BADANIA FIZJOGRAFICZNE
R. VIII – SERIA A – GEOGRAFIA FIZYCZNA (A68) str. 033–046

DOI 10.14746/bfg.2017.8.3

VERIFYING RELIABILITY OF SELECTED METEOROLOGICAL ELEMENTS IN THE GFS AND WRF MODELS IN POZNAŃ

SEBASTIAN KENDZIERSKI

Department of Climatology, Faculty of Geographical and Geological Sciences,
Adam Mickiewicz University in Poznań,
B. Krygowskiego 10, 61-680 Poznań, Poland

Abstract: The article presents the results of a study comparing the reliability of the GFS and WRF numerical models using different spatial resolutions. The study was carried out on the basis of a series of model and observational data for a period of 13 months from January 2015 to January 2016 with computational grid coordinates corresponding to the location of Poznań. Forecast quality was compared for basic meteorological parameters – air temperature, atmospheric pressure and wind speed. The quality of forecasts for the meteorological parameters included in the study decreased systematically in subsequent time steps in the case of both the WRF and the GFS models. It improved in the initial hours of simulation after a higher resolution was applied for the computational grid in the WRF model (1,75 km) as can be seen from the minimal Mean Error values and very high correlation. For forecast time t = 0–60 h, WRF generated better results than GFS, especially when it comes to the quality of forecasting the atmospheric pressure field.

Keywords: forecast verification, GFS, WRF, air temperature, air pressure, wind speed, Poznań

INTRODUCTION

Weather forecasts are of key importance for many aspects of human activity. Although many of those are unquantifiable (e.g. problems related to selecting adequate clothes), they are crucial from the perspective of the economy, impacting such sectors as energy, transport (land, maritime and air), agriculture, tourism, organisation of mass events and other undertakings which depend on weather conditions. Thus, even a small improvement in the quality of forecasts translates into more time for potential response as well as a better ability to adapt to and counter the effects of adverse atmospheric phenomena, bringing about a reduction in the costs of covering potential losses. Individual benefits related to weather forecasts are estimated at $2–10/month per person (Frei et al. 2014). The figures are many times higher in different sectors of the economy. In Switzerland, weather forecasts help reduce transport expenditure by 65.7–79.8 million Swiss francs annually (Frei et al. 2014). In other sectors of the economy the figures are often much higher (Frei et al. 2010). Improvements in the quality of weather forecasts generate long-term savings. In the US energy sector, for

example, they are estimated at about $1 million annually for each 1% of improvement in the forecast of air temperature (Teisberg et al. 2005).

Consequently, it is extremely important to implement evaluation processes and operational reviews of existing forecast models in order to select the best solutions. The process of verification consists in comparing forecasts with observations. In other words, numerical simulation is matched to the actual atmospheric conditions in a given period. Weather forecast verification should improve the quality of forecasts for end users through reviewing applied algorithms. Regardless of its focus, end user should also be aware of the forecast's limitations and the most probable range of possible results.

Currently, weather is forecast with the use of different numerical models. A large number of them, however, are based on boundary conditions received from the Global Forecast System (GFS). The National Oceanic and Atmospheric Administration (NOAA) publishes a basic verification of the model in operational time. In the case of the Weather Research and Forecasting Model (WRF), the situation is more complex (Skamarock et al. 2005). Due to the great number of possible configurations of this mesoscale model and its versatile applications (Werner et al. 2015; Rzepa 2012; Kryza et al. 2017), publications discussing it are less synthetic and often focus on the reliability of the model for specific parameters and phenomena, for example related to atmospheric convection (e.g. Schwartz et al. 2009). As regards the territory of Poland, there wasn't much attention given to the verification and comparison of the two models, whilst some available studies (Rzepa 2012) are based on their obsolete versions. Hence, this article tries to bridge the gap by verifying the GFS and WRF numerical models on the basis of one measurement point and a longer measurement series.

DATA AND RESEARCH AREA

The study is based on hourly weather values – air temperature, atmospheric pressure and wind speed – generated by two numerical models which were compared in the period from January 2015 to January 2016. Data from the GFS model developed by the US meteorological service was downloaded in the version in which it is posted on *meteomodel.pl*. Currently, the GFS model uses the resolution of 25 km. GFS has 64 unevenly distributed vertical layers. The analysis focused on its simulations from 00 and 12 UTC with maximum forecast time horizon of 240 hours.

Data from the high-resolution WRF model based on GFS boundary conditions were made available by the administrator of *meteoprognoza.pl* where forecasts of the model are published. The WRF model was activated in 4 configurations corresponding to increasing density of subsequent computational domains.

The WRF 1,75 km model was activated once per 24 hours (starting at 18.30 local time, GFS GRIBs from 12 UTC downloaded from 018 h, forecast from 06 UTC of the following day). The 3 km model was activated twice per 24 hours. The WRF 3 km model was active twice per 24 hours (1: starting at 6:15 local time, GRIBs from 18 UTC downloaded from 012 h, forecast from 12 UTC of the following day; 2: starting at 12:15 local time, GRIBs from 06 UTC downloaded from 018 h, forecast from 00 UTC of the following day). The WRF 12 km model was active twice per 24 hours (1: starting at 18:30 local time, GRIBs from 12 UTC downloaded from 018 h, forecast from 06 UTC of the following day; 2: starting at 6:30 local time, GRIBs from 06 UTC downloaded from 000 h, forecast from 12 UTC of the current day). The WRF 50 km model was active once per 24 hours (starting 18:30 local time, GFS GRIBs from 12 UTC, downloaded from 021 h, forecast from 03 UTC of the following day).

Observational data from the Poznań-Ławica weather station was obtained from the Institute of Meteorology and Water Management – National Research Institute. The results discussed in the article concern comparisons of forecasts generated by a given numerical model of a given resolution at computational grid points located nearest the Poznań-Ławica weather station (52°25'18"N, 16°49'25"E).

METHODS

Air temperature, atmospheric pressure and wind speed were verified on the basis of verification methods used for continuous meteorological parameters (Jolliffe et al. 2012). Among others, these include statistical measures assessing deviations of the model series from expected values. These are: mean error (ME) (1), mean absolute error (MAE) (2), root mean square error (RMSE) (3), mean squared error (MSE) (4), BIAS (b) (5) and the Pearson correlation coefficient (r) (6).

Mean error (1) measures the mean difference between forecast and observational data (Nurni 2003). It determines the mean forecast error assuming values from minus to infinity, the perfect score being ME = 0.

$$ME = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i) \tag{1}$$

Mean Absolute Error (2) measures the average magnitude of forecast errors in a given dataset and therefore is a scalar measure of forecast accuracy (Nurni 2003). Its perfect score is MAE = 0, potential values ranging from 0 to plus infinity.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i| \tag{2}$$

Root mean square error (3) is a quadratic scoring rule which gives the average magnitude of errors, weighted according to the square of the error (Stanski et al. 1989).It measures the mean magnitude of forecast error. Its best value is RMSE = 0, whilst potential values range from 0 to infinity.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}(\mathrm{F_i} - \mathrm{O_i})^2} \tag{3}$$

Mean squared error (4) is the squared difference between forecasts and observations. Due to the second power, the MSE and RMSE are much more sensitive to large forecast errors than the MAE (Nurni 2003). The perfect score for MSE is MSE = 0 with potential values ranging from 0 to plus infinity.

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{F_i} - \mathrm{O_i})^2 \tag{4}$$

Bias (5) measures the difference between the mean value obtained from measurements and simulation values. It demonstrates how the average forecast magnitude compares to the average observed magnitude. It does not measure the magnitude of the errors. It does not measure the correspondence between forecasts and observations, i.e., it is possible to get a perfect score for a bad forecast if there are compensating errors (http://www.cawcr.gov.au/projects/verification/). The best score is b = 1, potential values ranging from minus to plus infinity.

$$\mathrm{b} = \frac{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{F_i}}{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{O_i}} \tag{5}$$

The last statistical indicator used in the study is the Pearson correlation coefficient (6). It measures how forecast values correspond to observations. Its perfect score is 1, potential values ranging from –1 to 1.

$$r = \frac{\sum (F - \overline{F})\,(O - \overline{O})}{\sqrt{\sum (F - \overline{F})^2}\,\sqrt{\sum (O - \overline{O})^2}} \tag{6}$$

where:
F – forecast,
O – observation,
N – sample size verification.

Verification results are presented as tables containing data for the entire time span of the forecast and all partial time steps. In order to facilitate the interpretation of obtained errors, the distribution of errors in subsequent hours of the time step was also presented by means of box-plot charts. The charts show the distribution of the difference between forecasts and observations, median, percentiles and extremes in subsequent hours of the time step. Successive marks on box-plot charts show percentile values at levels 0.1, 0.25, 0.5, 0.75 and 0.9. Statistical computations were carried out with the use of the *verification* (NCAR – Research Applications Laboratory 2015) and *hydroGOF* (Zambrano-Bigiarini 2014) packages for system R. The results section of the study has been divided into three parts presenting verification results for air temperature, atmospheric pressure and wind speed, in that order.

RESULTS

Air temperature

Air temperature is the most often verified parameter in weather forecasts. In the case of long-term forecasts, air temperature is verified to determine the tendency of potential changes over longer periods of time. In short-term forecasts, we verify the exact value we may expect. Given the above premises and the incremental increase of forecast errors over time, long-term forecasts are fraught with errors in longer time horizons. Table 1 shows calculated values for studied statistics with a partial forecast horizon of the GFS model.

Table 1. Statistical results for air temperature according to GFS model for individual forecast time steps

t	ME	MAE	RMSE	MSE	BIAS	Correlation
0–48 h	–0.1°C	1.4°C	1.8°C	3.2°C	0.95	0.98
49–96 h	–0.1°C	1.7°C	2.2°C	4.9°C	0.95	0.97
97–144 h	–0.2°C	2.2°C	2.9°C	8.2°C	0.94	0.94
145–192h	–0.3°C	2.8°C	3.6°C	13.1°C	0.94	0.91
193–240 h	–0.2°C	3.4°C	4.4°C	19.5°C	0.94	0.86

As mentioned above, the longer the time horizon of the forecast, the more pronounced its error. Over the studied period, the GFS model shows low ME (Mean Error) values at all hours of the time step, but MAE (Mean Absolute Error) values increase gradually with each step of the forecast. The lowest values of 1.1°C were obtained for data assimilation time (t = 0 h) whereas the highest values – 2.3°C – appeared at the end of the forecast (t = 240 h) (Fig. 1). Forecast errors increase visibly from hour 100 of the time step.

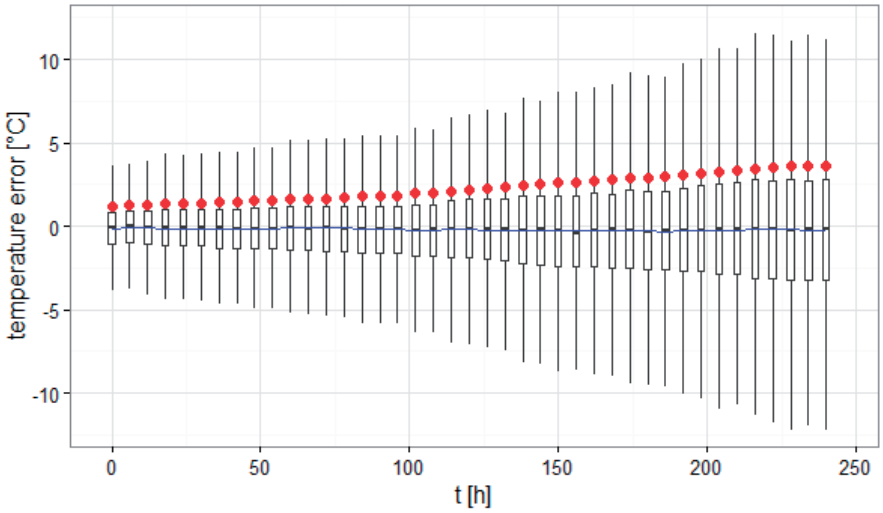

Fig. 1. Distribution of air temperature errors in individual forecast time steps for the GFS model

In the case of the WRF model, the size of the computational grid impacts the level of obtained forecast errors. In spite of more accurate representation of surface processes in smaller computational grids, the mean error was together with their resolution (ME is smallest for WRF-50 km). On the other hand, smaller computational grids translate into lower values of mean MAE, RMSE and MSE as well as better correlation between forecasts and observations (Tab. 2).

Table 2. Statistical results for air temperature according to the WRF model for forecast time step t = 0–60 h

Resolution	ME	MAE	RMSE	MSE	BIAS	Correlation
1.75 km	–0.7°C	1.5°C	2.0°C	4.4°C	0.93	0.97
3 km	–0.9°C	1.7°C	2.1°C	4.9°C	0.91	0.96
12 km	–0.2°C	1.7°C	2.2°C	5.0°C	0.98	0.95
50 km	–0.1°C	1.8°C	2.4°C	5.6°C	0.99	0.94

The distribution of errors depending on the time step of air temperature forecast is different in the cases discussed. The lowest error values are recorded in 1.75 km and 3 km grids in the initial hours of the forecast (Fig. 2). Mean Error and Mean Absolute Error values grow by leaps and bounds, without increasing linearly over longer periods. In extreme cases, maximum forecast errors reach almost 8°C. We may also observe a slight underestimation of temperature values in model simulation (i.e. the model is usually too "cold" compared to reality) especially in the case of WRF configuration for the 12 km grid. None of the applied computational grids showed error increase in subsequent time steps.

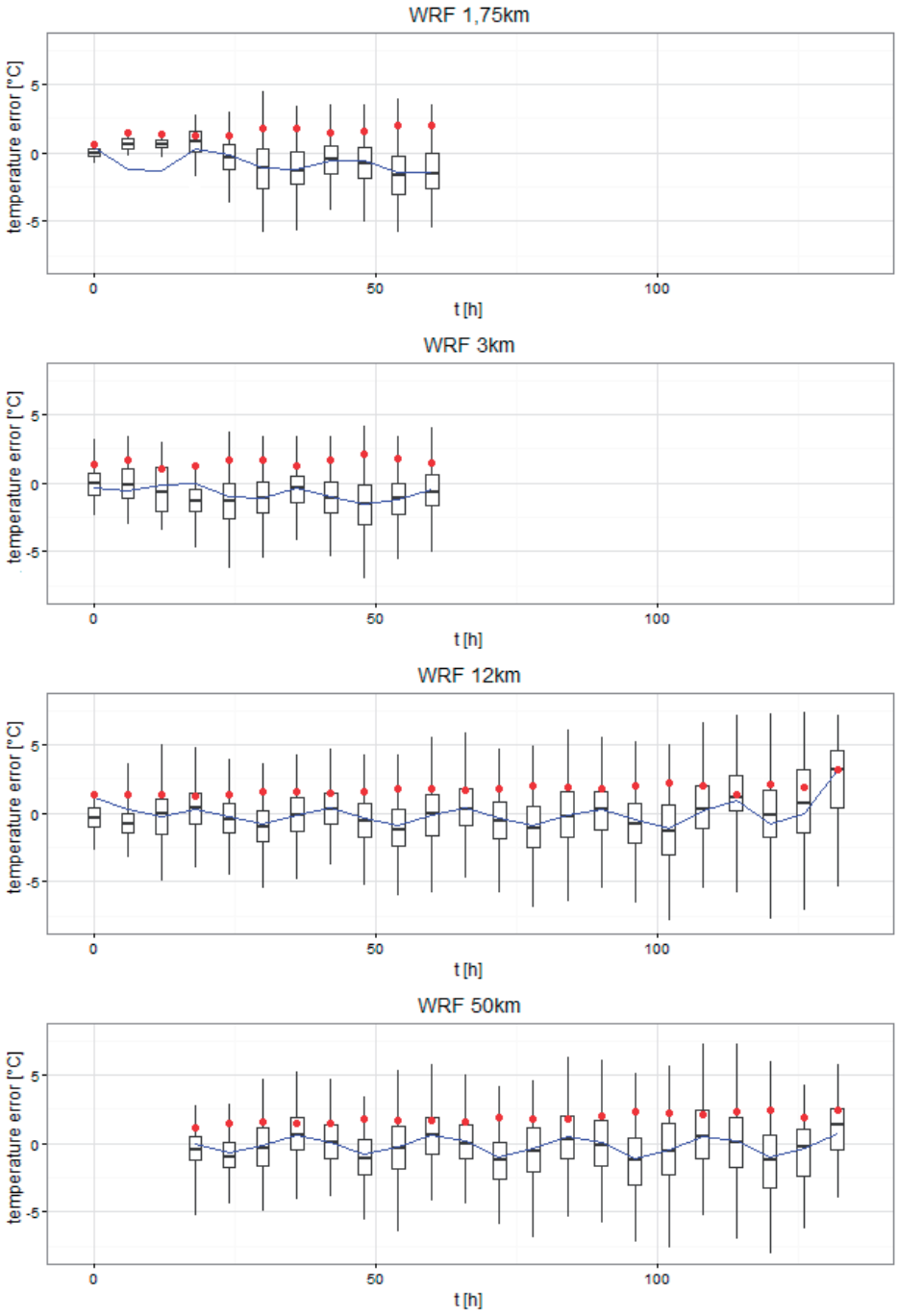

Fig. 2. Error forecast air temperature for WRF model in different resolutions in individual time steps

Air pressure

Table 3 shows verification results of atmospheric pressure (MSL) forecasts in Poznań. The GFS model displays major discrepancies between errors obtained for different forecast time horizons. In the initial hours of the time step, forecast error is relatively small, amounting to about 1.5 hPa for Mean Absolute Error. The values increase significantly after day four of the forecast, reaching maximum absolute error of 8 hPa for the time step of 240 h. In the case of atmospheric pressure forecast, Mean Error values stay at around 1 hPa regardless of the time horizon (Fig. 3). We need to highlight low correlation values of 0.55 for longer forecast times exceeding t = 192 h (Tab. 3).

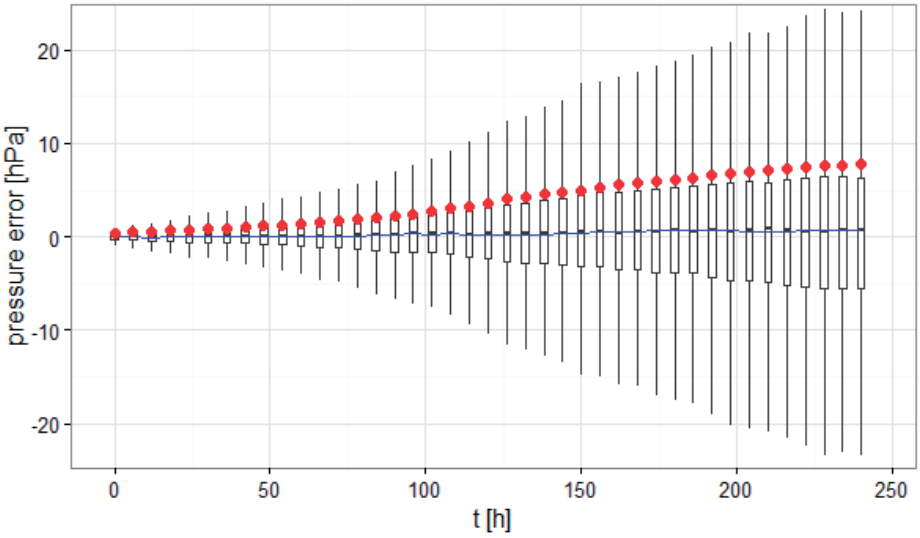

Fig. 3. Distribution of air pressure errors in individual forecast time steps for the GFS model

Table 3. Statistical results for air pressure according to GFS model for individual forecast time steps

t	ME	MAE	RMSE	MSE	BIAS	Correlation
0–48 h	0.0 hPa	0.7 hPa	1.1 hPa	1.3 hPa	1.00	0.99
49–96 h	0.1 hPa	1.8 hPa	2.4 hPa	5.7 hPa	1.00	0.97
97–144 h	0.2 hPa	3.6 hPa	4.9 hPa	23.8 hPa	1.00	0.87
145–192 h	0.6 hPa	5.7 hPa	7.7 hPa	59.1 hPa	1.00	0.69
193–240 h	0.7 hPa	7.2 hPa	9.6 hPa	91.9 hPa	1.00	0.55

Similar verification values for atmospheric pressure were obtained for the WRF model which produces very small errors for forecast times of 0–60 h

(Tab. 4). By increasing the resolution of the computational grid to 1.75 km, it was possible to eliminate the Mean Error almost totally. The same applies to Pearson correlation coefficient values of observed and forecast series which are very close to the levels of a perfect forecast. Other WRF configurations also display minimum levels of forecast errors.

Table 4. Statistical results for air pressure according to the WRF model for forecast time step t = 0–60 h

Resolution	ME	MAE	RMSE	MSE	BIAS	Correlation
1.75 km	0 hPa	1 hPa	1 hPa	2 hPa	1.00	0.99
3 km	1 hPa	1 hPa	2 hPa	2 hPa	0.99	0.98
12 km	0 hPa	1 hPa	2 hPa	3 hPa	1.00	0.97
50 km	–1 hPa	2 hPa	2 hPa	4 hPa	0.99	0.94

The distribution of the atmospheric pressure error expressed through other verification measures (see Tab. 4) in individual time steps is similar in different computational grids. It is smallest in the initial hours of the forecast. In the entire forecast, the longest period of low error levels can be observed in the WRF--12 model. In all cases, Mean Error values are stable, oscillating around 0 hPa (Fig. 4). Only WRF-50 km produces a negative Mean Error in the last hours of the forecast.

Wind speed

As wind speed is a very changeable meteorological element, it poses many problems when it comes to the possibility of forecasting it. We need to conclude, however, that Mean Error values diagnosed through verification results for the GFS model are relatively small, even if the correlation between forecasts and observations is quite poor (Tab. 5). On the other hand, diagnosed extreme values

Table 5. Statistical results for wind speed according to GFS model for individual forecast time steps

t	ME	MAE	RMSE	MSE	BIAS	Correlation
0–48 h	0.6 m/s	1.2 m/s	1.5 m/s	2.1 m/s	1.15	0.82
49–96 h	0.7 m/s	1.4 m/s	1.8 m/s	3.3 m/s	1.17	0.69
97–144 h	0.7 m/s	1.8 m/s	2.3 m/s	5.1 m/s	1.18	0.50
145–192 h	0.8 m/s	2.0 m/s	2.6 m/s	6.7 m/s	1.20	0.34
193–240 h	0.8 m/s	2.2 m/s	2.8 m/s	8.0 m/s	1.22	0.19

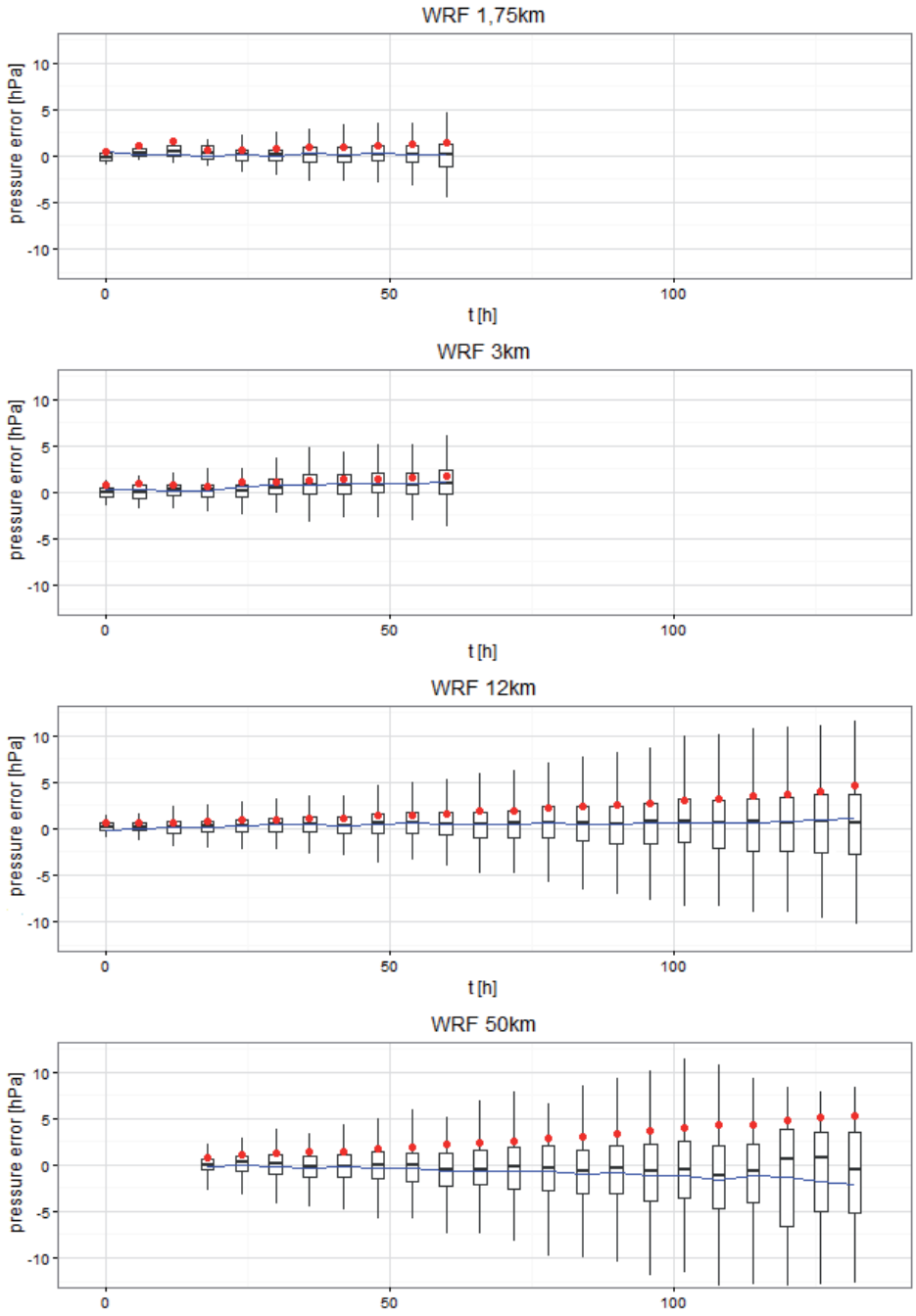

Fig. 4. Error forecast air pressure for WRF model in different resolutions in individual time steps

of wind speed forecast errors remain high in the entire period, which is demonstrated, among others, in high Mean Absolute Error values increasing gradually from 1.5 m/s to 2.3 m/s (Fig. 5). Mean Error values remain constant and their distribution shows that model values were overestimated compared to observations in the Poznań-Ławica weather station.

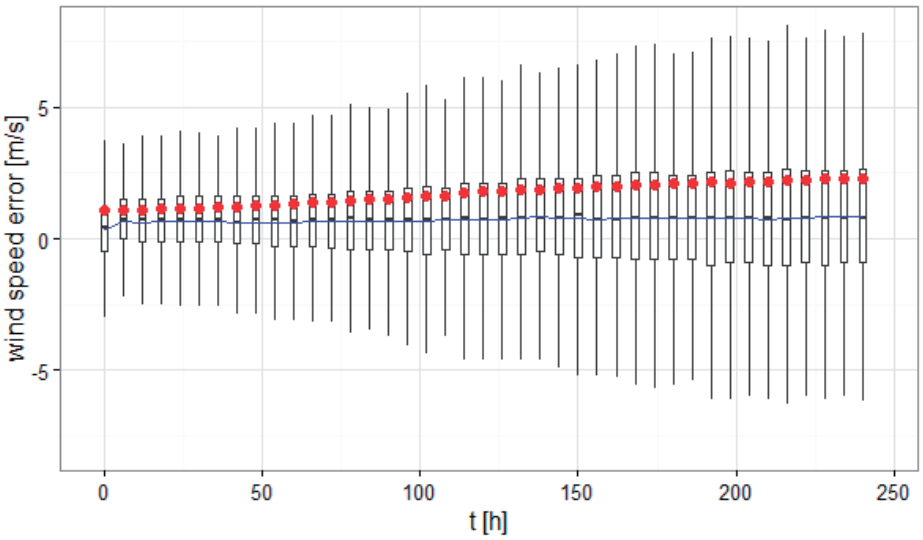

Fig. 5. Distribution of wind speed errors in individual forecast time steps for the GFS model

The distribution of test statistics for the WRF model shows great diversity in different resolutions of computational grids. Differences in results are most pronounced in the case of configurations with the smallest computational grids of the domains amounting to 1.75 km and 50 km (Tab. 6). In all cases, there is poor correlation between wind speed observations and forecasts.

Table 6. Statistical results for wind speed according to the WRF model for forecast time step t = 0–60 h

Resolution	ME	MAE	RMSE	MSE	BIAS	Correlation
1.75 km	0.4 m/s	1.2 m/s	1.5 m/s	2.5 m/s	1.23	0.74
3 km	0.7 m/s	1.3 m/s	1.7 m/s	3 m/s	1.20	0.77
12 km	1.1 m/s	1.7 m/s	2.2 m/s	4.9 m/s	1.34	0.67
50 km	1.3 m/s	1.8 m/s	2.3 m/s	5 m/s	1.4	0.64

Figure 6 shows error tendency in WRF wind speed forecasts for different computational grids. In the case of the 1.75 km model, smallest errors are observed in the initial hours of the forecast. When configured for a high spatial

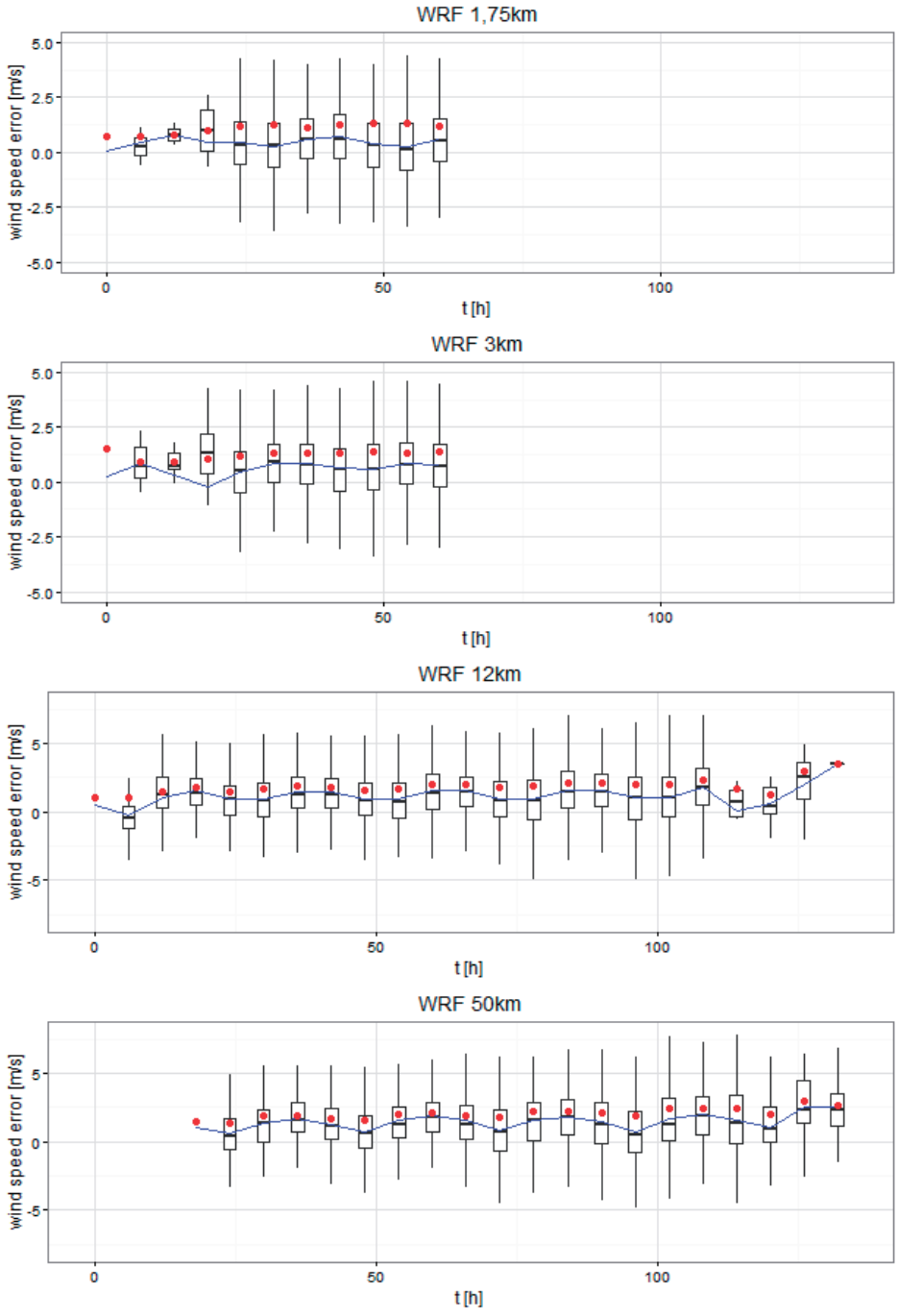

Fig. 6. Error forecast wind speed for WRF model in different resolutions in individual time steps

resolution grid, the WRF model produces better wind speed forecasts in the initial hours. WRF configurations for 12 km and 50 km resolution grids display similar error trends. ME values in all cases above 0 show overestimations of wind speed forecasts.

CONCLUSION

The article compares reliability of the GFS and WRF models in different configurations on the basis of selected meteorological elements for geographical coordinates corresponding to the location of Poznań. The idea of downscaling applied to different configurations of the WRF model does not necessarily improve weather forecast. This is confirmed by examples of air temperature verification. On the other hand, the quality of simulated atmospheric elements gets significantly better in line with increasing resolution of the model in the case of atmospheric pressure.

In the case of the WRF model, we may observe smaller errors in tested parameters for smaller sizes of computational grids (smallest WRF configuration errors appeared on the 1.75 km grid). It is particularly this grid which generates smallest mean errors of tested parameters and a high correlation with observational data. Compared to GFS, the advantage of using dynamic downscaling with the use of the WRF model, especially in smaller computational grids, is that it generates more accurate short-term forecasts. By taking more account of the local environment of the tested area and incorporating more assimilated data in the WRF model, it is possible to eliminate forecast errors to a large extent, which is particularly visible in the case of highly changeable meteorological elements such as the wind field.

There is a sharp drop in the quality of long-term forecasts after day four of the forecast. In the case of forecasts with a shorter time horizon (0–48 h), the best results were obtained in WRF 1.75 km for air temperature, WRF 1.75 km for atmospheric pressure and WRF 1.75 km for mean wind speed. Similar regularities can be observed in the following 24 hours of the forecast (48–96 h), a period which is often mentioned in the media for the benefit of end users.

Forecasts generated by the long-term GFS model get worse beyond day four in the forecast period. This is demonstrated by MAE values which become significantly higher for the studied parameters. The largest variety of values in the tested periods was recorded for atmospheric pressure forecasts. Forecasting wind speed in this model produces big extreme errors for the entire period of the forecast. In this case, the ME and MAE indicators are not much different in the studied periods.

The computation was partly performed with the use of resources provided by the Poznań Supercomputing and Networking Center (PCSS), grant No. 315.

BIBLIOGRAPHY

Frei T., 2010: *Economic and social benefits of meteorology and climatology in Switzerland,* Meteor. Appl., 17, 39–44.

Frei T., von Grünigen S. & Willemse S., 2014: *Economic benefit of meteorology in the Swiss road transportation sector*, Meteor. Appl., 21, 294–300; doi:10.1002/met.1329.

Jolliffe I.T., & Stephenson D.B. (eds), 2012: *Forecast verification: a practitioner's guide in atmospheric science*, John Wiley & Sons.

Kryza M., Wałaszek K., Ojrzyńska H., Szymanowski M.., Werner M. & Dore A.J., 2017: *High-resolution dynamical downscaling of ERA-Interim using the WRF regional climate model for the area of Poland. Part 1: Model configuration and statistical evaluation for the 1981–2010 period*, *Pure and Appl. Geophys.*, 174(2), 511–526; 10.1007/s00024–016–1272–5.

NCAR – Research Applications Laboratory, 2015: verification: Weather Forecast Verification Utilities. R package version 1.42 <https://CRAN.R-project.org/package=verification> [accessed: 2015–07–15].

Nurmi P., 2003: *Recommendations on the verification of local weather forecasts*, ECMWF Technical Memorandum, 430, 2003, s. 18.

Peck S.C. & Teisberg T.J., 1992: *CETA: a model for carbon emissions trajectory assessment*, The Energy Journ., 13(1), 55–77.

Rzepa M., 2012: *Numeryczne prognozy pogody dla Polski według modelu WRF i ocena ich sprawdzalności*, Uniwersytet Łódzki.

Schwartz C.S., Kain J.S., Weiss S.J., Xue M., Bright D.R., Kong F. & Coniglio M.C., 2009: *Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing*, Monthly Weather Rev., 137(10), 3351–3372.

Skamarock W.C., Klemp J.B., Dudhia J., Gill D.O., Barker D.M., Wang W. & Powers J.G., 2005: *A description of the advanced research WRF version 2 (No. NCAR/TN-468+ STR)*, National Center For Atmospheric Research Boulder Co Mesoscale and Microscale Meteorology Div.

Stanski H.R., Wilson L.J. & BurrowsW.R., 1989: *Survey of common verification methods in meteorology*, World Weather Watch Tech. Rept., No. 8, WMO/TD No. 358, WMO, Geneva, 114 pp.

Teisberg T.J., Weiher R.F. & Khotanzad A., 2005: *The economic value of temperature forecasts in electricity generation*, Bull. of the American Meteor. Soc., 86(12), 1765–1771.

Werner M., Kryza M., Ojrzyńska H., Skjøth C.A., Wałaszek K. & Dore A.J., 2015: *Application of WRF-Chem to forecasting PM10 concentration over Poland*, Int. J. Environ. Pollut., Vol. 58, No. 3/4, 280–292.

WWRP/WGNE Joint Working Group on Forecast Verification Research, 2017, <http://www.cawcr.gov.au/projects/verification/> [accessed: 2017–05–12].

Zambrano-Bigiarini M., 2014: *hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3–8.*, <http://CRAN.R-project.org/package=hydroGOF> [accessed: 2017–08–08].