GEOFIZIKA VOL. 35 2018

DOI: 10.15233/gfz.2018.35.5
Original scientific paper
UDC 551.509.313

Air temperature forecasts' accuracy of selected short-term and long-term numerical weather prediction models over Poland

Sebastian Kendzierski [1], *Bartosz Czernecki* [1], *Leszek Kolendowicz* [1] *and Adam Jaczewski* [2]

[1] Department of Climatology, Adam Mickiewicz University, Poznań, Poland

[2] Institute of Meteorology and Water Management – National Research Institute, Warszawa, Poland

Received 19 September 2017, in final form 17 January 2018

The article discusses the results of air temperature forecasts from four short-term and two long-term forecasts of numerical weather prediction models. The analysis covered the results of model simulations from January 2015 to January 2016 and compared them at 14 meteorological stations in Poland. The comparison was made based on the most commonly used measures for continuous parameters *i.e.*, *ME* (mean error), *MAE* (mean absolute error), *RMSE* (root mean square error), *MSE* (mean square error), *BIAS* and Pearson correlation. In the short time horizon, the best results in the context of the *MAE*, *RMSE*, *MSE* and correlation values were obtained by the Unified Model, although the diagnosed differences between the models are small. All models in the 0–72 h projection horizon reached a correlation of 0.95–0.97 and an *MAE* in the range of 1.5 °C to 2.1 °C. In the case of long-term forecasts, the HIRLAM model was slightly better than the GFS model. Clearly, in both cases, there is a marked decrease in quality after the fourth and in the following forecast lead days.

Keywords: verification, weather forecast, numerical weather prediction, NWP, long-term forecast, short-term forecast, air temperature, Poland

1. Introduction

Weather forecasts have enjoyed the interest of a wide variety of audiences over the years. Today's weather forecasts are practically an inseparable element in many areas of human activity. Starting from issues related to the right choice of clothing or planning activities outside the home, through the use of weather forecasts in agriculture, in the energy sector (both conventional and renewable), in road, air and sea transport, stock market analysis, insurance industry, industrial manufacturing, crisis management related to extreme weather events, and other related phenomena threatening human life and property (Anbarci et al., 2011).

Bearing in mind the above, high interest in publicly available weather forecasts should not come as a surprise. According to a report by Nielsen (Nielsen Audience Measurement, 2014), the viewership of the major television weather forecasts in Poland ranks among the five most viewed programmes and is only marginally lower than that of the main news bulletins. In recent years, a large group of weather forecast viewers also uses web services that provide more accurate and personalized weather information. According to Teisberg et al. (2005), an improvement in the air temperature forecast by 1% saves about $ 1 million a year in the US energy sector alone. Also in other sectors, the savings resulting from the use of specialized weather forecasts are significant. Frei (2010) estimates that in Switzerland, the savings associated with the use of weather forecasts in tourism, outdoor events and outdoor activities amount to $ 362 million per year. In road transport, these savings are estimated at $ 78–96 million, in the hydroelectric sector at about $ 98 million per year, while in the nuclear power sector they amount to $ 4 million per year.

Most of the values quoted above refer to specialized applications of weather forecasts covering a broad spectrum of atmospheric phenomena. In the meantime, most end-users are limited to obtaining information related only to the forecast near-surface air temperature (Keevallik et al., 2014) at a given location and in different time horizons. The effectiveness of individual weather sites, whose reliability may vary due to observational data assimilation schemes, the numerical model used, the configuration of the computational domain and the resolution of the grid, or the methods used to visualize and Model Output Statistic approach correct the model data is also often assessed from this perspective.

Therefore, on attempt has been made to determine the accuracy of the air temperature forecasts for selected, most popular short- and long-term numerical weather prediction (NWP) models operating on the Polish territory. Despite the lack of a universal criterion for evaluating the quality of forecasts (Jolliffe and Stephenson, 2012), the following error measures for continuous elements allow an objective comparison of the quality of forecasts based on the assumed priorities for each end-user group.

2. Data

The basis for the study is forecast data from various numerical models from January 2015 to January 2016. Detailed information on model resolution, model initiation time, time horizon, and source from where data was derived is provided in Tab 1.

All prognostic data concern the location of 14 measurement points located in Poland (Fig. 1). For these points, observational data of air temperature was

Table 1. Numerical models subjected to verification.

Model	Resolution	Initial time	Forecast length (hours)	Sources of data/Institution
GFS	25 km	00, 06, 12, 18	240	www.meteomodel.pl
HIRLAM	25 km	00, 12	240	www.yr.no
WRF	3 km	00, 12	72	www.meteoprognoza.pl
COSMO	7 km	00, 12	72	Institute of Meteorology and Water Management – National Research Institute (www.pogodynka.pl)
UM	4 km	00, 12	72	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw (www.meteo.pl)
GEM	25 km	00	42	www.meteomodel.pl

Note: GFS – Global Forecast System
HIRLAM – HIgh Resolution Limited Area Model
WRF – Weather Research and Forecasting model
COSMO – COnsortium for Small-scale MOdeling
UM – Unified Model
GEM – Global Environmental Multiscale model

obtained from meteorological stations at the height of 2 m above ground level (a.g.l.). Meteorological data in an hourly time resolution were obtained from the Institute of Meteorology and Water Management – National Research Institute, and have been used to verify the accuracy of each model.

Among the measurement points – two are located in the vicinity of the Baltic Sea (Świnoujście and Gdańsk-Świbno). They are situated at an altitude of 7 m above sea level (a.s.l.). Most stations are located in lowland areas that have been evenly matched. The station in Kraków, in the area at the border of the valley and the highlands, is located at 240 m a.s.l. Two stations are located in a mountain area – Jelenia Góra (station situated at an altitude of 350 m a.s.l.) and Zakopane (860 m a.s.l.).

The distribution of the observed hourly air temperature in the examined period for all measuring stations is shown in Fig. 2. The lowest temperature recorded was –21.3 °C in Jelenia Góra. The highest was 37.3 °C in Słubice. The highest percentage of observed values ranges from 0 °C to 20 °C.

The whole of 2015, according to the air temperature classification based on percentile values (Miętus et al., 2002), has shown itself to be extremely warm in most of the territory of Poland. Only the seafront strip was classified as anomalously warm. The average annual air temperature was 9.7 °C during this period (Climate Monitoring Bulletin, 2015). Considering the seasonality of air temperature changes, the MAM season (March, April, May) in northern Poland was very warm, while the southern part of the country was defined as slightly warm. The JJA season (June, July, August) all over Poland was classified as extremely warm, while the SON season (September, October, November) rated anoma-

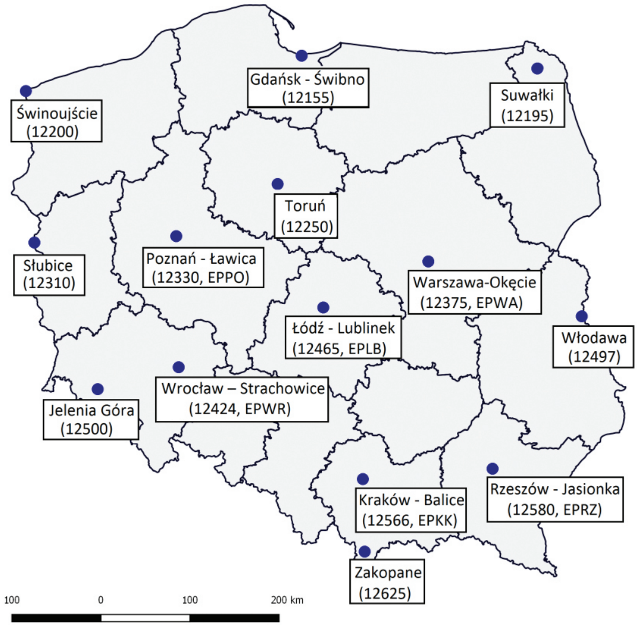

Figure 1. Location of meteorological stations in Poland for which numerical models were verified and observation data were obtained (name of the place, WMO station number and airport ICAO/ WMO acronym).

lously warm. January and February 2015 were classified as very warm in most areas of the country, only the coastal strip and mountain areas were defined as warm. In turn, January 2016 was cool throughout the area except for the seaside where it was described as very cool. The presented range of thermal classification that occurred during the analysed period can explain a small number of air temperature cases below 0 °C (Fig. 2).

3. Methodology

The evaluation of the accuracy of air temperature was based on verification methods for continuous meteorological parameters (Jolliffe and Stephenson, 2012). These include, but are not limited to, statistical measures for series of measurements that are subject to deviations from expected values. These are mean error (*ME*) (eq. 1), mean absolute error (*MAE*) (eq. 2), root mean square error (*RMSE*) (eq. 3), mean squared error (*MSE*) (eq. 4), *BIAS* (b) (eq. 5) and correlation (*r*) (eq. 6).

ME (eq. 1) measures the average difference between the forecast and the observation (Nurni, 2003). It defines the mean forecast error, the ideal result of

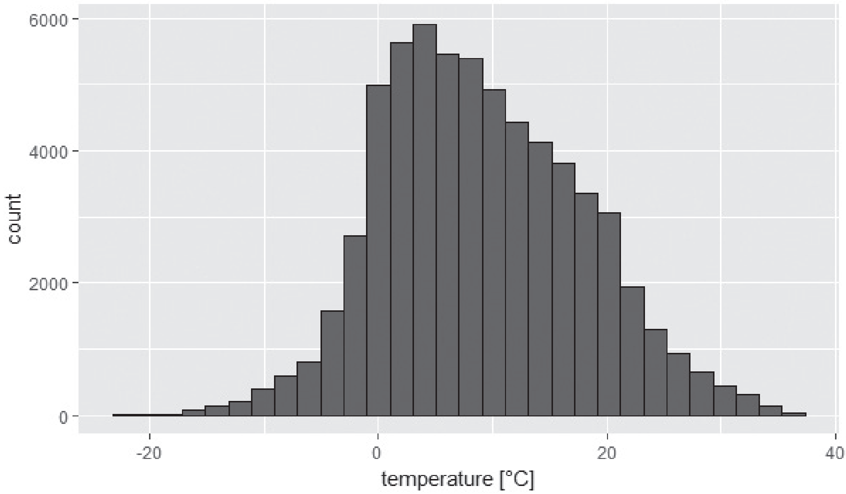

Figure 2. Air temperature frequency distribution in January 2015 – January 2016 observed at all measuring points.

the index is *ME* = 0. Forecasts that are on average too warm will exhibit positive value of this index, and negative for too cold forecasts (Wilks, 2011).

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

MAE (eq. 2) measures the average magnitude of forecast errors in a given dataset and therefore it is a scalar measure of forecast accuracy (Nurni, 2003). The ideal result of the index is *ME* = 0, its theoretical values range from 0 to infinity.

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

RMSE (eq. 3) is a quadratic scoring rule which gives the average magnitude of errors, weighted according to the square of the error (Stansky et al., 1989). It indicates the average magnitude of the forecast's error. The idealvalue is *RMSE* = 0, theoretical values range from 0 to infinity.

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

MSE (eq. 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 ideal *MSE* result is 0, and theoretical values range from 0 to infinity.

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

BIAS (eq. 5) gives the value of the difference between the mean value from measurements and the benchmark value. It answers the question how the average forecast magnitude compares to the average observed magnitude. It does not measure the magnitude of the errors (Linton et al., 1994) 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. The ideal result is $b = 1$.

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

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

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

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

The results cover the comparison of the temperature forecast determined by the closest numerical weather prediction grid to the station location. The results of the verification are presented in tabular form for the entire model (using the data for the whole period and for all forecast ranges). Long-term numerical models also show these statistics for each model start time. For a more complete picture of the results, they are presented using boxplot graphs, which show the distribution of the difference in predicted value from the one observed (difference between forecast and observation), median, percentiles, and extremes in subsequent forecast lead times. Boxplots have percentiles set at 0.1, 0.25, 0.5, 0.75 and 0.9. In addition, this graph is presented for each season (DJF, MAM, JJA, SON).

Verification and hydroGOF packages dedicated to the R programming environment (R Core Team, 2017) were used for the calculations. The numbers included in the maps show the verification results for individual stations for all studied numerical models. For long-term forecasts, Taylor diagrams were used (Taylor, 2001). They graphically describe the degree of consistency between forecast and observation. They show three statistics: Pearson's correlation coefficient, root mean square error (*RMSE*), and standard deviation. The whole study was divided into two parts, the first verified short-term numerical models (GEM, WRF, COSMO, UM) and additionally GFS and HIRLAM, while the second focused on long-term models (GFS and HIRLAM).

4. Verification of short term forecast

The first stage of the study was a comparison of forecast error results for different time intervals (0–24 h, 25–48 h, 49–72 h). Statistical results have also included GFS and HIRLAM models with a maximum forecast time of 72 h. Comparing the mean errors of these models, it can be noted that their values are below 0, which means that the air temperature forecast is slightly underestimated (Tab. 2). The smallest mean forecast errors (*ME*) are observed for the GFS model, while the largest for the WRF model. In turn, the smallest absolute mean error (*MAE*) is found in the UM model (1.5 °C), the largest in the COSMO model (2.2 °C). The smallest values of the *RMSE* and *MSE* indicators occur in the UM model and the highest in the COSMO model. For the *BIAS*, the GEM and COSMO models are at 0.97, UM 0.96 and WRF 0.95. Correlation of forecast values to observed values is high for all models, reaching values of 0.92 to 0.98 for the models tested.

Analyzing errors of numerical models for short-term forecasts one can see changes in their level in subsequent forecast time steps. In the analyzed nu-

Table 2. NWP models' accuracy statistics for air temperature for the forecast lead time t = 0–72 h.

MODEL	*t*	*ME*	*MAE*	*RMSE*	*MSE*	*BIAS*	*r*
WRF	0–24	–0.4	1.77	2.32	5.4	0.97	0.96
	25–48	–0.4	1.75	2.3	5.28	0.97	0.96
	49–72	–0.31	1.83	2.39	5.71	0.96	0.97
COSMO	0–24	–0.13	1.99	2.69	7.22	0.95	0.98
	25–48	–0.32	2.15	2.85	8.14	0.95	0.96
	49–72	–0.5	2.23	2.61	6.82	0.96	0.95
UM	0–24	–0.39	1.36	1.81	3.29	0.98	0.96
	25–48	–0.4	1.53	2.00	4.01	0.97	0.96
	49–72	–0.4	1.67	2.19	4.81	0.97	0.96
*GEM	0–24	–0.31	1.48	2.08	4.32	0.98	0.97
	25–42	–0.32	1.72	2.37	5.62	0.97	0.96
GFS	0–24	–0.1	1.58	2.19	4.78	0.97	0.94
	25–48	–0.08	1.7	2.31	5.34	0.96	0.95
	49–72	–0.08	1.7	2.32	5.38	0.96	0.92
HIRLAM	0–24	0.11	1.38	1.81	3.33	0.98	0.98
	25–48	0.08	1.59	2.15	4.62	0.97	0.98
	49–72	0.1	1.65	2.21	4.87	0.97	0.97

Note: *GEM – 0–42 h

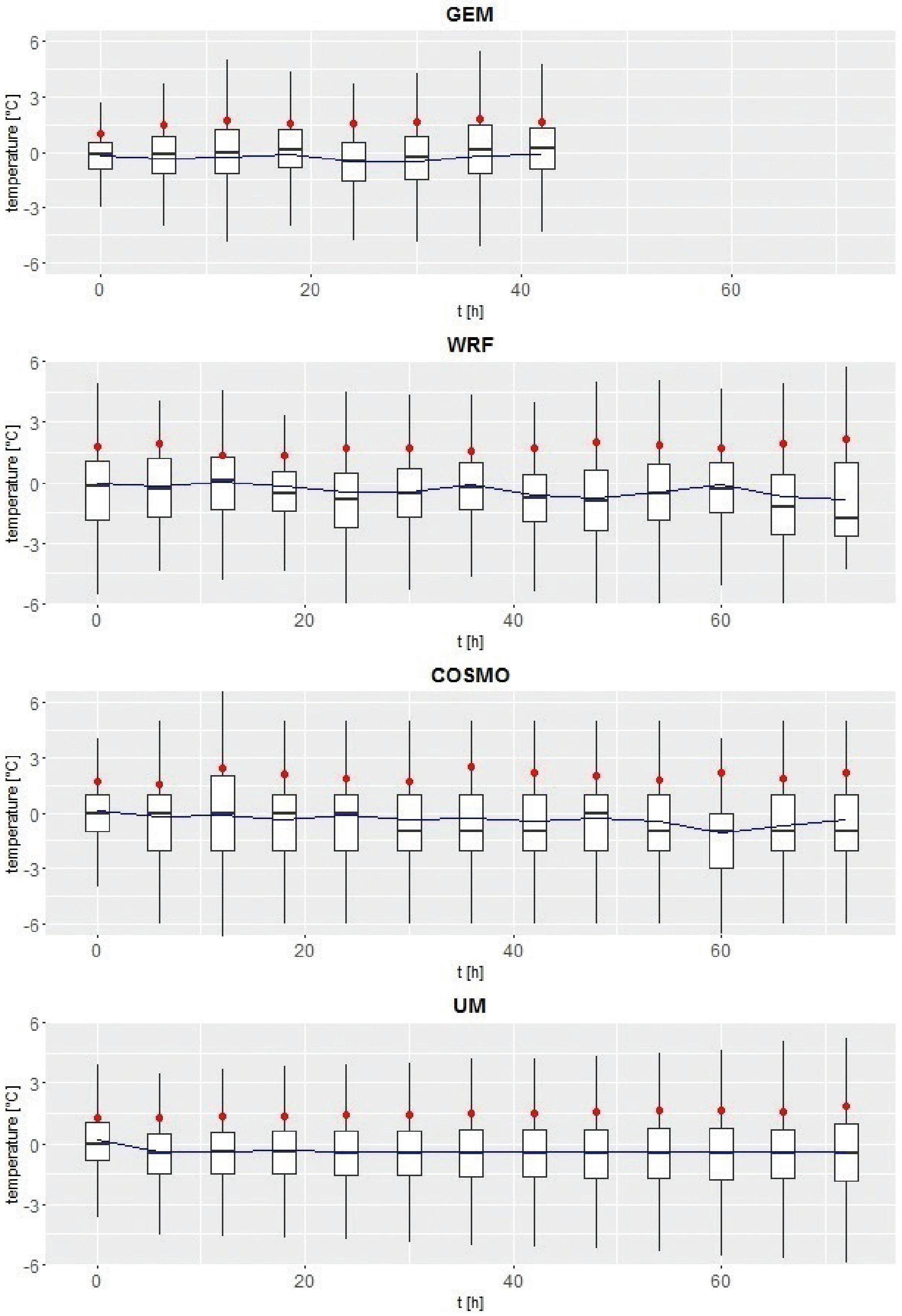

Figure 3. The error range for short-term predictions in the particular forecast lead times. *ME* is marked with a solid line, while *MAE* is marked with red dots.

merical models, the maximum difference in predicted and observed values is 6 °C, but this level does not occur during the last hours of the forecast time (Fig. 3). Only with the UM model one can see a progressive increase in the error value along with forecast time. Mean error values, which are indicated by a continuous line, are slightly fluctuating in the GEM, WRF and COSMO models. Only the UM model shows a constant mean error value with the forecast time lapse. Absolute error values (*MAE*s) are indicated with dots. As with the mean error, the values for GEM, WRF, and COSMO are time-varying, the UM model shows an upward trend over time. This graph also shows which numerical mod-

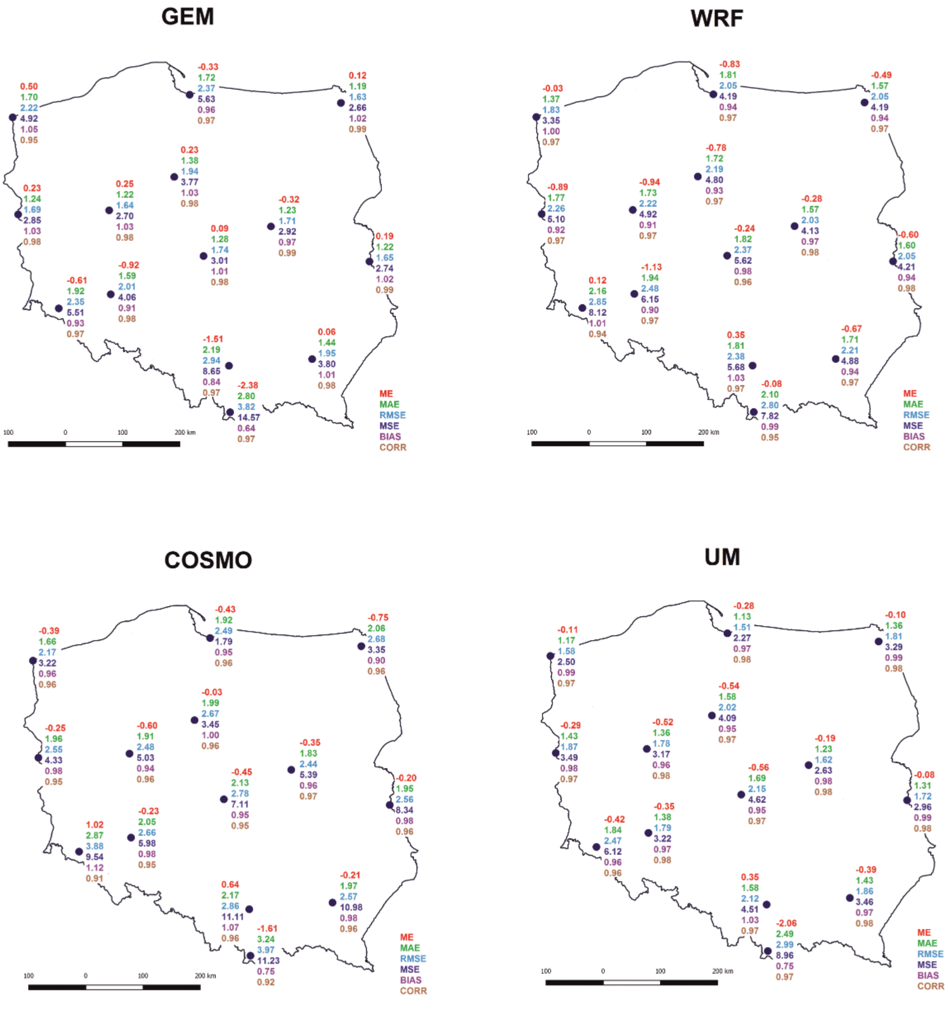

Figure 4. Values of the calculated verification statistics for short-term forecast at analysed stations.

els have the biggest errors at the start of the model – the largest are found in the WRF model, while the smallest in the GEM model.

When analyzing individual errors and statistical indicators, it can be noted that their spatial distribution is different for the analyzed numerical models (Fig. 4).

In the GEM model one can distinguish the stations for which the *ME* is the largest. These are stations in Zakopane (–2.4 °C) and Kraków (–1.5 °C). The smallest error values are found in Suwałki (0.1 °C) and Rzeszów (0.1 °C). MAEs are the lowest for stations located in central Poland, while the highest ones are recorded in Zakopane (2.8 °C) and Kraków (2.2 °C). *RMSE* and *MSE* show the smallest values in the latitudinal strip from stations in Słubice to Włodawa. The value of the bias reaches the smallest levels for the stations in Zakopane and Kraków; for other stations it is around 1. The correlation between the forecast and observed values is high for all stations and is in the range of 0.95 to 0.99. In conclusion, the GEM model has the best verification results for stations located in central Poland, slightly worse for Southern Poland, especially for Zakopane.

The second analyzed model – WRF has a number of settings in its properties which should take into account the occurrence of local phenomena affecting the quality of the forecast. This is particularly important for points located, e.g., in mountain areas or near the sea. *ME* for almost all points are not too high, but comparing with other models, *ME* is the largest. Only for the Wrocław, Poznań, Słubice and Gdańsk stations, the results indicate an underestimation of the forecast air temperature. *MAE* values begin from 1.5 °C for points located in northern Poland. They rise towards the south of the country, reaching there about 2 °C for the points in Jelenia Góra and Zakopane that might be related with more frequent in this areas occurrence of thermal inversion and foehn winds. Similar to the *MAE*, the *RMSE* and *MSE* indicators show the best values for points in northern Poland, while advancing to the south, there is a noticeable decrease in the quality of the forecast. The bias value reaches about 1 for all measuring points. Correlation of results is high in the range from 0.94 (Jelenia Góra) to 0.98 Włodawa. Statistical results for the WRF model show that the air temperature forecasts calculated by this model have small errors. In particular, it is visible for stations located in specific conditions – mountains (Jelenia Góra and Zakopane), where, compared with other numerical models, the quality of the forecast is high.

ME values for the COSMO model are the lowest for points in central Poland while increasing for the northern and southern stations. Almost all the results indicate an underestimation of the forecast temperature, indicating an overestimation only for the stations in Jelenia Góra and Kraków. The distribution of *MAE*, *RMSE*, *MSE* for the COSMO model shows spatial similarity. Their value is lower for stations located in the northern part of Poland. From north-west to south, a gradual increase in error rate is observed. The weakest forecasts are for the Zakopane and Jelenia Góra stations. The bias in the surveyed area ranges from 0.95 to 1.12. Correlation of results is high, especially for central stations. Only Jelenia Góra and Zakopane show a decrease to 0.91–0.92. In this model it is also possible to indicate stations where the quality of the forecast is high – they are stations located in northern and central Poland.

The last analyzed short-term numerical model is UM. With the exception of the station in Kraków, other stations show slight underestimation of the forecast. There is also a small variation in ME levels. It presents the lowest result for the point in Zakopane (–1.6 °C). As in previous models, the MSE and RMSE indicators are similar. The best values are noted for the northern and eastern points, while the weakest for the Zakopane and Jelenia Góra stations. The value of the bias indicator is comparable for all stations (around 1), except Zakopane where it reaches 0.75. Correlation of predicted to observed values is high for all measurement points (0.96–0.98).

Comparing all results, there is a narrow group of numerical models that have better predictive properties in mountain areas comparing to other models. Undoubtedly for the Zakopane and Jelenia Góra stations, the smallest errors are shown in the WRF model. Nonetheless, for some other stations located in other areas of Poland, the simulation results are not that robust in comparison to other models. This demonstrates the complexity and diversity in the way these models are calculated.

Verification charts presented in Fig. 5 show the distribution of predicted values to the observed values and make it possible to determine the ranges of temperatures showing the maximum differences. It can be seen that most short-term numerical models overestimate low air temperatures, especially below –10 °C. Only in the case of the UM model, no such regularities were found. In turn, all models point to a slight overestimation of the air temperature above 30 °C. All models in the temperature range of 0 °C to 25 °C do not indicate the deviation direction. The forecast temperature is closest to the observed temperature in this range.

5. Verification of long-term forecast

Statistical results for the entire numerical model, taking into account the entire period considered for all measurement points, are presented in Tab. 3. Comparing long-term numerical models for air temperature forecasts, better results for the HIRLAM model can be found. In all statistical analyses, this model performs better than the GFS model.

In the case of long-term forecasts, individual statistics were also broken down into available start-up times for the model as shown in Tab. 4. No substantial differences can be observed between the models' starting hours, so there is no clearly visible difference in the quality of forecasts as to the time of their start.

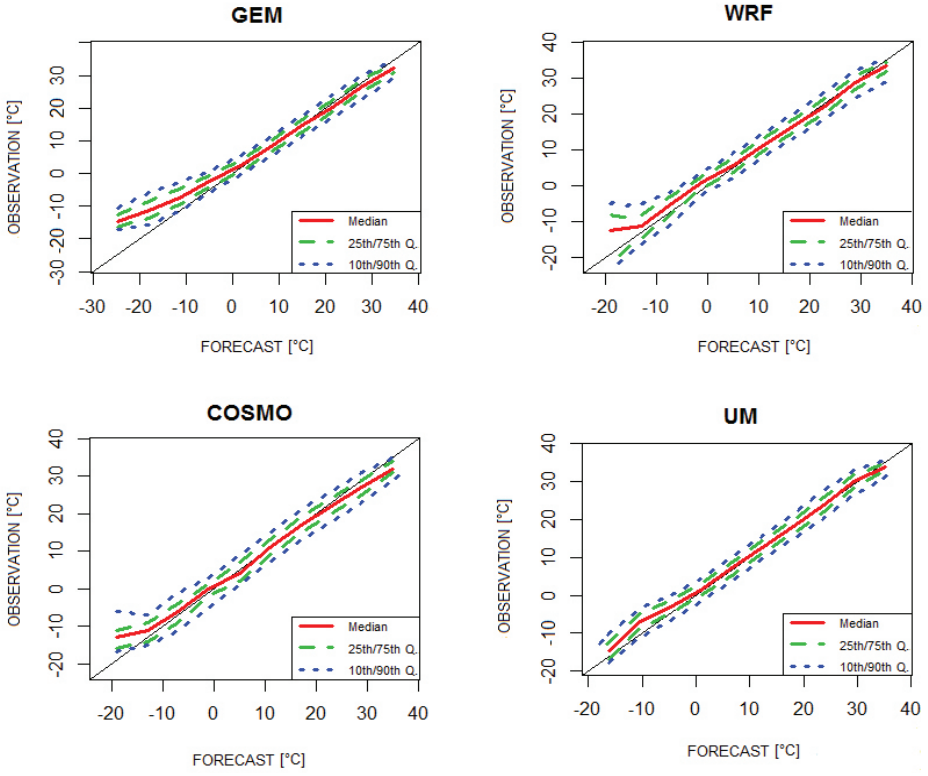

Figure 5. The distribution of forecast air temperature against in-situ data for short-term weather forecasts from January 2015 to January 2016.

Only for the GFS model at 12 UTC mean error (*ME*) is minimally lower than during the remaining hours.

The change in the quality of the air temperature forecast along with the forecast time elapsed is shown in Fig. 6. This data refers to the result of models for all measurement points.

For the GFS model, there is an increase in the value of the extremes of the predicted and observed temperature differences. Up to about the 100th forecast hour it is around 5 °C, later it systematically increases up to 12 °C in the 240th hour of the forecast.

The mean error (ME) values for the whole of this period do not show fluctuations. Its smallest value of –0.1 °C was reached in the 72th hour of the forecast. It gradually decreases from the 72nd hour. The lowest ME value was –0.3 °C in the 234th hour of forecast. The overall ME score for the GFS model indicates minimal underestimation of the air temperature forecast. The mean absolute error (MAE), as well as the ME, also shows a constant level change over the

Table 3. Statistical results of long-term forecast in different time horizons.

MODEL	*t*	*ME*	*MAE*	*RMSE*	*MSE*	*BIAS*	*r*
GFS	0–48	–0.09	1.63	2.24	5.00	0.97	0.94
	49–96	–0.07	1.91	2.56	6.53	0.96	0.95
	97–144	–0.17	2.38	3.17	10.02	0.93	0.93
	145–192	–0.23	2.93	3.87	14.94	0.90	0.93
	193–240	–0.25	3.52	4.60	21.85	0.86	0.92
HIRLAM	0–48	0.11	1.53	2.06	4.25	0.97	0.98
	49–96	0.09	1.82	2.42	5.84	0.96	0.98
	97–144	–0.04	2.31	3.07	9.42	0.94	0.97
	145–192	–0.18	2.97	3.92	15.34	0.90	0.95
	193–240	–0.15	3.40	4.46	19.91	0.87	0.94

Table 4. Statistical results for individual long-term NWP forecasts, split by model start times.

GFS

MODEL START	*ME*	*MAE*	*RMSE*	*MSE*	*BIAS*	*r*
00 UTC	–0.17	2.45	3.36	11.28	0.93	0.93
06 UTC	–0.19	2.45	3.37	11.37	0.93	0.93
12 UTC	–0.12	2.45	3.38	11.44	0.94	0.93
18 UTC	–0.17	2.48	3.4	11.58	0.93	0.92

HIRLAM

MODEL START	*ME*	*MAE*	*RMSE*	*MSE*	*BIAS*	*r*
00 UTC	0	2.41	3.32	11	1	0.93
12 UTC	–0.06	2.34	3.2	10.23	0.99	0.92

forecast time. It is marked with a dot in the graph. The value of MAE tends to increase. The smallest value is 1.5 °C and is reached at the start of the forecast. Then, up to about the 96th hour, it slightly rises up to 2 °C. Then a steady rise to a maximum of 3.73 °C is reached at the last hour of the forecast.

The HIRLAM model shows a stepwise distribution of the values of the temperature difference between the forecast and observed temperature extremes. The lowest of its values occur at the beginning of the forecast, then there is a big increase in the 6th hour. After this period there is a decline and then a steady increase of these levels is observed. Its maximum extreme values, around 12 °C, are observed at the 218th hour of the forecast. The *ME* indicator changes over

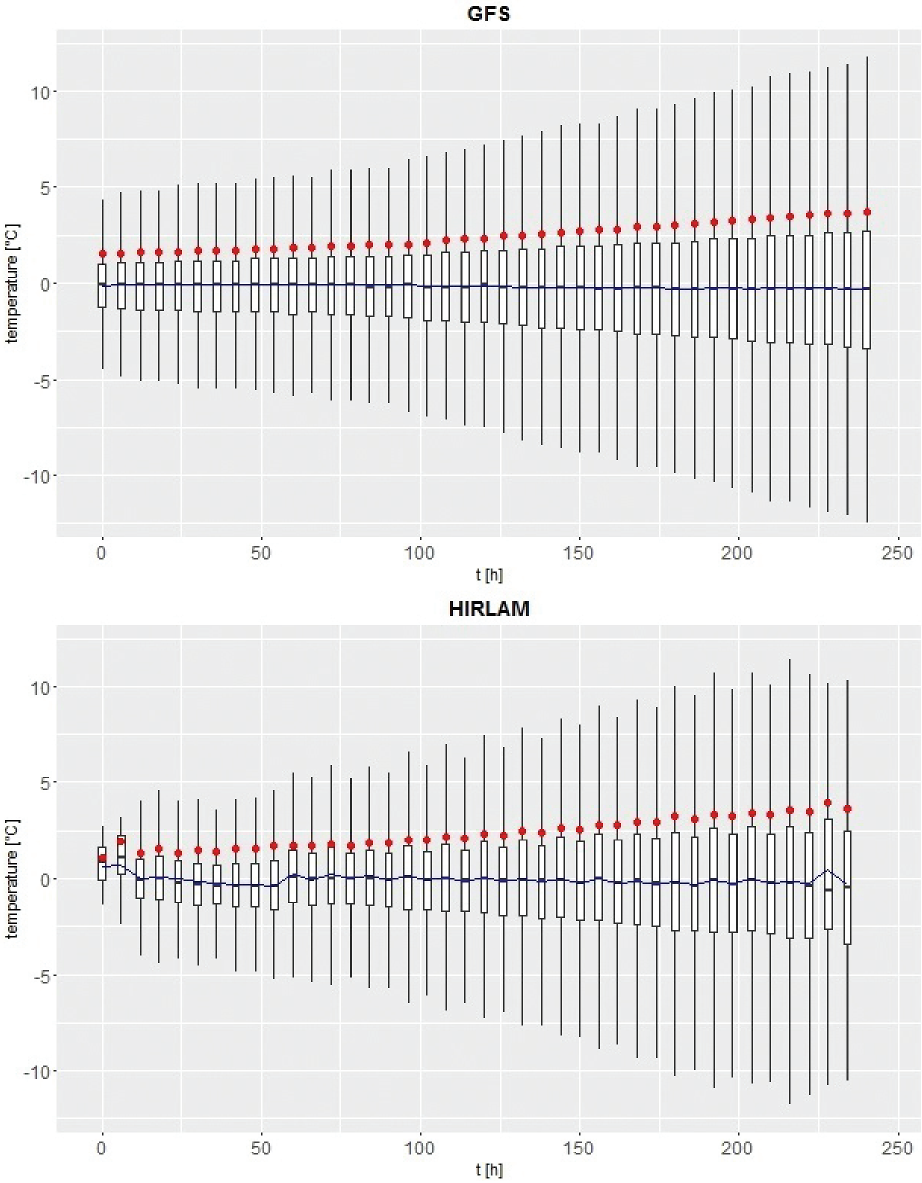

Figure 6. The air temperature error ranges for long-term NWP models in particular forecast lead times. The solid line marks *ME*, the red dot denotes *MAE*.

time. Initially, it rises to 0.68 °C in the 6th hour. Later it falls in the 54th hour (–0.4 °C). At the 60th hour, it reaches 0.2 °C, then it slowly drops to –0.3 °C during the last hour. Only during the 228th hour can a deviation be observed at 0.49 °C.

In general, the trend for *ME* is downward. The *MAE*, like *ME*, does not indicate a trend in the early hours of the forecast. At the 6th hour it reaches 2 °C, but in the next hour, it is marked down. From the 24th hour of the forecast, an upward trend may be observed, with its maximum level being reached at the 228th hour (4 °C).

Both models show similar trends in the statistical indicators in question. It is worth noting the first hours of the HIRLAM forecast, which, despite high fluctuations in the early hours, show smaller differences in temperature extremes than the GFS model. It offers better parameters for the purpose of short-term temperature forecasting. This model also shows smaller temperature errors. However, both models indicate a slight underestimation of the air temperature forecast.

ME in the stations under consideration is low. It is high only in Zakopane, *i.e.* –3.7 °C (Fig. 7). *MSE, RMSE* are at similar levels. The lowest values of these indicators are noted in northern Poland. Moving towards the south shows an increase in errors. The station with the greatest value is Zakopane. The bias indicator for all stations is about 1, only for Zakopane (0.37) it indicates the weakness of average forecasts. The value of the predicted-to-observed temperatures correlation is high for all stations and ranges from 0.90 to 0.94.

The lowest levels of *ME* in the case of the HIRLAM model are observed in central and eastern Poland, but for the other points, the values are not too far off. *MSE* and *RMSE* reach the lowest levels for the stations in Świnoujście and Gdańsk. They increase towards the south, reaching the highest values in the southern stations, especially in Kraków, Jelenia Góra, and Zakopane. The bias achieves similar values across the country, ranging from 0.94 to 1.03. Also for

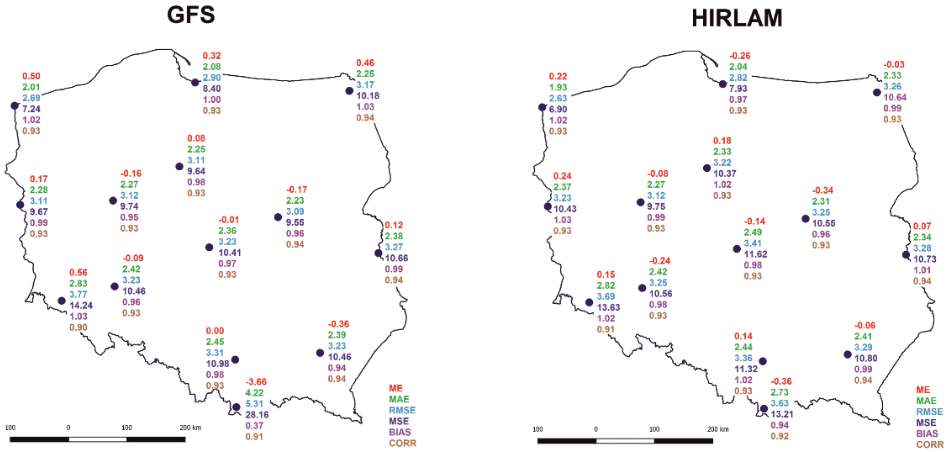

Figure 7. Values of calculated verification statistics for long-term numerical models at analyzed stations.

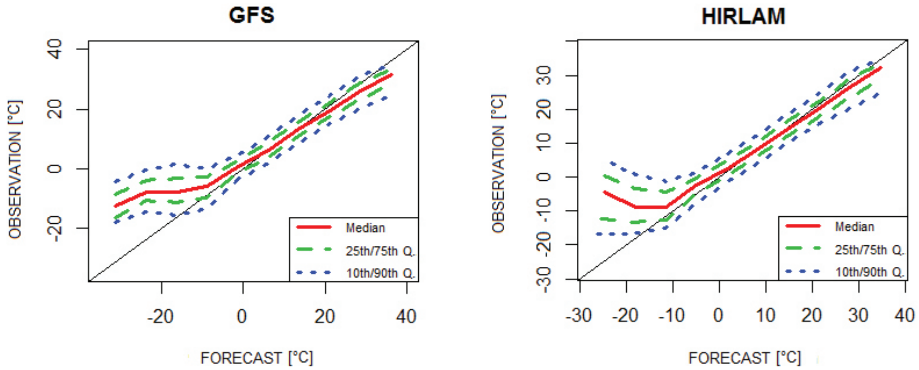

Figure 8. Course of observed and forecast air temperature values for long-term weather forecasts from January 2015 to January 2016.

all stations, the Pearson correlation is high and evenly distributed (0.91–0.94). By comparing long-term models, one can indicate stations where more accurate air temperature forecasts are calculated. These are undoubtedly stations located in the mountain area where the HIRLAM model shows minor errors. Also for most of the northern stations – Świnoujście, Gdańsk, Suwałki – this model showed lower error values. However, for points situated in the central strip from Słubice to Włodawa, the GFS model provided better performance as expressed by these indicators.

Both analysed long-term NWP models indicate big errors in the forecasting of low temperatures, significantly lowering the predicted values relative to the observed values (Fig. 8). For the GFS model, this is visible at temperatures below –5 °C. For the HIRLAM model below –10 °C. In the first model analysed for temperatures above 20 °C, we can also see a slight overestimation of temperature forecasts compared with observations. In the HIRLAM model, this is also visible,

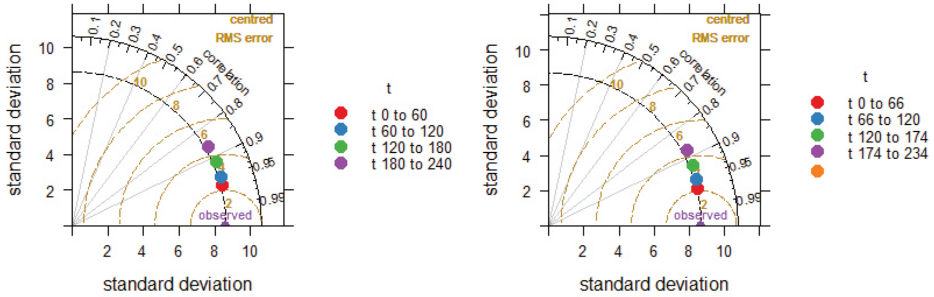

Figure 9. Taylor diagrams for long-term NWP forecast split by individual intervals of forecast lead time.

but to a lesser degree from 25 °C. The highest forecast efficiency is found in a temperature range of 0 °C to 20 °C (Fig. 8).

The Taylor (2001) diagrams for long-term numerical models in the different forecast time periods are shown in Fig. 9. The slightly better model performance is obtained in the case of HIRLAM, where, among other things, higher correlations of predicted values to those observed can be found.

6. Discussion and conclusions

Air temperature is undoubtedly the most frequently checked element of publicly available weather forecasts. In short-term forecasts, its greatest accuracy is important, while its long-term assessment can only show a trend. The detailed verification of the short-term numerical models does not give a clear answer to the question which one is the most reliable. In many cases, the UM model came top with the *MAE* of 1.5 °C being the lowest among the models tested and was the same as that obtained by Melonek (2011).

The analysis has shown that many NWP models, both short-term and long-term, have a problem with reliable predictions of low temperatures during the winter season. Also in this case, the exception is the UM model, which was best suited for cool episodes, and therefore it may indicate better matched parametrization schemes related to radiation and heat transfer.

For stations located in mountain and sub-mountain areas, it is important to adapt the model settings to local conditions by applying appropriate parametrization schemes and land use types (Skamarock et al., 2001; Powers et al., 2017). This type of dependency can be seen in the WRF model, which allows simulating many sub-scaled processes with respect to the boundary conditions of the GFS model. Taking into account the local conditions of the location of the area under investigation more carefully and using more computationally demanding parameterization schemes allows for significant elimination of forecast errors, although it should be noted at the same time that all the analysed short-term forecasts have high correlation values in the range of 0.92–0.98.

Long-term forecasts given by the GFS and HIRLAM models are characterized by a gradual and progressive increase in errors in the forecast air temperature. This is noticeable most strongly in the case of forecasts above the 4th day of the forecast. The models did not show a dependency between model start time and forecast quality. The biggest mean forecast errors occur in mountain and sub-mountain stations. In the case of long-term forecasts, the problem of low temperature overestimation and underestimation at temperatures greater than 30 °C is also observed, similarly to regularities found for short-term forecasts.

Acknowledgements – The authors wish to thank the Institute for Meteorology and Water Management – NRI for making COSMO data available and for providing observation data from meteorological stations.

In particular, the authors thank Piotr Djaków for providing GFS and GEM data, Juliusz Orlikowski (Gdańsk University of Technology) for making available data for the WRF model and the Interdisciplinary Centre for Mathematical and Computational Modelling of the University of Warsaw (ICM) for providing UM model data.

The computation was partly performed with the use of resources provided by the Poznań Supercomputing and Networking Center (PCSS), grant No. 356.

References

Anbarci, N., Boyd, J., Floehr, E., Lee, J. and Song, J. J. (2011): Population and income sensitivity of private and public weather forecasting, *Reg. Sci. Urban Econ.*, **41**,124–133, DOI: 10.1016/j.regsciurbeco.2010.11.001.

Frei, T. (2010): Economic and social benefits of meteorology and climatology in Switzerland, *Meteorol. Appl.*, **17**, 39–44, DOI: 10.1002/met.156.

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

Keevallik, S., Spirina, N., Sula, E. M. and Vau, I. (2014): Statistics of different public forecast products of temperature and precipitation in Estonia, *P. Est. Acad. Sci.*, **63**, 174–182, DOI: 10.3176/proc.2014.2.07.

Linton, O. and Nielsen, J. P. (1994): A multiplicative bias reduction method for nonparametric regression, *Stat. Probabil. Lett.*, **19**, 181–187, DOI: 10.1016/0167-7152(94)90102-3.

Melonek, M. (2011): Porównanie wyników weryfikacji modeli numerycznych prognoz pogody działających operacyjnie w ICM, *Infrastruktura i Ekologia Terenów Wiejskich*, (06), 31–42 (in Polish).

Miętus, M., Filipiak, J. and Owczarek, M. (2002): *Warunki termiczne na obszarze Wybrzeża i Pomorza w świetle wybranych klasyfikacji*. IMiGW.

NCAR – Research Applications Laboratory (2015): Verification: Weather forecast verification utilities. R package version 1.42., available at https://CRAN.R–project.org/package=verification

Nielsen Audience Measurement (2014): Programy-informacyjne-najchetniej-ogladane, available at http://www.press.pl/tresc/36624

Nurmi, P. (2003): *Recommendations on the verification of local weather forecasts*: *Technical memorandum: Number 430*, available at https://www.ecmwf.int/sites/default/files/elibrary/2003/11401-recommendations-verification-local-weather-forecasts.pdf

Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., Coen, J. L. and Gochis, D. J. (2017): The weather research and forecasting (WRF) model: Overview, system efforts, and future directions. *Bull. Am. Meteor. Soc.*, **98**, 1717–1737, DOI: 10.1175/BAMS-D-15-00308.1.

R Core Team (2017): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, available at https://www.R–project.org/

Skamarock W. C., Klemp, J. B. and Dudhia, J. (2001): Prototypes for the WRF (Weather Research and Forecasting) model, in: *Preprints, Ninth Conf. Mesoscale Processes, J11–J15*. AMS, Fort Lauderdale, FL.

Stanski, H. R., Wilson, L. J. and Burrows, W. R. (1989): *Survey of common verification methods in meteorology*. World Meteorological Organization, Geneva, pp. 114.

Taylor, K. E. (2001): Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.,* **106**, 183–7192, DOI: 10.1029/2000JD900719.

Teisberg, T. J., Weiher, R. F. and Khotanzad, A. (2005): The economic value of temperature forecasts in electricity generation, *Bull. Am. Meteor. Soc.*, **86**, 1765–1771, DOI: 10.1175/BAMS-86-12-1765.

Treder, W., Wójcik, K., Tryngiel-Gac, A., Klamkowski, K. and Michalczuk, L. (2011): Ocena jakości prognozowania pogody, *Infrastruktura i Ekologia Terenów Wiejskich*, **6**, 43–58 (in Polish).

Wickham, H. (2009): *ggplot2: Elegant graphics for data analysis*. Springer–Verlag, New York.

Wilks D. S. (2011): *Statistical methods in the atmospheric sciences*. International Geophysics Series, **100**. Elsevier, 676 pp.

Zambrano-Bigiarini, M. (2014): hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3–8, available at http://CRAN.R–project.org/package=hydroGOF

SAŽETAK

Točnost prognoza temperature zraka dobivenih odabranim kratko- i dugoročnim numeričkim modelima prognoze vremena iznad Poljske

Sebastian Kendzierski, Bartosz Czernecki, Leszek Kolendowicz i Adam Jaczewski

U članku se razmatraju rezultati prognoze temperature zraka pomoću četiri kratkoročna i dva dugoročna numerička modela prognoze vremena. Analiza je obuhvatila rezultate simulacija modela od siječnja 2015. do siječnja 2016., koji su uspoređeni s podacima 14 meteoroloških postaja u Poljskoj. Usporedba je izrađena na temelju najčešće korištenih mjera za kontinuirane parametre, *tj. ME* (srednja pogreška), *MAE* (srednja apsolutna pogreška), *RMSE* (korijen srednje kvadratne pogreške), *MSE* (srednja kvadratna pogreška), *BIAS* i Pearsonova korelacija. Za ovako kratak vremenski interval, u kontekstu vrijednosti *MAE*, *RMSE*, *MSE* i korelacije, najbolji rezultati dobiveni su ujedinjenim modelom, iako su utvrđene razlike među modelima male. Svi modeli su u prognostičkom vremenu od 0 do 72 h dostigli korelaciju od 0,95 do 0,97 i *MAE* u rasponu od 1,5 °C do 2,1 °C. U slučaju dugoročnih prognoza model HIRLAM bio je nešto bolji od GFS modela. Jasno je da u oba slučaja dolazi do znatnog smanjenja kvalitete nakon četvrtog i sljedećih prognostičkih dana.

Ključne riječi: verifikacija, prognoza vremena, numerička prognoza vremena, NWP, dugoročna prognoza, kratkoročna prognoza, temperatura zraka, Poljska

Corresponding author's address: Sebastian Kendzierski, Department of Climatology, Adam Mickiewicz University, Poznań, Poland; e-mail: wrf@amu.edu.pl