

## Forecasting daily meteorological time series using ARIMA and regression models\*\*

Małgorzata Murat<sup>1</sup>, Iwona Malinowska<sup>1</sup>, Magdalena Gos<sup>2</sup>, and Jaromir Krzyszcak<sup>2\*</sup>

<sup>1</sup>Department of Mathematics, Lublin University of Technology, Nadbystrzycka 38a, 20-618 Lublin, Poland

<sup>2</sup>Institute of Agrophysics, Polish Academy of Sciences, Doświadczalna 4, 20-290 Lublin, Poland

Received September 20, 2017; accepted January 31, 2018

**Abstract.** The daily air temperature and precipitation time series recorded between January 1, 1980 and December 31, 2010 in four European sites (Jokioinen, Dikopshof, Lleida and Lublin) from different climatic zones were modeled and forecasted. In our forecasting we used the methods of the Box-Jenkins and Holt-Winters seasonal autoregressive integrated moving-average, the autoregressive integrated moving-average with external regressors in the form of Fourier terms and the time series regression, including trend and seasonality components methodology with R software. It was demonstrated that obtained models are able to capture the dynamics of the time series data and to produce sensible forecasts.

**Key words:** regression models, forecast, time series, meteorological quantities

### INTRODUCTION

The prediction of the future courses of meteorological quantities on the basis of historical time series is important for agrophysical modelling (Lamorski *et al.*, 2013; Baranowski *et al.*, 2015; Murat *et al.*, 2016; Krzyszcak *et al.*, 2017a). All the crop production models are highly sensitive to climatic and environmental variations (Fronzek *et al.*, 2018; Pirttioja *et al.*, 2015; Porter and Semenov, 2005; Ruiz-Ramos *et al.*, 2018) and the temporal and space scaling properties of the weather time series should be considered when applying weather time series as the inputs to

these models (Hoffmann *et al.*, 2017; Krzyszcak *et al.*, 2017b; Walczak *et al.*, 1997). When taking into account the global warming effects on the processes occurring in the soil-plant-atmosphere system, the shifts in future weather patterns and the increase in frequency and magnitude of extreme events should be known (Lobell *et al.*, 2012; Semenov and Shewry, 2011; Sillmann and Roeckner, 2008). Increasing temperature and limited precipitation, which are responsible for drought incidence as a result of global warming, are posing serious threats to food security (Lobell *et al.*, 2013). The forecasting of these two quantities using statistical methods is, therefore, of great importance.

Many time series forecasting methods are based on the analysis of historical data. They assume that past patterns in the data can be used to forecast future events. In recent years, one of the most popular ways of time series modelling is autoregressive integrated moving-average (ARIMA) modelling. Its main aim is to carefully and rigorously study the past observations of a time series to develop an appropriate model which can predict future values for the series. It has three control constants *i.e.* irregular, trend and seasonal influence, which can control and manage influence of time segmentation through the specific time duration. In literature, ARIMA models have been widely used for various applications such as medicine, business, economics, finance and engineering. Moreover, ARIMA models have become, in last decades, a major tool in numerous meteorological applications to understand the phenomena of air temperature and precipitation. El-Mallah and Elsharkawy (2016) showed that the linear ARIMA model and the quadratic ARIMA model had the best overall performance in making short-term predictions of annual absolute

\*Corresponding author e-mail: jkrzyszcak@ipan.lublin.pl

\*\*This work was co-funded by the Polish National Centre for Research and Development as part of the GyroScan project contract No. BIOSTRATEG2/298782/11/NCBR/2016. We acknowledge the data providers in the European Climate Assessment & Dataset project (ECA&D) – for Lleida, the State Meteorological Agency. Data and metadata are available at [www.ecad.eu](http://www.ecad.eu). This work was also part of the FACCE JPI Knowledge Hub ‘Modelling European Agriculture with Climate Change for Food Safety’ project (MACSUR) contract FACCE JPI/06/2012).

**Table 1.** The basic characteristics of the sites and their agro-climatic conditions

Parameter	Site/Country			
	Dikopshof Germany (DE)	Jokioinen Finland (FI)	Lleida Spain (ES)	Lublin Poland (PL)
Latitude (°N)	50°48'29"	60°48'	41°42'	51°14'55"
Longitude (°E)	6°57'7"	23°30'	1°6'	22°33'37"
Altitude (m)	60	104	337	194
Environmental zone	Atlantic Central	Boreal	Mediterranean South	Continental
Köppen-Geiger climate classification	Cfb	Dfc	BSk	Dfb

temperature in Libya. Balyani *et al.* (2014) used ARIMA model in a 50-year time period (1955-2005) for Shiraz, south of Iran. Their modelling of temperature selected ARIMA as the optimal model. Furthermore, Anitha *et al.* (2014) used the seasonal autoregressive integrated moving average (SARIMA) model to forecast the monthly mean of the maximum surface air temperature of India. Their results showed that there is a trend in the monthly mean of maximum surface air temperature in India. Muhammet (2012) also used the ARIMA method to predict the temperature and precipitation in Afyonkarahisar Province, Turkey, until the year 2025, and found an increase in temperature according to the quadratic and linear trend models. Finally, Khedhiri (2014) studied the statistical properties of historical temperature data in Canada for the period 1913-2013 and determined a seasonal ARIMA model for the series to predict future temperature records.

Akpanta *et al.* (2015) adopted the SARIMA modelling of the frequency approach in analysing monthly rainfall data in Umuahia. Abdul-Azziz *et al.* (2013) and Afrifa-Yamoah (2016) forecasted monthly rainfall in several regions in Ghana with the use of the SARIMA models. The SARIMA models of the weekly and monthly rainfall time series of two selected weather stations in Malaysia were built by Yusof and Kane (2012) and in India by Dabral and Murry (2017). The mentioned ARIMA models have a good post-sample forecasting performance for yearly and monthly agrometeorological time series.

Another approach in forecasting the meteorological time series involves fitting regression models (RM) to time series including trend and seasonality components. The RM models are originally based on linear modelling, but they also allow parameters such as trend and season to be added to the data. In our study, the trend parameter will be fitted with polynomial function, and the season parameter will be estimated with Fourier series.

The aim of this paper is to examine the statistical properties of the daily mean air temperature and the precipitation time series from four different locations in Europe and to

develop predictive models to forecast the daily mean values of these quantities up to six years ahead, using the above-mentioned methods with external regressors in the form of Fourier terms.

#### MATERIALS

We decided to study four sites from northern, central and southern Europe in order to represent contrasting climatic conditions. Jokioinen in Finland was chosen for northern Europe and Lleida in Spain for southern Europe. For central Europe, two sites were chosen: Dikopshof – located in the west of Germany, and Lublin – in the east of Poland. The chosen sites represent boreal, Atlantic central, continental and Mediterranean south climates. The principal characteristics of these sites and their agro-climatic conditions are summarised in Table 1.

The Jokioinen site has a subarctic climate that has severe winters, no dry season, with cool, short summers and strong seasonality (Köppen-Geiger classification: Dfc, which means continental subarctic climate with the coldest month averaging below 0°C; 1 to 3 months averaging above 10°C and no significant precipitation difference between seasons). Lleida has a semi-arid climate with Mediterranean-like precipitation patterns (annual average of 369 mm), foggy and mild winters and hot and dry summers (Köppen-Geiger classification: BSk, which means dry cold semi-arid climate). Dikopshof represents a maritime temperate climate (Köppen-Geiger climate classification: Cfb, which means temperate oceanic climate with the coldest month averaging above 0°C, all months with average temperatures below 22°C and at least four months averaging above 10°C and no significant precipitation difference between seasons). There is significant precipitation throughout the year in the German site. The Lublin site has a warm summer continental climate (Köppen-Geiger climate classification: Dfb, which means warm-summer humid continental climate with the coldest month averaging below 0°C, all months with average temperatures below 22°C and at least four months averaging above 10°C and no

**Table 2.** Descriptive statistics of the whole daily, 31-year meteorological time series – from four stations in Germany (DE), Finland (FI), Poland (PL) and Spain (ES)

Meteorological variable	Site	Mean	Min	Max	Std	Median	Skewness	Kurtosis
Air temperature (°C)	Dikopshof (DE)	10.2	-16.8	28.9	6.8	10.5	-0.2	2.5
	Jokioinen (FI)	4.6	-33.4	25.0	9.3	4.7	-0.4	2.8
	Lleida (ES)	15.0	-8.3	33.1	7.6	14.7	0.0	2.1
	Lublin (PL)	8.7	-22.8	28.3	8.8	9.1	-0.2	2.4
Precipitation (mm day <sup>-1</sup> )	Dikopshof (DE)	1.7	0.0	75.4	3.8	0.0	4.5	38.1
	Jokioinen (FI)	1.7	0.0	79.1	3.9	0.1	5.0	49.3
	Lleida (ES)	0.9	0.0	83.6	3.8	0.0	7.2	75.7
	Lublin (PL)	1.5	0.0	61.6	3.9	0.0	5.7	49.7

Mean, min, max, standard deviation (Std) and median have the units corresponding to the units of meteorological variables; skewness and kurtosis are non-dimensional.

significant precipitation difference between seasons). The weather time series in all sites were measured with standard equipment, comparable for all stations. In the present study, we focus on an air temperature dataset collected on a daily basis from January 1, 1980 to December 31, 2010 (11 322 days). The descriptive statistics of the meteorological time series are presented in Table 2. The highest mean and median values of air temperature in the period of 31 years were observed at the Lleida station and the lowest at the Jokioinen station. The parameters of skewness and kurtosis of the analysed time series give information about differences in their statistical distributions. Air temperature is characterised by negative skewness and small kurtosis, which inform us that this distributions is left-tailed.

A completely different distribution shape can be observed for precipitation, with higher positive skewness and very high kurtosis values for all the stations. This means that this distribution is strongly right-tailed and has a very sharp peak and a fat tail.

## METHODS

A time series is an ordered sequence of values of a variable at equally spaced time intervals, *i.e.* hourly temperatures at a weather station. The main aim of time series modelling is to carefully analyse and rigorously process the past observations of the time series to develop an appropriate model which describes the inherent structure of the series. This makes it possible to explain the data in such a way as to facilitate prediction, monitoring, or control. There are several approaches to modelling series with a single seasonal pattern. Among these are exponential smoothing (Winters, 1960), seasonal ARIMA models (Box and Jenkins, 1970), state-space models (Harvey, 1989) and the innovations State Space Models (Hyndman

*et al.*, 2008). The ARIMA model was popularised by Box and Jenkins (1970) and Box and Tiao (1975). It is a combination of three mathematical models, using autoregressive, integrated, moving-average (ARIMA) models for time series data. An ARIMA ( $p, d, q$ ) model can account for temporal dependence in several ways. Firstly, the time series is  $d$ -differenced to render it stationary. If  $d = 0$ , the observations are modelled directly, and if  $d = 1$ , the differences between consecutive observations are modelled. Secondly, the time dependence of the stationary process  $\{X_t\}$  is modelled by including  $p$  autoregressive models. The equation for  $p$  is:

$$X_t = c + \sum_{i=1}^p \varphi_i x_{t-i} + \varepsilon_t, \quad (1)$$

where:  $c$  is the constant,  $\varphi$  is the parameter of the model,  $x_t$  is the value that observed at  $t$  and  $\varepsilon_t$  stands for random error. Thirdly,  $q$  are moving-average terms, in addition to any time-varying covariates. It takes the observation of previous errors. The equation for  $q$  is:

$$X_t = \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i}, \quad (2)$$

where:  $\theta_i$  is the parameter of the model,  $\varepsilon_t$  is the error term. Finally, by combining these three models, we get the ARIMA model. Thus, the general form of the ARIMA models is given by:

$$Y_t = c + \sum_{i=1}^p \varphi_i Y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-i}, \quad (3)$$

where:  $Y_t$  is a stationary stochastic process,  $c$  is the constant,  $\varepsilon_t$  is the error or white noise disturbance term,  $\varphi_i$  means autoregression coefficient and  $\theta_j$  is the moving average

coefficient. For a seasonal time series, these steps can be repeated according to the period of the cycle, whatever time interval. Usually, ARIMA models are described using the backward operator  $B$  defined as:

$$B_k(X_t) = X_{t-k} \quad t > k; t, k \in N, \tag{4}$$

where:  $k$  is the index denoting how many times backward operator  $B$  is applied to time series  $X_t$  characterised by time interval  $t$ , and  $N$  is the total number of time intervals. By employing the following notation:

$$\phi(z) = 1 - \sum_{i=1}^p \phi_i z^i, \quad \phi_p \neq 0, \tag{5}$$

$$\theta(z) = 1 - \sum_{i=1}^q \theta_i z^i, \quad \theta_q \neq 0, \tag{6}$$

the Eq. (1) can be written, respectively, as:

$$\phi(B)(1 - B)^d Y_t = c + \theta(B)\varepsilon_t. \tag{7}$$

The seasonal ARIMA  $(p, d, q) (P, D, Q)_m$  process noted also as SARIMA  $(p, d, q) (P, D, Q)_m$  is given by:

$$\Phi(B^m)\phi(B)(1 - B^m)^D(1 - B)^d Y_t = c + \Theta(B^m)\theta(B)\varepsilon_t, \tag{8}$$

where:  $m$  is the seasonal period,  $\Phi(z)$  and  $\Theta(z)$  are polynomials of orders  $P$  and  $Q$ , respectively, each containing no roots inside the unit circle. If  $c \neq 0$ , there is an implied polynomial of order  $d + D$  in the forecast function (Box *et al.*, 2008; Brockwell and Davis, 1991). To determine a proper model for a given time series data, it is necessary to carry out the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analysis, which reflect how the observations in a time series are interrelated. The plot of ACF helps to determine the order of Moving Average terms, and the plot of PACF helps to find Auto-Regressive terms.

The main task in SARIMA forecasting is selecting an appropriate model order; that is, if the values  $p, q, P, Q, D, d$ . If  $d$  and  $D$  are known, we can select the orders  $p, q, P$  and  $Q$  via one of the forecast measure error: the mean absolute error ( $MAE$ ), the root mean squared error ( $RMSE$ ) and the mean absolute scaled error ( $MASE$ ).  $MAE$  and  $RMSE$  are defined by the formulas:

$$MAE = \frac{1}{n} \sum_{t=1}^n |e_t|, \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}, \tag{10}$$

respectively, where  $n$  is the number of periods of time and  $e_t = y_t - f_t$  is the forecast error between the actual value  $y_t$  and the forecasted value  $f_t$ . The  $MAE$  is the average over the verification sample of the absolute values of the differences between the forecast and the corresponding observation. Moreover, the  $RMSE$  is the square root of the average squared values of the differences between forecast and the corresponding observation. These errors have the same units of measurement and depend on the units in which the data are measured. The  $MASE$  was proposed by Hyndman and Koehler (2006) for comparing forecast accuracies. The  $MASE$  is given by the formula:

$$MASE = \frac{MAE}{Q}, \tag{11}$$

where:  $Q$  is a scaling statistic, computed on the training data. For a non-seasonal time series, a useful way to define scaling statistics is to apply the mean absolute difference between the consecutive observations:

$$Q = \frac{1}{N-1} \sum_{j=2}^n |y_j - y_{j-1}|, \tag{12}$$

that is,  $Q$  is the  $MAE$  for naive forecasts, computed on the training data. The  $MASE$  is less than one if it arises from a better forecast than the average naive forecast computed on the training data. Conversely, it is greater than one if the forecast is worse than the average naive forecast computed on the training data. For a seasonal time series, a scaling statistic can be defined using the seasonal naive forecasts:

$$Q = \frac{1}{N-m} \sum_{j=m+1}^n |y_j - y_{j-m}|, \tag{13}$$

where the seasonal naive method accounts for seasonality by setting each prediction to be equal to the last observed value of the same season. The  $MASE$  is independent of the scale of the data, so it can be used to compare forecasts for data sets with different scales. When comparing forecasting methods, the method with the lowest  $MASE$  is the preferred one.

Sometimes the SARIMA model does not tend to give good results for the time series with a period greater than 200 years. In such a situation, the simplest approach is a regression with ARIMA errors, where the order of the ARIMA model and the number of Fourier terms is selected by minimising the  $RMSE$ ,  $MAE$  or  $MASE$ . In such models, external regressors in the form of Fourier terms are added to an ARIMA  $(p, d, q)$  model to account for the seasonal behaviour. We can consider ARIMA models with regressors as a regression model which includes a correction for autocorrelated errors. Hence, we can add ARIMA terms to the regression model to eliminate the autocorrelation and

further reduce the mean squared error. To do this, we re-fit the regression model as an ARIMA ( $p, d, q$ ) model with regressors, and specify the appropriate AR( $p$ ) or MA( $q$ ) terms to fit the pattern of autocorrelation we observed in the original residuals. To be more precise, we consider the following model:

$$y_t = c + \sum_{l=1}^K \left[ \alpha_l \sin \frac{2\pi lt}{m} + \beta_l \cos \frac{2\pi lt}{m} \right] + U_t, \quad (14)$$

where:  $U_t$  is an ARIMA process,  $\alpha_l$  and  $\beta_l$  are Fourier coefficients and  $m$  is a length of period. The value of  $K$  is chosen by minimising forecast error measures. For the purpose of this paper, this process will be noted as ARIMAF ( $p, d, q$ ) [K]. According to Hyndman (2010), the main advantages of this approach are as follows: (i) it allows any length seasonality for the data with more than one seasonal period, (ii) Fourier terms of different frequencies can be included, (iii) the seasonal pattern is smooth for small values of  $K$  and (iv) the short-term dynamics is easily handled with a simple ARMA error. The only real disadvantage (compared to a seasonal ARIMA model) is that the seasonality is assumed to be fixed (the pattern is not allowed to change over time), but in our situation, seasonality was remarkably constant (Fig. 1).

In this study, we also use two regression models in the basic form:

$$Y_t = b_t + s_t + \varepsilon_t, \quad (15)$$

where  $b_t$  and  $s_t$  represent the trend and the seasonal components of the time series at time  $t$ , respectively. In the first regression model (RMP), the trend in time series data is fitted with the use of a polynomial by including time as a predictor variable:

$$Y_t = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \dots + \gamma_n t^n + s_t + \varepsilon_t, \quad (16)$$

where the degree of polynomial  $n$  is chosen by minimising prediction errors.

In the second regression model (RMF), we apply the Fourier series to model the seasonal component in time series data as follows:

$$Y_t = b_t + \sum_{l=1}^L \left( \alpha_l \cos \frac{2\pi lt}{m} + \beta_l \sin \frac{2\pi lt}{m} \right) + \varepsilon_t, \quad (17)$$

where:  $L$  is chosen by minimising prediction errors. The results of the parameters of the proposed models are obtained from the output of RStudio integrated development environment for R version 0.97.551 (R Core Team, 2014).

## RESULTS AND DISCUSSION

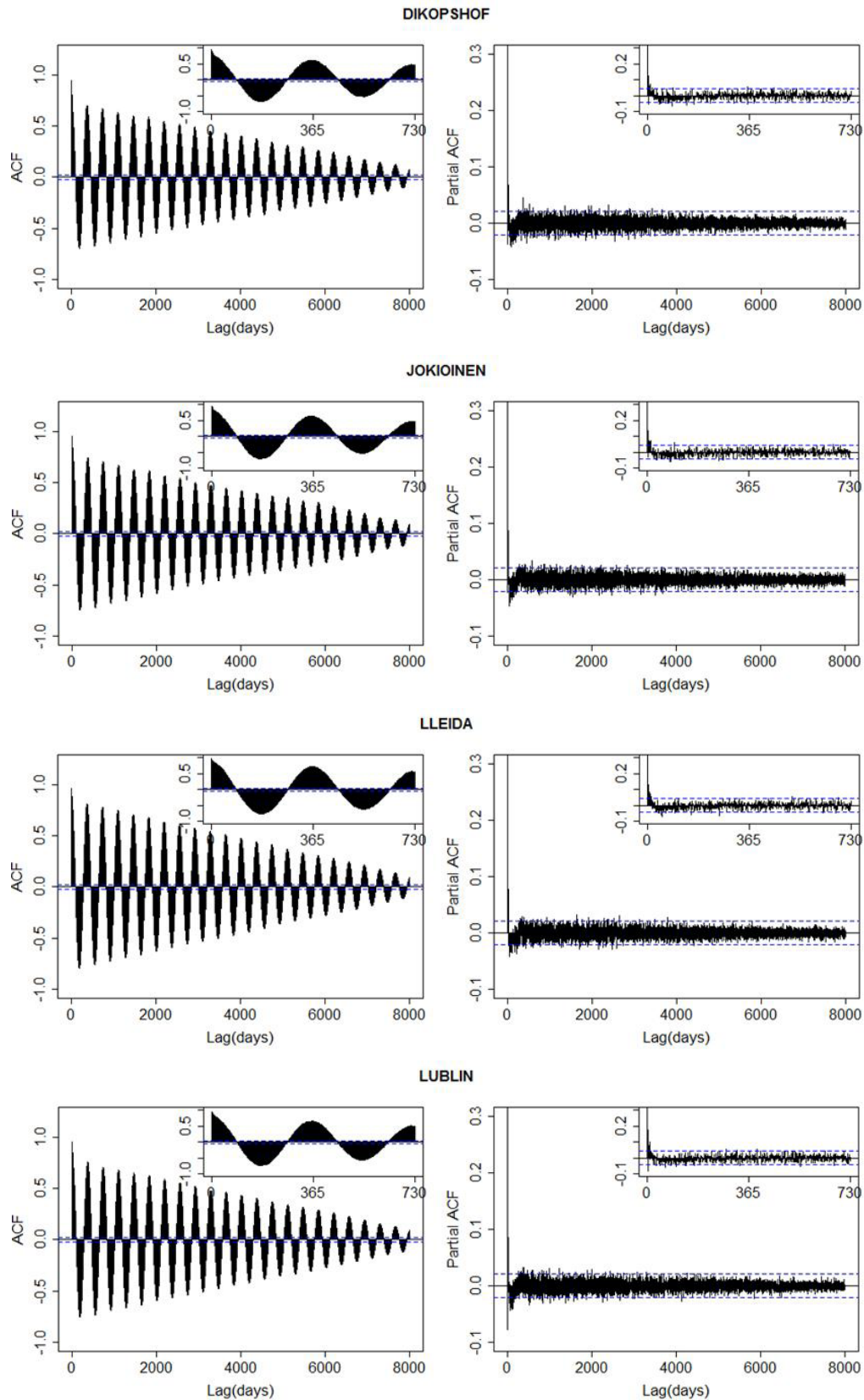
Air temperature and precipitation data recorded in four European sites from different climatic zones are used to create suitable SARMA, ARIMAF, RMP and RMF mo-

del. Several authors have done similar analysis in the last decade, albeit, most mainly considered ARIMA or SARIMA models for weekly, monthly or yearly time series. For example, Mahsin *et al.* (2012) analysed the monthly rainfall data of Dhaka district based on the SARIMA(0,0,1)(0,1,1)<sub>12</sub> model; Zakaria *et al.* (2012) used the weekly rainfall in the semi-arid Sinjar District at Iraq and found that SARIMA (3,0,2) (2,1,1)<sub>30</sub>, SARIMA (1,0,1) (1,1,3)<sub>30</sub>, SARIMA (1,1,2) (3,0,1)<sub>30</sub> and SARIMA (1,1,1) (0,0,1)<sub>30</sub> models were developed with the highest precision with regard to data obtained from four stations.

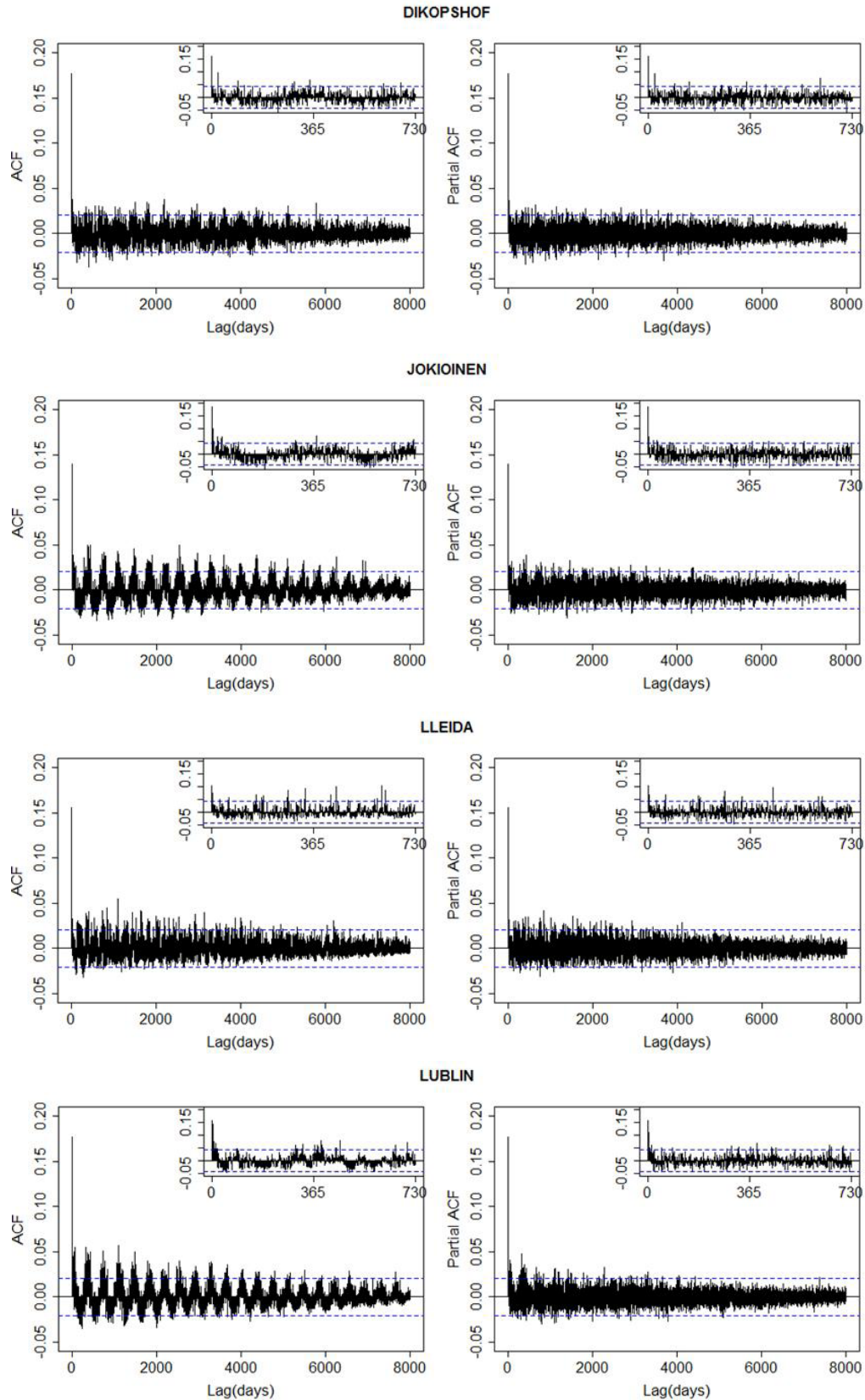
What is more, Abdul-Aziz *et al.* (2013) used the SARIMA(0,0,0)(2,1,0)<sub>12</sub> model for forecasting the monthly rainfall of the Ashanti Region of Ghana, while Ampaw *et al.* (2013) analysed the monthly rainfall data of the Eastern Region of Ghana and showed that SARIMA (0,0,0) (2,1,1)<sub>12</sub> model is the most accurate. The SARIMA (0,0,0) (1,1,1)<sub>12</sub> model has also been identified by Afrifa-Yamoah *et al.* (2016), as an appropriate model for predicting monthly average rainfall figures for the Brong Ahafo Region of Ghana. In addition, Yusof and Kane (2012) used rainfall data and showed that SARIMA(1,1,2)(1,1,1)<sub>12</sub> and SARIMA(4,0,2)(1,0,1)<sub>12</sub> models for two stations in Malaysia are adequate. What is more, Osarumwense (2013) used quarterly rainfall data and showed that the SARIMA (0,0,0) (2,1,0)<sub>4</sub> model was appropriate, Etuk *et al.* (2013) identified and established the adequacy of SARIMA (5,1,0) (0,1,1)<sub>12</sub> for modelling and forecasting the amount of monthly rainfall in Portharcourt, Nigeria. Anitha *et al.* (2014) selected the SARIMA(1,0,1)(0,1,1)<sub>12</sub> model to predict the monthly mean of maximum surface air temperature in India, while Balyani *et al.* (2014) found ARIMA (1,1,3) to be the optimal model for the annual temperature in Shiraz.

In addition to the aforementioned, annual surface absolute temperature from 16 stations situated on the coast of Libya were modelled with ARIMA (3,1,2) and ARIMA (3,2,3) by El-Mallah *et al.* (2016). Furthermore, an alternative model was run by Khedhiri (2014) for maximum and minimum mean temperature records for the Canadian province of Prince Edward Island. He suggested to fit maximum monthly temperature series to the SARIMA (2,0,1)(2,0,0)<sub>12</sub> model and to fit minimum temperature series to the SARIMA (1,0,1)(1,0,1)<sub>12</sub> model. Tanusree and Kishore (2016), based on knowledge of automatic ARIMA forecasting, selected SARIMA (1,0,0)(0,2,2)<sub>12</sub>, SARIMA (12,0,0)(0,1,1)<sub>12</sub>, SARIMA (0,0,10) (0,1,1)<sub>12</sub> and SARIMA (0,0,1)(0,1,1)<sub>12</sub> for the stations of Guwahati, Tezpur, Silchar and Dibrugarh (Assam, India). They found the selected models were adequate to represent the temperature data and could be used to forecast the upcoming temperature.

In our work, we studied data from 11,323 days in the period between January 1, 1980 and December 31, 2010. In our research, the data set was divided into a training set and a test set. All the observations from January 1, 1980 to December 31, 2004 were used as the training set and were



**Fig. 1.** Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for time series of mean daily air temperature for the studied stations. The larger plots contain lags covering the test time range of 25 years, whereas the smaller inside plots cover a range of 2 years.



**Fig. 2.** Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for time series of daily precipitation for the studied stations. The larger plots contain lags covering the test time range of 25 years, whereas the smaller inside plots cover a range of 2 years.

**Table 3.** Forecast accuracy measures of all forecasting approaches for a daily mean air temperature time series, with p-value derived from a Ljung-Box test

Site	Model	<i>RMSE</i>	Rank	<i>MAE</i>	Rank	<i>MASE</i>	Rank	p-value lag=365
Dikopshof	ARIMAF(3,0,1)[ <i>K</i> =1]	3.633	1	2.915	1	0.738	1	0.038
	RMF(1)	3.636	2	2.917	2	0.739	2	0 <sup>1</sup>
	RMP(3)	3.715	3	2.970	3	0.752	3	0 <sup>1</sup>
	SARIMA(0,0,1)(0,1,0) <sub>365</sub>	4.914	4	3.904	4	0.989	4	0 <sup>1</sup>
Jokioinen	ARIMAF(3,0,2)[ <i>K</i> =7]	4.536	1	3.563	2	0.768	2	0.648
	ARIMAF(2,0,3)[ <i>K</i> =7]	4.536	2	3.563	3	0.768	3	0.668
	RMF(7)	4.549	3	3.575	4	0.771	4	0 <sup>1</sup>
	RMP(1)	4.572	4	3.527	1	0.761	1	0 <sup>1</sup>
	SARIMA(3,0,1)(0,1,0) <sub>365</sub>	5.582	4	4.340	4	0.9367	4	0 <sup>1</sup>
Lleida	ARIMAF(3,0,3)[ <i>K</i> =3]	3.023	1	2.480	2	0.753	2	0.977
	RMF(3)	3.024	2	2.480	3	0.753	3	0 <sup>1</sup>
	RMP(1)	3.061	3	2.471	1	0.751	1	0 <sup>1</sup>
	SARIMA(0,0,1)(0,1,0) <sub>365</sub>	4.120	4	3.285	4	0.998	4	0 <sup>1</sup>
Lublin	ARIMAF(2,0,2)[ <i>K</i> =4]	4.156	1	3.309	1	0.734	1	0.558
	RMF(4)	4.173	2	3.3237	4	0.737	4	0 <sup>1</sup>
	RMP(2)	4.190	3	3.292	3	0.730	3	0 <sup>1</sup>
	RMP(1)	4.190	4	3.290	2	0.730	2	0 <sup>1</sup>
	SARIMA(0,0,1)(0,1,0) <sub>365</sub>	5.331	5	4.192	5	0.930	5	0 <sup>1</sup>

p-value < 0.01.

applied so as to fit the created statistical models for the mean air temperature and the precipitation. The data from January 1, 2005 to December 31, 2010 were designated as the test set and were used to assess the predictability accuracy of the fit. This approach gives the ability to compare the effectiveness of different methods of prediction.

Firstly, the plots of the considered series and their autocorrelation functions (ACF) and partial autocorrelation functions (PACF), plotted in Figs 1 and 2, were examined to establish the potential performances of SARIMA, ARIMA, ARIMAF, RMP and RMF models for the daily air temperature and the precipitation series. In handling such data, the mean and the variance of the mean air temperature time series are not functions of time, but rather are constants and their covariance of the *k*-th term and the (*k* + 365)-th term does not depend on time. The same behaviour is shown by the precipitation time series in Jokioinen and Lublin. This means that these eight series are stationary. Moreover, we use statistical testing to verify

the stationarity of considered series. In doing this, there are two different approaches: stationarity tests such as the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test that consider as null hypothesis that the series is stationary, and unit root tests, such as the Dickey-Fuller test and its augmented version, the augmented Dickey-Fuller test (ADF), or the Phillips-Perron test (PP), for which the null hypothesis is the contrary – that the series possesses a unit root and hence is not stationary. ADF and PP tests verified the stationarity of our series with *p*-value smaller than 0.01, which was also confirmed by the KPSS test. Therefore, the series did not require the trend differencing, so in all ARIMA models, we assumed *d*=0. Additionally, the ACF plots depict a sine wave and show spikes in the seasonal lags 365, 730 and 1095. This effect significantly supports the evidence of seasonality in the data sets (except for precipitation in Dikopshof and Lleida). The precise study of the plots given in Figs 1, 2b and 2d suggests the possibility of using SARIMA (*p*,*d*,*q*) (*P*,*D*,*Q*)<sub>*m*</sub> models with *m* = 365, *D* = 1,



**Table 4.** Forecast accuracy measures of all forecasting approaches for daily precipitation time series with p-value derived from a Ljung-Box test

Site	Model	<i>RMSE</i>	Rank	<i>MAE</i>	Rank	<i>MASE</i>	Rank	p-value lag=365
Dikopshof	ARIMA(2,0,1)	3.749	1	2.261	1	0.812	1	0.766
	ARIMA(1,0,3)	3.749	2	2.261	2	0.812	2	0.777
	RMP(1)	3.810	3	2.282	3	0.820	3	0 <sup>1</sup>
	ARIMAF(0,0,3)[ <i>K</i> =7]	3.581	1	2.227	4	0.812	4	0.469
	ARIMAF(3,0,1)[ <i>K</i> =7]	3.581	2	2.227	3	0.812	3	0.666
Jokioinen	RMF(7)	3.581	3	2.227	5	0.812	5	0 <sup>1</sup>
	RMP(2)	3.640	4	2.220	2	0.809	2	0 <sup>1</sup>
	RMP(1)	3.641	5	2.191	1	0.799	1	0 <sup>1</sup>
	SARIMA(3,0,0)(0,1,0) <sub>365</sub>	5.903	6	2.945	6	1.074	6	0 <sup>1</sup>
	ARIMA(3,0,3)	3.479	1	1.503	2	0.863	2	0.780
Lleida	RMP(1)	3.905	2	1.636	3	0.940	3	0 <sup>1</sup>
	RMP(4)	3.948	3	1.425	1	0.819	1	0 <sup>1</sup>
	ARIMAF(3,0,1)[ <i>K</i> =2]	3.996	1	2.142	5	0.890	5	0.04
	RMF(2)	3.996	2	2.141	4	0.889	4	0 <sup>1</sup>
Lublin	ARIMAF(2,0,1)[ <i>K</i> =3]	3.999	3	2.141	2	0.889	2	0.025
	RMF(3)	3.999	4	2.141	3	0.889	3	0 <sup>1</sup>
	RMP(1)	4.057	5	2.280	6	0.946	6	0 <sup>1</sup>
	RMP(3)	4.245	6	2.038	1	0.846	1	0 <sup>1</sup>
	SARIMA(1,0,1)(0,1,0) <sub>365</sub>	5.348	7	2.426	7	1.083	7	0 <sup>1</sup>

p-value < 0.01.

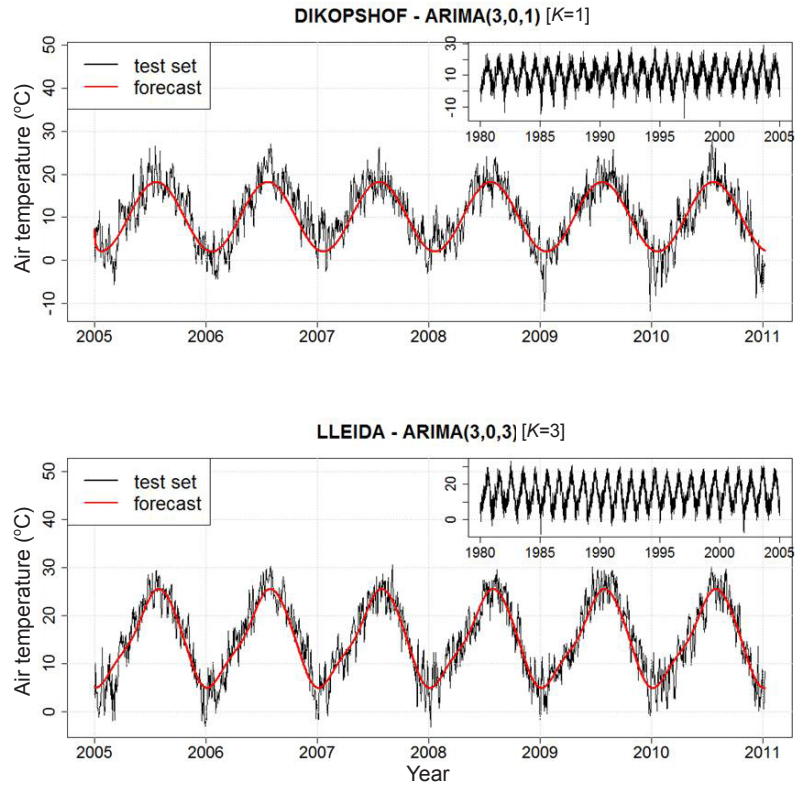
$Q = 0$ ,  $P = 0$  and  $p$ ,  $q$  changing from 0 to 3. Secondly, by minimising the forecast measure errors *RMSE*, *MAE* and *MASE*, we chose the best parameters of SARIMA models among the considered 16 models for each of the studied sites. SARIMA (0,0,1) (0,1,0)<sub>365</sub> and SARIMA (3,0,1) (0,1,0)<sub>365</sub> models were selected as the most appropriate from all 16 tested SARIMA models for air temperature, whereas the SARIMA (3,0,0) (0,1,0)<sub>365</sub> and the SARIMA (1,0,1) (0,1,0)<sub>365</sub> models were most appropriate from all 16 of the tested SARIMA models for precipitation.

In order to establish ARIMAF models, we tried parameters  $p$  and  $q$  between 0 and 3, while the number of the Fourier terms  $K$  varied between 1 and 10. Therefore, for each of seasonal series, we tested 160 cases. The models which were very close to the actual data were chosen by minimising the *RMSE*, *MAE* and *MASE*. Among those cases, we selected ARIMAF (3,0,1) [ $K=1$ ], ARIMAF

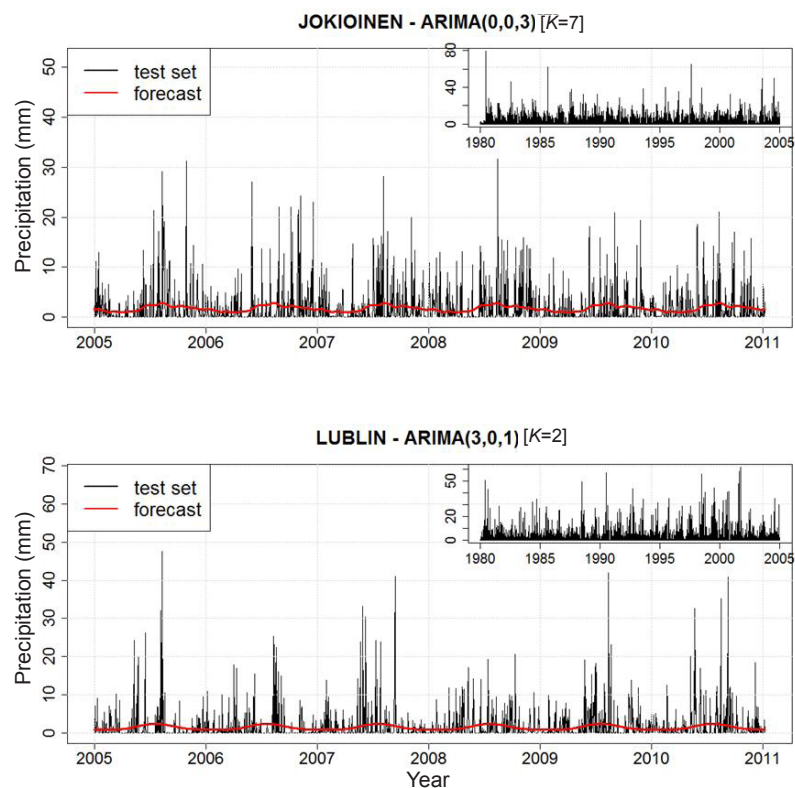
(3,0,2) [ $K=7$ ], ARIMAF (3,0,3) [ $K=3$ ] and ARIMAF (2,0,2) [ $K=4$ ] models for air temperature and ARIMAF (0,0,3) [ $K=7$ ] and ARIMAF(3,0,1) [ $K=2$ ] models for precipitation.

The precipitation time series for Dikopshof and Lleida did not show seasonal behaviour in ACF and PACF plots, and therefore are forecasted with use of ARIMA( $p,0,q$ ) models with  $p$ ,  $q$  changing from 0 to 3. In this case, we considered 16 different models for non-seasonal data sets. The ARIMA (2,0,2), ARIMA (1,0,3) and ARIMA (3,0,3) were selected as models in which forecasts can be built with the smallest *RMSE*, *MAE* or *MASE* – from all 16 tested ARIMA models.

To choose the best regression model, we ran the R software with  $n$  or  $K$  varying between 1 and 10. The smallest forecast errors gave RMP models with the polynomial order equal to 1, 2 or 3 for air temperature and 1, 2 or 4



**Fig. 3.** Smoothed time series of air temperature and respective forecasting results with the smallest *RMSE*. The smaller plot inside covers real data from the learning set.



**Fig. 4.** Smoothed time series of precipitation and respective forecasting results with the smallest *RMSE*. The smaller plot inside covers real data from the learning set.

for precipitation, depending on the site. Furthermore, the smallest forecast errors were achieved for the RMF models with polynomial orders of 1, 3, 4 or 7 for air temperature and 2, 3 or 7 for precipitation.

The accuracy of the selected models for air temperature is shown in Table 3 and for precipitation (Table 4). From the models listed in Tables 3 and 4, we selected the most adequate model which has the lowest forecast error when comparing predicted data using a suitable test set. The Ljung-Box test was also performed and the obtained p-values are shown in Tables 3 and 4. From the above analysis, it follows that the statistical model which is the most suitable model for the considered data sets depends on the climatic zones they come from. However, the obtained results show that the application of SARIMA, ARIMA, ARIMAF, RMP and RMF models to an air temperature and to a precipitation time series provide valuable insights into the studied data structures and their components, being a good basis for a satisfactory prediction.

Exemplary plots of the forecasts which produce the smallest *RMSE* and pass the Ljung-Box test are presented in Figs 3 and 4.

#### CONCLUSIONS

1. The statistical analysis shows that the daily mean temperature data from the considered four climatic zones exhibit similar behaviour and dynamics, although some statistical parameters differ considerably between these sites.

2. Air temperature and precipitation modelling and its forecasting pose a challenging task for handling any daily time series. In this study, we have shown that ARIMAF models can efficiently capture the course of the air temperature in the studied sites by producing the smallest forecast the root mean squared error and can better forecast long seasonal time series with high frequency.

3. The best fitting model for precipitation depends on the site. The Boreal and Continental precipitation time series are better described by ARIMAF models, while ARIMA models are more appropriate for the Central Atlantic and South Mediterranean data sets. The diagnostic checking confirms the adequacy of the models.

4. The selected models generate forecasts which are constructed on the basis of the learn test and compared to the six-year exact values on the independent test set (out-of-sample accuracy forecast errors).

5. It was demonstrated that in practice, for obtaining reasonable information about the overall forecasting error, one should use more than one measure and that the best model using statistical methodology could vary by changing the data. So, it is recommended to take into consideration all the time series models for any studied area and to assume weather parameters so as to choose the suitable model.

6. Although, the chosen models cannot predict the exact air temperature and precipitation, they can give us information that helps to establish strategies for proper planning of agriculture or can be used as a supplemental tool for environmental planning and decision-making.

#### ACKNOWLEDGEMENTS

We acknowledged the Finnish Meteorological Institute (FMI) for delivering the data for Jokioinen (Venäläinen *et al.* 2005), dr. Holger Hoffmann from INRES, University of Bonn in Germany for delivering the Dikopshof data and dr. Krzysztof Siwek from the Faculty of Earth Sciences and Spatial Management, Maria Curie-Skłodowska University in Lublin, for delivering Lublin data.

**Conflict of interest:** The Authors do not declare conflict of interest.

#### REFERENCES

- Abdul-Aziz A.R., Anokye M., Kwame A., Munyakazi L., and Nsawah-Nuamah N.N.N., 2013.** Modeling and forecasting rainfall pattern in Ghana as a seasonal arima process: The case of Ashanti Region. *Int. J. Humanities Social Sci.*, 3(3), 224-233.
- Afrifa-Yamoah E., Bashiru I.I. Saeed, and Karim A., 2016.** Sarima Modelling and Forecasting of Monthly Rainfall in the Brong Ahafo Region of Ghana. *World Environment*, 6(1), 1-9.
- Akpanta C.A., Okorie I.E., and Okoye N.N., 2015.** SARIMA Modelling of the frequency of monthly rainfall in Umuahia, Abia state of Nigeria. *American J. Mathematics Statistics*, 5, 82-87.
- Ampaw E.M., Akuffo B., Opoku L.S., and Lartey S., 2013.** Time series modeling of rainfall in new Juaben municipality of the Eastern region of Ghana. *Contemporary Res. Business Social Sci.s*, 4(8), 116-129.
- Anitha K., Boiroju N.K., and Reddy P.R., 2014.** Forecasting of monthly mean of maximum surface air temperature in India. *Int. J. Statistika Matematika*, 9(1), 14-19.
- Balyani Y., Niya G.F., and Bayaat A., 2014.** A study and prediction of annual temperature in Shiraz using ARIMA model. *J. Geographic Space*, 12(38), 127-144.
- Baranowski P., Krzyszczak J., Sławiński C., Hoffmann H., Kozyra J., Nieróbca A., Siwek K., and Gluza A., 2015.** Multifractal Analysis of Meteorological Time Series to Assess Climate Impacts. *Climate Res.*, 65, 39-52.
- Box G.E.P. and Jenkins G., 1970.** *Time Series Analysis: forecasting and control.* San Francisco, Holden-Day.
- Box G.E.P., Jenkins G., and Reinsel G., 2008.** *Time series analysis.* Wiley Press, New Jersey, USA.
- Box G.E.P. and Tiao G.C., 1975.** *Intervention Analysis with Applications to Economic and Environmental Problems.* *JASA*, 70, 70-79.
- Brockwell P.J. and Davis R.A., 1991.** *Time Series: Theory and Methods.* 2nd edition. Springer-Verlag, New York.
- Dabral P.P. and Murry M.Z., 2017.** Modelling and Forecasting of Rainfall Time Series Using SARIMA. *Environmental Processes*, 1-21.
- El-Mallah E.S. and Elsharkawy S.G., 2016.** Time-series modeling and short term prediction of annual temperature trend on Coast Libya using the box-Jenkins ARIMA Model. *Advances Res.*, 6(5), 1-11.
- Etuk H.E., Moffat U.I., and Chims E.B., 2013.** Modelling monthly rainfall data of portharcourt, Nigeria by seasonal box-Jenkins method. *Int. J. Sci.*, 2, 60-67.
- Fronzek S., Pirttioja N., Carter T.R., Bindi M., Hoffmann H., Palosuo T., Ruiz-Ramos M., Tao F., Trnka M., Acutis**

- M., Asseng S., Baranowski P., Basso B., Bodin P., Buis S., Cammarano D., Deligios P., Destain M.-F., Dumont B., Ewert F., Ferrise R., François L., Gaiser T., Hlavinka P., Jacquemin I., Kersebaum K.C., Kollas C., Krzyszcak J., Lorite I.J., Minet J., Minguez M.L., Montesino M., Moriondo M., Müller C., Nendel C., Öztürk I., Perego A., Rodríguez A., Ruane A.C., Ruget F., Sanna M., Semenov M.A., Slawiński C., Stratonovitch P., Supit I., Waha K., Wang E., Wu L., Zhao Z., and Rötter R.P., 2018. Classifying multi-model wheat yield impact response surfaces showing sensitivity to temperature and precipitation change. *Agricultural Systems*, 159, 209-224, doi: 10.1016/j.agsy.2017.08.004
- Harvey A., 1989. *Forecasting Structural Time Series Model and the Kalman Filter*. Cambridge University Press, New York.
- Hoffmann H., Baranowski P., Krzyszcak J., Zubik M., Slawiński C., Gaiser T., and Ewert F., 2017. Temporal properties of spatially aggregated meteorological time series. *Agric. Forest Meteorol.*, 234, 247-257, <https://doi.org/10.1016/j.agrformet.2016.12.012>
- Hyndman R., 2010. Forecasting with long seasonal periods. <http://robjhyndman.com/hyndsight/longseasonality>
- Hyndman R.J. and Koehler A.B., 2006. Another look at measures of forecast accuracy. *Int. J. Forecasting*, 22(4), 679-688.
- Hyndman R.J., Koehler A.B., Ord J.K., and Snyder R.D., 2008. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer-Verlag Inc., New York.
- Khedhiri S., 2014. Forecasting temperature record in PEI, Canada. *Letters in Spatial and Resource Sciences*, 9, 43-55, doi 10.1007/s12076-014-0135-x
- Krzyszcak J., Baranowski P., Hoffmann H., Zubik M., and Slawiński C., 2017a. Analysis of Climate Dynamics Across a European Transect Using a Multifractal Method, In: *Advances in Time Series Analysis and Forecasting* (Eds I. Rojas, H. Pomares, O. Valenzuela). Selected Contributions from ITISE 2016. Springer Int. Publishing, Cham., doi:10.1007/978-3-319-55789-2\_8.
- Krzyszcak J., Baranowski P., Zubik M., and Hoffmann H., 2017b. Temporal scale influence on multifractal properties of agro-meteorological time series. *Agric. Forest Meteorol.*, 239, 223-235.
- Lamorski K., Pastuszka T., Krzyszcak J., Slawiński C., and Witkowska-Walczak B., 2013. Soil water dynamic modeling using the physical and support vector machine methods. *Vadose Zone J.*, 12(4), <https://doi.org/10.2136/vzj2013.05.0085>.
- Lobell B.D., Sibley A., and Ortiz-Monasterio J.I., 2012. Extreme heat effects on wheat senescence in India. *Nature Climate Change*, 2, 186-189.
- Lobell D.B., Hammer G.L., Mclean G., Messina C., Roberts M.J., and Schlenker W., 2013. The critical role of extreme heat for maize production in the United States. *Nature Climate Change*, 3, 497-501.
- Mahsin M., Akhter Y., and Begum M., 2012. Modeling rainfall in Dhaka District of Bangladesh using time series analysis. *J. Mathematical Modelling Appl.*, 1, 67-73.
- Muhammet B., 2012. The analyse of precipitation and temperature in Afyonkarahisar (Turkey) in respect of box-Jenkins technique. *J. Academic Social Sci. Studies*, 5(8), 196-212.
- Murat M., Malinowska I., Hoffmann H., and Baranowski P., 2016. Statistical modeling of agrometeorological time series by exponential smoothing. *Int. Agrophys.*, 30(1), 57-66.
- Osarumwense O.I., 2013. Applicability of box Jenkins SARIMA model in rainfall forecasting: A case study of Port-Harcourt South South Nigeria. *Canadian J. Computing in Mathematics, Natural Sciences, Engineering Medicine*, 4(1), 1-4.
- Pirttioja N., Carter T.R., Fronzek S., Bindi M., Hoffmann H., Palosuo T., Ruiz-Ramos M., Tao F., Trnka M., Acutis M., Asseng S., Baranowski P., Basso B., Bodin P., Buis S., Cammarano D., Deligios P., Destain M.-F., Dumont B., Ewert F., Ferrise R., François L., Gaiser T., Hlavinka P., Jacquemin I., Kersebaum K.C., Kollas C., Krzyszcak J., Lorite I.J., Minet J., Minguez M.L., Montesino M., Moriondo M., Müller C., Nendel C., Öztürk I., Perego A., Rodríguez A., Ruane A.C., Ruget F., Sanna M., Semenov M.A., Slawiński C., Stratonovitch P., Supit I., Waha K., Wang E., Wu L., Zhao Z., and Rötter R.P., 2015. Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces. *Climate Research*, 65, 87-105, doi:10.3354/cr01322
- Porter J.R. and Semenov M.A., 2005. Crop responses to climatic variation. *Philosophical Trans. Royal Society B: Biological Sci.*, 360(1463), 2021-2035.
- Ruiz-Ramos M., Ferrise R., Rodríguez A., Lorite I.J., Bindi M., Carter T.R., Fronzek S., Palosuo T., Pirttioja N., Baranowski P., Buis S., Cammarano D., Chen Y., Dumont B., Ewert F., Gaiser T., Hlavinka P., Hoffmann H., Höhn J.G., Jurecka F., Kersebaum K.C., Krzyszcak J., Lana M., Mechiche-Alami A., Minet J., Montesino M., Nendel C., Porter J.R., Ruget F., Semenov M.A., Steinmetz Z., Stratonovitch P., Supit I., Tao F., Trnka M., de Wit A., and Rötter R.P., 2018. Adaptation response surfaces for managing wheat under perturbed climate and CO<sub>2</sub> in a Mediterranean environment. *Agricultural Systems*, 159, 260-274, doi: 10.1016/j.agsy.2017.01.009
- Semenov M.A. and Shewry P.R., 2011. Modelling predicts that heat stress, not drought, will increase vulnerability of wheat in Europe. *Scientific Reports*, 1, 66.
- Sillmann J. and Roeckner E., 2008. Indices for extreme events in projections of anthropogenic climate change. *Climate Change*, 86, 83-104.
- Tanusree D.R. and Kishore K.D., 2016. Modeling of mean temperature of four stations in Assam. *Int. J. Advanced Res.*, 4(12), 366-370.
- Walczak R.T., Witkowska-Walczak B., and Baranowski P., 1997. Soil structure parameters in models of crop growth and yield prediction. *Physical submodels. Int. Agrophysics*, 11, 111-127.
- Winters P.R., 1960. Forecasting sales by exponentially weighted moving averages. *Management Sci.*, 6, 324-342.
- Venäläinen A., Tuomenvirta H., Pirinen P., and Drebs A., 2005. A basic Finnish climate data set 1961-2000-description and illustration. *Finnish Meteorological Institute Reports 5*. Finnish Meteorological Institute, Helsinki, Finland.
- Yusuf F. and Kane I.L., 2012. Modelling monthly rainfall time series using ETS state space and SARIMA models *Int. J. Current Res.*, 4(9), 195-200.
- Zakaria S., Al-Ansari N., Knutsson S., and Al-Badrany T., 2012. ARIMA models for weekly rainfall in the semi-arid Sinjar district at Iraq. *J. Earth Sci. Geotechnical Eng.*, 2(3), 25-55.