European Geosciences Union General Assembly 2017 Vienna, Austria, 23-28 April 2017 EGU2017-3711

Dependence of long-term persistence properties of precipitation on spatial and regional characteristics

Hristos Tyralis, Panayiotis Dimitriadis, Theano Iliopoulou, Katerina Tzouka and Demetris Koutsoyiannis, National Technical University of Athens (itia.ntua.gr/1695)

1. Abstract

The long-term persistence (LTP), else known in hydrological science as the Hurst phenomenon, is a behaviour observed in geophysical processes in which wet years or dry years are clustered to respective long time periods. A common practice for evaluating the presence of the LTP is to model the geophysical time series with the Hurst-Kolmogorov process (HKp) and estimate its Hurst parameter *H* where high values of *H* indicate strong LTP.

We estimate *H* of the mean annual precipitation using instrumental data from approximately 1 500 stations which cover a big area of the earth's surface and span from 1916 to 2015. We regress the *H* estimates of all stations on their spatial and regional characteristics (i.e. their location, elevation and Köppen-Geiger climate class) using a random forest algorithm. Furthermore, we apply the Mann-Kendall test under the LTP assumption (MKt-LTP) to all time series to assess the significance of observed trends of the mean annual precipitation.

To summarize the results, the LTP seems to depend mostly on the location of the stations, while the predictive value of the fitted regression model is good. Thus when investigating for LTP properties we recommend that the local characteristics should be considered. Additionally, the application of the MKt-LTP suggests that no significant monotonic trend can characterize the global precipitation. Dominant positive significant trends are observed mostly in main climate type D (snow), while in the other climate types the percentage of stations with positive significant trends was approximately equal to that of negative significant trends. Furthermore, 50% of all stations do not exhibit significant trends at all.

2. Introduction

- Long-term persistence (LTP) is an inherent property of geophysical processes in which wet years or dry years are clustered to respective long time periods (Koutsoyiannis 2002).
- The LTP can be modelled with the Hurst-Kolmogorov process (HKp) and characterizes the magnitude of LTP (Koutsoyiannis 2003).
- Estimation of *H* is important in engineering practice (Lins and Cohn 2011).
- Uncertainty increases substantially when LTP is present (Koutsoyiannis 2006; Koutsoyiannis and Montanari 2007; Tyralis and Koutsoyiannis 2014).
- Significant trends under the independence assumption can be considered non-significant under the LTP assumption (Hamed 2008).
- A few studies examine the LTP properties of global precipitation (Fatichi et al. 2012; Sun et al. 2014; Iliopoulou et al. 2016). Evidence of LTP presence in annual precipitation records is inconclusive (O'Connell et al. 2015).

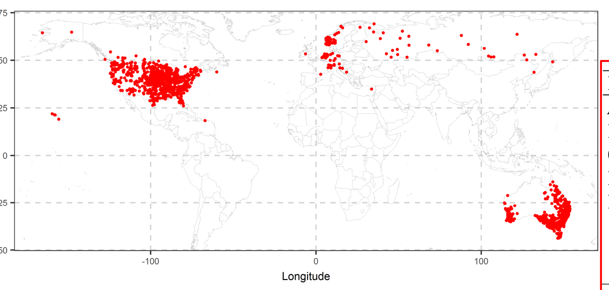
Here we:

- Estimate *H* of mean annual precipitation time series from instrumental measurements.
- Investigate possible relationships between *H* and station location features (latitude, longitude, elevation, climate type).
- Examine the importance of location features in predicting *H*.
- Predict *H* using location features as predictor variables.
- Estimate trends of mean annual precipitation and their significance.
- Perform an exploratory analysis on the trends coupled with station location features.

3. Data and methods

- Daily precipitation data from 1 535 stations (Menne et al. 2012a,b).
- Time-period of study: 1916-2015.
- Earth's surface coverage is limited to Australia, Europe, North America due to data availability.
- Daily time series imputation based on procedure described in Tyralis et al. (2017).
- Daily time series are transformed to mean annual time series.
- Estimation of *H* using the Maximum Likelihood Estimator (Tyralis and Koutsoyiannis 2011).
- Regression of *H* on predictor variables (longitude, latitude, xyz Cartesian coordinates, elevation, Köppen-Geiger climate class (Kottek et al. 2006)) using random forests (Breiman 2001), the cforest algorithm (Strobl et al. 2007) and linear regression.
- Estimation of trends and their significance using the Mann-Kendall test under the LTP assumption (MKt-LTP, Hamed 2008, Tegos et al. 2017).
- Application of methods using R packages (Breiman 2001 for the application of random forests, Strobl et al. 2007 for the application of the cforest algorithm, Kuhn 2008, Kuhn et al. 2016 for the optimization of the regression algorithms, Tyralis 2016 for the estimation of H and the application of the MKt-LTP).
- Further details and supplementary information can be found in Tyralis et al. (2017).

4. Stations location and Köppen-Geiger climate types



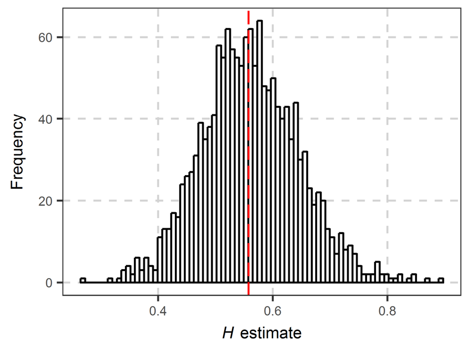
Köppen-Geiger climate types

Main climate		Precipitation		Temperature	
A	equatorial	W	desert	h	hot arid
B	arid	S	steppe	k	cold arid
C	warm temperate	f	fully humid	a	hot summer
D	snow	s	summer dry	b	warm summer
E	polar	w	winter dry	c	cool summer
		m	monsoonal	d	extremely continental
				F	polar frost
				T	polar tundra

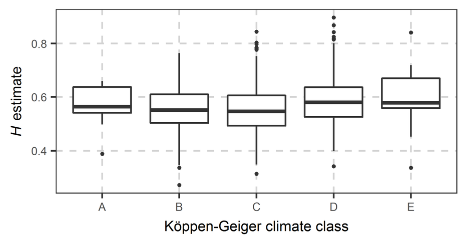
Climate class	Number of stations	Grouping 1	Grouping 2	Grouping 3
Am	5	A	A	Am
As	4	A	A	As
Aw	9	A	A	Aw
BSh	65	BS	B	steppe
BSk	223	BS	B	steppe
BWh	21	BW	B	BWh
BWk	6	BW	B	without dry season
Cfa	419	Cfa	C	without dry season
Cfb	206	Cfb	C	without dry season
Csa	41	Ca	C	summer dry
Csb	125	Csb	C	summer dry
Cwa	5	Ca	C	winter dry
Dfa	181	Dfa	D	without dry season
Dfb	148	Dfb	D	without dry season
Dfc	52	Dfc	D	without dry season
Dsb	8	Dsw	D	summer dry
Dsc	1	Dsw	D	summer dry
Dwb	3	Dsw	D	winter dry
Dwc	4	Dsw	D	winter dry
ET	9	E	E	polar tundra

- Regrouping of climate types, to increase the number of stations in each regrouped type.
- Grouping 1 includes types with low number of stations together, considering their main climate and precipitation type.
- Grouping 2 classifies stations according to their main climate.
- Grouping 3 is similar to that of Ragulina and Reitan (2017), who regrouped the stations according to precipitation conditions.

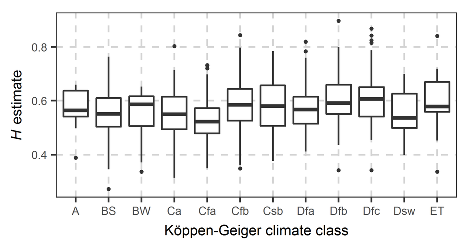
5. *H* estimate and climate type



- Median of *H* estimates equal to 0.56.
- A truncated normal distribution with support (0,1) seems to be a reasonable model for *H*.

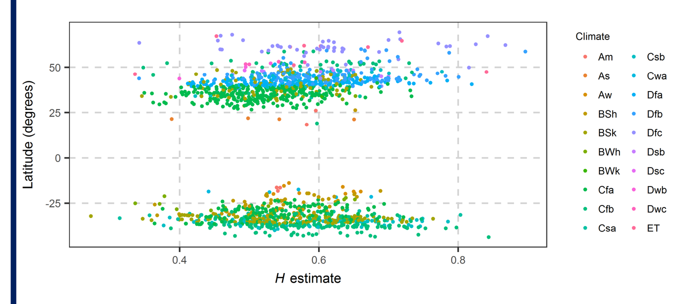


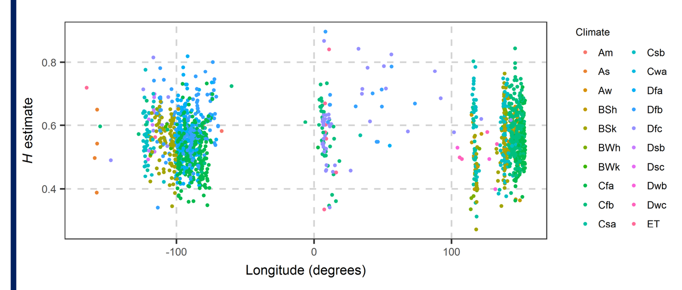
- *H* does not significantly vary with grouping 2.
- Its values are near to the median value 0.56.

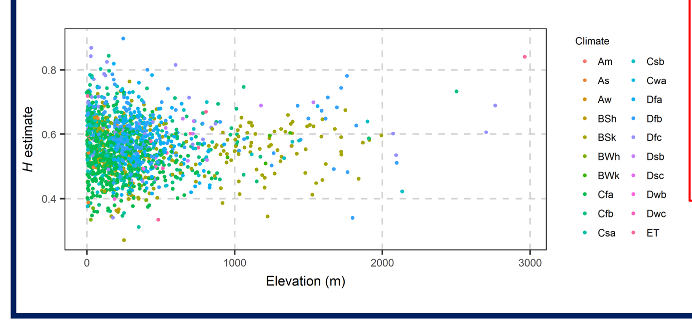


- Grouping 1 seems to be a better predictor, because of the higher variation of *H* between different climate classes.

6. *H* estimate and location characteristics







- Higher *H* values are observed for positive latitude, however no trend prevails.
- We do not observe any linear relationship between the two variables.
- We do not observe any clear linear relationship between *H* and the longitude.
- *H* is not linearly related to the elevation of each station.
- Absence of evidence for linear relationships.
- Therefore we must seek for possible relationships using non-linear algorithms.

7. Regression predictors and cross-validation

Combinations of predictor variables

Combination	Predictors
1	elevation
2	grouping 1
3	grouping 2
4	grouping 3
5	x, y
6	x, y, z
7	x, y, z, grouping 1
8	x, y, z, elevation
9	x, y, z, elevation, grouping 1
10	x, y, z, elevation, grouping 2
11	x, y, z, elevation, grouping 3
12	μ, σ
13	x, y, z, elevation, grouping 1, μ, σ
14	x, y, z, elevation, grouping 2, μ, σ
15	x, y, z, elevation, grouping 3, μ, σ
16	longitude
17	latitude
18	longitude, grouping 1
19	latitude, grouping 1
20	longitude, latitude
21	longitude, latitude, grouping 1
22	longitude, latitude, elevation
23	longitude, latitude, elevation, grouping 1
24	longitude, latitude, elevation, grouping 2
25	longitude, latitude, elevation, grouping 3
26	longitude, latitude, elevation, grouping 1, μ, σ
27	longitude, latitude, elevation, grouping 2, μ, σ
28	longitude, latitude, elevation, grouping 3, μ, σ

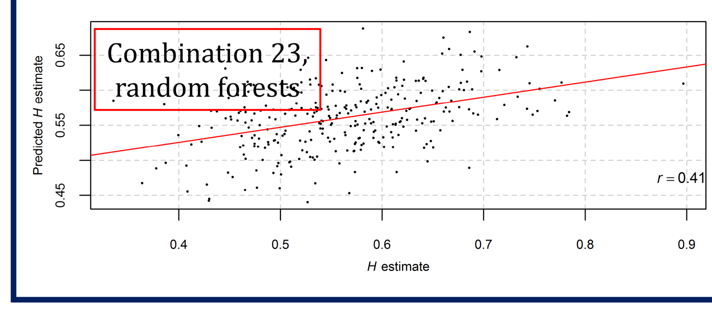
- Sample of 1 535 stations is split into 80% fitting set and 20% testing set.
- Performance of the linear models, random forests and the cforest algorithm are compared for each combination of predictors using the RMSE, MAE, MAPE and Pearson's *r* metrics.
- The metrics are calculated in the testing set.

Comb	Linear model				Random forests				cforest			
	RMSE	MAE	MAPE	*r*	RMSE	MAE	MAPE	*r*	RMSE	MAE	MAPE	*r*
1	0.086	0.068	0.124	0.01	0.096	0.075	0.137	0.02				
2	0.084	0.068	0.124	0.24	0.084	0.068	0.124	0.24	0.084	0.068	0.124	0.25
3	0.086	0.068	0.124	0.09	0.086	0.068	0.124	0.09				
4	0.088	0.069	0.126	−0.03	0.087	0.069	0.126	−0.03				
5	0.086	0.068	0.125	0.06	0.080	0.063	0.114	0.42				
6	0.086	0.068	0.123	0.11	0.079	0.061	0.110	0.44				
7	0.084	0.068	0.123	0.26	0.079	0.061	0.111	0.43				
8	0.086	0.068	0.123	0.11	0.077	0.059	0.107	0.47				
9	0.084	0.068	0.123	0.26	0.077	0.060	0.109	0.45				
10	0.085	0.067	0.122	0.17	0.077	0.059	0.108	0.46				
11	0.086	0.068	0.124	0.13	0.076	0.059	0.106	0.48				
12	0.086	0.068	0.124	0.07	0.091	0.071	0.130	0.09				
13	0.082	0.067	0.123	0.31	0.073	0.058	0.106	0.53				
14	0.085	0.067	0.122	0.21	0.073	0.058	0.105	0.52				
15	0.086	0.068	0.124	0.14	0.073	0.058	0.105	0.53				
16	0.086	0.068	0.124	0.05	0.098	0.077	0.141	0.14	0.084	0.067	0.121	0.28
17	0.086	0.069	0.124	0.00	0.097	0.078	0.142	0.09	0.087	0.070	0.127	0.19
18	0.084	0.068	0.125	0.24	0.091	0.072	0.132	0.25	0.081	0.064	0.117	0.37
19	0.084	0.068	0.125	0.24	0.091	0.071	0.130	0.19	0.082	0.064	0.116	0.34
20	0.086	0.068	0.123	0.12	0.080	0.062	0.113	0.42	0.078	0.061	0.110	0.43
21	0.084	0.068	0.124	0.25	0.082	0.063	0.116	0.38	0.080	0.062	0.114	0.39
22	0.086	0.067	0.123	0.13	0.077	0.060	0.109	0.45	0.078	0.061	0.110	0.43
23	0.084	0.068	0.124	0.25	0.079	0.062	0.113	0.41	0.080	0.062	0.114	0.39
24	0.085	0.067	0.122	0.17	0.078	0.061	0.111	0.43	0.080	0.062	0.114	0.38
25	0.086	0.067	0.123	0.14	0.078	0.061	0.110	0.43	0.079	0.061	0.112	0.40
26	0.082	0.067	0.122	0.32	0.074	0.059	0.108	0.51	0.077	0.061	0.111	0.46
27	0.085	0.067	0.122	0.20	0.075	0.059	0.108	0.50	0.077	0.060	0.109	0.46
28	0.086	0.068	0.123	0.15	0.075	0.059	0.108	0.50	0.076	0.060	0.109	0.47

8. 5-fold cross-validation, variable importance

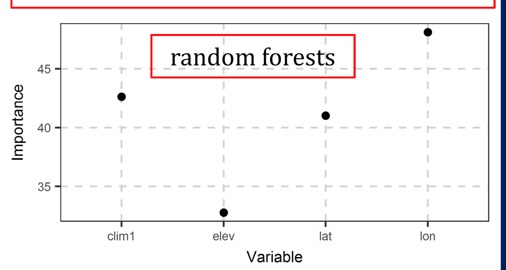
Results of the 5-fold cross-validation

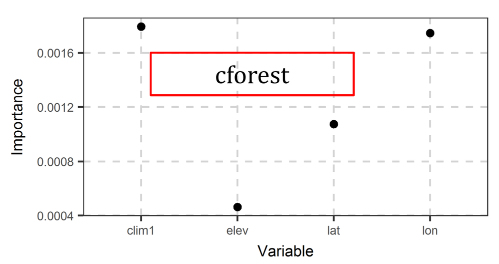
Method	Comb	Metric	Val 1	Val 2	Val 3	Val 4	Val 5	Mean
Random forests	2	RMSE	0.079	0.079	0.080	0.086	0.083	0.082
	2	*r*	0.35	0.28	0.28	0.24	0.30	0.29
	9	RMSE	0.074	0.071	0.074	0.081	0.076	0.075
	9	*r*	0.49	0.49	0.49	0.42	0.49	0.48
	16	RMSE	0.092	0.091	0.089	0.089	0.087	0.090
	16	*r*	0.22	0.19	0.22	0.32	0.31	0.25
	17	RMSE	0.095	0.094	0.090	0.102	0.096	0.095
	17	*r*	0.15	0.10	0.19	0.07	0.12	0.12
	18	RMSE	0.082	0.084	0.083	0.089	0.080	0.084
	18	*r*	0.40	0.33	0.35	0.33	0.44	0.37
	19	RMSE	0.092	0.089	0.087	0.096	0.088	0.090
	19	*r*	0.21	0.20	0.24	0.17	0.28	0.22
	20	RMSE	0.078	0.075	0.074	0.081	0.077	0.077
	20	*r*	0.45	0.44	0.49	0.43	0.48	0.46
	21	RMSE	0.078	0.073	0.075	0.082	0.076	0.077
	21	*r*	0.44	0.47	0.48	0.42	0.49	0.46
	23	RMSE	0.074	0.072	0.073	0.082	0.077	0.076
	23	*r*	0.49	0.48	0.50	0.41	0.48	0.47
Truncated normal		RMSE	0.085	0.082	0.084	0.089	0.088	0.086
		r	0.01	0.01	−0.01	−0.07	−0.01	−0.01



Variable importance

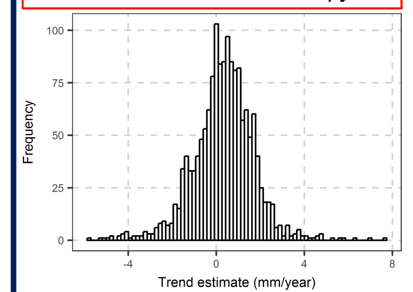
- Permutation importance measures were used.
- Permutation importance measures the mean decrease in classification accuracy after permuting each predictor variable in the trees of the trained model.
- Random forests and the cforest were applied to the dataset.



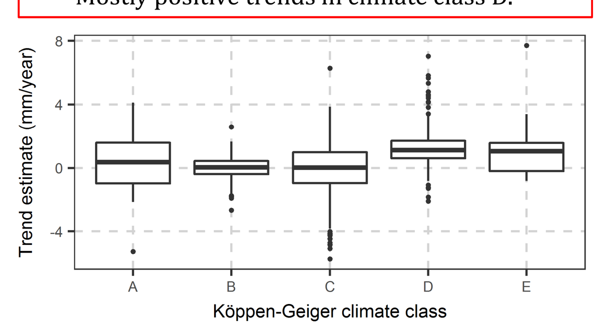


9. Significance of annual trend estimates

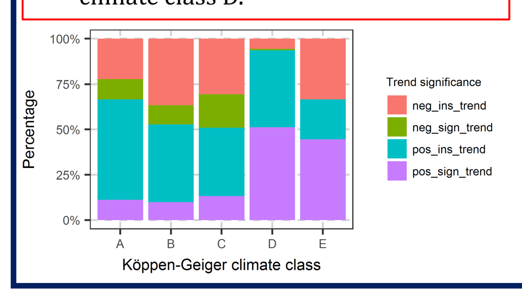
Median of trends: 0.36 mm/year



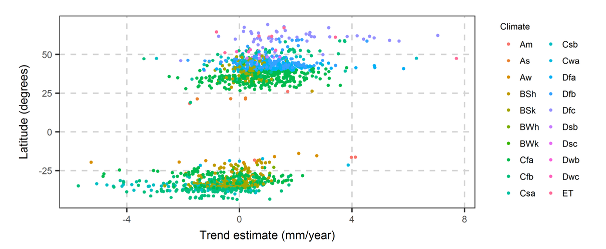
- Less uncertainty in the estimations of trends in climate classes B and D.
- Mostly positive trends in climate class D.



- Application of the MKt-LTP at a predefined significance level $a = 0.05$.
- pos, neg denote positive and negative respectively.
- sign, ins denote significant and insignificant respectively.
- Balance between positive and negative trends, excluding climate class D.
- Mostly positive significant trends in climate class D.



Mostly positive trends in the Northern hemisphere and negative trends in the Southern hemisphere.



10. Conclusions

- Median is $H = 0.56$ for the dataset of 1 535 mean annual precipitation time series for the time period 1916-2015.
- Result is consistent with Fatichi et al. (2012), Sun et al. (2014) and Iliopoulou et al. (2016).
- Location of stations is important in predicting *H*, followed by the climate type and elevation.
- However, the order of importance of the three former variables depends on the algorithm.
- The cforest algorithm estimates that the climate type is the most important, while due to its simultaneous handling of continuous and categorical variables can be considered more reliable than the random forests in estimating the variable importance.
- The combinations 6 and 20 of predictor variables, which include, respectively, the Cartesian coordinates and the geographic coordinates of the stations performs well in terms of the error metrics, but most importantly, their predictions had good correlation with the tested values.
- The inclusion of the climate type and the elevation (combinations 9, 23) improved further, albeit little, the performance of the random forests. However, this marginal improvement means that the information obtained from the geographic location of the station already includes the information of the climate type.

11. Conclusions

- The overall result is that the random forest algorithm can predict well the LTP of the mean annual precipitation, when the location characteristics are used as predictor variables while their performance is considerably better compared to the predictive ability of the simple distribution of H, particularly in terms of the correlation between the predicted and the estimated values.
- Therefore, the random forests can be used to predict *H* in locations without data or insufficient quantity of data and can serve as a substitute of spatial interpolation methods.
- Compared to spatial algorithms the random forests excel in combining information from distant locations through the common latitude, climate type and elevation variables, even if the spatial coverage is limited and non-uniform.
- Median value of the estimated trends is 0.36 mm/year.
- Dominant positive significant trends are observed mostly in main climate type D.
- In the other climate types the percentage of stations with positive significant trends is approximately equal to that of negative significant trends.
- In main climate types A-D 50% of the stations are characterized by insignificant trends.
- A limitation of our study is that the random forests algorithm can predict values only if given values of the predictor variables are within the range of the fitting set.
- Thus, the limited availability of data prohibits the generalization of the method to regions and Köppen-Geiger climate classes, which are not represented by the dataset.
- For more details see Tyralis et al. (2017).

References

Breiman, L., 2001. Random Forests. *Machine Learning*, 45 (1), 5–32. doi:10.1023/A:1010933404324
Fatichi, S., Ivanov, V.Y. and Caporali, E., 2012. Investigating Interannual Variability of Precipitation at the Global Scale: Is There a Connection with Seasonality?. *Journal of Climate*, 25, 5512–5523. doi:10.1175/JCLI-D-11-00356.1
Hamed, K.H., 2008. Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis. *Journal of Hydrology*, 349 (3–4), 350-363. doi:10.1016/j.jhydrol.2007.11.009
Iliopoulou, T., Papalexiou, S.M., Markonis, Y. and Koutsoyiannis, D., 2016. Revisiting long-range dependence in annual precipitation. *Journal of Hydrology*. doi:10.1016/j.jhydrol.2016.04.015
Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rybel, F., 2006. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15 (3), 259–263.doi:10.1127/0941-2948/2006/0130
Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal*, 47 (4), 573–595. doi:10.1080/02626660209492961
Koutsoyiannis, D., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences Journal*, 48 (1), 3–24. doi:10.1623/hysj.48.1.3.43481
Koutsoyiannis, D., 2006. Nonstationarity versus scaling in hydrology. *Journal of Hydrology*, 324 (1–4), 239–254. doi:10.1016/j.jhydrol.2005.09.022
Koutsoyiannis, D. and Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. *Water Resources Research*, 43, W05429. doi:10.1029/2006WR005592
Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28 (5), 1–26. doi:10.18637/jss.v028.i05
Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y. and Candan, C., 2016. caret: Classification and Regression Training. R package version 6.0-73. http://CRAN.R-project.org/package=caret.
Lins, H.F. and Cohn, T.A., 2011. Stationarity: Wanted Dead or Alive?. *Journal of the American Water Resources Association*, 47 (3), 475–480. doi:10.1111/j.1752-1688.2011.00542.x
Menne, M.J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R.S., Gleason, B.E. and Houston, T.G., 2012a. Global Historical Climatology Network - Daily (GHCN-Daily), Version 3.22. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ [access date:2016-09-02]
Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G., 2012b. An overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29, 897–910. doi:10.1175/JTECH-D-11-00103.1
O'Connell, P.E., Koutsoyiannis, D., Lins, H.F., Markonis, Y., Montanari, A. and Cohn, T.A., 2015. The scientific legacy of Harold Edwin Hurst (1880–1978). *Hydrological Sciences Journal*, 61 (9), 1571–1590. doi:10.1080/02626667.2015.1125998
Ragulina, G. and Reitan, T., 2017. Generalized Extreme Value's shape parameter and its nature for extreme precipitation using long-time series and Bayesian approach. *Hydrological Sciences Journal*. doi:10.1080/02626667.2016.1260134
Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8 (25). doi:10.1186/1471-2105-8-25
Sun, Q., Kong, D., Miao, C., Duan, Q., Yang, T., Ye, A., Di, Z. and Gong, W., 2014. Variations in global temperature and precipitation for the period of 1948 to 2010. *Environmental Monitoring and Assessment*, 186 (9), 5663–5679. doi:10.1007/s10661-014-3811-9
Tegos, A., Tyralis, H., Koutsoyiannis, D. and Hamed, K.H., 2017. An R function for the estimation of trend significance under the scaling hypothesis- application in PET parametric annual time series. *Open Water Journal*, 4 (1), 66–71.
Tyralis, H., 2016. HKprocess: Hurst-Kolmogorov Process. R package version 0.0-2. https://CRAN.R-project.org/package=HKprocess.
Tyralis, H., Dimitriadis, P., Koutsoyiannis, D., O'Connell, P.E., Tzouka, K. and Iliopoulou, T., 2017. On the long-range dependence properties of annual precipitation using a global network of instrumental measurements. Advances in Water Resources. In review.
Tyralis, H. and Koutsoyiannis, D., 2011. Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process. *Stochastic Environmental Research & Risk Assessment*, 25 (1), 21–33. doi:10.1007/s00477-010-0408-x
Tyralis, H. and Koutsoyiannis, D., 2014. A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. *Climate Dynamics*, 42 (11-12), 2867–2883. doi:10.1007/s00382-013-1804-y