Article

Predicting air quality from measured and forecast meteorological data: a case study in Southern Italy

Andrea Tateo [1], Vincenzo Campanaro [1], Nicola Amoroso[2,3], Loredana Bellantuono[3,4], Alfonso Monaco[3,5]*, Ester Pantaleo[3,5], Rosaria Rinaldi[6], Tommaso Maggipinto[3,5]

[1] Apulia Region Environmental Protection Agency (ARPA Puglia), C.so Trieste 27, Bari, 70126, Italy
[2] Dipartimento di Farmacia - Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Via A. Orabona 4, Bari, 70125, Italy
[3] Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Bari, Via A. Orabona 4, Bari, 70125, Italy
[4] Dipartimento di Scienze mediche di base, Neuroscienze e organi di senso, Università degli Studi di Bari Aldo Moro, Piazza G. Cesare 11, Bari, 70124, Italy
[5] Dipartimento Interateneo di Fisica M. Merlin, Universitá degli Studi di Bari Aldo Moro, Via G. Amendola 173, Bari, 70125, Italy
[6] Department of Mathematics and Physics E. De Giorgi, Universitá del Salento, via Arnesano, Lecce, 73100, Italy
* Correspondence: alfonso.monaco@ba.infn.it

1. Supplementary Material

1.1. Error reduction in WRF meteorological predictions

It is well known that meteorological forecasts are intrinsically biased, thus to reduce the prediction error we used a machine learning approach as suggested elsewhere [1]. For each variable, we estimated the prediction error defined as $VAR_{err} = VAR_{pred} - VAR_{obs}$ where VAR_{pred} is the predicted variable and VAR_{obs} the variable observed on ground.

To reduce the bias we used a temporal window of 30 days prior to the considered day. To estimate the prediction error for all variables we used a RF model; we used 11 features in this case: day, month, the 2 hourly cyclical component H_1 e H_2 defined by

$$H_1 = sin\left(\frac{h \cdot \pi}{24}\right)$$

$$H_2 = cos\left(\frac{h \cdot \pi}{24}\right)$$

and the 7 variables predicted by the WRF model. Thus VAR^*_{err}, the prediction error estimated by RF, can be used to correct the WRF forecasts: $VAR_{best} = VAR_{pred} - VAR^*_{err}$.

The performances have been estimated in terms of mean square error (MSE) and Correlation Coefficient. In Tables S1, S2, S3, and S4 the prediction performances of the meteorological variables with and without bias correction are compared. In general, our post processing approach reduces the forecast bias since, after its application, the MSE is close to zero. As regards the correlation between the predicted and observed values, we always obtain better performances or, when the correlation is already high (without post processing) as in the case of Temperature 2 m, comparable performances.

As an example, in figure S1 we report the MSE boxplots for the Wind Speed 10m on the domain d01 for the 1 to 24 hour forecasts. The boxplots show that the application of post processing in addition to reducing systematic error (zero-centered boxplot) also reduces the global error (narrower boxplot).

The same effect is observed for all the considered variables except for the 10m Wind Direction when, after post processing, a low correlation is observed. For this variable without correction the correlation was already very low (see figure S2).

In fact, as reported in table S2 for the wind direction 10 m the MSE after the correction is lower but the distribution, although more centered than zero, is not much narrower than

Wind Speed 10m							
1 to 24							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
-2.2	0.0	0.62	0.73	-1.9	0.0	0.61	0.73
25 to 48							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
-2.1	0.0	0.63	. 0.64	-1.8	0.0	0.63	0.64
49 to 72							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
-2.0	0.0	0.61	0.61	-1.7	0.0	0.60	0.60

Table S 1. Comparison between predictions of the wind speed 10m with and without bias correction on both domains, d01 and d02, in terms of MSE and correlation coefficient.

Wind Direction 10m							
1 to 24							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
40.0	-3.0	0.16	0.52	29.7	-3.2	0.25	0.57
25 to 48							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
41.0	-3.0	0.16	0.41	28.0	-1.7	0.28	0.49
49 to 72							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
40.9	-4.0	0.16	0.37	27.1	-3.0	0.27	0.45

Table S 2. Comparison between the wind direction 10m predictions with and without bias correction on both domains, d01 and d02, in terms of MSE and correlation coefficient.

those without post processing (figure S3). However, the effect of our post processing is to reduce the not very large prediction errors instead it does not seem to affect the gross prediction errors.

This is confirmed by the estimate of the Direction Accuracy (DACC) which calculates the percentage of errors in the forecast of the wind direction less than a reference angle. As shown in figure S4 it can be observed that the application of post processing further reduces the percentage of errors in absolute value below 120°.

Temperature 2m							
1 to 24							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
0.3	-0.1	0.95	0.97	0.4	-0.1	0.95	0.96
25 to 48							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
1.0	-0.1	0.95	0.95	1.1	-0.0	0.95	0.95
49 to 72							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
1.1	-0.1	0.94	0.94	1.2	-0.1	0.94	0.94

Table S 3. Comparison between the temperature 2m predictions with and without bias correction on both domains, d01 and d02, in terms of MSE and correlation coefficient.

Humidity 2m							
1 to 24							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
-20.3	0.1	0.58	0.77	-21.4	0.1	0.56	0.76
25 to 48							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
-23.8	0.1	0.57	0.72	-24.5	-0.1	0.56	0.71
49 to 72							
d01				d02			
MSE		Correlation		MSE		Correlation	
No Corr	Corr	No Corr	Corr	No Corr	Corr	No Corr	Corr
-24.0	0.1	0.56	0.69	-24.6	-0.0	0.54	0.69

Table S 4. Comparison between the humidity 2m predictions with and without bias correction on both domains, d01 and d02, in terms of MSE and correlation coefficient.

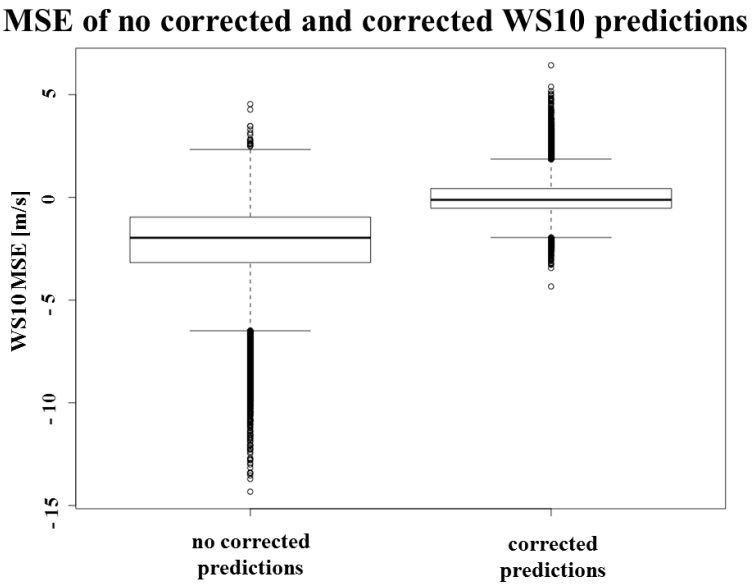

Figure S 1. Comparison between corrected and no corrected MSE of Wind Speed 10m predicted.

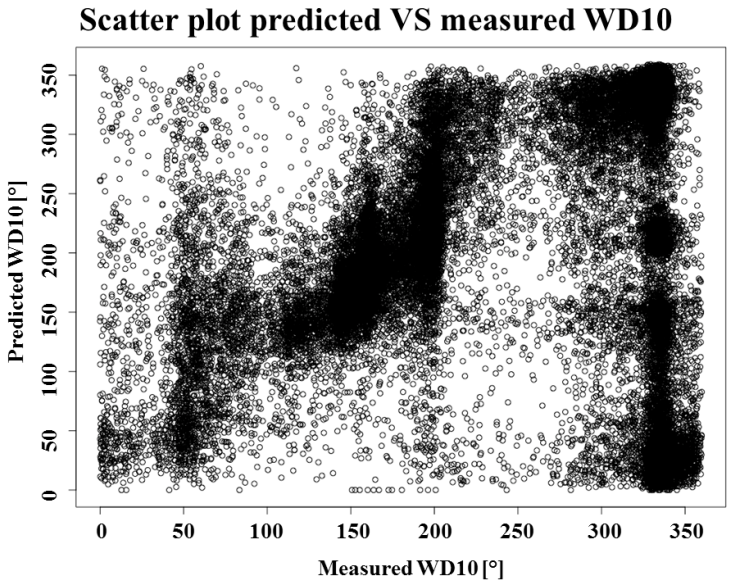

Figure S 2. Scatter plot to compare the wind direction predicted by the WRF model without post processing and the wind direction measured by the ground station.

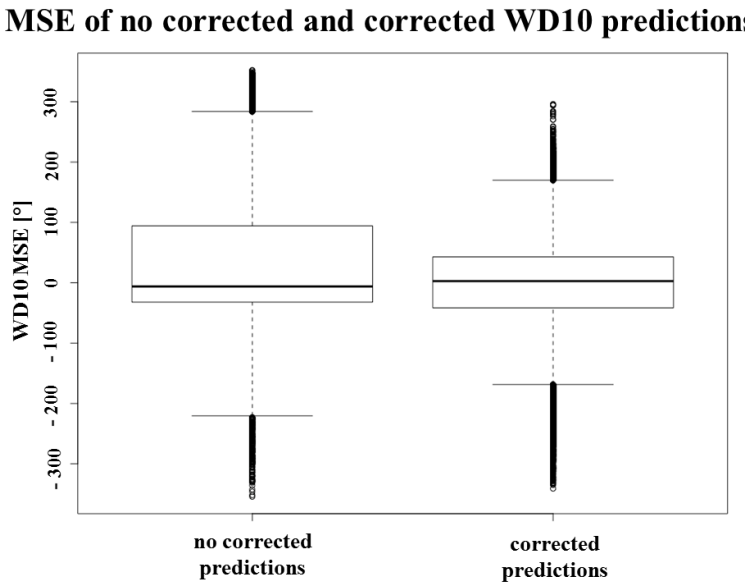

Figure S 3. Comparison between the MSE of the WD10 predicted with and without bias correction.

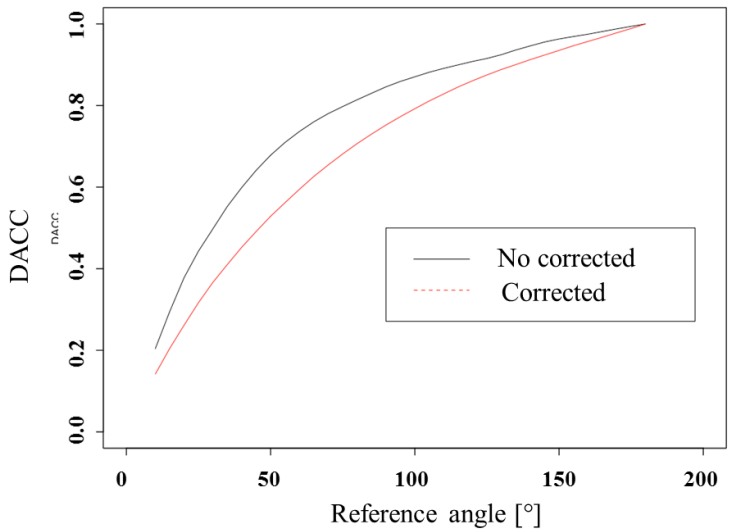

Figure S 4. Comparison between the Direction Accuracy for the prediction of Wind Direction 10 m with and without bias correction.

1.2. *Random Forest predictions performance for NO_2, CO, PM_{10}, and $PM_{2.5}$*

According to the results showing that RF models higher than 0.7) for NO_2, CO, PM_{10}, and $PM_{2.5}$, the relative scatter plots reported in the figures S5, S6, S7, confirm the goodness of RF predictions when compared with measured values. The results concern both higher and lower resolution for the first, second and third days. Moreover, the NO_2 scatter plots, being wider, confirm the lower RMSE if compared to that of the other three pollutants. Nonetheless, the time series reported in the figures S8, S9, S10, show how similar is the temporal trend of the predicted (black line) and measured (red line) values.

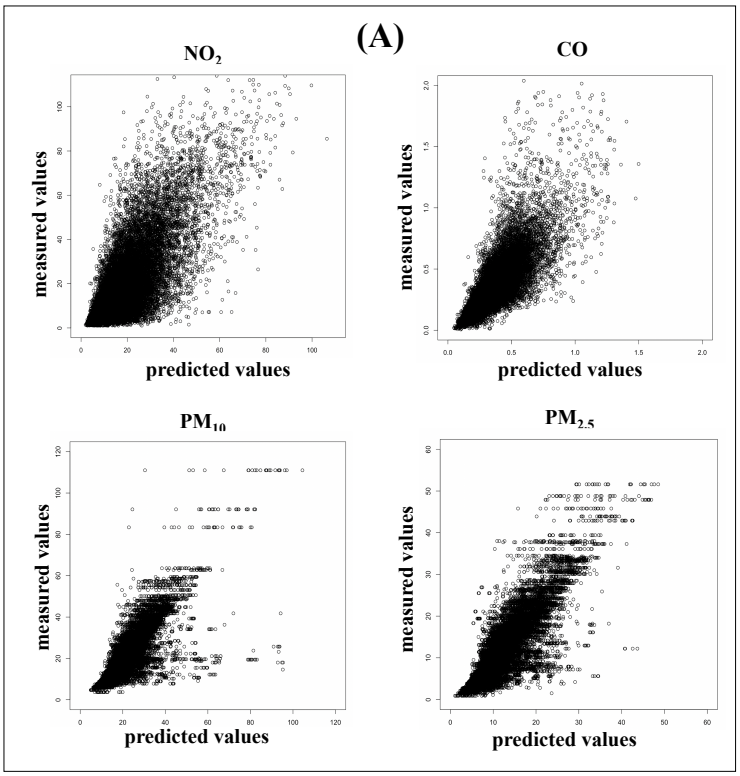

Figure S 5. Scatter plot to compare predicted and measured values for NO_2, CO, PM_{10}, $PM_{2.5}$ for the first day (+24) on the lower resolution domain d01 (A) and on higher resolution domain d02 (B).

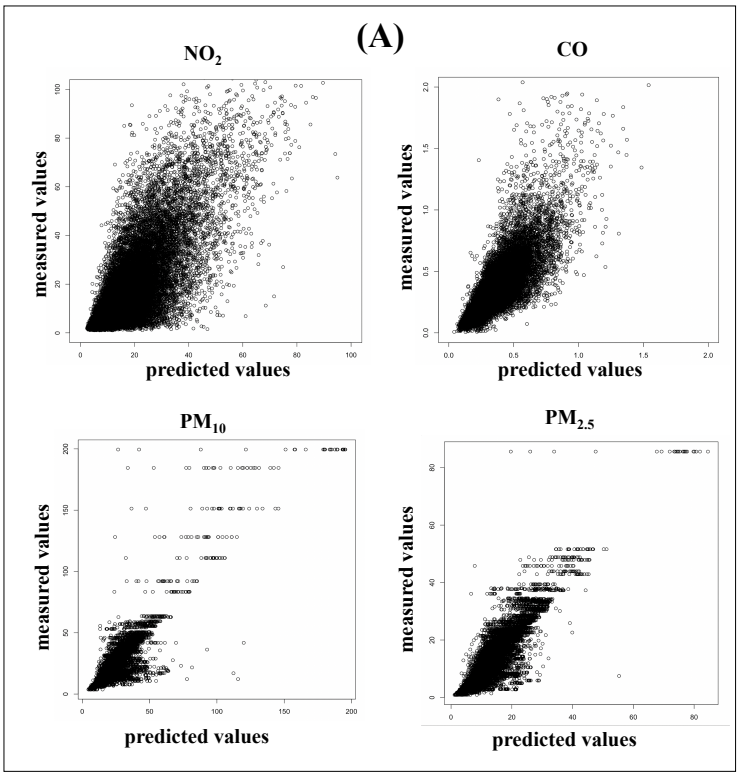

Figure S 6. Scatter plot to compare predicted and measured values for NO_2, CO, PM_{10}, $PM_{2.5}$ for the second day (+48) on the lower resolution domain d01 (A) and on higher resolution domain d02 (B).

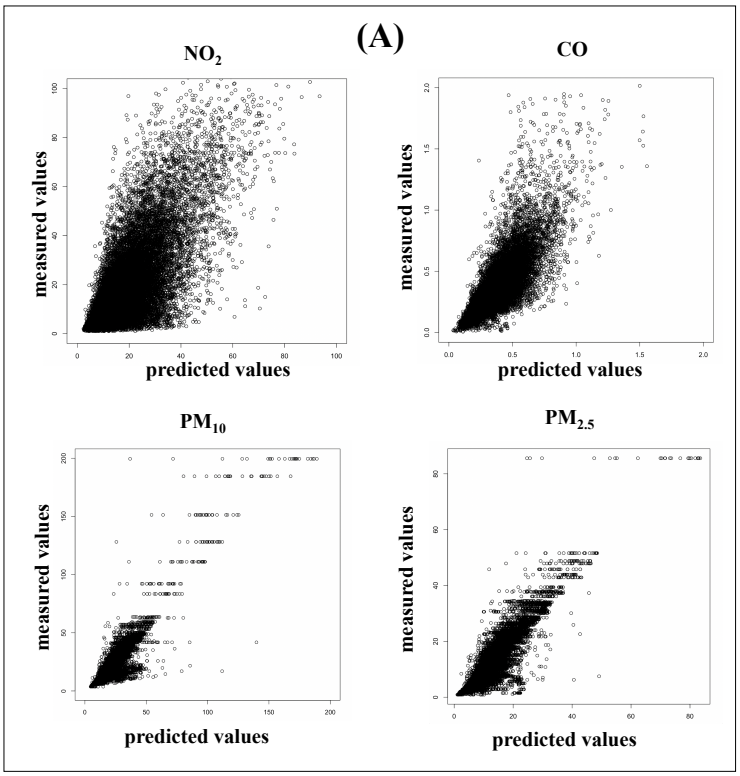

Figure S 7. Scatter plot to compare predicted and measured values for NO_2, CO, PM_{10}, $PM_{2.5}$ for the third day (+72) on the lower resolution domain d01 (A) and on higher resolution domain d02 (B).

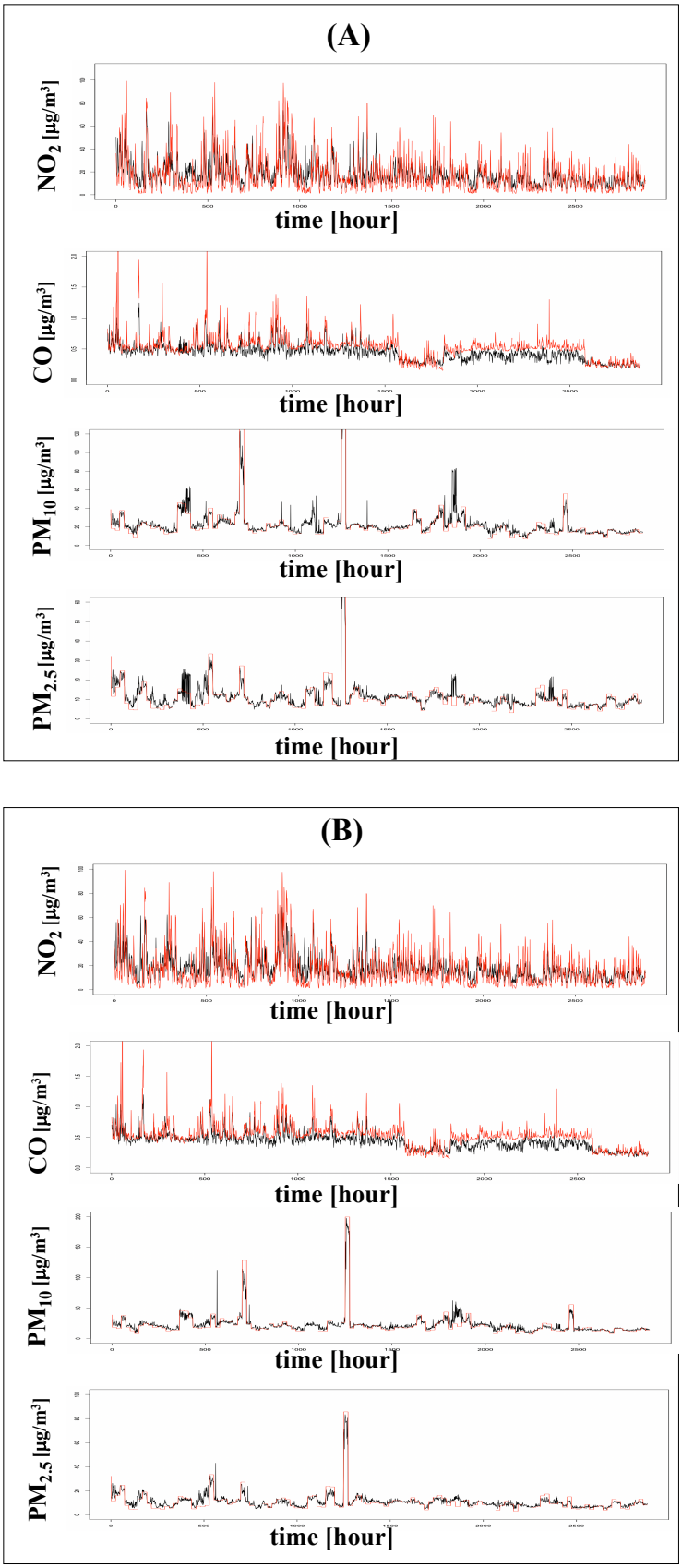

Figure S 8. Time series for NO_2, CO, PM_{10}, $PM_{2.5}$ for the first four months of the considered analysis period. The red line concerns the measured values. The black line concerns the RF predictions for the first day (+24). Panel (A) is for the lower resolution domain d01, while panel (B) is for the higher resolution domain d02.

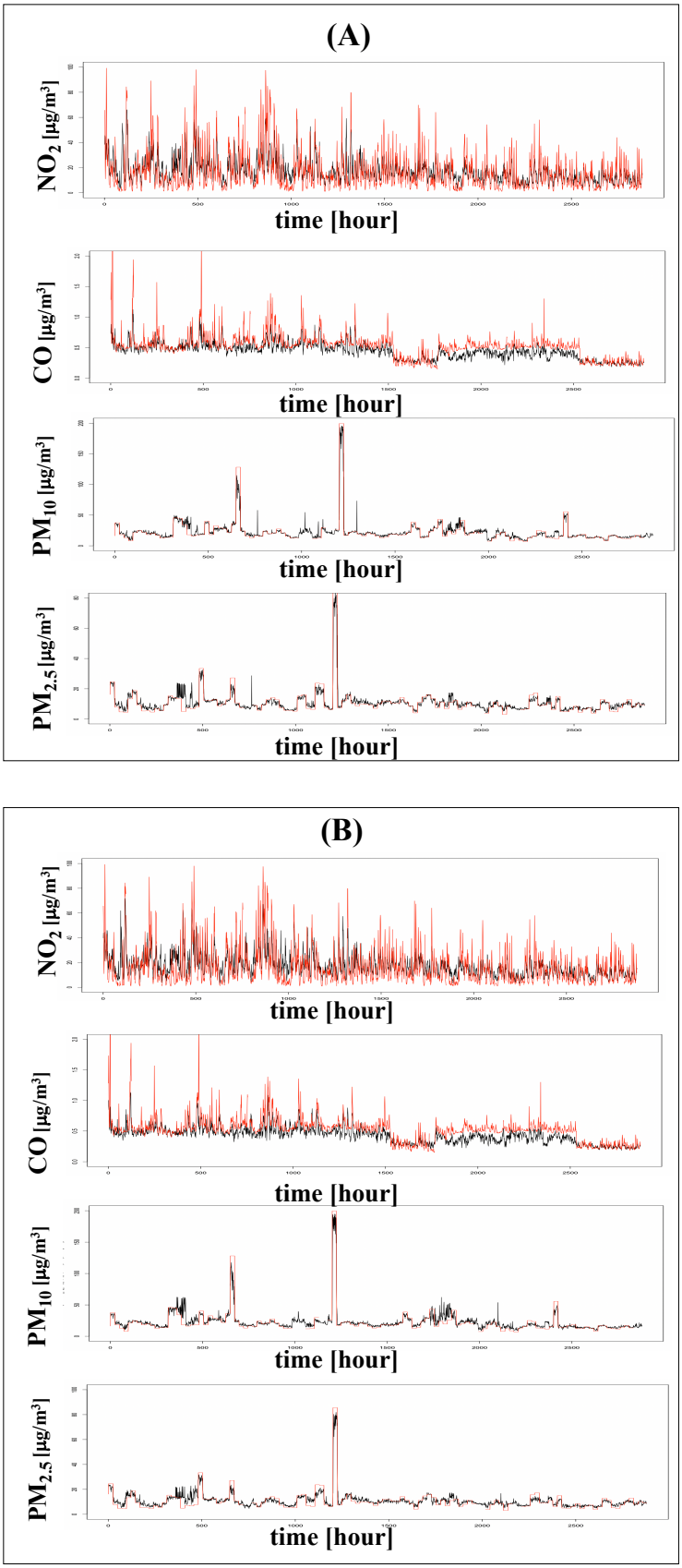

Figure S 9. Time series for NO_2, CO, PM_{10}, $PM_{2.5}$ for the first four months of the considered analysis period. The red line concerns the measured values. The black line concerns the RF predictions for the second day (+48). Panel (A) is for the lower resolution domain d01, while panel (B) is for the higher resolution domain d02.

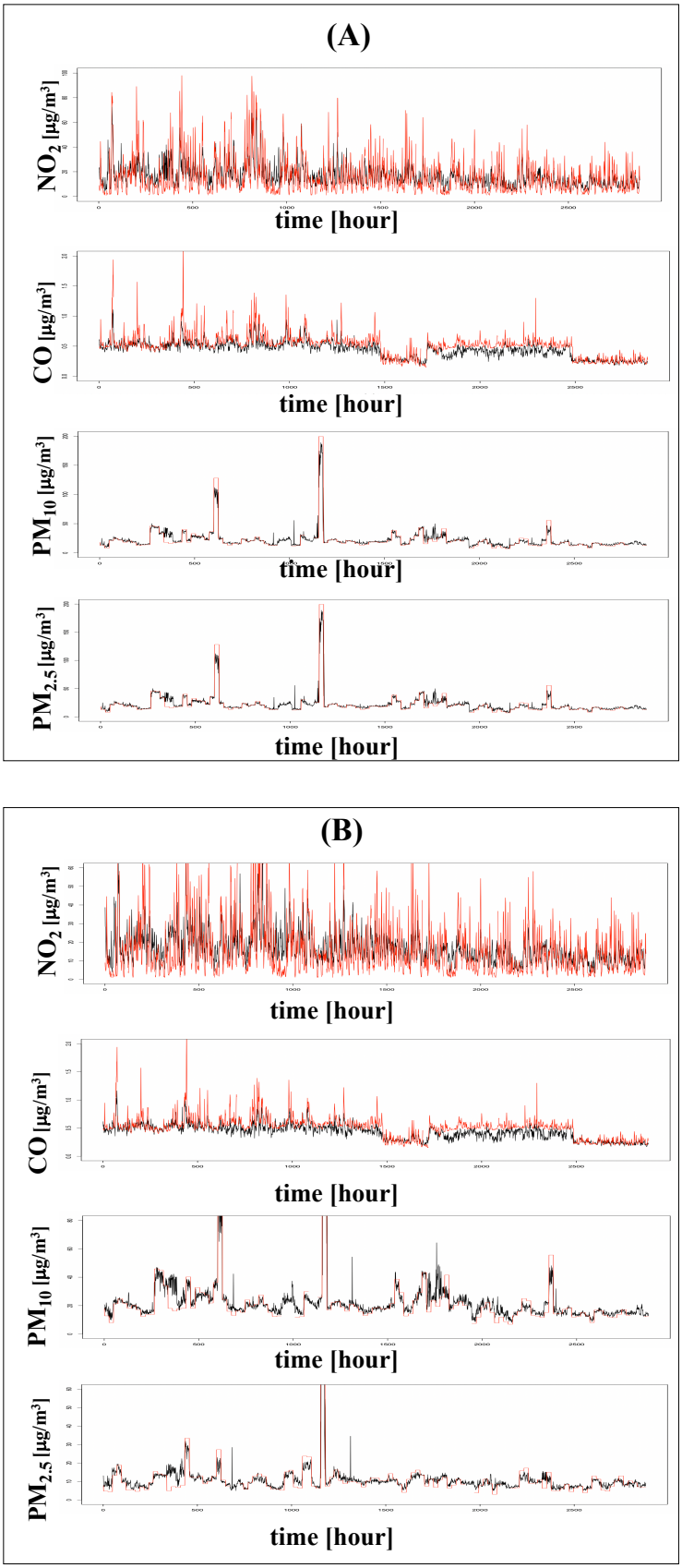

Figure S 10. Time series for NO_2, CO, PM_{10}, $PM_{2.5}$ for the first four months of the considered analysis period. The red line concerns the measured values. The black line concerns the RF predictions for the third day (+72). Panel (A) is for the lower resolution domain d01, while panel (B) is for the higher resolution domain d02

1. Tateo, A.; Miglietta, M.M.; Fedele, F.; Menegotto, M.; Monaco, A.; Bellotti, R. Ensemble using different Planetary Boundary Layer schemes in WRF model for wind speed and direction prediction over Apulia region. *Advances in Science and Research* **2017**, *14*, 95–102.