



COMPARISON OF CLASSICAL AND THEIL-KENDALL METHODS IN ASSESSING THE SIGNIFICANCE OF LINEAR TREND OF PRECIPITATION IN SOUTH-EASTERN POLAND

Stanisław Węglarczyk
Cracow University of Technology

Abstract

Two methods of linear trend estimation: the ordinary least squares (OLS, parametric) and Theil-Kendall (TK, nonparametric) are compared in the paper. The comparison was made using 65 time series of annual totals, P_a , and annual daily maximum, P_{max} , of precipitation, 30-year long each, recorded in the south-eastern part of Poland (the Upper Vistula catchment). The OLS and TK slope coefficients of trends revealed high similarity for both P_a and P_{max} series. The signs of slopes are the same for 64 sites for P_a and 63 sites for P_{max} with positive signs prevailing: the numbers of decreasing trends for P_a OLS and TK slopes were 3 and 4, respectively, and, for P_{max} , 13 for both OLS and TK slopes. In trend significance testing, both methods produced similar results for P_a time series: out of 16 significant trends, 13 were determined with both OLS and TK at the same sites. For P_{max} series such agreement was found for 4 trends out of 10.

Spatial distribution of significant trends showed a kind of clustering in certain parts of the investigated area.

Key words: annual precipitation total, annual daily maximum precipitation, Upper Vistula River, ordinary least squares linear regression, Kendall-Theil linear regression

INTRODUCTION

Climatic changes are the question of the worldwide, and not only of scientific, concern. They are evidenced by several indices, of which trends in time series of annual mean air temperature are the most known. Climatic changes influences can be expected to be also seen in other hydrometeorological variables such as precipitation (e.g., Gordon *et al.* 1992, Brath *et al.* 1999, Trenberth *et al.* 2003, Cebulska *et al.* 2007, Willems *et al.* 2012, Kendon *et al.* 2014, Razavi *et al.* 2016, Prosdocimi *et al.* 2014, Cebulska 2015, Meshram *et al.* 2016, Svoboda *et al.* 2016.) or streamflow (e.g., Douglas *et al.* 2000, Burn and Hag Elnur 2002, Wrzesiński D., 2009, Prosdocimi *et al.* 2014, Zhang *et al.* 2014).

To study the time changes of an investigated variable a trend analysis is usually applied. Ordinary least squares (OLS) linear regression method is undoubtedly a technique that is the most widely used. The drawback of the method (as all other parametric methods) is the requirement that the distributions of the residuals are known. Usually normal distribution is assumed. This is a strong assumption which may lead to false conclusions when the residuals distribution is highly skewed. False conclusions may be drawn also when outliers are present in the sample.

Nonparametric techniques of trend estimation are an alternative to parametric ones. They do not require the assumption of (normal) distribution of residuals and are robust to outliers. The Kendall-Theil nonparametric linear regression, based on using median instead of mean, is one of such methods.

The aim of this paper is to compare the OLS linear regression and the Kendall-Theil nonparametric linear regression applied to annual totals and annual daily maxima of precipitation recorded in the south-eastern part of Poland, within the Upper Vistula River catchment.

MATERIALS AND METHODS

Area and data

The area selected covers mainly the right part of the Upper Vistula River basin and comprises 65 precipitation gauging stations (Figure 1) with 30-year (1.01.1984 – 31.12.2013, 10958 days) daily precipitation totals time series provided by the Institute of Meteorology and Water Management – National Research Institute (IMGW-PIB). The selection criterion of a time series required that it should be an uninterrupted record of at least 30 years of daily observations.

As the map in Figure 1 shows, the precipitation stations are not distributed uniformly within the area. The spatial density of station locations decreases from west to east. A no-precipitation station area in the north-west part of the catchment area has been made intentionally and does not suggest the lack of stations there.

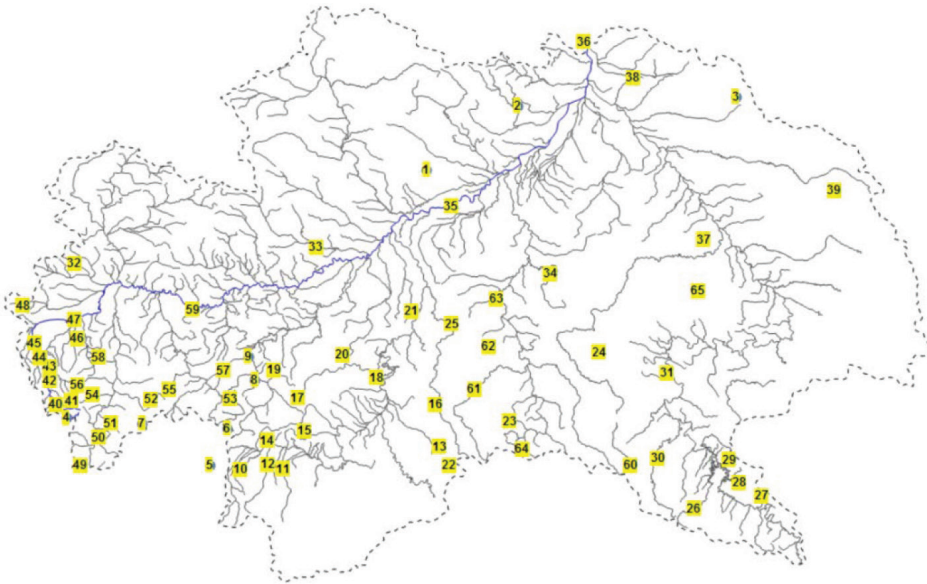


Figure 1. Location of precipitation gauging stations within the Upper Vistula River basin.

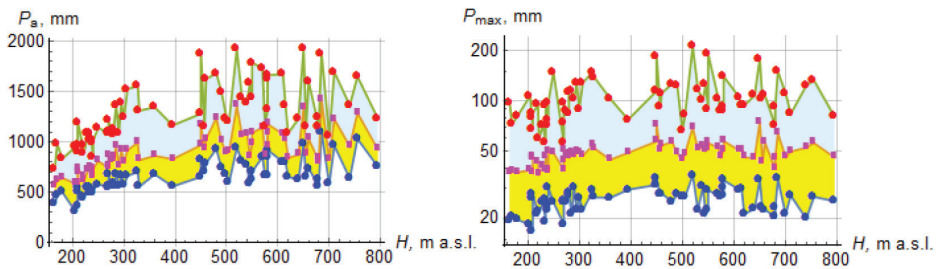


Figure 2. 30-year minimum, mean and maximum of annual precipitation totals, P_a , (left panel) and minimum, mean and maximum of annual maximum daily precipitation totals (log-scale), P_{max} , (right panel) versus gauging station elevation H

Daily precipitation time series at each of 65 stations were the basis for calculation the 30-year time series of annual totals, P_a , and 30-year time series of annual maximum daily precipitation, P_{max} . These time series were the input data for further analysis.

The elevation of gauging stations varied from 165 m to 795 m above sea level (Figure 2). The scattergrams of annual precipitation totals in the left panel of Figure 2 (all three 30-year statistics: minimum, mean and maximum) suggest the existence of trends in the range of elevations up to slightly over 300 m a.s.l.; further with elevation the graphs are not so suggestive. A similar information can be drawn for P_{\max} (right panel of Figure 2) up to ca. 400 m a.s.l. for 30-year minima and means. The course of all statistics shown in Figure 2 is less regular in the range of elevations exceeding 450 m a.s.l.

Figure 3 shows the time course of the P_a and P_{\max} series for all stations, exhibiting the spatial variability of precipitation in each year of record. Annual precipitation totals in the area range from 313.7 to 1928.4 mm, annual daily maxima: from 16.7 to 213 mm. Three years in the left panel of Figure 3 show the largest annual totals (1997, 2001 and 2010; the flood years). The right panel dramatically singles out the 2010 record.

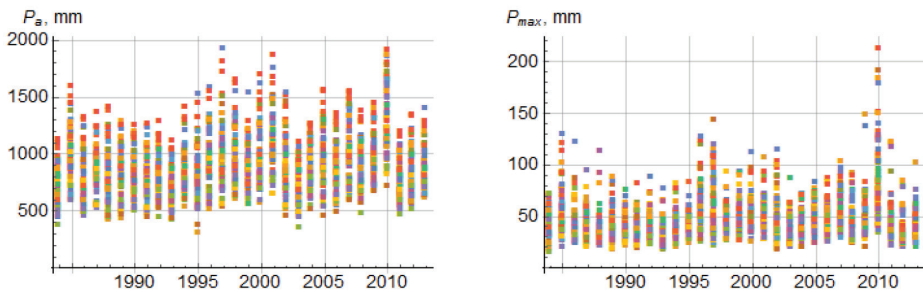


Figure 3. Time course of annual precipitation totals, P_a , and annual daily maxima, P_{\max} , in 65 precipitation gauging stations

Methods: trend estimation and testing

For each gauging station a linear model of time trend of annual precipitation totals, P_a , and annual daily maxima, P_{\max} , was assumed in the following form:

$$x(t) = \alpha_0 + \alpha_1 t \tag{1}$$

where $x(t)$ denotes the expected value of an annual times series of P_a or P_{\max} in year t . The regression coefficients were estimated with two methods: the ordinary least squares (OLS) and Kendall-Theil (TK) method, then trend significance was determined for each trend by testing the null hypothesis $H_0: \alpha_1 = 0$ for slope coefficients estimated with both methods.

The OLS regression parameters were calculated according to the well-known formulas:

$$\hat{\alpha}_{1,OLS} = \frac{\sum_{i=1}^n (x_i - \bar{x})(t_i - \bar{t})}{\sum_{i=1}^n (t_i - \bar{t})^2} \quad (2)$$

$$\hat{\alpha}_{0,OLS} = \bar{x} - \hat{\alpha}_{1,OLS} \bar{t} \quad (3)$$

where $n = 30$ is the sample length, i.e., the number of years of records, and x_i , $i = 1, \dots, n$, are the values of P_a or P_{\max} in year t_i . The significance of the OLS regression was determined by testing the hypothesis $H_0: \alpha_1 = 0$ with the use of the Student's t -test which requires the assumption of normality of the residuals. The procedure of testing with Student's t -test is well known, so it is not presented herein.

The assumption on the distribution of residuals is not required when regression coefficients are calculated by means of the Kendall-Theil nonparametric linear regression method (Theil, 1950; Sen, 1968; Helsel & Hirsch, 2002):

$$\hat{\alpha}_{1,KT} = \text{median}_{\substack{i=1, \dots, n-1 \\ j=i+1, \dots, n}} \left(\frac{x_j - x_i}{t_j - t_i} \right) \quad (4)$$

$$\hat{\alpha}_{0,KT} = \text{median}_{i=1, \dots, n} x_i - \hat{\alpha}_{1,KT} \text{median}_{i=1, \dots, n} t_i \quad (5)$$

The $\hat{\alpha}_{1,KT}$ coefficient (4), called the Theil slope estimate (Helsel & Hirsch, 2002), is the median of all $n(n-1)/2$ possible slope values, while the intercept, $\hat{\alpha}_{0,TK}$, is an analog of (3), with the mean replaced by the median.

Both slope coefficient estimates, $\hat{\alpha}_{1,OLS}$ and $\hat{\alpha}_{1,KT}$, are unbiased estimators of α_1 ; however their variances differ which means that their efficiencies differ. When the assumption of normality of residuals is met, the „OLS is slightly more efficient” (Helsel & Hirsch, 2002) than the TK estimator. When the distribution of residuals is skewed or outliers are present in the sample, the TK estimator, $\hat{\alpha}_{1,KT}$, is „much more efficient” than $\hat{\alpha}_{1,OLS}$ (Helsel & Hirsch, 2002). In summary, it can be said that the TK estimator is almost as efficient as the OLS one when the residuals are normally distributed, and much more efficient in the opposite case.

Testing the hypothesis $H_0: \alpha_1 = 0$ for the Kendall-Theil regression can be done by testing the hypothesis $H_0: \tau = 0$, where τ is the Kendall rank correlation coefficient (Sen, 1968; Helsel & Hirsch, 2002). The τ coefficient can be expressed as follows:

$$\tau = \frac{2S}{n(n-1)} \quad (6)$$

where S is the Kendall's S statistic

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i) \quad (7)$$

and

$$\text{sgn}(x) = \begin{cases} +1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases} \quad (8)$$

Having in mind the definition (6) of τ it is clear that testing hypothesis $H_0: S = 0$ is equivalent to testing $H_0: \tau = 0$ (Helsel & Hirsch, 2002).

RESULTS AND DISCUSSION

For each of 65 gauging stations, both methods: OLS and KT were applied to time series of annual precipitation totals, P_a , and annual daily maxima, P_{\max} . Regression coefficients (equations (2) through (4)) were estimated and the hypothesis $H_0: \alpha_1 = 0$ of no linear trend for both methods was tested.

The estimated values of slope coefficients $\hat{\alpha}_1$ calculated by the OLS (2) and KT (4) methods for all precipitation stations are shown in Figure 4.

The signs of the OLS and KT slope estimates are in almost perfect agreement both for P_a and P_{\max} : different signs occurred only for 1 and 2 stations, respectively. Almost all slope values are positive; for P_a these numbers are 62 and 61 for OLS and KT respectively; for P_{\max} the numbers are lesser, however, high, and equals 52 (80%) for both methods.

The OLS and KT slopes calculated for P_a exhibit high degree of similarity to each other (left panels of Figure 4). This similarity is more clearly visible in the right panels of Figure 4. The OLS slope estimates range between -0.934 and 12.7 mm/year; their TK counterparts lie between -2.04 and 11.4 mm/year. Relating these values to the 30-year average of P_a for each station gives the percentage values of the OLS slopes ranging from -0.107 %/year to 1.17 %/year, and the TK slopes from -0.213 to 1.05 %/year.

For annual maximum daily precipitation, P_{\max} , the differences between the OLS and KT slopes are relatively greater than those for P_a (left panels in Figures 4) and the values of the P_{\max} OLS and KT slopes are more dispersed about the 1:1 line (lower right panel in Figures 4). The OLS estimates for P_{\max} range between -0.418 and 1.28 mm/year; the TK slope values lie between -0.355 and 0.957 mm/year. The respective P_{\max} percent slopes (i.e., related to the 30-year average of P_{\max} for each station) range from -0.908 %/year to 2.36 %/year for OLS, and from -0.701 to 1.43 %/year for TK. As could be expected, the dispersion of P_{\max} percent slopes is much greater than that for P_a .

To see whether the OLS residuals are normally distributed, the Shapiro-Wilk (Shapiro and Wilk 1965) test for normality of residuals of the P_a and P_{\max} OLS trend was performed (Figure 5). At the 5% level, the P_a OLS residuals meet the normality requirement in all but 4 stations; the image for P_{\max} is to large

extent opposite: in 33 cases the OLS residuals do not meet the normality requirement. This is not unexpected as mean skew coefficients for the P_a is 0.61 while that for P_{max} is more than twice larger (1.28).

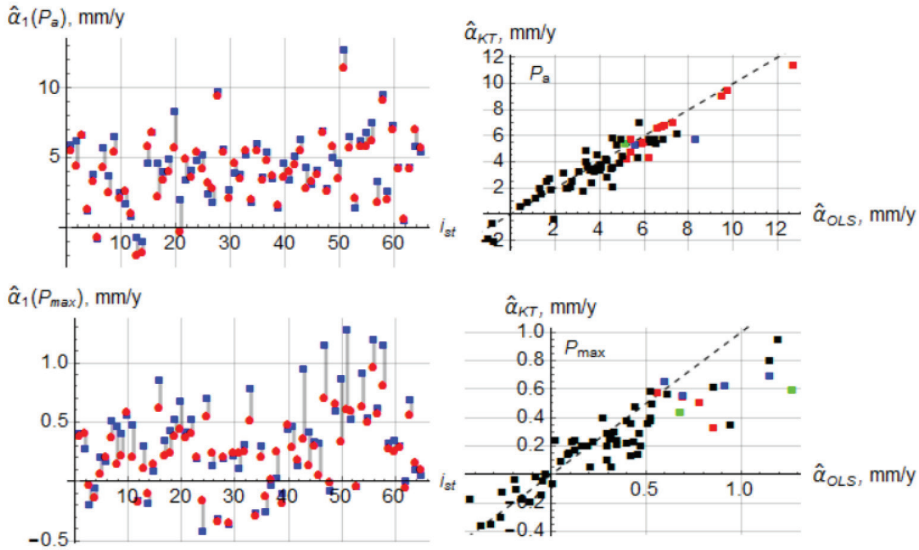


Figure 4. Estimates of slope coefficients $\hat{\alpha}_i$ calculated by the OLS (red discs in left panel) and KT (blue squares in left panel) methods for all precipitation stations. The colours in the right panel are explained in the Figure 6. The station number, i_{st} , corresponds to the numbers in Figure 1.

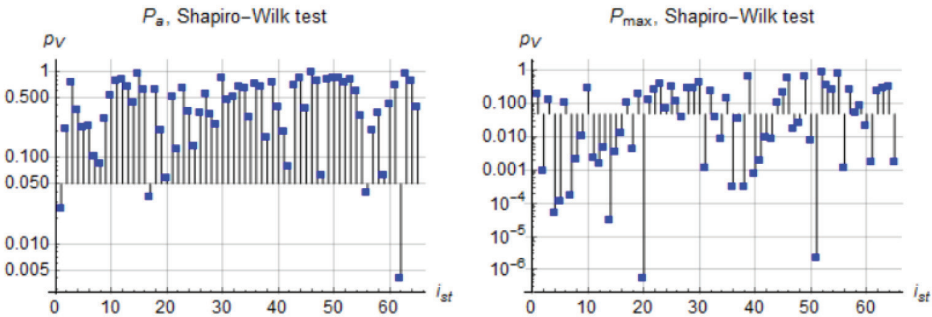


Figure 5. The results of the Shapiro-Wilk test for normality of the OLS residuals. The station number, i_{st} , corresponds to the numbers in Figure 1.

Statistical significance of slope coefficients $\hat{\alpha}_1$ calculated by the OLS (2) and KT (4) methods for all P_a and P_{max} series is shown in details in Figure 6 and summarized in Table 1.

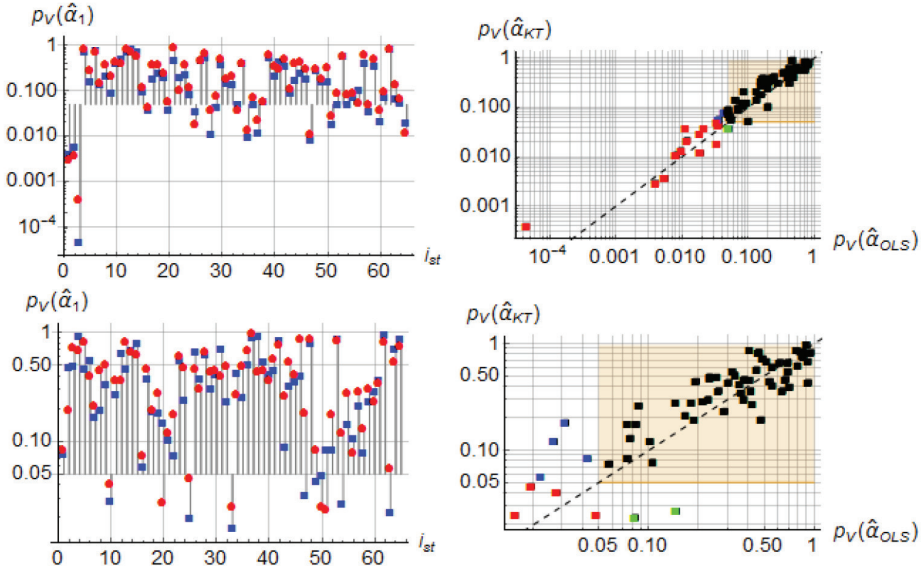


Figure 6. Comparison of the p -values of the OLS and KT slope coefficient estimates for all precipitation stations. Left panels: red discs – OLS, blue squares – KT. Right panels: the colours denote the significance at the 5% level: black – both OLS and KT not significant; red – both OLS and KT significant; blue – OLS significant, KT not significant; green – OLS not significant, KT significant. The marked area depicts the ‘not significant’ results.

No negative slopes are significant at the 5% level. The higher variability P_{max} series reflects in the higher variability of the significance results. This also concerns the comparison of the p -values for the OLS and TK. Nevertheless, as is shown in Table 1, it can be stated that the agreement between the OLS and TK significance testing results is very high: it occurred in 95.4% cases for P_a and in 90.8% cases for P_{max} .

Comparing the results of the Shapiro-Wilk tests with OLS significance test and taking into account the high level of concordance between the OLS and TK slopes estimates, it is a bit surprising a some kind of insensitivity of the OLS method to not meeting the normality requirement. This may be explained by not high skewness of the series.

It is interesting to see the spatial distribution of the signs of the OLS and TK slopes and their significance results.

Table 1. Numbers of concordant and discordant pairs of the OLS and TK slopes resulting from significance testing for 65 P_a and P_{max} precipitation series

OLS slope – TK slope significance at 5% level	Number of cases for P_a		Number of cases for P_{max}	
	Out of 65	%	Out of 65	%
OLS slope not significant and TK slope not significant	49	75.4%	55	84.6%
OLS slope significant and TK slope significant	13	20.0%	4	6.2%
OLS slope significant and TK slope not significant	2	3.1%	4	6.2%
OLS slope not significant and TK slope significant	1	1.5%	2	3.1%

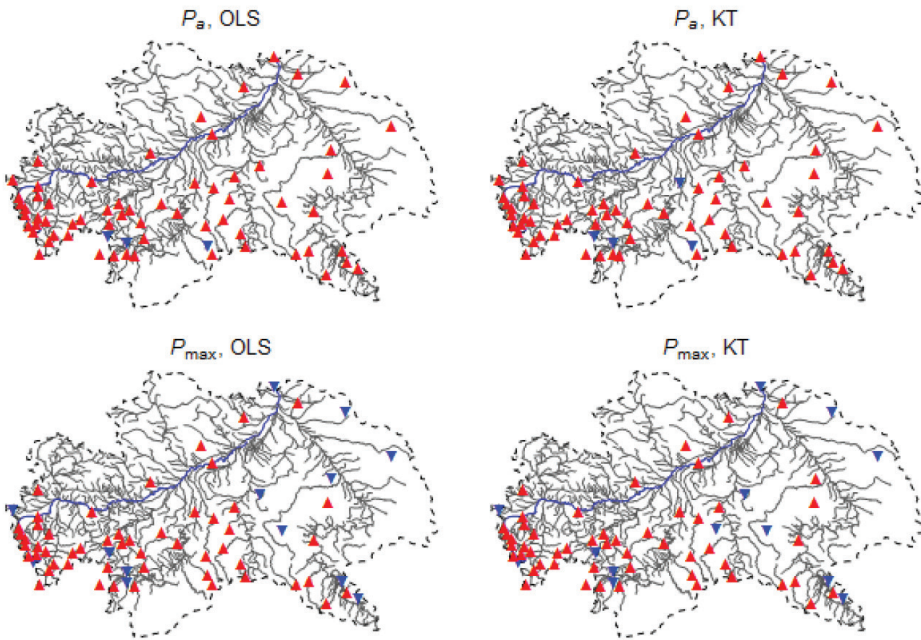


Figure 7. Significant and not significant increasing (red triangles) and decreasing (inverted blue triangles) linear trends calculated by the OLS and TK methods for annual precipitation totals, P_a , and annual daily maxima, P_{max} .

The directions of all 65 OLS and TK trends, independently of their significance, within the studied area are compared in Figure 7. Of all trends, increasing ones are prevailing, especially for annual totals, P_a .

The P_a time series show in most sites increasing trends: only 3 sites exhibit decreasing trends according to the OLS estimation and 4 sites according to the TK. The former 3 trends are among the latter 4 so the OLS and TK slope signs differ in only 1 case.

The number of decreasing trends of P_{max} time series is greater: 13 for OLS and 13 for TK. Both methods show high agreement almost as good as for P_a : in 63 out of 65 sites the signs of OLS slope is the same as for TK. Compared to those for P_a the P_{max} sites with decreasing trends are situated mainly in the eastern part of the area.

All statistically significant linear trends are increasing for both OLS and TK methods and for P_a and P_{max} time series. Their spatial distribution is shown in Figure 8.

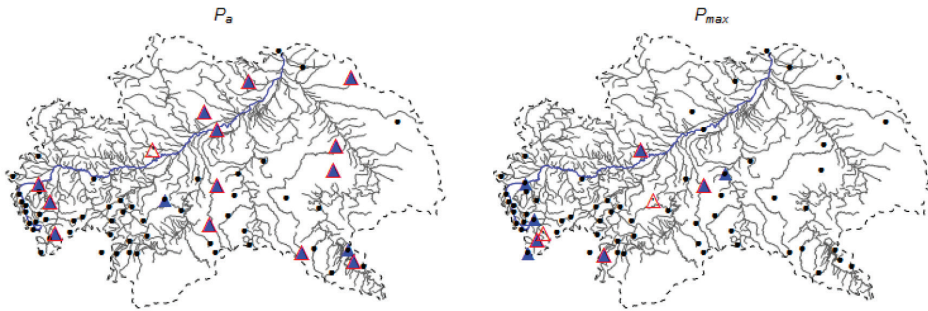


Figure 8. Significant (at the 5% level) linear trends calculated by the OLS (in blue) and KT (in red) methods for annual precipitation totals, P_a , and annual daily maxima, P_{max} . Small dots denote sites with not significant trends.

It is interesting that significant trends seem to cluster in the area. The P_a significant trends show approximately linear clustering along the 3 south-north lines situated in the western, central and eastern parts of the area. For P_{max} , clustering can be seen in the western part; the rest trend is situated in the vicinity of the geometric centre of the catchment.

FINAL REMARKS AND CONCLUSIONS

Two methods of linear trend estimation: the ordinary least squares (OLS, parametric) and Theil-Kendall (TK, nonparametric) were compared for 30-year time series of annual totals, P_a , and annual daily maximum, P_{max} , of precipitation recorded in 65 stations in the south-eastern part of Poland. Independently of their significance, the slope coefficients calculated by both methods revealed high similarity for both type of time series. The signs of slope coefficients were

the same for 64 sites for P_a and 63 sites for P_{max} with positive signs prevailing. Decreasing trends were few: 3 and 4 for P_a OLS and TK slopes, respectively, and, for P_{max} , 13 for both OLS and TK slopes. In trend significance testing, both methods produced similar results for P_a time series: out of 16 significant trends, 13 were found with both OLS and TK in the same precipitation stations. For P_{max} series such agreement was found for 4 trends out of 10.

Spatial distribution of significant trends showed a kind of clustering in certain parts of the investigated area.

ACKNOWLEDGMENTS

The paper has been prepared within the research problem Ś-1/201/2016/DS.

REFERENCES

- Brath, A., Castellarin, A., Montanari, A. (1999). Detecting non stationarity in extreme rainfall data observed in Northern Italy, in: Proceedings of EGS – Plinius Conference on Mediterranean Storms, Maratea, 219–231
- Burn, D.H., Hag Elnur, M.A. (2002). Detection of hydrologic trends and variability. *Journal of Hydrology*, 255, 107–122
- Cebulska M., Twardosz R., Cichocki J. (2007). Zmiany rocznych sum opadów atmosferycznych w dorzeczu górnej Wisły w latach 1881 – 2030, [w:] K. Piotrowicz, R. Twardosz (red.), *Wahania klimatu w różnych skalach przestrzennych i czasowych*, Instytut Geografii i Gospodarki Przestrzennej UJ, Kraków, 383 – 390.
- Cebulska M. (2015). Wieloletnia zmienność maksymalnych opadów dobowych w Kotlinie Orawsko–Nowotarskiej (1984-2013), *Czasopismo Inżynierii Lądowej, Środowiska i Architektury*, 2015, z. 62, nr 3/I, 49-60
- Douglas E.M., Vogel R.M., Kroll C.N. (2000). Trends in floods and low flows in the United States: impact of spatial correlation, *Journal of Hydrology* 240 (2000) 90–105
- Gordon H. B., Whetton P. H., Pittock A. B., Fowler A. M., Haylock M. R. (1992). Simulated changes in daily rainfall intensity due to the enhanced greenhouse effect: implications for extreme rainfall events, *Climate Dynamics* (1992) 8 : 83-102
- Helsel D.R., Hirsch R.M.(2002). *Statistical Methods in Water Resources*, U.S. Geological Survey, Techniques of Water-Resources Investigations Book 4, Chapter A3
- Kendon E.J., Roberts N. M., Fowler H. J., Roberts M. J., Chan S. C., Senior C.A. (2014). Heavier summer downpours with climate change revealed by weather forecast resolution model, *Nature Climate Change* 4, 570–576 (2014) doi:10.1038/nclimate2258

Meshram, S.G., Singh, V.P., Meshram, C. (2016). Long-term trend and variability of precipitation in Chhattisgarh State, India, *Theor. Appl. Climatol.*, doi:10.1007/s00704-016-1804-z

Sen P.K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau, *Journal of the American Statistical Association*, Vol. 63, No. 324. (Dec., 1968), pp. 1379-1389

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, Part I, In the Proceedings of the Royal Netherlands Academy of Sciences **53** (1950), pp 386-392.

Prosdocimi I., Kjeldsen T. R., Svensson C. (2014). Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK. *Natural Hazards and Earth System Sciences*, 14, pp. 1125-1144.

Razavi T., Switzman H., Arain A., Coulibaly P. (2016). Regional climate change trends and uncertainty analysis using extreme indices: A case study of Hamilton, Canada, *Climate Risk Management xxx* (2016) xxx-xxx (article in press)

Shapiro S. S., Wilk M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples) *Biometrika*, Vol. 52, No. 3/4. (Dec., 1965), pp. 591-611.

Svoboda V., Hanel M., Máca P., Kyselý J. (2016). Projected changes of rainfall event characteristics for the Czech Republic. *Journal of Hydrology and Hydromechanics*, 64, DOI: 10.1515/johh-2016-0036

Trenberth K. E., Dai A., Rasmussen R. M., Parsons, D. B. (2003). The changing character of precipitation, *Bull. Am. Meteorol. Soc.* 84, 1205–1217.

Willems P., Arnbjerg-Nielsen K., Olsson J., Nguyen V.T.V. (2012). Climate change impact assessment on urban rainfall extremes and urban drainage: Methods and shortcomings, *Atmospheric Research* 103, 106–118

Wrzesiński D. (2009). Tendencje zmian przepływu rzek polski w drugiej połowie XX wieku. *Badania Fizjograficzne Seria A – Geografia Fizyczna*; 60, 147-162

Zhang Q., Gu X., Singh V.P., Xiao M., Xu C-Y. (2014). Stationarity of annual flood peaks during 1951-2010 in the Pearl River basin, China. *Journal of Hydrology*, 519, 3263-3274.

Prof. dr hab. Stanisław Węglarczyk
Cracow University of Technology
Institute of Water Engineering and Water Management
ul. Warszawska 24
31-155 Kraków
e-mail: sweglar@pk.edu.pl

Received: 04.10.2016

Accepted: 20.10.2016