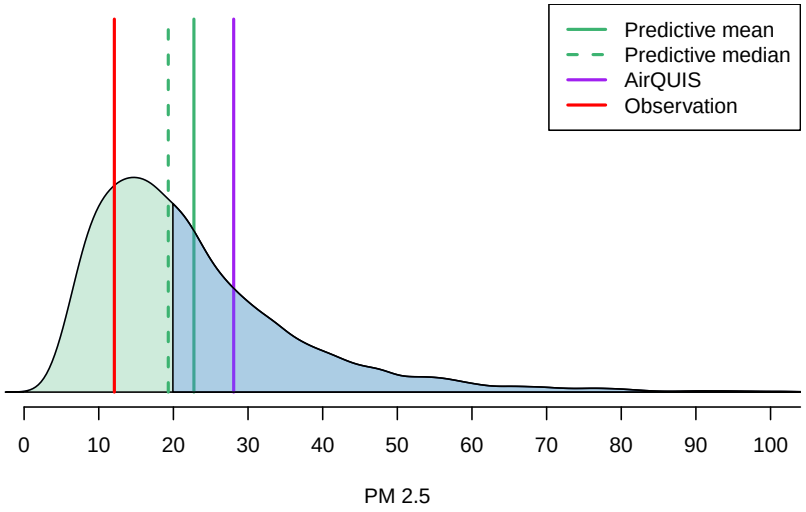
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Data assimilation and statistical post-processing for numerical air quality predictions



Note no SAMBA/49/14

Authors
Gunnhildur Högnadóttir Steinbakk
Thordis Thorarinsdottir
William Lahoz
Sam-Erik Walker

Date December 16, 2014

The authors

Gunnhildur Högnadóttir Steinbakk and Thordis L. Thorarinsdottir are Senior Research Scientists; William Lahoz is Senior Research Scientist and Sam-Erik Walker is Research Scientist at the Norwegian Institute for Air Research.

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norwegian Institute for Air Research

NILU, the Norwegian Institute for Air Research (Norsk institutt for luftforskning) is a non-profit private research organization. NILU is involved in approximately 250 projects each year for governments, industry, and national and international organizations. It has approximately 180 employees (of which approximately 75 scientists) and has accredited laboratories for chemical analyses and monitoring instruments. Its annual turnover is approximately 110M NOK.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Data assimilation and statistical post-processing for numerical air quality predictions**
Authors	**Gunnhildur Högnadóttir Steinbakk** <gunnhildur@nr.no> **Thordis Thorarinsdottir** <thordis@nr.no> **William Lahoz** <William.A.Lahoz@nilu.no> **Sam-Erik Walker** <Sam-Erik.Walker@nilu.no>
Date	December 16, 2014
Publication number	SAMBA/49/14

Abstract

This is a joint report based on work by NILU and NR on the use of data assimilation and statistical post-processing tools to improve the air quality prediction in the context of the Bedre Byluft programme. The objective is to improve Bedre Byluft air quality prediction at both the short and long-term (hours and days to weeks), in particular for meteorological conditions associated with high pressures, which historically have been difficult to predict. We present the tools; present first results and quantify the performance of the tools; and outline further work needed to make these tools operational.

Keywords	Bias correction, predicktive distribution, air pollution forecasts, communicating uncertainty
Target group	Anyone
Availability	Open
Project	Mulige løsninger for å kombinere deterministiske beregninger med stokastiske modeller for å forbedre prognosene for luftforurensning og spredning av lokale utslipp
Project number	220621
Research field	Statistics for climate, environment and health
Number of pages	66

Utvidet sammendrag

Meteorologisk Institutt, Norsk institutt for luftforskning (NILU) og Statens vegvesen har utviklet et prognoseverktøy for beregning og spredning av luftforurensning i de største norske byene på vinterstid gjennom FoU prosjektet Bedre Byluft (Ødegaard et al., 2013). Den numeriske spredningsmodellen for luftforurensing som inngår i Bedre Byluft systemet kalles EPISODE. Prognosemodellen kjøres hver natt i vinterhalvåret og produserer daglige prognoser for 1–48 timer fram i tid.

Denne rapporten er et samarbeid mellom NILU og Norsk Regnesentral (NR) for å forbedre de numeriske prognosene av NO_2, PM_{10} og $PM_{2,5}$ i Bedre Byluft prosjektet ved å undersøke to tilnærminger: Den ene kombinerer en skjevhetskorrigering av prognosene og dataassimilering for å forbedre inngangsdata til EPISODE. Det andre arbeidet er en statistisk etter-prosessering av utgangsprognosene fra EPISODE, som korrigerer de opprinnelige prognosene basert på tidligere avvik mellom målinger og prognoser. Hovedfunnene oppsummeres i dette kapittelet, mens resten av rapporten er på engelsk og inneholder en mer detaljert beskrivelse av analysene og de matematiske modellene.

Forbedring av inngangsdata

På oppdrag fra Statens Vegvesen (Vegdirektoratet) har NILU utviklet metoder for å forbedre EPISODE modell-prognosene i Bedre Byluft, spesielt i perioder med uforutsett stagnerende meteorologiske forhold. Rapporten beskriver resultatene av en test av metodene i Oslo for en periode sist vintersesong.

Som ledd i Bedre Byluft prosjektet gjennomfører NILU hver dag i vinterhalvåret (1. okt–1. mai) 48 timers prognoseberegninger med spredningsmodellen EPISODE for flere norske byer og tettsteder i samarbeid med Meteorologisk Institutt. I de fleste tilfeller gir modellen en rimelig bra overensstemmelse med målte konsentrasjoner på målestasjonene.

I noen situasjoner hver vintersesong kan det imidlertid oppstå forhold med mer stagnerende luftmasser og svakere vind enn det som varsles av prognosemodellen, noe som ofte resulterer i beregnede konsentrasjoner som er lavere, og til dels mye lavere, enn det som observeres. I slike situasjoner er det viktig å kunne korrigere de beregnede konsentrasjonene for de første 24 timene av 48 timers prognoseperioden for å bringe de mer i overenstemmelse med målte verdier, og som grunnlag for en forbedret prognose neste dag.

I denne delen av arbeidet er det utviklet to metoder for å forbedre EPISODE modell prognosene de første 24 timene basert på målinger: (1) Bias-korreksjon; og (2)

data-assimilasjon ved å bruke ensemble Kalman filter.

Metodene er blitt testet ved å bruke målinger og 24 timers modellprognoser for hver av de tre komponentene NO_2, PM_{10} og $PM_{2.5}$ i Oslo for perioden 2.– 8. desember 2013. Resultatene viser at begge metodene fungerer bra, og at de klarer å forbedre modellprognosene sammenlignet med målinger på de fleste av målestasjonene i Oslo i testperioden.

Etter-prosessering av luftkvalitetsprognoser

I 2013 utviklet NR en stokastisk prototypmodell for avvik mellom prognoser for og målinger av luftkvalitet i Oslo for komponentene NO_2, PM_{10} og $PM_{2,5}$, samt grovfraksjonen PM_{10}–$PM_{2,5}$. Analysen viste at de statistiske etterjusterte prognosene traff betydelig bedre enn de originale prognosene (Steinbakk et al., 2013). Idéen er at vi kan lære av historiske avvik mellom observasjoner og tilhørende prognoser i målestasjonene, og deretter bruke denne kunnskapen til å korrigere de opprinnelige 1–48 timers prognosene. Analysen i denne rapporten er basert på det samme datagrunnlaget som i Steinbakk et al. (2013), det vil si fra 10 målestasjoner i Oslo fra vintersesongene 2011-2012 og 2012-2013.

Osloregionen har flere målestasjoner enn de andre byene i Norge som omfattes av Bedre Byluft. I dette arbeidet har vi derfor undersøkt effekten av å etterjustere de originale prognosene basert på data fra få målestasjoner. Det vil si at vi kun har brukt en liten andel av målestasjonene i Oslo for å trene opp parametere i den statistiske prototypmodellen og sammenlignet de justerte prognosene med data fra målestasjoner som ikke ble brukt til å trene modellen. Denne tilnærmingen gjentas så for ulike grupper av målestasjoner. I hovedsak har vi brukt tre stasjoner for å trene modellen, men appendiks D viser i tillegg valideringsresultater oppsummert i tabeller ved å bruke to stasjoner.

Prognosene fra EPISODE i Bedre Byluft foreligger rundt klokken seks om morgenen en gang i døgnet. For å validere resultatene trenger vi et par dager i begynnelsen av sesongen for å trene opp modellen. Vi trener derfor opp modellen fram til klokken seks den femte dagen i observasjonsperioden for hver vintersesong. Deretter sammenligner vi de justerte 1-24 timers prognosene med observasjoner i målestasjoner som ikke er de samme som ble brukt til å trene modellen. Dette gjentar vi så daglig for hele perioden, slik at treningssettet blir lengre utover vintersesongen. Vi analyserer vintersesongene hver for seg.

Vi evaluerer prognosenes treffsikkerhet ved å sammenligne observerte verdier mot tilhørende justert prognose med ulike deskriptive mål, og sammenligner så tilsvarende resultater med de originale prognosene. De deskriptive målene vi har brukt er *kvadratroten av gjennomsnittlig kvadrert feil* (RMSE) og lineær *korrelasjonen* (COR) mellom observert verdi og prognose, samt *gjennomsnittlig absolutt*

feil (MAE). Lavere verdier av RMSE og MAE indikerer bedre tilpasning mellom prognose og observasjon. RMSE og MAE er lik 0 dersom prognosene og observasjonene er helt like. Korrelasjonen COR er et tall mellom -1 og 1 hvor 1 er perfekt korrelasjon.

Selv ved å bruke få stasjoner til å trene modellen, blir treffsikkerheten til de etterjusterte prognosene nesten alltid bedre enn de originale prognosene. Korrelasjonene for de etterjusterte prognosene for NO_2 er alltid bedre enn de originale prognosene, men resultatene for RMSE blir dårligere. At forbedringen i korrelasjonen er så mye bedre enn RMSE, kan tyde på at forurensningsnivået i modellen for NO_2 ikke er helt korrekt tilpasset.

I tillegg til å forbedre prognosene kan dette rammeverket beskrive usikkerheten til prognosene som en fordelingsfunksjon. En sannsynlighetsfordeling vil gi oss mulighet til å beregne ulike størrelser, for eksempel å anslå sannsynligheten for at forurensningnivået vil overskride en gitt verdi. Vanlige valg av punktestimat i en sannsynlighetsfordeling er forventningen og medianen i fordelingen. Medianen er ikke lik sensitiv for ekstreme verdier som forventningen og kan derfor være en mer robust tilnærming, som kan være et viktig moment i en eventuell implementering av metoden i Bedre Byluft.

Generelt viste denne analysen at medianestimatet i den prediktive sannsynligheten for NO_2 traff bedre enn forventningsestimatet, men for svevestøv var forventningsestimatet bedre. I en operasjonalisering burde det undersøkes nærmere hvilket estimat som er best å bruke. Fordelingen til en luftforurensningskomponent vil typisk være skjev med tung hale mot høyre (høyere forurensningsverdier). I et videre arbeid bør også usikkerhetsmodellen for NO_2 kalibreres bedre, slik at modellen gir en realistisk beskrivelse av usikkerheten.

Anbefalinger ved en operasjonalisering og videre arbeid

Bias-korreksjonsmetoden uten dataassimilering kan, på samme måte som den statistiske prototypmodellen, brukes til å etterjustere prognosene fra EPISODE. Den statistiske prototypmodellen korrigerer også for bias i de originale prognosene, men justerer for autokorellasjon (avhengigheter i tid) i prognosefeilen i tillegg. Fokuset for bias-korreksjonsmetoden var å forbedre prognosene under spesielle værsituasjoner (stagnerende meteorologiske forhold) som for eksempel oppstod i Oslo 2.–8. desember 2013. Begge tilnærmingene viser forbedring i prognosene etter justering, men er testet ut på data fra Oslo-regionene for ulike tidsperioder.

Både bias-korreksjonsmetoden og den statistiske prototypmodellen kan implementeres som en uavhengig modul som justerer de originale prognosene fra EPISODE en gang i døgnet. Metodene krever ikke ensembler av modellbereg-

nede verdier, da den kan anvendes i kombinasjon med de eksisterende EPISODE modellberegningene i det nåværende Bedre Byluft systemet. En slik modul vil derfor være beregningsmessig rask og kreve minimalt med regnekraft. Det er imidlertid flere praktiske problemstillinger vi må ta hensyn til for å få et operasjonelt system. Blant annet bør måleverdiene til forurensningskomponentene kvalitetssikres fortløpende.

Implementasjon av dataassimilasjon ved bruk av ensemble Kalman filter vil kreve større endringer av beregningssystemet, siden vi da må utføre flere parallelle kjøringer med EPISODE modellen (f.eks. minst 5-10) for hver varslingsperiode. Dette vil kreve større endringer i modelloppsettet og skriptsystemet, og vil i tillegg kreve mer regneressurser. Dataassimilering spiller imidlertid en viktig rolle i arbeidet med å forbedre spredningsmodeller som EPISODE ved å sammenligne og kombinere slike modeller med tilgjengelige observasjoner.

Den statistiske etter-prosesseringen kan beskrive usikkerheten til prognosene ved en (prediktiv) sannsynlighetsfordeling, istedenfor et punktestimat. Dersom usikkerhetsvurderinger skal brukes i et operasjonelt system som Bedre Byluft, har dette arbeidet vist at den prediktive sannsynlighetsfordeling til NO_2 bør kalibreres bedre mot empiriske data. Den prediktive fordelingen til svevestøvskomponentene har derimot vist seg å passe mye bedre til de empiriske dataene. I denne sammenhengen kan det også være lurt å undersøke om medianestimatet er mer robust for endringer og store avvik i prognosefeilen enn forventningsestimatet.

Denne rapporten diskuterer også hvordan usikkerhet kan kommuniseres i kart. Figur 9 viser et tenkt eksempel med sannsynligheten for å overskride luftforurensningnivået med en gitt terskel. Det nåværende rammeverket gir bare mulighet for å beregne usikkerheten i hvert gridpunkt. Dersom vi vil beregne usikkerheten på tvers av et område, for eksempel langs en vei, trenger vi å utvide modellen til å beskrive korrelasjonsstrukturen mellom gridpunktene.

Contents

1 Introduction **9**

2 Bias correction and data assimilation **10**
2.1 Background 10
2.2 Bias correction 12
2.3 Data assimilation 15
2.4 Results 17
2.4.1 NO_2 18
2.4.2 PM_{10} 22
2.4.3 $PM_{2.5}$ 26

3 Statistical post-processing **30**
3.1 Data 31
3.2 Model for statistical post-processing 31
3.3 Parameter estimation 32
3.4 Probabilistic forecasts 33
3.5 Evaluation of predictive performance 33
3.6 Results 35

4 Communicating uncertainty **40**

5 Discussion and concluding remarks **44**
5.1 Discussion 44
5.2 Future work 46

References **47**

A Time series plots of observed and uncorrected model concentrations **50**

B Time series plots of observed and corrected model concentrations **57**

C Algorithm for computing the predictive distribution **64**

D Results for two neighbouring measurement stations **65**

1 Introduction

Through the research project Bedre Byluft (Ødegaard et al., 2013), the Norwegian Meteorological Institute (MET Norway), the Norwegian Institute for Air Research (NILU) and the Norwegian Road Administration (VVS) provide numerical predictions for air quality in the largest cities of Norway. In this context, NILU has developed an integrated air quality management system (AirQUIS)[1] containing a suite of tools for monitoring and predicting air quality. One of the AirQUIS modules is the numerical dispersion model EPISODE which calculates spatial distribution of hourly concentrations of NO_2 and of particulate matter, PM, for sizes less than 2.5 microns and 10 microns, $PM_{2.5}$ and PM_{10} (Slørdal et al., 2003).

Each winter season, as part of the Bedre Byluft project, NILU applies the EPISODE model to forecast air pollution in several Norwegian cities, in cooperation with Met Norway. While there is generally a reasonable correspondence between observed and modelled concentrations, some biases may be observed. For example, occasionally, unforeseen stagnant meteorological conditions may occur, where observed wind speeds are much lower than forecast, which leads to predicted concentrations being much lower than observed (sometimes as much as between a quarter and a half of observed values).

This report presents a joint work between NILU and the Norwegian Computing Center (NR) to improve the numerical predictions of NO_2, $PM_{2.5}$, and PM_{10} within Bedre Byluft by investigating two complementary approaches. A data assimilation technique combined with bias correction aims to improve the input data that enters into the EPISODE model, while the goal of the other approach, a statistical post-processing of the EPISODE model output, is to correct the model output by utilizing recent forecast errors.

NILU has developed a bias correction procedure and a data assimilation method using the ensemble Kalman filter, EnKF (Evensen, 2007; Sakov and Oke, 2008). Both of these methods have been tested, and results are shown, using observations and forecast model data from Oslo for NO_2, $PM_{2.5}$, and PM_{10} during a period with stagnant meteorological conditions from the last Bedre Byluft season (2013-2014). The improvement is typically in the range 20-80 % for root mean squared errors, and an increase of 0.2-0.3 for the correlations overall. The description of the methods and the corresponding results are given in Section 2. The EnKF data assimilation method readily provides the uncertainty of the predicted field, and provides diagnostics that allow to test assumptions about the error characteristics of the input model and observational fields. We discuss in

1. http://www.nilu.no/airquis/what_is_airquis.htm

Section 4 this role of uncertainty.

The statistical post-processing method developed by NR in 2013 provides adjusted forecasts which improved the root mean squared forecast error by 10-20% and yield 10-40% better correlation for the particulate matter (Steinbakk et al., 2013). In Section 3 we investigate whether the gain of post-processing is still significant when only a few air quality observation stations are available. As most of the cities in Norway have very few monitoring stations, this is an important test to see whether post-processing of forecasts should be implemented in cities other than Oslo.

The post-processing framework can provide fully probabilistic forecasts in the form of predictive distributions. Additionally, we explore methods for assessing an appropriate uncertainty distribution in Section 3. Predictive exceedance probabilities can be, for example, an important decision-making tool for deciding if pollution measures such as limiting traffic are necessary or not.

The use of ensembles and statistical post-processing for describing forecast uncertainties, are now routinely used in many applications for numerical predictions, such as weather forecasting. Section 4 illustrates and discuss how uncertainties and other results derived from our predictive probabilistic prototype model can be presented in maps. Section 5 provides conclusions and outlines further work required to make operational the tools described.

2 Bias correction and data assimilation

In Section 2.1, we give a general introduction and background to data assimilation including the need for bias correction. In Sections 2.2 - 2.3, we describe in more detail the bias correction and data assimilation methods that NILU have implemented. Section 2.4 shows the results of applying these methods using observed and model data from Oslo for each of the three species NO_2, PM_{10} and $PM_{2.5}$ from the last Bedre Byluft season (2013-2014).

2.1 Background

Observations are essential to estimate the state of the Earth System, including air quality in an urban environment. Observations have two key limitations. The first one is that they contain errors — these can be systematic (also called bias), random, and of representativeness (see Cohn, 1997; Lahoz et al., 2010a; Ménard, 2010). Averaging reduces the random errors, but not the systematic errors. Model bias is the systematic difference or error between model values and a underlying

estimate of the system state, e.g., provided by observations, at a given time, or over a given time period. It is important to correct the bias before applying data assimilation, since the EnKF and other data assimilation methods, generally assume that observations and model values are unbiased. The second limitation is that observations have spatio-temporal gaps (Lahoz et al., 2010a). It is necessary to fill in the gaps in the information provided by observations: (i) to make this information more complete, and more useful; and (ii), to provide information at a regular scale to allow easier quantification of physical processes. One can use information at an irregular scale to quantify physical processes, but this procedure is more tractable when done at a regular scale.

To fill in the gaps in the observations a model is needed (Lahoz et al., 2010a). This model can be simple, e.g., linear interpolation, or take account of the system's behaviour. For example, the model could be a chemistry-transport model (CTM), such as EPISODE (Slørdal et al., 2003), incorporating a suite of chemical equations. The model extends the observations, fills in the observational gaps and allows one to organize, summarize, and propagate the information from observations. The model, like the observations, also exhibits gaps in space and time.

We require methods that fill in the observational information gaps in a way that makes use of quantitative concepts for combining information. For example, by finding the state that minimizes a "penalty function" calculated from observational information and prior information (e.g., from a model forecast). We can think of the model used for the forecast as an intelligent interpolator of the observational information. A methodology that allows this intelligent interpolation is data assimilation (Kalnay, 2003; Lahoz et al., 2010b). It has strong links to a number of mathematical disciplines, including control theory and Bayesian estimation (Nichols, 2010).

Data assimilation adds value to the observations by filling in the observational gaps, and adds value to the model by constraining it with observations (Lahoz et al., 2010a) This allows self-consistent and realistic representation of the Earth System on a regular grid. Data assimilation provides methods for combining in an objective way observations and models with different spatio-temporal characteristics and errors: local footprint vs. quasi-global footprint; local coverage vs. global coverage; differences in sampling frequency; and errors arising from matching different spatio-temporal scales. The weather forecasting agencies provide an example of how data assimilation combines heterogeneous observational and model information (Kalnay, 2003). Dee et al. (2014) discusses the success of data assimilation in providing weather forecasts. The result of data assimilation, combining observational and model information and their errors, is termed the "analysis." In data assimilation, we never know precisely the observations, mod-

els, and analyses have errors, so we need to estimate them. This means we must state the data assimilation problem in probabilistic terms (see, e.g., Cohn, 1997).

The objective combination of information from a model and from observations can be formulated mathematically using Bayesian estimation ideas (Rodgers, 2000). Although Bayesian estimation defines a systematic and rigorous approach to data assimilation (Evensen, 2007; Rodgers, 2000), its full-scale implementation in many areas, including weather forecasting, is impossible, chiefly due to the size of the problem. The typical dimension of current weather forecasting models is $\sim 10^7$ elements, while the number of observations available over 24 h is $\sim 10^6$–10^7 (Lahoz et al., 2007). As a result, error covariance matrices for the model and observational information have $\sim 10^{14}$ elements. Thus, in many practical applications it is necessary to make simplifying assumptions to the data assimilation methodology.

Simplification of the data assimilation methodology follows two main lines: (i) statistical linear estimation, principally involving variational methods (e.g., the 4-D variational method,4D-Var) and sequential methods (e.g., the Kalman filter, KF), and (ii) ensemble assimilation (e.g., the ensemble Kalman filter, EnKF). In the context of statistical linear estimation, 4D-Var and the KF methods are two different algorithms for determining the BLUE (Best Linear Unbiased Estimate), and they are equivalent only under the condition of linearity. Ensemble assimilation is a form of Monte-Carlo approximation that attempts to estimate the probability distribution functions (PDFs) using a finite number of elements. In the EnKF (Evensen, 2003), used in this project, a Monte-Carlo ensemble of short-range forecasts is used to estimate the forecast error in the KF. In the EnKF, the size of the analysed ensembles typically lies between a few tens to a few hundreds of model states. The estimation becomes more accurate as the ensemble size increases. The EnKF is attractive as, for example, it requires no derivation of a tangent linear operator or adjoint equations and no integrations backward in time, as for 4D-Var (Evensen, 2003). Several authors (e.g., Kalnay et al., 2007; Lorenc, 2003) have compared 4D-Var and the EnKF, with an emphasis on their suitability for weather forecasting.

2.2 Bias correction

The bias correction procedure implemented here first estimates the model bias at a set of observation stations, before interpolating these station, or point, biases to other regions of the model area using spatial interpolation.

The model bias, for a given species, station and time point (hour), is estimated by calculating the average of the differences between observed and model calculated concentrations over a given time period, chosen to be close to the time point of

interest. Since air quality measurements are generally much more accurate than modelled values (e.g., observed NO_2 typically has an accuracy of about $\pm 5\%$), they can be used as reference against which model bias is estimated. Thus, for each station s_i we estimate the model bias $\hat{B}_{s_i,t_0}$ at the time point (hour) of interest t_0 by

$$\hat{B}_{s_i,t_0} = \frac{1}{T}\sum_{t=t_1}^{t_2}\{y_{s_i,t} - M_{s_i,t}\}; T = t_2 - t_1 + 1 \quad (1)$$

for $i = 1, \ldots, m$, where m is the number of observation stations, and where $y_{s_i,t}$ and $M_{s_i,t}$ denote, respectively, the observed and model calculated values at station i and time (hour) t, for $t_1 \leq t \leq t_2$, a period of length $T = t_2 - t_1 + 1$ (in time units).

The interval length T of the bias estimator should be chosen long enough to estimate the model bias, rather than short-term differences or errors between observed and model calculated values; yet should be chosen short enough to be able to estimate the bias, at the time point t_0, with reasonable accuracy. Thus, we want the average of the observed and modelled values to be representative of the time point of interest.

In this work we use Eq. (1) with $t_1 = 13$ and $t_2 = 23$ (i.e., $T = 11$ hours) to estimate the station biases at midnight ($t_0 = 24$) between the first and second day of forecasting.

Given biases estimated at each station as above, we can use spatial interpolation to estimate, or predict, the model bias at any other point in the model domain. For this purpose, we use ordinary kriging (Cressie and Wikle, 2011; Matheron, 1963). Thus, the model bias at an arbitrary spatial location (point) s_0 and time point (hour) t_0 (here $t_0 = 24$) is calculated by

$$\hat{B}_{s_0,t_0} = \sum_{i=1}^{m} w_i \hat{B}_{s_i,t_0} \quad (2)$$

where w_i is the weight attached to the bias at station s_i, $i = 1, \ldots, m$, for predicting the model bias at the point s_0. We calculate the weights separately by solving a linear system of equations for each point of interest. In ordinary kriging, the weights $w_1 \ldots, w_m$ always sum up to one, which makes the predictor in Eq. (2) unbiased for any spatially constant mean field of bias values.

To define the linear system of equations, and solve for the weights, we need to specify the correlations between estimated biases at the stations, and between stations and the spatial location s_0 of interest. Here, the following exponential correlation function is used

$$\rho(s_i, s_j) = \exp\left(-\frac{||s_i - s_j||_2}{\delta}\right) \quad (3)$$

where $||s_i - s_j||_2$ denotes the spatial distance between the points s_i and s_j, and δ is a scaling constant. If δ is small, station biases will only have a local influence close to each station; while if δ is large, station biases will influence a larger region around each station. Far away from the stations, however, the bias predicted by Eq. (2) will be close to the average of the station biases.

In this work, we use $\delta = 3$ km in the correlation function Eq. (3). Both smaller (down to 1 km) and larger (up to 5 km) values of this parameter were also tested, but it was decided to settle, at least tentatively now, for this value as a reasonable compromise between a small and large influence of the various estimated station biases. Ideally, we should estimate this parameter from data, i.e., using correlations between the biases calculated at each station.

The spatial bias predictor (or interpolator) in Eq. (2) is mainly used to estimate or predict the model bias at each grid point of the EPISODE 3D model grid (at the lowermost layer near the ground) at midnight ($t_0 = 24$) between the first and second day of forecasting. We then calculate new bias-adjusted model values at this time point by adding the estimated bias values to the model grid values M_G as follows

$$\hat{M}_G(i, j, 1) = M_G(i, j, 1) + \hat{B}_{s_0,t_0} \tag{4}$$

where s_0 represents the midpoint of grid cell (i, j), for $i = 1, \ldots, n_x$, $j = 1, \ldots, n_y$, and where n_x and n_y denote the number of grid cells in the E-W (East-West) and S-N (South-North) directions, respectively. We then use these new bias-adjusted model grid values as initial concentrations in the EPISODE model for the next 48-hour forecast.

The spatial bias predictor (or interpolator) in Eq. (2) is also used to estimate or predict the model bias at midnight ($t_0 = 24$) at other receptor points of interest, e.g., at observation stations where observations (and thus station biases) are missing, or at any other receptor points used by EPISODE (e.g., building locations, etc.).

We then calculate new bias-adjusted model receptor values $\hat{M}_{r,i}$ at this time point by adding the estimated bias values to the model receptor values M_r as follows

$$\hat{M}_{R,i} = M_{R,i} + \hat{B}_{s_0,t_0}, \tag{5}$$

where s_0 successively represents each receptor point s_i, for $i = 1, \ldots, n_r$, and where n_r denotes the total number of receptor points.

Finally, all receptor biases (including station biases) calculated at midnight are added to the EPISODE model receptor concentrations for time points (hours) 7-23 (7 am to 11 pm) of the current day, and to time points (hours) 1-6 (1 am to 6 am) used by the next 48-hour forecast. Thus, we assume that the model bias is constant over the time period ranging from 7 am (current day) to 6 am (next day).

2.3 Data assimilation

The main application of data assimilation in this work is to update the bias-corrected EPISODE model grid concentrations at midnight between the first and second day of forecasting, using observations from all available stations at that time point. To this end we use the EnKF method (Evensen, 2007; Sakov and Oke, 2008).

In the EnKF, we operate with a set, or ensemble, of N different model states $x_t^{(i)}, i = 1, \ldots, N$, (ideally) representing a (discrete) probability distribution of the underlying true (but unknown) system state at each time point T over a given period of time $t = 1, \ldots, T$. We then use the observations to update the ensemble at selected points in time, which (generally) will have the effect of reducing the uncertainty in the ensemble, bringing the ensemble (in particular the ensemble mean) closer to the true (but unknown) system state.

In our application of the EnKF, we define the model state x_t as a vector of the 522 (29×18) EPISODE model grid concentrations (lowermost layer of the 3D grid) at each time point (hour) during the 2 days (48 hour) forecasting period, since this is the output of the model that we wish to update using all available observations (the update being at midnight). To apply the EnKF, we therefore need to create an ensemble of N different model grid concentrations in the EPISODE model over each forecasting period, representing a realistic discrete probability distribution for the true state (true ground level grid concentrations) at each time point (hour).

We do this by creating corresponding ensembles of emissions (home heating and traffic) and background concentrations as input data to the EPISODE model. We generate these ensembles by randomly perturbing the deterministic model input data using Monte Carlo random draw procedures to simulate model (input data) uncertainties. The meteorological data used to force the model (i.e., the HARMONIE model fields), are currently not perturbed. Therefore, these data are unchanged for all ensemble members.

We currently use a Gaussian distribution with a 20% relative error standard deviation for the perturbation of emissions from both home heating and traffic for all species (NO_2, PM_{10} and $PM_{2.5}$). We perform the perturbations on cube-root transformed emission data, before being transformed back to the original scales. For traffic, the perturbations are used both for area and line source emissions. Emission values, which are exactly zero, are not perturbed.

For background concentrations (where the deterministic model values are based on the MACC (Monitoring Atmospheric Composition and Climate)[2] ensemble mean) we use a Gaussian distribution with 5% relative error standard deviation

2. http://www.gmes-atmosphere.eu

for all species (NO_2, O_3, PM_{10}, $PM_{2.5}$), again on cube-root transformed data before transforming the data back to the original scales. Background concentrations, which are exactly zero, are not perturbed.

The EPISODE model is then run (in parallel) N times (with N different sets of input data of emissions and background concentrations) over each 48-hour forecasting period, to propagate and update the ensemble of model states (ground level grid concentrations). In these runs, we do not apply extra perturbations (i.e., stochastic physics errors) in the EPISODE model. The ensemble size chosen for this work is $N = 7$. We mainly choose this number of ensemble members to perform the current calculations reasonably fast; we can easily increase the number of ensemble members later.

At midnight ($t_0 = 24$) between the first and second day of forecasting, all available observations at this time point (hour) are used by the EnKF to update the model state, i.e., the ground level grid concentrations. We then use the updated model state as a set of improved initial concentrations in the EPISODE model for the next 48-hour forecast run (operationally in Bedre Byluft this starts at around 5 am).

Thus, if $x^f_{t_0}$ represents the forecast ensemble mean of model ground level grid concentrations at time t_0 ($t_0 = 24$), the updated, or assimilated, ensemble mean of these grid concentrations $x^a_{t_0}$ (at the same time point) is given by

$$x^a_{t_0} = x^f_{t_0} + K\left\{y_{t_0} - H(x^f_{t_0})\right\} \tag{6}$$

where K is the Kalman gain matrix (Evensen, 2007; Sakov and Oke, 2008); y_{t_0} is the vector of observations at time t_0 (at the various stations); and $H(\cdot)$ is the non-linear observation operator linking the model state with the observations. Generally one may consider $H(x)$ as the vector of expected observations (at the various stations) if x represents the true model state.

In this work, we define the observation operator as

$$H(x^f_{s_i,t_0}) = x^f_{s_i,t_0} + L^f_{s_i,t_0}, \tag{7}$$

where $x^f_{s_i,t_0}$ represents the ground level grid concentration, and $L^f_{s_i,t_0}$ represents the ensemble average of the sub-grid scale line source model contribution, at each observation station s_i at time t_0 for $i = 1, \ldots, m$, where m is the number of stations. It is important and necessary to add sub-grid scale line source concentrations to the grid concentrations in Eq. (7) since most of the stations in Oslo are roadside stations.

The new updated, or assimilated, receptor concentrations at each receptor point s_i for $i = 1, \ldots, n$ (here receptor points also include all observation stations) at

time t_0 (midnight) are given by

$$H(x^a_{s_i,t_0}) = x^a_{s_i,t_0} + L^f_{s_i,t_0}.$$

In the numerical implementation of the data assimilation procedure, the observation operator variances σ^2_R implicitly used in Eq. (6) are defined so that the following equation

$$\sigma^2_R + \sigma^2_M = \sigma^2_\chi$$

is satisfied, where σ^2_M denotes model ensemble variance, and σ^2_χ denotes the chi-square analysis variance. We calculate the latter two variances as follows

$$\sigma^2_M = \frac{1}{N-1}\sum_{k=1}^{N}\left\{x^{f,(k)}_{s,t_0} - \bar{x}^{f,(\cdot)}_{s,t_0}\right\}^2;\ \ \sigma^2_\chi = \frac{1}{T-1}\sum_{t=t_1}^{t_2}\left\{y_{s,t} - \bar{x}^{f,(\cdot)}_{s,t}\right\}^2$$

with $t_1 = 13$, $t_2 = 23$ and $T = t_2 - t_1 + 1 = 11$ (this is the same set up as for the bias estimator defined in Section 2.2).

The above data assimilation procedure is local as it is only used to update grid and receptor concentrations in the EPISODE model at midnight (between the first and second day of the 48-hour forecasting period). We do no extrapolations to other time points (hours), unlike what is done in the bias correction procedure.

2.4 Results

In this section we show the results of applying the bias correction and data assimilation methods described in Section 2 using observations and EPISODE model 2 day (48 hour) forecasting data from Bedre Byluft in Oslo for the week 2 – 8 December 2013 (Monday - Sunday). In this report, we focus on the improvement of model concentrations during the first 24 hours of each 48-hour forecasting period during this week.

This period was chosen as it represents a period of (unforeseen) stagnant meteorological conditions, with lower observed wind speed as compared to model simulations, especially during the first two days of the period (2 – 3 December). This results in too low modelled concentrations compared with observations. It is particularly during such periods that the bias correction and data assimilation methods are important to use to bring modelled concentrations in line with observations.

Results of improvements in the EPISODE model concentrations as compared with observations at stations in Oslo are shown separately for each of the three species NO_2, PM_{10} and $PM_{2.5}$ in Sections 2.4.1 - 2.4.3, respectively.

2.4.1 NO_2

Figure 1 shows maps of the EPISODE model ground level grid concentrations of NO_2 in Oslo on 2 December 2013 at 24h (midnight), i.e., after 24 hours of the model run for the 2 day (48 hours) forecasting period 2–3 December 2013. This is used as the initial concentrations in the EPISODE model for the next 2 day (48 hours) forecasting period starting at the same time point (hour), i.e., 3 December 2013 at 0h.

The upper left map shows the original grid concentrations before we apply bias correction and data assimilation, while the upper right map shows the grid concentrations after applying the bias correction. The middle right map shows the difference between these two concentration fields. As can be seen this results in a fairly large increase in the grid concentrations over central parts of the city, which are the result of interpolating estimated model biases at the various stations in Oslo during this day (from 13h – 23h).

The middle left map shows the grid concentrations after we apply both bias correction and data assimilation, while the bottom left map shows the impact of the data assimilation procedure in this case. As can be seen, applying data assimilation results in (only) a slight further increase in the concentrations. The bottom right map shows the difference between the bias corrected and assimilated field and the original field, which is very similar to the difference between the bias corrected field and the original field.

Table 1 shows original and modified (after applying bias correction and data assimilation) root mean square errors (RMSE) between observed and EPISODE model receptor concentrations of NO_2 at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013.

As can be seen from Table 1, bias correction works reasonably well overall, resulting in improvements in the RMSE both in absolute terms, and in percentage terms, at all stations in Oslo, except at Kirkeveien and Smestad, where the bias correction leads to a higher RMSE. The reason for the higher RMSE at Smestad after bias correction is that on some days (especially 7 December) observed concentrations are much higher than modelled concentrations during most of the day, most likely due to too strong wind in the model at this station. This leads to a high positive estimated bias, while closer to midnight, the model values are more in line with observations. This leads to an increase in model bias after correction, rather than a decrease, and thus to a higher overall RMSE at this station. Use of data assimilation, in addition to bias correction, however, generally leads to significant improvements in the RMSE metric, except at the stations Hjortnes and RV4 Aker sykehus, where there is only a slight increase.

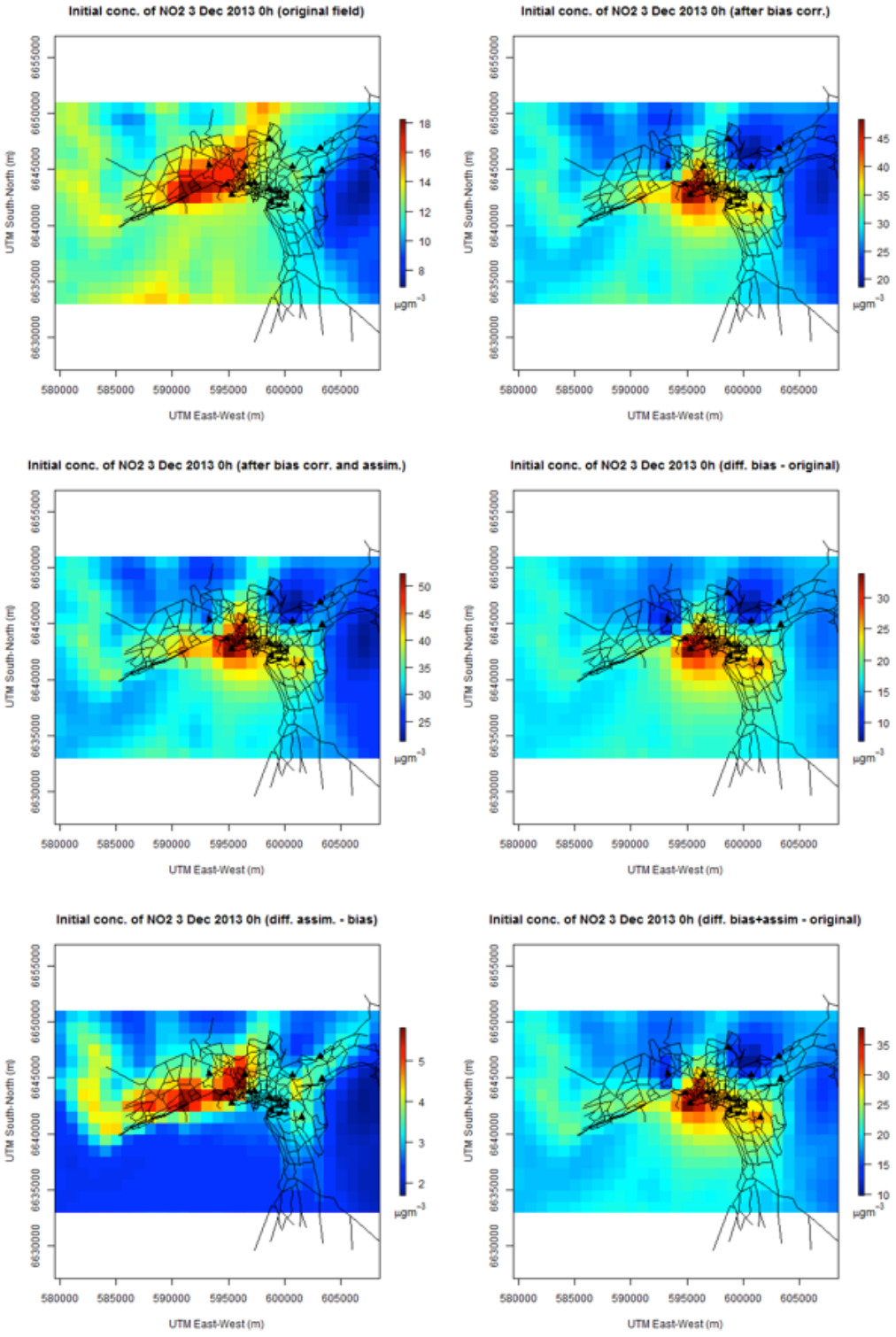


Figure 1. Maps of EPISODE model initial concentrations of NO_2 in Oslo (ground level) on 3 December 2013 at 0h (midnight). Units: μgm^{-3}. Upper left: uncorrected; Upper right: after bias correction; Middle left: after bias correction and data assimilation; Middle right: difference between bias correction and uncorrected; Bottom left: difference between data assimilation and bias correction; Bottom right: difference between bias correction + data assimilation concentration fields and uncorrected concentration fields.

Table 1. Original and modified root mean square errors (RMSE) between observed and EPISODE model concentrations of NO_2 at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013 (Mon. – Sun.). Negative values in the improvement column (identified by "Impr.") indicate the procedure worsens the field representation.

Station	Original model predictions	After bias correction		After bias correction and data assimilation	
	RMSE μgm^3	RMSE μgm^3	Impr. %	RMSE μgm^3	Impr. %
Alnabru	90.8	37.2	59.0	34.0	62.6
Bygdøy Alle	92.3	67.1	27.3	41.4	55.2
Grønland	100.9	56.0	44.5	52.3	48.2
Hjortnes	77.9	55.6	28.6	62.4	19.9
Kirkeveien	52.8	57.3	-8.6	35.6	32.6
Manglerud	75.7	42.2	44.3	36.6	51.7
RV4 Aker sykehus	57.6	39.2	31.9	39.6	31.2
Smestad	65.1	90.6	-39.2	73.8	-13.5
Sofienbergparken	-	-	-	-	-
Åkebergveien	87.1	45.4	47.9	38.1	56.3

Table 2 shows original and modified correlations between observed and EPISODE model receptor concentrations of NO_2 at stations in Oslo, based on values at 24h (midnight) during the week.

As can be seen from Table 2, bias correction works reasonably well overall, resulting in improvements in correlation at all stations in Oslo. Use of data assimilation, in addition to bias correction, generally leads to further improvements in correlation, except at station Hjortnes. A reduction in correlation due to data assimilation at a few stations is something that must be expected statistically since the calculated correlations are based on only 7 values (at midnight each day). It is more significant that we see an overall improvement in correlations.

As an example of the overall improvement in EPISODE model receptor concentrations after the combination of bias correction (with extrapolation of the estimated bias to all hours during the first day of forecasting) and data assimilation, we show results for NO_2 at station Bygdøy Allé in Figure 2. The time series in this figure compare the observed (blue curve) and EPISODE model receptor concentrations (red curve) for the week 2 – 8 December 2013.

In Figure 2, we show in the top panel observations and original model concentrations without corrections, while observations and model concentrations after bias correction and data assimilation are shown in the bottom panel. As can be seen from the figure, the modified model values (after bias correction and data

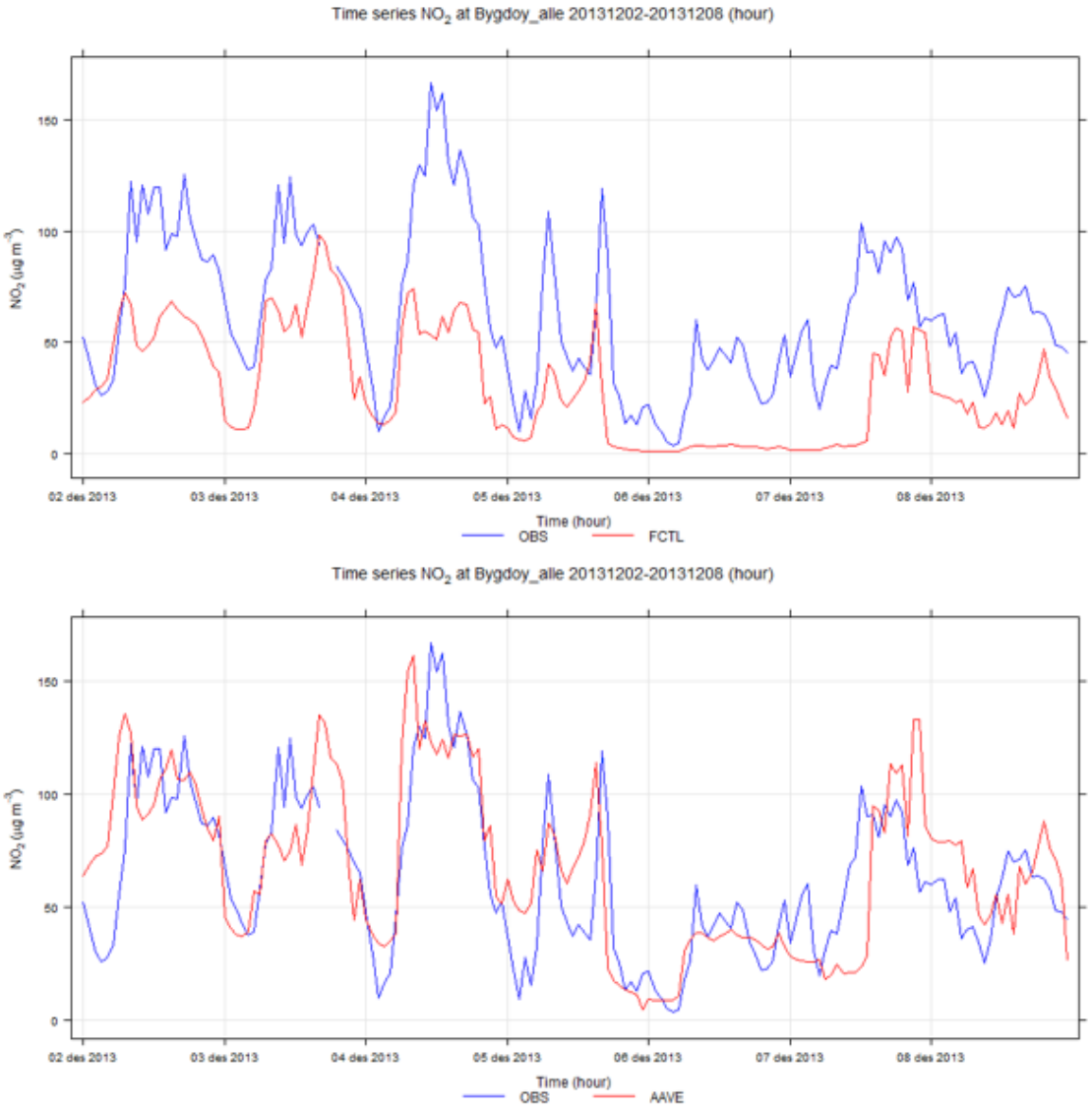


Figure 2. Time series plots of observed (blue) and EPISODE model concentrations (red) of NO_2 at Bygdøy Allé for the week 2 - 8 December 2013 (Mon. - Sun.). Original model values without corrections (top panel), and after bias correction and data assimilation (bottom panel). Units: $\mu g m^{-3}$.

Table 2. Original and modified correlations between observed and EPISODE model concentrations of NO_2 at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013 (Mon. – Sun.).

Stations	Original model predictions	After bias correction	After bias correction and data assimilation
	Correlation	Correlation	Correlation
Alnabru	0.89	0.96	0.97
Bygdøy Alle	0.69	0.72	0.91
Grønland	0.71	0.81	0.92
Hjortnes	0.50	0.90	0.80
Kirkeveien	0.72	0.77	0.90
Manglerud	0.30	0.71	0.79
RV4 Aker sykehus	0.75	0.87	0.91
Smestad	0.28	0.29	0.43
Sofienbergparken	-	-	-
Åkebergveien	0.61	0.82	0.98

assimilation) are overall much closer to the observed values. Similar figures for the other stations in Oslo, for this and the other two species, are provided in Appendices A and B, for the uncorrected and corrected model values, respectively.

2.4.2 PM_{10}

In this section, similar plots and tables of results as for NO_2 in Section 2.4.1, are given for PM_{10}.

Figure 3 show maps of the EPISODE model ground level grid concentrations of PM_{10} in Oslo on 2 December 2013 at 24h (midnight), i.e., after 24 hours of model run for the 2 day (48 hours) forecasting period 2 – 3 December 2013.

In Table 3 we show the original and modified (after applying bias correction and data assimilation) root mean square errors (RMSE) between observed and EPISODE model receptor concentrations of PM_{10} at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013.

As can be seen from Table 3, bias correction works reasonably well overall, resulting in good improvements in the RMSE both in absolute terms, and in percentage terms, at all stations in Oslo, except at Bygdøy Allé and Smestad, where the bias correction leads to a higher RMSE. Again, the reason for the much higher RMSE after bias correction at these stations is that on some days (especially 7 December) observed concentrations are much higher than modelled except close to midnight. Again, this is most likely due to too strong winds in the model at this station, leading to an increase in model bias after correction, and thus to a higher

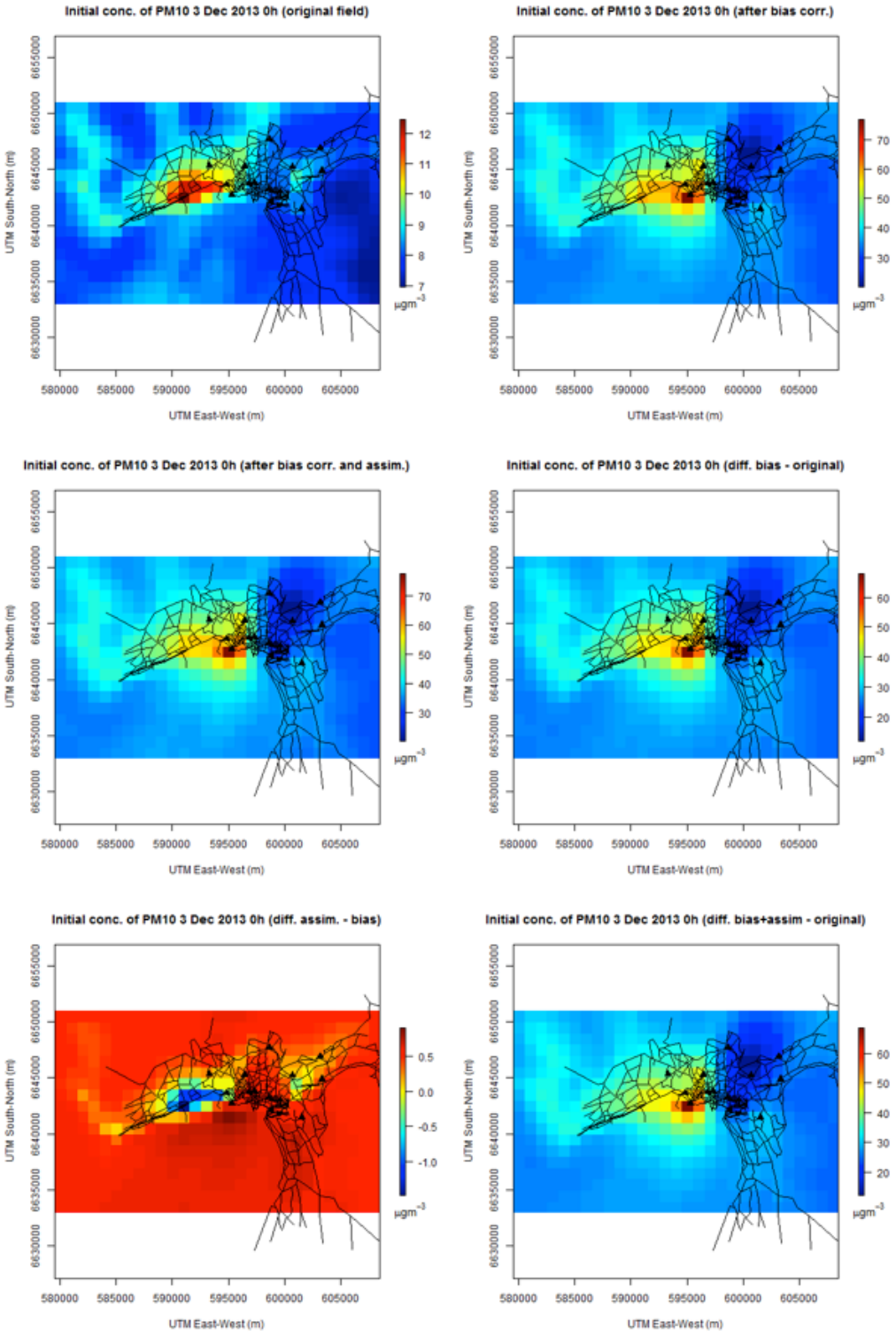


Figure 3. Maps of EPISODE model initial concentrations of PM10 in Oslo (ground level) on 3 December 2013 at 0h (midnight). Units: µgm-3. Upper left: uncorrected; Upper right: after bias correction; Middle left: after bias correction and data assimilation; Middle right: difference between bias correction and uncorrected; Bottom left: difference between data assimilation and bias correction; Bottom right: difference between bias correction + data assimilation concentration fields and uncorrected concentration fields.

Table 3. Original and modified root mean square errors (RMSE) between observed and EPISODE model concentrations of PM_{10} at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013 (Mon. – Sun.). Negative values in the improvement column indicate the procedure worsens the field representation.

Station	Original model predictions	After bias correction		After bias correction and data assimilation	
	RMSE μgm^3	RMSE μgm^3	Impr. %	RMSE μgm^3	Impr. %
Alnabru	93.8	46.8	50.1	32.9	65.0
Bygdøy Alle	34.7	61.5	-77.5	45.6	-31.5
Grønland	-	-	-	-	-
Hjortnes	127.6	57.7	54.8	57.7	54.8
Kirkeveien	82.8	20.4	75.3	22.2	73.2
Manglerud	44.4	24.9	43.9	25.3	43.2
RV4 Aker sykehus	45.4	21.5	52.6	16.2	64.3
Smestad	19.5	67.1	-244.8	49.4	-154.1
Sofienbergparken	104.2	76.5	26.6	74.3	28.7
Åkebergveien	39.1	20.9	46.6	13.8	64.6

overall RMSE. Use of data assimilation, in addition to bias correction, however, again leads to overall improvements in the RMSE metric, except at the stations Kirkeveien and Manglerud, where there is a slight increase.

Table 4 shows original and modified correlations between observed and EPISODE model receptor concentrations of PM_{10} at stations in Oslo, again based on values at 24h (midnight) during the week.

As can be seen from Table 4, bias correction works reasonably well overall, resulting in improvements in correlation at all stations in Oslo, except at Bygdøy Allé and Åkebergveien. Use of data assimilation, in addition to bias correction, generally leads to further improvements in correlation, except at station Bygdøy Allé. Again, a reduction in correlation due to bias correction or data assimilation at a few stations is something that must be expected statistically since the calculated correlations are based on only 7 values (at midnight each day). More significant is that we see an overall improvement in correlations.

We show in Figure 4 time series plots of observed (blue curve) and EPISODE model receptor concentrations (red curve) of PM_{10} at station Hjortnes for the week 2 – 8 December. As can be seen from the figure, the corrected model values are overall much closer to the observed values. Similar figures for the other stations in Oslo, for this and the other two species, are provided in Appendices A and B, for the uncorrected and corrected model values, respectively.

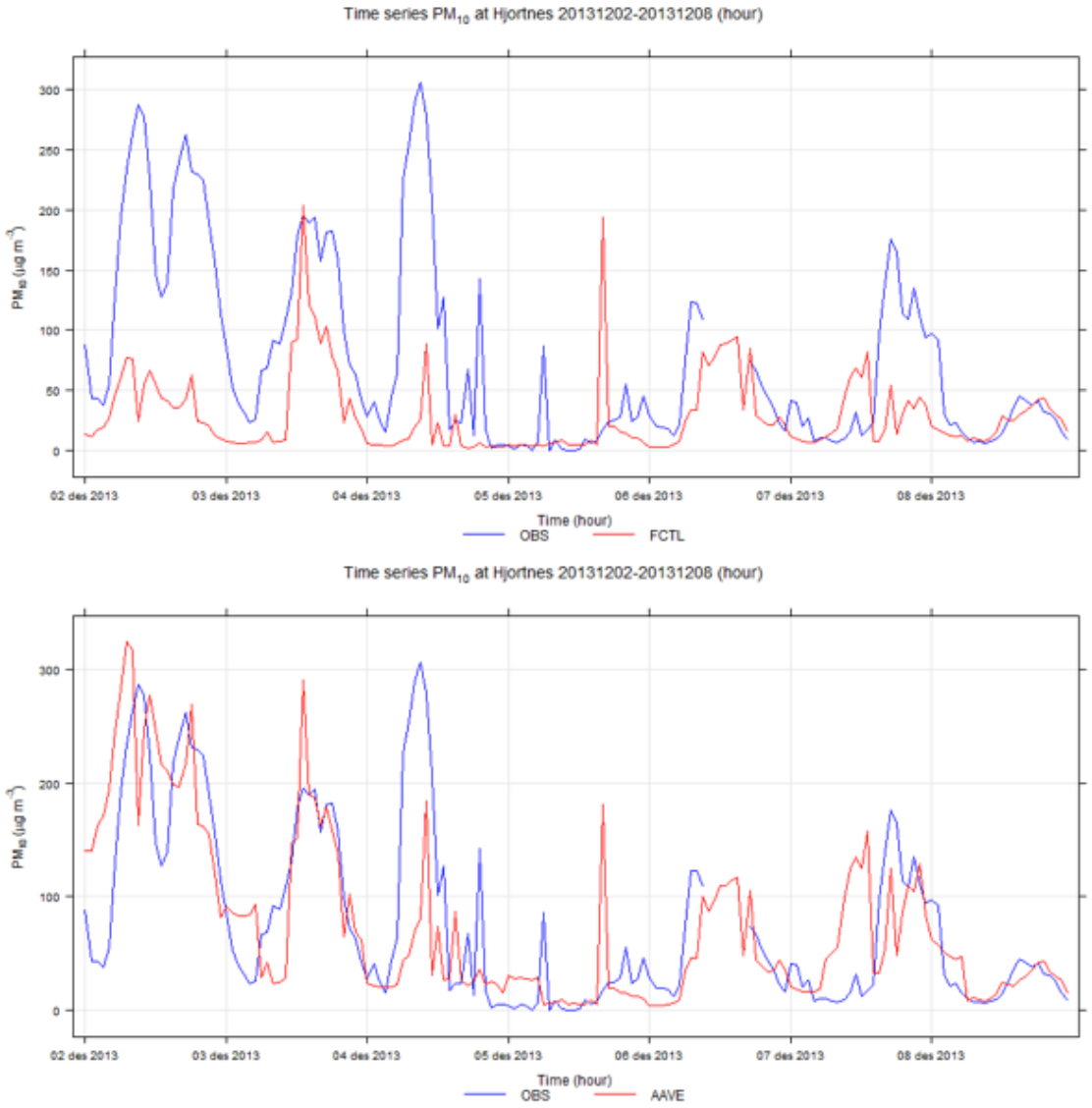


Figure 4. Time series plot of observed (blue curve) and EPISODE model concentrations (red curve) of PM_{10} at Hjortnes for the week 2 - 8 December 2013 (Mon. - Sun.). Original model values without corrections (top panel), and after bias correction and data assimilation (bottom panel). Units: $\mu g m^{-3}$.

Table 4. Original and modified correlations between observed and EPISODE model concentrations of PM_{10} at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013 (Mon. – Sun.).

Stations	Original model predictions	After bias correction	After bias correction and data assimilation
	Correlation	Correlation	Correlation
Alnabru	0.81	0.88	0.97
Bygdøy Alle	0.91	0.88	0.83
Grønland	-	-	-
Hjortnes	0.34	0.84	0.84
Kirkeveien	0.64	0.96	0.98
Manglerud	-0.11	0.66	0.67
RV4 Aker sykehus	0.64	0.92	0.95
Smestad	0.73	0.82	0.84
Sofienbergparken	0.67	0.82	0.95
Åkebergveien	0.99	0.86	0.98

2.4.3 $PM_{2.5}$

In this section, similar plots and tables of results as given for NO_2 and PM_{10} in Sections 2.4.1-2.4.2, are given for $PM_{2.5}$.

Figure 5 shows maps of the EPISODE model ground level grid concentrations of $PM_{2.5}$ in Oslo on 2 December 2013 at 24h (midnight), i.e., after 24 hour of the model run for the 2 days (48 hours) forecasting period 2 – 3 December 2013.

Again, there is an increase in grid concentrations over the central parts of the city. Applying data assimilation results in a further increase in the concentrations. The difference between the bias corrected and assimilated field and the original field, is now bigger than the difference between the bias corrected and original fields.

Table 5 shows the original and modified (after applying bias correction and data assimilation) root mean square error (RMSE) between observed and EPISODE model receptor concentrations of $PM_{2.5}$ at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013.

As can be seen from Table 5, bias correction works reasonably well overall, resulting in improvements in the RMSE both in absolute and percentage terms, at all stations in Oslo, except at Smestad, where the bias correction leads to a higher RMSE. Again, the reason for the higher RMSE after bias correction at this station, is because at some days (especially on 4 December and 7 December) observed values are much higher than modelled during most of the day. Again this most likely due to too strong winds in the model at this station, leading to an increase in model bias after correction at midnight. Use of data assimilation, in addition

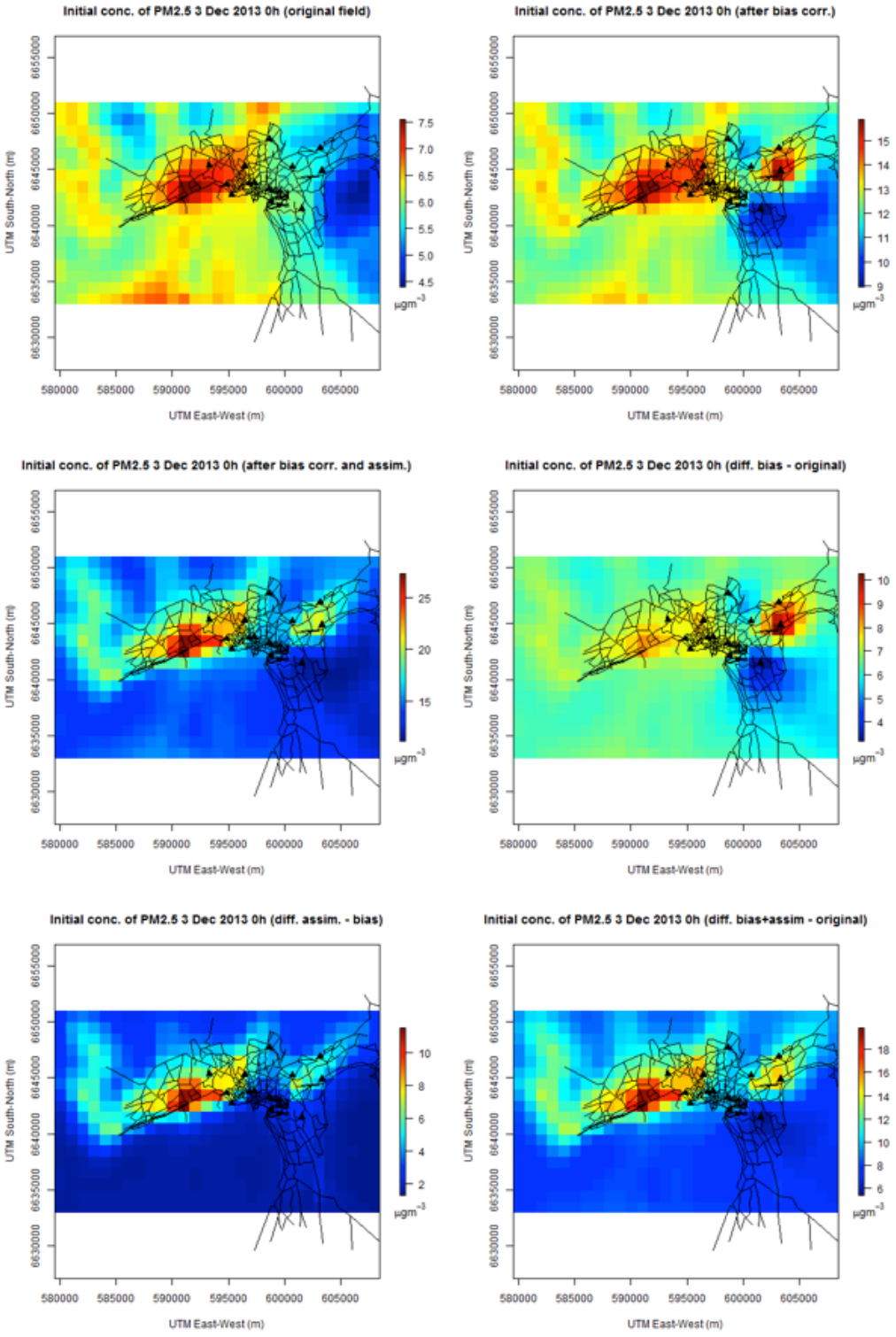


Figure 5. Maps of EPISODE model initial concentrations of $PM_{2.5}$ in Oslo (ground level) on 3 December 2013 at 0h (midnight). Units: µgm-3. Upper left: uncorrected; Upper right: after bias correction; Middle left: after bias correction and data assimilation; Middle right: difference between bias correction and uncorrected; Bottom left: difference between data assimilation and bias correction; Bottom right: difference between bias correction + data assimilation concentration field and uncorrected concentration field.

Table 5. Original and modified root mean square errors (RMSE) between observed and EPISODE model concentrations of $PM_{2.5}$ at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December (Mon. – Sun.). Negative values in the improvement column indicate the procedure worsens the field representation.

Station	Original model predictions	After bias correction		After bias correction and data assimilation	
	RMSE μgm^3	RMSE μgm^3	Impr. %	RMSE μgm^3	Impr. %
Alnabru	47.2	18.4	61.0	10.1	77.9
Bygdøy Alle	37.9	26.0	31.4	18.3	51.8
Grønland	-	-	-	-	-
Hjortnes	22.6	16.9	25.3	15.1	33.4
Kirkeveien	38.1	24.1	36.7	16.1	57.7
Manglerud	13.1	5.8	55.7	6.9	47.4
RV4 Aker sykehus	32.8	19.4	40.9	14.5	55.8
Smestad	7.1	19.6	-174.7	21.6	-202.8
Sofienbergparken	56.4	41.6	26.3	36.7	34.9
Åkebergveien	33.7	21.6	36.0	16.7	50.5

to bias correction, however, again leads to overall improvements in the RMSE metric, except at the stations Manglerud and Smestad, where there is a slight increase.

Table 6 shows original and modified correlations between observed and EPISODE model receptor concentrations of $PM_{2.5}$ at stations in Oslo, again based on values at 24h (midnight) during the week.

As can be seen from Table 6, bias correction works reasonably well overall, resulting in improvements in correlation at all stations in Oslo, except at Smestad. Use of data assimilation, in addition to bias correction, generally leads to further improvements in correlation, except at station Smestad. Again, since the data set is quite small (only 7 values), reduced correlations at a few stations after bias correction or data assimilation is something that must be expected statistically. More significant is the overall improvement in correlations.

We show in Figure 6 time series plots of observed (blue curve) and EPISODE model receptor concentrations (red curve) of $PM_{2.5}$ at station Kirkeveien for the week 2 – 8 December 2013. As can be seen from the figure, the corrected model values are overall much closer to the observed values. Similar figures for the other stations in Oslo, for this and the other two species, are provided in Appendices A and B, for the uncorrected and corrected model values, respectively.

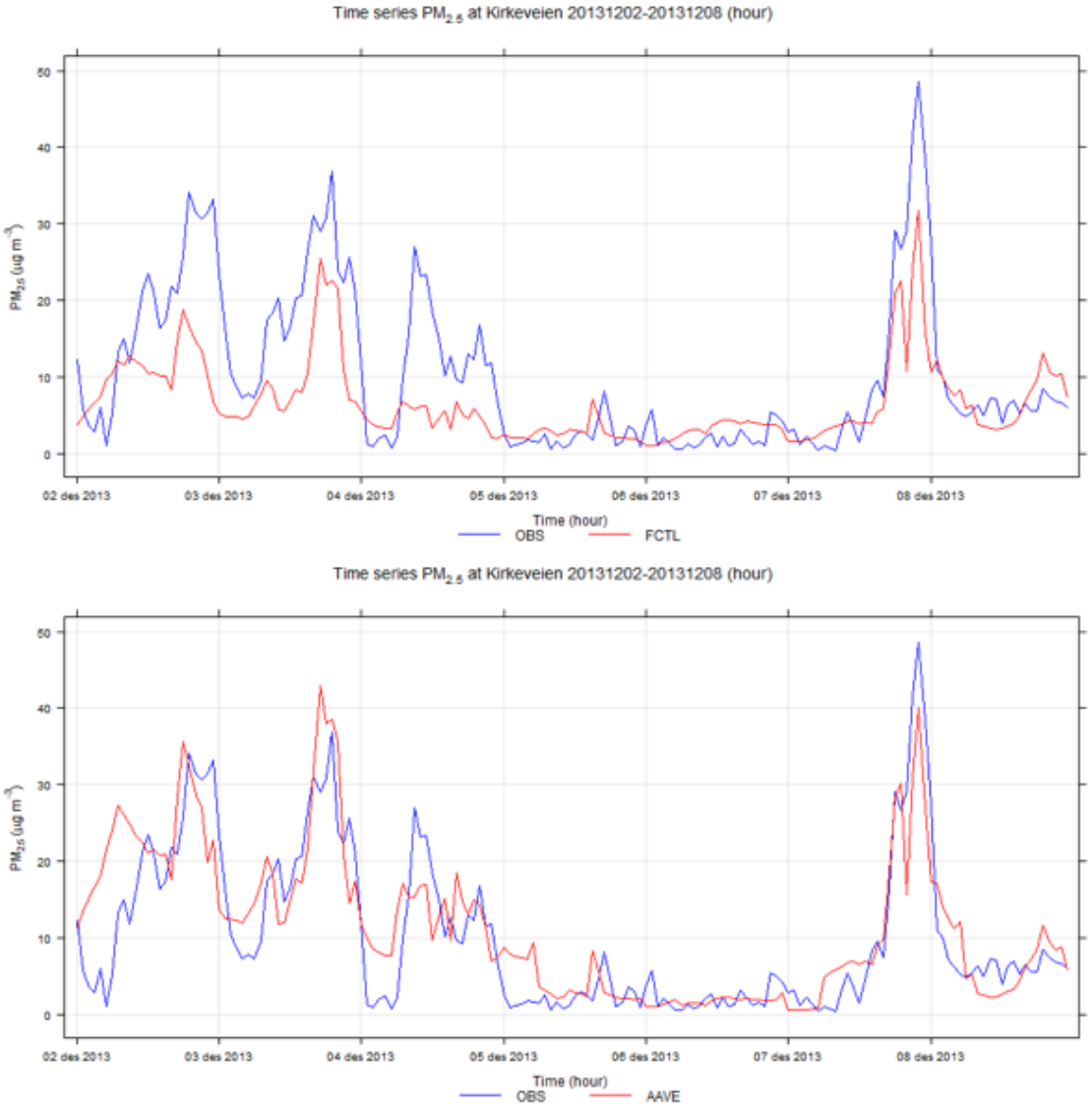


Figure 6. Time series plot of EPISODE model concentrations of $PM_{2.5}$ at Kirkeveien for the week 2 - 8 December 2013 (Mon. – Sun.). Original model values without corrections (top panel), and after bias correction and data assimilation (bottom panel). Units: μgm^{-3}.

Table 6. Original and modified correlations between observed and EPISODE model concentrations of $PM_{2.5}$ at stations in Oslo based on values at 24h (midnight) during the week 2 – 8 December 2013 (Mon. – Sun.).

Stations	Original model predictions	After bias correction	After bias correction and data assimilation
	Correlation	Correlation	Correlation
Alnabru	0.80	0.95	0.97
Bygdøy Alle	0.72	0.81	0.90
Grønland	-	-	-
Hjortnes	0.71	0.80	0.88
Kirkeveien	0.82	0.95	0.99
Manglerud	0.96	0.98	0.98
RV4 Aker sykehus	0.85	0.94	0.97
Smestad	0.93	0.76	0.74
Sofienbergparken	0.72	0.94	0.98
Åkebergveien	≈1.00	≈1.00	≈1.00

3 Statistical post-processing

NR has developed a prototype model for statistical post-processing of the air quality forecasts within AirQUIS (Steinbakk et al., 2013). This work was based on output data from the deterministic dispersion model EPISODE and observations from 11 stations in Oslo from the winter seasons 2011-2012 and 2012-2013. In general, the statistical post-processing significantly improved the predictive performance of the numerical model.

Other cities in Norway typically have only a few observation stations which makes post-processing difficult due to lack of data. To investigate the performance of the post-processing proposed by Steinbakk et al. (2013) in this situation, we apply a cross-validation scheme to the Oslo data in which observed data are assumed available at a few locations only within the region of interest. That is, we use a small subset of the available stations in Oslo as a training data set to learn the parameters of the statistical post-processing model and compare the adjusted results against observations at stations not used for the training. This procedure is then repeated for different groups of observation stations.

The air quality forecasts within AirQUIS used in this analysis are deterministic such that at each location, the forecast is given by a point value without associated uncertainty. The post-processing method of Steinbakk et al. (2013), on the other hand, provides full predictive distributions and thus includes procedures for appropriate uncertainty assessment as described below. The availability of the

Table 7. The measurement stations in Oslo.

Name	Type	Variables
Kirkeveien	Road	NO_2, PM_{10}, $PM_{2.5}$
Smestad	Road	NO_2, PM_{10}, $PM_{2.5}$
RV4	Road	NO_2, PM_{10}, $PM_{2.5}$
Bygdøy Allé	Road	NO_2, PM_{10}, $PM_{2.5}$
Alnabru	Road	NO_2, PM_{10}, $PM_{2.5}$
Hjortnes	Road	NO_2, PM_{10}, $PM_{2.5}$
Manglerud	Road	NO_2, PM_{10}, $PM_{2.5}$
Åkebergveien	Road, background	NO_2, PM_{10}, $PM_{2.5}$
Sofienbergparken	Background	PM_{10}, $PM_{2.5}$
Grønland	Background	NO_2

full predictive distributions is especially valuable for threshold exceedances, as it allows us to estimate the predictive probability of exceeding any given threshold.

3.1 Data

The data consist of forecasts from EPISODE and corresponding measurements of NO_2, $PM_{2.5}$, and PM_{10} at different measurement sites in Oslo for the winter seasons 2011-2012 and 2012-2013, see Table 7. For the first winter season, data are available from 15th of January to 13th of April 2012, while the data from the second season cover the period from 5th of October 2012 to 30th of April 2013. The coarse fraction PM_c is given by the difference $PM_{10} - PM_{2.5}$. In this work we focus on the 1–24 hour ahead predictions of EPISODE, but the statistical post-processing can also be applied to the 25–48 hour ahead predictions.

Negative observed values sometimes exist in the data set due to measurement error. Since our analysis is performed on a logarithmic scale, small negative observed values are truncated such that $PM2.5$ and PM_c have 0.5 as the minimum value, while 2 and 1 are used as minimum values for NO_2 and PM_{10}, respectively.

3.2 Model for statistical post-processing

Denote by $y_{j,t}$ the logarithm of the true concentration of an air pollutant located at a geographical point j on a grid G and a time point t. Further, let $\hat{y}_{j,t}$ be the corresponding deterministic prognosis on a logarithmic scale. The relationship between the true concentration and the prognosis can be written as

$$y_{j,t} = \beta_0 + \beta_1 \hat{y}_{j,t} + n_{j,t}, \quad (8)$$

where β_0 and β_1 are unknown parameters, and $n_{j,t}$ is the error term. The parameters β_0 and β_1 are intended to adjust for systematic biases between the true concentration and the deterministic predictions. The error term in Eq. (8) is a function of an underlying process l_t given by

$$n_{j,t} = l_t + \epsilon_{j,t},$$

with the error terms $\epsilon_{j,t}$ assumed independently normally distributed with mean zero and variance σ_t^2. The underlying process l_t is common for all grid points and is assumed to follow an autoregressive process

$$l_t = \phi_1 l_{t-1} + \phi_{24} l_{t-24} - \phi_1 \phi_{24} l_{t-25} + a_t. \quad (9)$$

Here, the error a_t is an independent Gaussian process with mean zero and standard deviation τ_t, and ϕ_1 and ϕ_{24} are unknown parameters. The underlying error l_t is thus a function of the error one hour ahead and the error 24 and 25 hours earlier, where the parameters ϕ_1, ϕ_{24}, and $\phi_1\phi_{24}$ indicate the degree of dependence on past errors. This model form is based on that of a multiplicative seasonal autoregressive model (Box and Jenkins, 1976).

3.3 Parameter estimation

The model in Eq. (8) for the statistical post-processing involves unknown parameters and processes that are estimated from the observations and the deterministic prognoses at m measurement stations located on a subset of the grid G. At a given time point, say T, we fit the statistical model to historical data up to time T. Then, we use the estimated statistical model to compute the predictive distributions 1-48 hours ahead.

The estimation of the unknown model parameters is performed in two steps. In a first step, the regression parameters are estimated by ordinary least square estimation using the observations and the corresponding prognoses at all the m measurement stations. Secondly, we use the resulting estimates in the model in Eq. (8) to estimate the terms in the error process $n_{t,j}$. As we have computed the error $\hat{n}_{t,j}$ at each observation site, we compute the common underlying error process at each time point as

$$\hat{l}_t = \frac{1}{m} \sum_{j=1}^{m} \hat{n}_{t,j}.$$

Finally, we estimate the unknown parameters ϕ_1 and ϕ_{24} in equation Eq. (9) using $\hat{l}_t$ as the error process (using standard statistical software, for instance the function `"Arima"` in the statistical software `R`).

3.4 Probabilistic forecasts

Our predictive distribution on a logarithmic scale is normal with mean $\hat{\mu}_{j,t|T}$ and variance $\widehat{sd}^2_{j,t|T}$, see Appendix C for details. Here, the sub-script $j, t|T$ indicates a predicted value at a grid point j and a time point t given information up to time point T. A random variable from the predictive distribution for an air pollutant on the original scale is thus given by

$$Y_{j,t|T} = \exp(\hat{\mu}_{j,t|T} + \widehat{sd}_{j,t|T} Z_{t,j}), \tag{10}$$

where $Z_{t,j}$ is a standard normal variable. Given the estimates $\hat{\mu}_{j,t|T}$ and $\widehat{sd}_{j,t|T}$, it is straightforward to obtain a sample of any size from the predictive distribution by first sampling standard normal variates and then transforming them according to Eq. (10). Two examples of such distributions are shown in Figure 7, illustrating that the predictive distribution of an air pollutant on the original scale is non-symmetric with a heavy, right tail. An algorithm for estimating the unknown parameters in the predictive distribution in Eq. (10) and simulating its distribution, is given in the Appendix C.

Different quantities and estimates can be computed from the full probabilistic distribution for an air pollution component, such as its mean, quantiles and standard deviation. Following Steinbakk et al. (2013) we use the mean of the full predictive distribution as our adjusted predictions. In addition, we investigate the predictive performance of the median estimate. Furthermore, a full probabilistic distribution can provide exceedance probabilities as demonstrated in Section 4.

3.5 Evaluation of predictive performance

The deterministic prognosis system in AirQUIS provides daily prognosis about six o'clock in the morning during the winter. This time point is denoted by T. We evaluate the post-processed predictions by estimating the model on (historical) data for a time period up to time T and compare the post-processed predictions to actual observations for $t = T + 1h, \ldots, T + 48h$. Thus, we use data from $t = 1, \ldots, T$ to train the model (training data set) and assess how the predictions fit to real data up to 48 hours ahead (validation data set). This procedure is repeated daily, resulting in a longer training set as the season progresses. To have a sufficiently long training set to estimate the model at the beginning of the season, the validation starts after 14 days (i.e. $T = 14 \cdot 24$ hours) each season.

We apply a cross-validation scheme by assuming available data at a few locations only. Thus, the model is trained or estimated on a small subset of the available stations in Oslo, and then we compare these post-processed results against observations at stations not used for training. The scheme is repeated for groups of air quality stations.

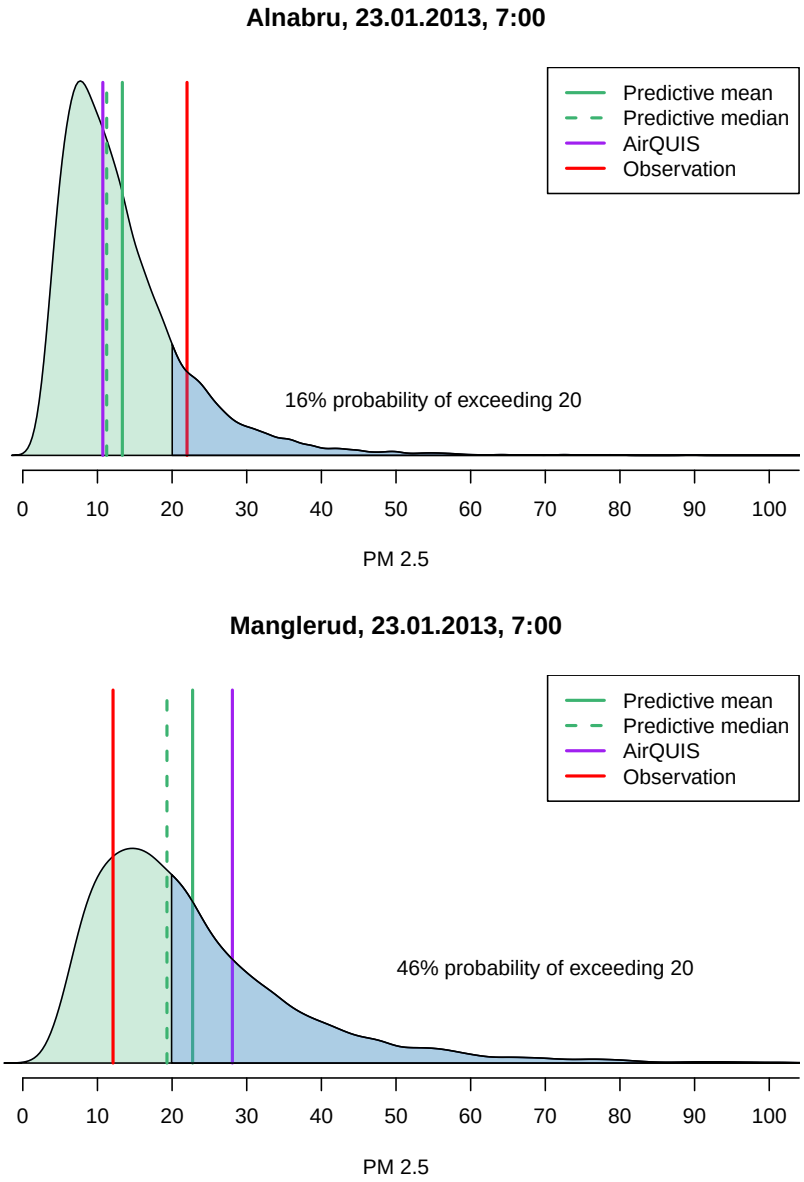


Figure 7. Two examples of full predictive distributions of hourly $PM_{2.5}$ concentrations using the statistical post-processing framework of Steinbakk et al. (2013). The full distributions are indicated in green with the part exceeding a threshold of 20μm^{-3} indicated in blue.

The following descriptive measures are used to compare the prognosis $\hat{y}_{j,t}$ and the observations $y_{j,t}^{\text{obs}}$, $t = 1, \ldots, n$ at each measurement station j:

- Square root of the mean squared error

$$\text{RMSE}_j = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_{j,t}^{\text{obs}} - \hat{y}_{j,t})^2}.$$

- Correlation coefficient between the observed values and the prognoses:

$$\text{COR}_j = \frac{1}{n-1}\sum_{t=1}^{n}\frac{\left(y_{j,t}^{\text{obs}} - \bar{y}_{j,\cdot}^{\text{obs}}\right)\left(\hat{y}_{j,t} - \bar{\hat{y}}_{j,\cdot}\right)}{s_j^{\text{obs}} s_j},$$

  where $\bar{y}_{j,\cdot}^{\text{obs}}$ and $\bar{\hat{y}}_{j,\cdot}$ are the means of all observations and prognosis, respectively, over all time points n, while s_j^{obs} and s_j are the corresponding standard deviations.

- Mean absolute error

$$\text{MAE}_j = \frac{1}{n}\sum_{t=1}^{n}|y_{j,t}^{\text{obs}} - \hat{y}_{j,t}|.$$

Following Gneiting (2011), we use the predictive mean of the post-processed forecast when the predictive performance is measured by the RMSE and the predictive median under the MAE. The correlation coefficients are calculated using both the predictive mean and the median.

3.6 Results

Here, we show the results from the statistical post-processing based on only a small sub-group of the measurement stations in Table 7, using the cross-validation scheme described in Section 3.5. Results from the statistical post-processsing method based on all measurement stations in Oslo are presented in Steinbakk et al. (2013).

We show the performance of the statistical post-processing at the stations RV4, Åkebergveien and Hjortnes, where Åkebergveien is almost regarded as a background station due to low traffic (annual average daily traffic is about 7000). These three target stations are left out when estimating the models based on sub-groups of neighbouring sites different from the target stations. The adjusted predictions at the target stations are then compared to their corresponding observations. The two groups of neighbouring stations that are used for training the model are:

Group 1 Sofienbergparken, Kirkeveien, Manglerud

Group 2 Smestad, Bygdøy Allé, Alnabru

Further results where the parameter estimation is based on data from only two stations are given in Appendix D.

Hence, we estimate the model based on data from one of these two groups and validate the 24 hours ahead predictions and the observations at the target stations, with the results summarised in Table 8. The values in the parentheses are the percentage improvements compared to the same measure computed with the original prognoses. The results for group 1 and group 2 are also compared to the results using all stations (see the column "All st."). The adjusted predictions are here defined as the mean of the predictive distribution.

The adjusted root mean squared forecast errors for the particulate matters, PM, are always better than (or at least as good as) the original predictions except for $PM_{2.5}$ at RV4 in season 2 based on group 2. In this case, the estimated variances were higher in season 2 than in season 1. The mean estimate is a function of the standard deviations on a logarithmic scale, resulting in a greater difference at the original scale. A more robust estimate in this case might be the median estimate discussed below, since the median prediction is not so sensitive to outliers as the mean.

Figure 8 shows a histogram of the RMSE values and the correlations between predictions and observations based on all possible unique combinations of groups of three stations (training data sets) for $PM_{2.5}$ at RV4 (all together 54 combinations). The red vertical line indicates the same measure based on the original prognosis. The correlation is almost always better than the original prognosis for all groups of stations and the same holds for RMSE in season 1. We also see that the RMSE for group 2 in season 2 are amongst the higher ones of all the combinations of groups of validation data set in Figure 8, but that the adjusted prognoses are better on average. In this context, we should also mention that the improvements for $PM_{2.5}$ in season 2 at RV4 based on all the other stations was amongst the smallest of all the target stations (i.e., 5% in Steinbakk et al. (2013)).

The RMSE for the adjusted predictions of NO_2 in Table 8 shows that RMSE in season 2 is better than in season 1. The adjusted prognosis based on all the other measurement stations for these three target stations had higher RMSE than the original prognosis in season 1. The correlations, on the other hand, show improvements compared to the original prognosis similarly to the results based on all stations, which may indicate that the level of air pollution concentration is not very well fitted.

Note that in Steinbakk et al. (2013) the RMSE for the original prognosis for $PM_{2.5}$ at Åkebergveien in season 1 was 7.6 (see Table A.3 in Steinbakk et al. (2013)) which is much lower than the value 9.2 in Table 8 of the current report. The reason

Table 8. Root mean squared error (RMSE) and correlation (COR) for prognosis of PM_{10}, $PM_{2.5}$, PM_c and NO_2 for the post-processed prognosis (mean value estimate) and based on different groups of neighbouring stations and the original prognosis from AirQUIS. The numbers in the parentheses indicate improvement (in %) compared to the original prognoses.

PM_{10}	RMSE				COR			
Season 1	Orig.	Gr. 1	Gr. 2	All st.	Orig.	Gr. 1	Gr. 2	All st.
RV4	23.0	18.8 (18%)	18.9 (18%)	18.8 (18%)	0.59	0.67 (12%)	0.66 (11%)	0.67 (13%)
Åkeberg.	25.9	23.8 (8%)	23.1 (11%)	23.0 (11%)	0.49	0.53 (9%)	0.54 (11%)	0.56 (15%)
Hjortnes	59.0	32.5 (45%)	31.9 (46%)	32.3 (45%)	0.29	0.36 (24%)	0.36 (25%)	0.36 (25%)
PM_{10} Season 2								
RV4	20.3	18.1 (11%)	18.4 (10%)	18.0 (11%)	0.52	0.61 (16%)	0.61 (17%)	0.62 (19%)
Åkeberg.	19.5	16.4 (16%)	16.5 (16%)	16.0 (18%)	0.47	0.64 (35%)	0.64 (35%)	0.66 (39%)
Hjortnes	39.5	28.3 (28%)	27.4 (31%)	28.3 (28%)	0.29	0.43 (45%)	0.44 (49%)	0.44 (50%)
$PM_{2.5}$ Season 1								
RV4	7.9	5.4 (31%)	7.0 (10%)	5.5 (30%)	0.57	0.63 (11%)	0.57 (1%)	0.63 (11%)
Åkeberg.	9.2	8.5 (8%)	8.7 (5%)	8.4 (9%)	0.57	0.65 (12%)	0.58 (1%)	0.65 (13%)
Hjortnes	11.5	6.6 (42%)	8.9 (22%)	7.0 (39%)	0.42	0.49 (18%)	0.42 (0%)	0.49 (17%)
$PM_{2.5}$ Season 2								
RV4	6.1	5.1 (16%)	7.6 (-25%)	5.8 (5%)	0.46	0.52 (13%)	0.47 (1%)	0.52 (13%)
Åkeberg.	8.9	8.2 (8%)	8.3 (8%)	7.9 (12%)	0.49	0.60 (22%)	0.55 (12%)	0.61 (25%)
Hjortnes	8.5	6.2 (28%)	8.5 (0%)	6.8 (20%)	0.42	0.48 (15%)	0.42 (0%)	0.47 (12%)
PM_c Season 1								
RV4	19.4	17.0 (12%)	16.8 (14%)	16.6 (15%)	0.61	0.68 (12%)	0.69 (14%)	0.70 (15%)
Åkeberg.	22.2	20.9 (6%)	20.6 (7%)	19.9 (11%)	0.48	0.53 (11%)	0.54 (13%)	0.57 (20%)
Hjortnes	51.7	33.7 (35%)	30.0 (42%)	31.7 (39%)	0.29	0.36 (23%)	0.38 (31%)	0.36 (25%)
PM_c Season 2								
RV4	18.2	16.6 (9%)	16.9 (7%)	16.4 (10%)	0.52	0.60 (16%)	0.60 (17%)	0.62 (20%)
Åkeberg.	15.3	12.9 (16%)	14.8 (3%)	12.6 (18%)	0.45	0.66 (46%)	0.63 (38%)	0.69 (52%)
Hjortnes	35.0	25.9 (26%)	24.3 (31%)	26.5 (24%)	0.27	0.41 (50%)	0.45 (65%)	0.41 (50%)
NO_2 Season 1								
RV4	27.9	31.0 (-11%)	32.7 (-17%)	28.8 (-3%)	0.59	0.60 (2%)	0.62 (5%)	0.62 (5%)
Åkeberg.	28.5	35.8 (-26%)	40.3 (-41%)	35.5 (-25%)	0.50	0.58 (16%)	0.56 (12%)	0.60 (19%)
Hjortnes	38.8	42.9 (-11%)	44.0 (-13%)	39.6 (-2%)	0.45	0.48 (7%)	0.50 (11%)	0.49 (8%)
NO_2 Season 2								
RV4	28.6	25.5 (11%)	27.3 (4%)	24.6 (14%)	0.63	0.66 (4%)	0.66 (5%)	0.68 (7%)
Åkeberg.	31.6	22.8 (28%)	33.2 (-5%)	27.3 (14%)	0.64	0.71 (11%)	0.66 (2%)	0.71 (10%)
Hjortnes	32.1	29.7 (8%)	29.7 (7%)	28.5 (11%)	0.63	0.65 (2%)	0.66 (5%)	0.67 (6%)

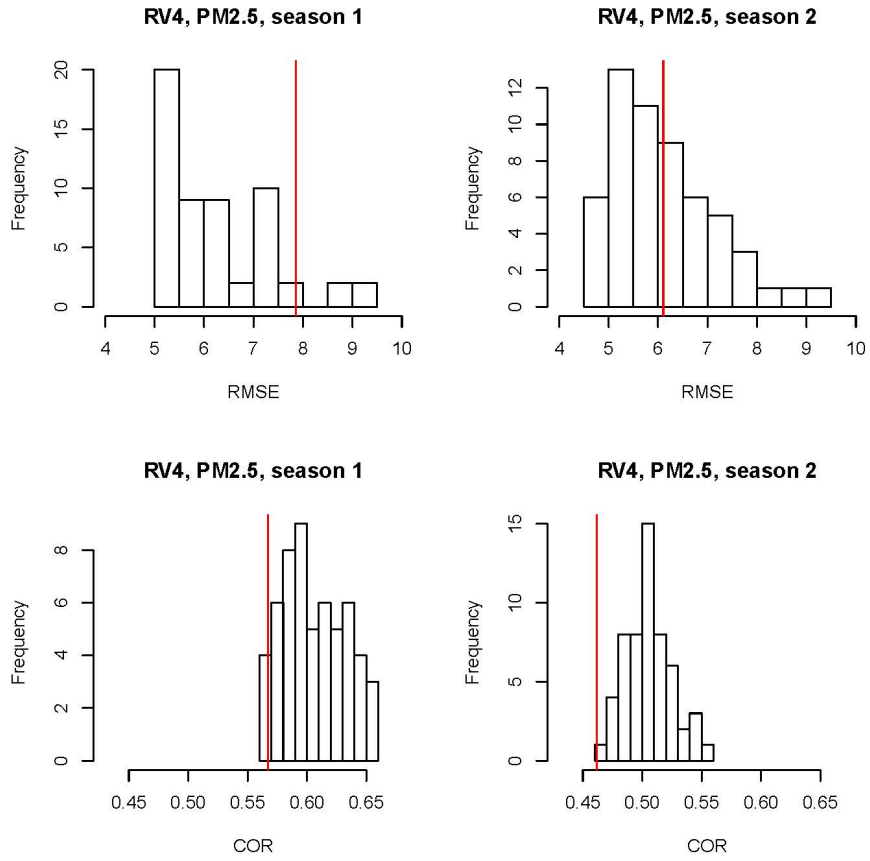


Figure 8. Histogram of root mean squared error (RMSE) (top row) and correlation (COR) (bottom row) between post-processed prognosis (mean value estimate) and data for $PM_{2.5}$ at RV4, for parameter estimation based on all possible combinations of groups of three stations. The red vertical line indicates the corresponding result for the original 24 hour prognosis.

is that the results in this report and in Steinbakk et al. (2013) are based on slightly different data sets. Steinbakk et al. (2013) presented an additional local method only useful for grid points that included measurement stations. For comparing the regional model presented in this report and the local model, Steinbakk et al. (2013) used exactly the same data points for both model frameworks, where the other local method could not predict for all time points.

An alternative measure is to use the median of the predictive distribution as our estimate rather than the mean. The median is a generally a more robust estimate as it is less sensitive to outliers than the mean. This might have an effect here as the predictive distributions are usually skewed with a heavy upper tail, see Figure 7.

The results for the median prognoses are given in Table 9. As mentioned above, the predictive performance is here measured by the MAE rather than the RMSE.

Table 9. Mean absolute error (MAE) and correlation (COR) for prognosis of PM_{10}, $PM_{2.5}$, PM_c and NO_2 for the post-processed median prognosis based on different groups of neighbouring stations and the original prognosis from AirQUIS. The numbers in the parentheses indicate improvement (in %) compared to the original prognoses.

PM_{10}	MAE			COR		
Season 1	Orig.	Gr. 1	Gr. 2	Orig.	Gr. 1	Gr. 2
RV4	23.0	20.0 (13%)	19.8 (14%)	0.59	0.67 (12%)	0.66 (11%)
Åkeberg.	25.9	25.3 (2%)	24.8 (4%)	0.49	0.54 (11%)	0.55 (12%)
Hjortnes	59.0	32.0 (46%)	31.2 (47%)	0.29	0.36 (25%)	0.36 (26%)
PM_{10} Season 2						
RV4	20.3	18.7 (8%)	18.5 (9%)	0.52	0.61 (17%)	0.61 (17%)
Åkeberg.	19.5	17.8 (9%)	17.4 (11%)	0.47	0.64 (35%)	0.64 (35%)
Hjortnes	39.5	29.3 (26%)	28.7 (27%)	0.29	0.43 (46%)	0.44 (49%)
$PM_{2.5}$ Season 1						
RV4	7.9	5.3 (33%)	5.3 (33%)	0.57	0.64 (12%)	0.62 (9%)
Åkeberg.	9.2	9.5 (-3%)	8.9 (3%)	0.57	0.65 (13%)	0.63 (10%)
Hjortnes	11.5	5.6 (51%)	6.2 (46%)	0.42	0.50 (19%)	0.47 (11%)
$PM_{2.5}$ Season 2						
RV4	6.1	4.6 (24%)	5.0 (18%)	0.46	0.53 (14%)	0.45 (-3%)
Åkeberg.	8.9	9.2 (-2%)	9.2 (-3%)	0.49	0.60 (22%)	0.53 (9%)
Hjortnes	8.5	6.1 (29%)	6.5 (24%)	0.42	0.49 (16%)	0.40 (-5%)
PM_c Season 1						
RV4	19.4	19.2 (1%)	20.9 (-8%)	0.61	0.68 (12%)	0.69 (13%)
Åkeberg.	22.2	22.6 (-1%)	24.1 (-8%)	0.48	0.53 (11%)	0.53 (12%)
Hjortnes	51.7	30.1 (42%)	30.4 (41%)	0.29	0.35 (23%)	0.38 (30%)
PM_c Season 2						
RV4	18.2	18.3 (0%)	18.9 (-4%)	0.52	0.59 (15%)	0.58 (13%)
Åkeberg.	15.3	14.2 (7%)	15.0 (2%)	0.45	0.66 (46%)	0.60 (33%)
Hjortnes	35.0	28.0 (20%)	28.0 (20%)	0.27	0.42 (52%)	0.44 (62%)
NO_2 Season 1						
RV4	27.9	26.8 (4%)	28.4 (-2%)	0.59	0.61 (4%)	0.63 (6%)
Åkeberg.	28.5	25.5 (10%)	32.7 (-15%)	0.50	0.59 (18%)	0.57 (13%)
Hjortnes	38.8	37.1 (5%)	38.9 (0%)	0.45	0.50 (10%)	0.51 (13%)
NO_2 Season 2						
RV4	28.6	26.0 (9%)	24.8 (13%)	0.63	0.66 (5%)	0.67 (5%)
Åkeberg.	31.6	17.3 (45%)	24.9 (21%)	0.64	0.71 (10%)	0.66 (2%)
Hjortnes	32.1	33.8 (-5%)	28.5 (11%)	0.63	0.65 (3%)	0.66 (5%)

The MAE is less sensitive to outlying prediction errors than the RMSE such that single instances in which a forecast performs poorly have less influence on the overall average performance than under the RMSE. However, we see that the correlation results based on the predictive median are equal to the correlation results based on the predictive mean in Table 8.

Overall, the relative performance of the median prediction is better than that of the mean prediction for NO_2 while the relative improvement compared to the original prognosis is better for the mean prediction for the particulate matters. For a potential operationalization of the method, it should thus be investigated further whether the mean or the median is the best measure more to determine the most likely forecast.

4 Communicating uncertainty

The air quality forecasts within AirQUIS may be associated with a varying degree of uncertainty due to, e.g., the meteorological conditions, the day of the week or the hour of the day, and hence are not complete without a description of their uncertainty.

Observations have errors which are characterized as random (also known as precision), systematic (also known as bias) and of representativeness (or representativity). We sometimes term the sum of these errors the accuracy. A property of random errors is their reduction when averaged. This is not the case of systematic errors; commonly, we subtract them from an observation if known. The representativeness error is associated with differences in the resolution of observational information and the resolution of the model interpreting this information.

Models also have errors. These errors arise through the construction of models, as models can be incomplete due to a lack of understanding or due to omission of processes to make the problem tractable; and through their imperfect simulation of the "real world", itself sampled by observations or measurements. Thus, information, whether in the form of observations or models has errors, and we must consider them. In data assimilation, the observations, models, and analyses have errors, never known precisely; we must estimate them. This means we must state the data assimilation problem in probabilistic terms (see, e.g., Cohn, 1997).

Typically, there are biases between different observations types, and between the observations and the model. Ménard (2010) discusses bias estimation in data assimilation. These biases vary in space and time, and it is a major challenge to estimate and correct them. Despite this, and mainly for pragmatic reasons, in

data assimilation we often assume that the errors in the observations and the background or model are unbiased. For numerical weather prediction, however, many assimilation schemes now incorporate a bias correction, which from the point of view of general estimation theory is the proper way to deal with biased data. There are various techniques to correct observations by removing the bias (e.g, Dee and da Silva, 1998); Dee (2005) reviews the treatment of biases in data assimilation systems.

The statistical post-processing approach of Steinbakk et al. (2013) returns fulls predictive distributions and the predicted most likely value may thus be supplemented with information regarding the corresponding predictive uncertainty as shown in Figure 7. That is, at a given location or in a fixed grid cell, we have access to supplementary information such as prediction intervals or the predictive probability of exceeding a given threshold.

This information may then, to a certain extent, be expanded to the entire region covered by the EPISODE model output. A simulated example of such information for $PM_{2.5}$ over the Oslo region is shown in Figure 9. Here, the most likely forecast for each grid cell given by the mean prediction is supplemented with the corresponding predictive uncertainty as represented by the standard deviation of the local predictive distribution. As expected, the forecast is highly non-stationary in space with the largest concentrations predicted along the busiest roads through the city. The uncertainty follows a similar spatial pattern revealing a strong spatial variability in the variance of the forecast.

The bottom plot in Figure 9 presents the point-wise probability of the realized value exceeding a threshold of 20 μm^{-3}. Again, we see a similar spatial pattern as in the two previous plots, the exceedance probability is higher than 20% over most of the city with the highest values over the busiest roads reaching approximately 45%. This aligns with the top plot showing the predicted mean values within the highlighted region ranging from approximately 15 μm^{-3} to 25 μm^{-3}. The simulated example in Figure 9 is just to illustrate how probabilites can be visualised in a map. A realistic threshold for $PM_{2.5}$ has to be chosen according to guidelines developed through interaction between the scientific and regulatory communities.

In addition to the local uncertainty information conveyed in Figure 9, many applications require uncertainty information regarding derived or composed quantities which again requires multivariate probabilistic forecasts with a physically coherent spatial structure. Examples of such applications include the predicted maximum or minimum value within a region, and the probability of exceeding a threshold at least once or everywhere within a region. The region of interest may, for instance, consist of a stretch of road such as the Ring 3 in Oslo.

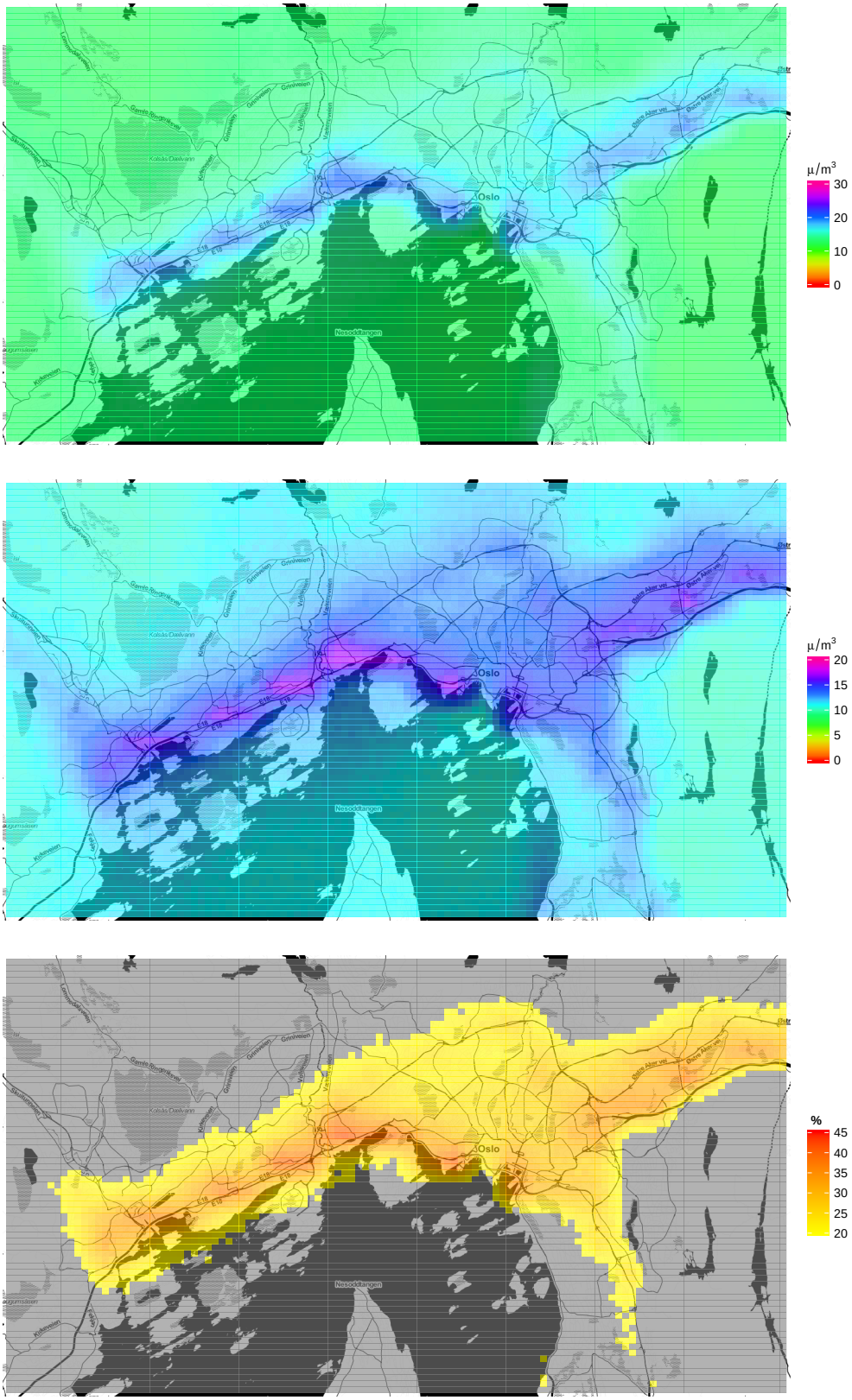


Figure 9. Simulated example of post-processed mean predictions of $PM_{2.5}$ over the entire Oslo region (top plot), the associated predicted standard deviation (centre plot), and the probability of exceeding a threshold of 20 μm^{-3} in each grid point (bottom plot).

Multivariate physically coherent forecasts are required in applications in various fields and several alternative approaches have been proposed. This subject is, e.g., a very active area of research within weather forecasting, see Schefzik et al. (2013) for a recent review. A physically based approach consists of running the deterministic numerical model multiple times using slightly varying initial and boundary conditions or, potentially, alternative parameterizations of the numerical model (Palmer, 2002). This results in a so-called ensemble forecast where each ensemble member is considered an equally likely representation of the future state. In this approach, the ensemble Kalman filter method described in Section 2 is commonly used to generate equally likely perturbations of the initial and boundary conditions.

Statistical techniques can be used to generate multivariate probabilistic forecast based on a single output from a numerical model. Gel et al. (2004) propose a geostatistical method to perturb the output of a numerical model rather than the input. The method may then be combined with a local post-processing technique such as that proposed by Steinbakk et al. (2013) in order to obtain post-processed forecast fields (Feldmann et al., 2015). While such an approach is computationally much more efficient than having to perform repeated runs of the numerical model, the original version of the technique as proposed by Gel et al. (2004) requires an observational dataset that provides a good representation of the spatial forecast error structure. Alternatively, the spatial structure may be assumed given as is the case in the bias correction method described in Section 2.2. The highly non-stationary structure apparent in Figure 9 suggests that this might not be appropriate for general regions. However, it might provide an attractive option when the region of interest can be assumed homogeneous, e.g. a stretch of road.

A third option to obtain physically coherent probabilistic forecast fields is to combine the local predictive distributions shown in Figure 7 with a multivariate structure learned from past observational data or the output of the numerical model without assuming a specific multivariate statistical model. This approach was first proposed by Clark et al. (2004) who used a large observational dataset to find a representative subset in which the underlying conditions may be assumed similar to the current conditions. The multivariate structure is then given by the empirical structure in the subset of interest and it can be combined with the post-processed marginal distributions using a copula structure. Schefzik et al. (2013) consider the case where a large forecast ensemble is available and the multivariate structure may be obtained directly from the ensemble forecast.

In the current setting, we have neither a large forecast ensemble nor a large, representative observational database. However, it might be feasible to combine the

approaches of Clark et al. (2004) and Schefzik et al. (2013) in that an appropriate similarity measure may be applied in order to select a subset of past forecast cases that represent states similar to the current state. This would yield an arbitrarily large ensemble which could be combined with the probabilistic marginal predictions of Steinbakk et al. (2013) to yield post-processed forecast fields as described in Schefzik et al. (2013). Such an approach would be computationally extremely efficient. However, it would require the existence of a database of past outputs from the numerical model over the entire forecast region.

5 Discussion and concluding remarks

5.1 Discussion

The two complementary approaches, a data assimilation technique combined with bias correction and a statistical post-processing of model output, presented in this report, significantly improve the predictive performance of the numerical forecasts within Bedre Byluft. The first approach has been tested on a short period in the winter season 2013-2014 with a focus on a specific meteorological condition, namely stagnant conditions ocurring during wintertime. The statistical post-processing is tested on a longer, but different, period than the one used in the first approach. The bias correction approach can be used alone without data assimilation, and can in this context be seen as a post-processing of the output from EPISODE similar to the statistical post-processing. Note that the latter method also corrects for bias in the original forecasts, based on an alternative approach, but adjusts for auto-correlation in the forecast error as well.

The data assimilation technique combined with a bias correction (Section 2), has been developed for improving EPISODE model concentrations. This approach has been tested using observations and 2 day (48 hour) EPISODE model forecasting data from Bedre Byluft in Oslo for the week 2 – 8 December 2013. The results, where we focus on the first 24 hours of the 48 hour forecasting period, show that both data assimilation and bias correction work reasonably well in that they manage to improve model concentrations as compared to observations at most stations in Oslo during this period. The amount of improvement is typically in the range of 20 - 80 % for the RMSE, and 0.2 - 0.3 for the correlations overall.

Since the model concentrations produced by EPISODE are highly biased during large parts of the test period, due to the unforeseen stagnant meteorological conditions, the bias correction method leads to the largest improvements in the modelled concentrations. However, use of data assimilation applied to the bias corrected model values, also helps to improve the modelled concentrations fur-

ther.

Except for Oslo, the other cities in Norway typically have only a few measurement stations for air quality. To assess the usefulness of post-processing in such situations, we have tested the performance of the statistical post-processing proposed by Steinbakk et al. (2013) assuming data is available only at a few locations in Oslo (Section 3). We tested the approach by applying a cross-validation scheme using a small subset of stations as a training set to fit the statistical post-processing model and then compared the results against stations not used for training the model. This procedure was repeated for different groups of measurement stations.

Even with only a few stations to train the model, the predictive performance of the post-processed prognosis for the particulate matters were almost always better than the original prognosis. The post-processed prognosis for NO_2 showed improvement in correlation compared to the original prognosis, but showed poorer fit under the RMSE disgnostic. In general, the relative predictive performance using the median prediction is better than that of the mean prediction of NO_2 compared to the original prognosis, while the mean prediction shows greater relative improvement for the particulate matters, $PM_{2.5}$ and PM_{10}. Whether to use mean or median predictions should be investigated further for a potential operationalization of the statistical post-processing method. The predictive performance of the post-processed NO_2-predictions showed much higher improvements in correlation than in RMSE compared to the original prognosis, which may indicate that the level of NO_2 concentration is not appropriately fitted.

The statistical post-processing of air quality forecast from EPISODE proposed by Steinbakk et al. (2013) provides a full predictive distribution, rather than merely a forecast of the most likely value. A full predictive distribution allows us to estimate the predictive probability of exceeding any threshold of interest. For a potential operationalization, we suggest further work on an improved calibration of the full predictive distributions for NO_2 to better fit the empirical data distributions. The predictive distributions of particulate matters have, on the other hand, shown to be more appropriate for describing the uncertainty throughout this study.

The methods data assimilation, bias correction and statistical post-processing, presented in this report, can all be implemented in the operational Bedre Byluft forecasting system. This is most easily done for the bias correction and statistical post-processing, since these procedures can be used together with the single deterministic EPISODE model run used in the current system. Thus, this can be regarded as an independent module in the Bedre Byluft forecast system. We expect that such a module would be fast and computationally efficient.

Implementation of data assimilation using the Ensemble Kalman Filter requires more changes to the Bedre Byluft system, since we then will need to be able to run parallel runs with the EPISODE model for the ensemble members (at least 5-10) during each forecasting period. Thus, this will require changes in the current model set up and script system, and additional computational resources. However, there is very good experience of making data assimilation operational for example, in the weather forecasting community.

In general, there are some technical issues relating to a possible operationalization that need to be considered, such as data flow and data quality. The statistical post-processing module, for instance, would rely on daily data of good quality entering the system in real time.

5.2 Future work

NILU and NR suggest the following areas of further work, some of which involve further collaboration between NILU and NR:

- Compare the capabilities of the two methods (NILU bias correction and data assimilation system; NR statistical post-processing) to communicate uncertainties associated with the air quality forecasts in Bedre Byluft. The period or periods selected should be sufficiently challenging and long to provide significant and robust results.

- Extend the statistical post-processing method of Steinbakk et al. (2013) to include a spatial correlation structure for a more flexible uncertainty assessment. If only homogeneous regions, such as sections of roads, are of interest, a first approach might consider a fixed spatial correlation structure. For a general spatial model over the entire forecast region, we suggest an application of the methods proposed by Clark et al. (2004) and Schefzik et al. (2013) given that a database of past model outputs is available.

- Implementation of a first version of the bias correction and EnKF data assimilation system (for improved initial conditions) as a part of the operational AirQUIS forecast system in Bedre Byluft. Perform further testing of the EnKF assimilation system using the implemented system in Oslo (and elsewhere in Norway), including analysis of the quality and robustness of the ensemble system, and evaluation of the analyses against independent observations. The system would be set up to allow extension to other cities in Norway.

- Apply the statistical post-processing method developed by NR to the new improved AirQUIS forecasts. In weather forecasting, we have seen that combining ensemble predictions with statistical post-processing gives a better predictive performance than each individual method, see e.g. Schefzik et al.

(2013). We thus assume that similar results will hold here. We should also investigate the effect of different meteorological conditions in improving the performance of the statistical post-processing method.

- Test the possibility of incorporating aspects of the statistical methods developed at NR into the NILU data assimilation system to process the results from data assimilation, that is, analyses and forecasts. This might involve improving the production of ensemble members, and testing assumptions such as the Gaussianity of the probability distribution functions.

- Improve aspects of the EPISODE model implemented in the data assimilation system. These aspects include: (i) implementing suggested improved traffic volume curves into NILU's AirQUIS emission database system (this is a part of the EPISODE model); and (ii) analysing systematic errors and biases in modelling of particulate matter, PM, in Oslo, and implementing improved correction algorithms for this in the model output. Improving the EPISODE model would be beneficial to the data assimilation work at NILU.

References

Box, G. E. P. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holdend-Day.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5:243–262.

Cohn, S. E. (1997). An introduction to estimation theory. *J. Meteorol. Soc. Jpn.*, 75:257–288.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken, NJ, John Wiley & Sons.

Dee, D. P. (2005). Bias and data assimilation. *Q. J. R. Meteorol. Soc.*, 131:3323–3343. doi:10.1256/qj.05.137.

Dee, D. P., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A. J., and Thépaut, J.-N. (2014). Toward a consistent reanalysis of the climate system. *Bull. Amer. Meteorol. Soc.* DOI:10.1175/BAMS-D-13-00043.1.

Dee, D. P. and da Silva, A. (1998). Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, 124:269–295.

Evensen, G. (2003). The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.*, 53:343–367. doi:10.1007/s10236-003-0036-9.

Evensen, G. (2007). *Data Assimilation: The Ensemble Kalman Filter*. Springer, Berlin; Heidelberg.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*. In press, doi: http://dx.doi.org/10.1175/MWR-D-14-00210.1.

Gel, Y., Raftery, A. E., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting. *Journal of the American Statistical Association*, 99:575–590.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, UK.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., and Ballabrera-Poy, J. (2007). 4d-var or ensemble kalman filter? *Tellus*, 59A:758–773. doi:10.1111/j.1600-0870.2007.00261.x.

Lahoz, W. A., Errera, Q., Swinbank, R., and Fonteyn, D. (2007). Data assimilation of stratospheric constituents: a review. *Atmos. Chem. Phys.*, 7:5745–5773. 10.5194/acp-7-5745-2007.

Lahoz, W. A., Khatattov, B., and Ménard, R. (2010a). Data assimilation and information. In Lahoz, W. A., Khatattov, B., and Ménard, R., editors, *Data Assimilation: Making Sense of Observations*, pages 3–12. Springer, Berlin.

Lahoz, W. A., Khatattov, B., and Ménard, R. (2010b). *Data Assimilation: Making Sense of Observations*. Springer, Berlin. doi:10.1007/978-3-540- 74703-1.

Lorenc, A. C. (2003). The potential of the ensemble kalman filter for nwp – a comparison with 4d-var. *J. R. Meteorol. Soc.*, 129:3183–3203. doi: 10.1256/qj.02.132.

Matheron, G. (1963). Traité de Géostatistique Appliqué, Tome II: le Krigeage. Editions Technip, Paris (Mémoires du Bureau de Recherches Géologiques et Minières, No 24).

Ménard, R. (2010). Bias estimation. In Lahoz, W. A., Khatattov, B., and Ménard, R., editors, *Data Assimilation: Making Sense of Observations*, pages 113–135. Springer, Berlin. doi:10.1007/978-3-540-74703-1_6.

Nichols, N. K. (2010). Mathematical concepts of data assimilation. In Lahoz, W. A., Khatattov, B., and Ménard, R., editors, *Data Assimilation: Making Sense of Observations*, pages 13–39. Springer, Berlin; Heidelberg.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128:747–774.

Rodgers, C. D. (2000). *Inverse Methods for Atmorpheric Sounding: Theory and Practice*. World Scientific, London.

Sakov, P. and Oke, P. R. (2008). Implications of the form of the ensemble transformation in the ensemble square root filters. *Monthly Weather Review*, 136:1042–1053.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640.

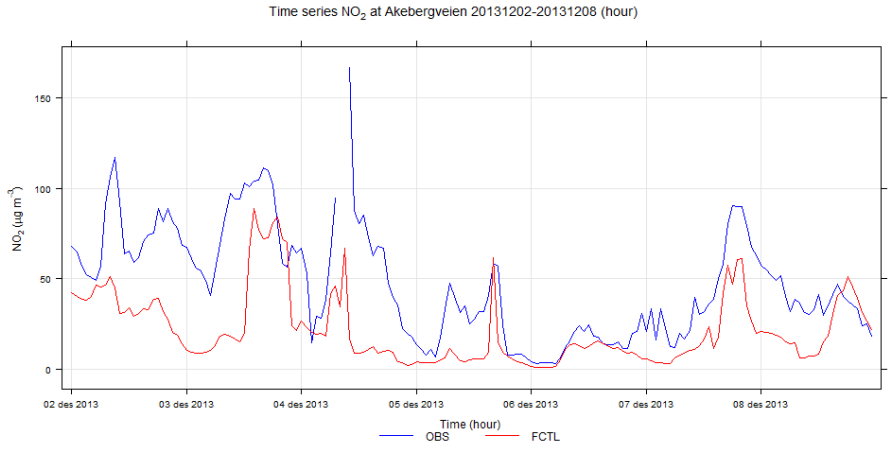
Slørdal, L. H., Walker, S. E., and Solberg, S. (2003). The Urban Air Dispersion Model EPISODE Applied in AirQUIS 2003. Technical Report NILUTR 12/2003, Norwegian Institute for Air Research, Kjeller. Available online at: http://www.nilu.no.

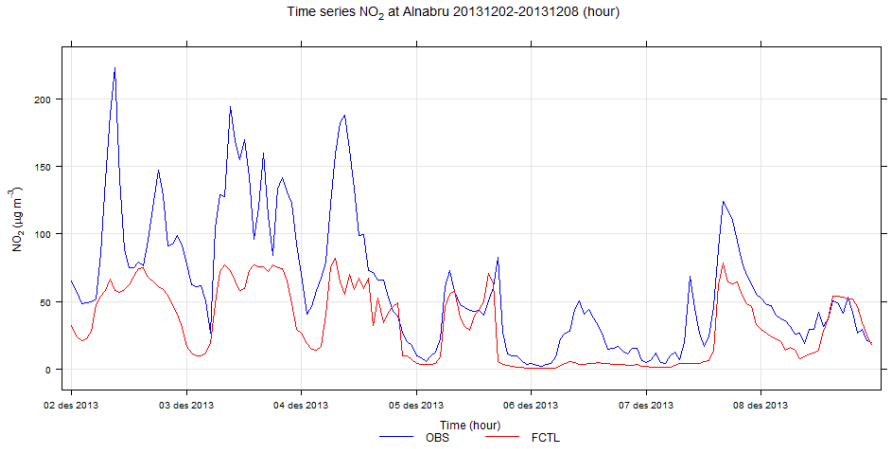
Steinbakk, G. H., Aldrin, M., and Thorarinsdottir, T. (2013). Statistiske metoder for korreksjon av deterministiske luftforurensningsprognoser. Technical Report SAMBA/38/13, Norwegian Computing Center.

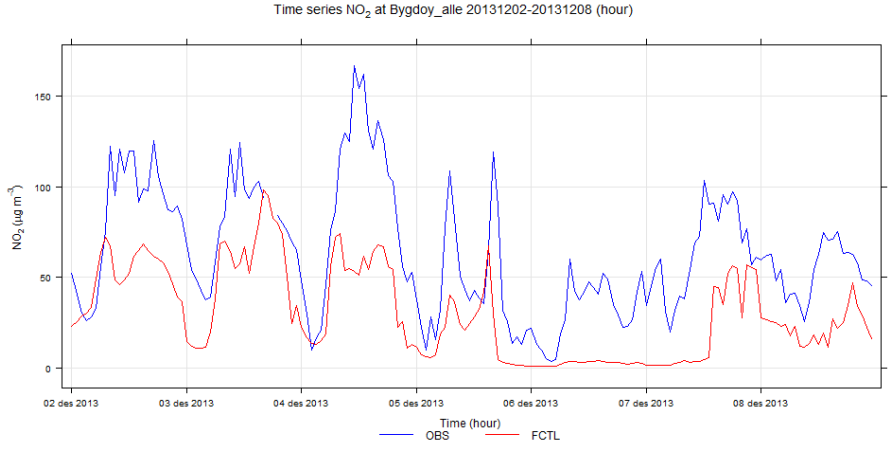
Ødegaard, V., Gjerstad, K. I., Slørdal, L. H., Abildsnes, H., and Olsen, T. (2013). Bedre byluft - prognoser for meteorologi og luftkvalitet i norske byer vinteren 2011 -2012 . Technical Report 10, Meteorologisk institutt.

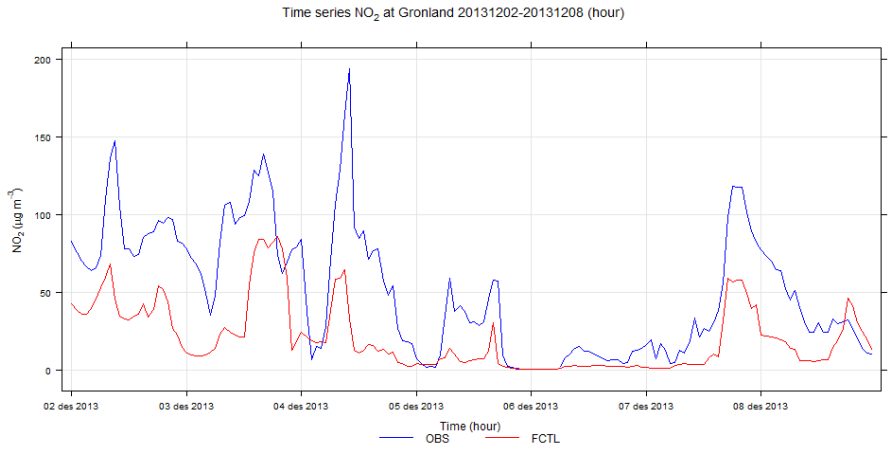
A Time series plots of observed and uncorrected model concentrations

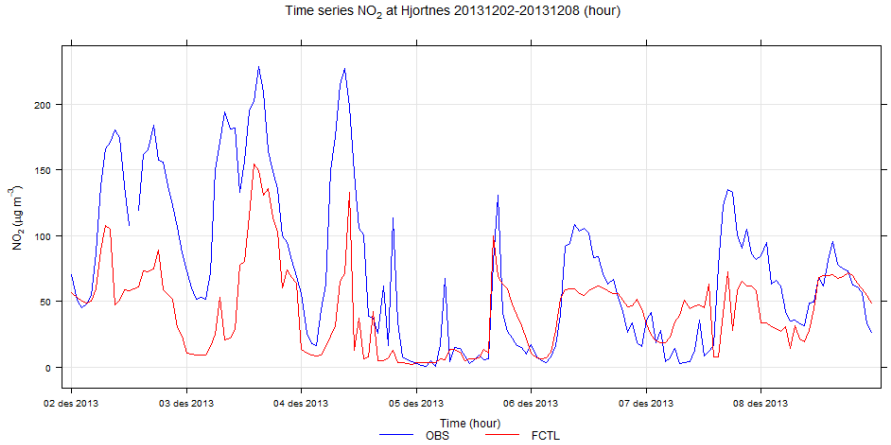
The figures in this appendix show observed (blue curve) and uncorrected model concentrations (red curve) at each station in Oslo for the period 2–8 December 2013 for each of the three species NO_2, PM_{10} and $PM_{2.5}$.

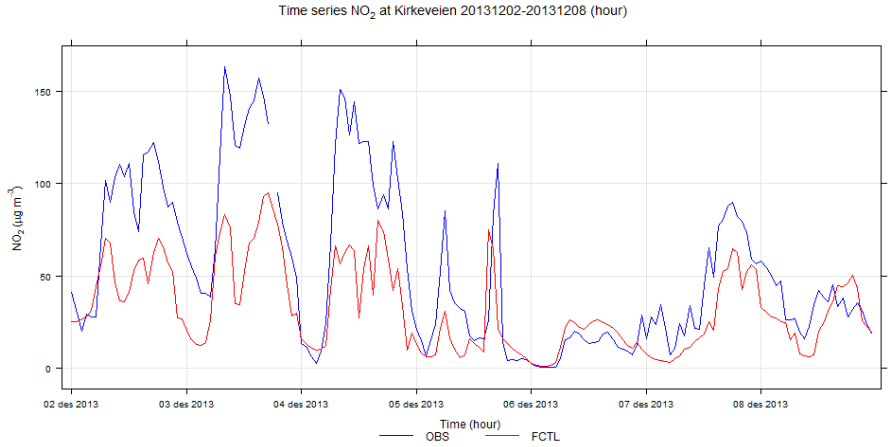


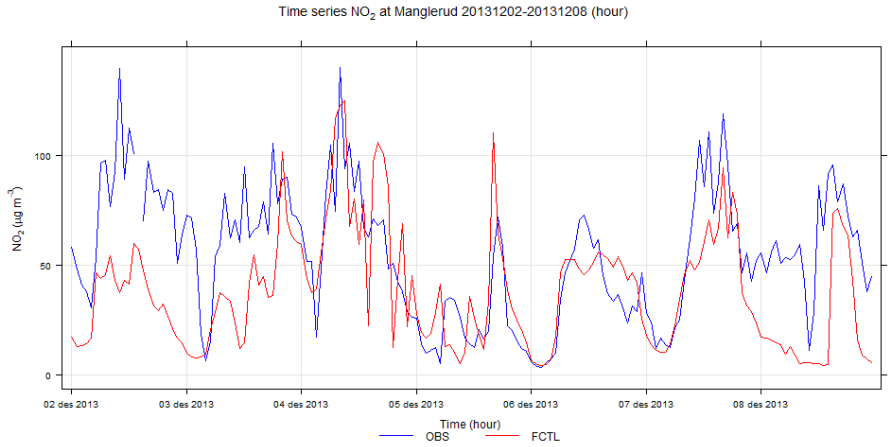


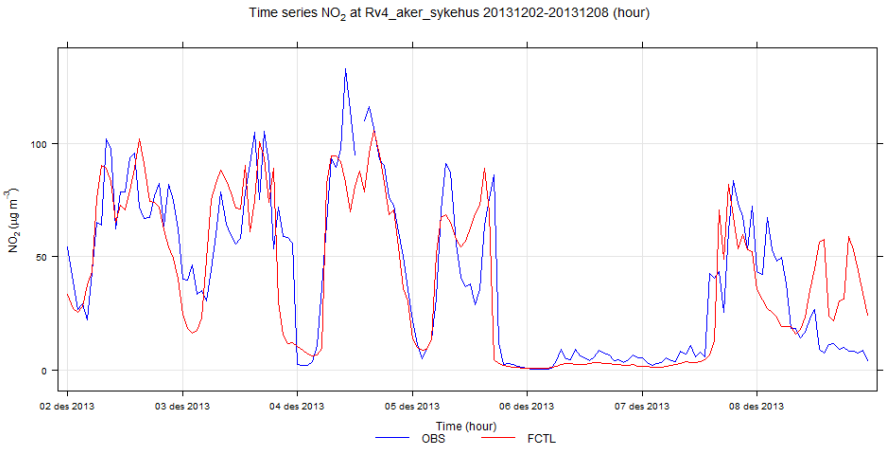


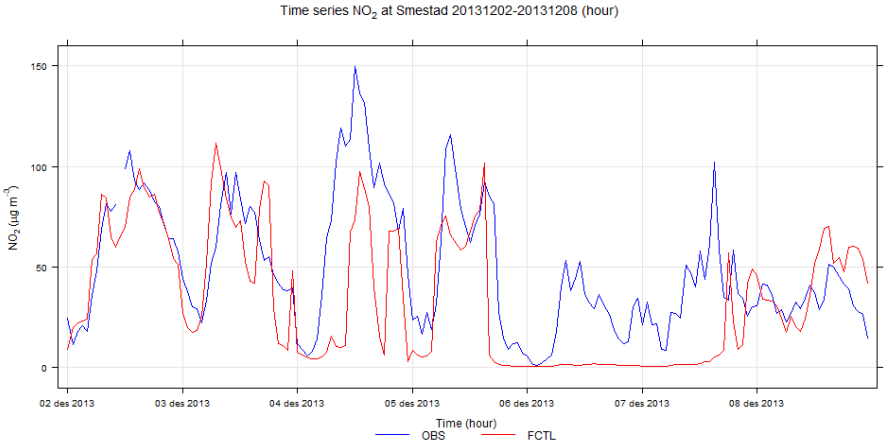

Time series NO2 at Gronland 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

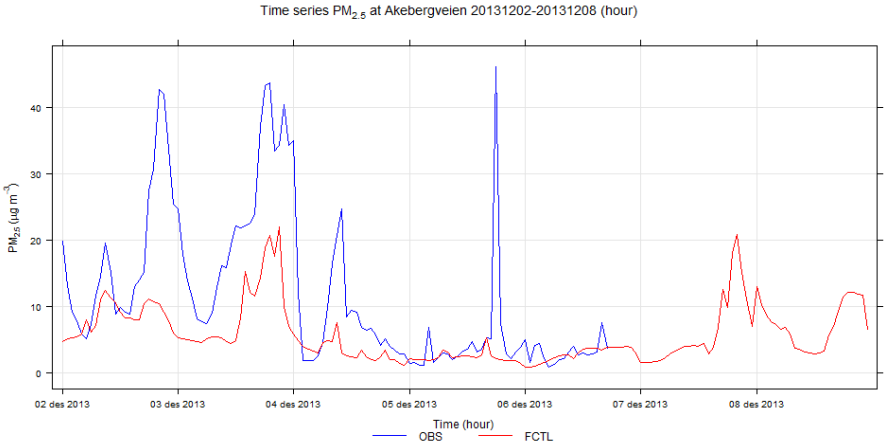

Time series NO2 at Hjortnes 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

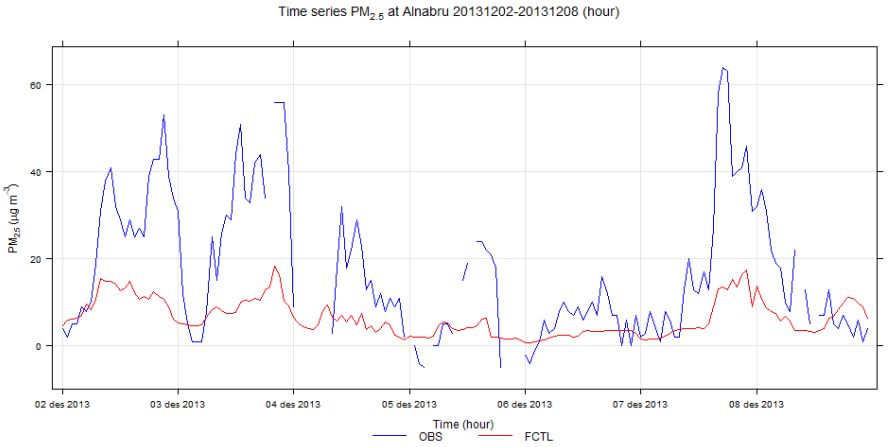

Time series NO2 at Kirkeveien 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

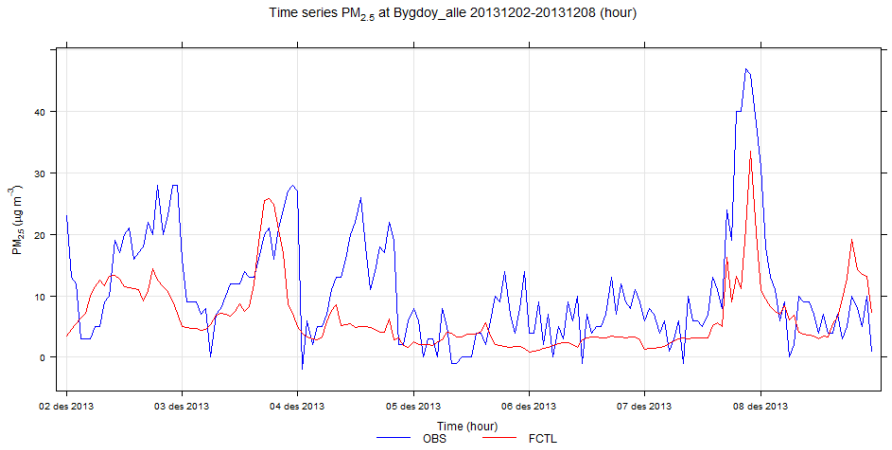

Time series NO2 at Manglerud 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

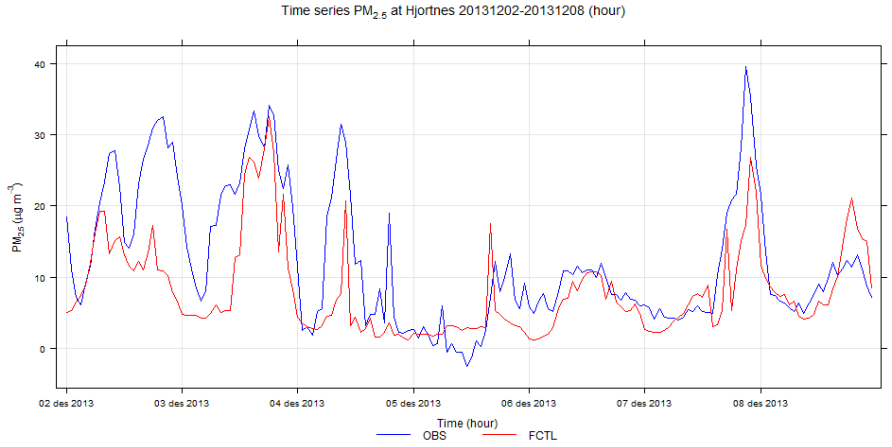

Time series NO2 at Rv4_aker_sykehus 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

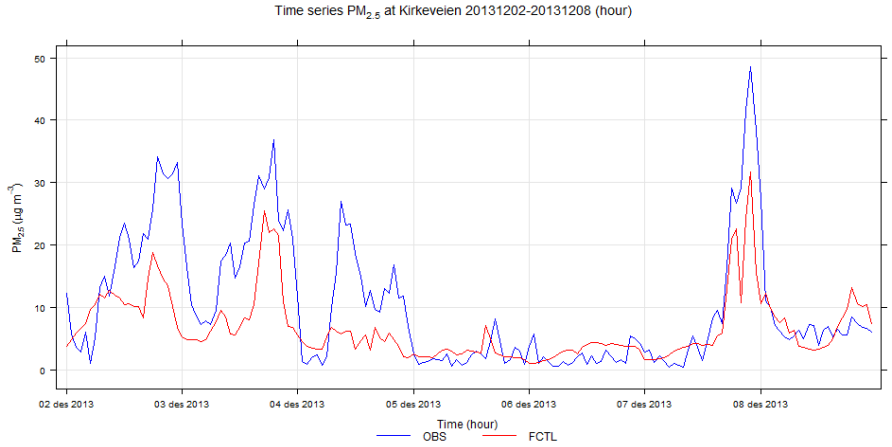

Time series NO2 at Smestad 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

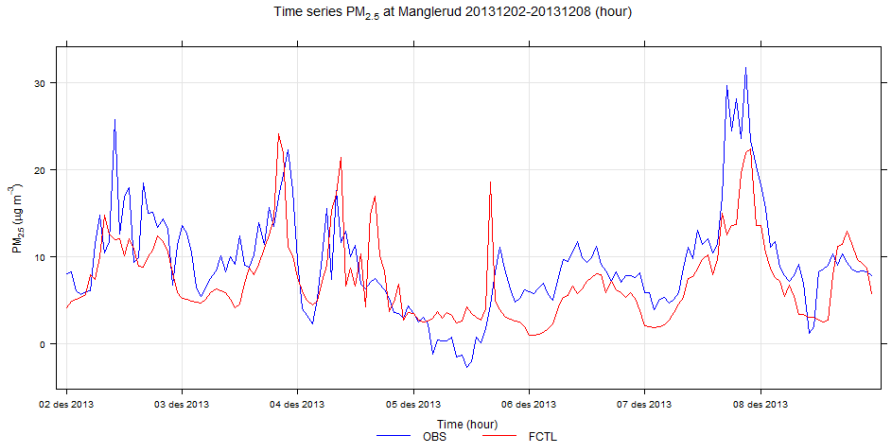

Time series PM2.5 at Akebergveien 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

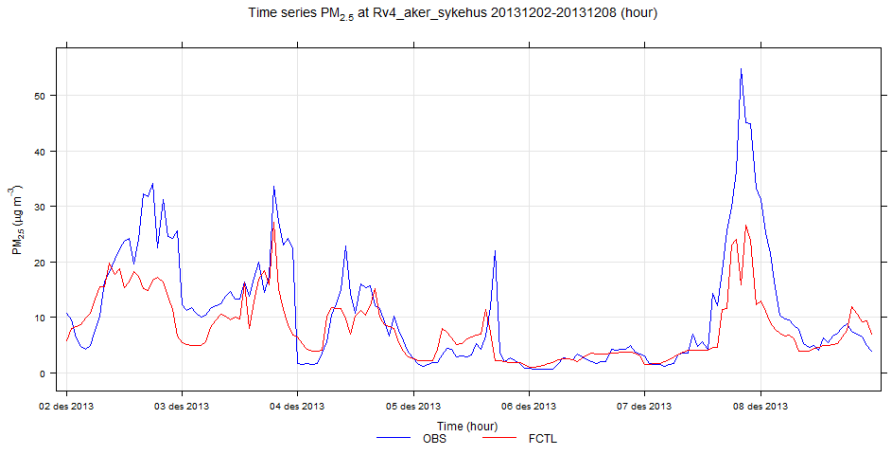

Time series PM2.5 at Alnabru 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
20
40
60
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

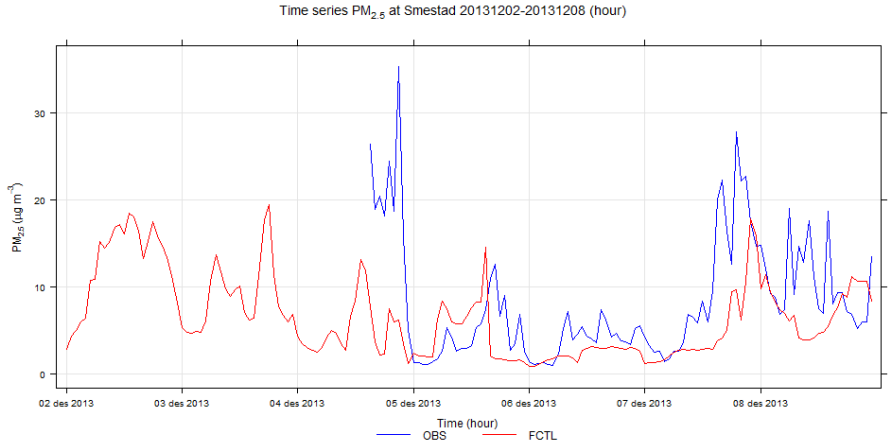

Time series PM2.5 at Bygdoy_alle 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

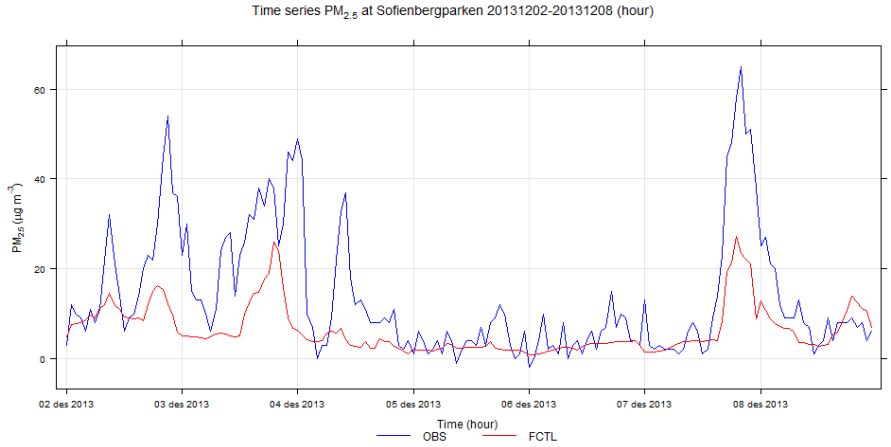

Time series PM2.5 at Hjortnes 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

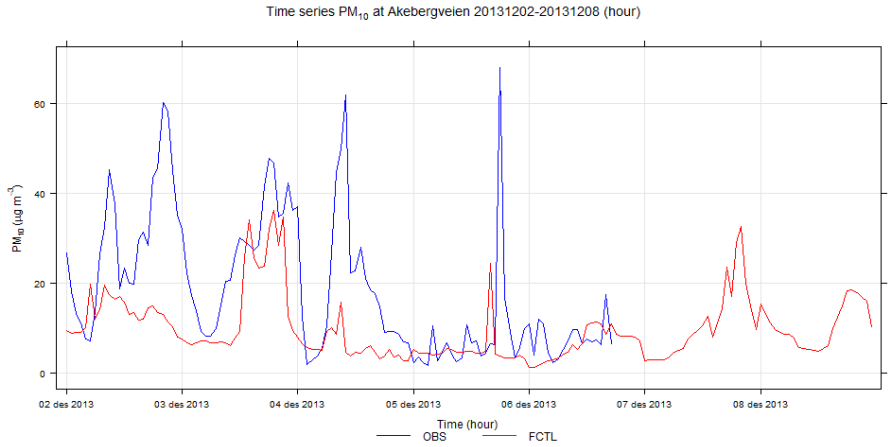

Time series PM2.5 at Kirkeveien 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
50
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

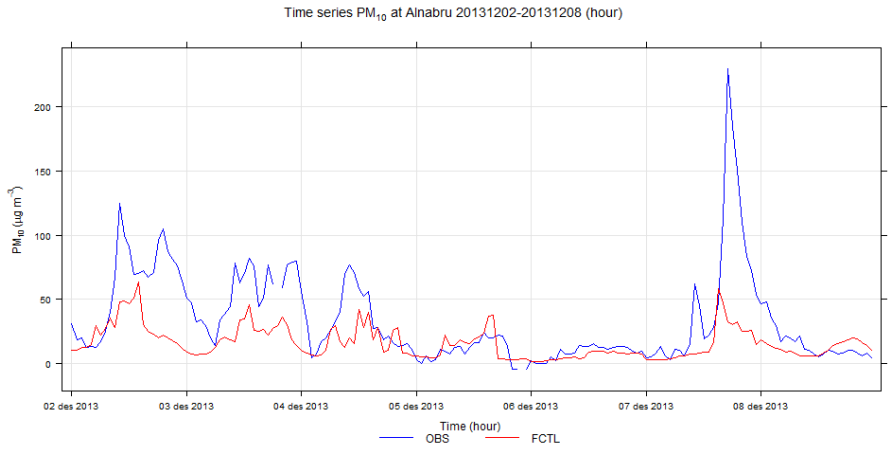

Time series PM2.5 at Manglerud 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

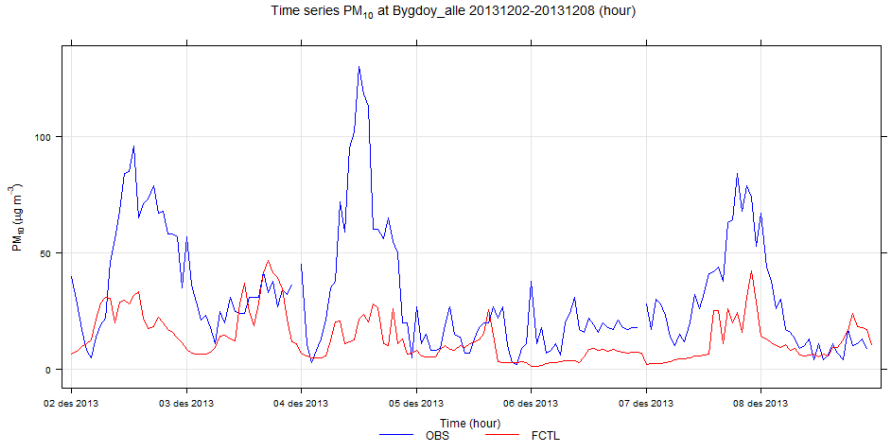

Time series PM2.5 at Rv4_aker_sykehus 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
50
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

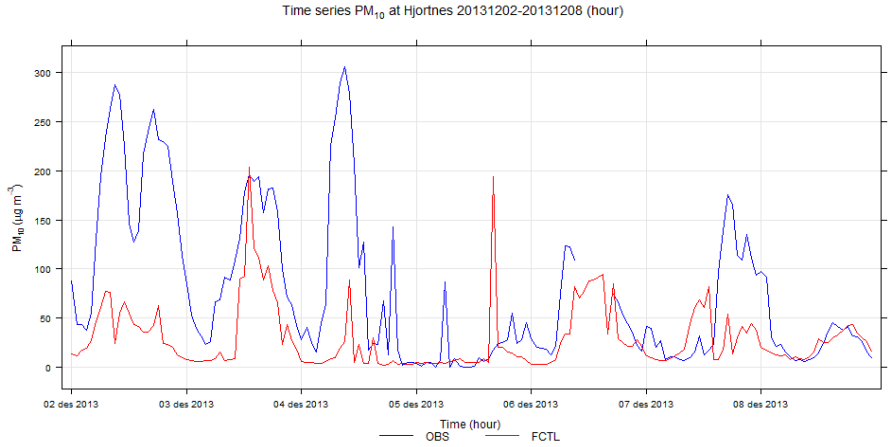

Time series PM2.5 at Smestad 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

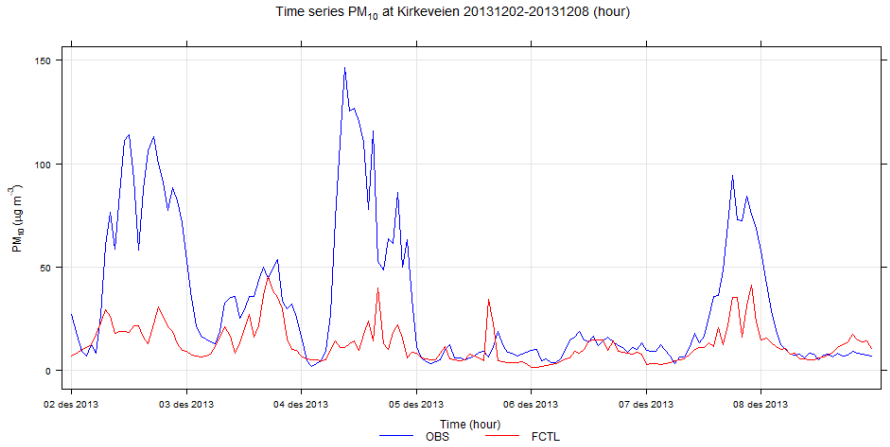

Time series PM2.5 at Sofienbergparken 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
20
40
60
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL


Time series PM10 at Akebergveien 20131202-20131208 (hour)
PM10 (µg m-3)
0
20
40
60
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

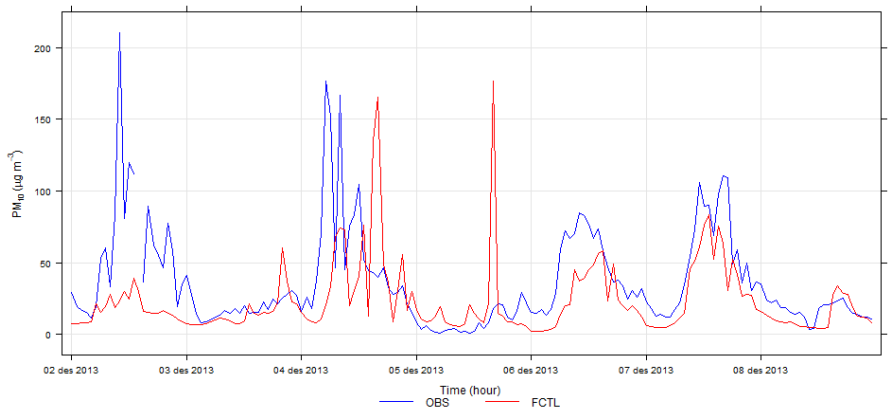

Time series PM10 at Alnabru 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL


Time series PM10 at Bygdoy_alle 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

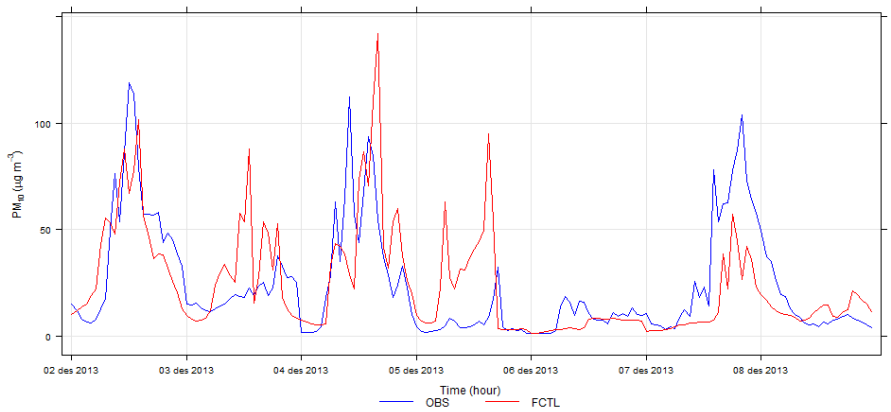

Time series PM10 at Hjortnes 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
200
250
300
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL


Time series PM10 at Kirkeveien 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
FCTL

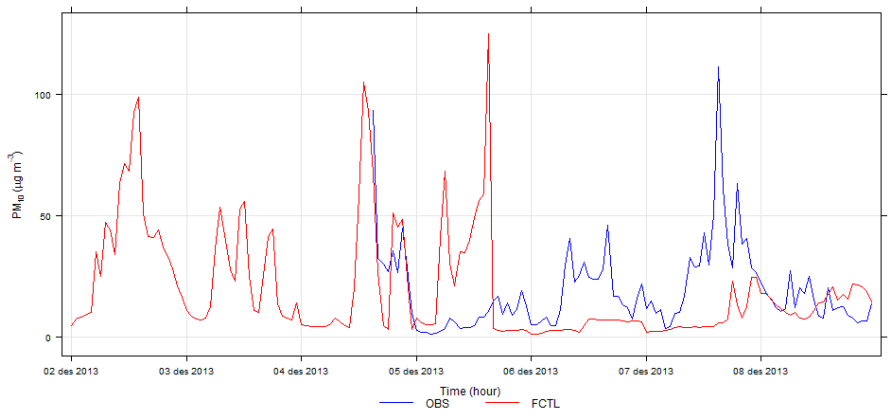
Time series PM_{10} at Manglerud 20131202-20131208 (hour)



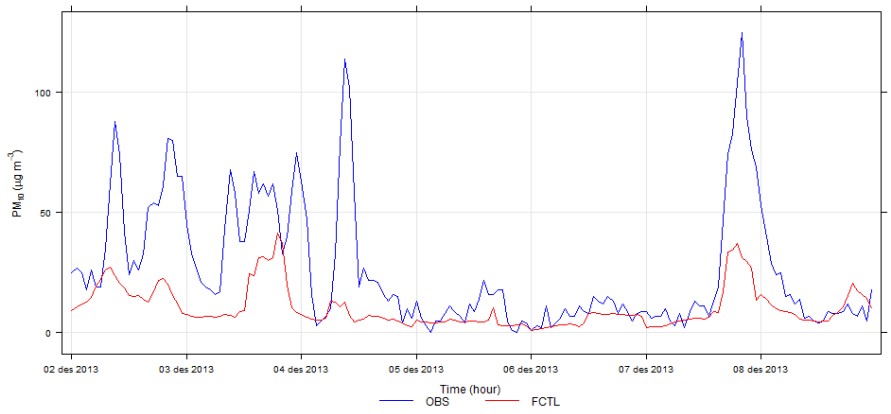
Time series PM_{10} at Rv4_aker_sykehus 20131202-20131208 (hour)



Time series PM_{10} at Smestad 20131202-20131208 (hour)

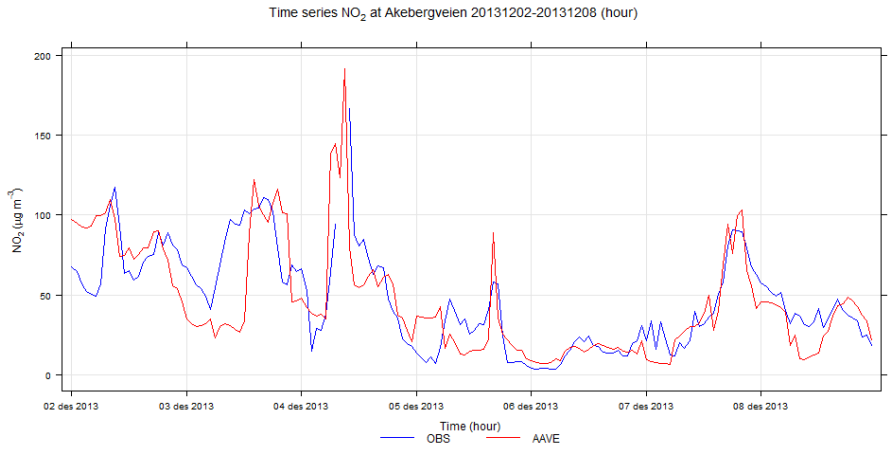


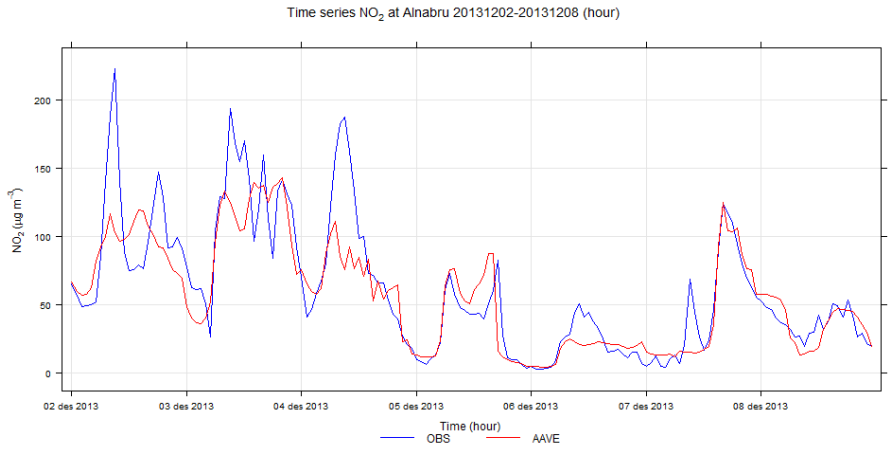
Time series PM_{10} at Sofienbergparken 20131202-20131208 (hour)

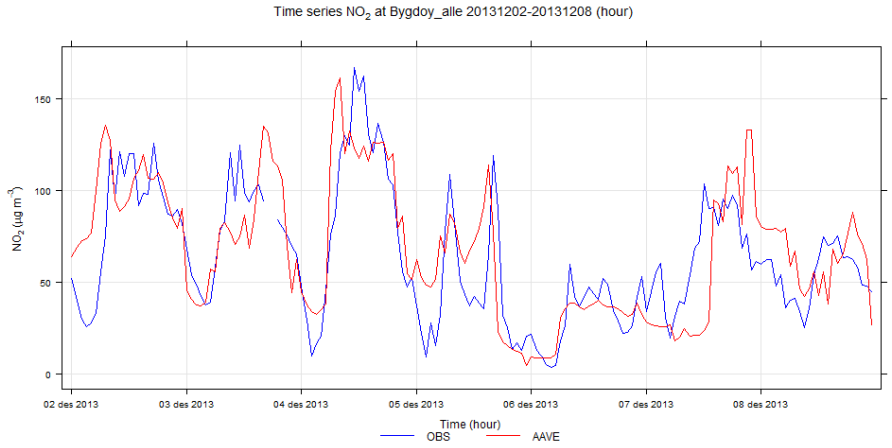


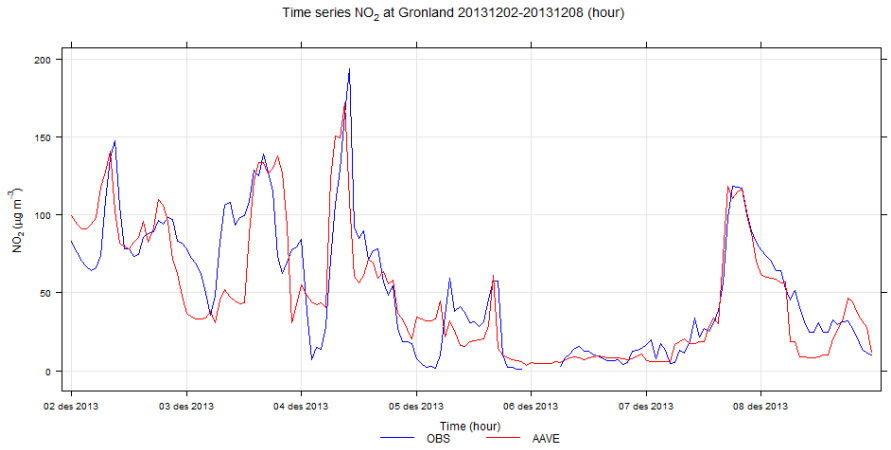
B Time series plots of observed and corrected model concentrations

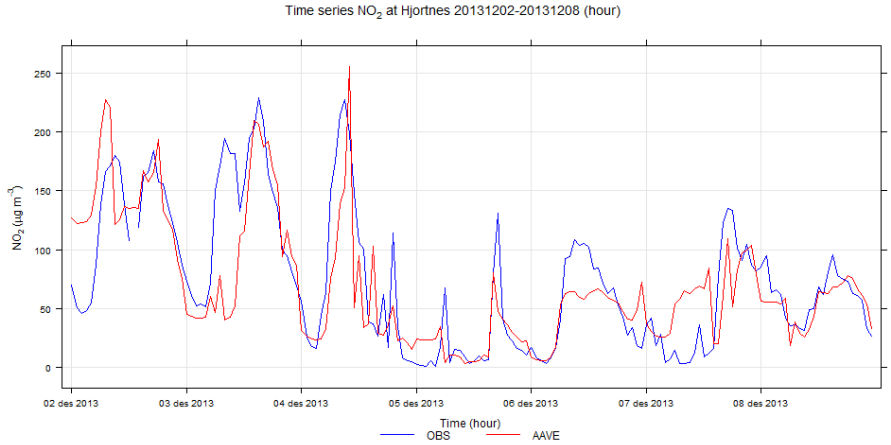
The figures in this appendix show observed (blue curve) and corrected (using both bias correction and data assimilation) model concentrations (red curve) at each station in Oslo for the period 2–8 December 2013 for each of the three species NO_2, PM_{10} and $PM_{2.5}$.

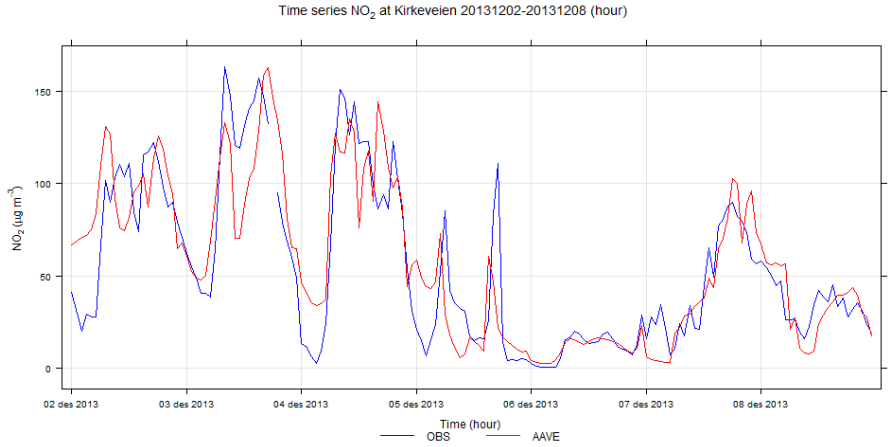


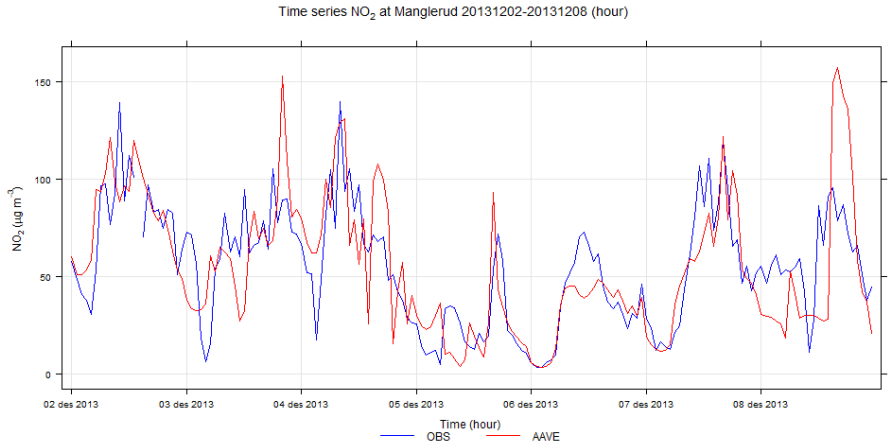


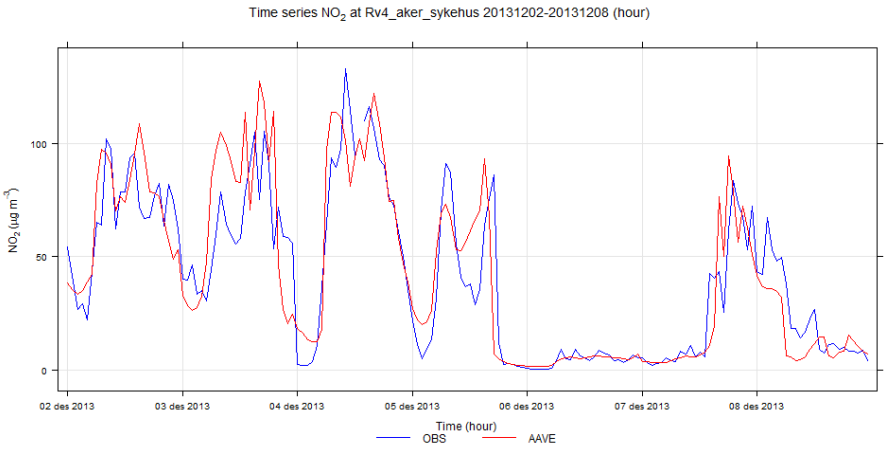


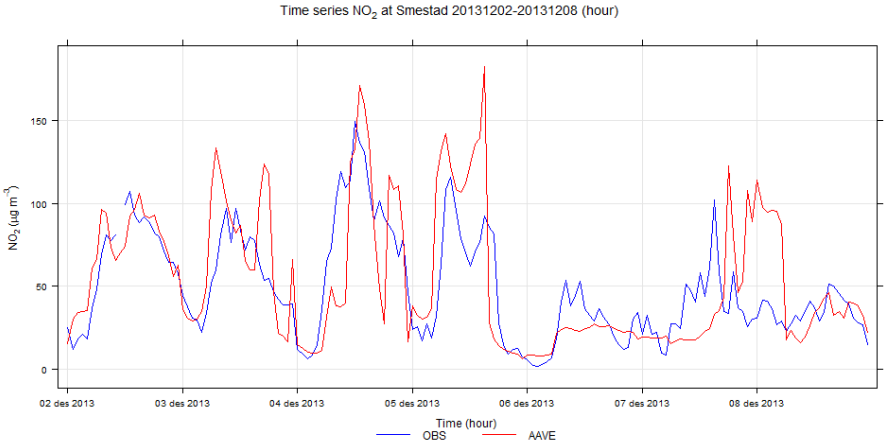

Time series NO2 at Gronland 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

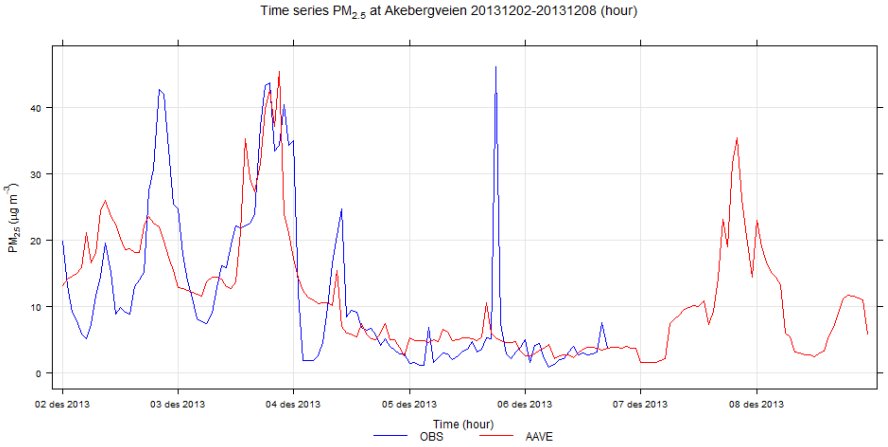

Time series NO2 at Hjortnes 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
200
250
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

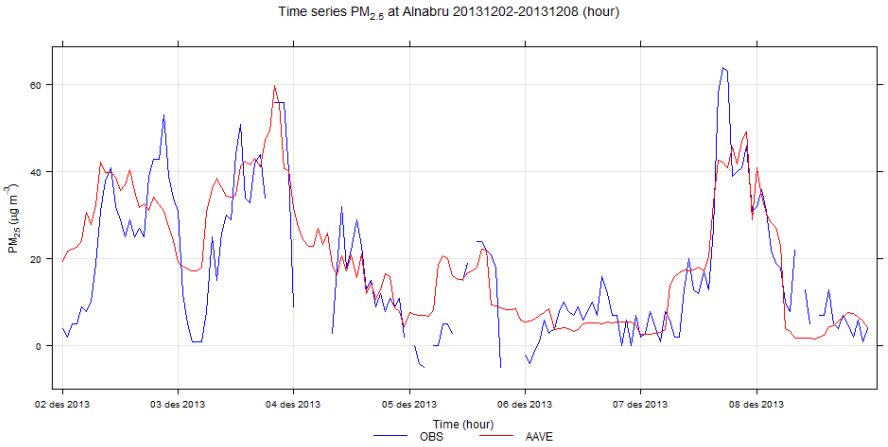

Time series NO2 at Kirkeveien 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

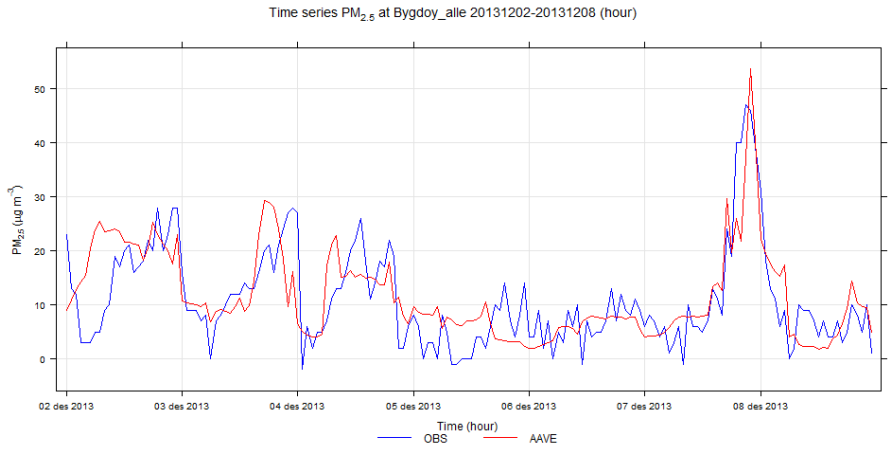

Time series NO2 at Manglerud 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

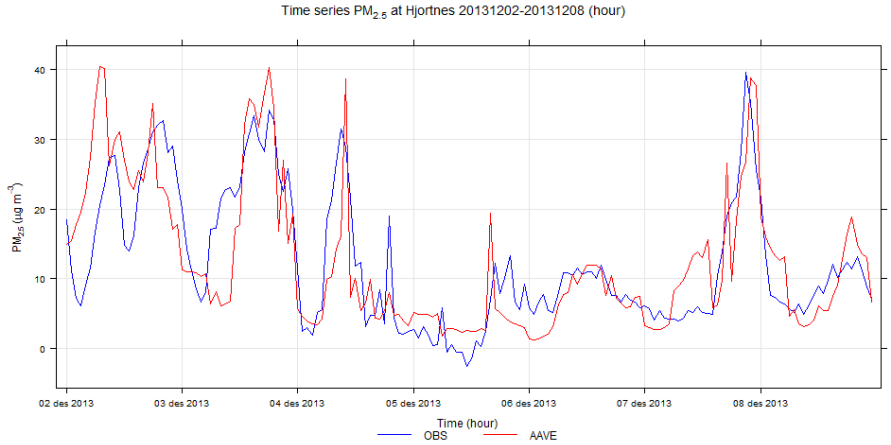

Time series NO2 at Rv4_aker_sykehus 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

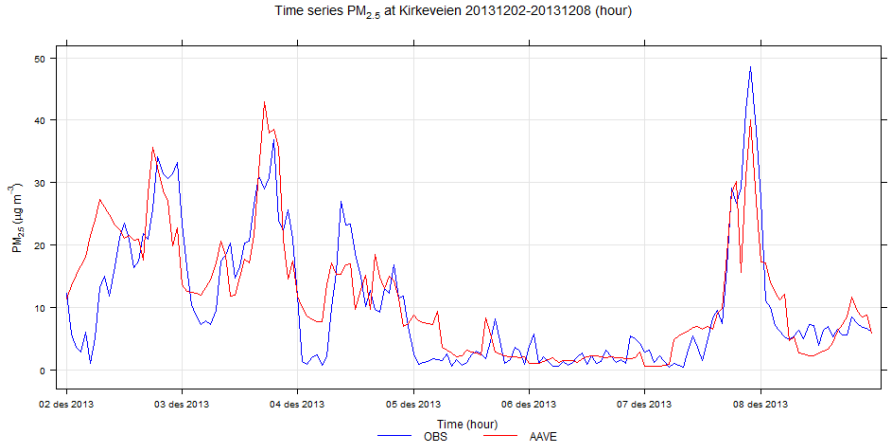

Time series NO2 at Smestad 20131202-20131208 (hour)
NO2 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

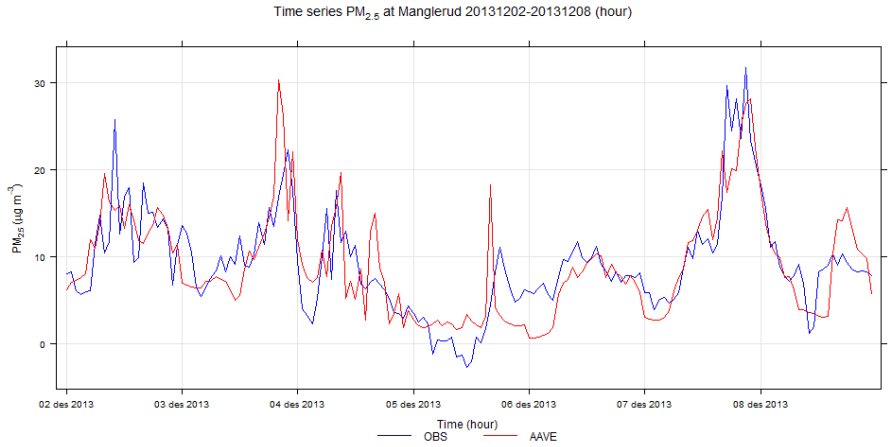

Time series PM2.5 at Akebergveien 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

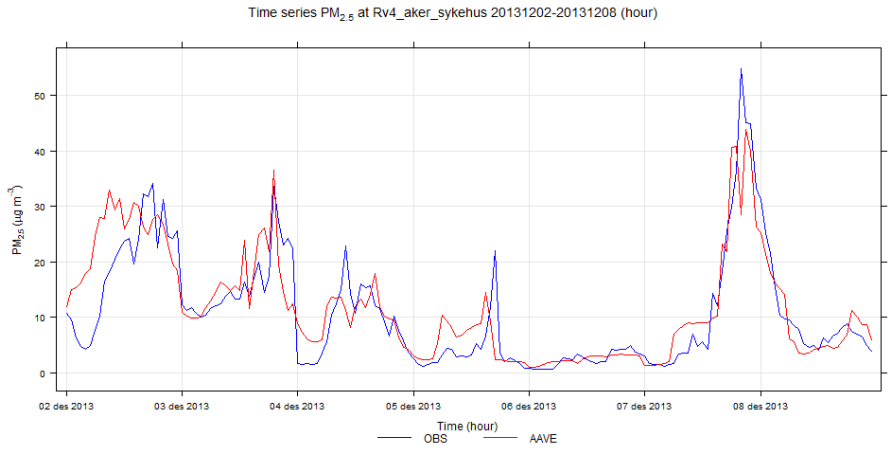

Time series PM2.5 at Alnabru 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
20
40
60
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

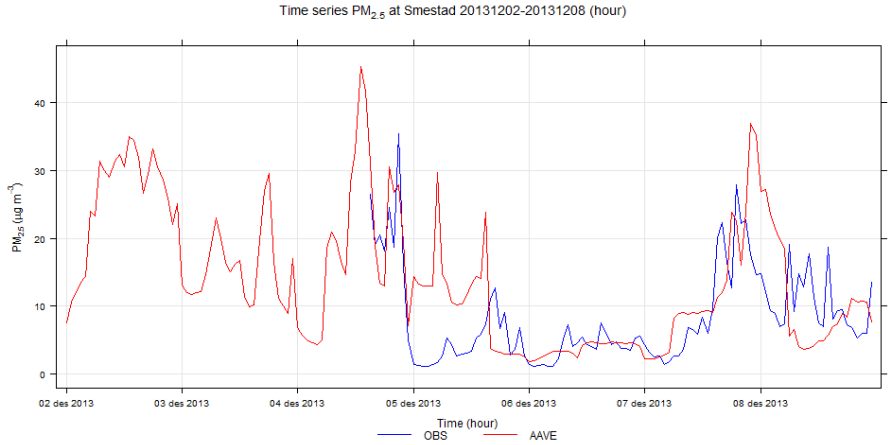

Time series PM2.5 at Bygdoy_alle 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
50
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

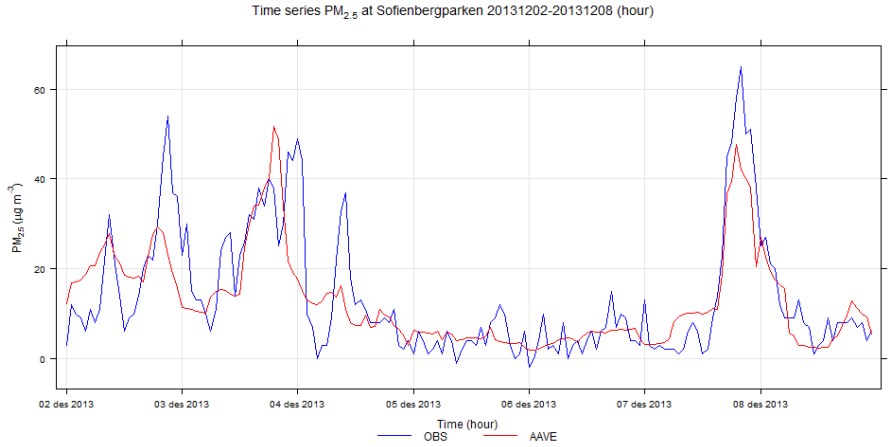

Time series PM2.5 at Hjortnes 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

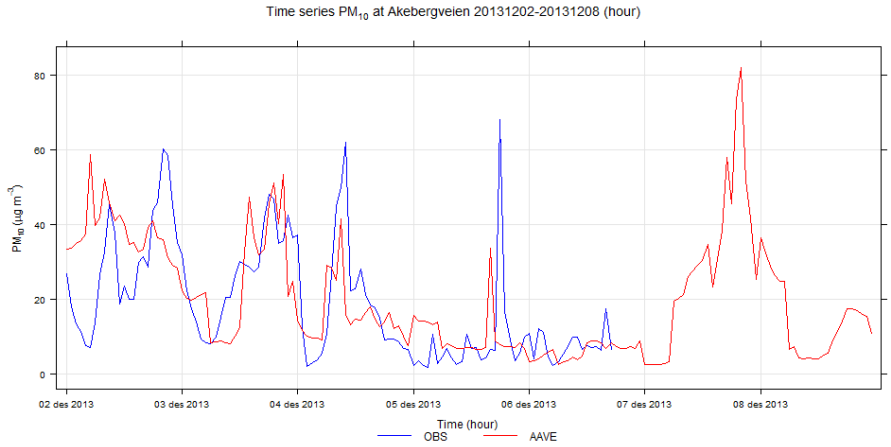

Time series PM2.5 at Kirkeveien 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
40
50
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

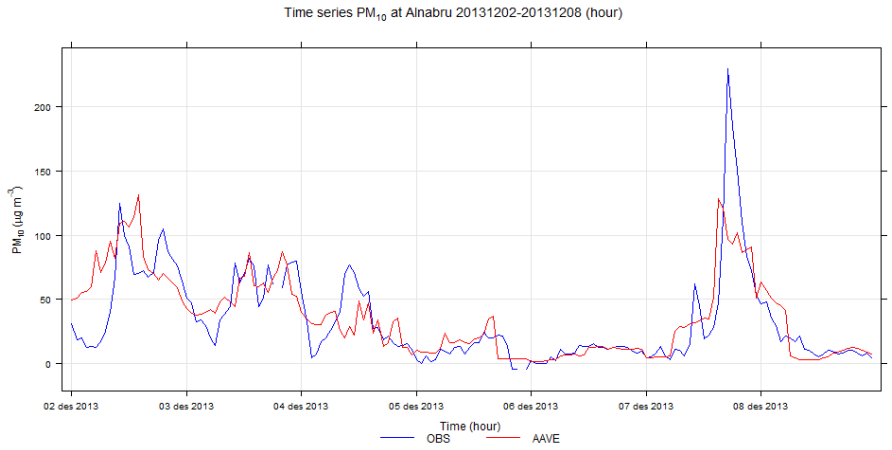

Time series PM2.5 at Manglerud 20131202-20131208 (hour)
PM2.5 (µg m-3)
0
10
20
30
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

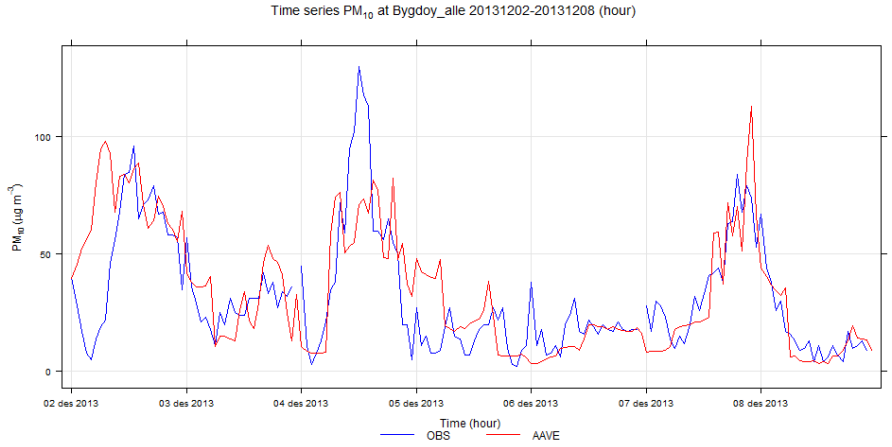

Time series PM2.5 at Rv4_aker_sykehus 20131202-20131208 (hour)
PM25 (µg m-3)
0
10
20
30
40
50
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

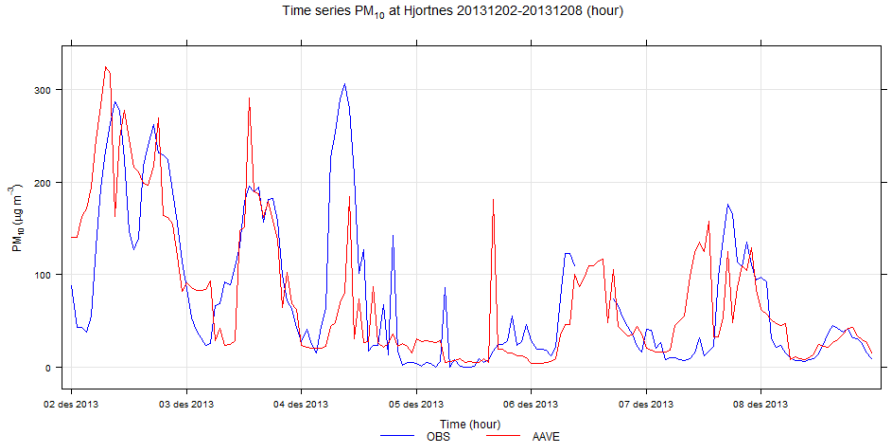

Time series PM2.5 at Smestad 20131202-20131208 (hour)
PM25 (µg m-3)
0
10
20
30
40
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

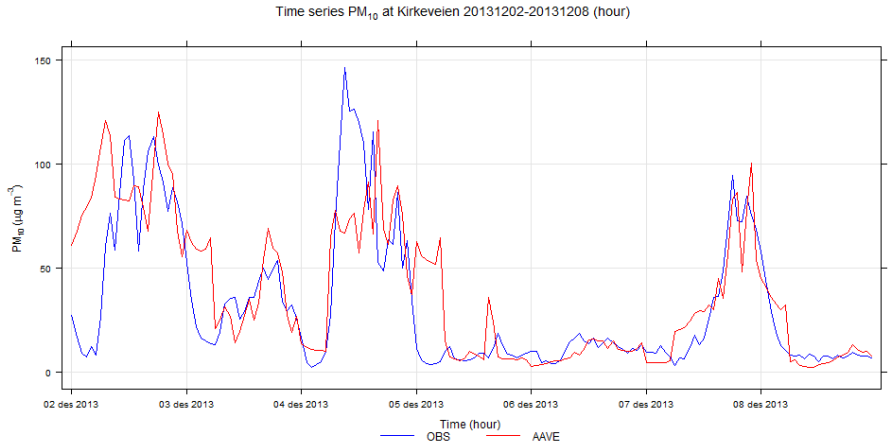

Time series PM2.5 at Sofienbergparken 20131202-20131208 (hour)
PM25 (µg m-3)
0
20
40
60
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

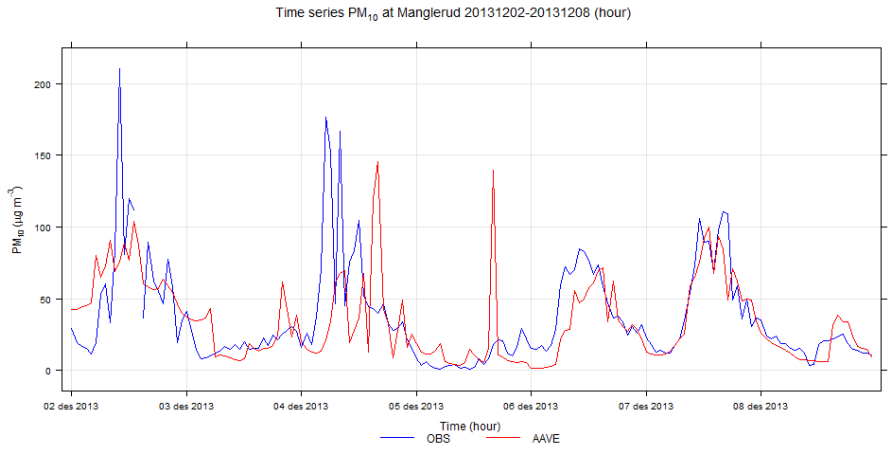

Time series PM10 at Akebergveien 20131202-20131208 (hour)
PM10 (µg m-3)
0
20
40
60
80
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

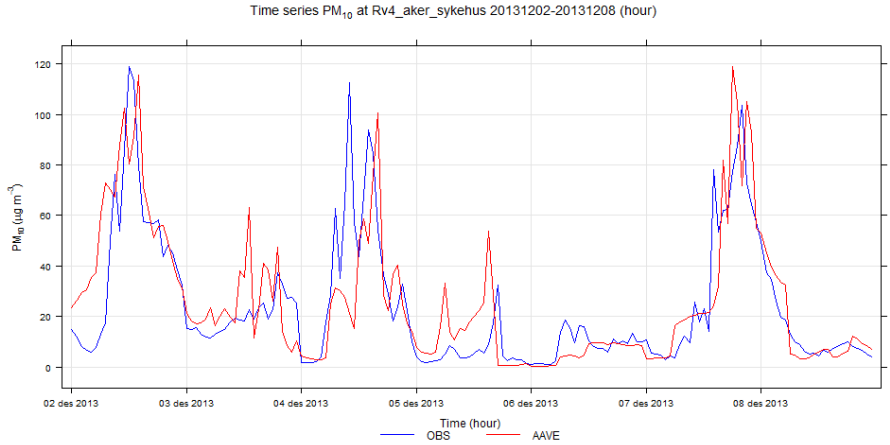

Time series PM10 at Alnabru 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

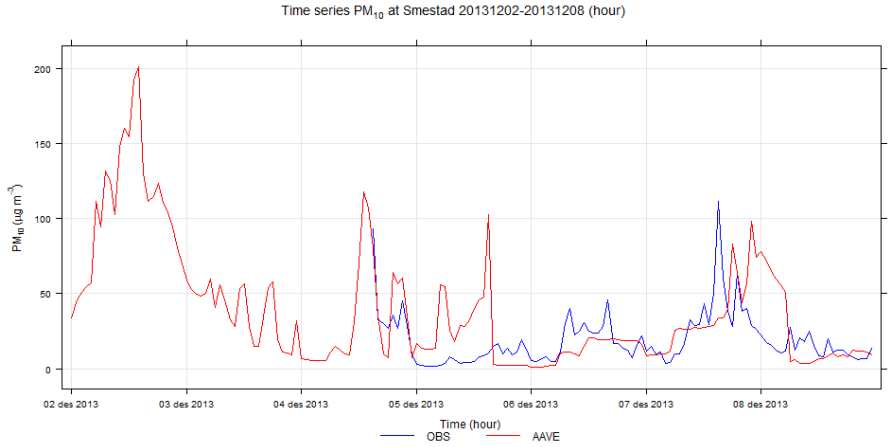

Time series PM10 at Bygdoy_alle 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

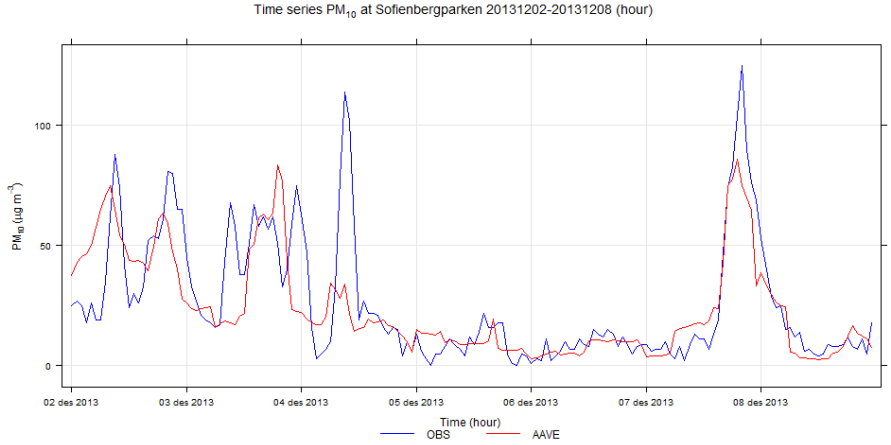

Time series PM10 at Hjortnes 20131202-20131208 (hour)
PM10 (µg m-3)
0
100
200
300
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE


Time series PM10 at Kirkeveien 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE


Time series PM10 at Manglerud 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE


Time series PM10 at Rv4_aker_sykehus 20131202-20131208 (hour)
PM10 (µg m-3)
0
20
40
60
80
100
120
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE


Time series PM10 at Smestad 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
150
200
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE


Time series PM10 at Sofienbergparken 20131202-20131208 (hour)
PM10 (µg m-3)
0
50
100
02 des 2013
03 des 2013
04 des 2013
05 des 2013
06 des 2013
07 des 2013
08 des 2013
Time (hour)
OBS
AAVE

C Algorithm for computing the predictive distribution

The algorithm for computing the predictive probability distribution in Eq. (10) is given in Algorithm 1. We estimate the model for time points $t = 1, \ldots, T$ and simulate the calibrated predictive distribution for time points $t = T + 1h, \ldots, T + 48h$. The deterministic prognoses are given in a grid G, where the m measurement stations are located in a unique subset G^{obs} of G.

Model adaptation based on historical data

- Compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in (8) based on all observations $y_{j,t}^{\text{obs}}$ and their corresponding prognoses $\hat{y}_{j,t}$ for time points $t = 1, \ldots, T$ and grid points $j \in G^{\text{obs}}$.
- Compute the error term $\hat{n}_{j,t} = y_{j,t}^{\text{obs}} - \hat{\beta}_0 - \hat{\beta}_1 \hat{y}_{j,t}$ for $j \in G^{\text{obs}}$ and $t = 1, \ldots, T$.
- Estimate the underlying process $\hat{l}_t = \frac{1}{m} \sum \hat{n}_{j,t}$ for all time points $t = t_0, \ldots, T$.
- Estimate the parameters ϕ_1 and ϕ_{24} from $\hat{l}_t$.

Predictive distribution at all grid points G for $t = T + 1, \ldots, T + 48$:

- Compute $\hat{l}_{t|T} = \hat{\phi}_1 \hat{l}_{t-1|T} + \hat{\phi}_{24} \hat{l}_{t-24|T} - \hat{\phi}_1 \hat{\phi}_{24} \hat{l}_{t-25|T}$.
- Estimate $n_{j,t}$ by $\hat{n}_{j,t|T} = \hat{l}_{t|T}$
- Compute the estimated mean $\hat{\mu}_{j,t|T} = \hat{\beta}_0 + \hat{\beta}_1 \hat{y}_{j,t|T} + \hat{n}_{j,t|T}$
- Compute the estimated variance $\widehat{sd}^2_{j,t|T} = \hat{\tau}_{t|T} + \hat{\sigma}_{t|T}$.
- Simulate $Y_{j,t} \sim \exp\{\mathcal{N}(\hat{\mu}_{j,t|T}, \widehat{sd}_{j,t|T})\}$ where $\mathcal{N}$ is the normal distribution

Algorithm 1. Algorithm for simulating predictive distribution of an air pollutant.

We need the prediction variances, given as $\hat{\tau}_{t|T} + \hat{\sigma}_{t|T}$, to compute the predictive distribution on the original scale. Here, the sub-script $j, t|T$ indicates a predicted value at a grid point j and a time point t given information up to time point T. The first term is the prediction error of $\hat{l}_{t|T}$ that can be estimated by a standard statistical software for auto regressive models. The other term is the estimated variance of $\epsilon_{j,t}$ which we compute as the empirical variance of $(\hat{n}_{j,t} - \hat{l}_t)$:

$$\hat{\sigma}_{t|T} = \frac{1}{N_e - T} \sum_{t=1}^{T} \sum_{j \in G^{\text{obs}}} I_{jt} (\hat{n}_{j,t} - \hat{l}_t)^2.$$

Here, $I_{j,t}$ is an indicator function given as

$$I_{j,t} = \begin{cases} 1 & \text{if } \hat{n}_{j,t} \text{ is known} \\ 0 & \text{if } \hat{n}_{j,t} \text{ is missing} \end{cases}$$

and $N_e = (\sum_j \sum_t I_{j,t})$, which means we are only counting those time points without missing values.

D Results for two neighbouring measurement stations

The model is estimated based on data from two groups containing two stations:

Group 1 Sofienbergparken, Kirkeveien

Group 2 Kirkeveien, Manglerud

The results are summarized in the Tables D.1-D.4.

Table D.1. RMSE and COR for PM_{10} using the mean estimate. The training data sets in group 1 and group 2 contain two stations.

PM10	RMSE				COR			
Season 1	Orig.	Gr. 1	Gr. 2	All st.	Orig.	Gr. 1	Gr. 2	All st.
RV4	23.0	19.0 (17%)	18.5 (19%)	18.8 (18%)	0.59	0.65 (9%)	0.68 (14%)	0.67 (13%)
Åkeberg.	25.9	23.9 (7%)	24.0 (7%)	23.0 (11%)	0.49	0.50 (3%)	0.53 (8%)	0.56 (15%)
Hjortnes	59.0	34.7 (41%)	32.5 (45%)	32.3 (45%)	0.29	0.34 (16%)	0.37 (28%)	0.36 (25%)
Season 2								
RV4	20.3	18.1 (11%)	18.5 (9%)	18.0 (11%)	0.52	0.61 (16%)	0.59 (12%)	0.62 (19%)
Åkeberg.	19.5	16.5 (15%)	16.8 (14%)	16.0 (18%)	0.47	0.63 (33%)	0.62 (30%)	0.66 (39%)
Hjortnes	39.5	28.5 (28%)	28.5 (28%)	28.3 (28%)	0.29	0.42 (42%)	0.43 (45%)	0.44 (50%)

Table D.2. RMSE and COR for $PM_{2.5}$ using the mean estimate. The training data sets in group 1 and group 2 contain two stations.

PM2.5	RMSE				COR			
Season 1	Orig.	Gr. 1	Gr. 2	All st.	Orig.	Gr. 1	Gr. 2	All st.
RV4	7.9	5.8 (26%)	5.1 (35%)	5.5 (30%)	0.57	0.61 (7%)	0.64 (13%)	0.63 (11%)
Åkeberg.	9.2	8.6 (6%)	9.1 (1%)	8.4 (9%)	0.57	0.60 (5%)	0.64 (12%)	0.65 (13%)
Hjortnes	11.5	7.1 (38%)	5.8 (50%)	7.0 (39%)	0.42	0.46 (10%)	0.51 (22%)	0.49 (17%)
Season 2								
RV4	6.1	5.6 (8%)	4.7 (23%)	5.8 (5%)	0.46	0.50 (8%)	0.53 (15%)	0.52 (13%)
Åkeberg.	8.9	8.1 (10%)	8.8 (2%)	7.9 (12%)	0.49	0.58 (19%)	0.59 (20%)	0.61 (25%)
Hjortnes	8.5	6.6 (23%)	5.9 (31%)	6.8 (20%)	0.42	0.47 (13%)	0.49 (17%)	0.47 (12%)

Table D.3. RMSE and COR for PM_{10}-$PM_{2.5}$ using the mean estimate. The training data sets in group 1 and group 2 contain two stations.

PMc	RMSE				COR			
Season 1	Orig.	Gr. 1	Gr. 2	All st.	Orig.	Gr. 1	Gr. 2	All st.
RV4	19.4	22.2 (-14%)	17.7 (9%)	16.6 (15%)	0.61	0.63 (4%)	0.67 (10%)	0.70 (15%)
Åkeberg.	22.2	25.3 (-14%)	21.6 (3%)	19.9 (11%)	0.48	0.48 (1%)	0.54 (13%)	0.57 (20%)
Hjortnes	51.7	55.9 (-8%)	34.4 (34%)	31.7 (39%)	0.29	0.30 (5%)	0.38 (31%)	0.36 (25%)
Season 2								
RV4	18.2	16.7 (8%)	18.9 (-3%)	16.4 (10%)	0.52	0.59 (14%)	0.57 (10%)	0.62 (20%)
Åkeberg.	15.3	13.2 (14%)	14.1 (7%)	12.6 (18%)	0.45	0.64 (42%)	0.64 (41%)	0.69 (52%)
Hjortnes	35.0	25.8 (26%)	28.2 (19%)	26.5 (24%)	0.27	0.40 (46%)	0.39 (44%)	0.41 (50%)

Table D.4. RMSE and COR for NO_2 using the mean estimate. The training data sets in group 1 and group 2 contain two stations.

NO2	RMSE				COR			
Season 1	Orig.	Gr. 1	Gr. 2	All st.	Orig.	Gr. 1	Gr. 2	All st.
RV4	27.9	31.2 (-12%)	36.4 (-31%)	28.8 (-3%)	0.59	0.57 (-4%)	0.60 (2%)	0.62 (5%)
Åkeberg.	28.5	33.9 (-19%)	45.9 (-61%)	35.5 (-25%)	0.50	0.54 (8%)	0.57 (12%)	0.60 (19%)
Hjortnes	38.8	45.4 (-17%)	49.7 (-28%)	39.6 (-2%)	0.45	0.41 (-10%)	0.50 (11%)	0.49 (8%)
Season 2								
RV4	28.6	25.7 (10%)	27.0 (6%)	24.6 (14%)	0.63	0.65 (3%)	0.66 (4%)	0.68 (7%)
Åkeberg.	31.6	22.0 (30%)	29.4 (7%)	27.3 (14%)	0.64	0.70 (9%)	0.69 (7%)	0.71 (10%)
Hjortnes	32.1	30.5 (5%)	30.4 (5%)	28.5 (11%)	0.63	0.63 (0%)	0.63 (0%)	0.67 (6%)