Proposal / Contract no.: 248488
Project start: February 1, 2010
Project end: January 31, 2013

UncertWeb

The *Uncert*ainty Enabled Model *Web*

SEVENTH FRAMEWORK PROGRAMME

THEME FP7-ICT-2009-4

ICT for Environmental Services and Climate Change Adaptation

Deliverable 6.2

Probabilistic forecasting of air quality in Oslo

Title of Deliverable	Probabilistic forecasting of air quality in Oslo
Deliverable reference number	UncertWeb D6.2
Related WP and Tasks	WP6, Task 6.2 and Tasks 2.2, 2.3 and 7.2
Type of Document	Public
Authors	Sam-Erik Walker
Date	30/1/2012
Version	1.0

Project coordinator

Dr. Dan Cornford

Aston University, United Kingdom

E-mail: d.cornford@aston.ac.uk

http://www.uncertweb.org

Revision History

Version	Date	Changes	Authors
0.0	5/12/2011	Created document	Sam-Erik Walker
0.1	20/1/2012	First draft to reviewers	Sam-Erik Walker
0.2	26/1/2012	Review	Lorenzo Bigagli
1.0	30/1/2012	Final revision	Sam-Erik Walker

Related task(s):

Task 2.2 User requirements analysis and the international context

Active partners: CNR, AST, UOM

Task 2.3 Chaining and publication methodology and services

Active partners: CNR, AST, UOM

Task 7.2 User requirements for creating the uncertainty enabled model Web

Active partners: NILU, UOM, CNR

Legal Notices

Executive Summary

The present document is the UncertWeb Deliverable 6.2 “Probabilistic forecasting of air quality in Oslo”.

In this report we present the first version of a probabilistic air quality forecasting system for Oslo, which has been developed and implemented as a test case in the UncertWeb project. The aim of the application is to provide probabilistic forecasts (up to 3 days) and nowcasts (current and coming 2-3 hours) of air quality in Oslo. To achieve this, a model chain is set up delivering output from a synoptic scale weather forecasting system (ECMWF) to a mesoscale meteorological model (TAPM) which in turn provides urban scale meteorological fields as input to an urban and local scale air quality model for Oslo (EPISODE). Additional input to the EPISODE model includes emissions and background concentrations using the GEMS/ MACC ensemble of regional scale models.

First preliminary results with the system are presented in the report and are quite encouraging in that ensemble of forecasts of meteorology and air pollution concentrations compares well with available observations at stations in Oslo. Generally, model uncertainty is represented in the system in the form of ensembles of data and stored using the UncertWeb developed netCDF-U format. In developing the system further we will focus on using this format in the communication between various models in the model chain and in the web integration of the system.

Contents

1 Introduction

The present report is the UncertWeb Deliverable 6.2 “Probabilistic Forecasting of Air Quality in Oslo”.

The idea behind this case study is to explore and demonstrate the use of encodings, formats, processing and tools of visualization, developed as part of the UncertWeb project, for a chain of model components in a newly developed system for probabilistic forecasting of air quality in Oslo, Norway. Forecasting air pollution involves modelling highly complex processes in space and time so should be quite suitable as a test case in UncertWeb, where the focus is on representing and communicating uncertainties across a linked chain of model components.

This document is structured in six main sections. Section 1 is the current introduction, which also includes definitions and acronyms (abbreviations) used throughout the report. In Section 2, an overall description of the model chain which constitutes the probabilistic air quality forecasting system is given. Then, in Section 3, each component of this model chain is described in more detail, with special emphasis on each models input and output and representation of their uncertainties. Section 4 contains some preliminary results using the present implementation of the forecast system, while Section 5 contains a discussion of some UncertWeb issues related to this case study, such as communication of models and how the air quality forecast system can be made available as a web service. Some main conclusions are finally given in Section 6.

1.1 Definitions

Model web / Model chain: a chain of model components connected through web service interfaces.

Model component: a representation of a process or series of processes implemented as computer code, often called a simulator.

Model input: a value, or series of values (if a field), that must be provided to the model component to evaluate the model (likely to include parameters in the model component, initial and boundary conditions for the model component)

Model output: a value, or series of values (if a field), that is produced by the model when it is evaluated

Observation component: a set of observations of reality that might be used as model inputs, but also might be used to compare against model outputs. The distinction between an observation component and a model input is that an observation component is used to describe observations (which are collected and stored – i.e., there is some data there) which might be used in the processing chain somewhere, often for output validation. Model inputs are descriptions of what the models want to get as their inputs, which might not always be directly comparable to the observations we have available.

1.2 Acronyms

AQ: Air Quality
AR: Auto Regressive
CHIMERE: Multi-scale model for air quality forecasting and simulation from INERIS and CNRS
CMAR: CSIRO Marine and Atmospheric Research
CNRS: Centre National de la Recherche Scientifique
CSIRO: The Commonwealth Scientific and Industrial Research Organisation

CTM: Chemical Transport Model
ECMWF: European Centre for Medium-Range Weather Forecasts
EMEP: European Monitoring and Evaluation Programme (also a regional CTM from met.no)
EURAD-IM: EURopean Air pollution Dispersion-Inverse Model extension (from RIU)
FMI: Finnish Meteorological Institute
GEMS: Global and regional Earth-system (atmosphere) Monitoring using Satellite and in-situ data
GFDL: Geophysical Fluid Dynamics Laboratory
GRIB: GRIdded Binary
INERIS: L'Institut National de l'EnviRonnement Industriel et des riSques
KNMI: Koninklijk Nederlands Meteorologisch Instituut
LOTOS-EUROS: A regional scale CTM from KNMI and TNO
MACC: Monitoring Atmospheric Composition and Climate
MATCH: Multiscale Atmospheric Transport and Chemistry Model (from SMHI)
met.no: Norwegian Meteorological Institute
MOZART: Model for OZone And Related chemical Tracers (from NCAR, GFDL and MPI-Met)
MPI-Met: Max Planck Institute for Meteorology
NCAR: National Centre for Atmospheric Research
NetCDF-U: Network Common Data Form Uncertainty Conventions
NILU: Norwegian Institute for Air Research
NO: Nitrogen Oxide
NO_2: Nitrogen dioxide
NO_x: Nitrogen Oxides (usually NO + NO2)
O_3: Ozone
PAQFS: Probabilistic Air Quality Forecasting System
PM10: Particles with diameter less than 10 μm
RIU: Rheinisches Institut fuer Umweltforschung
SD: Standard deviation
SILAM: System for Integrated ModeLling of Atmospheric composition (from FMI)
SMHI: Swedish Meteorological and Hydrological Institute
TAPM: The Air Pollution Model (from CSIRO, Australia)
TNO: Nederlandse organisatie voor Toegepast Natuurwetenschappelijk Onderzoek

2 Overview of model chain

The aim of this application is to provide probabilistic forecasts (up to 3 days) and nowcasts (current and coming 2-3 hours) of air quality in Oslo. To achieve this, a model chain has been set up delivering output from a synoptic scale weather forecasting system (ECMWF) to a mesoscale meteorological model (TAPM) which in turn provides urban scale meteorological fields as input to an urban and local scale air quality model (EPISODE). Additional input to the EPISODE model includes emissions and background concentrations. This chain provides the 1-3 day (0-72 h) forecast. In addition a short term (3 hour) forecast will be delivered later in the project that will combine near real time observations with modelled air quality using data assimilation.

To provide probabilistic forecasts, methods are employed where input data to the models TAPM and EPISODE are provided as ensembles. A detailed flow chart of the Oslo PAQFS is shown in Figure 1.

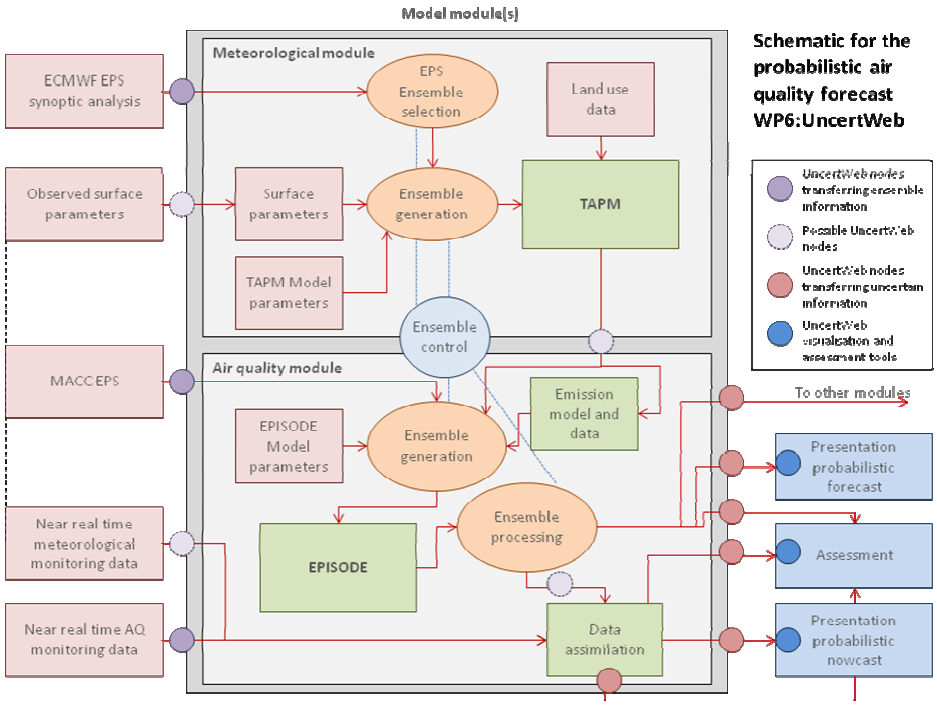

Figure 1: Flow chart of the probabilistic air quality forecasting system for Oslo.

As seen from this figure, the system is mainly divided into a meteorological module (top) and an air quality module (bottom). The main model in the meteorological module is the TAPM model (green box). This model primarily takes as its input ensembles of synoptic scale meteorological forecast data from ECMWF and produces corresponding ensembles of forecasted urban scale meteorological fields to be used by the air quality model EPISODE. The TAPM model also reads in land use data such as topography, soil and vegetation types, etc. for the area from a built-in (global) database. Some surface parameters such as sea surface temperature and deep soil temperature and moisture can be supplied via local observations.

The main model in the air quality module is the EPISODE model (green box). In addition to ensembles of meteorological data produced by TAPM, the main input data to EPISODE are corresponding ensembles of emission data for Oslo and its surroundings produced by the separate emission model (green box), and ensembles of background concentrations provided by the GEMS/MACC model system. The latter is an ensemble of 7 regional scale air quality models for Europe providing the necessary initial and boundary conditions for EPISODE.

Finally, as described above, for short term forecasting the output from the EPISODE model is to be combined with near real-time meteorological and AQ observations using a separate data assimilation model (green box).

In addition, both the TAPM and EPISODE model has a number of more or less uncertain model parameters which also influences the results from running these models.

A more detailed description of the separate parts of the model chain is given in Section 3.

The system is implemented on a Windows PC using a set of Windows batch scripts operating on a given set of directories. This is depicted in Figure 2.

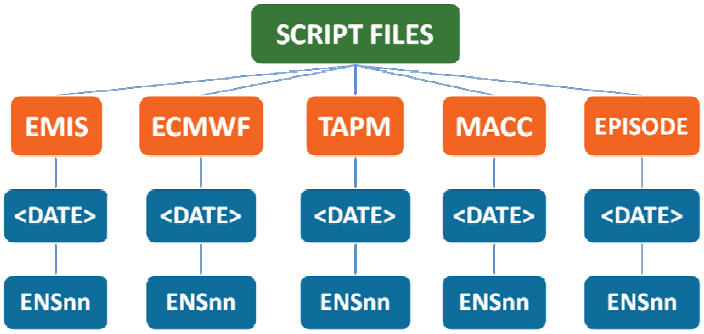

Figure 2: Directory structure of the Oslo PAQFS with script files controlling the system at the top.

As we see from the figure, the system consists of a main directory at the top, containing all scripts controlling the system, with a set of subdirectories underneath, one for each sub part of the system. Each subdirectory is organized in the same way, with a set of directories, one for each base date (<date>), and underneath there, one for each ensemble member nn (ENSnn), where nn is a number ranging from 00 (for control (original) data), to 50 which is the maximum number of ensemble members. All files involved in a particular part of the system for a given date and ensemble member, are stored in the respective directory.

The scripts are organized in a similar manner. At the top, there is a main script (main.bat) controlling the system as a whole, calling a dedicated subscript for each part of the system, for performing various tasks such as generating input data, running a given model, converting the output data to other formats such as e.g., NetCDF and deleting input data or results etc.

Each script operates on a set of input arguments such as:

- Choice of operation to be performed
- Base date to be used (format: yyyymmdd)
- Control (0) or ensemble members to be used (01-50)
- Compound selected (no2, no, nox, pm10)

The top main script can operate on a sequence of dates and number of ensemble members.

The following is an example of how we use the script to create synoptic data for the TAPM model and then run this model, followed by an EPISODE model run for NO_2, for all base dates in the period 2 – 5 January 2011 using 15 ensemble members:

- main tapm_create_synoptic 20110102 20110105 15
- main tapm_run_model 20110102 20110105 15
- main episode_run_model 20110102 20110105 15 no2

Every operation that can be performed in parallel will be performed in parallel by the main script, such as e.g., copying and creating files, running models, converting output data to other formats, deleting results etc. On our current Windows PC with 4 physical CPUs we operate on 8 ensemble members at a time, using the hyper-threading feature of Windows 7 OS, which we have found to be optimal, not only for the model runs but also for more disk intensive operations such as creating and copying files, reading and writing data, deleting files etc.

In addition to the directories described above, the system contains two directories common to all parts of the system: A base directory containing static resource data utilized by the various parts of the system; and a bin directory containing FORTRAN executables and R scripts.

3 Description of model components

In this chapter, various model components of the Oslo PAQFS are described in more detail.

3.1 Emissions

Emission data from air pollution sources in Oslo and its surroundings is input to the urban and local scale air quality model EPISODE. The two main source groups of air pollution in Oslo are domestic heating and traffic. Domestic heating is mainly associated with emissions from the burning of oil and wood during the winter season, while traffic is mainly associated with emissions from car traffic (light and heavy-duty vehicles).

In order to run the EPISODE model for various compounds such as NO_2, PM10 etc., emissions from the above two source groups are needed for these compounds. In addition, since NO_2 is a photo-chemical reactive pollutant, emissions of NO are also needed. Emission data is mainly defined as gridded data using the EPISODE grid for Oslo which is a 30 × 30 km^2 grid with resolution 1 × 1 km^2 covering the city of Oslo and some of its surroundings. This is shown in Figure 3.

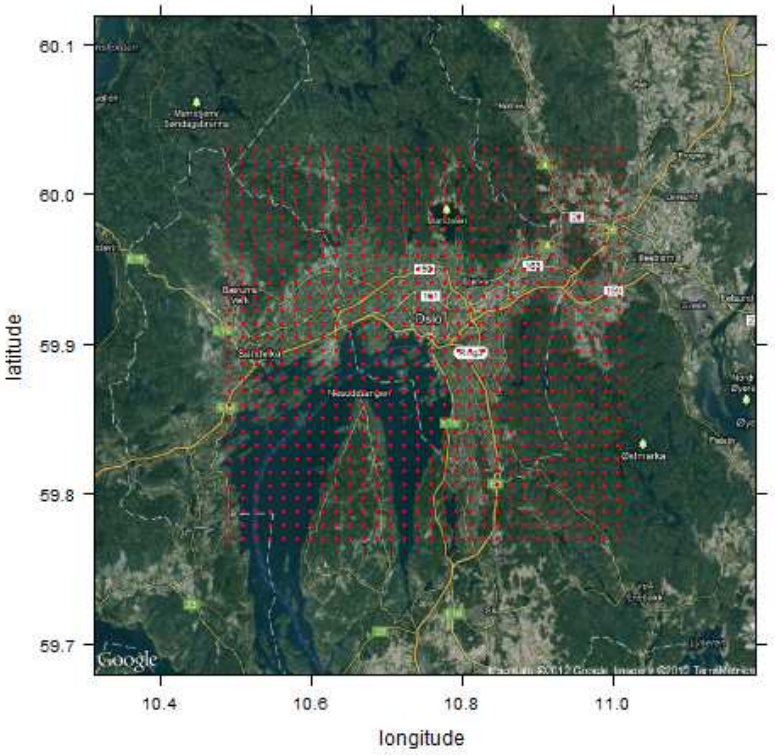

Figure 3: EPISODE 30 × 30 km^2 grid for Oslo with resolution 1 × 1 km^2. The red dots represent the midpoints of each grid cell.

Emissions from individual traffic line sources, such as individual (major) roads or streets in Oslo, will later be added to this emission database. This will be used by the EPISODE model

to more accurately calculate concentrations at the many roadside observation stations in Oslo, which will be important for the proper use of data assimilation in connection with this model.

For the Oslo PAQFS, uncertainties in the emission data need to be addressed. This is done here first for the gridded emission data. Later, uncertainties associated with individual line sources also need to be defined.

For the gridded emission data, the main idea is to represent uncertainties in the form of ensembles (or samples) of conceived or underlying true emission values. There are many ways this can be done, see e.g., [Cressie & Wikle 2011] for a recent and good overview of such spatio-temporal modelling. In our approach here, true gridded emission values are modelled by stochastically perturbing the corresponding deterministically calculated values.

To this end, let $Q_{ij,t}$ denote the calculated emission of one of the compounds for grid cell i, j and time (hour) t, and let $Q_{ij,t,true}$ denote the corresponding true emission. This latter quantity is then tentatively modelled as follows:

$$Q^{(\lambda)}_{ij,t,true} = Q^{(\lambda)}_{ij,t} + \varepsilon_{ij,t}; \quad \varepsilon_{ij,t} = \phi\varepsilon_{ij,t-1} + \sigma^2\eta_{ij,t}; \quad \eta_t \sim N(0,V) \tag{1}$$

where λ is the Box-Cox power transformation parameter; ϕ is an AR(1) parameter handling possible time dependency; σ is the standard deviation of the AR(1) noise process of random Gaussian variables $\eta_{ij,t}$ (and subsequently $\varepsilon_{ij,t}$); and V is the variance-covariance matrix of the vector η_t of $\eta_{ij,t}$ values for $i = 1,...,n_x$, $j = 1,...,n_y$ (where $n_x = n_y = 30$). The assumption here is that, after a suitable Box-Cox power transformation (see below), the transformed true and model calculated emission values follows the model in (1) with Gaussian errors ε and η. Here ε is modelled as dependent both in time and space, while η is modelled as dependent in space only, with V as its spatial covariance (correlation) matrix. Furthermore, we assume for simplicity here that η is stationary and isotropic so that the covariance or correlation between η-variables only depends on the horizontal distance between the grid cells.

The Box-Cox power transformation is here defined by:

$$Q^{(\lambda)} = \begin{cases} \dfrac{Q^{\lambda}-1}{\lambda} + \lambda & \text{for} \quad 0<\lambda\le 1 \\ \ln(Q) & \text{for} \quad \lambda=0 \end{cases} \tag{2}$$

where $\lambda = 1$ corresponds to untransformed values, and $\lambda = 0$ corresponds to log-transformed values (representing the limit as $\lambda \to 0$).

A good choice for V is to use the Matérn class of covariance functions to model the spatial dependencies, i.e.,

$$V(s_1,s_2) = \frac{1}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}2^{\nu-1}} \left(\kappa\|s_1 - s_2\|\right)^{\nu} K_{\nu}\left(\kappa\|s_1 - s_2\|\right) \tag{3}$$

where s_1 and s_2 are two arbitrary points in d-dimensional space (here $d = 2$); $\|\cdot\|$ denotes the Euclidian distance norm; Γ is the Gamma function; and where K_{ν} is the modified Bessel

function of the second kind and with order ν. This latter parameter is a smoothness parameter determining the differentiability (or smoothness) of the Gaussian field of η-values. The parameter can be set to any positive real number, but one of the following values is often used:

- $\nu = 1/2$ i.e., an exponential covariance function
- $\nu = 1$ slightly smoother than exponential
- $\nu = 2$ even smoother
- $\nu = \infty$ a Gaussian covariance function, which is infinitely, continuously differentiable

The parameter κ in (3) is a scaling parameter related to the range ρ through the relation:

$$\rho = \frac{\sqrt{8\nu}}{\kappa}. \tag{4}$$

The interpretation is that at distances larger than ρ, correlation becomes negligible, i.e., less than about 0.1.

Ideally, the parameters λ, ϕ, σ, ν and κ (or ρ) should be estimated from data. This is, however, difficult since we have no (direct) observations of emission data.

Currently, the following values are used (tentatively) in the Oslo PAQFS:

- $\lambda = 1/3$ corresponding to using a cube-root transformation of emissions
- $\phi = 0.7$ which means that temporal correlations become negligible after 6-7 hours
- $\nu = 1$ corresponding to spatial covariance slightly smoother than exponential
- $\kappa \approx 1.0$ which means that spatial correlations becomes negligible at $\rho \approx 3$ km

Parameters should probably be set differently for the two source group domestic heating and traffic, and for the various compounds.

The marginal standard deviation σ is defined using the following formula:

$$\sigma = \frac{p}{100}\sqrt{1-\phi^2}\,Q_p^{\lambda} \tag{5}$$

where Q_p is the emission level at which the deterministically calculated value has a p% relative error standard deviation as compared to the true emission value. At higher levels the percentages will be smaller; at lower levels they will be higher. The values Q_p and p is determined either from uncertainties associated with the emission data, or set subjectively. Currently we set $p = 10\%$ and use $Q_{10\%} = Q_{\max}$ where $Q_{\max}$ is the highest achievable emission value from the given source group and compound.

The above stochastic model is currently used to draw ensembles of 30 x 30 fields of gridded emissions from domestic heating and traffic and for each of the compounds NO_2, NO, NO_x, and PM10.

Grid cells with zero calculated emissions are interpreted as areas with no inhabitants, so true emissions are defined to be exactly zero there for each ensemble member.

Some emission data depends on meteorological parameters, such as e.g., emission of PM10 from re-suspension of road dust which depends on precipitation, and emission from domestic heating which depends on air temperature. Such specific links in uncertainties between meteorology and emissions will be added to the uncertainty model later.

An R script [R 2.14.1 2011] has been made to store the original (control) and ensemble emission data from the source groups domestic heating and traffic and for various compounds such as NO_2, NO, NO_x and PM10 into a single netCDF-U formatted file [Bigagli et al, 2011], [UW-D2.2 2011]. Such files are written separately for each base date containing the 1-3 days (0-72h) forecasted emission data.

Each ensemble emission variable is stored in the NetCDF file using the following extra local attribute:

`ref = http://www.uncertml.org/samples/random.`

A coordinate variable for the ensemble members is also created in the NetCDF file with the following extra attribute:

`ref = http://www.uncertml.org/samples/realisation.`

The use of these attributes are in accordance with the principles for storing ensemble data in netCDF-U formatted files as laid out in [Bigagli et al, 2011] and [UW-D2.2 2011]. Later this file may be used to communicate uncertainty in the emission data to the EPISODE model (for a discussion of this, see Section 5).

3.2 ECMWF ensemble

Synoptic scale weather forecasts from ECMWF provides the necessary initial and boundary conditions for the mesoscale meteorological model TAPM, which in turn provides the urban scale meteorological fields for the urban and local scale air quality model EPISODE. The ECMWF ensemble forecast (http://www.ecmwf.int) consists of a global analysis forecast and an ensemble forecast of 50 ensemble members produced by various perturbations of the initial conditions provided by the global analysis. In our system the former is used as the control forecast, while the latter is used as the ensemble meteorological forecast.

The downloading of the ECMWF data is done in a semi-automatic way using a Korn shell script. The data are in GRIB format and consists of 2D and 3D fields of meteorological data for a forecasting period of 3 days (0-72h) from each base date. The GRIB files are combined and converted to NetCDF formatted files using a utility program (g2n.exe) made by Peter Hurley, CMAR-CSIRO. Temporally the resolution of the data is every 6 hour with each value representing an instantaneously hourly mean value at that time point. Spatially the data are gridded with a horizontal resolution of 0.5° × 0.5° and with 10 vertical layers at pressure levels ranging from 100 – 1000 hPa. The domain for the data is defined as the area between 65° N, 0° E and 55° N, 20° E with 40 × 20 × 9 grid points. Currently, data has been downloaded for the period 1 – 31 January 2011.

The ECMWF data will have a varying uncertainty depending on the synoptic situation. For a 3-day forecast meteorological data can range from being quite certain to very uncertain. The

variability is (supposedly) represented by the spread in the ensemble values for each meteorological parameter (wind, temperature etc.). Currently, the uncertainty represented by the ensemble is communicated to the TAPM model using the set of separately generated NetCDF files.

3.3 TAPM

The TAPM model [Hurley, 2008] takes as its input the synoptic scale weather forecast data from ECMWF as described in the previous section, and produces forecasted urban scale meteorological fields for the air quality model EPISODE. The model operates in a nested fashion, taking the large scale synoptic meteorological data from ECMWF as initial and boundary conditions for the outermost (largest) grid and producing gridded fields of meteorology for the innermost (smallest) grid covering the city of Oslo. To this end, four nesting levels are used, from the largest grid covering an area of 800 × 800 km^2, through intermediate grids covering areas of 320 × 320 km^2 and 120 × 120 km^2, to the smallest grid covering an area of 40 × 40 km^2. This is depicted in Figure 4 with the outer boundaries of the four TAPM grids shown in yellow (the smallest red square is the EPISODE 30 × 30 km^2 grid for Oslo (see Section 3.5)).

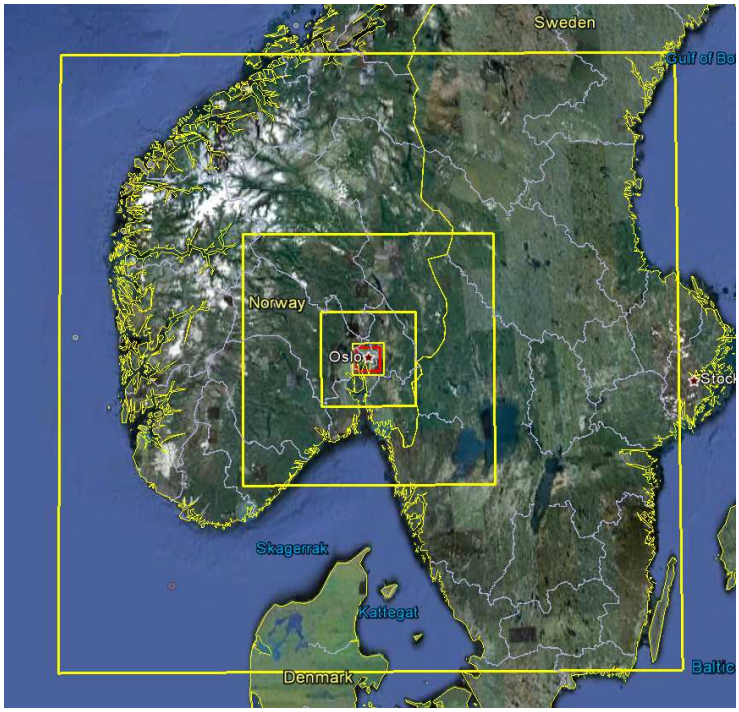

Figure 4: Boundaries of the four nested grids used by the TAPM model (in yellow).

All the grids have the same number of grid points which is 40 × 40 horizontally and 25 levels vertically. They are also centred on a common point, which is a selected central point in Oslo with latitude, longitude 59.90° N, 10.75° E. The horizontal resolution of the four grids is thus 20 km, 8 km, 3 km and 1 km. The vertical levels range from 10 m to 8000 m for all four grids.

In order to run the TAPM model, the ECMWF control and ensemble NetCDF files must be converted to a special set of synoptic files used by the TAPM model. This is done by a utility program (n2s.exe) made by Peter Hurley, CMAR-CSIRO. The TAPM model is then run separately for control data and for each set of ensemble data. For each of these, the TAPM

model produces a number of output files, the most important of which is a binary file (test01a.out) containing all meteorological fields for the innermost grid. In addition, files containing precipitation in the form of rainfall (test01.rfl) and snowfall (test01a.sfl) are produced. Another special conversion program (tapm2episode.exe) reads these files and produces meteorological field files in binary format that the EPISODE model can read (*.fld). These files consist of 30 × 30 km^2 gridded data and 20 levels vertically (part of the TAPM innermost grid).

Table 1 lists the meteorological parameters and other data which is output from the TAPM model and which the EPISODE model reads.

Symbol	Description	Unit	Dim.
u, v, w	Wind	ms^{-1}	3
T	Temperature	K	3
θ	Potential temperature	K	3
RH	Relative humidity	%	3
TKE	Turbulent kinetic energy	m^2s^{-2}	3
u_*	Friction velocity	ms^{-1}	2
w_*	Convective velocity scale	ms^{-1}	2
$HMIX$	Mixing height	m	2
T_{screen}	Temperature at ground level (2 m)	K	2
$T_{surface}$	Temperature at surface	K	2
RH_{screen}	Relative humidity at ground level (2 m)	%	2
EHF	Evaporative heat flux	Wm^{-2}	2
SHF	Sensible heat flux at surface	Wm^{-2}	2
TSR	Total solar radiation	Wm^{-2}	2
$NETRAD$	Net radiation	Wm^{-2}	2
θ_*	Potential temperature scale at surface	K	2
PVT	Potential virtual temperature scale at surface	K	2
$PREC$	Precipitation	mmh^{-1}	2
z_{EPI}	3D EPISODE z above ground	m a.g.l.	2
z_{topo}	Topography	m a.s.l.	2

Table 1: Meteorological parameters and other data, which is output from the TAPM model and read by the EPISODE model.

The values in column named "Dim." in the table indicate whether the variable is 3D or 2D. Communicating the height of vertical layers (z_{EPI}) and topography (z) used by the TAPM model is important for the accurate use of the meteorological data in the EPISODE model.

An R script [R 2.14.1 2011] has been made to store (some of) the control and ensemble data given in Table 1 into a netCDF-U formatted file [Bigagli et al, 2011], [UW-D2.2 2011]. The following variables are currently written to this file:

- Horizontal wind (u, v) at the lowest level (10 m)
- Temperature at ground level (2 m) (T_{screen})
- Vertical temperature gradient (dT/dz)
- Relative humidity at ground level (2 m) (RH_{screen})

- Precipitation (*PREC*)
- Topography (z_{topo})

Each ensemble meteorological variable is stored in the NetCDF file using the following extra local attribute:

```
ref = http://www.uncertml.org/samples/random.
```

A coordinate variable for the ensemble members is also created in the NetCDF file with the following extra attribute:

```
ref = http://www.uncertml.org/samples/realisation.
```

The use of these attributes are in accordance with the principles for storing ensemble data in netCDF-U formatted files as laid out in [Bigagli et al, 2011] and [UW-D2.2 2011]. Later this file may alternatively be used to communicate uncertainty in the meteorological data to the EPISODE model (for a discussion of this, see Section 5).

In addition to the ECMWF meteorological data, the TAPM model also uses a global land use database in order to describe the surface characteristics. This database can be manually altered for the situation or site of interest but this is not done in our present use of this model. The land use data is used to supply surface parameters for the model typical for the land use type used. The data are static so that any errors in them will lead to errors or bias in the model, but they are currently not treated as uncertain input data in our system.

There are, however, a number of surface parameters linked to the land use data that need to be supplied as initial conditions for the model. These include sea surface temperature, deep soil temperature, deep soil moisture, and snow and ice cover (depth). In the use of the TAPM model these are currently set to their default (climatological) values. Some of them are, however, more uncertain than others, so we may wish later to vary some or all of these. In particular, we may include perturbations of sea surface temperature, deep soil temperature and deep soil moisture. Accounting for uncertainty in these parameters, in addition to the meteorological data, will lead to increased variability (uncertainty) in the TAPM model ensemble output. Another possibility is to include observations of these, which instead may lead to decreased uncertainty.

For more information about the TAPM model see: http://www.cmar.csiro.au/research/tapm.

3.4 GEMS/MACC model ensemble

Background concentrations (boundary conditions) of NO_2, O_3 and PM10 for the urban scale air quality model EPISODE in the Oslo PAQFS are provided by the GEMS/MACC model ensemble (http://gems.ecmwf.int/d/products/raq). This is an ensemble of air pollution models predicting air quality at the regional scale in Europe.

Table 2 lists the names, institutions involved and resolution characteristics of the 7 models which were part of this model ensemble as of January 2011, which is the basis for our current calculations.

Each model in the GEMS/MACC model ensemble provides a European wide 3 day (0-72h) hourly average concentration forecast. Key input data to these models such as emission data,

meteorological forcing, chemical boundaries and observations assimilated are common to all models. The models differ, however, when it comes to horizontal and vertical resolution (as shown in Table 2), and also regarding their formulations and parameterizations.

Model name	Institution(s) involved	Horizontal resolution	Vertical levels and top
CHIMERE	INERIS, CNRS	25 km	8 levels up to 500 hPa
EMEP	met.no	0.2°	20 levels up to 100 hPa
EURAD-IM	RIU	15 km	23 levels up to 10 hPa
LOTOS-EUROS	KNMI, TNO	15 km	4 levels up to 3.5 km
MATCH	SMHI	0.2°	40 levels up to 100 hPa
MOZART	NCAR, GDFL, MPI-Met	120 km	60 levels (top N/A)
SILAM	FMI	0.2°	46 levels up to 100 hPa

Table 2: GEMS/MACC model ensemble as of January 2011.

Gridded values for each of the models in the ensemble are interpolated to the selected central Oslo latitude, longitude coordinate which is 59.90° N, 10.75° E. This gives a single background concentration value for Oslo from each of the models in the ensemble for each forecast hour (0-72h). Currently only ground level concentrations from the models are used as background concentrations (boundary conditions) for the EPISODE model.

Not all models in the GEMS/MACC ensemble calculate all compounds, and some models may not be available at a given date and time, so occasionally there can be quite few members in the ensemble for a given compound. Instead of relying upon a variable member ensemble to provide the uncertainty in the background concentrations, we decided instead to construct a fixed sized ensemble by stochastically perturbing the background concentrations from each available model. Hence, if n_e is the number of ensemble values to be produced (≤ 50), and n_m is the number of models available (≤ 7), this system will produce approximately n_e/n_m perturbed ensemble members for each available model.

Each ensemble member represents a time series of possibly true background concentrations $B_{t,true}$ and is calculated as follows:

$$B_{t,true}^{(\lambda)} = B_{t,k}^{(\lambda)} + \varepsilon_{t,k}; \quad \varepsilon_{t,k} = \phi\varepsilon_{t-1,k} + \eta_{t,k}; \quad \eta_{t,k} \sim N\left(0, \sigma_{\eta,k}^2\right) \tag{6}$$

where $B_{t,k}$ is the background concentration value from GEMS/MACC model number k (1-7) for hour t (0-72); λ is the Box-Cox power transformation parameter (see (2)); ϕ is an AR(1) parameter handling possible time dependency; and $\sigma_{\eta,k}$ is the standard deviation of the AR(1) noise process of random Gaussian variables $\eta_{t,k}$ (and subsequently $\varepsilon_{t,k}$). So, the assumption here is similar to that in (1), i.e., that after a suitable Box-Cox power transformation the transformed background concentrations follows (6) with Gaussian errors $\varepsilon_{t,k}$ and $\eta_{t,k}$.

The parameters λ, ϕ and $\sigma_{\eta,k}$ in (6) should ideally be estimated from data, i.e., by comparing GEMS/MACC model values with observations at regional background stations close to Oslo. This will be attempted later in the project.

Currently, the parameters are set to tentative values for each compound as given in Table 3

(equal for all models).

Compound	λ	ϕ	$\sigma_{\eta,k}$	% relative error SD at level
NO_2	1/3	0.7	0.58	33% at 15 μgm^{-3}
O_3	1/3	0.7	0.61	25% at 40 μgm^{-3}
PM10	1/3	0.7	0.25	17% at 9 μgm^{-3}

Table 3: Tentative background concentration ensemble model parameters.

Here $\lambda = 1/3$ again corresponds to using a cube-root transformation of each background concentration value. This is often found to give approximately a Gaussian distribution of air pollution concentrations (which originally are typically skewed to the right). The percentage relative error standard deviations as given in the table are based on some preliminary studies of the spread of model values for the different compounds in the GEMS/MACC ensemble. Based on a percentage relative error of p % at a given level $B_{p\%}$, the $\sigma_{\eta,k}$ -values in the above table are calculated as follows:

$$\sigma_{\eta,k} = \frac{p}{100}\sqrt{1-\phi^2}\,B_{p\%}^{\lambda}. \tag{8}$$

At higher levels the percentage relative error standard deviations will be smaller, while at lower levels they will be higher. This is depicted in Figure 5 showing the percentage relative error standard deviations as a function of level for each of the three compounds NO_2 (left), O_3 (middle) and PM10 (right). It must be emphasized, however, that these numbers and curves are preliminary and may be revised going forward based on more thorough analysis.

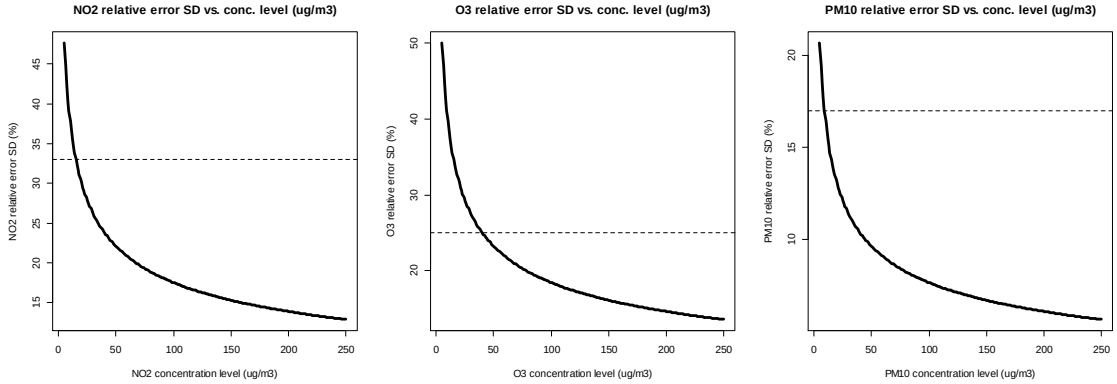

Figure 5: Relative error standard deviation in modelled background concentrations as a function of background concentration level for NO_2 (left), O_3 (middle) and PM10 (right).

For each base date the resulting background concentration ensemble is currently written to a set of ASCII files which are used by the EPISODE model. The GEMS/MACC ensemble median values are stored as control values (ensemble member 0) in the PAQFS.

An R script [R 2.14.1 2011] has been made to store the original (control) and ensemble background concentration data for the compounds NO_2, O_3 and PM10 into a single netCDF-U formatted file [Bigagli et al, 2011], [UW-D2.2 2011]. Such files are written separately for each base date containing the 1-3 days (0-72h) forecasted data.

For each compound, a background concentration variable is stored in the NetCDF file using the following extra local attribute:

`ref = http://www.uncertml.org/samples/random.`

A coordinate variable for the ensemble members is also created in the NetCDF file with the following extra attribute:

`ref = http://www.uncertml.org/samples/realisation.`

The use of these attributes are in accordance with the principles for storing ensemble data in netCDF-U formatted files as laid out in [Bigagli et al., 2011] and [UW-D2.2 2011]. Later this file may be used to communicate uncertainty in the background concentrations data to the EPISODE model (for a discussion of this, see Section 5).

3.5 EPISODE

Air quality forecasts for Oslo are performed using the EPISODE model [EEA MDS, 2012], which is a model developed by NILU for calculating dispersion of air pollution at the urban and local scale. The model takes as its input, emissions, meteorology and background concentrations as described in Sections 3.1 - 3.4 and calculates hourly average (ground level) concentrations of NO_2, NO, O_3, NO_x and PM10 in the 30 × 30 km^2 grid shown in Figure 3. Vertically the model has 20 layers, using approximately the same levels as the TAPM model from 10 m close to the ground up to a model top at 3750 m.

The EPISODE model contains a traffic model for calculating hourly average (ground level) concentrations more accurately close to major roads and streets in a city. This is important in order to be able to compare model output concentrations with air quality observations at the many roadside stations in Oslo.

Currently, input data is generally communicated to the model using a set of specially formatted binary (*.fld) and ASCII (*.asc) files. Model output concentrations are also currently produced using such formats.

Uncertainties in the input to EPISODE is represented in the form of ensembles of emissions, meteorology and background concentrations, so uncertainties in the output from the model is naturally represented by a corresponding set of ensemble output concentrations.

An R script [R 2.14.1 2011] has been made to store control and ensemble generated gridded and receptor concentrations of NO_2, NO, O_3, NO_x and PM10 into one or more netCDF-U formatted files [Bigagli et al, 2011], [UW-D2.2 2011], which are then written separately for each base date containing the 1-3 days (0-72h) forecasted concentration values.

For each compound, a concentration variable is stored in the NetCDF file using the following extra local attribute:

`ref = http://www.uncertml.org/samples/random.`

A coordinate variable for the ensemble members is also created in the NetCDF file with the following extra attribute:

`ref = http://www.uncertml.org/samples/realisation.`

The use of these attributes are in accordance with the principles for storing ensemble data in netCDF-U formatted files as laid out in [Bigagli et al., 2011] and [UW-D2.2 2011]. Later this file may be used to communicate uncertainty in the output concentrations from the EPISODE model (for a discussion of this, see Section 5).

4 Preliminary results

In this chapter, some preliminary results with the Oslo PAQFS are shown.

4.1 ECMWF/TAPM

Figure 6-Figure Figure 9 show 24h meteorological forecasts at 3 January 2011 at the central met station Valle Hovin in Oslo using the combined ECMWF and TAPM model systems. In each of these figures, the blue curve constitutes the observed values, while the red and orange curves represent the ensemble mean and median values respectively. The upper and lower green curves are the ensemble 5 percentile and 95 percentile curves so the area between these curves represents a 90% central prediction interval. The black curve is the original (control) values, while the grey curves represent the individual members of the ensemble (50).

Figure 6 displays the air temperature at 2 m above ground, while Figure 7-Figure 9 show vertical air temperature difference between 25 m and 8 m (an indicator of atmospheric stability), and wind speed and wind direction at 25 m above ground level, respectively.

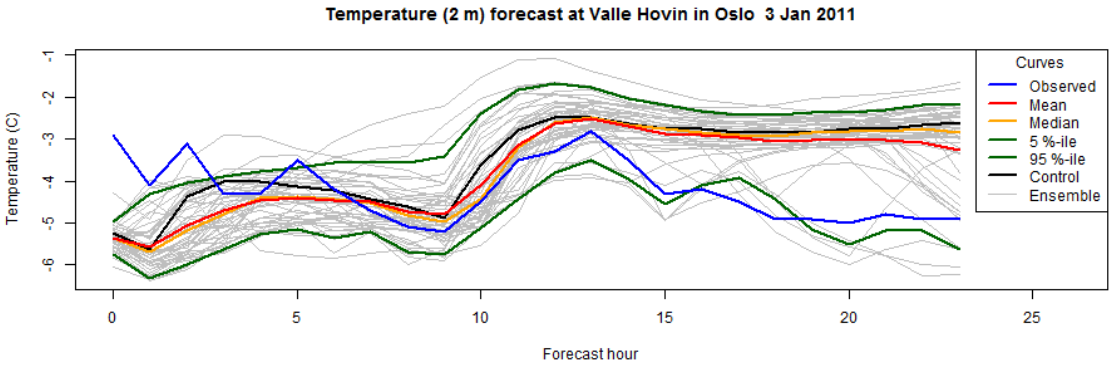

Figure 6: A 24 h forecast for 3 January 2011 of air temperature at 2 m above ground at station Valle Hovin in Oslo using the ECMWF forecast data and TAPM model. Unit: °C. Curves shown are: Observations (blue); TAPM ensemble mean (red), ensemble median (orange), 90% prediction interval (green), control run (black) and individual ensemble members (grey).

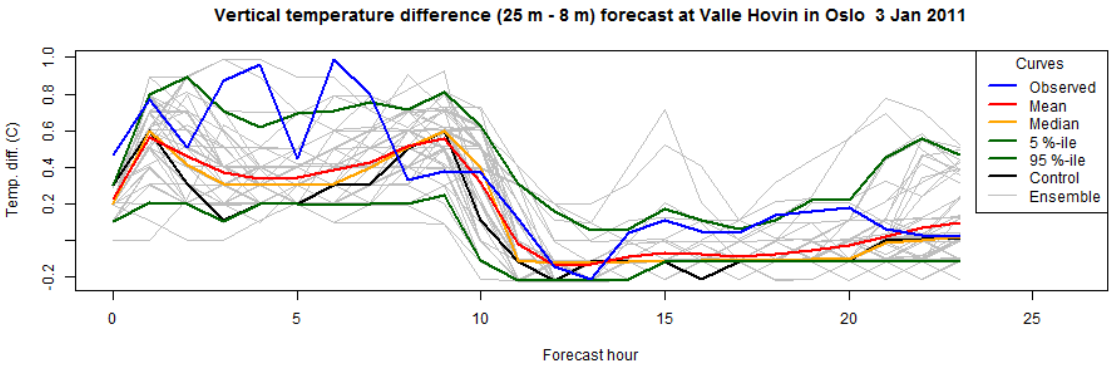

Figure 7: A 24 h forecast for 3 January 2011 of vertical air temperature difference between 25 m and 8 m above ground at station Valle Hovin in Oslo using the ECMWF forecast data and TAPM model. Unit: °C. Same curves and colours as used in Figure 6.

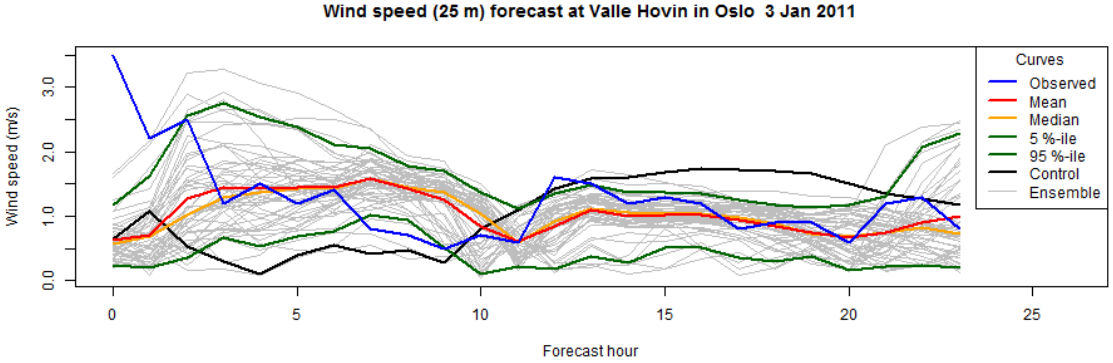

Figure 8: A 24 h forecast for 3 January 2011 of wind speed at 25 m above ground at station Valle Hovin in Oslo using the ECMWF forecast data and TAPM model. Unit: ms^{-1}. Same curves and colours as used in Figure 6.

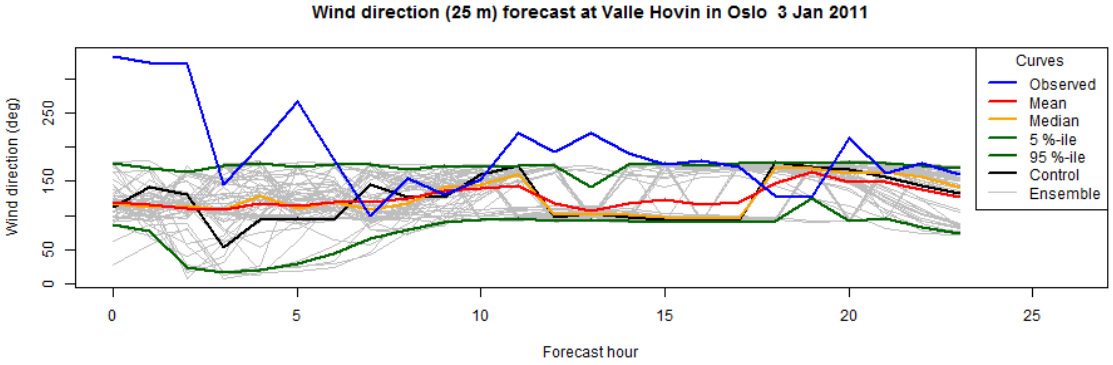

Figure 9: A 24 h forecast for 3 January 2011 of wind direction at 25 m above ground at station Valle Hovin in Oslo using the ECMWF forecast data and TAPM model. Unit: °. Same curves and colours as used in Figure 6.

If the probabilistic system is calibrated correctly, the observations (blue curve) should be in the green interval around 90% of the time. In practice this is difficult to achieve fully, but the results presented in these figures, although being preliminary, are quite encouraging, since in a relatively large fraction of the time (hours) observations are captured in this interval.

Notice also how the model uncertainty actually varies with time, illustrated by the variable spread of the ensemble (grey curves) and corresponding prediction interval (green curves).

4.2 GEMS/MACC model ensemble

Figure 10-Figure Figure 12 show 72h background concentration forecasts for Oslo based on the GEMS/MACC model ensemble and stochastic perturbation model as outlined in Section 3.4. Data from the GEMS/MACC model ensemble for January 2011 has been kindly provided by Françoise Chéroux at Météo-France.

Figure 10 displays background concentrations of NO_2, while Figure 11 and Figure 12 show background concentrations of O_3 and PM10 respectively.

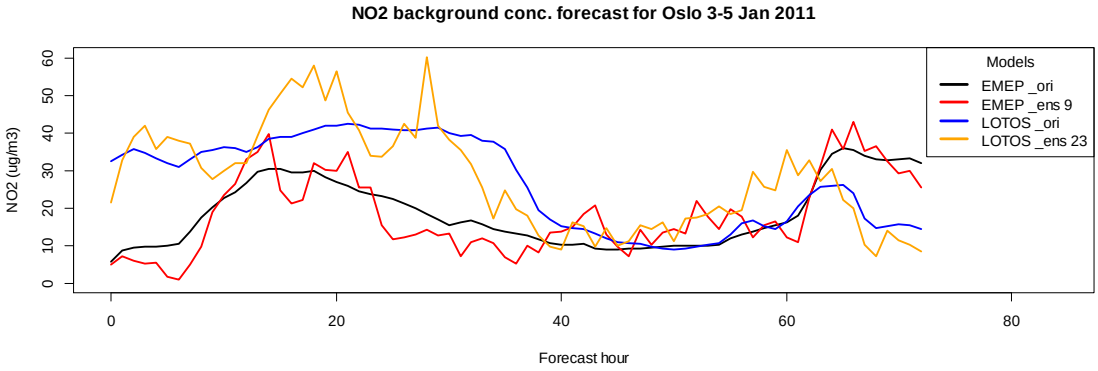

Figure 10: Original (GEMS/MACC model) and stochastically perturbed (Oslo PAQFS) forecasts of background concentrations of NO_2 for Oslo 3-5 January 2011. Unit: µgm^{-3}. Curves shown are: EMEP model (black); EMEP model perturbed (red); LOTOS-EUROS model (blue); LOTOS-EUROS model perturbed (orange).

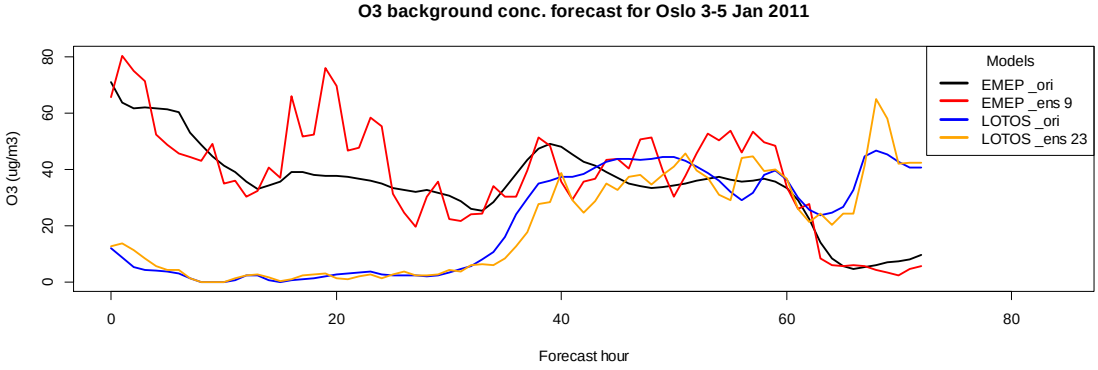

Figure 11: Original (GEMS/MACC model) and stochastically perturbed (Oslo PAQFS) forecasts of background concentrations of O3 for Oslo 3-5 January 2011 (0-72h). Unit: µgm^{-3}.
Same curves and colours used as in Figure 10.

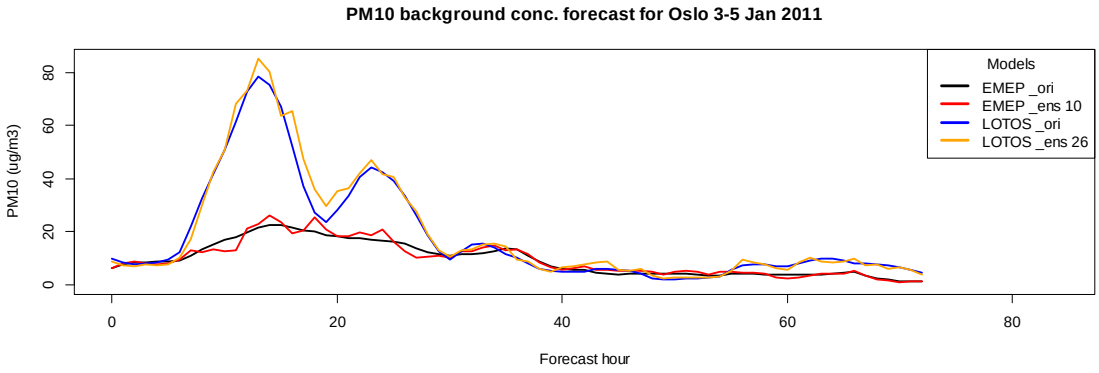

Figure 12: Original (GEMS/MACC model) and stochastically perturbed (Oslo PAQFS) forecasts of background concentrations of PM10 for Oslo 3-5 January 2011 (0-72h). Unit: µgm^{-3}. Same curves and colours used as in Figure 10.

In these figures only the EMEP and LOTOS-EUROS models is shown, and only with one of the stochastically perturbed members generated for each of these models. For this period all GEMS/MACC models were available for all compounds (except for the MOZART model for PM10), so since the complete ensemble has 50 members, 7-8 stochastically perturbed time series of background concentrations were produced for each of the MACC models.

4.3 EPISODE

Figure 13-Figure 15 show 24h air quality forecasts at 3 January 2011 at the roadside stations Kirkeveien and Åkerbergveien in Oslo using the EPISODE model. The colours in the figures have the same interpretation as the ones used in the Section 4.1.

Figure 13 displays ground level concentrations of NO_x at Kirkeveien, while Figures 14 and 15 show concentrations of NO_x at Åkerbergveien and PM10 at Kirkeveien, respectively.

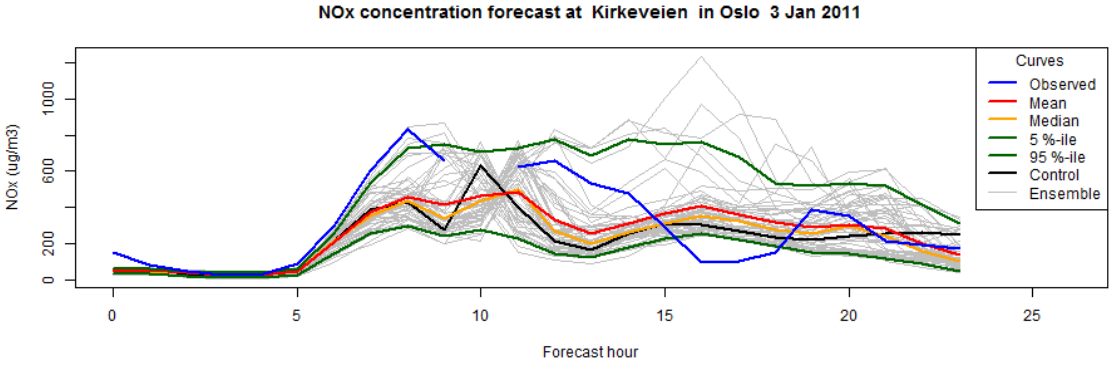

Figure 13: A 24 h forecast for 3 January 2011 of NO_x at 2 m above ground at station Kirkeveien in Oslo using the EPISODE model. Unit: µgm^{-3}. Curves shown are: Observations (blue); EPISODE ensemble mean (red), ensemble median (orange), 90% prediction interval (green), control run (black) and individual ensemble members (grey).

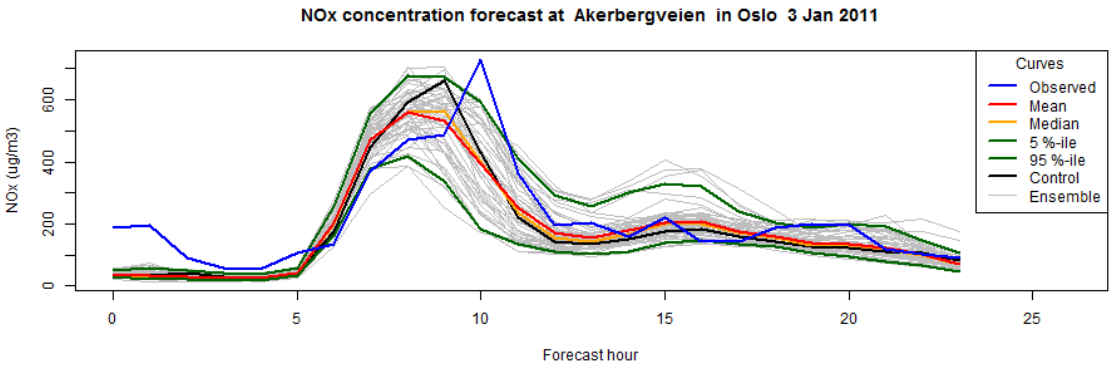

Figure 14: A 24 h forecast for 3 January 2011 of NO_x at 2 m above ground at station Åkerbergveien in Oslo using the EPISODE model. Unit: μgm^{-3}. Same curves and colours as used in Figure 13.

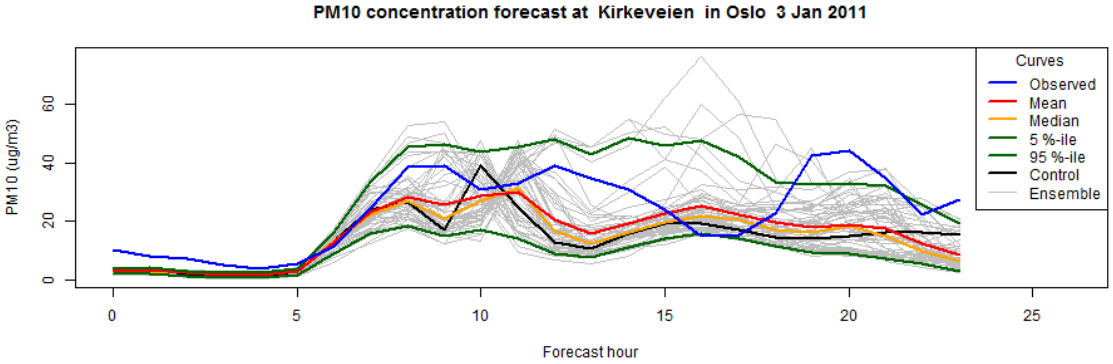

Figure 15: A 24 h forecast for 3 January 2011 of PM10 at 2 m above ground at station Kirkeveien in Oslo using the EPISODE model. Unit: μgm^{-3}. Same curves and colours as in Figure 13.

Again, if the probabilistic forecast system is calibrated correctly, the observations (blue curve) should be in the 90% central prediction interval (between the green curves) around 90% of the time. Although the results presented in these figures are quite preliminary, they are quite encouraging, since in a relatively large fraction of the time (hours) observations are actually captured in this prediction interval.

Also, notice how the model uncertainty varies over time, illustrated by the variable spread of the ensemble (grey curves) and the corresponding prediction interval (green curves). At night time there is very little activity and concentrations are very low with low uncertainty. Then during day time, especially during morning rush hours, concentrations become larger, with an attached larger uncertainty depicted by the increase in the spread of the ensemble.

Figure 16 shows the ensemble mean gridded concentrations of NO_x over Oslo at 3 January 9h. In the figure only the innermost 20 × 20 km^2 part of the EPISODE grid is shown, using a colour scale going from clean air (dark blue) to high concentrations shown in red and brown.

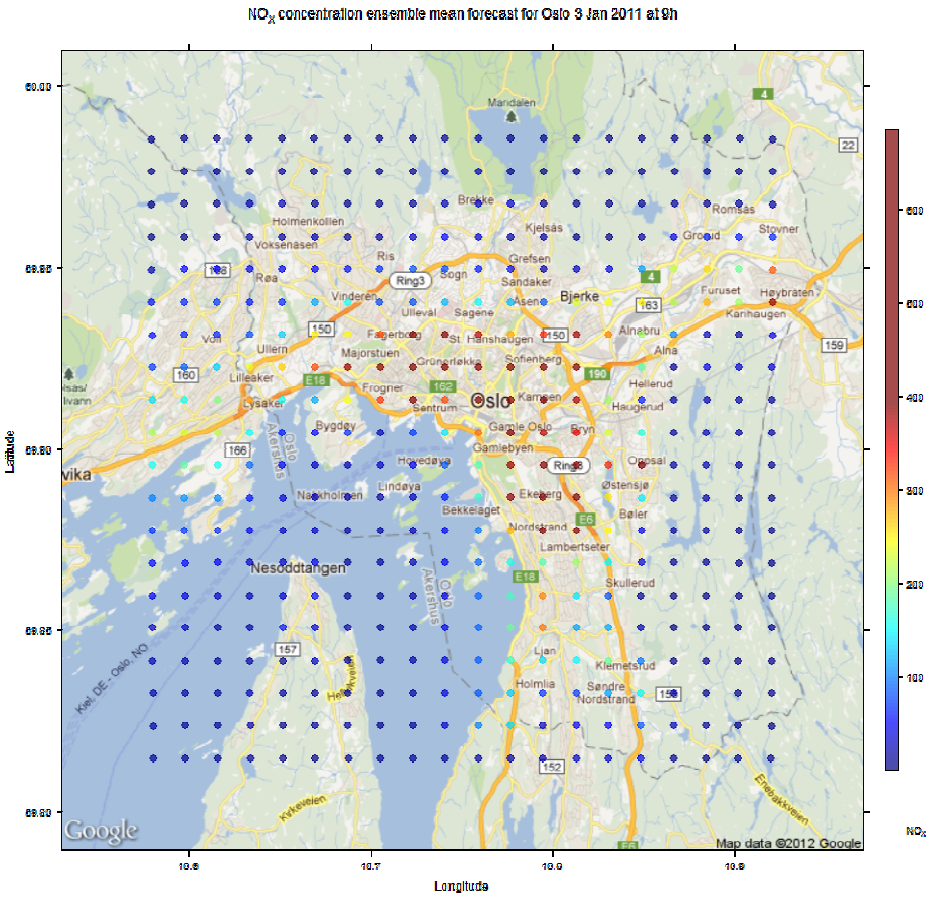

Figure 16: Ensemble mean ground level NO_x concentration forecast for Oslo at 3 January 2011 9h using the EPISODE model. Unit: μgm^{-3}.

As we see from this figure, concentrations are highest over the central part of Oslo.

5 Model communication and web integration

In this chapter we describe how the various models in the Oslo PAQFS model chain currently communicate uncertainties, and briefly suggest how this can be improved in later versions of the system. In the last section we discuss how the present system can be integrated in the UncertWeb web infrastructure.

5.1 ECMWF to TAPM

The ECMWF ensemble meteorological data are downloaded in the GRIB format. Data are then converted to a set of NetCDF files (one for each ensemble member) and then further to a special set of synoptic data files used by the TAPM model. This cannot be changed since we have no access to the code or internal parts of either of these systems. The only possibility is to change the intermediate NetCDF format used for individual ensemble members to a single netCDF-U file containing all members. This is, however, not a great improvement in the communication of uncertainty since this single netCDF-U file again has to be converted into individual sets of synoptic data for each running version of the TAPM model. It may, however, be practical to have the ECMWF ensemble meteorology in a single netCDF-U file for data archiving purposes, or for communication of these data to other potential users, e.g., through the use of project developed web interfaces.

5.2 Land use data for TAPM

Land use data (surface characteristics) in ensemble form are communicated to the TAPM model via a set of simple ASCII files. Again, this cannot be changed since we have no access to the program code. So the way we currently communicate uncertainties in land use data to TAPM cannot be changed.

5.3 TAPM to EPISODE

Ensemble meteorological data are produced by the TAPM model in the form of a set of binary output files and converted to a corresponding set of binary input files for EPISODE using the conversion program tapm2episode. An R script exists for converting the binary input files for EPISODE into a single netCDF-U file containing all the ensemble meteorology. Later this single file could be used to communicate uncertainties in meteorological data to EPISODE. Since several instances of the model are run in parallel using different ensemble data, each model instance will just need to pick out its own data from this single file (which then may be cached).

5.4 Emission data for EPISODE

Currently a set of specially formatted binary files created by the emission model is used to communicate emission ensemble data to the EPISODE model. As for the meteorology, an R-script exists for converting these binary files to a single common netCDF-U file. Later, this could also be changed so that the netCDF-U file was produced by the emission model and used directly by EPISODE. Again, each instance of the EPISODE model running in parallel will need to pick out its own data from this single file (which also may be cached).

5.5 GEMS/MACC to EPISODE

Data from GEMS/MACC are currently downloaded as separate NetCDF files, one for each hour of the forecast. These NetCDF files are then converted into a set of simple ASCII files, one for each ensemble member to be read by the EPISODE model. An R-script exists for converting these data into a single netCDF-U file. Again, it should be possible to directly convert the downloaded data into one common netCDF-U file and communicate this to the EPISODE model. Again the EPISODE model, running in parallel for each ensemble member, would have to read this single (cached) file and pick out only the data it needs.

5.6 Output from EPISODE

This is currently produced as a set of specially formatted binary files, one for each ensemble member. An R-script exists for converting these data into a single netCDF-U file. Again, it should be possible to change the EPISODE model code so that it produces NetCDF output files, rather than in the aforementioned formats. However, since each instance of EPISODE is run in parallel for the different ensemble members, it will not be possible to produce a single netCDF-U file directly. Rather, each model instance has to write out its own results to, say, a set of NetCDF files, which can then be combined into a single netCDF-U file when all model runs are finished. Again, for data archiving purposes, or for communication of these data to other potential users, e.g., through the use of web interfaces, it will be advantageous to have the data stored as a single netCDF-U file.

5.7 Web integration

In order to integrate the PAQFS in the UncertWeb infrastructure, a Web Processing Service (WPS) is currently being set up at NILU to provide some of the functionality of this model system on the Web. We envisage one or more web services to provide the functionality of creating emission data, meteorological data, background concentration data, and air pollution data in the form of 1-3 days (0-72h) forecasts and 2-3 hour updated nowcasts. The data will be provided through these services with uncertainty encoded as ensembles of data values using the UncertWeb developed netCDF-U format as described above. Interacting with this air quality service, external users will be able to select start and end date for the simulations, the number of ensemble members to be used, and which compounds to be calculated. In a further step the air quality model service will be coupled with the Albatross Web services developed in WP7 [UW-D7.2 2012] in order to assess the exposure of individuals to air pollution.

6 Conclusions

We have in this report presented the first version of the Oslo PAQFS, which is developed and implemented as a test case in the UncertWeb project. The first preliminary results with the system show that the ensembles of forecasts of meteorology and air pollution concentrations compares well with corresponding observations at various stations in Oslo. Model uncertainty is represented in the system in the form of ensembles and is encoded and stored in files using the netCDF-U format. Developing the system further will focus on integrating the system in the UncertWeb infrastructure as an air quality service using WPS.

References

[Bigagli 2011]

Bigagli, L., Nativi, S. “NetCDF Uncertainty Conventions (NetCDF-U)”. OGC Discussion Paper, 2011 (to be published).

[Cressie & Wikle 2011]

Cressie, N., Wikle C.K. “Statistics for spatio-temporal data”. John Wiley & Sons, Inc., New Jersey, 2011.

[EEA MDS 2012]

European Environment Agency, EIONET, Model Documentation System, AirQUIS-EPISODE model. http://pandora.meng.auth.gr/mds/mds.php.

[Hurley 2008]

Hurley, P.J. “TAPM V4. Part I: Technical description”, CSIRO Marine and Atmospheric Research Paper No. 25, 59 pp.

[R 2.14.1 2011]

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[UW-D1.2 2011]
UncertWeb Consortium, “UncertML best practice proposal”, Deliverable 1.2, 2011.

[UW-D2.2 2011]
UncertWeb Consortium, “Service Frameworks for modelling resources”, Deliverable 2.2, 2011.

[UW-D4.1-6.1 2010]
UncertWeb Consortium, “Consolidated requirements for service chains within UncertWeb”, Deliverable 4.1-6.1, 2010.

[UW-D7.2 2012]
UncertWeb Consortium, “Prototype UncertWeb activity model chain”, Deliverable 7.2, 2012.

[UW-D8.1 2011]
UncertWeb Consortium, “Report on integration requirements in UncertWeb”, Deliverable 8.1, 2011.