Integration of modern statistical tools for the analysis of climate extremes into the web-GIS "CLIMATE"

A A Ryazanova[1,2], I G Okladnikov[1,2,3] and E P Gordov[1,2,3]

[1]Institute of Monitoring of Climatic and Ecological Systems SB RAS, Tomsk, Russia
[2]V.E. Zuev Institute of Atmospheric Optics SB RAS, Tomsk, Russia
[3]Institute of Computational Technologies, Tomsk Branch, Tomsk, Russia

E-mail: raa@scert.ru, igor.okladnikov@gmail.com, gordov@scert.ru

Abstract. The frequency of occurrence and magnitude of precipitation and temperature extreme events show positive trends in several geographical regions. These events must be analyzed and studied in order to better understand their impact on the environment, predict their occurrences, and mitigate their effects. For this purpose, we augmented web-GIS called "CLIMATE" to include a dedicated statistical package developed in the R language. The web-GIS "CLIMATE" is a software platform for cloud storage processing and visualization of distributed archives of spatial datasets. It is based on a combined use of web and GIS technologies with reliable procedures for searching, extracting, processing, and visualizing the spatial data archives. The system provides a set of thematic online tools for the complex analysis of current and future climate changes and their effects on the environment. The package includes new powerful methods of time-dependent statistics of extremes, quantile regression and copula approach for the detailed analysis of various climate extreme events. Specifically, the very promising copula approach allows obtaining the structural connections between the extremes and the various environmental characteristics. The new statistical methods integrated into the web-GIS "CLIMATE" can significantly facilitate and accelerate the complex analysis of climate extremes using only a desktop PC connected to the Internet.

1. Introduction

Climate extremes may have significant effects on natural and artificial terrestrial ecosystems. However, their impact is complex, depends on numerous factors, and has not yet been fully understood so far. A detailed knowledge of the statistics of climate extremes and spatio-temporal patterns of the essential climate variables causing such events is important for estimating the potential damages and, hence, sustainable planning and managing the ecosystem-related resources. It is well-known that since 1970 both the average global temperature and total precipitation amount are increasing [1]. Also, the frequency of occurrence and magnitude of the events directly related to the extreme values of these characteristics (showers, hurricanes, droughts, and heat waves) have clear growth trends [1]. Detailed analyses and understanding of their impact on the environment are needed for the forecasting of such extremes and mitigation of their effects. Such studies must employ modern statistical methods for the analysis of extreme values [2-6]. To characterize climate extremes, data from various sources (meteorological stations, reanalysis products, remote sensing data) covering various spatial and temporal ranges and scales are used. To simplify the use of extreme value statistics

tools and handle large volumes of such data, we augmented thematic web-GIS called "CLIMATE" developed for cloud storage processing of geophysical datasets [7, 8]. The integration of these tools into the web-GIS "CLIMATE" and some results are presented in this paper.

2. Methods and approaches

The system "CLIMATE", based on web- and GIS-techniques, is part of a hardware and software complex for cloud storage analysis of climate data. This complex includes various climatological and meteorological datasets, as well as dedicated interactive tools for their search, sampling, processing, and visualization. Utilization of this system significantly simplifies and accelerates the handling of huge volumes of spatial climate datasets, allowing users without specific IT knowledge to perform remotely processing and analysis of data using only a modern desktop PC connected to the Internet. The web-GIS allows calculating the basic and complex statistical properties of time series of meteorological and climatological characteristics [7-9], including average, minimum, maximum, dispersion, and a set of basic WMO climate change and extreme indices (http://etccdi.pacificclimate.org/) and hydrothermal coefficients.

Powerful, more sophisticated alternatives accounting for trends in extremes (like time-dependent extreme value statistics [10, 11] or quantile regression [12, 13]) are available but have not yet been systematically exploited. Since complex climate extremes (e.g., of temperature and precipitation) are of particular relevance for ecosystems, the well-known copula approach for describing multivariate probability distributions [14, 15] provides a promising technique, which has also not yet been used in this context. These advanced methods for climate data analysis significantly extend the current functionality of the system "CLIMATE" and seem very promising for climate research applications.

2.1. Time-dependent statistics of extremes

A statistical description of extreme events, such as extreme precipitation or temperature, can be done using the concepts of extreme value statistics (EVS). Unlike in the early stages of EVS [16], we can now take advantage of well-developed approaches to describe univariate extremes with non-stationary, i.e. time-dependent, probability distribution functions (PDFs). EVS provides two major routes to a probabilistic description of extremes: *the block maxima approach* to describe the probability distribution of the intensity of maxima of blocks (e.g., monthly or annual maximum of a meteorological variable) with a generalized extreme value distribution (GEV), and the *peak-over-threshold approach*, modelling excesses over a threshold with a generalized Pareto distribution. For example, you can use the block maxima approach and a time-dependent model to describe the seasonal variability of precipitation maxima [10]. A well-written introduction to EVS covering both the routes and non-stationary concepts is given by Coles [17].

A software implementation of EVS in the R language is represented by the class "fevd" from the software package "extRemes" [18, 19]. The input arguments for an instance of this class depend on the selected approach. If the *block maxima approach* is chosen, the maximum values of the analyzed meteorological characteristics for a selected time range (e.g., month or year) are given. If the *peak-over-threshold approach* is chosen, daily or maximum values for selected time range values are given, but an additional input argument "threshold" must be given also. Input argument "type" allows us to select PDF (by default GEV is used). The input argument "method" allows us to select estimation parameters (by default "Maximum Likelihood Estimation" – MLE is used). The arguments "location.fun", "scale.fun", "shape.fun", and "threshold.fun" allow us to set time dependencies of PDF parameters. The result of a calculation is a set of estimated parameters of a given PDF (in the case of GEV it is a vector with location, scale, and shape parameters). A more detailed description can be found in [18].

The PDF location, scale, and shape parameters by themselves provide little information on the probability of occurrence of extreme events. A physically more meaningful and also more relevant quantity for risk assessment is the probability of an observed variable exceeding a certain level. These levels can be calculated from a parameterized GEV and are frequently expressed as return levels r_T for

a certain return period T. r_T is defined as the level which is exceeded on average every T blocks, i.e., with probability $1/T$ [10]. r_T is calculated using «return.levels» class with a list of class «fevd» instances as the major argument. There are ad-on arguments: «return.period», a numeric vector of desired return periods and «qcov», a numeric matrix with rows the same length as q and columns equal to the number of parameters (+ 1 for the threshold, if a POT model). This gives any covariate values for a nonstationary model and is not used in the stationary case. «qcov» is used to obtain effective return levels.

As a result, an object of class "return.level" is returned, which is either a numeric vector (stationary models) of length equal to the «return.period» argument giving the return levels or a matrix of a dimension equal to either n by np or q by np, where n is the length of the data used to fit the model and np is the number of return periods, and q is the number of rows of «qcov», if supplied. The returned value also includes useful attributes describing how the return levels are to be estimated. A detailed description of the «return.levels» class can be found in [18].

2.2. Quantile regression

The analysis of trends in meteorological observations is one of the most common activities in climate change studies. Quantile regression provides a well-defined statistical framework for estimating the rate of change not only in the mean as in ordinary regression, but in all parts of the data distribution. It was first introduced in econometrics by [5]. Now quantile regression is used in various geoscience contexts [12, 20-24].

Given a random variable Y with a cumulative continuous distribution function $F_Y(y)$, the quantile is defined as the value $Q_Y(\tau)$ such that $P[Y \leq Q_Y(\tau)] = \tau, 0 \leq \tau \leq 1$. The quantile function $Q_Y(\tau)$ is defined from the cumulative distribution function $F_Y(y)$ as $Q_Y(\tau) = F_Y^{-1}(\tau)$. Then, considering the conditional distribution of Y given $X = x$, the conditional quantile function $Q_{Y|X}(\tau; x)$ verifies $P\left[Y \leq Q_{Y|X}(\tau; x)|X = x\right] = \tau$. Whereas ordinary regression is based on the conditional mean function $E[Y|X = x]$ and minimization of the respective residuals, quantile regression is based on the conditional quantile function and minimization of the sum of asymmetrically weighted absolute residuals $\sum_{i=1} \rho(\tau)|y_i - Q_{Y|X}(\tau; x = x_i)|$, where ρ is the tilted absolute value function.

Another advantage of quantile regression is that it is non-parametric, i.e. the distribution of the investigated variable is not assumed to belong to any parametric population. Moreover, quantile regression is a robust method and weakly affected by the outliers and peak (extreme) values. Further details can be found in [25, 26].

A software implementation of quantile regression in the R language is represented by the class "qr" from the software package "quantreg" [27]. To calculate the quantile trends, the input argument "formula" allows us to set a formula describing a quantile model (e.g., in the simplest case we can set a linear time dependence of the meteorological values; in a more complicated case the model can be extended with seasonal components). The quantile values of interest are set between 0 and 1 in the input vector "tau". The results are returned as a vector of quantile regression model estimated parameters. A detailed description of "qr" class can be found in [27].

2.3. Copula approach

Due to a possible nonlinear dependence between the variables, the probability distributions of multivariate random variables are generally more complex than their univariate counterparts. One of the promising approaches to this problem is the use of copulas. It has become popular in recent years, especially in econometrics, finance, risk management, or insurance. The copula approach is a simple and straightforward method to find parametric descriptions of multivariate non-normally distributed random variables. Although it is a fast growing field of statistics, it is still little known in climate research, where often non-normally distributed random variables like precipitation, wind speed, cloud cover, humidity, etc. are involved. An advantage of this approach is the ability to catch various covariance structures while keeping proper parametric descriptions of the margins.

Unlike the univariate extreme value theory, here the possible limiting distribution functions of extreme values cannot be captured by a finite-dimensional parametric family of functions. The study of multivariate extremes splits into descriptions of the marginal distribution and the dependence structure. In order to characterize the limit behavior of multivariate extremes, it has been shown that weak convergence of the multivariate distribution is equivalent to weak convergence of the marginals as well as the copula function, provided that the marginals are continuous [3]. This is the reason why the copula approach is very popular in modeling multivariate extremes [28].

Possible applications of the copula approach include different typical selections of pairs of variables, like one meteorological characteristic at different locations or different meteorological characteristics at the same location, as far as they can be expressed in the copula framework [14]. There is no general procedure for selecting the copula class.

A software realization of the copula approach technique is represented in the R language as the dedicated stand-alone software package "copula" [29, 30].

3. Results

The above-described software tools for statistical analysis of time series of spatial climatological datasets have been integrated into the web-GIS "CLIMATE", thus allowing analysis of current and future extreme climate events. The algorithms from the packages "extRemes" [18, 19], "quantreg" [27], and "copula" [29, 30] implementing time-dependent statistics of extremes, quantile regression and the copula approach were used as the basis for these tools. These packages are written in the R language (https://www.r-project.org/, [31]) and provide a flexible API for the integration into a third-party software. The computing core of the web-GIS "CLIMATE" is written in the GNU Data Language (GDL, http://gnudatalanguage.sourceforge.net/), and currently there is lack of universal wrappers providing the execution of R procedures inside a GDL program. To overcome this obstacle, the fact that GDL has built-in Python (https://www.python.org/) interface was used and special software adapters connecting GDL and R were developed in Python. These adapters utilize the dedicated Python-package RPy2 [32] and provide a two-way data and instruction transfer between the web-GIS computing core and procedures in R packages (Figure 1). As a result, the computing core acquired the ability to call new procedures implemented in the R language to perform complex statistical analysis of climate data.

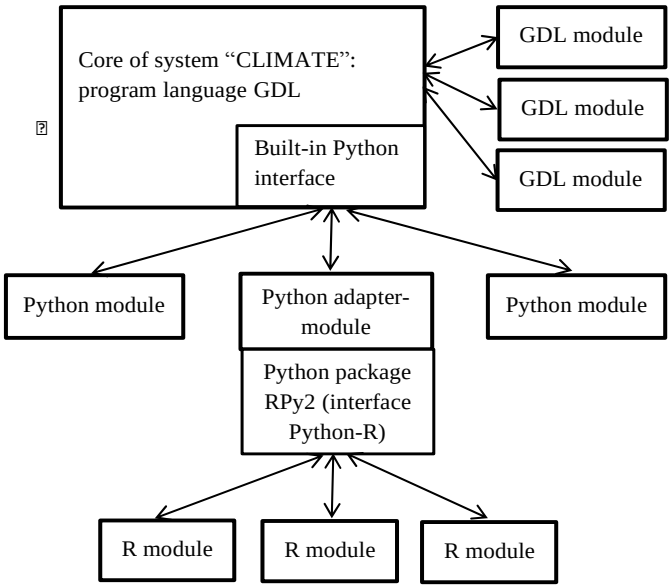

Figure 1. Diagram of interactions between computing core and R-modules.

To provide an intuitively clear and user friendly graphical interface of the online tools as well as to provide effective processing and data handling, the corresponding metadata describing new processing routines were inserted during the integration process into a web-GIS dedicated metadata database. This database [33] contains spatial and temporal characteristics of the available spatial datasets, their physical locations, and the parameters of software procedures for data analysis and visualization. Thanks to this database it has become possible to extend the processing abilities of the web-GIS and to edit the GUI content without any modification of the source codes of the web-GIS components. All new abilities have become available to the users immediately after the metadata insertion, thus providing a rapid and sustainable development of the web-GIS "CLIMATE".

To illustrate the extended web-GIS "CLIMATE"'s functionality, the software module utilizing the package "extRemes" was used to calculate 100-year return levels of maximum precipitation conditioned on the July month on ECMWF ERA Interim [34] (Figure 2a) and on APHRODITE JMA [35] data (Figure 2b) for Southern Siberia (52.5-60° N, 75-95° E).

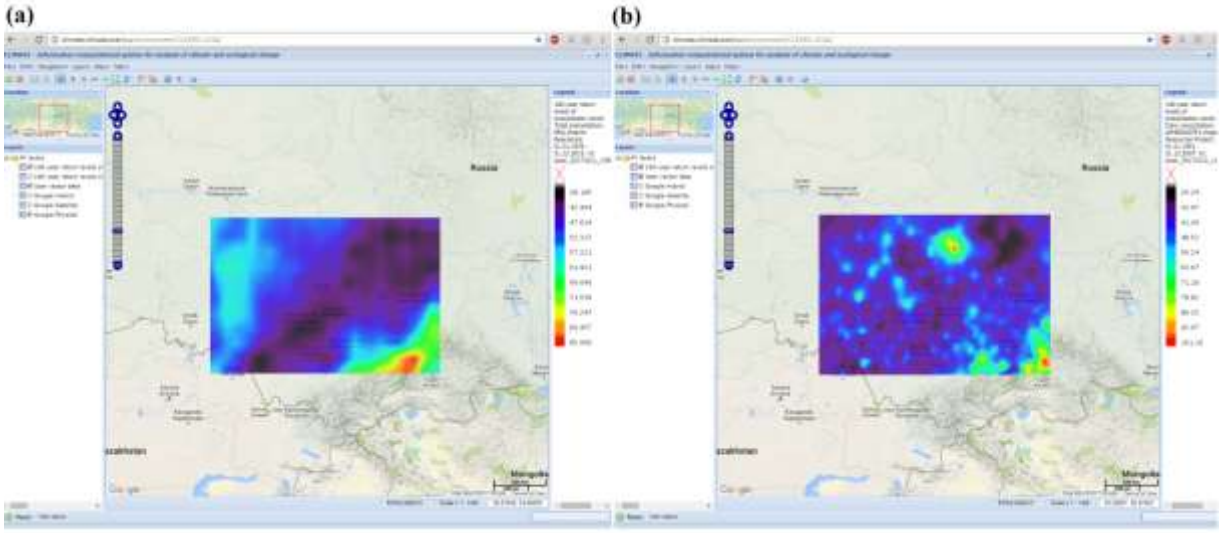

Figure 2. Screen shots of obtained cartographic layers for 100-year return levels of maximum precipitation conditioned on the July month of their occurrence for Southern Siberia: a) on ECMWF ERA Interim data, 0.75x0.75 horizontal grid, 1979-2012, b) on APHRODITE JMA data, 0.25x0.25 horizontal grid, 1951-2007.

The results obtained are in good agreement with the information available in the literature [10].

4. Conclusions

The integration of modern statistical packages for extreme analysis developed in the R language into the web-GIS "CLIMATE" significantly extended its functional abilities. Thanks to it, new powerful climate data analysis techniques, such as time-dependent statistics of extremes, quantile regression and copula approach, have become available to the users of the web-GIS "CLIMATE". These techniques allow performing detailed analysis of various extreme climate events, to estimate the magnitude of their impact, and to discover the structural interconnections between the extremes and the environment characteristics.

A practical application of the earlier developed metadata database showed that its usage improves the scalability and flexibility of computations and facilitates new processing procedures to the web-GIS "CLIMATE".

The results obtained show that the new interactive tools for the extreme climate events analysis can be useful for decision makers and specialists working in affiliated sciences, to make socio-economic

impact assessment, ecological impact assessment, adaptation strategies, science policy administration, and other climate related activities. These tools can provide reliable climate related characteristics required for studies of economic, political, and social consequences of global climate change at the regional level.

Acknowledgements
This work was supported by SB RAS under Basic Research Program no. IX.138.2.1.

References

[1] Stocker T F, Qin D, Plattner G-K, Tignor M, Allen S K, Boschung J, Nauels A, Xia Y, Bex V and Midgley P M 2013 *Climate Change 2013*: *The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge: Cambridge University Press and NY, USA) p 1535
[2] Friederichs P 2007 *An Introduction to Extreme Value Theory* (Bonn: Meteorological Institute University of Bonn, COPS, Summer School)
[3] Beirlant J, Goegebeur Y, Segers H and Teugels J 2004 *Statistics of Extremes: Theory and Applications* (Wiley Series in Probability and Statistics)
[4] Embrechts P, Kluppelberger C and Mikosch T 1997 *Modelling Extremal Events for Insurance and Fincance* (Berlin: Springer)
[5] Koenker R and Basset 1978 Regression quantiles *Econometrica* **46** 33–50
[6] Nelsen R 2006 *An Introduction to Copulas* (New York: Springer, 2nd edn)
[7] Gordov E P, Okladnikov I G, Titov A G, Bogomolov V Yu, Shulgina T M and Genina E Yu 2012 Geo-information system for investigation of regional climatic changes and first results obtained *Atmos. Ocean. Opt.* **25** (2) 137–43
[8] Gordov E P, Shiklomanov A, Okladnikov I G, Prusevich A and Titov A G 2016 Development of Distributed Research Center for analysis of regional climatic and environmental changes *IOP Conf. Series: Earth and Environmental Science* **48** doi:10.1088/1755-1315/48/1/012033
[9] Riazanova A A, Voropay N N, Okladnikov I G, Gordov E P 2016 Development of computational module of regional aridity for web-GIS "Climate" *IOP Conf. Series: Earth and Environmental Science* **48** doi:10.1088/1755-1315/48/1/012032
[10] Rust H, Maraun D and Osborn T J 2009 Modelling seasonality in extreme precipitation *Europ Phys J ST* **174** 99–111
[11] Katz R W, Parlange M B and Naveau P 2002 Statistics of extremes in hydrology *Advances in Water Resources* **25** 1287–1304
[12] Barbosa S M, Scotto M G and Alonso A M 2011 Summarising changes in air temperature over Central Europe byquantile regression and clustering *Nat Hazards Earth Syst Sci* **11** 27–3233
[13] Sterin A M and Timofeev A A 2014 Specific features of estimates of surface air temperature trends in the Russian Federation obtained using quantile regression *Proceedings of RIHMI-WDC* **178**
[14] Schölzel C and Friederichs P 2008 Multivariate non-normally distributed random variables in climate research – introduction to the copula approach *Nonlin Proc Geophys* **15** 761–72
[15] Salvadori G and De Michele C 2004 Frequency analysis via copulas: Theoretical aspects and applications to hydrological events *Water resources research* **40**
[16] Fisher R A and Tippett L H C 1928 Limiting Forms of the Frequency Distribution of the Largest or Smallest Members of a Sample *Proc. Cambridge Phil. Soc*. **24** 180–90
[17] Coles S G 2001 *An Introduction to Statistical Modelling of Extreme Values* (London: Springer)
[18] Gilleland E 2016 Package "extRemes" The Comprehensive R Archive Network (CRAN) https://cran.r-project.org/web/packages/extRemes/extRemes.pdf
[19] Gilleland E and Katz R W 2016 extRemes 2.0: An Extreme Value Analysis Package in R *Journal of Statistical Software* **72** (8) doi: 10.18637/jss.v072.i08

[20] Koenker R and Schorfheide F 1994 Quantile spline models for global temperature change *Climatic Change* **28** (4) 395–404
[21] Cade B and Noon B 2003 A Gentle introduction to quantile regression for ecologists *Front. Ecol. Environ.* **1** 412–20
[22] Baur D, Saisana M and Schulze N 2004 Modelling the effects of meteorological variables on ozone concentration – a quantile regression approach *Atmos. Environ.* **38** 4689–99
[23] Elsner J B, Kossin J P and Jagger T H 2008 The increasing intensity of the strongest tropical cyclones *Nature* **455** 92–5
[24] Barbosa S M 2008 Quantile trends in Baltic sea-level *Geophys. Res. Lett.* **35** L22704 doi:10.1029/2008GL035182
[25] Koenker R and Hallock K 2001 Quantile Regression *J. Economic Perspect.* **15** 143–56
[26] Koenker R 2005 *Quantile regression* (New York: Cambridge University Press)
[27] Koenker R, Portnoy S, Tian P, Zeileis A, Grosjean P and Ripley B D 2017 Package "quantreg" The Comprehensive R Archive Network (CRAN) https://cran.r-project.org/web/packages/quantreg/quantreg.pdf
[28] Renard B and Lang M 2007 Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology *Adv. Water Resour.* **30** 897–912
[29] Hofert M, Kojadinovic I, Maechler M, Yan Y 2017 Package "Copula" The Comprehensive R Archive Network (CRAN) ftp://cran.r-project.org/pub/R/web/packages/copula/copula.pdf
[30] Yan J 2007 Enjoy the Joy of Copulas: With a Package copula *Journal of Statistical Software* **21** 4
[31] Gilleland E 2011 *Using R to Analyze Extremes* National Center for Atmospheric Research (Boulder, Colorado, U.S.A)
[32] Documentation for RPy2 http://rpy2.readthedocs.io/en/version_2.8.x/index.html
[33] Okladnikov I G, Gordov E P and Titov A G 2016 Development of climate data storage and processing model *IOP Conf. Series: Earth and Environmental Science* **48** doi:10.1088/1755-1315/48/1/012030
[34] Dee D P *et al.* 2011 The ERA-Interim reanalysis: configuration and performance of the data assimilation system Quarterly *Journal of the Royal Meteorological Society* **137** (656) Part A 553–97
[35] APHRODITE JMA, http://www.chikyu.ac.jp/precip/data/APHRO_V1003R1_readme.txt