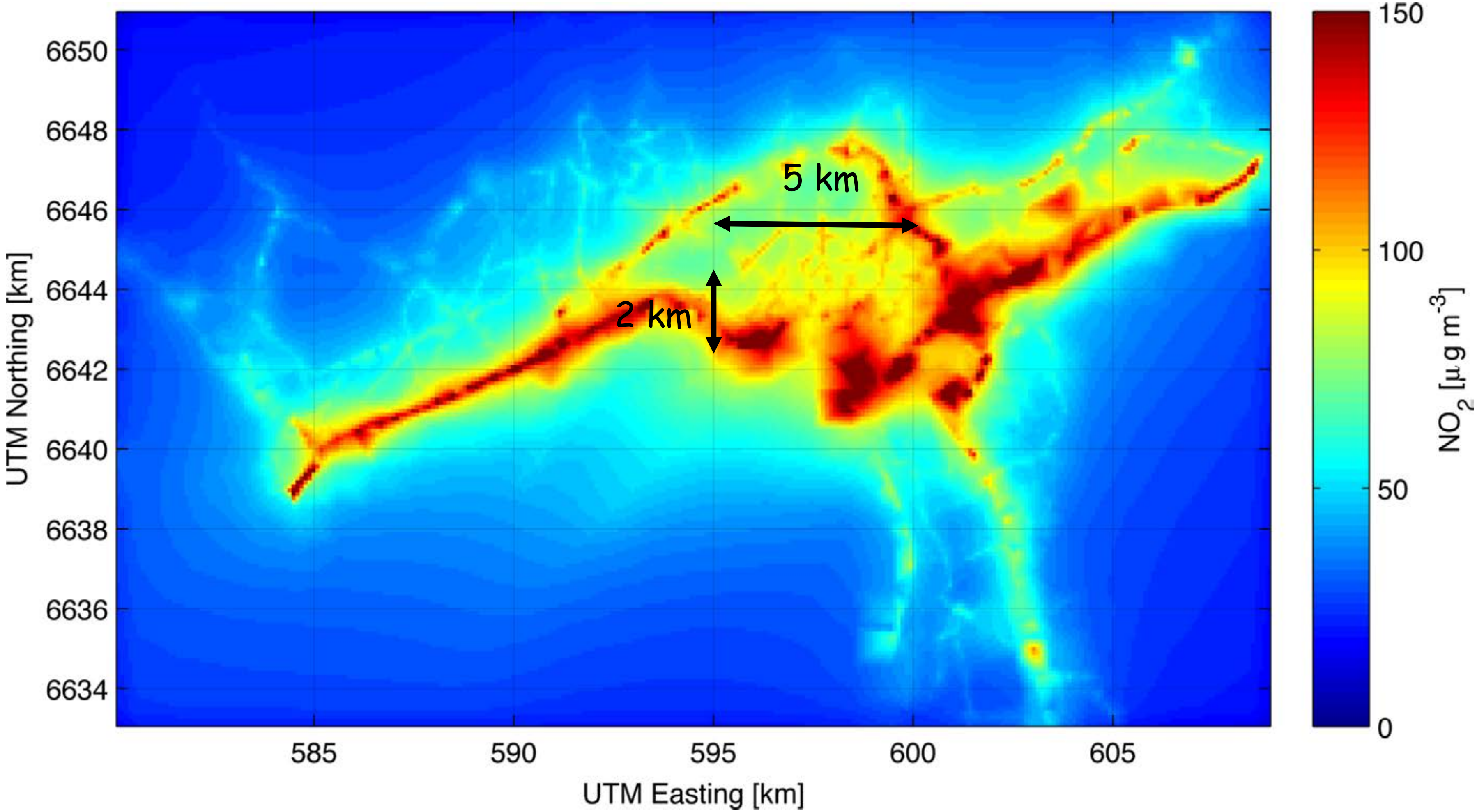
Noisy information, visualization, errors, models, algorithms, different spatio-temporal scales, merging observations at different scales

and privacy…

Extra slides...

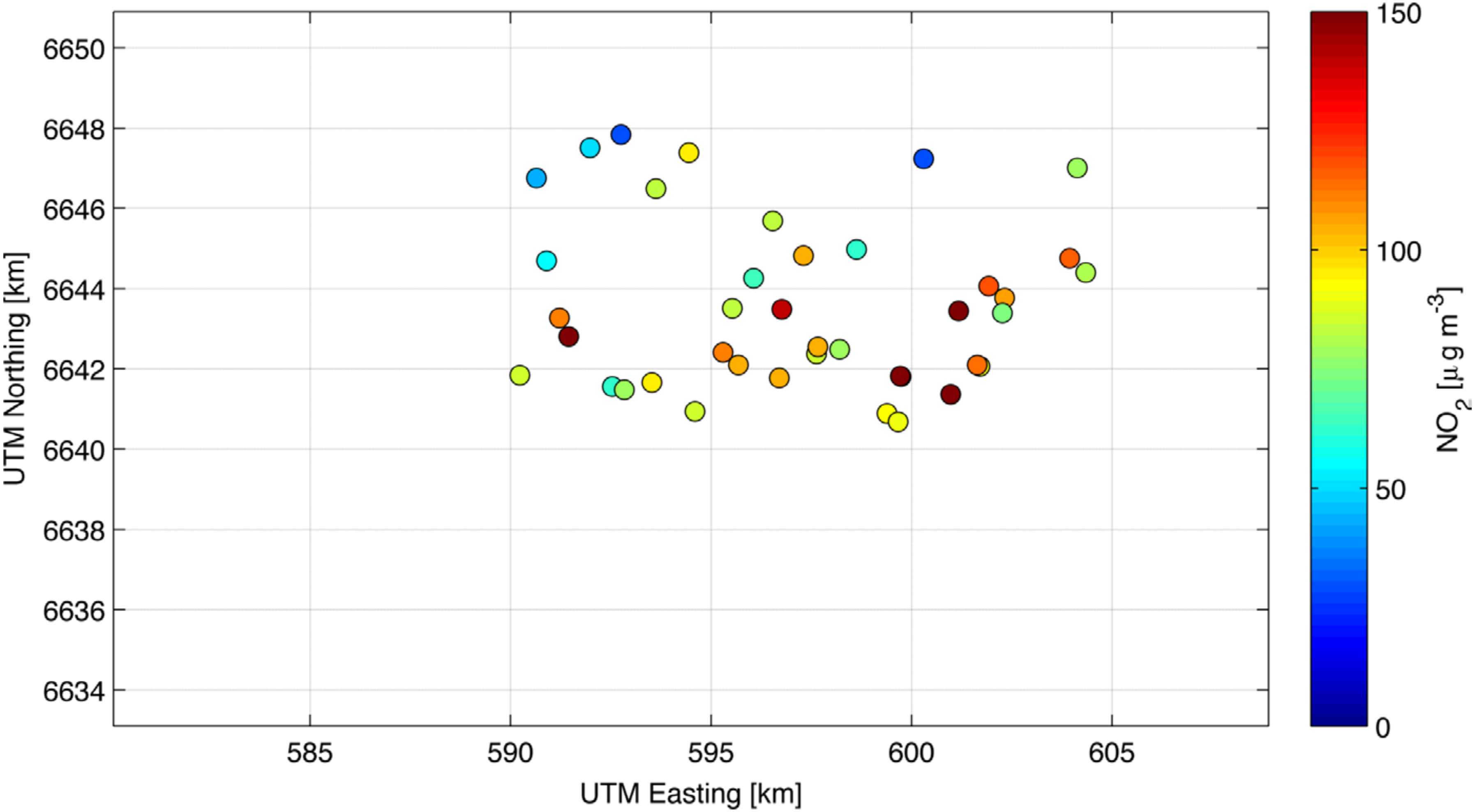
Data fusion

E.g. Oslo: Model information (auxiliary data)



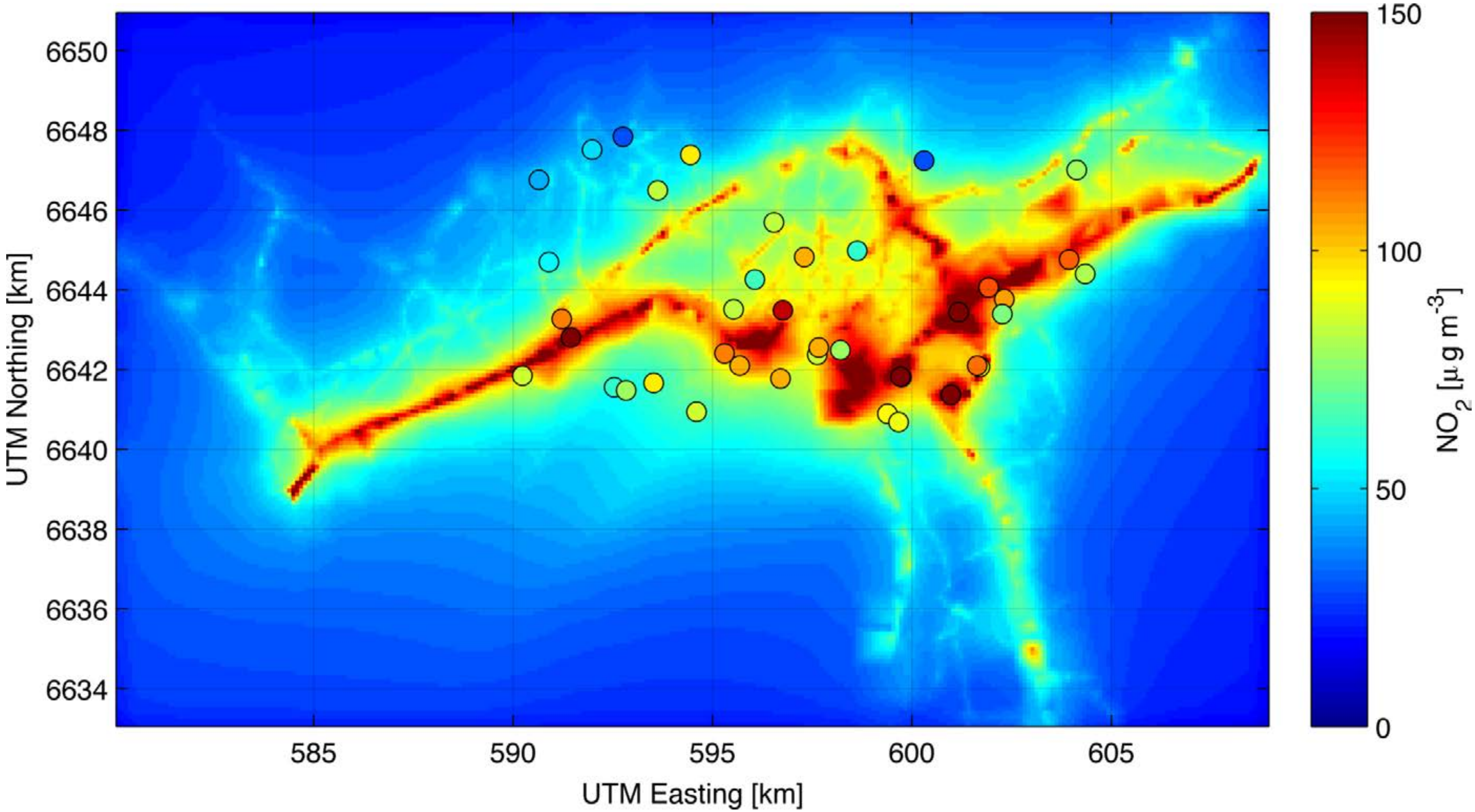
Average NO_x concentrations over Oslo region (2008) provided by EPISODE air pollution dispersion model (*Slørdal et al., 2008*). Methodology for high-resolution model output developed by Bruce Denby at NILU.

E.g. Oslo: Observations



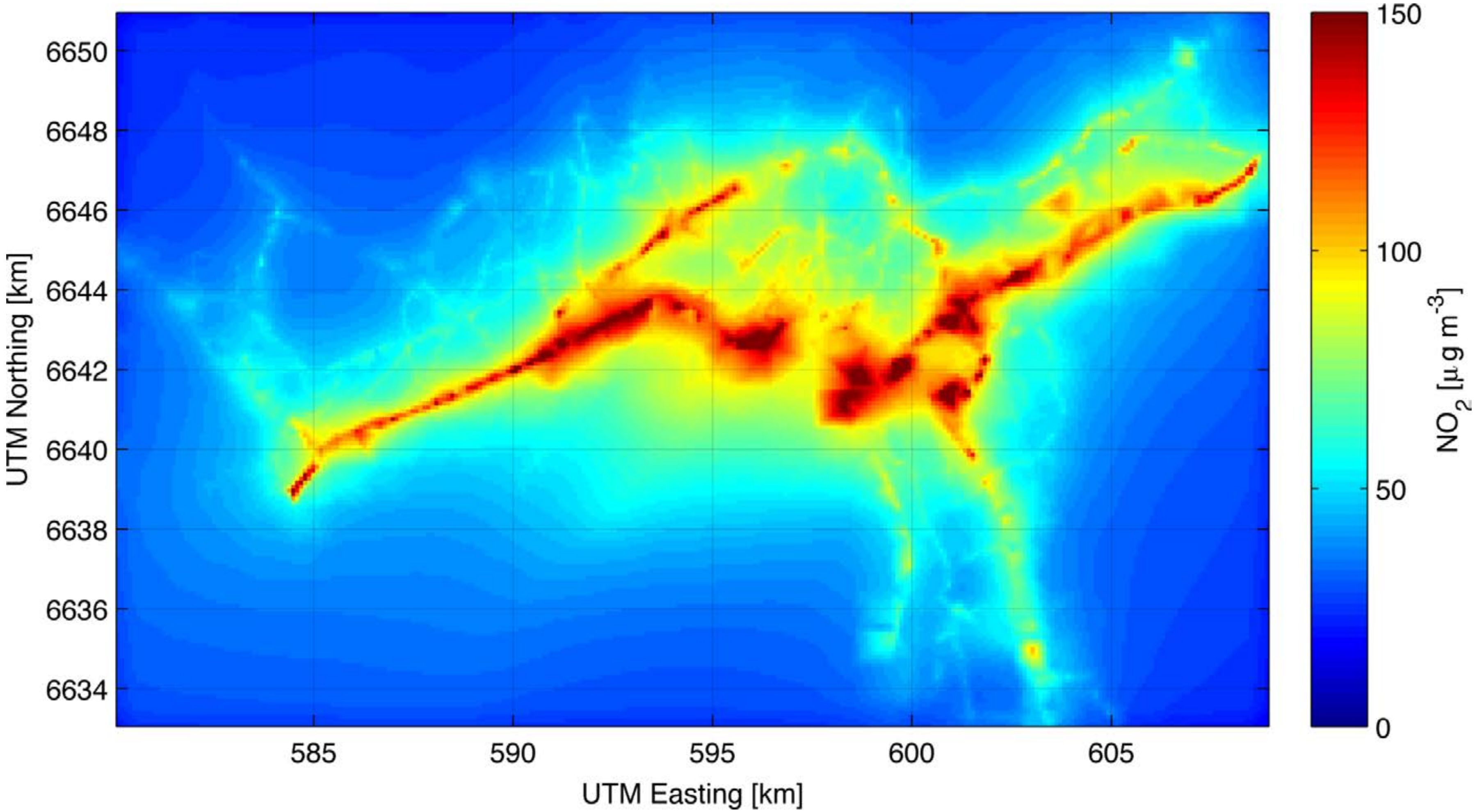
Synthetic observations of NO_2 concentrations generated over Oslo.

E.g. Oslo: Model plus observations



Model data (auxiliary information) & synthetic observations over Oslo.
Note observations agree well with model information in some areas but show significant discrepancies in other areas.

E.g. Oslo: Fused estimate



Fused product of NO_2 concentrations over Oslo, combining information from the EPISODE dispersion model & observations.

PM2.5 : Fields at 2400 2-Dec-2012