Model selection and verification for ensemble based probabilistic forecasting of air pollution in Oslo, Norway

Sam-Erik Walker*
Dept. of Mathematics, Univ. of Oslo, Oslo, Norway – samerikw@math.uio.no

Gudmund Horn Hermansen
Dept. of Mathematics, Univ. of Oslo, Oslo, Norway – gudmunhh@math.uio.no

Nils Lid Hjort
Dept. of Mathematics, Univ. of Oslo, Oslo, Norway – nils@math.uio.no

Abstract

In this paper, we discuss building time series models for forecasting of air pollution during wintertime conditions in Oslo, Norway, using ensembles of air pollution model data. Since such ensembles becomes increasingly available as part of regular air quality forecast modelling, it is important to build properly calibrated statistical models utilising such data. In particular, we focus on model selection using the Akaike and Bayesian information criteria, and verification of the forecasts using Probability Integral Transform (PIT) histograms and Brier scores. Three time series models are considered, using ensemble mean values as a primary covariate in a linear regression setting explaining observations, and modelling the residual errors as an autoregressive process, using either a constant variance; a time-varying (heteroscedastic) variance only depending on the ensemble variances; or as a combination of both. We show that for the limited, although representative, data analysed, the model incorporating both terms, seems to have an edge according to the model selection criteria and forecast verification tools used. Finally, we briefly discuss the possibility of introducing more focused model selection criteria for these types of models and data.

Keywords: Time series; Information criteria; Brier score; Air quality.

1. Introduction

In this paper, time series models for 24-hour (next day) forecasting of air pollution (mainly from traffic and domestic heating) during wintertime conditions in Oslo, Norway are discussed, using ensembles of air pollution dispersion model data. The data are taken from the recent EU FP7 project UncertWeb (Walker et al., 2013), and consist of modelled ensembles of ground level hourly average concentrations of nitrogendioxide (NO_2) in Oslo for a period in January 2011. The model runs are based on the European Centre for Medium-range Weather Forecasts (ECMWF)[1] 50 ensemble member forecasts; MACC (Monitoring Atmospheric Composition and Climate)[2] regional scale ensemble forecasts; and ensembles of urban emission data. Measurements of NO_2 at a single but representative station in Oslo (Kirkeveien) are used as observational data.

2. Data and models

Fig. 1 shows time series of 24-hour (next day) ensemble air pollution model forecasts and observations of NO_2 at station Kirkeveien in Oslo for the period 3 – 15 January 2011 (312 hour-values). In the figure, the blue curve represents observations, while the red and orange curves represent the mean and median values of the 50 ensemble predictions, respectively. The lower and upper green curves indicate 90% central prediction intervals, based on the 0.05 and 0.95 quantiles of the ensemble data. The individual ensemble predictions are shown by the grey curves.

[1] http://www.ecmwf.int
[2] http://www.gmes-atmosphere.eu

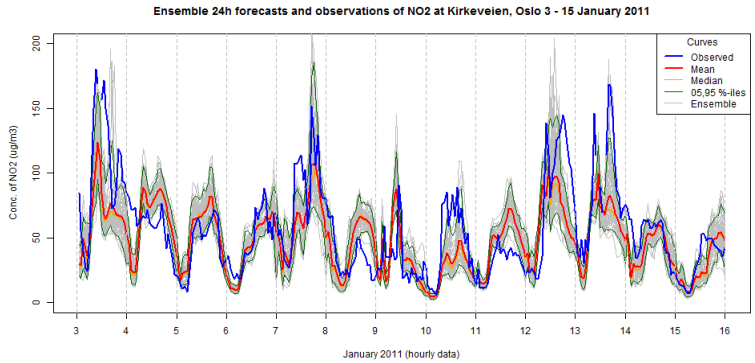

Fig. 1: Ensemble 24-hour (next day) forecasts and observations of NO_2 at station Kirkeveien in Oslo for the period 3 – 15 January 2011 (312 hour-values). Unit: μgm^{-3}.

Generally, there is a good correspondence between ensemble forecasts and observations, although when viewed as a probability distribution, the ensemble forecasts are clearly not calibrated, being somewhat under-dispersed and biased as compared with the observations.

Fig. 2 shows histograms of the ensemble mean values (left) and the observations (right).

As seen from the figure, both distributions are skewed to the right (this is typical for air pollution data), here most notably for the observations.

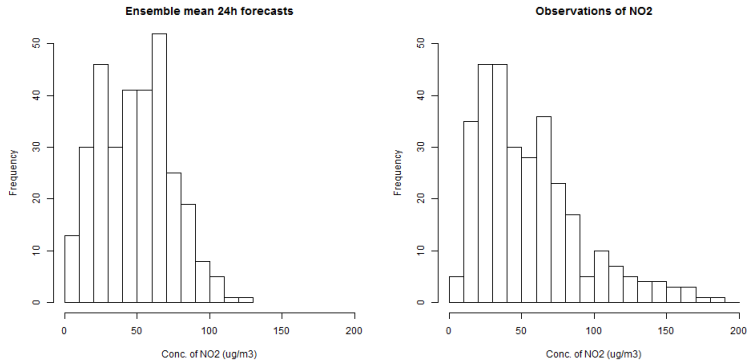

Fig. 2: Histograms of ensemble mean values (left) and observations (right) of NO_2 at station Kirkeveien in Oslo for the period 3 – 15 January 2011 (312 hour-values). Unit: μgm^{-3}.

Fig. 3 shows scatter plots of observations (y) vs. ensemble mean values (x) (left), and ensemble standard deviations (y) vs. ensemble mean values (x) (right).

As seen from the figure (left), there is a clear linear relationship between observations and ensemble mean values, as expected. The spread in the observations tend to increase with increasing level of the ensemble mean values. Such a level-spread relationship is very typical for air pollution data, and can also be seen in the figure to the right, where ensemble standard deviations also increase with ensemble mean values, with some notable deviations.

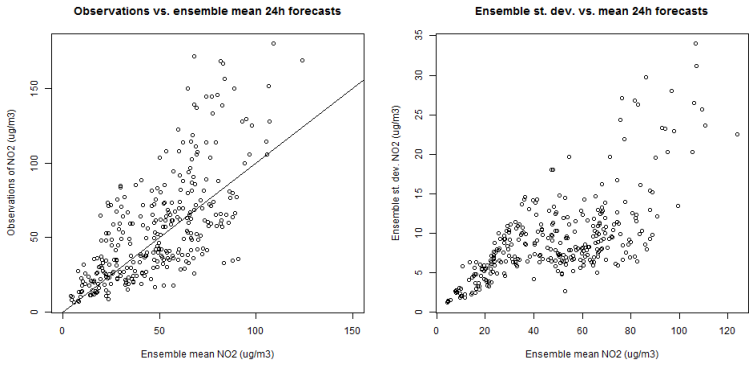

Fig. 3: Scatter plots of observations of NO_2 (y) vs. ensemble mean values (x) (left) (with a 45° line); and ensemble standard deviations (y) vs. ensemble mean values (x) (right). Unit: μgm^{-3}.

Based on this, we suggest using the available ensemble data to develop a time series forecasting model based on modelling the relationship between observations (response variable) and ensemble mean and standard deviation values (covariates) as follows:

$$y_t^{(\lambda)} = \beta_0 + \beta_1 x_t^{(\lambda)} + \sigma_t \varepsilon_t;\ \varepsilon_t = \phi \varepsilon_{t-1} + \eta_t;\ \eta_t \sim N(0,1);\ t = 1,...,T\ . \tag{1}$$

Here $y_t^{(\lambda)}$ represents a Box-Cox transformed value of the observation y_t at time (hour) t, with $t = 1,...,T$, where T is the total number of hours. The Box-Cox transformation with parameter λ is defined by

$$y_t^{(\lambda)} = \begin{cases} \left(y_t^{\lambda} - 1\right)/\lambda + \lambda & \text{for} \quad 0 < \lambda \le 1 \\ \ln y_t & \text{for} \quad \lambda = 0 \end{cases}.$$

For $\lambda = 1$, $y_t^{(\lambda)}$ represents untransformed values, i.e. $y_t^{(1)} = y_t$, while for $\lambda = 0$ we obtain log-transformed values. In (1), β_0 and β_1 are coefficients in a linear regression relationship between $y_t^{(\lambda)}$ and $x_t^{(\lambda)}$, where $x_t^{(\lambda)}$ represents the ensemble mean of Box-Cox transformed ensemble values at time (hour) t, for $t = 1,...,T$, using the same transformation, with parameter λ, as for the observations. Transforming both observations and ensemble data helps to make the error terms ε_t in (1) become symmetric and Gaussian. The error terms are modelled as an AR(1) process using the parameter ϕ, with η_t representing independent standard Gaussian noise terms. The standard deviations σ_t of the error terms are generally modelled in (1) as a time-varying quantity.

Three models, M0a, M0b and M1, are considered here, based on the following three ways of defining σ_t in (1):

$$\text{M0a: } \sigma_t^2 = \sigma^2;\quad \text{M0b: } \sigma_t^2 = \gamma^2 s_t^2;\quad \text{M1: } \sigma_t^2 = \sigma^2 + \gamma^2 s_t^2 \tag{2}$$

where σ and γ are two additional parameters to be estimated from data (using maximum likelihood together with the other parameters, including the parameter λ, in (1)), and where s_t, for $t = 1,...,T$, represents hourly standard deviation values of the Box-Cox transformed ensemble values (using the

same parameter λ). This allows for defining time-varying variance in the models M0b and M1, using standard deviation values of the transformed ensemble, while model M0a uses a constant variance for all hours. The idea is to investigate if there is any information left in the ensemble variances after the Box-Cox transformation (which generally tend to stabilize the variances of the error terms in (1)), which can be be exploited by models M0b and M1.

The forecasting model defined in (1)-(2) above is similar to a model suggested for post-processing of meteorological ensemble data given in Gneiting (2014). Our model differs in that we use a Box-Cox transformation for correcting the skewness of the air pollution data.

3. Results

Estimates of parameters with standard errors, together with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, for each of the three models M0a, M0b and M1, based on ensemble 24-hour forecasts and observations of NO_2 at station Kirkeveien in Oslo for a training period 3 January – 8 January 2011 (144 hour-values) is shown in Table 1.

Table 1: Estimated parameters with standard errors, and AIC and BIC values, for each of the three models M0a, M0b and M1 based on ensemble 24-hour forecasts and observations of NO_2 at station Kirkeveien in Oslo for a training period 3 January – 8 January 2011 (144 hour-values).

n=144	**Model M0a**		**Model M0b**		**Model M1**	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
β_0	2.857	0.904	5.618	1.058	2.755	1.059
β_1	0.533	0.105	0.542	0.062	0.544	0.105
ϕ	0.873	0.043	0.768	0.054	0.872	0.042
λ	0.203	0.095	0.505	0.066	0.188	0.128
σ	0.510	0.194	-	-	0.342	0.180
γ	-	-	1.577	0.215	0.309	0.165
AIC	-1120.0		-1194.9		-1110.7	
BIC	-1134.9		-1209.8		-1128.5	

For all models, the regression parameter $\beta_1 > 0$ as expected, and the error terms ε_t are positively correlated, with ϕ in the range 0.77 – 0.87. For models M0a and M1, the Box-Cox transformation is not far from being logarithmic, with $\lambda \approx 0.2$, while for model M0b, $\lambda \approx 0.5$, corresponding to a square-root transformation. According to both the AIC and the BIC criterion, model M1 is the best model, with highest AIC and BIC values. In this model, variance is modelled as a combination of a constant variance σ^2, and a time-varying ensemble based variance $\gamma^2 s_t^2$, with about equal weights given to both here, σ and γ having values 0.342 and 0.309 respectively (the s_t values have here been scaled to have mean value equal to one in order to make it possible to compare σ and γ).

Fig. 4 shows probabilistic (marginal) 24-hour forecasts based on model M1 and observations of NO_2 at station Kirkeveien in Oslo for the test period 9 – 15 January 2011 (168 hour-values). In the figure, the blue curve represents observations, while the red and orange curves represent the mean and median values of the predictive distributions, respectively. The lower and upper green curves indicates 90% central prediction intervals, based on the 0.05 and 0.95 quantiles of the predictive distributions, while corresponding 50% central prediction intervals, based on the 0.25 and 0.75 quantiles, are indicated by the two dotted green curves. As we see from the figure, observations seem to be captured reasonably well by the probabilistic forecasts during this period, with actual coverage probability being 85% of observations within the 90% prediction intervals.

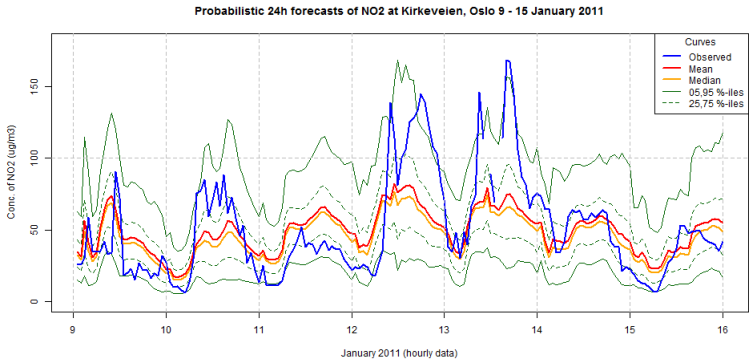

Fig. 4: Probabilistic 24-hour forecasts based on model M1 and observations of NO_2 at station Kirkeveien in Oslo for the test period 9 – 15 January 2011 (168 hour-values). Unit: µgm^{-3}.

The Probability Integral Transform (PIT) histogram (Wilks, 2006) of probabilities $p_t = F_t(y_t)$, using the forecast distributions F_t and observations y_t as in Fig. 4, is shown in Fig. 5.

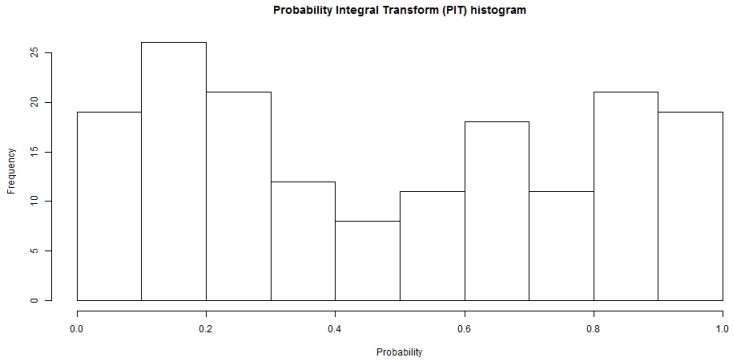

Fig. 5: Probability Integral Transform (PIT) histogram based on the forecast distributions and observations of NO_2 at station Kirkeveien in Oslo for the test period 9 – 15 January 2011 (168 values).

Ideally, if our probabilistic predictions are properly calibrated, this histogram should have a uniform shape. As seen from the figure, the histogram is slightly U-shaped, with fewer probability values (observations) in the middle part than on either side, which indicates that the probabilistic predictions are slightly too narrow. This is also evident if we look at Fig. 4, where we see that observations tend to fall on either side of the forecast distributions, with fewer cases in the middle. In fact, the actual coverage probability of observations is only 37% within the 50% central prediction intervals.

The Brier score[3] (Wilks, 2006), based on the binary event of exceeding (or not) a limit value of 100 µgm^{-3} each hour, which is half of the official hourly limit value for NO_2 in Norway (indicated by the horizontal dotted green line in Fig. 4), is shown for each of the three models M0a, M0b and M1 in Table 2. Also included is the reliability, resolution and climatology (uncertainty) decomposition parts of the Brier score. The latter corresponds to a climatological constant prediction each hour, using the actual frequency of observations exceeding the limit value during the test period, which is 11.3%, as a constant probability of exceedance each hour. As seen from the table, all models have quite low values (close to zero) of the Brier score, with model M1 slightly better than the other two. The reliability part is also close to zero for all models, indicating that all models are quite reliable in predicting

[3] Brier score = $\frac{1}{T}\sum_{t=1}^{T}\left\{I(y_t > y_{\lim}) - (1 - F_t(y_{\lim}))\right\}^2$ = Reliability – Resolution + Climatology

exceedances of the limit value during the test period (with model M0b the best model). Regarding the resolution part, where higher values are better, model M1 is slightly better than the other two.

Table 2: The Brier score, together with its reliability, resolution and uncertainty (climatological) decomposition parts, for each of the three models M0a, M0b and M1 based on the probabilistic 24-hour forecasts and observations of NO_2 at station Kirkeveien in Oslo, for the binary event of exceeding (or not) a limit value of 100 μgm^{-3} each hour during the test period 9 January – 15 January 2011.

n=168	**Model M0a**		**Model M0b**		**Model M1**	
Parameter	Estimate	% of Climatology	Estimate	% of Climatology	Estimate	% of Climatology
Brier score	0.0872	86	0.0905	89	0.0830	82
Reliability	0.0263	26	0.0236	23	0.0314	31
Resolution	0.0405	40	0.0345	34	0.0499	49
Climatology	0.1014	100	0.1014	100	0.1014	100

All three models improves on the climatological Brier score (0.1014) with scores of 86%, 89% and 82% of this value The reliability part is 26%, 23% and 31% of the climatological Brier score for the three models, which means that probabilistic predictions are 25-30% less reliable than the perfectly reliable climatological prediction. Resolution however, is 40%, 34% and 49% of the climatological score, implying a 35-50% better resolution than the climatological forecasts, which has no (zero) resolution. The best model is overall model M1 with 49% improved resolution.

5. Conclusions and further work

In this paper, three time series models based on ensemble model data for probabilistic forecasting of air pollution in Oslo, Norway were presented. Based on limited, although representative, observations at one station in Oslo (Kirkeveien), the model using both the ensemble means and variances seem to have an edge according to the Akaike and Bayesian model selection criteria, and also according to a calculated Brier score for the binary event of exceeding (or not) an hourly limit value of 100 μgm^{-3}. Going forward, we would like to explore the possible use of other and more specific criteria for model selection, most notably the use of the so-called Focused Information Criterion (FIC), the theory of which has recently been extended to include stationary time series models in Hermansen and Hjort (2014). Of particular research interest would be to extend the use of FIC to time series models that includes regression terms and time-varying (heteroscedastic) variance.

Acknowledgements

This work was done as part of the current FocuStat project at the Dept. of Mathematics at the Univ. of Oslo (http://www.mn.uio.no/math/english/research/projects/focustat/). The ensemble forecast data was kindly provided by the Norwegian Institute for Air Research (NILU) (http://www.nilu.no), via the UncertWeb project (http://www.uncertweb.org), supported by the European Commission under the 7th Framework Programme Theme FP7-ICT-2009-4. We also wish to thank the Agency for Urban Environment in the municipality of Oslo for providing the observational data.

References

Gneiting, T. (2014) Calibration of medium-range weather forecasts. ECMWF TM 719, 18 March 2014.

Hermansen, G. H., Hjort, N. L. (2015) Focused information criteria for regression models with dependent errors. Dept. of Mathematics, Univ. of Oslo.

Walker, S.-E., Denby, B., Pross, B. & Cornford, D. (2013). Validation of UncertWeb: Air quality forecasting. UncertWeb Consortium Deliverable 6.3. Birmingham, UK. http://www.uncertweb.org.

Wilks, D. S. (2006) Statistical Methods in the Atmospheric Sciences (2nd ed.). Academic Press, London.