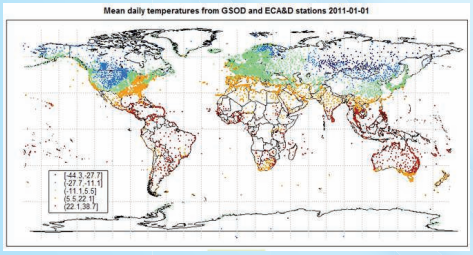
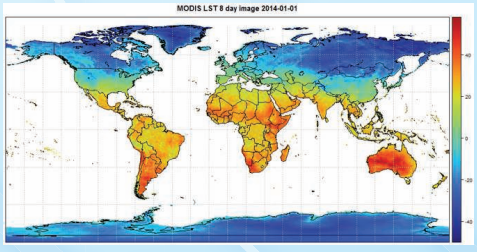
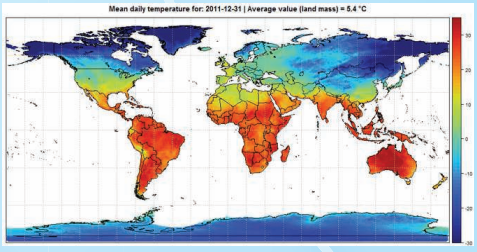
Proceedings of DailyMeteo.org/2014 Conference

Belgrade, Serbia 26-27 June 2014.







dailymeteo.org/2014

DailyMeteo.org/2014
Abstracts, extended abstracts and full papers
of the DailyMeteo.org/2014 Conference
Belgrade, Serbia, 26-27 June 2014

Edited by
Branislav Bajat
Milan Kilibarda

For the publisher
Dusan Najdanović

Design and prepress
Dosije studio doo Belgrade

Printed by
Dosije studio doo, Belgrade
www.dosije.rs

Book Circulation
50 copies

ISBN 978-86-7518-169-9

Contents

Statistical Modelling of Space-Time Variability

[extended abstact]

Gerard B.M. Heuvelink
Soil Geography and Landscape group
Wageningen University
Wageningen, The Netherlands
gerard.heuvelink@wur.nl

***Abstract*—**Many environmental variables, such as precipitation, temperature and radiation, vary both in space and time. The space-time variability of these variables is governed by physical laws, which are often characterised by partial differential equations. However, these equations can be very complex and their parameters and initial and boundary conditions are often very poorly known. This makes it extremely difficult to obtain practically useful solutions. In such case, statistical modelling offers an alternative. Statistical models are no replacement for mechanistic models because they give less insight into governing processes and cannot easily be extrapolated, but they are easier to implement, calibrate and run. Provided that the observation density is sufficiently large, they often yield sufficiently accurate predictions of the space-time variable at unobserved points. Geostatistics offers a rich methodology for statistical modelling and prediction of spatially distributed variables. The basic approach is to treat the variable of interest as a sum of a deterministic trend and a zero-mean stochastic residual. The trend is often taken as a linear combination of explanatory variables that must be known spatially exhaustively, while the stochastic residual is usually assumed to be normally distributed and stationary. It will typically also be spatially correlated, as characterised by a semivariogram. With this model, predictions at unobserved locations can be made using kriging, which also quantifies the prediction error variance. Extension of the geostatistical model to the space-time domain can be done in various ways. One is to consider the spatial variable at multiple time points, deriving a geostatistical model at each of these time points and characterising the correlation between variables at different time points through a cokriging approach. However, the disadvantage of this approach is that it only addresses the variable at the selected times and not in between, and that modelling is cumbersome when the number of time points is moderate or large. A more attractive alternative is to include time as a third dimension and model space-time variability by means of a spatio-temporal trend and a space-time stochastic residual. Once this model has been defined and calibrated it can be used to predict and simulate at any point in space and time, hence producing a 'movie' of the spatial distribution over time. In recent years many advances have been made in developing theoretically sound space-time statistical models. The difficulty is in the space-time stochastic residual, because the associated covariance model must include zonal and geometric anisotropies. Popular representations of the space-time covariance structure are the sum-product model and the sum-metric model. Fitting of these models to real-world data sets and using these models for space-time prediction and simulation has greatly improved in recent years due to advances in the *spacetime* and *gstat* packages in R. The main problem with defining valid space-time covariance structures is that these must be semi-positive definite, which is difficult to prove. If, however, the space-time covariance structure is derived from an explicit model of the space-time variable, such as through a space-time auto-regressive moving average (ARMA) model or a so-called state-space model, then the semi-positive definiteness is guaranteed by construction. In such case, spatio-temporal prediction may be done using the Kalman filter and Kalman smoother, which, as does kriging, calculate the conditional probability distribution of a target variable given conditioning data. The attractive property of space-time ARMA and state-space models is also that these bridge the gap with mechanistic modelling of space-time variability. This is because the ARMA and discrete state-space approach may be interpreted as discrete approximations of stochastic partial differential equations. There is yet a lot to be discovered in this research area, and if software development can go hand in hand with theoretical developments we may see major steps forward in the years to come. All statistical approaches described above are explained in this lecture and illustrated with real-world applications.

Spatial and Temporal Support of Meteorological Observations and Predictions

[abstract]

Edzer Pebesma
Institute for Geoinformatics
University of Münster

***Abstract*—**Support refers to the physical size of the area, volume, and/or temporal duration of a measured or predicted data value. Support of measurements is often related to the physical constraints: we cannot directly observe the temperature of a square kilometre, not even of an area of 100 m2; rainfall measurements also usually refer to devices with a catchment area of less than 1 m^2.

By choosing measurement sites carefully, we hope, by the idea of representativity, that measured values carry more information about their surroundings than when they were not chosen with the same care. Representativity could reflect the notion that we would like to be able to measure average values over larger areas, as local extreme conditions are typically avoided.

Nevertheless, measured value and local or regional averages will differ. Geostatistical theory allows for predicting linearly aggregated (mean) values by regularizing (averaging) semivariances, and by predicting nonlinearly aggregated values by simulation. The type of aggregation (function), the aggregation predicate (target support), and the variability of the predictant all play a role here.

Aggregation is the process of deriving a single number from a collection of numbers. The aggregation function may be simple such as in the case of mean or max, it may also be complex, e.g. computing catchment discharge from spatially distributed precipitation values. The aggregation predicate is the spatial area and/or temporal period over which aggregation takes place. Aggregation may be useful to (i) match data that is collected at a coarser support (ii) increase accuracy of predictions, and (iii) smooth out local, or short-term variability.

When we want to aggregate over a continuous area but do not have exhaustive (continuous) measurement data available for this area, a model for the observation data is required to fill the area with missing data in with predictions. Typical models are stationary covariance models, as used in geostatistics. When, in these models, we assume the mean function to depend on external variables with a different support (e.g. derived from satellite imagery, or elevation data), we introduce a bias that depends on the difference of the external variable at the support we have it and that, at the support that would match that of the primary observation data. We will discuss where this bias comes from, and how it may be dealt with.

Spatiotemporal Trends in Climate within Redwood Range

[abstract]

Miguel Fernandez
Berkeley
1612 Lincoln St. Berkeley CA 94703

***Abstract*— Redwood (Sequoia sempervirens), once a widely distributed species and now limited to a narrow 50km belt along the west coast of North America, provides many ecosystem services, including reservoirs of unique biodiversity and high rates of carbon sequestration. Although our knowledge of the spatial distribution and ecophysiological traits of the species are relatively advanced, the spatiotemporal trends in climate within the species range are still unknown. Part of the reason is the complexity of the climate system, where fine scale sharp coastal energy/moisture gradients are associated with wind-driven upwelling of cold water in the coast of California. Coastal upwelling can limit increases in coastal temperatures, decoupling the system from synoptic conditions. Taking advantage of a very fine resolution time series (PRISM), for the years 1950 to 2012, we evaluated the nature of historic climate trends. We applied standard non-parametric statistics (e.g., Mann-Kendall and Theil-Sen) to evaluate the magnitude and the significance of spatio-temporal climate trends on a cell-by-cell basis. Our results characterize the environmental heterogeneity in climatic trends within the redwood range over the past 60 years, identifying areas of recent significant changes as well as areas of relative climate stability that can be used to inform natural resource management and planning in the face of global change.**

Application of Space-Time Kriging for Monthly Precipitation Values of Lake Van Basin in Turkey

[extended abstact]

Pınar Aslantaş
Yuzuncu Yil University, Faculty of Agriculture, Landscape Architecture Department
Van, Turkey
pinarbostan@yyu.edu.tr

Abstract*— Precipitation is an important climatic variable that varies both in space and time. Like other climatic, meteorological, hydrologic and environmental variables, precipitation is measured from specific locations. Predictions at the locations that have no measurement values are obtained with interpolation techniques. Space-time interpolation techniques which use variables that vary both in space and time have received increasing attention. In this study space-time kriging is performed by combining spatial and temporal information of precipitation. The aim of the study is to apply space-time Universal kriging (ST-UK) method to monthly precipitation values measured from meteorological stations at the Lake Van Basin and predict precipitation at each spatial and temporal location. Lake Van is the largest lake of Turkey and located at the far-east part of the country. Lake Van is one of the largest closed drainage basins of the world. The Lake Van basin includes the lake and neighboring districts. The area of Lake Van Basin is used as study area in the study. The area of the basin is about 16.000 km^2 and area of lake is about 3800 km^2. Space-time information of precipitation is obtained from ten meteorological stations located over the basin and measurements are recorded for 1981-2010 years. Elevation is used as secondary information that was obtained by resampling the 3 arc second SRTM (the Shuttle Radar Topography Mission) (approximately 90 m spatial resolution) to 1 km spatial resolution using the Nearest Neighbor algorithm. Monthly precipitation values are analyzed and predicted over 1*1 km spatial resolution grid. One-fold cross-validation is used to assess accuracy performance of space-time kriging technique. In this way, R-square and RMSE (Root Mean Square Error) are calculated and evaluated for each prediction maps.

Keywords*— Precipitation, meteorological station, Space-time Universal Kriging, Lake Van Basin

I. INTRODUCTION

Spatial kriging methods have been used for many years to predict variables at unmeasured locations in many disciplines. The first geostatistics and spatial kriging applications started in mining and geology. The variables used in these sciences can often be assumed constant in time. After understanding the usefulness and reliability of kriging in these disciplines, it was also introduced to many other disciplines within the earth and environmental sciences, such as meteorology, climatology, agronomy, soil science, hydrology, etc. Generally variables in these sciences vary both in time and space. Therefore the requirement of kriging methods for space-time interpolation is raised [1]. If the data have been measured in different time and space locations, then more data may be used for prediction, and this allows obtaining more accurate predictions, helps to parameter estimation and helps to define spatial and/or temporal auto-correlation in measurements [2]. In case of space-time kriging, to predict the value of the variable of interest at a specific location and time, past and future measurements are used to predict on the specified time. This may add more complexity to the kriging procedure but may help to gain more accurate results.

In this study space-time Universal kriging (ST-UK) method is applied to monthly precipitation values measured from 10 meteorological stations from 1981 to 2010 over the Lake Van basin of Turkey. The aim is to discuss applicability of space-time kriging methods on monthly precipitation values by using limited number of meteorological station.

II. STUDY AREA AND DATA

A. Study Area

The study area is Lake Van Basin that is located at the far east part of Turkey (Figure 1). The area of basin is about 16.000 km^2. Lake Van basin has a high topography. The high mountains are located at the northern and southern parts of the basin. The mean elevation of basin is about 2200-2400 m., minimum elevation is about 1500 m. and maximum elevation is approximately 4000 m (Figure 2).

Lake Van which is the biggest lake of country is located at the basin (Figure 2). The Lake is a depression state in the middle of high mountains. Lake has a surface of 3574 km^2, length of shoreline is 505 km, and a volume of 607 km^3. The lake stands at 1650 m. above sea level. The Lake is a closed lake without any significant outflow. With a maximum depth of 451 m and a volume of 607 km^3, it ranks fourth in water content among all the closed lakes of the world [3].

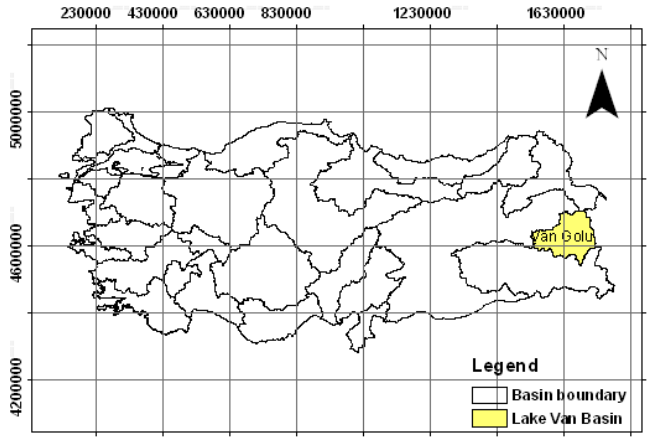


Fig. 1. Location of Lake Van Basin on Turkey

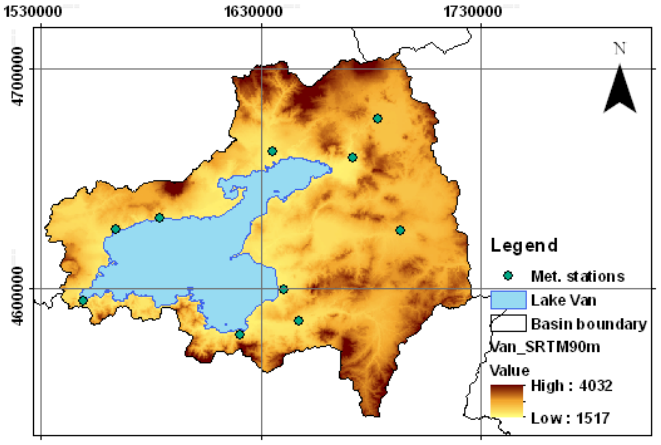


Fig. 2. Location of Lake Van Basin on Turkey Lake Van, SRTM (90m) of basin, and distribution of meteorological stations over basin.

B. Data

The precipitation data used in this study were obtained from the Turkish State Meteorological Service. The primary dependent data source was monthly precipitation measured at 10 meteorological stations between 1981 and 2010. The spatial distribution of stations is not fairly uniform over the basin; when looking the overall distribution condensed placement can be seen near the Lake (Figure 2). The highest monthly precipitation is between 200-280 mm and is measured generally at the October, November, March and April. As independent data source, an elevation map with 1 km spatial resolution was used (Figure 2). It was obtained by resampling the 3 arc second SRTM (the Shuttle Radar Topography Mission) (approximately 90 m spatial resolution) to 1 km spatial resolution.

It is observed in many studies that secondary information can often improve the spatial interpolation of environmental variables [4], [5], [6], [7].

III. Methodology

Monthly precipitation predictions are made on spatio-temporal framework. Each month is interpolated separately. Space-time universal kriging (ST-UK) method is used to obtain predictions over the basin. Elevation is used as secondary variable. For accuracy assessment R-square and RMSE performance measures are used.

A. Space-time Kriging

Consider a variable z which varies in the spatial (s) and time (t) domain. Let z be observed at n space-time points (s_i, t_i), i=1, ..., n. These measurements constitute a space-time network of observations. However it is practically impossible to measure data point z at each spatial and temporal point. In order to obtain a complete space-time coverage, interpolation of z is required. The aim of space-time interpolation is to predict $z(s_0, t_0)$ at an unmeasured point (s_0, t_0), which is a node of a space-time grid. To predict z at these nodes, it is assumed to be a realization of a random function Z which has a known space-time dependence structure. Next $Z(s_0, t_0)$ is predicted from the observations and using the assumed space-time model [1].

The random function Z can be defined with a deterministic trend *m* and a zero-mean stochastic residual *V* as follows (1):

$$Z(s,t) = m(s,t) + V(s,t) \quad (1)$$

The deterministic trend *m* represents large-scale variations whereas the stochastic component *V* represents small-scale variations [1].

B. Cross-validation

Ten-fold cross-validation method was used to evaluate the performances of the space-time interpolation technique [8], [9]. For this purpose, the total dataset comprising all measurements was randomly split in ten (approximately) equally sized sub-datasets. For each sub-dataset, the remaining 90% of the data was used as a training set to calibrate the space-time prediction model and make predictions of monthly precipitation at the sub-datasets that was set aside, and which comprises the test or validation dataset. In this way, predictions at the test dataset locations were compared with the observed data for each of ten test datasets. Every measurement was used once in test datasets. Performance assessment was done by comparing the Root Mean Squared Error (*RMSE*), and *R-square.*

IV. Results and Discussion

Space-time Universal kriging is performed for each month separately. Only the results for the January month are represented in this paper (Figures 3). As seen from the Figure, prediction maps have less detail and have similar values for consecutive time periods. Nevertheless, predictions are obtained for each spatial and temporal framework. This undesirable situation is resulted because of using few observations. In Addition, the meteorological stations do not have the complete observations for every month.

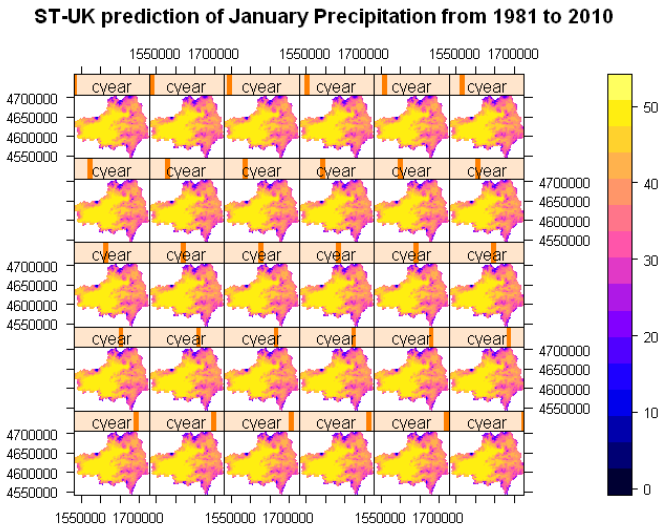


Fig. 3. Space-time Universal kriging prediction maps.

V. CONCLUSIONS

In this study space-time kriging method was applied to predict monthly precipitation of the Lake Van Basin, Turkey. Measurements obtained from ten meteorological stations were used for 1981-2010 time period. Secondary variable that vary in space but are static in time (elevation) was used by the space-time Universal kriging method. ST-UK method resulted with reasonable prediction values at space; however prediction values for each time scale are very similar to each other. Thus from this study, it is understood that using limited number of observations at space-time kriging gives unsatisfactory results with regard to temporal framework.

ACKNOWLEDGMENT

I would like to thank to Yuzuncu Yil University, Coordination of Scientific Research Projects for supporting me financially during this study.

REFERENCES

[1] Heuvelink, G. B. M., Griffith, D. A., "Space–Time Geostatistics for Geography: A Case Study of Radiation Monitoring Across Parts of Germany", *Geographical Analysis* 42 (2010) 161–179.

[2] Gething, P.W., Atkinson, P.M., Noor, A.M., Gikandi, P.W., Hay, S.I., Nixon, M.S., "A local space-time kriging approach applied to a national outpatient malaria data set", Computers & Geoscienses 33, 2007, pp 1337-1350.

[3] Degens, E.T., Wong, H.K., Kempe, S., Kurtman, F., "A Geological study of Lake Van, Easten Turkey", Geologische Rundschau 73, 2, pp 701-734, 1984.

[4] Bostan, P.A., Heuvelink, G.B.M., Akyurek, S.Z., "Comparison of Regression and Kriging Techniques for Mapping the Average Annual Precipitation of Turkey", International Journal of Applied Earth Observation and Geoinformation 19, pp 115-126, 2012.

[5] Lloyd, C.D., 'Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain', Journal of Hydrology 308, pp 128-150, 2005.

[6] Hofierka, J., Parajka, J., Mitasova, H., Mitas, L., 'Multivariate interpolation of precipitation using regularized spline with tension', Transactions in GIS 6 (2), pp 135–150, 2002.

[7] Boer, E. P. J., Beurs, K. M., Hartkamp, A. D., 'Kriging and thin plate splines for mapping climate variables', International Journal of Applied Earth Observation and Geoinformation 3 (2), pp 146-154, 2001.

[8] Gilardi, N., Bengio, S., 'Local machine learning models for spatial data analysis', Journal of Geographic Information and Decision Analysis, volume 4, number 1, pp 11-28, 2000.

[9] Rigol-Sanchez. J. P., Chica-Olmo, M., Abarca-Hernandez, F., 'Artificial neural networks as a tool for mineral potential mapping with GIS', International Journal of Remote Sensing, volume 24, number 5, pp 1151-1156, 2003.

Experiences with Interpolation of Daily Values of Various Meteorological Elements in the Czech Republic

[abstract]

Petr Stepanek
Global Change Research Centre AS CR,
Department of climate modelling and scenarios development
Brno, Czech Republic

***Abstract*— Several methods for interpolation of daily values of various meteorological elements are compared for the area of the Czech Republic. Maps were generated for the period 1961-2010 using IDW and various kriging methods. Suitable settings for air temperature, relative humidity, wind speed, sunshine duration and precipitation were found. For the task, ArcView, RAP (http://www.striz.info/rap/) and R software were linked to ProClimDB software (www.climahom.eu) for enabling automation of the calculation process. The experiences with the data processing are presented.**

Indicator Kriging vs. Sequential Indicator Simulation in Mapping Probabilities of Precipitation Occurrence

[full paper]

Jelena Pandžić
Department of Geodesy and Geoinformatics
Faculty of Civil Engineering, University of Belgrade
Belgrade, Serbia
jpandzic@grf.bg.ac.rs

***Abstract*—Estimation and simulation are two forms of geostatistical prediction used to assess spatial distribution of a continuous variable (e.g. precipitation) sampled at a finite number of locations. They can be considered as two optimization problems differing in optimization criteria: estimation means minimizing a local error variance and simulation strives to reproduce global statistics (variogram and histogram) of a variable. This paper compares indicator kriging (IK) to sequential indicator simulation (SIS) on the example of mapping probabilities of precipitation occurrence on the territory of the Republic of Serbia. Indicator means that no statements on spatial distribution of the original variable values are made, but rather on spatial distribution of probabilities that the original variable values exceed (or not) some threshold value. The data of four distinctive months (February, June, August and October) in 2009 were chosen as a basis for the prediction. One of the aims was to emphasize the smoothing effect of kriging on stochastic surface which illustrates spatial variability of a certain phenomenon. Although significant similarities between corresponding kriging and averaged simulation maps could be noticed, it is evident that in some cases simulation is much more careful when it comes to prediction of spatial variability, avoiding statements on the existence of the areas of extreme probabilities for something to happen.**

Keywords—indicator kriging; sequential indicator simulation; precipitation occurrence probabilities

I. INTRODUCTION

Precipitation, as one of the most important climate elements, is a typical example of a continuous, spatially variable phenomenon which requires for conclusions on its spatial distribution to be derived, based on the data sampled at various locations. Possible spatial distribution of precipitation is thus obtained by applying different geostatistical prediction methods. The prediction occurs in two forms, as an estimation and as a simulation [1], whereby the estimation gives one map as a result and the simulation gives greater number (usually about hundred) of maps. The map obtained by using estimation is statistically speaking the best linear unbiased estimate (BLUE) of the variable spatial distribution, whereas maps obtained via simulation illustrate equally probable spatial distribution of the observed variable. In that sense, estimation and simulation can be considered as two optimization problems differing in optimization criteria: estimation means minimizing a local error variance and simulation strives to reproduce global statistics (variogram and histogram) of a variable [2].

The aim of this paper was to assess spatial distribution of the probabilities for the occurrence of a certain amount of precipitation on the territory of the Republic of Serbia in the year 2009. The prediction was done via two different geostatistical methods, indicator kriging (IK) and sequential indicator simulation (SIS), whose results were afterwards compared to each other. Apart from this comparison, the obtained results were used for the verification of the current knowledge on the spatial distribution of rainfall on the territory of the Republic of Serbia.

II. INDICATOR KRIGING

Mathematical description of a spatial variation of any variable, which is the essence of kriging applications, is performed using the sum of the three main components [3]:

$$Z(x) = m(x) + \varepsilon'(x) + \varepsilon'' \quad (1)$$

with:

$Z(x)$ – being a value of a random function,

$m(x)$ – being a deterministic function that describes the so-called structural component, i.e. trend,

$\varepsilon'(x)$ – being a stochastic (random) component that is spatially correlated and represents the remainder of the structural component, also known as a regionalized variable,

ε'' – being a residual error, i.e. spatially uncorrelated noise.

Based on different approaches to treating some of the components of spatial variation of a variable, especially the trend, one can distinguish between kriging variants: simple, ordinary, universal, regression, indicator kriging, cokriging, etc. Unlike the other kriging variants which give an estimated value of a variable of interest, i.e. of a spatial attribute at a certain location as a result, indicator kriging provides

information on the probability that a variable value at a certain location exceeds some predefined limiting value, i.e. threshold [4].

Applying indicator kriging requires for binarization of the original data to be done. This means that all the values of the observed continuous variable must be transformed to the so-called indicators. Every original variable value is replaced with a value of 1 or 0 depending on whether the observed value is below or above defined threshold [5]. Mathematically, this nonlinear transformation can be represented by the formula:

$$i(x, z_k) = \{1, \text{for } z(x) \leq z_k\} \text{ or } \{0, \text{for } z(x) > z_k\} \qquad (2)$$

with:

$z(x)$ – being a measured variable value at the point x,

z_k – being a boundary value, i.e. threshold,

$i(x, z_k)$ – being a transformed variable value (indicator) at the point x for the given threshold value z_k.

Bearing in mind that normality compliance anyhow disappears with binarization of the data, indicator kriging does not require for the original variable values to comply with the normal distribution [6]. Indicator kriging is also an especially efficient way to limit the effect of extremely big values or outliers on the results of prediction due to the fact that variable values are assigned the same indicator as other values that are above the set threshold regardless of the absolute difference [7]. Yet, this limitation of the extremes is at the same time a serious shortcoming of the indicator kriging method. Namely, since indicator kriging aims at minimizing a local error variance, the so-called smoothing of a stochastic surface representing variable spatial distribution is done, which ultimately leads to losing the information on the original spatial variability of a sample used in the prediction. Thus, kriging is not a particularly suitable prediction method in situations where extreme variable values or significant local variations of variable values are present. In those cases, simulation is preferably used.

Although it was originally used for mapping mineral resources, today more and more possible (and successful!) applications of indicator kriging arise: the application in the area of water quality assessment [8], precipitation mapping [6, 9], drawing conclusions about prevalence of certain diseases like schistosomiasis in humans [10], just to mention a few.

III. Sequential Indicator Simulation

Geostatistical simulation is a spatial extension of Monte Carlo simulation concept, whose goal differs to a significant extent from the estimation goal, i.e. from the goal of kriging application. The essence of geostatistical simulation is reproducing variance of the original data, in one-dimensional sense (through histogram), as well as in space (via variogram). Generally, all realizations (simulations) differ from one another and each individual simulation is worse estimate than that obtained by applying the appropriate kriging method. Nevertheless, averaging large number of simulations leads to a good estimate, ultimately to the one gained from geostatistical interpolation, i.e. kriging [11].

Besides reproducing histogram and spatial variability of the data, simulation can honor the data themselves, i.e. take into account concrete variable values which condition gaining some (unknown) variable value at a certain location. This type of simulation is called *conditional simulation*.

The choice of a simulation method largely depends on the nature of the variable whose spatial distribution is to be simulated, so it can be distinguished between [11]:

- pixel-based methods (nonparametric, Gaussian and fractal methods) and
- object-based methods (point processes, Boolean methods).

Sequential indicator simulation together with p-field simulation is the most frequently used nonparametric simulation method. Nonparametric methods are the result of the indicator approach in geostatistics, which means using indicators for conducting structural analyses suitable for describing spatial distribution of some categorical variable or continuous variable transformed into categorical one based on predefined threshold values [11].

Algorithm of sequential indicator simulation consists of the following steps [11, 12]:

1. original variable values are transformed into indicators (every original value is replaced with the indicator vector containing only digits 1 and 0, which defines affiliation of the original value to a certain class),
2. order by which grid cells (in which indicator variable values are to be simulated) will be visited, is defined by random choice,
3. in the first cell, k probabilities that the unknown variable value at that location belongs to each of the k defined classes are calculated (probabilities are conditioned by the set of known indicator variable values in the neighborhood of the observed cell),
4. based on calculated probabilities, conditional probability distribution function (cpdf), i.e. conditional cumulative distribution function (ccdf) is determined for the observed grid cell,
5. number between 0 and 1 is picked by random choice – that number represents probability based on which one, by inspecting the corresponding ccdf, determines the class the unknown variable value at the observed location belongs to; indicator vector for that grid cell is then filled by giving the value of 1 to a class the cell belongs to and 0 to all the other classes,
6. simulated indicator vector for the observed location is added to the set of known values which condition simulation of values in the next grid cell,
7. steps 3-6 are repeated for all grid cells, whereby the cell visiting order was defined in step 2.

IV. Precipitation Regime in Serbia and Data Used

The territory of the Republic of Serbia is characterized by two precipitation regimes, continental and Mediterranean, whereby the greater part of Serbia belongs to the continental regime. Continental regime means that the greatest amount of precipitation occurs in May and June, while the least occurs in February and October. Areas that belong to the Mediterranean regime, which is the case with the southwestern part of Serbia, experience a rainier period in November, December and January and a drier one in August [4, 13]. These facts were the reason why February, June, August and October 2009 were chosen as time periods of interest for the prediction of probabilities for the occurrence of a certain amount of precipitation on the territory of the Republic of Serbia.

Comparison of the results obtained by using two distinctive prediction methods, indicator kriging and indicator sequential simulation, was done, as well as the verification of the current knowledge on the spatial distribution of rainfall throughout Serbia. For this purpose, the data from relatively uniformly distributed weather stations at the territory of the Republic of Serbia were used. Geographic coordinates, elevation and cumulative monthly precipitation amount during the aforementioned four months of the year 2009 were provided for 191 stations in total. Prediction was done for each grid cell, whereby the territory of the Republic of Serbia was gridded with the resolution of 1 km × 1 km.

V. Methodology

Applying indicator kriging and sequential indicator simulation requires binarization of the original variable values. For the purpose of the transformation of the original data into indicators, median of every dataset (one dataset per month) was chosen as a threshold value for a particular case of prediction. For example, in the case of cumulative monthly precipitation amount in June 2009 at the territory of the Republic of Serbia median had a value of 119.5 mm, which means that the value of 1 was assigned to every variable value less than (or equal to) 119.5 mm, while zeros were assigned to the variable values exceeding the threshold value. Medians of cumulative monthly precipitation amounts for distinctive months in 2009 are given in Table I.

TABLE I. Medians

Month	February	June	August	October
Median [mm]	57.8	119.5	43.8	105.0

The transformed data were used to calculate experimental variograms which were afterwards modeled by variograms based on different mathematical functions. Unlike the variogram model for February 2009 which utilized exponential function, variogram models for June, August and October 2009 were based on using spherical function. The same variogram model was used both in SIS and for the estimation by the means of IK, and this was the case with all four months.

Simulation and kriging were implemented through R software environment using its packages (particularly gstat, rgdal and RSAGA). Sequential indicator simulation was completely conducted according to the procedure previously described in the paper. The total of 100 realizations of possible spatial distribution of precipitation occurrence probabilities was created for every distinctive month. When conducting simulation, the maximum number of 20 nearest points (i.e. grid cells) was used for conditioning prediction of an unknown variable value at a certain grid cell. Introducing this restriction was inevitable since otherwise simulation would last for a very long time period (if it could be completed at all) because of taking into account all known variable values at a particular moment.

Fig. 1 depicts four characteristic realizations (25^{th}, 50^{th}, 75^{th} and 100^{th}) obtained during applying SIS on the data of June 2009. All realizations, of course, consist only of grid cells with values of either 1 or 0, meaning that the precipitation amount is either within previously defined limits or not. The most probable spatial distribution of probabilities for the occurrence of a precipitation amount that does not exceed the chosen threshold value was obtained by simply averaging all of the realizations, i.e. simulations for a particular month.

For every month, the averaged map obtained through sequential indicator simulation was compared to the corresponding map created by utilizing the indicator kriging algorithm. Again, functions implemented in R packages were used for performing the necessary calculations. The total of four IK maps was created, one for every observed month. It is important to notice that, because of some prediction errors, the occurrence of probabilities outside the range of [0,1], which seems impossible to common sense, is rather usual. These probabilities have to be corrected, i.e. all probabilities greater than 1 (100%) have to be replaced with the value of 1 and all negative probabilities have to be replaced with the value of 0. This was done for all obtained "problematic" variable values and the final IK maps were immediately afterwards created.

VI. Results and Discussion

Maps given in Fig. 1 substantiate claims that every simulation results in different predicted variable values. The most evident difference can be seen when comparing Fig. 1a) to Fig. 1d), i.e. 25^{th} to 100^{th} realization. Nevertheless, all maps given in Fig. 1, although binarized, look very much like the averaged map of all realizations given in Fig. 2b) (map on the right). The similarity is most obvious in the area of western Serbia, where white color prevails in all of the shown realizations and the lightest grey in the averaged map, which points to the small possibility that precipitation amount in that area does not exceed the chosen threshold value.

Fig. 2 shows final maps of probabilities for the occurrence of a certain amount of precipitation on the territory of the Republic of Serbia in February, June, August and October 2009. The maps on the left are the result of applying indicator kriging method during prediction, while the maps on the right are the averaged maps obtained from sequential indicator simulation. In every map, dark-grey color corresponds to the interval [0.8,1], i.e. [80%,100%] and suggests that there is a great possibility that the monthly precipitation amount at an observed location does not exceed the threshold value. On the other hand, white color that corresponds to the interval [0,0.2]

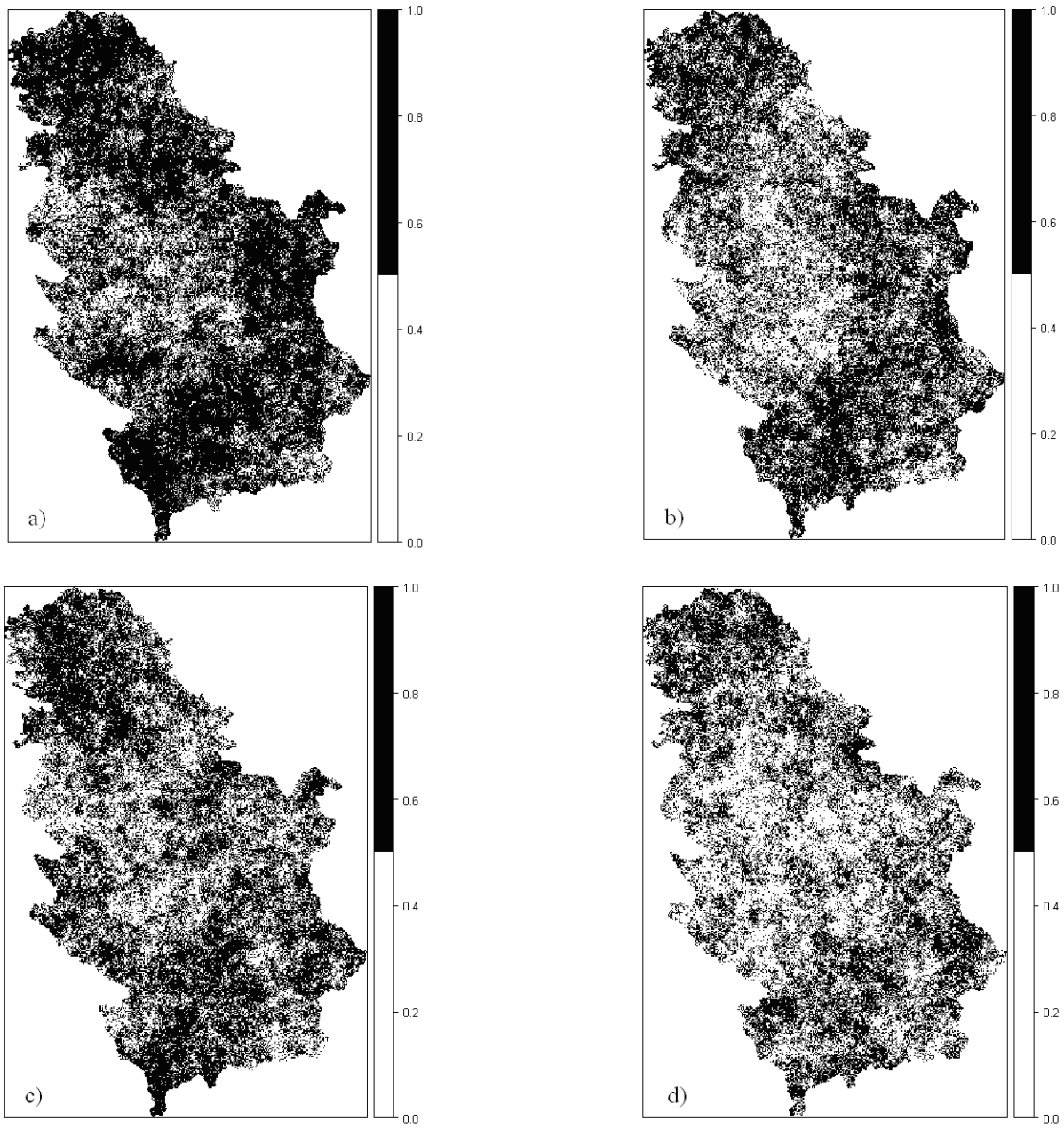


Fig. 1. Characteristic realizations of SIS for June 2009: a) 25th, b) 50th, c) 75th and d) 100th realization.

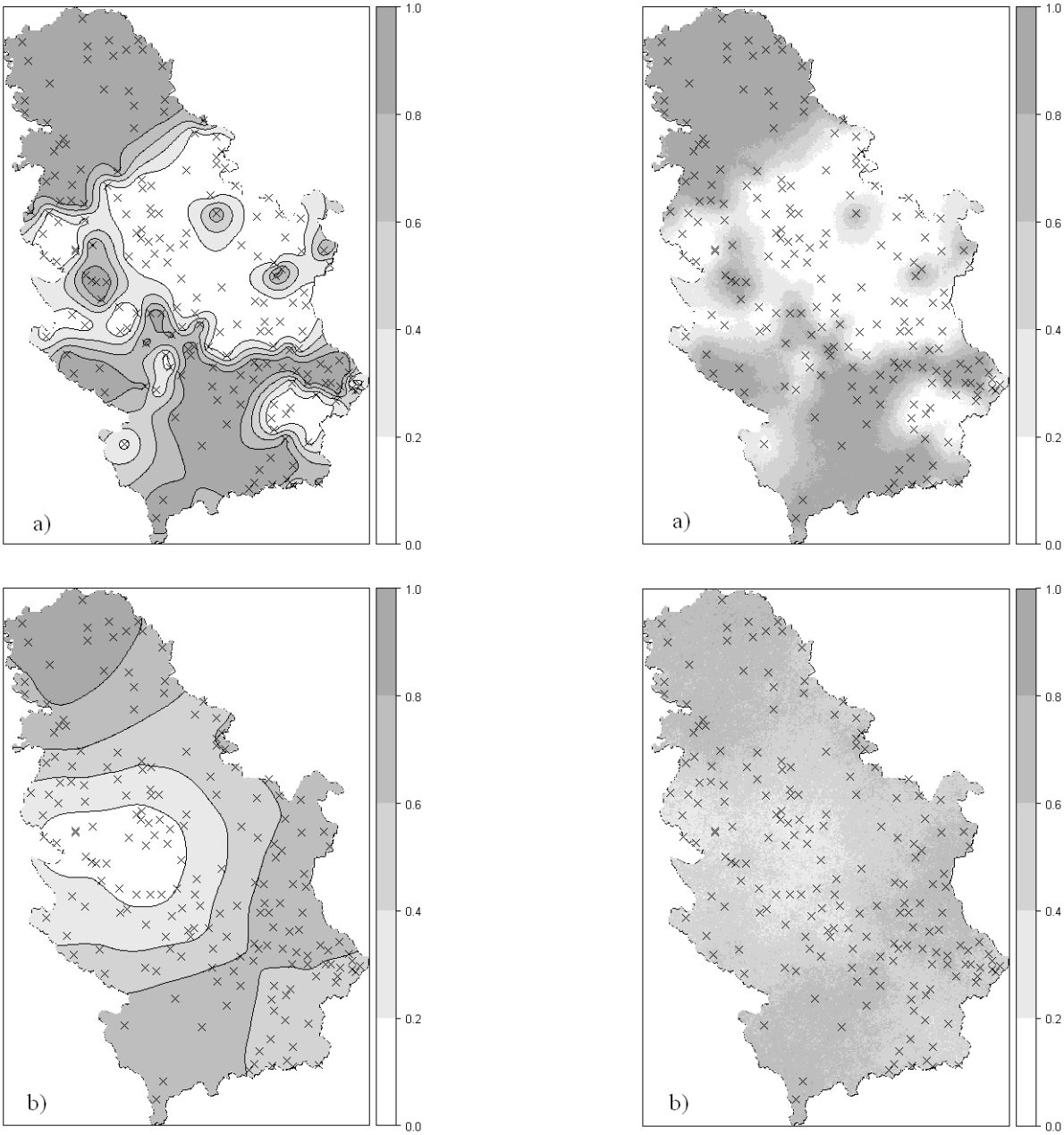
points to the fact that it is pretty unlikely the precipitation amount is within given limits, in other words, it suggests that there is a huge possibility that the precipitation amount at an observed location exceeds the threshold value [4].

By comparing every IK map to its corresponding SIS map, the basic difference between estimation and simulation reflecting in smoothing of the variations, can be seen. While the maps of the best linear unbiased estimates feature only a few probability zones (see Fig. 2, maps on the left, especially Fig. 2b) and Fig. 2c)) based on which probability maps in vector form could be derived, that is not the case with the maps obtained in simulation process (Fig. 2, maps on the right). SIS maps generally depict more pronounced variations and although several areas of different probability ranges could be distinguished, points from other ranges stay in those areas. This implies that creating a probability distribution map based on the simulated values in vector form would not be of a great use since it would be pretty unreadable because of a huge number of polygons or simply many pieces of information would be lost because of generalization, which would ultimately lead to gaining a map similar to IK map (map with smoothed variations). By comparing maps for February 2009 (Fig. 2a)), it could be noticed that the SIS map looks very much like the corresponding IK map, more than in case of other months. But, although in this case creating a vector map based on the results of SIS seems to be achievable, variations on the SIS map are still considerably less smoothed than those on the IK map,

which, in the case of the aforementioned creation of a vector map, would ultimately lead to polygons with very rough edges.

Despite the obvious differences between corresponding IK and SIS maps dealing with smoothing of the stochastic surface of the variations, the aforementioned maps undoubtedly point to a very similar distribution of the probabilities for the occurrence of a certain amount of precipitation on the territory of the Republic of Serbia. Moreover, the obtained maps given in Fig. 2 verify current knowledge on the spatial distribution of rainfall throughout Serbia. Northern parts of Serbia, the Morava valley and the territory of Metohija are the regions with relatively low precipitation amounts [14], which is confirmed by dark colors that those regions feature in all of the maps given in Fig. 2 (especially IK maps for June and October). The rainiest regions of Serbia are located in the west which is again obvious from all of the maps since these areas are colored in white or light-grey. Considerably higher precipitation amounts in the western part of the country are caused by a cold front and showers brought by cold air masses coming from the Atlantic and western Europe [15]. It is the reason why the western parts of the country receive more precipitation than the eastern ones, although they are located at the same latitude.

An important thing to notice is that, in some cases, the prediction based on simulation is not that exclusive as the one based on estimation, in the sense that, in the SIS maps, areas with probabilities in the ranges of [0.2,0.4], [0.4,0.6] and [0.6,0.8] prevail, with only a few points with the probability in the range of [0,0.2] (western and central Serbia) or [0.8,1] (northern Serbia). This could most obviously be seen in Fig. 2b) and 2c) (maps on the right). Again, the exception is the map shown in Fig. 2a) (map on the right), i.e. the SIS map for



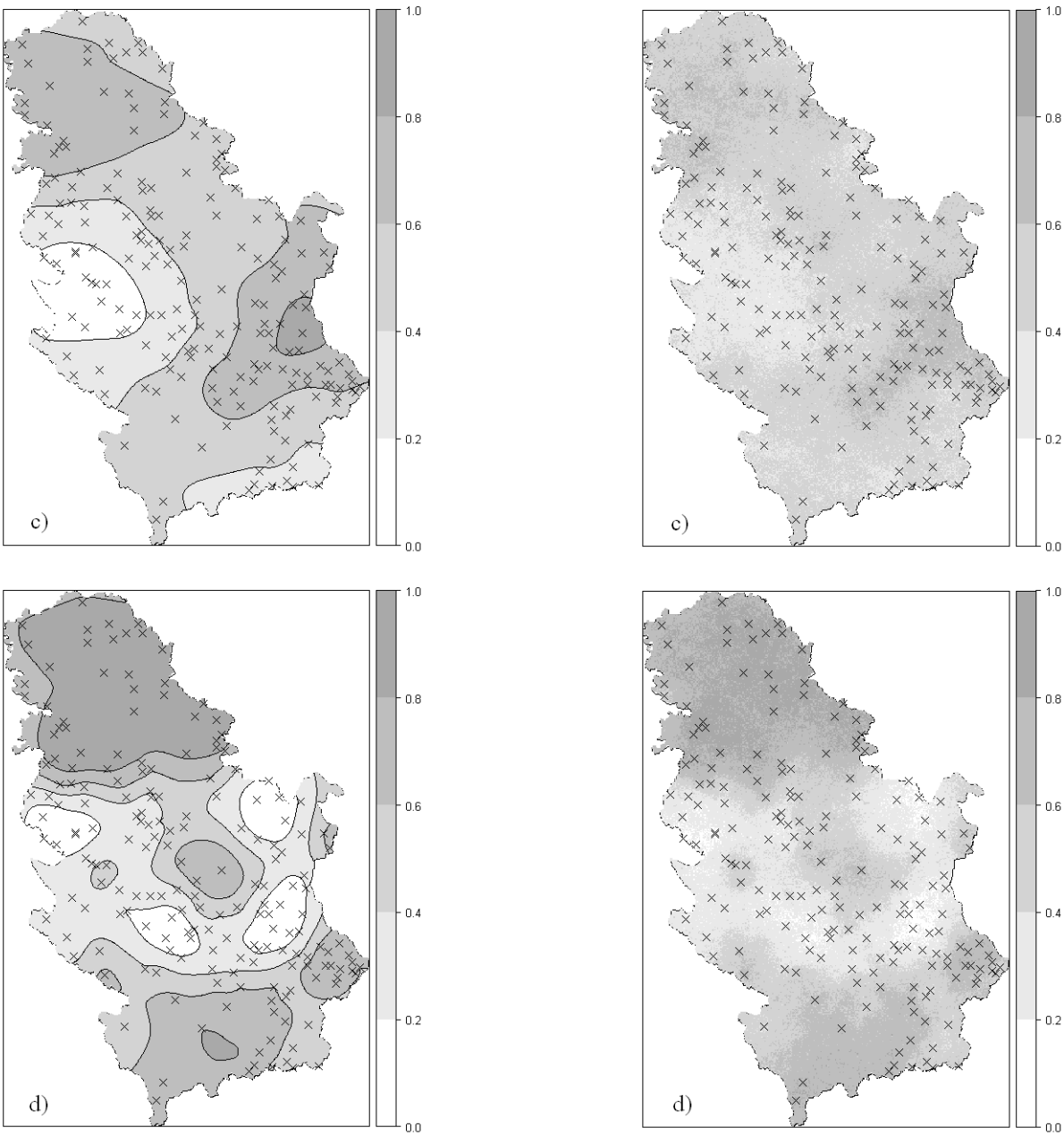


Fig. 2. Results of IK (left) and SIS (right) for: a) February, b) June, c) August and d) October 2009.

February 2009, but this disagreement with the previous statements probably appear due to the nature of the data. The fact that the areas with extreme probabilities (whether big or small) occur only sporadically in resulting maps, suggests that the prediction obtained through simulation is much more moderate regarding the values it gives as a result, than the one obtained using indicator kriging.

VII. Conclusion

Climate variables, among which precipitation as well, vary in space and time. Those irregular variations cannot be adequately described by simple mathematical functions, therefore more complex methods, including geostatistical prediction, are required. Geostatistical prediction of unknown variable values can be done through estimation or simulation. The main difference between these two ways of prediction lies in the fact that by using estimation only one and that is the best linear unbiased estimate of spatial distribution of an observed variable is obtained, while simulation means generating several different distributions that are equally probable. Obtaining the best estimate requires smoothing of stochastic surface of the variations to be done. This is not the case with simulation, because, unlike estimation, it reproduces the variability of an observed variable visible from the available sample (histogram and variogram). Simulation gives a great number of the so-called realizations, out of which the map of the most probable spatial distribution of an observed variable is derived by simple

averaging. The final map depicts local variations as well, which usually stay overlooked in the case of estimation.

Within this paper prediction of probabilities for the occurrence of a certain amount of precipitation on the territory of the Republic of Serbia was done by using two prediction methods: indicator kriging and sequential indicator simulation. Prediction was done for February, June, August and October 2009 and, since the original data were given in the form of cumulative monthly precipitation amounts at different locations throughout Serbia, transformation of the data into indicators had to be done. Medians of the available datasets were chosen as respective threshold values.

All the maps given in Fig. 2, although obtained by using different prediction methods and thus differing in the degree of smoothing of stochastic surface of the variations, have something in common. They all correspond to the well-known spatial distribution of precipitation in Serbia, thereby identifying the northern parts of the country, the Morava valley and Metohija as regions in which the occurrence of a smaller precipitation amount (the amount within the limits of the defined threshold) is more likely to happen. The western parts of Serbia were confirmed to feature greater probabilities of abundant precipitation occurrence, i.e. precipitation which by amount exceeds the limits defined when conducting geostatistical prediction.

ACKNOWLEDGMENT

The author would like to thank the Ministry of Education, Science and Technological Development of the Republic of Serbia for financially supporting her work through Contract No. TR 36009. Gratitude also goes to Ms. Jelena Luković from the Faculty of Geography, University of Belgrade, for providing the author with the precipitation data.

REFERENCES

[1] Y. Zhang, Introduction to geostatistics - course notes [online]. Laramie: University of Wyoming, Department of Geology and Geophysics, 2011. Retreived from http://geofaculty.uwyo.edu/yzhang/files/Geosta1.pdf [19.04.2014].

[2] P. Goovaerts, "Estimation or simulation of soil properties? An optimization problem with conflicting criteria," Geoderma, vol. 97, no. 3, pp. 165-186, 2000.

[3] P.A. Burrough and R.A. McDonnell, Principles of Geographic Information Systems, Serbian translation by B. Bajat and D. Blagojević. Beograd: Građevinski fakultet, 2006.

[4] J. Pandžić, B. Bajat and J. Luković, "Mapping probabilities of precipitation occurrence on the territory of the Republic of Serbia by the method of indicator kriging," Bulletin of the Serbian Geographical Society, vol. 93, no. 2, pp. 23-40, 2013.

[5] E.H. Isaaks and R.M. Srivastava, Applied Geostatistics. New York: Oxford University Press, 1989.

[6] P.M. Atkinson and C.D. Lloyd, "Mapping precipitation in Switzerland with ordinary and indicator kriging," Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 65-76, 1998.

[7] I. Glacken and P. Blackney, "A practitioners implementation of indicator kriging," in Proceedings of the Symposium "Beyond Ordinary Kriging: Non-Linear Geostatistical Methods in Practice," J. Vann, Ed. Perth: The Geostatistical Association of Australasia, 1998, pp. 26-39.

[8] R. Tolosana-Delgado, Simplicial indicator kriging: presentation [online]. Wuhan: China University of Geosciences, 2007. Retreived from http://www.sediment.uni-goettingen.de/staff/tolosana/extra/Wuhan-talk-4.pdf [22.04.2014].

[9] X. Sun, M.J. Manton and E.E. Ebert, "Regional rainfall estimation using double-kriging of raingauge and satellite observations," BMRC research report no. 94. Melbourne: Australian Government, Bureau of Meteorology, 2003.

[10] R.J.P.S. Guimarães et al., "Use of indicator kriging to investigate schistosomiasis in Minas Gerais State, Brazil," Journal of Tropical Medicine [online], vol. 2012, 2012. Retreived from http://www.readcube.com/articles/10.1155/2012/837428?locale=en [22.04.2014].

[11] J. Vann, O. Bertoli and S. Jackson, "An overview of geostatistical simulation for quantifying risk," in Proceedings of the Symposium "Quantifying Risk and Error." Perth: The Geostatistical Association of Australasia, 2002.

[12] J.J. Gómez-Hernández, "Indicator conditional simulation of the architecture of hydraulic conductivity fields: application to a sand-shale sequence," in Proceedings of the Symposium "Groundwater management: Quantity and Quality," A. Sahuquillo, J. Andreu and T. O'Donnell, Eds. Wallingford, Oxfordshire: IAHS Press, 1989, pp. 41-51.

[13] Republic Hydrometeorological Service of Serbia - RHMS, Padavinski režim u Srbiji [online], 2013. Retreived from http://www.hidmet.gov.rs/podaci/meteorologija/latin/Padavinski_rezim_u_Srbiji.pdf [10.04.2014].

[14] V. Ducić and M. Radovanović, Klima Srbije. Beograd: Zavod za udžbenike i nastavna sredstva, 2005.

[15] M. Unkašević and I. Tošić, "A statistical analysis of the daily precipitation over Serbia: trends and indices," Theoretical and Applied Climatology, vol. 106, pp. 69–78, 2011.

Spatial Prediction Using Uncertain Variogram

[full paper]

Jan Caha
Department of Geoinformatics,
Faculty of Science, Palacký University in Olomouc
17. listopadu 50, 771 46, Olomouc, Czech
jan.caha@upol.cz

Lukáš Marek, Vít Pászto
Department of Geoinformatics,
Faculty of Science, Palacký University in Olomouc
17. listopadu 50, 771 46, Olomouc, Czech
lukas.marek@upol.cz, vit.paszto@gmail.com

***Abstract*— Uncertainty of results is often as important as results themselves for any type of prediction. This is especially true for methods of prediction that contain epistemic uncertainty, which often takes the form of specification of parameters for the prediction method. These parameters are usually determined by an expert knowledge and perceived as granted, however, their selection is often a matter of opinion and several different solutions are possible. Each of such solutions can provide different prediction. Fuzzy prediction models can be used to handle epistemic uncertainty in models and they provide results in the form of uncertain prediction, which can be used to obtain most likely prediction along with minimal and maximal values of prediction. We decided to apply and test the usability of the fuzzy prediction model, which was developed and described by Loquin and Dubois in Kriging with Ill-Known Variogram and Data (2010).**

In this article we study the influence of epistemic uncertainty of variogram parameters (sill, range, and nugget) selection on results of spatial interpolation of two datasets. Particularly, the well-known and well described meuse data set is used as one of data sources, while the second data set is the collection of mean atmospheric pollution measurements (PM_{10}) in the Czech Republic in 2013. Three possible variograms are selected for each dataset. Modal variogram is constructed as the most likely optimal variogram, while minimal and maximal variograms provide bounds for possible realizations of variograms. The fuzzy surface construction is based on optimisation scheme given by Loquin and Dubois (2010) [8]. The fuzzy surface is than compared against surfaces with simulated parameters from the range specified by minimal and maximal variograms in order to determine its usefulness as boundaries of uncertainty caused by user's selection of variogram parameters. These predictions are further studied. Although the validity of the presented optimisation scheme is not fully proved, the usability and analysis of errors still show only up to 6% of acceptable errors out of 5 000 simulations. That proves the suitability of the procedure for the spatial prediction based on kriging with uncertain variogram.

Keywords—fuzzy surface, variogram, uncertainty, spatial prediction

I. INTRODUCTION

Every system under study contains two types of uncertainty. Aleatory uncertainty has its origins in inherent randomness of the system, while epistemic uncertainty is a result of lack of knowledge [5]. Epistemic uncertainty is often met in the form of fixed values of parameters of the model that actually are not exactly known. Values of such parameters quite often depend partially on the data, and partially on expert knowledge. Whenever expert knowledge is used, it is possible that more than one solution exists and through that fact an epistemic uncertainty is introduced to the model [9]. Probabilistic representations of uncertainty are quite often used for handling the epistemic uncertainty. Even though this approach is successful, it has been criticized for requiring too detailed knowledge about uncertainty [4]. However, such knowledge is usually not available to the user, so alternative representations of the uncertainty would be more suitable for use [11,5]. The alternative theories for epistemic uncertainty representation are evidence theory, possibility theory and fuzzy set theory [4].

The problematic of epistemic uncertainty affects all models used for predictions, and spatial prediction methods are no exception to that fact. Every method used for spatial interpolation has a set of parameters that are adjusted and based on expert knowledge of the data. Very often, the influence of these parameters on spatial prediction is not discussed and the prediction based on exact values of the parameters is considered as certain. However, the selection of these parameters can affect the result quite significantly [9,1].

In this paper we study the approach to handle the epistemic uncertainty in the kriging interpolation method presented in [8]. The authors provided method that leads to creation of fuzzy surface because such surface does incorporate the uncertainty of the interpolation parameters. The method for creating the surface is potentially computationally very demanding. An optimisation scheme, that overcome this problem was proposed by the authors [8] as well, but this optimisation algorithm has not been studied and verified so far. In this research we tested two datasets to create fuzzy surface using the optimisation algorithm and perform experiments to verify whether it provides bounds of the solutions that would be obtained without the optimisation.

II. FUZZY NUMBERS

Fuzzy numbers are special cases of fuzzy sets that represent vague, imprecise or ill-known values [3,7]. Like a fuzzy set a fuzzy number is defined by a membership function, which specifies membership degree for each element x from the universe X. The membership function of fuzzy number $\tilde{A}$ is

usually denoted as $\mu_{\tilde{A}}(x)$. Fuzzy number has to be a normal convex fuzzy set, with at least piecewise continuous membership function that is defined on the universe of real numbers [3,7]. Fuzzy numbers are proven to be well suited for calculation with imprecise values in situation when uncertainty of the value is not result of variability [3,7,11]. Fuzzy number than forms bounds around uncertain value and allows further processing of such vague value by means of fuzzy arithmetic.

Triangular fuzzy numbers (TNF) are of special interest for the purpose of this research because of three main reasons; (1) TNF is simple models of uncertain numbers, (2) TNF is specified by three values $[min, modal, max]$ and (3) TNF has linear membership function between these values:

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 \text{ if } x < min \text{ or } x > max \\ \frac{x-min}{modal-min} \text{ if } min \leq x \leq modal \\ \frac{\max - x}{\max - modal} \text{ if } modal < x \leq max \end{cases} \quad (1)$$

The TNF is completely defined by these three values [3]. $Modal$ value specifies the most likely value of the uncertain number while min and max forms bounds for possible realizations of the uncertain value. This can also be called range of fuzzy number.

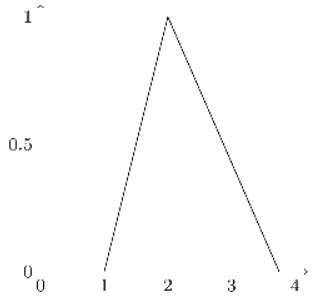


Fig. 1. Example of a fuzzy number represntin uncertain value, that can be described as approximatelly 2.

III. Fuzzy Surface

Classic surface is a surface, where for each coordinate pair x, y exists one value of z that is associated with this location. The value of z specifies the height of the surface at the given location. On fuzzy surface, each location x, y has associated with fuzzy number $\tilde{Z}$, that represents possible set of values that the surface can take at the given location [2]. Such model naturally contains uncertainty of the surface (Fig. 2).

There are two main approaches to the creation of fuzzy surfaces. The first method starts with uncertain data (specified as fuzzy numbers) and extends the interpolation process by using extension principle [2]. Alternative approach uses crisp data, which are much more common, and specifies parameters for interpolation as fuzzy numbers, which leads to result being also fuzzy number [1,8].

IV. Method of Fuzzy Surface Creation

The process of the creation of fuzzy surfaces based on the kriging with uncertain variogram was originaly presented in [1] and lately improved by [8,10]. The approach is based on the premise that the selection of variogram parameters (nugget, sill and range) depends mainly on the expert's opinion and therefore there exist several possible solutions. In such situation, each of those parameters can be specified by the expert as a fuzzy number with specific $min, modal$ and max values. This is especially useful in situations when the shape of the experimental variogram is ambiguous and it is difficult to fit the theoretical variogram [1]. According to [8] there is a little difference between manual and automated fitting of variograms. In each case the uncertainty is present in the selection of parameters. This is one point where the epistemic uncertainty is present in the spatial prediction done by kriging.

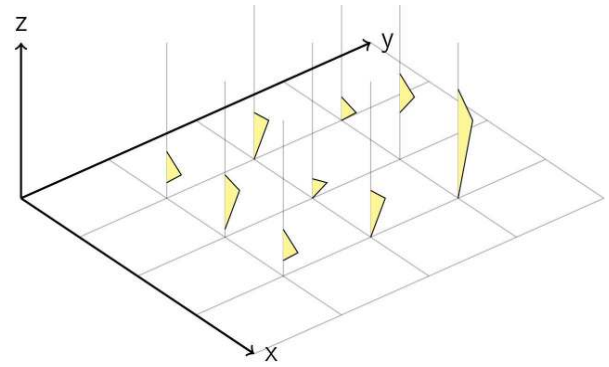


Fig. 2. Example of a fuzzy surface

Loquin and Dubois [8] also point out another issue that is completeness of the data set that is used to construct the variogram. Since this dataset is only "representative" sample of the complete dataset, it is quite possible that it is incomplete and it is definitely ill-known. So it is reasonable to consider the variogram as an uncertain representation of the reality that is most likely not complete and most likely missing some part of information.

Based on mentioned premises, the method for the calculation of kriging with fuzzy variograms was proposed [1,8]. The variogram is considered as fuzzy (Fig. 3) because of its parameters (sill, range and nugget) are specified as triangular fuzzy numbers. The calculation with fuzzy numbers is done by means of fuzzy arithmetic [3,7] that is based on the extension principle[15]. However, as noted by several authors [3], the direct use of extension principle is very complicated and computationally very demanding. This is especially true in the case of the kriging calculation, which is computationally expensive on its own. Because of these facts, it is useful to have optimisation scheme that would simplify the calculation. Such optimisation practices exist for fuzzy arithmetic [3].

Three predictions at each location x, y need to be calculated to obtain fuzzy surface. $Modal$ value of the TNF is easy to obtain as it can be obtained directly from the modal variogram. But calculation of min and max values is more problematic since their calculation requires solving of a global optimisation problem [8]. The optimisation scheme for kriging with fuzzy variogram that allows avoidance of this computationally complicated task was proposed in [8].

A. The Optimisation Scheme

Suppose that there are values of sill, range a nugget specified as fuzzy numbers. So there are three variables that

have modal, minimal and maximal value. The objective of fuzzy kriging is to calculate $\tilde{Z}$ at each location. $Modal$ value is obtained directly by calculating kriging with $modal$ values of sill, range and nugget. Min and max values of $\tilde{Z}$ should be obtained from all possible combinations of sill, range and nugget over their ranges [8]. However this step is rather complicated and problematic, as it requires solving of global optimization problem, so preliminary optimization, that provides estimates of the min and max values, can be done [8]. According to the authors [8] it has been empirically observed that the bounds of the prediction $\tilde{Z}$ are formed for extreme combinations of kriging parameters. That means that practically only 2^3 calculations of kriging need to be done, which lowers the computational load significantly. Minimum and maximum from these calculations are quite likely to be practical bounds of $\tilde{Z}$, however this fact is only based empirical observation and is not universally valid [8].

V. Case Studies

The case studies are designed to verify the practical usefulness of the optimisation scheme presented in [8]. The experiment consists of a creation of fuzzy surface based on uncertain variograms and the optimisation scheme. Then this fuzzy surface is compared with results of probabilistic metaheuristic method: simulated annealing [8], in order to find out if some combination of kriging parameters provides estimates outside of range of the fuzzy surface. The question to be verified is the reliability of the optimisation scheme result.

VI. Case Studies: Data

In this contribution, the influence of epistemic uncertainty of variogram parameters selection on results of spatial interpolation in two datasets was studied. Firstly, the well-known and well described meuse dataset that is freely available e.g. together with R package gstat [12] was used. The concentration of zinc in the soil (or its logarithm to be more precise) was selected as studied characteristic. Secondly, we decided to use the collection of mean atmospheric pollution measurements (coarse particles - PM10) in the Czech Republic in 2013. This dataset was collected by the network of measuring stations owned and managed by Czech Hydrometeorological Institute[1]. Data are published in the form of linked html tables[2] that are possible to parse using suitable R packages (e.g. RCurl [13], XML [14], etc.). Original dataset contains records from 113 measuring stations. Unfortunately, only 86 of them were suitable for further computations due to their completeness.

VII. Variograms

Variogram (or semivariogram) is usually a crucial part of geostatistical analysis. The variogram plots semivariance as a function of distance and therefore, it describes how similarity (spatial autocorrelation or spatial dependence) decreases with the distance. Its main characteristics, except the type of fitted model, are nugget, sill and range. Nugget parameter describes measurement errors or variance in lower scales; sill is the total nonspatial variance of the data set and (practical) range is a distance at which the semivariance is close to 95% of the sill [6]. They are usually set as initial parameters during the fitting of the theoretical model of variogram that comply with the experimental variogram. An experimental variogram is a plot showing how one half the squared differences between the sampled values (semivariance) changes with the distance between the point-pairs. It is usually expected to see smaller semivariances at shorter distances and then a stable semivariance (equal to the global variance) at longer distances [6].

TABLE I. Parameters of fitted theoretical variogram models

	Meuse			PM_{10}		
	Min	***Mod***	***Max***	***Min***	***Mod***	***Max***
Fitted Model	Sph	Sph	Sph	Gau	Gau	Gau
Nugget	0.00	0.06	0.15	50	53	55
Partial sill	0.50	0.55	0.60	90	114	130
Range	800	900	1 000	0.20	0.24	0.50

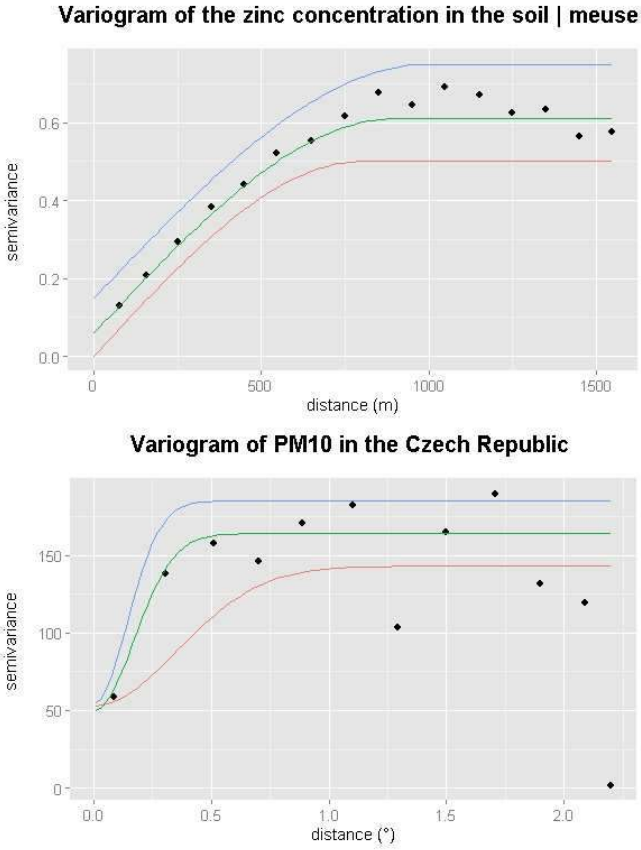


Fig. 3. Triplets of variograms of meuse data (top) and PM10 particles in the Czech Republic in 2013 (bottom). Minimal variograms represents a bottom boundary (red line), modal variograms (green line) is the optimal option and maximal variogram (blue line) bounds upper values

Three possible variograms are selected for each dataset in order to create fuzzy surfaces. Modal variogram is constructed as the most likely optimal variogram while minimal and maximal variograms provide bounds for possible realizations of variograms. All triplets of variograms are depicted in Figure

[1] Map of the stations network is available at http://goo.gl/59aEIV

[2] Updated tabular data reports from measuring stations for year 2013 are available at http://goo.gl/WjQcL9

3. Particular values of parameters for bot variograms are then shown in the Table I.

VIII. Optimization And Simulations

The fuzzy surface is constructed by following procedure. The modal value of the surface is calculated as kriging using the selected variogram model with modal values of nugget, sill and range. Then krigings for all the combinations of min and max values of sill, range and nugget are calculated to obtain the values of min and max of $\tilde{Z}$ at prediction locations. From these 8 (2^3) kriging surfaces the minimal and maximal values at each prediction location are selected to form the bounds of the fuzzy number $\tilde{Z}$.

Because the optimization procedure cannot guarantee that the resulting fuzzy surface contain all the possible values, it should be combined with probabilistic method to verify its completeness. Random realization of the surface is generated by calculating the kriging using random values of the kriging parameters selected from the range specified by their min and max value. Basically, this is an application of the Monte Carlo method.

Subsequently, each randomly generated surface is compared to the fuzzy surface whether the z value of the random surface lies outside of the interval specified by $[min, max]$ of $\tilde{Z}$. If that statement is true, then the value is considered as an "error". In this case, the error means that the value obtained from simulation violates the optimization scheme. It points to situations where the optimization scheme failed to predict bounds of the $\tilde{Z}$. However, not every such error is significant, as some of them could be smaller than the precision of the input data and thus they cannot be considered real errors.

IX. Results

During the optimization process, we carried out up to 5 000 simulations of possible variogram realizations for both datasets. Thresholds of simulations were given by predefined variograms. Each variogram simulation generated certain percentage of erroneous realizations that deviated from given thresholds. In fact, the amount of deviated realization allows the evaluation of method's usability. The overall statistical description of errors is provided in Table II. The distribution of errors is then depicted in Figure 4. Both results consist of three main parts. Firstly, the overall number (its percentage) of errors was evaluated, then systematic errors (given by the scale of the analysis) were calculated so the percentage of purely "real errors" was analysed, and lastly we computed the ratio of real to systematic errors, which describes the number of real errors falling on one systematic errors, i.e. the lower is the ratio the better is the optimization process. Systematic errors were defined as records, which values are under the original resolution of primary data (e.g. number of decimal places, scale of the data, etc.). Shapes of distributions of errors belonging to used datasets are not very similar. The PM10's errors distribution of probability is highly positively skewed and significantly leptokurtic, while the meuse datasets errors are slightly negatively skewed and rather platykurtic. The PM10 dataset also shows generally higher average values of both overall and real errors as well as the error ratio.

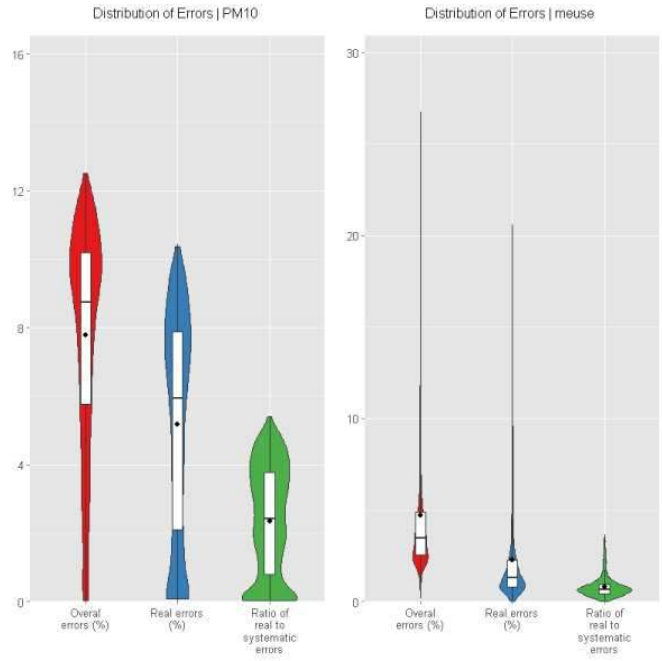


Fig. 4. Violin plots of erros appeared during the optimization process - PM10 (left part), meuse (right)

TABLE II. Descriptive statistics of errors appeared during the optimization process

[a]	Meuse			PM_{10}		
	OE (%)	***RE(%)***	***RRSE***	***OE (%)***	***RE(%)***	***RRSE***
Min	0.13	0.00	0.00	0.02	0.00	0.00
Max	26.91	20.59	3.64	12.52	10.44	5.41
Mean	4.69	2.26	0.78	7.78	5.17	2.33
Median	3.46	1.29	0.67	8.73	5.93	2.41
Std. deviation	3.90	2.94	0.57	3.15	3.20	1.61
IQR	2.39	1.48	0.50	4.45	5.80	2.98
Skewness	2.66	3.07	1.97	-0.81	-0.33	-0.05
Kurtosis	7.82	10.47	4.95	-0.33	-1.25	-1.31

[a.] OE represents overall percentage of errors appeared during the optimization process, RE is the percentage of real errors (i.e systematic errors are not included) and RRSE is the ratio of real to systematic errors

In the Fig. 5 the profile of the fuzzy surface is shown. It is clearly visible from the profile that the different variograms that compose the fuzzy variogram, actually model different relations among input data.

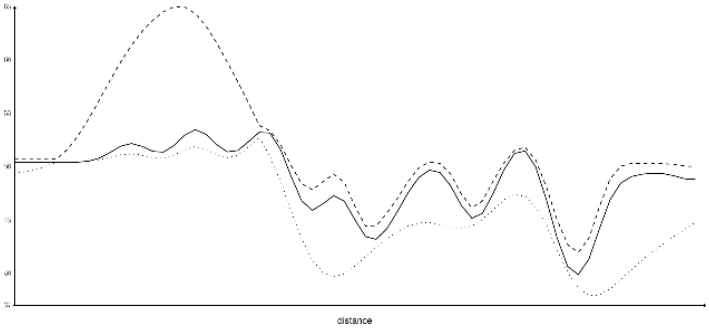


Fig. 5. Profile of the fuzzy surface of the PM_{10} in the Czech Republic. The profile is along 49°21´N. Full line shows modal value of the fuzzy surface, dotted line shows the minimal value and dashed line shows the maximal value.

So the fuzzy surface created by method proposed by [1,8] is actually containing several relations to its neighbouring points in comparison to classic kriging that only contain one specific relation between input points. By this way the uncertainty of the relationships between points in space is implemented into to the surface model.

Figure 6 is then showing three realizations of the kriging interpolation with usage of lower (minimal), upper (maximal) and optimal variogram for the PM_{10} dataset. In fact these three krigings represent optimal and bounding surfaces of resulting fuzzy surface, which profile is depicted in Fig. 5.

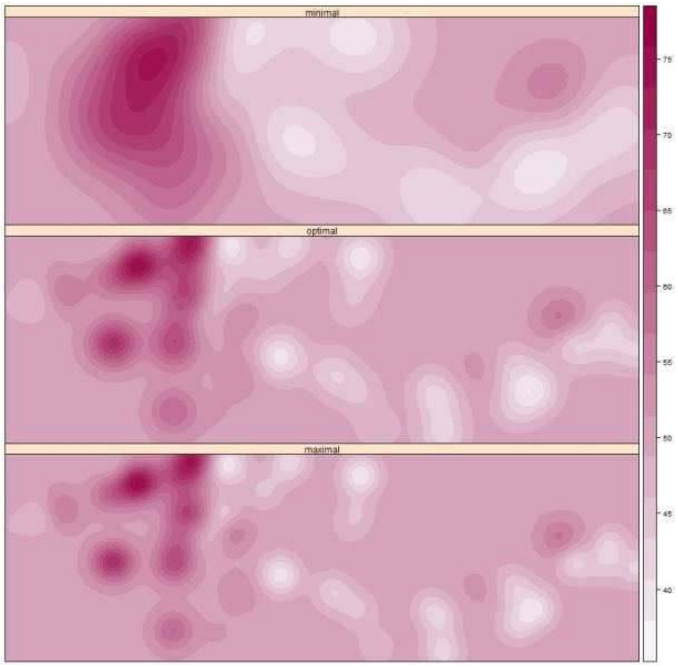


Fig. 6. The ordinary kriging interpolations of PM_{10} in the Czech Republic in 2013 based on triplets of variograms, minimal variogram (top), optimal (centre) and maximal variogram (bottom)

X. Discussion and Conclusions

According to the results show in Fig. 4 and Table II the optimisation scheme for the creation of a fuzzy surface can be viewed as a good instrument that is suitable for initial estimates and from our point of view, it should be also sufficient for most of other practical applications. Considering the fact that the fuzzy surface as created by this approach models user's uncertainty about variogram's parameters, then the percentage of "errors" (Tab. II) can be considered as rather small. The amount of calculation time that is saved is also notable. Given these facts, the optimisation scheme [8] can be thought as the useful tool for creating fuzzy surfaces. So far this approach [1,8] is the only one that is able to model epistemic uncertainty of the kriging parameters that is semantically valid [5,11].

Further research should be focused on using this approach for practical studies that would provide surfaces together with the uncertainty estimation. The use of such surfaces is crucial for decision making, because it allows the uncertainty of the surface to be propagated to the subsequent analysis.

Acknowledgement

The authors gratefully acknowledge the support by] the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

References

[1] A. Bardossy, I. Bogardi, W. E. Kelly, "Kriging with imprecise (fuzzy) variograms". Theory, Mathematical Geology, 1990, vol. 22, no. 1, pp. 63–79

[2] P. Diamond, "Fuzzy Kriging", Fuzzy Sets and Systems, 1989, vol. 33, no. 3, pp. 315–332.

[3] M. Hanss, "Applied Fuzzy Arithmetic: An Introduction with Engineering Applications", Berlin, Springer-Verlag, 2005.

[4] J. C. Helton, J. D. Johnson, W. L. Oberkampf, C. J. Sallaberry, "Sensitivity analysis in conjunction with evidence theory representations of epistemic uncertainty", Reliability Engineering & System Safety, 2006, vol. 91, no. 10-11, pp. 1414–1434.

[5] J. C. Helton, J. D. Johnson, "Quantification of margins and uncertainties: Alternative representations of epistemic uncertainty", Reliability Engineering & System Safety, 2011, vol. 96, no. 9, pp. 1034–1052.

[6] T. Hengl, "A Practical Guide to Geostatistical Mapping", Office for Official Publications of the European Communities, Luxembourg, 2009.

[7] A. Kaufmann, M. M. Gupta, "Introduction to Fuzzy Arithmetic", New York, Van Nostrand Reinhold Company, 1985.

[8] K. Loquin, D. Dubois. "Kriging with Ill-Known Variogram and Data", In: A. Deshpande, A. Hunter (eds), Scalable Uncertainty Management SE - 5, Berlin, Springer, 2010, p. 219–235.

[9] K. Loquin, D. Dubois. "Kriging and Epistemic Uncertainty: A Critical Discussion", In: R. Jeansoulin, O. Papini, H. Prade, S. Schockaert (eds), Methods for Handling Imperfect Spatial Information, Berlin, Springer, 2010, p. 269–305.

[10] K. Loquin, D. Dubois, "A fuzzy interval analysis approach to kriging with ill-known variogram and data", Soft Computing, 2011, vol. 16, no. 5, pp. 769–784.

[11] M. Oberguggenberger, "The mathematics of uncertainty: models, methods and interpretations", In: W. Fellin, H. Lessmann, M. Oberguggenberger, R. Vieider (eds), Analyzing Uncertainty in Civil Engineering, Berlin, Springer, 2005.

[12] E. J. Pebesma, "Multivariable geostatistics in S: the gstat package", Computers & Geosciences, 2004, vol. 30, p. 683-691.

[13] D. Temple Lang, "RCurl: General network (HTTP/FTP/...) client interface for R. R package version 1.95-4.1.", 2013, http://CRAN.R-project.org/package=RCurl

[14] D. Temple Lang, "XML: Tools for parsing and generating XML within R and S-Plus. R package version 3.98-1.1.", 2013, http://CRAN.R-project.org/package=XML

[15] L. A. Zadeh, "Concept of a Linguistic Variable and Its Application to Approximate Reasoning-I", Information Sciences, 1975, vol. 8, no. 3, pp. 199–249.

In order to ensure the reproducible research, authors decided to publish all the necessary data along with the used source code of analyses. They are available at *https://github.com/JanCaha/DailyMeteo2014-paper*.

Regression Kriging with GLM in Predicting Average Annual Precipitation in Serbia (1961-1990)

[full paper]

Milutin Pejović[1] Zagorka Gospavić[2], Branko Milovanović[3]
Faculty of Civil Engineering
Department for Geodesy and Geoinformatics
Belgrade, Serbia
[1]mpejovic@grf.bg.ac.rs; [2]zaga@grf.bg.ac.rs; [3]milovano@grf.bg.ac.rs

***Abstract*— GLM (Generalized Linear Model) is a widely used regression technique that represents the generalization of standard regression linear models. GLM is a flexible framework for modeling and the analysis of the variety of data coming from the exponential distribution family, which is often the case in experiments related to meteorological processes. In this study, GLM is used as a part of the hybrid interpolation technique, "Regression Kriging", to estimate the trends in average annual precipitation in Serbia for the period 1961 - 1990. R package GSIF, primarily designed for the modeling of soil phenomena, was used to carry out the whole estimation. In order to evaluate the quality of the estimation, besides standard diagnostic procedures, results were compared to the results obtained through multiple regression and standard regression Kriging.**

Keywords—GLM, Linear models; Regression kriging; precipitations.

I. INTRODUCTION

The application of geostatistical methods for the creation of precise meteorological maps is still completely unexplored. Besides many studies in which the possibility of geostatistical methods for climate mapping was investigated, new needs and techniques offer new possibilities for searching even better ways of creating better maps. Precipitation is a very complex phenomenon, with daily appearance depending on many external factors. This makes daily precipitation very difficult to model. However, on average scale, several external factors, like elevation, and the geographical location, are recognized as important in the geostatistical modeling of precipitations [1][2]. There are two similar studies that may be said to stand out for their comprehensive survey of comparison techniques for precipitation mapping. Both studies compare the interpolation techniques applied to annual and monthly rainfall data, which are of particular interest to this paper, and both sets of data are collected on relatively low density gauges network. First, the older study was done by Goovaerts, P. [3]. He showed that the stochastic technique, which takes into account spatial correlation (ordinary kriging), yields a more accurate prediction in comparison to the more simple deterministic techniques, and also in comparison to regression techniques, which take into account elevation as an external factor. But, none of the compared techniques takes into account both spatial correlation and external factors, at the same time. The second study, the newer one, is a work by Moral J.F. [4]. He also compared three main geostatistical techniques (ordinary kriging, simple kriging and universal kriging) with three more complex algorithms (collocated cokriging, simple kriging with varying local means, and regression kriging). These three techniques incorporate external factors in two different ways - deterministic (regression kriging) and stochastic (collocated cokriging). He showed that more complex methods yield more accurate results, but also need a more demanding analysis and computation, especially collocated cokriging. In their introduction chapters, these two works also give a comprehensive review of relevant references.

Over the last few years, an interest in the spatio-temporal climate analysis has increased significantly. For example, Hengl et. al. [5] have created a framework for the prediction of daily temperatures. They also used regression kriging, but the version enriched with temporal components (Spatio-Temporal Regression Kriging - STRK). Kilibarda. et.al [6], tested the performance of STRK for the prediction of mean, max. and min. temperatures on the area of the whole world, in the spatial resolution of one kilometer, and daily temporal resolution. The principle of STRK is the same, but its implementation requires a complex analysis of the spatio-temporal correlation of temperatures. The application of spatio-temporal interpolation requires the availability of predictor and response variables in both spatial and temporal sense, and for the whole area and period.

The increased availability of external (auxiliary) variables in raster format, favors regression kriging as the most appropriate interpolation technique. Also, there have been many studies showing superior performance of Regression kriging over other interpolation techniques [7] [4]. However, the most interesting advantage of RK is the possibility of the implementation of various regression techniques in the regression part of RK. All these techniques have the same role in RK - to create the trend of spatial phenomena as good as possible.

In this study, we use Generalized Linear Models (GLM) as the regression part of RK to interpolate average annual precipitations for the period 1961 -1990. The GLM is a well known and recommended regression technique. The conceptual framework of GLM allows the analysis and modeling of a wide range of phenomena [8]. GLM represents a generalization of ordinary linear regression that allows modeling wide range of data with error distribution other than normal. The first interpretation of GLM was given by Nelder and Wedderburn, (1972) [9]. A very detailed explanation of this method, applied to spatial data, is given in study by C. A. Gotway and W. W. Stroup, (1997) [10]. This exhaustive work covered the analysis of the performance of GLM in the analysis of spatially correlated treatments, and also in prediction.

The aim of this work is to explain what benefits the GLM brings to Regression Kriging, in order to produce a precise precipitation map. We used a simple linear model as the base for investigating what kind of consequences may occur in the results, if requirements related to a linear model are not exactly fulfilled, and how can GLM overcome them. The model involves three often used co-variables, DEM (digital elevation model) and location, expressed in easting and northing coordinates), that are linearly related to the values of precipitation observations.

II. MATERIALS AND METHODS

A. Motivation

A common procedure in the statistical modeling of any phenomenon starts with examining the distribution of data. Due to its linear formulation, the geostatistical techniques from the Kriging family give the best results, if the data are normaly distributed. However, in reality it is not so often. One approach to overcoming this problem is the usage of the regression kriging. Regression kriging usually combines multiple linear regression with simple kriging on residuals [7]. These residuals come from a model which is very strict about assumptions related to the linear model theory (normality, linearity, variance homogeneity and independency). Venables, B. et.al [11] pointed out that variance heterogeneity and non-normality could bring increased uncertainty into prediction, for points with extreme and unusual positions in the predictor space. For this purpose, a wide range of fitted diagnostic procedures has been developed [12]. If all assumptions are met, these residuals have to play the role of a stationary, spatially correlated and approximately normal distributed variable. The violation of anyone assumption can cause an unusual distribution of residuals, and also to make problems with the interpretation of model parameters. In order to overcome this problem, it is common to transform the response or the predictor variables. But doing this kind of transformation has several drawbacks. GLM allows analysis results, in sense of mean parameters, in the same scale as the measured response, unlike the transformed data, for which is recommended to stay in transformed scale [13]. Lane, P.W. (2002), in his work related to soil data [14], showed the main advantages and drawbacks of doing transformation.

GLM enables the modeling of mean by retaining the concept of additive explanatory effects, which can be expressed on a transformed scale and, at the same time cover a wide range of variance behavior by using distribution functions from the exponential family. Therefore, GLM is relaxed of strict assumptions related to ordinary linear models. But, the crucial issue of applying GLM is to choose appropriate error distribution and link functions, which defines the relation between mean and linear predictors. It assumes having some knowledge about the phenomena that is being investigated. The purpose of this work is to attempt to explore the performance of GLM in determining the trend of average precipitation data over a long period, as the base for doing the prediction by the means of the regression kriging. GSIF R package allows doing the prediction in an automated way, but the structure of the output data allows for the analysis of all results, obtained in each step. Such an easy way for creating maps through Regression kriging with GLM is the main motivation for doing this analysis.

B. Data and study area

We used the same dataset as in the study done by Bajat et.al. 2012. [15] It consists of two different datasets. The first one contains the rain gauge stations data (their spatial coordinates - northing, easting and altitudes, and associated average annual precipitation values as a target variable. The second set presents a publicly available digital elevation model (DEM) as an auxiliary variable. DEM is derived from reducing the ASTER model for the territory of Serbia to the 1km spatial resolution.

Spatial autocorrelation analysis, conducted on the target variable, in study [15], reveals significant clustering and spatial autocorrelation for the whole pattern of observation points. Overlaping of autorcorrelated clusters and topographic units also justified the usage of DEM as a covariable.

C. Regression Kriging and GLM

Regression Kriging combines two conceptually different types of techniques - deterministic regression and stochastic kriging technique. If the measured values of the target variable are given as $Z(s_i)$, $i=1\ldots n$, where s_i represents spatial location and n the number of realized measurements, then the system of equations, for estimating values of target variables $\hat{Z}(s_0)$ is:

$$\hat{Z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \tag{1}$$

$$\hat{Z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} w_i(s_0) \cdot e(s_i); \quad q_0(s_0) = 1 \tag{2}$$

where $\hat{m}(s_0)$ is the fitted deterministic part, $\hat{e}(s_0)$ is the interpolated residual, $\hat{\beta}_k$ are estimated deterministic model coefficients, $q_k(s_0)$ the values of predictors in point s_0, w_i are ordinary Kriging weights resolved by the spatial structure of residuals $e(s_i)$ [7].

Deterministic part $\hat{m}(s)$ is linear regression model (LM), which can be both simple and multiple. The easiest formulation of that model is:

$$E\left[z_{(s)} \right] \equiv \hat{m}(s) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s), \qquad z_{(s)} \sim N(\hat{m}_{(s)}, \sigma^2) \tag{3}$$

Regression coefficients $\hat{\beta}_k$ could be obtained through different fitting methods, like ordinary least squares (OLS) or generalized least squares (GLS).

Deterministic parts can also be estimated within a GLM framework. By comparing it with LM, GLM model can be written as:

$$E\left[z_{(s)} \right] \equiv \hat{m}(s) = \gamma(\sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s)), \qquad z_{(s)} \sim \textit{Exponetial family distribution} \tag{4}$$

where γ_i is any smooth monotonic function, called the mean function. Generalized linear models extend linear models by allowing non-linearity in the model structure (trough to mean function), and much more flexibility in the specification of the distribution of the response variable. An important feature of the Exponential distribution family is the variance function, which is, for some distributions, a function of the mean and one more parameter (scale parameter): $\text{var}(y_i) = V(\mu_i)\phi$. The inference of GLM is based on the theory of maximum likelihood estimation. The most likely values for the parameters are those that cause the likelihood to be as large as possible. In the majority of GLM applications, the likelihood is maximized by iteratively re-weighted least squares (IRLS) algorithm.

Model diagnostic for GLM is performed by using the residuals, in the same way as for linear models [13]. However, in the regression kriging framework, the prediction is based on response (raw) residuals. In case of the Gaussian family and log link function, these two sets of residuals are equal, and comparable between a standard linear model (LM) and GLM, because $\text{var}(y_i) = 1$.

In order to make a comparison between these two trend models, and to make an inference about the influence of violated assumptions in the linear model, several diagnostic tools and measures of accuracy were used. Residuals plot with spread level [13] plot were used in order to investigate the fulfillment assumptions of the linear model. R^2 (coefficient of determination) and RMSE (residual mean squared error) were used to compare the global accuracy. The estimated procedure for R^2 in GLM framework is slightly different from that in OLS (ordinary least squared) used in the standard linear model, but the logic is the same (for GLM R^2 is based on the relation of residual and null deviance: $R^2 = 1 - \frac{D_1}{D_0}$).

The effects of the unusual observation are investigated by examining hat values and the relation between the hat values [13] and the residual differences of the two models.

D. Software

All computations are carried out by R software and related packages. R is the system for statistical computation and graphics, which provides, among other things, programming facilities, high-level graphics, interfaces to other languages, and debugging facilities [16]. R is free and open source software, under the terms of the GNU General Public License. R is organized as a collection of packages designated for specific tasks.

All diagnostic methods used for the fitting of linear models are implemented in "car" package. With the "car" package, the diagnostics of methods, standard linear models and GLM is made possible. Analyzing and modeling the data from a spatial point of view has been done in several well tested packages for spatial analysis: sp, gstat, spatstat, GSIF.

III. RESULTS AND DISCUSSION

The linear model created to define the relation between the predictors and a response variable is the same in LM and GLM. Such a linear model (linear predictor) consists of average annual precipitation, as response variables and three predictor variables: Easting, Northing and Elevation. These three predictors represent a standard and a proven, useful set for the modeling of precipitations. According to the right skew of the response variable, that simple log transformation is chosen for the linear model , which could be useful for making the data normal, and also for making the error variance more stable. In GLM framework, these data characteristics are used in the decision-making about error distribution and link function. Due to the nature of the average data and the rightly skewed distributions, we are accustomed to using the "Gaussian" family and the "log" link function.

The statistically significant influence of all predictors is shown (for all predictors) in each model (table 1).

In order to examine whether the the requirements related to standard linear model assumptions were fulfilled the residual plot was created (Figure 1). If all assumptions were met, all these graphs should have a pattern with no particular trend. The presence of moderate systematic features in these patterns indicates the violation of linearity and homoscedasticity assumptions. These shortcomings can be fixed by applying some of the transformations to the predictors. The aim of this study is exactly to examine how these moderate violations of requirements affect the estimation, in comparison to GLM.

The estimates of the coeficient of determination R^2 and RMSE (residual mean squared error) favor GLM as a better solution. Small advantage, in the sense of RMSE, goes to GLM, but R^2 seems to be of a significantly different accuracy, again in favor of GLM.

Table 1 – Summary statistics for two models. (Modified output from Stargazer R package [18])

Response	log(mmperm2) OLS	(mmperm2) glm: gaussian, link = log
East	-4.229e-07***	-5.316e-07***
Std. Error	-5.86E-08	-5.98E-08
North	-1.645e-07***	-1.204e-07**
Std. Error	-4.88E-08	-4.99E-08
H	3.077e-04***	3.280e-04***
Std. Error	-1.13E-05	-8.68E-06
Constant	7.453***	7.289***
Std. Error	-0.256	-0.26
Observations	1,014	1,014
R2	0.59	0.66
Adjusted R2	0.591	
RMSE	95.48	94.11

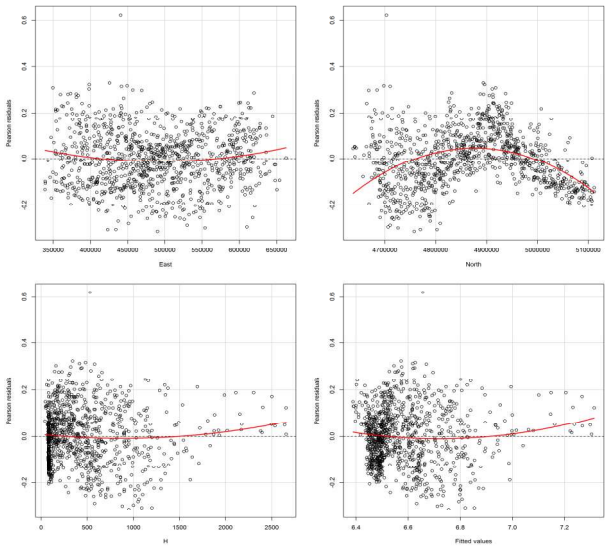


Fig. 1. *Residual plot of ordinary linear model.*

Spatial distribution of residuals from the two models looks very similar. The spatial correlation of residuals could be modeled by almost the same variogram (table 2).

Table 1 – Estimated parameters of fitted variogram models of residuals from two model.

LM residual variogram			GLM residual variogram		
model	psill	range	**model**	psill	range
Nug	2,296.90	0	**Nug**	2,339.90	0
Exp	8,031.69	68,500.67	**Exp**	7,993.86	71,535.98

The Residual difference map (Figure 2) reveals interesting patterns. This map shows that LM has made larger residuals on the points located in the south-west part of Serbia. Considering that it is mainly a mountainous area, and also the area of extreme observation values, it can be concluded that these two models produce the greatest difference in areas with extreme values of observations and predictors (Figure 2). To prove this statement, a map of hat-values of observations has been created.

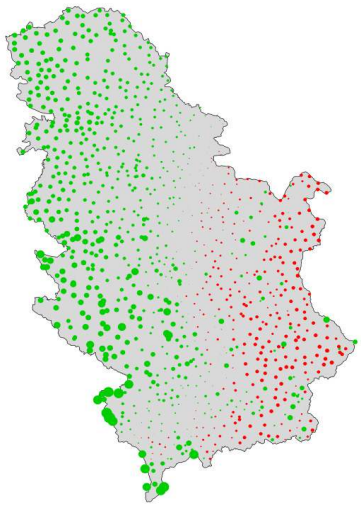


Fig. 2. *Residual difference map*

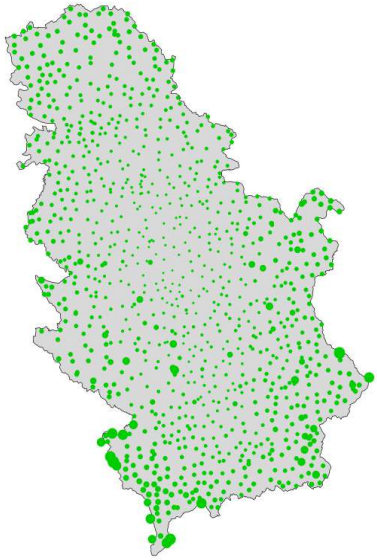


Fig. 3. *Map of hat-values of observation for linear model*

Spatial distribution of hat values (Figure 3) coincides with the spatial distribution of residual differences. This is also confirmed by the coefficient of correlation, which shows a moderate positive correlation (r=0.6). The obtained results indicate that the observations which are relatively far from the center of the predictor space, produce a different influence on these two models.

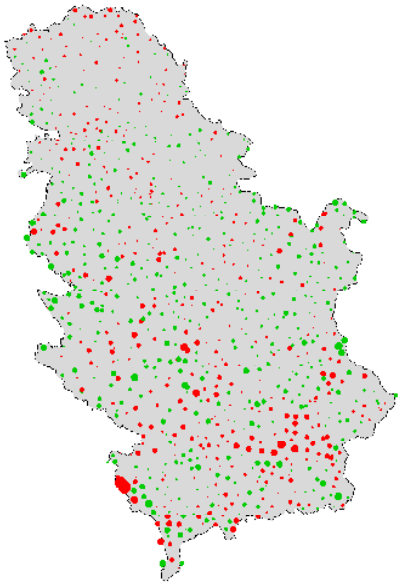


Fig. 4. *Kriging residuals difference*

Obtained results show that these two models provide different input data for Kriging interpolation. Kriging interpolation, applied to obtained residuals, with almost the same variograms, has almost completely eliminated the existing differences between two regression models. On the obtained Kriging residuals difference map (Figure 4), no systematic pattern can be recognized. Both models gave a very high and similar estimation coefficient of determination - R^2 ($R^2_{GLM} = 0,86$; $R^2_{LM} = 0,84$ obtained from the residuals estimated in the process of cross-validation). These results imply a very good performance of Regression kriging, used independently from the regression methods. The final result, expressed in the map of predicted average precipitation, has remained resistant to the moderate violation of requirements related to standard linear models.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science of the Republic of Serbia (Contracts No. TR 36009).

.

REFERENCES

[1] Diodato, N. 'The influence of topographic co-variables on the spatial variability of precipitation over small regions of complex terrain', *International Journal of Climatology* **25**, 351-363, 2005.

[2] Lloyd, C. 'Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain', *Journal of Hydrology* **308**, 128-150, 2005.

[3] Goovaerts, P. 'Geostatistical approaches for in corporating elevation into the spatial interpolation of rainfall', *Journal of Hidrology* **228**, 113-129, 2000.

[4] Moral, J. F. (2010), 'Comparison of different geostatistical approaches to map climate variables: application to precipitation', *International Journal of Climatology* **30**, 620-631.

[5] Hengl, T.; Heuvelink, G. B. M.; Perčec Tadić, M. & Pebesma, E. J. (2012), 'Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images', *Theoretical and Applied Climatology* **107**, 265-277.

[6] Kilibarda, M.; Hengl, T.; Heuvelink, G. B. M.; Gräler, B.; Pebesma, E.; Perčec Tadić, M. & Bajat, B. (2014), 'Spatio-temporal interpolation of daily temperatures for global land areas at 1km resolution', *Journal of Geophysical Research: Atmospheres* **119**, 2294–2313.

[7] Hengl, T.; Heuvelink, G. B. M. & Rossiter, D. M. 'About regression-kriging: From equations to case studies', *Computers & Geosciences* **33**, 1301–1315, 2007.

[8] Dobson, A. J. *An Introduction to Generalized Linear Models*, Chapman & Hall/CRC, 2002.

[9] Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society* **135**, 370-384.

[10] Gotway, C. A. & Stroup, W. W. 'A Generalized Linear Model Approach to Spatial Data Analysis and Prediction', *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 157-178, 1997.

[11] Venables, W. & C.M., D. 'GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research', *Fisheries Research* **70**, 319–337, 2004.

[12] Fox, J. & Weisberg, S. *An R Companion to Applied Regression*, SAGE Publications, Inc., 2011.

[13] Lindsey, J. & Jones, B. 'Choosing among generalized linear models applied to medical data', *Statistics in Medicine* **17**, 59-69, 1998.

[14] Lane, P. 'Generalized linear models in soil science', *European Journal of Soil Science* **53**, 241-251, 2002.

[15] Bajat, B.; Pejović, M.; Luković, J.; Manojlović, P.; Ducić, V. & Mustafić, S. 'Mapping average annual precipitation in Serbia (1961–1990) by using regression kriging', *Theoretical and Applied Climatology*, 1—13, 2012.

[16] R Development Core Team. 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria. 2012.

[17] Hlavac, Marek 'stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables'. R package version 5.0. http://CRAN.R-project.org/package=stargazer ,2014.

Modeling Resilience to Climate Change in Space and Time

[full paper]

Slobodan P. Simonovic

Professor, Department of Civil and Environmental Engineering
Director of Engineering Works, Institute for Catastrophic Loss Reduction
University of Western Ontario, London, Ontario, Canada

There are practical links between flood risk management, climate change adaptation and sustainable development leading to reduction of flood risk and re-enforcing resilience as a new development paradigm. There has been a noticeable change in flood management approaches, moving from disaster vulnerability to disaster resilience; the latter viewed as a more proactive and positive expression of community engagement with flood risk management. As flood hazard is increasing, at the same time it erodes resilience, therefore climate change has a magnifying effect on the flood risk. In the past, standard disaster management planning emphasized the documentation of roles, responsibilities and procedures. Increasingly, these plans consider arrangements for prevention, mitigation, preparedness and recovery, as well as response. However, over the last ten years substantial progress has been made in establishing the role of resilience in sustainable development. Multiple case studies around the world reveal links between attributes of resilience and the capacity of complex systems to absorb disturbance while still being able to maintain a certain level of functioning. Building on emergency planning experience, there is a need to focus more on action-based resilience planning to strengthen local capacity and capability, with greater emphasis on community engagement and a better understanding of the diversity, needs, strengths and vulnerabilities within communities. Floods do not impact everyone in the same way. It is clear that the problems associated with sustainable human wellbeing in calls for a paradigm shift. Use of resilience as an appropriate matrix for investigation arises from the integral consideration of overlap between: (a) physical environment (built and natural); (b) social dynamics; (c) metabolic flows; and (d) governance networks. This paper provides an original systems framework for quantification of resilience. The framework is based on the definition of resilience as the ability of physical and social systems to absorb disturbance while still being able to continue functioning. The disturbance depends on spatial and temporal perspectives and direct interaction between impacts of disturbance (social, health, economic, and other) and adaptive capacity of the system to absorb disturbance.

I. INTRODUCTION

The terms 'floods', 'flooding', 'flood hazard' and 'flood risk' cover a very broad range of phenomena [1]. Among many definitions of floods that do not incorporate only notions of *inundation* and *flood damage* for the purpose of this paper I will stay with the definition provided by [2] that a flood is a body of water which rises to overflow land which is not normally submerged. This definition explicitly includes all types of surface inundation but flood damage is addressed only implicitly in its final three words. Both, inundation and damage occur on the great range of scale.

The term such as 'flood risk' and 'flood losses' are essentially our interpretations of the negative economic and social consequences of natural events. Human judgment is subject to value systems that different groups of people may have and therefore these terms may be subject to different definitions. The flood risk, at various locations, may increase by human activity – like inappropriate land use practices. Also, the flood risk may be reduced by flood management structures and/or effective emergency planning. The real flood risk therefore, stems from the likelihood that a major hazardous event will occur unexpectedly and that it will impact negatively on people and their welfare [3]. Flood hazards result from a combination of physical exposure and human vulnerability to flooding. Physical exposure reflects the type of flood event that can occur, and its statistical pattern, at a particular location. The human vulnerability reflects key socio-economic factors such as the number of people at risk on the floodplain, the extent of flood defense works and the ability of the population to anticipate and cope with flooding. In this paper the formal definition of *flood risk* is a combination of the chance of a particular event, with the impact that the event would cause if it occurred. Flood risk therefore has two components – the chance (or probability) of an event occurring and the impact (or consequence) associated with that event. The consequence of an event may be either

desirable or undesirable. A convenient single measure of the importance of a flood risk is given by:

$$Risk = Probability \times Consequence \quad (1)$$

If any of the two elements in (1) increases or decreases, then risk increases or decreases respectively.

How we manage flood risk? In many countries flood risk management is evolving from traditional approaches based on design standards to the development of risk-based decision-making, which involves taking account of a range of loads, defense system responses and impacts of flooding [4]. The difference between a risk-based approach and other approaches to design or decision making, is that it deals with outcomes. The World Meteorological Organization is promoting the principal of integrated flood management - IFM – [5] that has been practiced at many places for decades. An Integrated Flood Management plan should address the following six key elements: (i) Manage the water cycle as a whole; (ii) Integrate land and water management; (iii) Manage risk and uncertainty; (iv) Adopt a best mix of strategies; (v) Ensure a participatory approach; and (vi) Adopt integrated hazard management approaches.

Flood risk management is a part of all social and environmental processes aimed at minimizing loss of life, injury and/or material damage. [6] and [7],[1] advocate systems view of flood risk management processes in order to address their complexities, dynamic character and interdisciplinary needs of management options. A primary emphasis of systems analysis in flood risk management is on providing an improved basis for effective decision-making. A large number of systems tools, from simulation and optimization to multi-objective analysis, are available for formulating, analyzing and solving flood risk management problems.

In order to apply a continuous improvement approach to flood risk management it is essential to have a way or thinking – a model – of what is being managed. The system in our focus is a social system. It describes the way floods affect people. The purpose of describing the system is to help clarify the understanding and determine best points of systems intervention.

The flood risk management system comprises four linked subsystems: individuals, organizations and society, nested within the environment. Individuals are the actors that drive organizations and society to behave in the way they do. They are decision makers in their own right, with a direct role in mitigation, preparedness, response and recovery from flooding. Organizations are the mechanism people use to produce outcomes that individuals cannot produce. Organizations are structured to achieve goals. Structure defines information and/or resource flows and determines the behavior of the organization. The concept of society is different from those of individuals and organizations, being more difficult to put boundaries around. In general, society itself is a system of which individuals and organizations are subsets and contains the relationships people have with one another, the norms of behavior and the mechanisms that are used to regulate behavior. The environment includes concrete elements such as water and air, raw materials, natural systems, etc. It also encompasses the universe of ideas, including the concept of 'future'. This concept is important in considering flood risk management - it is the expectation of future damages and future impacts that drives concern for sustainable management of flood disasters. Six management principles are presented by [1].

A change to proactive flood risk management requires an identification of the risk, the development of strategies to reduce that risk, and the creation of policies and programs to put these strategies into effect. Flood risk management is a part of all social and environmental processes aimed at minimizing loss of life, injury and/or material damage. A systems view of flood risk management is recommended in order to address the complexity, dynamic character and interdisciplinary needs of management options. A primary emphasis of systems analysis in flood risk management is on providing an improved basis for effective decision-making. A large number of systems tools, from simulation and optimization to multi-objective analysis, are available for formulating, analyzing and solving flood risk management problems. The main objective of this book is to present a variety of systems tools for flood risk management.

Recognizing the progress in flood risk management and recognizing the needs of vulnerable communities, the United Nations and its partners at the World Conference on Disaster Reduction (WCDR) in Kobe City in January 2005, came up the "Hyogo Framework for Action 2005-2015: Building the Resilience of Nations and Communities to Disasters". This was the introduction of resilience thinking as a replacement for flood risk management. Governments around the world have traditionally planned large-scale, centralized infrastructure flood protection systems that aim to control variables and reduce uncertainties. There is growing awareness that a transition toward sustainable alternatives is necessary if systems are to meet society's future water needs in the context of drivers such as climate change and variability, demographic changes, environmental degradation, and resource scarcity. However, there is minimal understanding of how to transition from flood risk management to building flood resilience and how to operationalize resilience thinking as one component of strategic planning for such change to facilitate the transition to a sustainable water future [8].

II. RESILIENCE QUANTIFICATION FRAMEWORK

There are many definitions of resilience [9], from general: (i) The ability to recover quickly from illness, change or misfortune; (ii) Buoyancy; (iii) The property of material to assume its original shape after deformation; (iv) Elasticity; to

ecology–based [10]: (i) The ability of a system to withstand stresses of 'environmental loading'; to hazard–based [11]: (i) Capacity for collective action in response to extreme events; (ii) The capacity of a system, community, or society potentially exposed to hazards to adapt, by resisting or changing, in order to reach and maintain an acceptable level of functioning and structure; (iii) The capacity to absorb shocks while maintaining function; (iv) The capacity to adapt existing resources and skills to new situations and operating conditions. The common elements of these definitions include: (i) minimization of losses, damages and community disruption; (ii) maximization of the ability and capacity to adapt and adjust when there are shocks to systems; (iii) returning systems to a functioning state as quickly as possible; (iv) recognition that resilient systems are dynamic in time and space; and (v) acknowledgements that post-shock functioning levels may not be the same as pre-shock levels.

Resilience is a dynamic process, but for measurement purposes is often viewed as static phenomena [12]. In this paper a flood resilient community is a sustainable network of physical (constructed and natural) systems and human communities (social and institutional) that possess the capacity to survive, cope, recover, learn and transform from flood events by: (i) reducing failure probabilities; (ii) reducing failure consequences (for example material damage); (iii) reducing time to recovery; and (iv) creating opportunity for development and innovation from adverse impacts. Numerous institutions, organizations, and elements in the urban environment contribute to community flood resilience, for example water and power lifelines, acute-care hospitals, and organizations that have the responsibility for emergency management. Improving the resilience of critical lifelines is critical for overall community resilience. These organizations are essential for community functioning; they enable communities to respond, provide for the well-being of their residents, and initiate recovery activities when disasters strike [13]. For example, since no community can cope adequately with a flood disaster without being able to provide emergency care for injured victims, hospital functionality is crucial for community resilience. Water is another essential lifeline service that must be provided to sustain disaster victims.

The quantification framework recommended by [9] following [12] has two qualities: inherent (functions well during non-flooding periods); and adaptive (flexibility in response during flood events) and can be applied to physical environment (built and natural), social systems, governance network (institutions and organizations), and economic systems (metabolic flows). An original space-time dynamic resilience measure (STDRM) of Simonovic and Peck is designed to capture the relationships between the main components of resilience; one that is theoretically grounded in systems approach, open to empirical testing, and one that can be applied to address real-world problems in various communities.

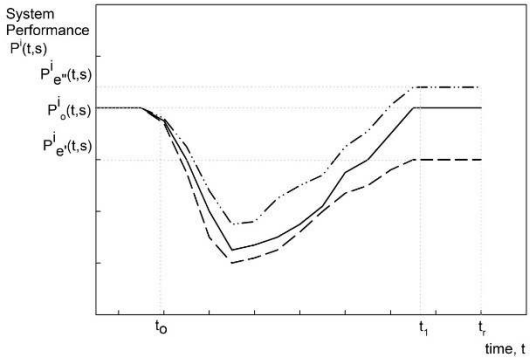


Fig. 1. System performance.

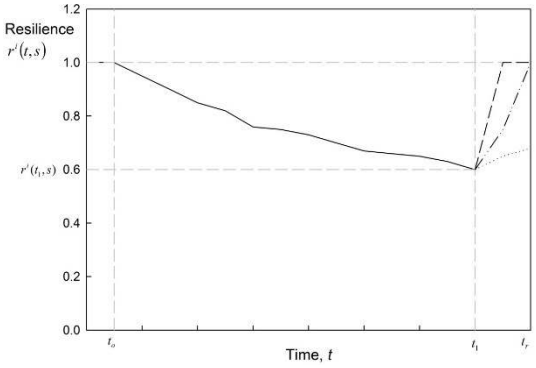


Fig. 2. System resilience

STDRM is based on two basic concepts: level of system performance and adaptive capacity. They together define resilience. The level of system performance integrates various impacts (i) of flood on a community. The following impacts (units of resilience (ρ^i)) can be considered: physical, health, economic, social and organizational, but the general measure is not limited to them. Measure of system performance $P^i(t,s)$ for each impact (i) is expressed in the impact units (physical impact may include for example length [km] of road being inundated; health impact may be measured using an integral index like disability adjusted life year (DALY); and so on). This approach is based on the notion that an impact, $P^i(t,s)$, which varies with time and location in space, defines a particular resilience component of a community, see Figure 1 adapted from [9]. The area between the initial performance line $P_0^i(t,s)$ and performance line $P^i(t,s)$ represents the loss of system resilience, and the area under the performance line $P^i(t,s)$ represent the system resilience ($\rho^i(t,s)$). In Figure 1, t_0 denotes the beginning of the flood event, t_1 the end, and t_r the end of the flood recovery period.

In mathematical form the loss of resilience for impacts (i) represents the area under the performance graph between the beginning of the system disruption event at time (t_0) and the

end of the disruption recovery process at time (t_r). Changes in system performance can be represented mathematically as:

$$\rho^i(t,s) = \int_{t_o}^{t}\left[P_0^i - P^i(\tau,s)\right]d\tau \qquad (2)$$

Where $t \in [t_0, t_r]$

When performance does not deteriorate due to disruption, $P_o^i(t,s) = P^i(t,s)$ the loss of resilience is 0 (i.e. the system is in the same state as at the beginning of disruption). When all of system performance is lost, $P^i(t,s) = 0$, the loss of resilience is at the maximum value. The system resilience, $r^i(t,s)$ is calculated as follows:

$$r^i(t,s) = 1 - \left(\frac{\rho^i(t,s)}{P_o^i \times (t - t_o)}\right) \qquad (3)$$

As illustrated in Figure 1, performance of a system which is subject to a flood (disaster event) drops below the initial value and time is required to recover the loss of system performance. Disturbance to a system causes a drop in system resilience from value of 1 at t_o to some value $r^i(t_1,s)$ at time t_1, see Figure 2. Recovery usually requires longer time than the duration of disturbance. Ideally resilience value should return to a value of 1 at the end of the recovery period, t_r (dashed line in Figure 2); and the faster the recovery, the better. The integral STDRM (over all impacts (*i*)) is calculated using:

$$R(t,s) = \left\{\prod_{i=1}^{M} r^i(t,s)\right\}^{\frac{1}{M}} \qquad (4)$$

Where, M is the total number of impacts.

The calculation of STDRM for each impact (i) is done at each location (s) by solving the following differential equation:

$$\frac{\partial \rho^i(t)}{\partial t} = AC^i(t) - P^i(t) \qquad (5)$$

Where, AC^i represents adaptive capacity with respect to impact i.

The STDRM integrates resilience types, dimensions and properties by solving for each point in space (*s*):

$$\frac{\partial R(t)}{\partial t} = AC(t) - \prod_i P^i(t) \qquad (6)$$

The implementation of the presented framework is proceeding by using system dynamics simulation approach together with spatial analysis software [14], [15] in the form of Coastal Megacity Resilience Simulator (CMRS).

III. APPLICATION

The presented resilience framework is being implemented on large cities in low-lying deltaic environments (Vancouver, Canada; Manila, Philippines; Lagos, Nigeria; and Bangkok, Thailand) selected for consideration under the project "Coastal Cities at Risk: Building Adaptive Capacity for Managing Climate Change in Coastal Megacities" supported by the International Research Initiative on Adaptation to Climate Change of the Canadian International Development Research Centre [16].

In this paper some basic information is provided for the implementation in Vancouver, Canada. Vancouver is a coastal megacity and can be considered as a network of three interdependent subsystems: (i) the natural subsystem; (ii) the socio-economic subsystem; and (iii) the administrative and institutional subsystem. Each of the three subsystems is characterized by its own elements and is surrounded by its own environment. For the purpose of the project, coastal megacity resilience is caused by the interaction between society and climate change caused hazards (project focus is on precipitation, floods and seal level rise).

The five major impacts that are being considered in the STDRM include: physical impacts, economic impacts, social impacts, health impacts and organizational impacts. They are being individually modeled for Vancouver in order to describe the local conditions.

The Coastal Megacity Resilience Simulator (CMRS) is data intensive. A very detailed description of each of the five impacts considered within the tool and detailed temporal and spatial scales require serious data support [9].

ACKNOWLEDGMENT

The author is thankful for the research financial support provided by IDRC and research assistance provided by Ms. Angela Peck, a PhD candidate at the University of Western Ontario.

REFERENCES

[1] Simonovic, S.P. (2012). *Floods in a Changing Climate – Risk Management,* Cambridge University Press, Cambridge.

[2] Ward, R.C. (1978). *Floods: A Geographical Perspective*, Macmillan, London.

[3] Smith, K., and R. Ward (1998). *Floods: Physical Processes and Human Impacts,* John Wiley & Sons, New York.

[4] Sayers,P.B., J.W. Hall and I. C. Meadowcroft, (2002). "Towards risk-based flood hazard management in the UK", *Civil Engineering*, Proceedings of ICE, 150:36–42, Paper 12803.

[5] WMO (2009). *Integrated Flood Management: Concept Paper*, WMO-No.1047, pp.32, Geneva.

[6] Mileti, D.S. (1999). *Disasters by Design*, Joseph Henry Press, Washington,USA.

[7] Simonovic, S.P., (2011). *Systems Approach to Management of Disasters: Methods and Applications.* John Wiley & Sons Inc., Hoboken,New Jersey, USA.

[8] Ferguson, B. C., R. R. Brown, and A. Deletic, (2013). A diagnostic procedure for transformative change based on transitions, resilience, and institutional thinking. Ecology and Society 18(4): 57..

[9] Simonovic, S.P., and A. Peck, (2013). "Dynamic Resilience to Climate Change Caused Natural Disasters in Coastal Megacities: Quantification Framework ", *British Journal of Environment & Climate Change,* **3**(3): 378-401.

[10] Gunderson, L.H, and C.S. Holling, editors, (2001). Panarchy: understanding transformation in human and natural systems. Island Press, Washington.

[11] UNISDR, United Nations International Strategy for Disaster Reduction, (2014). UNISDR Terminology on Disaster Risk Reduction. Geneva: United Nations; Last accessed May 2, 2014, http://www.unisdr.org/we/inform/terminology.

[12] Cutter, S.L., I. Barnes, M. Berry, C. Burton, E. Evans, and E. Tate, (2008). "A place-based model for understanding community resilience to natural disasters". Global Environmental Change, 18:598-606.

[13] Bruneau, M., S.E. Chang, R.T. Eguchi, G.C. Lee, T.D. O'Rourke, and A.M. Reinhorn, (2003). "A framework to quantitatively assess and enhance the seismic resilience of communities". Earthquake Spectra, 19(4):733-752.

[14] Srivastav, R.K. and S.P. Simonovic, (2014). "Generic Framework for Computation of Spatial Dynamic Resilience". Water Resources Research Report no. 085, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 82 pages. In print.

[15] Peck, A. and S.P. Simonovic, (2014). "Spatial System Dynamics (SSD): Coupling System Dynamics and Geographic Information Systems", Water Resources Research Report no. 086, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 42 pages. In print.

[16] IDRC, International Development Research Centre, (2014). Accessed May 3, 2014. Available: http://coastalcitiesatrisk.org/wordpress/.

Local Meteorological Simulation to Define Critical Areas for Agricultural Production

[abstact]

Boris Mifka
Ecological Department
Teaching Institute of Public Health
NZZJZ
Rijeka, Croatia

Dr. Maja Žuvela Aloise
Section Climate Variability and Modeling
Central Institution for Meteorology and Geodynamics
ZAMG
Vienna, Austria

Agricultural production in semi-arid Mediterranean regions is extremely sensitive to impacts of global climate change, particularly to possible increase in excessive heat and drought events. Assessment of climatic trends needed for development of long-term planning strategies in coastal regions with complex topography, such as the mid-Adriatic islands, poses a problem since these areas are too small to be distinguished by the regional climate models. The purpose of this project was to test a novel method to obtain fine-scale climatic information using the local-scale climate model MUKLIMO_3 developed by the German Weather Service (DWD). We performed numerical simulations of atmospheric conditions for potential days with excessive heat load for three mid-Adriatic islands: Brac, Hvar and Korcula. Additionally,idealized simulations according to typical local winds relevant for these islands are considered. The model uses a grid with a horizontal resolution up to 100 m and CORINE land use data and ASTER GDEM orography as input. In combination with the so-called "cuboid method", it was possible to calculate climatological indices such as mean yearly number of summer days based on observational data from the last 30-year climatic period from six local meteorological stations. Model results were compared to observational values for climatic indices from local meteorological stations which gave good agreement of more than 84% at most of the stations and even 97.5 % at three of them. In order to make a better validation of the model which would account gradients of temperature and relative humidity of air according to spatial changes due to topography and land use, a measurement campaign was organised at islands Hvar and Korčula from 7/8/2013 till 14/8/2013. The validation and improvement of the model's results is still in process. Climatological charts of heat load during summer period are used to reveal critical areas for agriculture and wildland fires.

Keyword; microclimate modelling, semi-arid regions, complex topography, land-use, MUKLIMO_3, cuboid method, climate indices, mobile measurements, wildland fires.

Mapping of Maximum Snow Load Values for the 50-Year Return Period for Croatia

[full paper]

Melita Perčec Tadić, Ksenija Zaninović, Renata Sokol Jurković
Department for climatological research and applied climatology
Meteorological and Hydrological Service
Croatia
melita.percec.tadic@cirus.dhz.hr

***Abstract*—Snow load is defined as a product of snow density, snow depth and gravitational acceleration. It is an important climatic element that is, together with minimum and maximum temperatures and wind load, a part of the national annex for the application of standards for the design of structures. In particular, the characteristic snow load, defined as the maximum snow load at the ground for the 50-year return period, has to be estimated and a map has to be supplied as a part of national annex. The method for the estimation of the characteristic snow load at the station locations is presented, followed by the geostatistical mapping procedure for estimating this parameter for the whole Croatian territory.**

I. INTRODUCTION

The Croatian Standards Institute publishes standards for the design of structures with national annexes. One of these contains characteristic snow load, s_k defined as the maximum snow load values for the 50-year return period. The existing standards for structural design from year 2000 have to be updated according to the European standard EN 1991-1-3:2003, Eurocod 1: Actions on structures – Part 1-3: General actions – Snow load, with the new s_k map in the national annex. The details of the mapping procedure of the s_k together with preparation and description of snow depth, snow density and snow load data are presented in this article. Careful estimation of the snow load values is important in order to avoid both unnecessary construction costs but also the risk of structure failure [13, 9].

Depending on the measuring tool, snow density is calculated from snow depth and snow weight or from snow water content. Measuring snow density nor snow water content is not standard on meteorological stations so it is less frequently performed [7, 21, 20]. Reference [21] estimates that measuring snow water equivalent takes ~20 times as long as measuring depth since it is recommended to repeat the procedure few times to ensure the best accuracy. They convert large number of snow depth measurement into snow water equivalent to be able to access worldwide snow water resources. The snow density model they developed used day of the year, snow depth and climate classes to predict snow density. Simple, but inadequate rule of 1:10 ratio, reflecting an assumed snow density of 100 kgm^{-3}, has been used in USA and Canada to estimate density. Reference [20] proposed the neural networks algorithm that predicts density based on monthly solar radiation, vertical profiles of temperature and humidity and external compaction (surface wind speed and liquid equivalent precipitation) to forecast the snow dept from the quantitative precipitation forecast. Some European countries do not even measure the snow depth which makes the snow load estimation even more challenging. Is Spain [9], they assume that the total precipitation is in the form of snow (neglecting the sleet) for the days with mean daily temperatures $\leq 0°C$. This can be used directly to calculate the snow load as a product of precipitation, water density and gravitational acceleration, avoiding the snow density problem.

The work that was intended to produce a common scientific basis which can be accepted by all European countries involved in the drafting of Eurocodes has been published in University of Pisa Report [5]. It revealed that each country used their own traditional simplified density model, from the most simple single constant values in countries with oceanic or Mediterranean climate or flat terrain (in kgm^{-3}: Belgium-150, France-150, Greece-125, Ireland-156.82, Luxembourg-150, Netherlands-100, UK-156.82, Denmark-200) through more elaborated regional values in Sweden.

II. THEORY AND DATA

A. Data

Most of the precipitation and climatologic stations in Croatia measure snow depth with a snow stick, while only 13 stations from the 1971–2000 period had snow density data. A spring balance or a snow sampler is used at eleven stations, except in Skrad and Zavižan where the standard Hellmann cutter tube is used. With spring balance, the mass and depth of snow is measured and density is calculated directly as the mass to volume ratio. Using the Hellmann cutter tube seeks for additional measurement of the snow water content from the melted snow sample.

Measured snow densities lower than or equal to 30 kgm^{-3} as well as higher than and equal to 600 kgm^{-3} were discarded. Since [8] discuss so called “wild snow”, a low density new

snow of 40 kgm^{-3} as unique for some regions, it was proposed that density lower than 30 kgm^{-3} is very likely an outlier for our area. Old snow may reach a density of 400–500 kgm^{-3} (www.avalanche-center.org/Education), and firn even 600 kgm^{-3}. Since firn is typical for glaciers that do not exist in Croatia the densities equal or larger than 600 kgm^{-3} has been considered as outliers also. This is confirmed also in [21] where it is stated that [14] found that the maximum density of seasonal snow approaches 600 kgm^{-3}.

For the size of the Croatian territory of 56 594 km^2 the representative spatial density of the data would be around 150 stations [10], to access the map spatial resolution of 1 km, which presents a challenge since 118 station measure snow depth but only 13 of those measure snow density. Hence, the question was how to estimate the snow density also on the rest of the 105 stations measuring only snow depth. Preparing the first characteristic snow loads for Croatian standards in 2001 [24] the typical density was determined as the mean value of mean and median of all data and mean and median of 20% of situations with maximum snow load in every winter.

B. *Snow load estimation*

Snow load is the weight of snow cover on the surface of 1m^2 (2.1).

$$\begin{aligned} s\left[Nm^{-2}\right] &= m_s g / A\left[kgms^{-2} / m^2\right] \\ &= \rho_s h_s A g / A \\ &= \rho_s h_s g \end{aligned} \qquad (2.1)$$

where m_s[kg] is mass of snow, g[ms^{-2}] is gravitational acceleration, A[m^2] is surface area, ρ_s[kgm^{-3}] is snow density and h_s[m] is snow depth. Using (2.1) the snow load is calculated as a product of snow density, snow depth and the gravitational acceleration.

C. *The snow density model*

Snow density has been measured on a small number of stations and their spatial coverage is not adequate for the purpose of mapping snow load because they cover mainly the lowland, continental part of Croatia and only a smaller part of the mountain area, and there is no information on density on the coast. To overcome this and to use additional information from the 105 stations with snow depth measurements, the snow density models based on snow depths have been built, through linear regression of snow density (ρ_s) on snow depth (h_s).

D. *Maximum annual snow load estimation*

This series of annual maximum snow loads is the basis for the estimation of the annual maximum snow loads for the 50-year return period s_k, by means of the generalized extreme value (GEV) theory. These values are referred to as characteristic snow loads. Most of the stations had at least 20 years of available data. During the work on the establishing the snow load zones on the map of the United States it is confirmed that adding a big snow year to data developed from periods of record exceeding 20 years will usually not change 50-year return values much [2] so that criteria is adopted here also. Only eight stations with shorter data series were included in the analysis because of their location in data sparse or mountain areas, where there are also very few stations.

E. *Mapping – regression kriging*

Finally, the estimated annual maximum snow loads for the 50-year return period s_k at 118 stations constitute the input for the geostatistical procedure of mapping this parameter for the Croatian territory by the regression kriging (RK) method [16, 11]. It uses correlation with multiple environmental predictors through regression and spatial autocorrelation of the targeted values through kriging for estimation of values at the new locations. RK has been applied in mapping numerous climatic parameters for the territory of Croatia [25], which has been explained in detail in [18]. The RK framework typically consists of four steps: (1) the deterministic part of the variation is modelled using the auxiliary maps; (2) the residuals are modelled for spatial autocorrelation (variogram); (3) predictions and prediction errors are computed using the RK model and (4) the accuracy of the predictions is evaluated using cross-validation (CV). The RK is applied to estimate the ln(sk) for the whole territory of Croatia and finally, the exponential back transformation is used for the estimation of the s_k field.

ME and RMSE have also been used for assessing the accuracy of the regression kriging through the leave-one-out cross-validation (LOOCV) technique as implemented in the gstat package [17]. Leave-one-out cross validation visits a data point, and predicts the value at that location by leaving out the observed value, and proceeds with the next data point.

Significant predictors for the snow load regression model were selected using the stepwise method of the *base* package in *R* [22, 19]. Regression kriging has been performed in the *gstat* package [17, 15] in the *R* open-source environment for statistical computing and visualization.

III. RESULTS

A. *Snow depth and snow density – spatial differences*

The first step in the analysis was the establishment of the linear density models based on daily snow depth and density data from 13 meteorological stations. Since snow depth and density change with altitude and season, the stations have been divided into three subsets according to station altitude and the regression equations are also built on monthly basis [7]. Inference in the snow depth data showed that inland stations below 200 m can expect around 20 cm of snow with the maximum between 56–79 cm. Stations between the altitudes of 300 m (in the intermediate region between the continental and the Gorski kotar region) and 600 m (in the Lika region) have a mean snow depth of 29 cm, while the maximum may be as high as 120 cm. Higher stations in the Gorski kotar region (higher than 600 m) have a mean snow depth of 38 cm, and a maximum of 160 cm. The highest Croatian meteorological station, Zavižan, has on average 93 cm of snow. 50% of the snow depths on Zavižan are between 44 and 130 cm while the maximum is 320 cm.

Mean snow density at lower altitude stations is 182.3 kgm^{-3}, at two higher stations (300–600 m) the average is 195.8 kgm^{-3}, and in the Gorski kotar region (600–1000 m) the average value is 228.6 kgm^{-3}. The highest mean snow density value is the one at the Zavižan station (351.5 kgm^{-3}). The differences in maximum densities between stations are smaller than the differences in mean values, comparing the coefficients of variation. For the stations at different altitudes, the differences in mean snow densities are less pronounced than the differences in mean snow depth, based on the coefficients of variation (also discussed in [7]).

These differences in snow density, and especially in snow depth, led to the definition of borders for the three altitude ranges: low-altitude stations (0–600 m), higher altitude stations (600–1000 m) and the highest Zavižan station on Velebit Mountain (1594 m) and the three sets of regression equations.

B. Snow depth and snow density – temporal differences

With regard to the annual course of snow densities for the lower altitude stations it may be seen that the median is quite stable, around 200 kgm^{-3}, with slightly larger densities in January, February or March. For the higher altitude stations the density shows an increase from the beginning of the winter season to maximum density in March. For the Zavižan station the increase in density is apparent through the entire winter season. This was the motivation for distinguishing the regression models in terms of density on a monthly basis as well. For some months and altitude ranges no linear regression could be found between snow depth and density, so in snow load calculations the average density as calculated from the data is used.

The base density increases from the beginning of the winter season to the end (March for low and higher altitude stations and April for the Zavižan station), which corresponds to increased compaction of snow due to settling and ripening [7]. Some occasional snowfall may also occur in April at low and higher stations and in May on Zavižan, but the density is again lower than at the beginning of the winter. Comparing for every altitude region the density (for several snow depths) with mean temperatures on the monthly bases revealed the similarity in increase of the density through the winter season with the maximum on the last month that has mean temperature below or around 0°C. This maximum density around 0°C is typical for the new snow also [8].

Comparing individual months, base density also increases from lower altitude stations to Zavižan. The negative slope of the regression curve is the consequence of the large spread of snow density for low snow depths, because there are two types of snow data in the data set: fresh, not so dense snow, but also older, partly melted and denser snow. As denser, probably more mature and thin snow prevails in the data set, a slightly negative slope coefficient of the regression line results. For the Zavižan station, the spread of snow density for low snow depths is also large, but since there is also a substantial quantity of deep snow with high density, this results in a positive slope in regression line.

The regression models were also tested on a monthly basis for a single altitude class to compare them with those for the lowest altitude range (alt < 1400 m) in [7]. In that case, the regression had a positive slope similar to [7] that seems to be typical for the higher altitudes. The negative slope of the regression line is discussed in [23] where they confirm based on observations on snow density and snow depth that higher densities (except for mountainous regions) are usually associated with shallow snow covers and extreme wind and temperature conditions (i.e. prairies).

C. Accuracy of the regression models

The snow density regression models were tested for the accuracy of the snow load calculation through the ME and RMSE where they proved to be capable of reproducing the measured maximum annual snow loads at those 13 stations. ME is between -0.1 kNm^{-2} for lower altitudes and -0.4 kNm^{-2} in Zavižan. RMSE is between 1.15 kNm^{-2} and 0.78 kNm^{-2}.

D. Characteristic snow loads at meteorological stations

According to the Final Report I [5] for the stability of structures, there is a demand for excluding exceptional snow loads from the GEV analysis. Those exceptional snow loads have been defined as isolated and very infrequent snowfalls where the resulting snow load is significantly greater than the loads in the general body of snow load data. Exclusion of these exceptional snow loads from the data set of a particular station reduces the average annual snow load at the 25 coastal stations by 0.2 kNm^{-2} and by 0.35 kNm^{-2} at the five continental stations with exceptional snow loads. For most of the stations, data are close to the Gumbel family of the GEV distributions. In coastal areas of Croatia snow is a rather rare event and maximum snow load cannot be calculated using theoretical distributions; hence, the maximum snow load for a 50-year return period was estimated as the 98th percentile of empirical distribution.

E. Characteristic snow load map

From the histogram of the characteristic snow load data, the deviation from the normal distribution has been identified. For this type of data distribution, the logarithm (ln) transformation is the most appropriate [3, 12] for data normalisation prior to multiple regression analysis. In addition, some authors [9] use the power function to define the relationship between snow load and altitude. Taking the logarithm of this type of function leads to linear dependence of $\ln(s_k)$ with altitude, which also supports the use of ln transformation in this research.

Several possible combinations of predictors were investigated and finally the regression model for $\ln(s_k)$ was selected that explained 75% (adjusted coefficient of determination, R_a^2) of the spatial variability of the snow load data. The most important predictor was the one mainly influenced by the weighted distance from the sea, reflecting the large maritime as well as the influence of altitude on the values of this climate element on Croatian territory.

In the second step of the regression kriging framework, the exponential variogram model with the parameters C0=0.05

(nugget), C1=0.30 (sill) and a=16423 m (range) was fitted to the residuals.

The final step in the mapping process is the calculation of the prediction map, that is, the characteristic snow load map and the prediction variance map. When predicting the ln transformed value, the prediction has to be back transformed to get the final prediction, but the prediction variance map cannot be back transformed, so it may give us limited information on mapping accuracy. The localized kriging based on 90 nearest meteorological stations was applied for prediction. This number is estimated according to the leave-one-out cross-validation procedure, as the one for which the average residual value is the lowest, and accuracy the highest.

The map of characteristic snow load is presented with 14 snow load classes of unequal width, ranging from 0–14 kNm^{-2} (Fig. 1). The largest part of Croatian territory can expect a snow load between 0.5–1.5 kNm^{-2} (60% of the area). The coastal areas expect a snow load of up to 0.5 kNm^{-2} (19% of the territory). Snow load between 1.5–2.0 kNm^{-2} is expected on 7% of the territory. Only 4% of the territory expects snow loads between 2.0–2.5 kNm^{-2}, while snow loads larger than 2.5 kNm^{-2} are expected on 9% of Croatian territory. The highest snow load values between 12.0–14.0 kNm^{-2} are expected on Velebit Mountain, where the snow load at the Zavižan station was measured at 12.0 kNm^{-2}.

The uncertainty of the prediction [11] is larger at the country borders and in data sparse areas according to the kriging variance map of the $\ln(s_k)$ (not shown).

Mapping accuracy was tested with the ME and RMSEr, that is, RMSE normalised by standard deviation of the data. As a rule of thumb [11], the prediction is accurate if RMSEr < 40%. In that case, the model explains more than 85% of variability (1-$RMSEr^2$) at the validation points. The ME of the applied regression kriging model was 0.02 and RMSEr=0.31, which corresponds to a 90% accuracy at the validation points; hence this RK model can be considered very accurate for the spatial prediction of the characteristic snow load values on the territory of Croatia.

IV. Discussion and Conclusions

The purpose of this study was to supply a characteristic snow load map of Croatia as an important national annex for the application of standards for the design of structures. The work consists of three main steps: building a snow density model based on snow depths, estimation of the snow loads and characteristic snow loads at station locations and prediction of the characteristic snow loads on the regular grid and at unvisited locations, that is, producing a map of characteristic snow load. In building a snow density model, daily data of snow densities and snow depths were available. Snow density models were built taking into account three important factors affecting snow density: season, snow depth and altitude.

Snow depth is more important factor for the variability of the snow load spatial distribution on Croatian territory, compared with the influence of the snow densities and measured using the coefficient of variation of those two factors between the stations. The regional differences in climate conditions affect snow depth spatial distribution, whose spatial patterns are similar to the temperature and precipitation patterns on the territory of Croatia [18]. As in the case of these two climatic variables, the most influential climatic factors are the altitude and distance from the coast. For example, Zavižan and Gorski kotar are the coldest regions with the highest precipitation and consequently with the highest snow depths. In the Lika region it is not as cold and precipitation is lower, as well as the snow depths, while warmer compared to mountains, continental parts of Croatia (at lower altitudes) has even less precipitation and snow fall. Due to warmer winters, there is snow fall only on rare occasions on the Adriatic coast.

The highest predicted snow load values in the regression kriging framework may be considered quite high, and in certain other research, [9] the prediction of the snow load at altitudes higher than the altitude of the highest station was set to the value measured at that station. We decided to retain the RK prediction values.

The characteristic snow load map for Croatia has been visually compared with the neighbouring Slovenian map [1]. The impression is that the values at the border match and that the values at similar altitudes match on the two maps. Together with all applied accuracy measures (ME, RMSRr) this also inspires confidence that the presented method can successfully predict the snow loads at locations where only snow depths are measured applying the snow density model and, furthermore, that the method can predict snow loads at unvisited locations with the RK model.

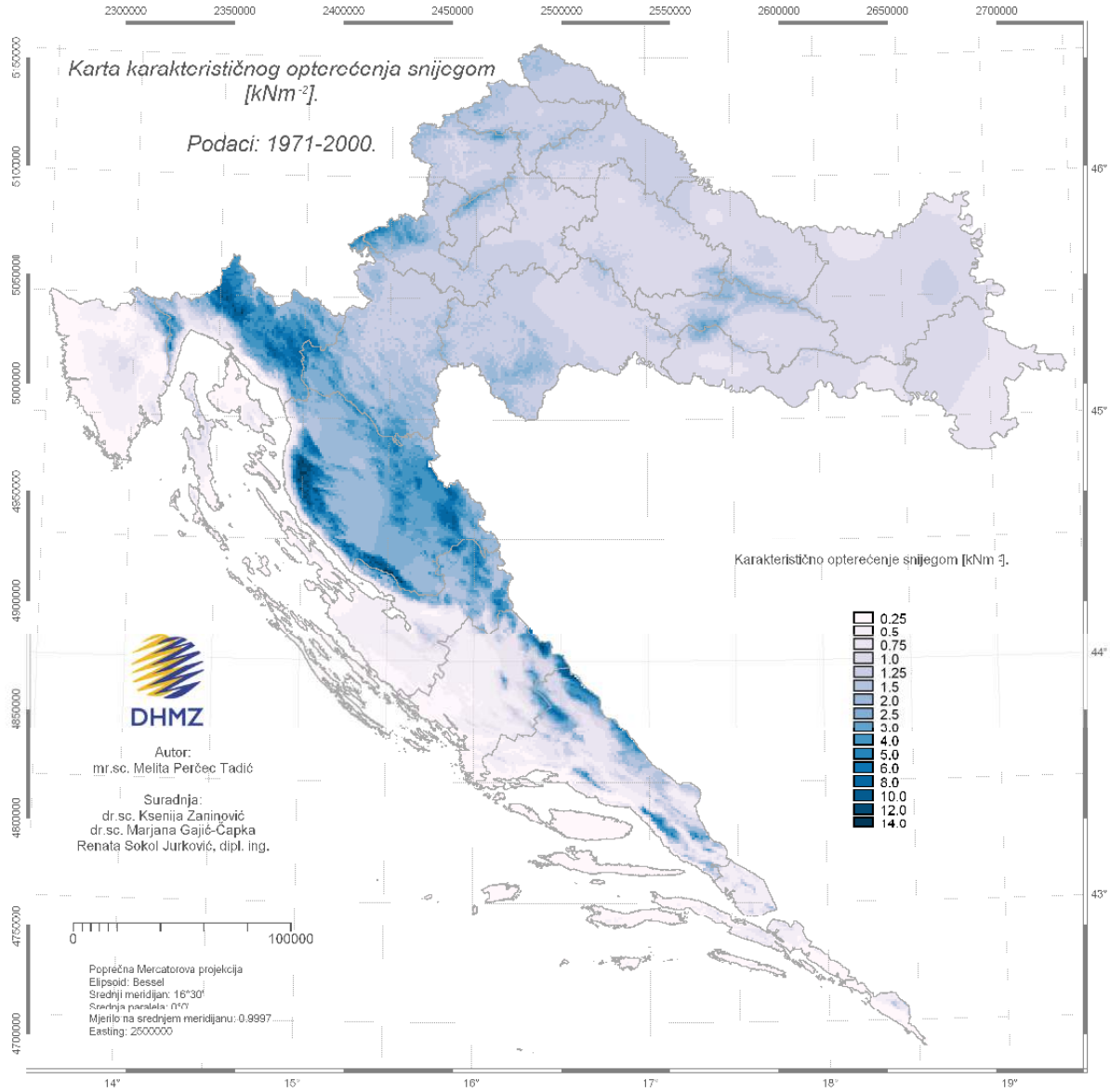


Fig. 1. Map of the characteristic snow load [kNm^{-2}].

REFERENCES

[1] ARSO (accessed on Oct. 20th 2011.) The Environmental Agency of the Republic of Slovenia. www.arso.gov.si, http://meteo.arso.gov.si/uploads/probase/www/climate/image/en/by_variable/snow/max_snow_load_50years_51_05.png

[2] ASCE/SEI 7-05, Minimum Design Loads for Buildings and Other Structures. American Society of Civil Engineers, 2006.

[3] PA. Burrough and RA. McDonnell, Principles of Geographical Information Systems. Oxford University Press, Oxford, 2004.

[4] HJ. Critchfield, General Climatology. Prentice.Hall, Inc. Englewood Cliffs, New York, 1960.

[5] DGIII-D3, Scientific support activity in the field of structural stability of civil engineering works, Snow loads, Final report I. Commission of the European Communities, DGIII-D3. Contract 500269, 1996.

[11] T. Hengl, A practical guide to geostatistical mapping. University of Amsterdam, Amsterdam, 2009.

[6] NR. Draper, H. Smith, Applied regression analysis, 3rd edn. Wiley, New York, 1998

[7] T. Jonas, C. Marty, J. Magnusson, Estimating *the snow water equiv*alent from snow depth measurements in the Swiss Alps. Jou. Hydrol. 378:161–167, 2009.

[8] A. Judson and N. Doesken, Density of Freshly Fallen Snow In the Central Rocky Mountains. BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY .81-7:1577-1587, 2000.

[9] MJ. Luna, A. Morata, A. Chazarra and A. Almarza, Mapping Of Snow Loads On The Ground In Spain. Geographical Information Systems and Remote Sensing: Environmental Applications. *(Proceedings of the International Symposium held at Volos, Greece, 7–9 November 2003), 2005.*

[10] T. Hengl, Finding the right pixel size. Computers and Geosciences, 32(9):1283-1298, 2006

[12] T. Hengl, G. Heuvelink, A. Stein, A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120:75–93, 2004.

[13] MJ. O'Rourke, Snow Load on Buildings, American Scientist, 85, 64-70, 1997.

[14] WSB Paterson, The Physics of Glaciers. 2nd ed. Pergamon Press, 1981.

[15] EJ. Pebesma, Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691, 2004.

[16] EJ. Pebesma, The role of external variables and GIS databases in geostatistical analysis. T GIS 10(4):615–632, 2006.

[17] EJ. Pebesma and CG. Wesseling, Gstat: a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences* Vol. 24, No. 1:17–31, 1998.

[18] M. Perčec Tadić, Gridded Croatian climatology for 1961–1990. Theor Ap*pl Climatol Vol 102, No*. 1–2:87–103, 2001.

[19] C. Reimann, P. Filzmoser, RG. Garrett, R. Dutter, Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley and Sons, Ltd., 2008.

[20] JP. Roebber, SL. Bruening, DM.Schultz and JV.Cortinas, Improving snowfall forecasting by Diagnosing Snow density. Weather and Forecasting. 18:264:287, 2003.

[21] M. Sturm, B. Taras, GE. Liston, C. Derksen, T. Jonas, J. Lea, Estimating Snow Water Equivalent Using Snow Depth Data and Climate Classes. Jou. Hydrom. 11:1380–1394, 2010.

[22] WN. Vanables, DM. Smith and R Development Core Team, An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Version 2.11.1 (2010-05-31), 2010.

[23] GP. Williams and LW. Gold, Snow Density And Climate. Transactions, Engineering. Institute Of Canada. VOL. 2, NO. 2:91-04, 1958.

[24] K. Zaninović, M. Gajić-Čapka, B. Androić, I Džeba; D. Dujmović, Determination of typical snow load (in Croatian, Abstract in english), Građevinar, 53, No. 6, 363-378, 2001.

[25] K. Zaninović (ed.), M. Gajić-Čapka, M. Perčec Tadić, M. Vučetić, J. Milković, A. Bajić, K. Cindrić, L. Cvitan, Z. Katušin, D. Kaučić, T. Likso, E. Lončar, Ž. Lončar, D. Mihajlović, K. Pandžić, M. Patarčić, L. Srnec, V. Vučetić, Klimatski atlas Hrvatske / Climate atlas of Croatia 1961–1990, 1971–2000. Državni hidrometeorološki zavod, Zagreb, 2008.

Rainfall Variability and NAO, Spatial Pattern

[extended abstact]

Jelena Luković[1], Branislav Bajat[2], Dragan Blagojević[2], Milan Kilibarda[2].
[1]Faculty of Geography, University of Belgrade, Belgrade, Serbia
jlukovic@gef.bg.ac.rs
[2]Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia

***Abstract*—This study examines spatial pattern of relationship between annual and seasonal rainfall in Serbia and North Atlantic Oscillation in the period 1961-2009. Correlation analysis is done to test relationship between parameters while spatial analysis involved autocorrelation analysis. Correlation coefficients were plotted on Google maps using a *plotGoogleMaps* package. Results showed negative mainly statistically significant correlations at annual and winter scale.**

Keywords—rainfall; NAO; correlation; spatail pattern;

I. INTRODUCTION

This paper examines annual and seasonal rainfall variability in Serbia and North Atlantic Oscillation (NAO) in the period from 1961 until 2009.

The North Atlantic Oscillation is one of the dominant parameters of global climate. The term itself has been introduced by Sir Gilbert Walker in 1920. Traditionally has been defined as sea level pressure difference between subtropical anticyclone over the Azores and Sub polar depression over the Iceland. It has considerable impact on winter weather conditions in Europe and some parts of North American continent. This pressure difference is normal condition which becomes more intensive during winter. There are positive and negative phase of NAO. During the positive phase winters in North Europe are warmer and wetter with less precipitation then usually. At the other hand, negative phase is followed by colder winters in North Europe and higher precipitation amount in Southern Europe. There are two types of NAO indices: one given by Rogers (1984) that presents pressure difference between Iceland (Akureyri) and Azores (Ponta Delgada) and second one introduced by Hurrell (1995), describing pressure difference between Iceland (Stykkisholmur) and Portugal (Lisbon). Portis et al. (2001) have introduced "mobile" NAO index that change in space depending on season. It shows higher correlations with the intensity of western winds over the mid-latitudes in North Atlantic, then traditional NAO indices.

Still, there is no unique scientific agreement about the mechanism of the NAO origin. What is proved is that it is atmospheric phenomenon which is the result of the ocean-atmosphere interaction. Possible explanations are very wide from natural causes to atnthropogenic.

The impact of NAO can be seen at different levels weekly, monthly and decadal. Sometimes weekly and monthly oscillations could be caused by polar stratospheric circulation (Baldwin, Dunkerton 2001). Decadal and annual NAO changes could also be determined by anomalies of ocean surface temperature and could have significant impact on temperature and precipitation in Europe (Hurrell, 1995).

Hurrell и Van Loon (1997) have got that NAO in its extreme phase, after 1980 had impact on surface warming over Europe during winter as well as temperature decrease over the North West Atlantic. The impact on precipitation can be seen in dry weather conditions in Southern Europe and in Mediterranean region and wet in North Europe and Scandinavia. The authors also noticed that NAO can influence the storm tracks, moving it towards the north. They highlighted that NAO should be investigated in the upper parts of the troposphere in order to find out regional pattern of the changes caused by NAO.

Ducić et al. (2007) have investigated connection between ENSO and NAO indices and rainfall in Serbia in the period from 1951 until 2000. Trend analysis for decadal values has shown statistically significant result for two stations out of 20. With the application of cluster analysis all stations are grouped into three clusters. For each cluster has been calculated correlation coefficient with ENSO and NAO indices. Some stations have shown statistical significance. Such high R value could be possible explained by indirect mechanism of ENSO impact on NAO index (Harrison and Larkin, 1998).

On the base of the results given by Pohlmann и Latif (2005) it has been presumed that the impact of the Atlantic on precipitation in Serbia is more intensive during summer while precipitation in winter is influenced by Atlantic and Indo-Pacific. NAO impact can be determined in stations with continental regime, while ENSO impact could be determined in those with Mediterranean one (Ducić et al. 2007).

The aim of this study is to examine relationship between NAO and annual and seasonal rainfall in Serbia in the period from 1961 to 2009.

II. DATA SETS AND METHODOLOGY

A. *Data sets*

In this paper, the data were collected from 63 meteorological stations provided by the Hydro-Meteorological Service of Serbia. The data is compiled from 29 synoptical stations and 34 climatological stations. The weather station network is spatially distributed quite well. However, it should

be indicated that station networks in mountainous areas are sparse and of uneven distribution due to the lack of measurements in these areas. The data, compiled from the period between 1961 and 2009, is quality controlled in terms of correction of misprints, relocation history and missing values (WMO 2002). In order to ensure data quality 63 weather stations are with complete series. Annual and seasonal rainfall data were calculated for each station and seasons were considered as following: spring (March–May), summer (June–August), autumn (September–November), and winter (December–February).

As a parameter of North Atlantic Oscillation NAO index[1], defined as a air pressure difference between Iceland and Azores, has been used at annual and seasonal scale.

B. *Methodology*

After visual inspection the precipitation time series were subject to analysis correlation determination, followed by a standard procedure of hypothesis testing in order to assess the statistical significance of the results.

Furthermore, depending on the data distribution, either the parametric or nonparametric method may be used for correlation detection. In general, nonparametric methods perform better relative to their parametric counterparts for abnormal distributions. Due to the fact that a preliminary analysis has shown the presence of a skewed distribution in some of our precipitation time series, the nonparametric method of the *Kendall's tau* test

The global autocorrelation index *Moran's I* (O'Sullivan and Unwin 2003) and local autocorrelation index like *Getis-Ord Gi** statistics (Getis and Ord 1992) were used for detecting spatial patterns in the distribution of meteorological stations by considering both their locations and associated correaltion values. The calculated spatial autocorrelation indices measure and test how observed locations are clustered /dispersed in space with respect to their attribute values.
*Getis-Ord Gi** statistics is used to detect possible non-stationarity of the data; i.e. clustering patterns in specific sub-regions. In addition, *Hot Spot Analysis* (Lee and Wong, 2005) incorporating *Getis-Ord Gi** statistics was used to provide more insight into how the locations with high and low levels of estimated correlations are clustered.

We used a recently developed software package called *plotGoogleMaps*[2] to get better insight in the spatial distribution of calculated correlation coeficents in Serbia. The software integrates Asynchronous JavaScript, XML (AJAX), and the Google Maps Application Programming Interface (API) service to produce HyperText Markup Language (HTML) file map mashups (web maps) that maintain high-resolution Google Map images as background data. The tool *plotGoogleMaps* is developed in the open source R software language (R-project, 2011), and is designed to automatically create web maps by combining the users' data and Google Maps layers (Kilibarda and Bajat 2012).

[1] http://www.cdc.noaa.govn/Pressure/Timeseries/nao.long.data.
[2] http://cran.r-project.org/web/packages/plotGoogleMaps/index.html

The *plotGoogleMaps* software offers many advantages when compared to other classical graphic device environments. The high quality of the background Google layers make better abstractions of geographical reality and allow the user to explore data spatially with a variety of interactive controls (navigation control, pan, zoom, attribute info windows, etc). This package promotes the creation of interactive maps in user friendly environments where the map is stored in the HTML format.

III. RESULTS

Results of the analysis have shown generally negative correlations between rainfall and NAO index both on annual and seasonal level (Fig. 1 and 2). Annually 17 stations showed statistically significant correlation coefficients, mostly located in eastern parts of Serbia. Results obtained for winter season showed statistically significant correlations at 55 stations, evenly distributed over Serbia. For other seasons significant correlations are not calculated that is why maps for are not produced.

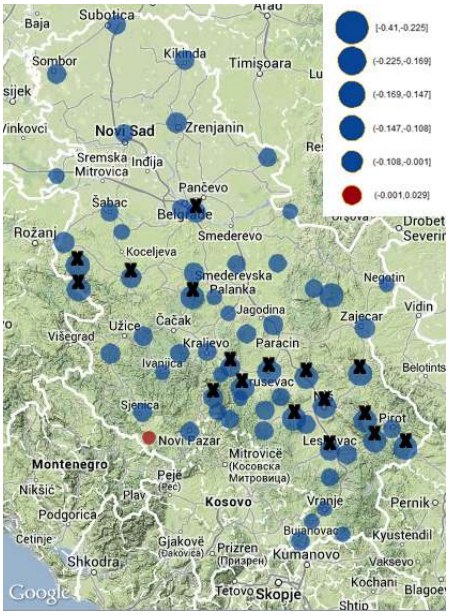


Fig. 1.Spatial pattern of correlation coefficients between annual rainfall and NAO (1961-2009). Stations showing significant correlations are marked with X.

Hurrell and Van Loon (1997) found NAO having impacts on drier winter conditions over southern Europe and Mediterranean area and wetter over northern Europe and Scandinavia. Authors suggested that NAO may influence path of the storms, moving them towards the north. They also suggested that this should be particularly studied in upper layers of troposphere in order to recognize regional pattern of changes caused by NAO.

TABLE I. Moran's I autocorrelation statistics and Z value in Serbia.

	Moran I	***Z***	***p***	**comment**
NAO annual	-0.02	0.02	0	The pattern is neither clustered nor dispersed.
NAO winter	-0.08	-1.27	0	The pattern is neither clustered nor dispersed.
NAO spring	-0.05	-0.61	0	The pattern is neither clustered nor dispersed.

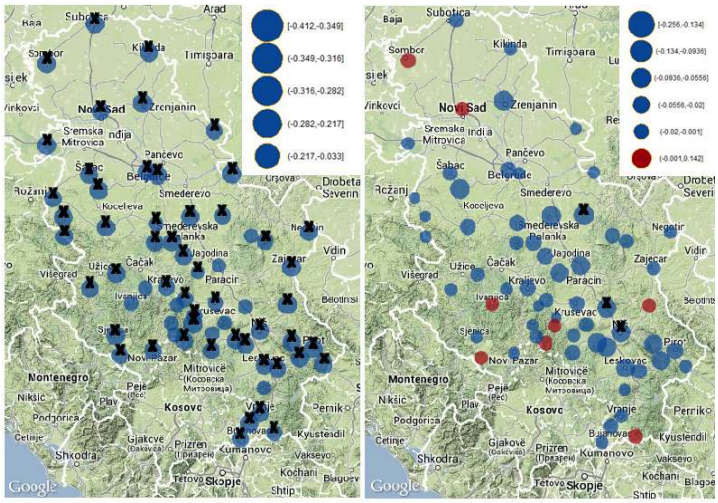


Figure 2. Spatial pattern of correlation coefficients between winter rainfall and winter NAO (left) and spring rainfall and spring NAO (right) in the period 1961-2009. Stations showing significant correlations are marked with X.

In order to test spatial clustering of calculated correlations autocorrelation analysis is applied at annual, winter and spring level. However, results are not showing any significant clustering (Table 1).

IV. Conclusion

In this paper we have studied relationship between annual and seasonal rainfall in Serbia and North Atlantic Oscillation for the period 1961-2009. Focus of the research was given to spatial pattern of this connection. Therefore, apart from the correlation analysis, autocorrelation statistics has been applied. Results have shown the strongest correlation for winter season. Both annual and winter scale showed negative relationship between the investigated parameters. For winter season 55 stations showed negative statistically significant correlations. All values are plotted on Google maps using a *plotGoogleMaps* package. Results for Serbia fit in to those obtained for Europe (Hurrell and Van Loon, 1997).

Acknowledgment

This study was supported by the Serbian Ministry of Education and Science, under grants No. III, 43007, III 47014, TR 36035 and TR 36020.

References

[1] V. Ducić, B.Milovanović, J. Luković, Temperature changes on the Balkan Peninsula in the period of satellite observation and possible volcanic influence. Fourth International Conference- Global changes and regional challenges, Sofia University "St. Kliment Ohridski", Faculty of Geology and Geography, 20-22 April 2007, Sofia, Bulgaria, Proceedings, 2007.

[2] A. Getis, and J.K Ord., The analysis of spatial association by use of distance statistics. Geografiska Annaler, 24(3):189–206, 1992.

[3] D. E. Harrison and N. K. Larkin, El Nino-Southern Oscillation sea surface temperature and wind anomalies. Rev. Geophys., no. 36, pp. 353-399, 1998.

[4] J. Hurrell, Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. Science, no. 269, pp. 676-679, 1995.

[5] J. W. Hurrell, and Harry Van Loon, Decadal Variations in climate associated with the North Atlantic oscillation, Climatic Change 36, pp. 301–326, 1997.

[6] D. O'Sullivan, and D Unwin, Geographical Information Analysis, John Wiley & Sons, New Jersey, p 436, 2003.

[7] M. Kilibarda, B. Bajat, PlotGoogleMaps: the R-based web-mapping tool for thematic spatial data. *Geomatica*, 66:37-49, 2012.

[8] J Lee, and D.W.S Wong, Statistical analysis of geographic information with ArcView GIS and ArcGIS. Wiley, New York, p 446, 2005.

[9] H. Pohlmann and M. Latif, Atlantic versus Indo-Pacific influence on Atlantic-European climate. Geophysical Research Letters, vol. 32. L23710, 2005.

[10] D.H. Portis, J.E. Walsh, M. E. Hamly and P.J. Lamb, Seasonality of the North Atlantic Oscillation. Journal of Climate, 13, pp. 2069-2078, 2000.

[11] J.C. Rogers, The association between the North Atlantic oscillation and the Southern oscillation in the Northern Hemisphere. Mon. Wea. Rev., 112, pp. 1999-2015, 1984.

Future Changes in Drought Characteristics in Serbia

[extended abstact]

Aleksandra Kržič
Republic Hydrometeorological Service of Serbia
RHMSS/SEEVCCC
Belgrade, Serbia
aleksandra.krzic@hidmet.gov.rs

Vladimir Djurdjević, Ivana Tošić
Institute of Meteorology
University of Belgrade-Faculty of Physics
Belgrade, Serbia
vdj@ff.bg.ac.rs; itosic@ff.bg.ac.rs

***Abstract*— In this study, possible future changes in drought characteristics in Serbia were analyzed using two drought indices, the Standardized Precipitation Index (SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI) on time scale of 12 months.**

Study showed that the SPEI index is more suitable for drought monitoring and projections because it includes evaporative demand. According to the EBU-POM model projections, it is very likely that the drought frequency and its severity will increase in the future.

Keywords*— *drought characteristics; SPI; SPEI; Serbia

I. INTRODUCTION

The most common tools for monitoring drought conditions are drought indices. Most of them are based solely on precipitation, some are based on precipitation and evapotranspiration, while others are related to runoff and vegetation conditions [1]. Some of the drought indices are Palmer Drought Severity Index (PDSI), Standardized Precipitation Index (SPI), Standardized Precipitation Evapotranspiration Index (SPEI), Crop Moisture Index (CMI), Keetch-Byram Drought Index (KBDI), etc.

Numerous studies used these indices to analyze drought characteristics, e.g. in Turkey [2], Greece [3], Iberian Peninsula [4], Portugal [1], Czech [5], Serbia [6, 7, 8]. These studies are based on observed data sets. In [9], drought statistics based on the PRUDENCE multi-model approach are estimated. According to [9], British Isles would experience more intense short-term droughts but less severe longer duration events. The Mediterranean was identified by [10] as a particularly vulnerable region to global climate change.

In this study, we will analyze the present and future changes in drought characteristics in Serbia using the SPI and SPEI indices estimated from the regional climate model [11, 12, 13].

II. DATA AND METHODOLOGY

A. Data

Observed dataset used for model verification comprises monthly values of air temperature (29 stations) and accumulated precipitation (30 stations) for the period 1961-1990 and 2001-2010. Stations are equally distributed throughout the country. Technical and quality control of these measurements were made by the Republic Hydrometeorological Service of Serbia (RHMSS).

Outputs from the atmosphere-ocean two-way coupled regional climate model, the EBU-POM [11, 12], are used as well. The atmospheric part of the EBU-POM presents Eta/NCEP model and ocean part, Princeton Ocean Model (POM). The atmospheric model was initialized and forced by lateral boundary conditions using fields from the coupled atmosphere-ocean general circulation model SINTEX-G. The atmospheric model domain covers the greater part of European region (Fig. 1) with horizontal resolution of 0.25°. The ocean model horizontal resolution over the Mediterranean was 0.2°.

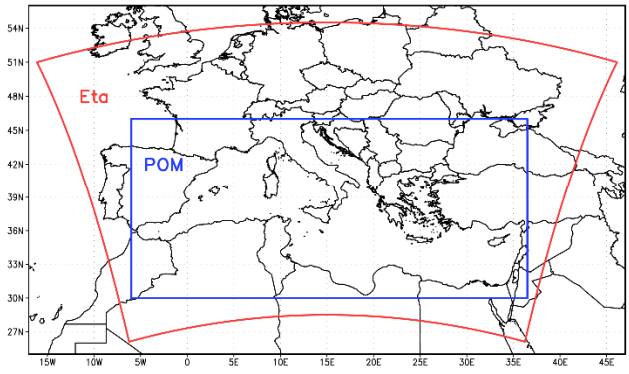


Fig. 1. Domain of the EBU-POM regional climate model

Regional integrations were performed as 30 years' time slices of global model experiments for two periods and climate change scenarios: 1961-1990 – the reference period, and 2071-2100 – projections for the A1B and A2 scenarios. Chosen scenarios are known as 'medium' and 'high' forcing scenarios. For the A1B atmospheric concentrations of CO2, at the end of 21st century, are ~1.8 and for the A2 ~2.2 times higher than the present value of ~390 ppm.

B. Indices

In this study, we analyzed the main characteristics of drought using the Standardized Precipitation Index (SPI) and

the Standardized Precipitation Evapotranspiration Index (SPEI) on time scale of 12 months. The 12-month time scale was chosen because it is proven to be relatively good correlated with soil moisture and river discharge. The SPI is a popular index because of its simplicity, only precipitation data are needed. The National Meteorological and Hydrological Services around the world are encouraged to use the SPI in order to characterize meteorological droughts [14]. On the other side, SPEI can account for the possible effects of temperature variability and temperature extremes (through evapotranspiration and the water supply-demand relation) in the context of global warming [11].

Because indices are standardized, comparing climatic conditions of areas with different hydrological regimes is allowed. The strength of the anomaly is classified as set out in Table I [15].

TABLE I. CLASSIFICATION OF INDICES

SPI/SPEI Values	**Drought category**
< -2.326	Exceptional drought
-2.325 to -1.645	Extreme drought
-1.644 to -1.282	Severe drought
-1.281 to -0.935	Moderate drought
-0.934 to -0.524	Minor drought
-0.525 to 0.524	Near normal

III. RESULTS

Both indices obtained with the observed data sets, mapped using open source SAGA-GIS software and ordinary kriging method, show normal moisture conditions for the period 2001-2010 (Fig. 2).

Looking at the SPI values (Fig. 3), moisture conditions in the period 2071-2100 will be normal with the exception of southwest Serbia and A1B scenario (minor drought). It is in accordance with a more distinct decrease of the precipitation amount in Serbia [12] for A1B (13 mm/season) than for the A2 scenario (6 mm/season). For the period 2071-2100 and both scenarios, significant influence of temperature increase and precipitation decrease is evident on the spatial distribution of SPEI index (Fig. 4). The whole considering area is in the category extreme to exceptional drought.

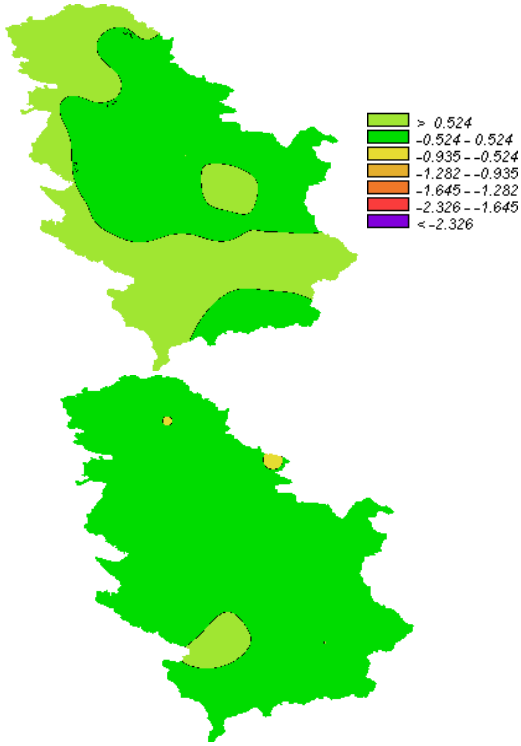


Fig. 2. Spatial distribution of the SPI12 (upper) and SPEI12 (bottom) for the period 2001-2010 based on observational data sets

IV. CONCLUSIONS

The regional climate model EBU-POM reproduces well natural moisture conditions. Study showed that the SPEI index is more suitable for drought monitoring and projections because it includes evaporative demand. According to the EBU-POM model projections and looking at the SPEI values, it is very likely that the drought frequency and its severity will increase in the future, thereby enhancing the associated impacts.

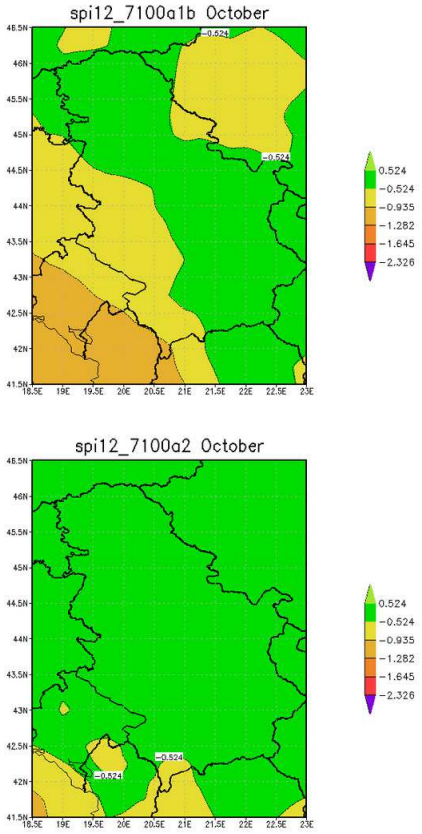


Fig. 3. Spatial distribution of the SPI for the period 2071-2100 and both scenarios A1B and A2 based on the EBU-POM data sets

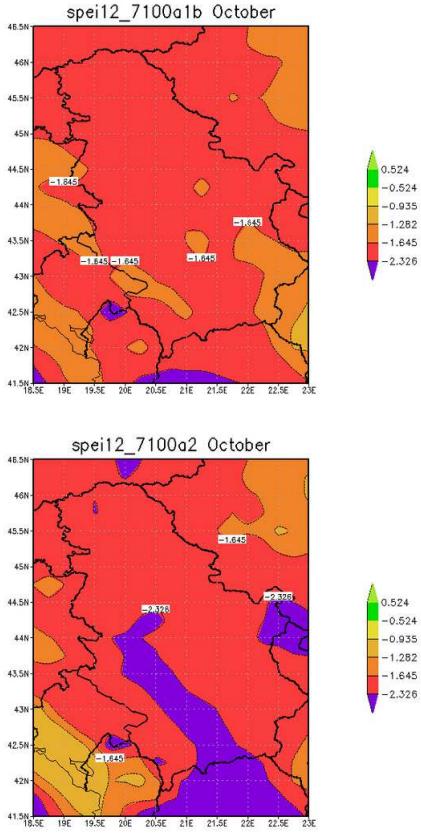


Fig. 4. Spatial distribution of the SPEI for the period 2071-2100 and both scenarios A1B and A2 based on the EBU-POM data sets

ACKNOWLEDGMENT

Authors greatly acknowledge data provided by the Republic Hydrometeorological Service of Serbia.

REFERENCES

[1] A.A. Paulo, R.D. Rosa, L.S. Pereira, "Climate trends and behaviour of drought indices based on precipitation and evapotranspiration in Portugal," Nat. Hazards Earth Syst. Sci. 12, pp. 1481–1491, 2012.

[2] S. Sirdas, Z. Sen, "Spatio-temporal drought analysis in the Trakya region, Turkey," Hydrol. Sci. J. 48, pp. 809-820, 2003.

[3] A. Loukas, L. Vasiliades, "Probabilistic analysis of drought spatiotemporal characteristics in Thessaly region, Greece," Nat. Hazards Earth Syst. Sci. 4, pp. 719-731, 2004.

[4] S.M. Vicente-Serrano, I.J. Lopez-Moreno, A. Drumond, L. Gimeno, R. Nieto, E. Moran-Tejeda, J. Lorenzo-Lacruz, S. Begueria, J. Zabalza, "Effects of warming processes on droughts and water resources in the NW Iberian Peninsula (1930–2006)," Clim. Res. 48, pp. 203–212, 2011.

[5] V. Potop, L. Türkott, V. Kožnarová, M. Možný, "Drought episodes in the Czech Republic and their potential effects in agriculture," Theor Appl Climatol 99, pp. 373–388, 2010.

[6] M. Gocić, S. Trajković, "Analysis of precipitation and drought data in Serbia over the period 1980–2010," J. Hydrol. 494, pp. 32–4, 2013.

[7] M. Gocić, S. Trajković, "Spatiotemporal characteristics of drought in Serbia," J. Hydrol. 510, pp. 110–123, 2014.

[8] I. Tošić, M. Unkašević, "Analysis of wet and dry periods in Serbia," Int. J. Climatol. 34, pp. 1357–1368, 2014.

[9] S. Blenkinsop, H.J. Fowler, "Changes in drought frequency, severity and duration for the British Isles projected by the PRUDENCE regional climate models," J. Hydrol. 342, pp. 50–71, 2007.

[10] X. Gao, F. Giorgi, "Increased aridity in the Mediterranean region under greenhouse gas forcing estimated from high resolution simulations with a regional climate model," Global Planet. Change 62, pp.195–209, 2008.

[11] V. Djurdjević, B. Rajković, "Development of the EBU-POM coupled regional climate model and results from climate change experiments," in Advances in Environmental Modeling and Measurements, D.T. Mihailović and B. Lalić, Eds. Nova Science Publishers Inc, 2010, pp. 23-32.

[12] V. Djurdjević, B. Rajković, "Verification of a coupled atmosphere-ocean model using satellite observations over the Adriatic Sea," Ann. Geophys. 26, pp. 1935–1954, 2008.

[13] A. Kržič, I. Tošić, V. Djurdjević, K. Veljović, B. Rajković, "Changes in climate indices for Serbia according to the SRES-A1B and -A2," Clim. Res. 49, pp. 73-86, 2011.

[14] M. Hayes, M. Svoboda, N. Wall, M. Widhalm, "The Lincoln declaration on drought indices," Bull. Amer. Meteorol. Soc. 92, pp. 485-488, 2011.

[15] T.B. McKee, N.J. Doesken, J. Kleist, "The relationship of drought frequency and duration to time scales," Eighth Conf. on Applied Climatology, Anaheim, CA, American Meteorological Society, 179–184, 1993.

Fitting Theoretical Distributions to Rainy Days for Eastern Cape Drought Risk Assessment

[full paper]

Dušan Sakulski[1,2] Andries Jordaan[1], Lukić Tin[3], Cinde Greyling[1]
[1]Disaster Management Training and Education Centre (DiMTEC), Faculty of Natural and Agricultural Sciences, University of the Free State, Bloemfontein, South Africa (dsakulski2@gmail.com)
[2]Department of Environmental Engineering and Occupational Safety, Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[3]Department of Geography, Tourism and Hotel Management, Faculty of Natural Sciences, University of Novi Sad, Novi Sad, Serbia

***Abstract*— Dry spells of varying severities are regular occurrences in South Africa, with a climate that varies from sub-tropical in the eastern part to semi-arid and arid in the western part with a mean and highly variable precipitation. To have a better understanding of dry spells' spatial and temporal characteristics it is necessary, together with rainfall amount, to have information about spatial and temporal distribution of number of rainy days. Several distributions were qualified for the analysis of monthly total of rainy days for every catchment. Rice, Log-Logistic, Singh Maddala, Log-Normal, Extreme Value, Frechet, and Rayleigh probability distributions were applied to fit a rainfall data for spring, summer, autumn and winter seasons, for the selected Eastern Cape Province's catchments. Goodness-of-the-fit selection was done based on the Anderson-Darling test. Rainfall data source was the Water Research Commission's WR2000, for the period January 1950 – December 2000. As a final conclusion it can be stated that the Singh-Maddala distribution fits all-four-seasons' number of rainy days data for all Eastern Cape secondary catchments.**

Keywords*— *rainfall, rainy days, probability distributions, goodness-of-fit, Eastern Cape, dry spells.

I. Introduction and study area

The Eastern Cape is situated in the south-eastern part of South Africa, covering 13,9% of South-Africa's land cover with a size of approximately 170 000 square kilometres. This province includes coastlines, temperate forests, rolling hinterland and semi-desert landscapes. In the southern parts mountains and hills are common and in the Karoo flat topography is found [1]. The mountainous area in the north form part of the Great Escarpment and further south, between East London and Port Elizabeth, the Cape Folded Mountains start [2]. The Eastern Cape coastline of approximately 800km stretches from the north near Port Edward, down south to Tsitikamma [3].

Characterised by diverse natural beauty, the Eastern Cape houses eight of the nine biomes indigenous to South Africa: nama-Karoo, grassland, succulent Karoo, forest, savanna, fynbos, Albany thickets and wetlands [2]. The diverse fauna ranges from benthic macro-fauna, invertebrates to large ungulates [3]. The ecological diversity is also characterised by a great diversity of climates with the coastal region ranging from subtropical conditions prevalent at the KwaZulu-Natal border, to the Mediterranean climate at the Western Cape boarder. The Karoo has long hot summers and moderate winters, and the mountainous areas towards Lesotho and the Free Sate often experience snowfall in the winter [2].

With more than 300 sunny days per year, the Eastern Cape province has more sunshine days than any other province in the country. Regular rainfall is common in Great Escarpment areas and the lowland coastal belt can have rain all year round, whereas the regions west of Port Elizabeth experience winter rainfall. The Karoo receives little rain [1]. The province's average rainfall is between 808mm and 1000mm per year [3].

In a precipitation trend study for the period 1910 – 2004, Kruger [4] found that there was a significant decrease in the annual precipitation of the southeastern areas of the Eastern Cape. The northwest and northern areas showed a significant increase. Furthermore, it was found that the annual consecutive dry days during the driest season for the Eastern Cape, significantly increased. However, the winter precipitation showed no significant trends, therefore it could be concluded that the increase in consecutive dry days is due to more extreme precipitation with longer dry periods rather and a decrease in winter precipitation. Generally, it was found that daily rainfall has become more extreme.

For the period 1960 – 2010, a significant increase in rain days across the majority of the Eastern Cape confirms the possibility of more extreme daily rainfall, however, a pattern of drying was projected [5]. Johnston et al. [6] predicts an annual rainfall decrease of about 100mm in the larger part of the Eastern Cape. Whereas an increase is more likely to be expected to the east of the province [7].

II. Data and methods

Data for the analysis originate from the Water Research Commission funded study "Water Resources 2000" (WR2000) [8]. Gauging (point) stations' daily meteorological data have been used for South Africa. As a result, areal data were derived per quaternary catchment, including rainfall, maximum and minimum temperature, soil water content, soil moisture deficit, saturated drainage, vapor pressure deficit, minimum and maximum relative humidity, solar radiation and reference evapotranspiration. A fifty years' time span of the daily WR2000 data ranges from 01 January 1950 until 31 December 1999.

For the purpose of this study a number of rainy days per secondary catchment are used, based on the total number of rainy days for every month for every quaternary catchment as a sub-catchment of a particular secondary catchment, from the WR2000 database. Only days having one or more mm of rain are taken into consideration (rainfall >= 1mm).

The following continuous statistical distributions were selected for the analysis of monthly total of rainy days for every secondary catchment:

TABLE I. TABLE 1: THE SEVEN SELECTED DISTRIBUTIONS

Rayleigh (RLD)	Singh Maddala (SMD)
$\begin{cases} 1-e^{-\frac{x^2}{2\sigma^2}} & x>0 \\ 0 & \text{True} \end{cases}$	$\begin{cases} 1-\left(1+\left(\frac{x}{b}\right)^a\right)^{-q} & x>0 \\ 0 & \text{True} \end{cases}$
Rice (RCD)	Log-Logistic (LLD)
$\begin{cases} \frac{e^{-\frac{x^2+\alpha^2}{2\beta^2}}\,\text{xBesselI}\left[0,\frac{x\alpha}{\beta^2}\right]}{\beta^2} & x>0 \\ 0 & \text{True} \end{cases}$	$\begin{cases} \frac{x^{\gamma+1}\gamma\sigma^{\gamma}}{\left(1+\left(\frac{x}{\sigma}\right)^{\gamma}\right)^2} & x>0 \\ 0 & \text{True} \end{cases}$
Log Normal (LND)	Extreme Value (EVD)
$\begin{cases} \frac{e^{-\frac{(\mu+\text{Log}[x])^2}{2\sigma^2}}}{\sqrt{2\pi x\sigma}} & x>0 \\ 0 & \text{True} \end{cases}$	$\frac{e^{-e^{\frac{-x-\alpha}{\beta}}+\frac{-x+\alpha}{\beta}}}{\beta}$
Frechet (FCD)	
$\begin{cases} \frac{e^{-\left(\frac{x}{\beta}\right)^{-\alpha}}\alpha\left(\frac{x}{\beta}\right)^{-1-\alpha}}{\beta} & x>0 \\ 0 & \text{True} \end{cases}$	

Goodness-of-fit, for every catchment, for every of above probability distributions, was performed using the Anderson-Darling test, following the likelihood ratio statistics between the hypothetic distribution and the empirical distribution function. The main reason for selected Anderson-Darling test is the fact that, despite comparing fitted and theoretical (assumed) distribution, it gives more weight to the tail of the distribution then the Kolmogorov-Smirnov or Chi-Squared test [9].

The Anderson-Darling test is defined as:

H_0: The data follow a specified distribution,

H_a: The data do not follow the specified distribution,

α: Significance level,

Test statistic: The Anderson-Darling test statistic is defined as A^2=-N-S where

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N}\left[\ln F(Y_i) + \ln\left(1 - F(Y_{N+1-i})\right)\right]$$

where, F is the cumulative distribution function of the specified distribution and Yi are the ordered data.

To perform all the above-mentioned analysis a programming language Mathematica version 9 [10] was used. Mapping component was done by the Quantum GIS [11].

III. results and discussion

Figure 1 shows temporal distribution of a number of rainy days statistics, for all catchments. For the period of fifty years (1950-1999) there were no significant oscillations in the annual number of rainy days.

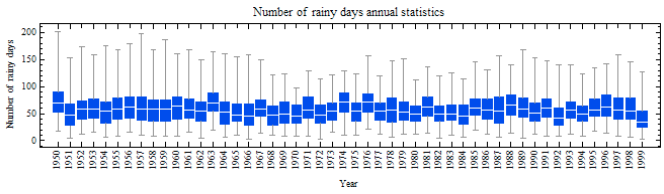


Fig. 1. Number of rainy days annual statistics for all catchments

Seasonal oscillations are shown on Fig. 2. The highest variability is for the summer months (blue colour), and partially for the spring (green colour) and autumn (brown colour), and the lowest variability is for the winter months (red colour).

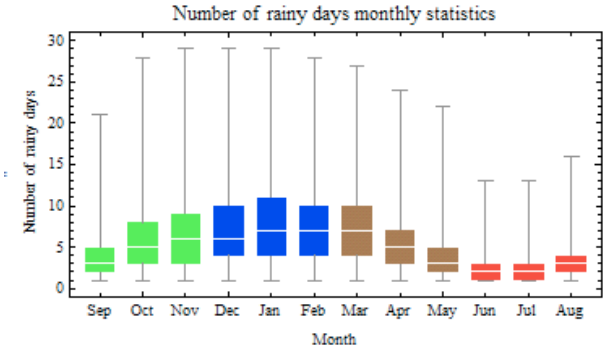


Fig. 2. Monthly statistics for the number of rainy days for all catchments

A. *Spring*

Two continuous probability distributions, Singh-Maddala and Log-Normal, are slightly more represented in fitting the number of rainy days during Spring season (September, October and November), as shown on Fig. 3, according to their maximum Anderson-Darling goodness-of-the-fit test. According to Table I, three out of seven probability

distributions significantly fit everyone of 58 catchments: Singh-Maddala, Rice, and Extreme-Value distributions. Out of those three distributions, Singh-Maddala distribution has the highest Anderson-Darling values (first quartile, median and third quartile). 75% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.793, and 50% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.9.

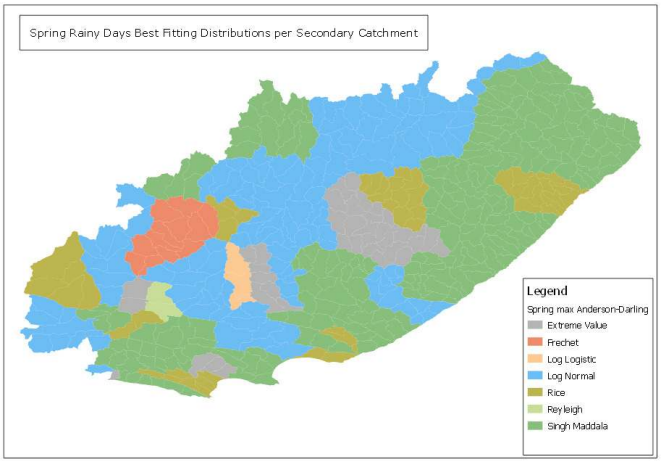


Fig. 3. Spring rainy days probability distributions having max. Anderson-Darling value

TABLE II. TABLE 2: SPRING RAINY DAYS ANDERSON-DARLING VALUES STATISTICS

No. of Catchments	Distribution	First Quartile	Median	Third Quartile
58	Singh-Maddala	0.793	0.908	0.976
56	Log-Normal	0.571	0.847	0.923
58	Rice	0.65	0.825	0.932
58	Extreme-Value	0.494	0.773	0.88
16	Log-Logistic	0.512	0.644	0.838
49	Frechet	0.127	0.273	0.446
18	Rayleigh	0.174	0.227	0.531

As shown on Fig. 3, seven probability distributions have fitted to rainy days data. For the practical purpose, question is can, for some catchments, second best distribution replace best-fit distribution, but still have a high Anderson-Darling value? As from Table I, the best candidate is a Singh-Maddala distribution.

Figure 4. shows an Anderson-Darling goodness-of-the-fit values for the Singh-Maddala distribution, for the Spring season. Only 25% of all catchments have goodness-of-the-fit values below 0.793, and the Median value is very high (0.908). There is enough reasons to conclude that, for the practical purposes, a Singh-Maddala distribution fits Spring number of rainy days data with high confidence.

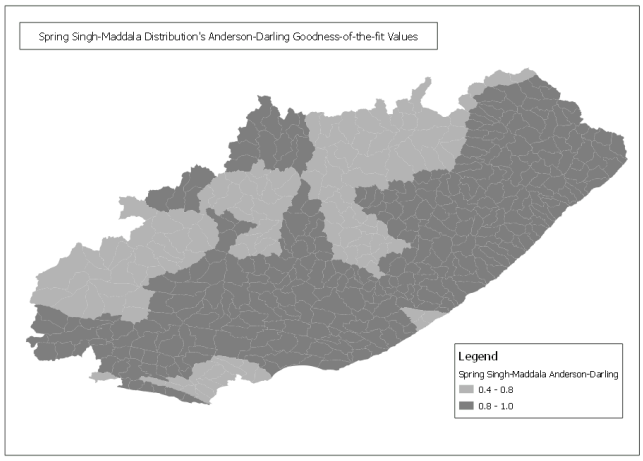


Fig. 4. Spring Singh-Maddala goodness-of-the-fit values

B. *Summer*

During summer season (December, January, February) two distributions dominate in fitting number of rainy days: Singh-Maddala and Rice distributions, according to their maximum Anderson-Darling coefficient values (Fig. 5). According to Table II, four out of seven probability distributions significantly fit everyone of 58 catchments: Singh-Maddala, Rice, Log-Normal, and Extreme-Value distributions. Out of those four distributions, Singh-Maddala distribution has the highest Anderson-Darling values (first quartile, median and third quartile). 75% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.878, and 50% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.92.

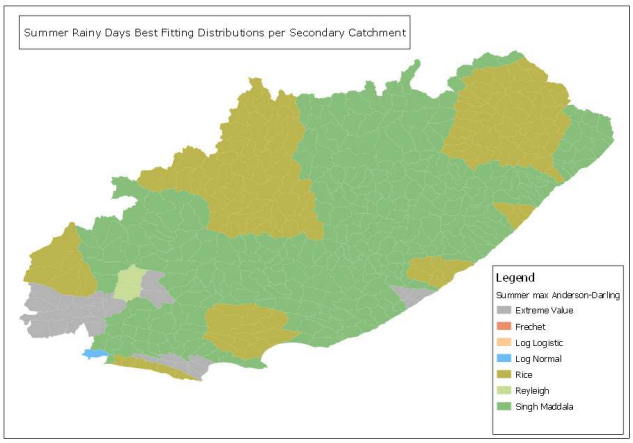


Fig. 5. Summer rainy days probability distributions having max. Anderson-Darling value

TABLE III. TABLE 3:SUMMER RAINY DAYS ANDERSON-DARLING VALUES STATISTICS

No. of Catchments	Distribution	First Quartile	Median	Third Quartile
58	Singh-Maddala	0.878	0.92	0.979
58	Rice	0.738	0.901	0.967
9	Log-Logistic	0.69	0.712	0.89
58	Log-Normal	0.382	0.607	0.767
58	Extreme-Value	0.389	0.563	0.869
19	Rayleigh	0.143	0.231	0.8
48	Frechet	0.067	0.108	0.187

Figure 6. shows an Anderson-Darling goodness-of-the-fit values for the Singh-Maddala distribution, for the Summer season. Only 25% of all catchments have goodness-of-the-fit values below 0.878, and the Median value is very high (0.92). There is enough reasons to conclude that, for the practical purposes, a Singh-Maddala distribution fits Summer number of rainy days data with high confidence.

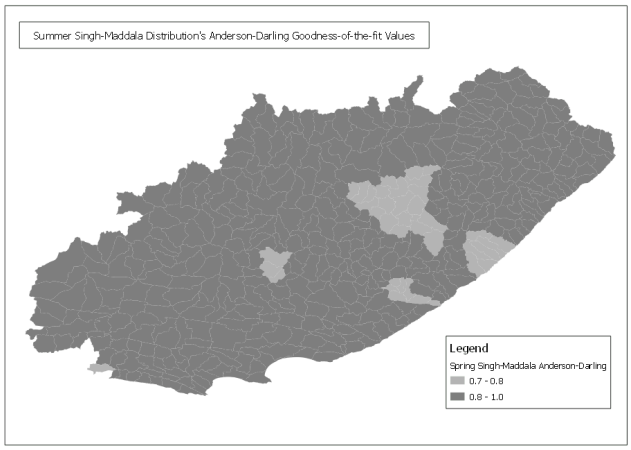


Fig. 6. Summer Singh-Maddala goodness-of-the-fit values

C. Autumn

A Singh-Maddala distribution is dominant in fitting the number of rainy days during Autumn season (March, April, and My), as shown on Fig.7, according to the Anderson-Darling goodness-of-the-fit test. Other three distributions (Rice, Extreme-Value, and Log-Normal) also fit everyone of 58 catchments, but their Anderson-darling goodness-of-the-fit test values are lower than Singh-Maddala distribution, acoording to Table III (75% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.82, and 50% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.92).

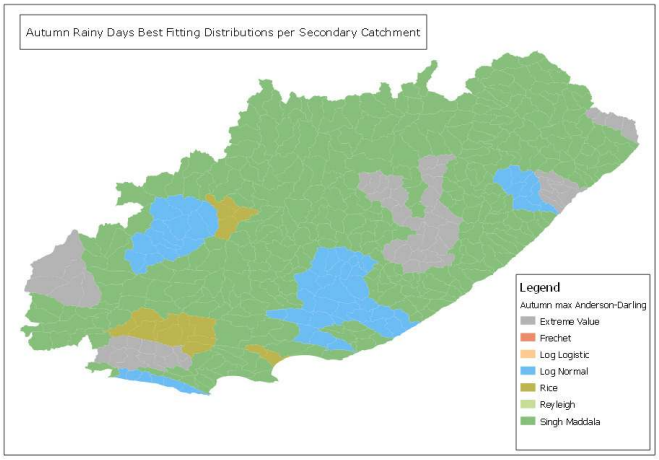


Fig. 7. Autumn rainy days probability distributions having max. Anderson-Darling value

TABLE IV. AUTUMN RAINY DAYS ANDERSON-DARLING VALUES STATISTICS

No. of Catchments	Distribution	First Quartile	Median	Third Quartile
58	Singh-Maddala	0.821	0.924	0.982
58	Rice	0.689	0.812	0.935
58	Extreme-Value	0.714	0.805	0.948
58	Log-Normal	0.651	0.799	0.917
8	Log-Logistic	0.613	0.69	0.809
54	Frechet	0.163	0.226	0.412
22	Rayleigh	0.085	0.109	0.181

Figure 8. shows an Anderson-Darling goodness-of-the-fit values for the Singh-Maddala distribution, for the Autumn season. Only 25% of all catchments have goodness-of-the-fit values below 0.821, and the Median value is very high (0.924). There is enough reasons to conclude that, for the practical purposes, a Singh-Maddala distribution fits Autumn number of rainy days data with high confidence.

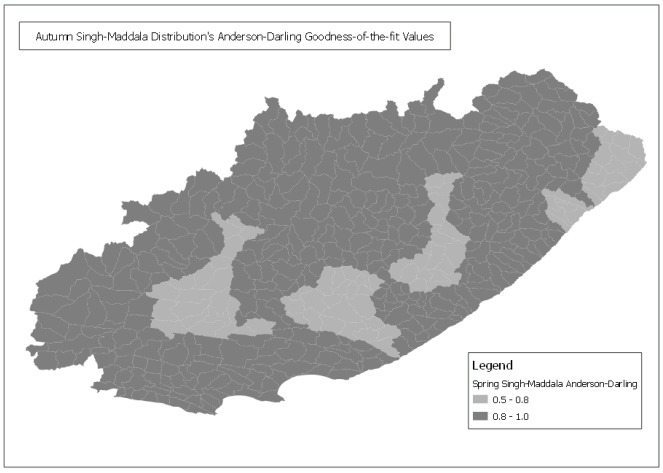


Fig. 8. Autumn Singh-Maddala goodness-of-the-fit values

D. Winter

During winter months (June, July, and August) a dominant distribution fitting number of rainy days is Singh-Maddala, followed by Log-Normal distribution (Fig. 9). According to Table IV, four out of seven probability distributions significantly fit everyone of 58 catchments: Singh-Maddala, Rice, Log-Normal, and Extreme-Value distributions. Out of those four distributions, Singh-Maddala distribution has the highest Anderson-Darling values (first quartile, median and third quartile). 75% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.88, and 50% of catchments have the Anderson-darling goodness-of-the-fit parameter above 0.96.

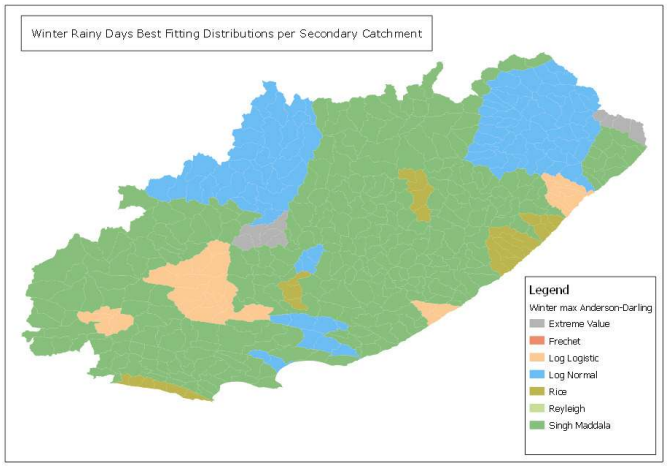


Fig. 9. Winter rainy days probability distributions having max. Anderson-Darling value

TABLE V. WINTER RAINY DAYS ANDERSON-DARLING VALUES STATISTICS

No. of Catchments	Distribution	First Quartile	Median	Third Quartile
58	Singh-Maddala	0.888	0.956	0.975
40	Log-Logistic	0.691	0.842	0.93
58	Extreme-Value	0.658	0.824	0.917
58	Log-Normal	0.59	0.773	0.917
58	Rice	0.548	0.748	0.947
30	Rayleigh	0.085	0.235	0.44
54	Frechet	0.098	0.2	0.337

Figure 10. shows an Anderson-Darling goodness-of-the-fit values for the Singh-Maddala distribution, for the Winter season. Only 25% of all catchments have goodness-of-the-fit values below 0.793, and the Median value is very high (0.908). There is enough reasons to conclude that, for the practical purposes, a Singh-Maddala distribution fits Winter number of rainy days data with high confidence.

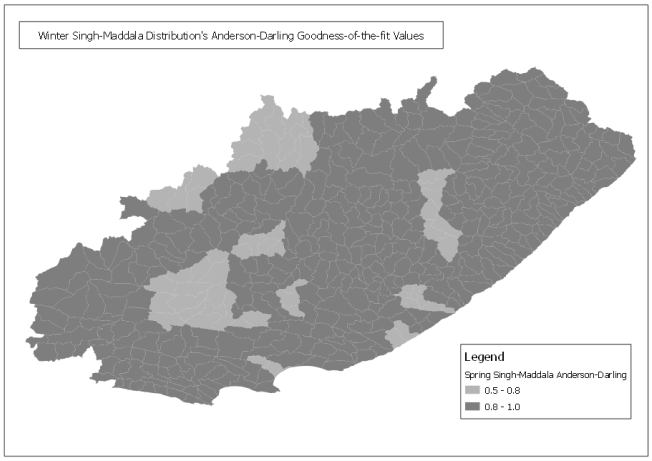


Fig. 10. Winter Singh-Maddala goodness-of-the-fit values

As a final conclusion it can be stated that the Singh-Maddala distribution fits all-four-seasons' number of rainy days data for all Eastern Cape secondary catchments.

References

[1] Innovation Eastern Cape. Provincial growth and development plan: Eastern Cape. [Online] Available: http://www.innovationeasterncape.co.za/download/ec_growth_dev_plan_2004_2014.pdf [5 June 2014].

[2] M. Hamann, and V. Tuinder, Introducing the Eastern Cape: A quick guide to its history, diversity and future challenges. Stockholm Resilience Centre: Stockholm, 2012.

[3] Department of economic affairs, environment & tourism. Environmental implementation plan: Eastern Cape Province. [Online] http://www.enviroleg.co.za/acts/National%20Environmental%20Management/REGS/151-03%20Eastern%20Cape%20EIP.pdf [5 June 2014].

[4] A.C. Kruger, Observed trends in daily precipitation indices in South Africa: 1910 – 2004. *International journal of climatology*, 26 pp. 2275-2285. 2006.

[5] Department of Environmental Affairs, Full technical report on climate trends and scenarios for South Africa. [Online] Available: https://www.environment.gov.za/sites/default/files/docs/climate_trends_scenarios.pdf. June 2014

[6] P. Johnston, S. Hachigonta, L.M. Sibanda, and T.S. Thomas, Southern African agriculture and climate change: A comprehensive analysis – South Africa. [Online] Available: http://www.ifpri.org/sites/defaults/files/publications/aacccs_southafrica_note.pdf [10 June 2014].

[7] Coastal and Environmental Services. Eastern Cape climate change response strategy. Department of Economic Development and Environmental Affairs: East London. 2014.

[8] Water Research Commission of South Africa. "Water Resources 2000" (WR2000)

[9] J. Zhang and Y. Wu, Likelihood-ratio tests for normality, Computational Statistics & Data Analysis. 49, no. 3, 709-721, 2005.

[10] Wolfram Research: Mathematica. (www.wolfram.com).

[11] Quantum GIS, an Open Source GIS software (www.qgis.org).

Use of Bio-Physical Indicators to Map and Characterize Coping Strategies of Households to Rift Valley Fever Outbreaks in Ijara District

J. Kiplimo
International Livestock Research Institute, P. O. Box 30709-00100, Nairobi, Kenya
Jomo Kenyatta University of Agriculture & Technology
Department of Geomatic Engineering & Geospatial Information Systems, Nairobi, Kenya
j.kiplimo@cgiar.org

A. Notenbaert
International Livestock Research Institute, P. O. Box 30709-00100, Nairobi, Kenya

B. Bett
International Livestock Research Institute, P. O. Box 30709-00100, Nairobi, Kenya

Waithaka H.E.
Jomo Kenyatta University of Agriculture & Technology
Department of Geomatic Engineering & Geospatial Information Systems, Nairobi, Kenya

Abstract— Extended and above normal rainfall across (semi-) arid Africa after a warm phase of El Niño creates conditions favorable for outbreak of Rift Valley Fever (RVF); a vector borne disease. There have been two major epizootics in the Horn of Africa in 1997/98 and 2006/07 that had the highest mortality to both humans and livestock.

Our study, at Ijara District- Kenya, characterizes coping strategies used by communities in high and low risk areas to make themselves less vulnerable to effects of adverse climate variability and in consequent RVF outbreaks.

RVF outbreaks resolved to division level werecollated to bio-physical factors that were significantly associated with the outbreaks. Geostatistical analyses were used to identify RVF risk areas. At selected risk areas, focus group discussions (FGDs) involving the local communities, community health workers, and veterinary officers were used to characterize coping strategies that were employed in recent RVF outbreak. Solonetz, luvisols and vertisols and areas below 1000m were significant. Low areas, fairly flat with a 0 – 15% slope rise having these soil types have higher risk compared to the other areas. The low and high RVF risk areas were approximately split halfway across district, northwards and southwards respectively.

From the FGDs, actions taken by communities at high risk areas were strategic while those at low risk areas used reactive, ad hoc coping strategies.

Communities at high risk areas would cope better to adverse climate variability and extended disease burden compared to those at low risk areas who lack knowledge of some of those strategies.

More needs to be done to understand climate variability, disease ecology of RVF, community awareness and facilitationas there are at times the whole district is affected by the RVF.

Keywords— Rift Valley Fever (RVF), Coping Strategies, Geostatistical Analysis, Focus Group Discussion (FGD), Communities, Risk

I. INTRODUCTION

Coping refers to the ability of people, organizations, systems or communities, using available skills and resources to face and manage (or strategize towards) adverse conditions arising from endemic and epidemic diseases, emergencies or disasters. Coping strategies are engaged as a livelihood diversification by many rural communities; pastoralists and agro-pastoralists (in the arid and semi-arid regions) included at certain periods of the year to avert extensive losses in main sources of livelihood. Some of the periods that cause need of development of coping strategies are the unreliability and uncertainty of climate, and increased disease burden that would greatly impacts ecosystem services, farming and livestock keeping which are the main sources of income and food to many in the rural areas [1]; [2].

Our study area Ijara District, Kenya is part of the Arid and Semi-Arid Lands (ASAL) which is mainly inhabited by pastoralists. The pastoralists (mainly livestock keepers of cattle, shoats and, donkeys) here have overtime had a livelihood dependent on the intricate knowledge they have gained to adapt and sustain equilibrium in the natural and socio-economic environment[3].

Literature [3];[4]reveals that the extent and exercise of knowledge however is dependent on coping strategies/ household characteristics for timely decisions in climate uncertainty and outbreak of diseases, diversity of livestock (some animals are hardier than others), mobility (ability to utilize the rangelands and other eco-system services without depletion), diversification of livestock species and breeds (so as to have a wider range of animal products, even to sale in terms of crisis), maximization of stock numbers (helps ensure survival despite loses during droughts and disease outbreaks) and splitting or redistribution of herds (so that in the event of extensive losses one can recall from the other members).

This study focuses on characterization of some of the coping strategies that households have practiced over time to make themselves less exposed to adverse climate conditions and be at a less risk of getting exposed to Rift Valley Fever (RVF). RVF is a severe viral disease that over time has been observed to cause huge socio-economic losses and death to both humans and animals in hundreds of thousands ([5]; [6]. There are some bio-climatic and bio-physical indicators that have been found to be related and also to be drivers of the RVF disease. By way of geostatistical analysis this risk factors were analyzed to develop a risk map that depicted areas at high risk and low risk of RVF within our study area. Coping strategies at these areas were elicited by way of focus group discussions.

II. LITERATURE REVIEW

RVF is a viral disease that is acute, causing fever in both animals (such as cattle, buffalo, sheep, goats, and camels) and humans. The disease is associated with mosquito-borne epidemics during years of prolonged and above normal rainfall with extensive localized flooding. The flooding allows buried mosquito eggs, usually of the genus Aedes, to be brought to surface to hatch. The eggs having been infected from previous RVF outbreak hatch into mosquitoes that transfer the virus to the livestock they feed on. As other genus of mosquito populations and those uninfected build up, the virus spread is amplified once they start to feed on the infected livestock and bite other uninfected livestock plus humans. Humans could also acquire the disease when they are exposed to blood or other body fluids from RVF infected animals. Exposure of this kind may happen in the cause of tending to animals delivering, slaughtering, sickly or ingesting RVF contaminated meat or milk [7]; [8].

RVF primarily affects livestock and can cause deaths in large numbers; a situation referred to as an epizootic as was the case in 1950/51 where there was an estimated death in hundreds of thousands of sheep. The RVF virus can at times lead to an epidemic among people exposed to animals with RVF. Most people do not exhibit RVF symptoms while others have been observed to experience mild illness associated fever and liver abnormalities. Some people experience more serious symptoms like fever, weakness, back pain, dizziness, and others suffer extreme weight loss. However, in some advanced cases victims suffer hemorrhagic fever, encephalitis (inflammation of the brain) that at times led to seizures, coma, and eye disease that in some few cases have led to permanent eye-sight loss [8].

The climatic factors which predispose the emergence of RVF from the areas in which the infection persists during inter-epidemic periods have been well documented [9]. Epidemics always follow heavy and prolonged, often unseasonal, above normal rainfall. Such were the conditions that occurred prior to the 1997/98 outbreak in the arid/ semi-arid East Africa, in association with the El Niño event that resulted in a serious epidemic. The outbreak was of a huge geographic extent and impact as it led to ban on livestock trade across all countries in the Horn of Africa. From the 1997/98 outbreak more research has been put to understand the disease drivers and dynamics surrounding the RVF disease and outbreak. Some of these primary indicators have been useful in the development of early forecasting systems that could help set up measures that would reduce, or in some instances avert impending epidemics [9].

Rift Valley Fever outbreaks occur several years apart, irregularly, but when it does occur losses are counted in thousands of livestock in the arid and semi-arid areas where farmers have little capacity of buying the vaccine and in time before the outbreak[10]. In most cases the farmers lack ability to identify symptoms of the RVF and by the time they get to know there is an outbreak a month after, thus administered vaccination helps very little. Some drugs like Ribavirin have shown hope in treatment of RVF in animals already exposed. Very few pastoralists are able to afford or access the drug [11].

At onset of RVF outbreak, some households resort to selling most of their livestock to avert losses. This is in retrospect makes the actual prices for the livestock actually plunge as there are many livestock suppliers. The resultant value greatly undermines the household coping strategy to fend for family and with very little chance also of being able to restock same number of livestock after the outbreak. Media coverage of human suffering and livestock loss in the period of outbreak greatly affects the image of people and society of those affected. After the outbreak it will take a longer period for people to readily trade and interact with those from the affected areas thus jeopardizing their development and recovery[12].

Global climate change comes with varied environmental stresses/ shocks that make communities vulnerable at various points of getting exposed dependent on how well one is prepared to face the climate variability ([14][15];[16][17];[13];[15]) observe that different shocks require different coping reactions and strategies. This has been seen in many communities where there is difficulty in differentiating active coping strategies (like increase in home food production, vaccination, restricting livestock movement, obtaining of a supplementary job and formal borrowing) and weak coping strategies (like sale of assets, introducing child labor, reducing/ restriction of food intake, reducing education expenses, postponing healthcare, restructuring family to live in a smaller house, reliance on external support e.g. friends and family, NGOs).

Weak coping strategies may offer instant solution in light of impending need but in the long run seen as a poverty trap that a family/ community might fall into. The ability of a household to adopt active coping strategies majorly dependent

on the household characteristics to differentiate those that are poverty driven and those that comes from shocks. There is a lot to be done in quantification and understanding of coping strategies amongst pastoral and agro- pastoral households. More needs to be done in terms of understanding the existing environments, climate and diseases, livelihoods, structure and characteristics of households and their choice of coping strategies in view of the imminent uncertain future[2].

III. METHODS AND MATERIALS

Research design and methodology for the study involved identification and mapping of bio-physical factors that are significant to outbreak of Rift Valley Fever (RVF). The mapped places provided areas that are at risk of RVF. Characterization of coping strategies was done by way of focus group discussions at the RVF risk areas, see Figure 1. Characterization was done by employment of in-depth participatory epidemiology techniques and methods. Participatory Epidemiology (PE) uses a combination of practitioner communication skills and participatory methods to improve involvement of animal keepers in the understanding and analysis of their livelihoods, animal disease problems, design, implementation and evaluation of disease control programmes and policies[18].

Contrary to conventional data collection methods, PE has gained popularity in that with limited research time, financial resources and well trained personnel in rural settings it is possible to prod for common issues and come up with general perceptions for action oriented research to common challenges facing a community. The approach used for our study was use of focus group discussion that was semi structured discussion made around guided conversational themes.

Main tools used in analysis of the project work are R statistical software[19] and ArcGis10x[20]. The R package provided a platform where it was possible to identify which of the bio-physical variables are significant to outbreak of the RVF cases in Kenya divisions. The significant variables were input in ArcGis10x where a friction data of the combined variables was created. The friction data output was a risk map which was then reclassified to depict areas of high risk and low risk across the study area, see Figure 1. Within Ijara district, locations with relative high and low probability were identified.

These risk map areas informed where focus group discussions were to be held in order to characterize the coping strategies which communities used. FGDs were held in both high and low risk areas in Ijara District. The group sessions were noted to a book and flip chart, recorded with a voice recorder, GPS location of the FGD taken, and also photos. The FGD data for each group was then tabulated in Microsoft Excel 2010 to decode and understand the coping strategies amongst the different communities.

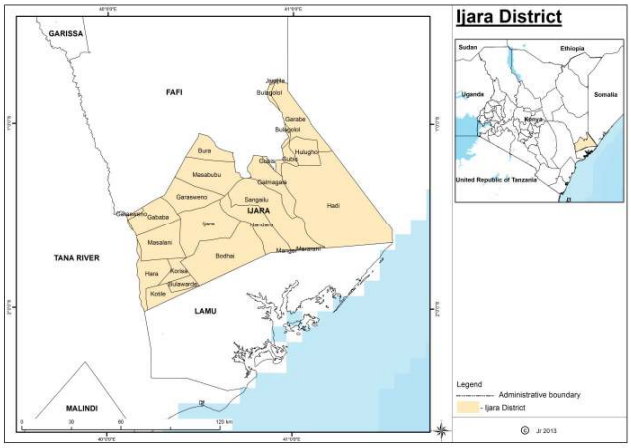


Fig. 1. Study Area- Ijara District zoomed in. Ijara is one of the eleven districts in North Eastern Province, Kenya

A. *Characterization of the study site*

Latest bio-physical data (landcover, soil and elevation) was prepared in ArcGis to ensure the data was in one geographic coordinate systemand in one kilometre resolution for the whole country. Landcover data was obtained from the Global Landcover (GlobCover) 2009 database. The GlobCover 2009 was aninitiative of the European Space Agency that was generated from the Envisat MERIS sensor that is able to produce fine scale data in the year 2009. The landcover classes considered for this analysis were twenty two in total, consistent with the UN Land Cover Classification System (LCCS) [21].

Soil data was obtained from the Harmonized World Soil Database (HWSD) which is a comprehensive database of the world for all soil types as per current FAO classification system [22]. Elevation data was obtained from the Shuttle Radar Topographic Mission (SRTM) 90m [23]. Rainfall data was from TRMM for the period December 1997 to July 2013. The data is a satellite product of NASA and JAXA services that was launched in November 1997 to provide rainfall data at a resolution of 0.25degrees[24][25][26][27].

Records of RVF outbreak cases as confirmed by the Department of Veterinary Services (DVS) for the period 1912 to 2007 were obtained from the DVS and Centers for Disease Control (CDC), [28]. All the outbreak cases can be traced to the province, district and division of outbreak but hardly georeferenced to the epicenter of the outbreak. An inventory of the confirmed RVF outbreak cases traced and reported in all locations, hospitals were aggregated to division level. The division thus was settled for as the unit of analysis. Irrespective of the number of cases, both in human and animalsa division reported, it was resolved that a division with an outbreak was a ‘case’ and no outbreak as ‘no case’. It was further noted that even if a division reported a single outbreak over 30 years then it was evidence enough that the division has conducive ecological conditions to facilitate a future outbreak.

The division being unit of analysis, spatial analyst within ArcGis10x was used to perform summary statistics on the bio-physical indicators so as to determine by way of majority

coverage which feature type was most dominant. The dominant feature was then assigned to the division. An entry of all the divisions by case and bio-physical data was then prepared for input for further analysis in R statistical software. Geostatistical Analysis

B. Geostatistical Analysis

The aim at this stage was to determine the factors that affect the probability of RVF occurrence. The R package was chosen for its capability to support analysis of discrete data and continuous data. Discrete data for this case was report of outbreak cases in each of the divisions while continuous data was the biophysical data. This was a very powerful feature of the software as it gave opportunity to identify which of the bio-physical indicators were significant for there to be an outbreak in the divisions that reported RVF cases. Generalized Linear Models(GLMs) offer a weighted linear regression that can be used to obtain parameter estimates of non- linear observations as is the case of 'our' RVF outbreaks. A variety of GLMs exist like the 'normal', 'binomial', 'poisson' and 'gamma'[29]. Generalized Linear Model (GLM) with a Logit link was used for this analysis. The model suited our study as the cases of RVF at the division level were binomial; 'present' or 'absent'. The logit link in the model allows for regression of presence/ absence to the explanatory variables; bio-physical data that are qualitative data.

A four stage approach was used; all the bio-physical data was run using the model to see which was significant. An auto-correlation test was run at this stage to check for bio-physical data that significantly influenced each other. Single data was then run using the model to see which class within the biophysical data was quite significant to occurrence of the cases.

C. RVF Risk Map

From the geostatistical analysis it was possible to make observations on which of the bio-physical data and of which type are quite significant to outbreak of RVF. The bio-physical data were then pre-coded for a risk index e.g. 1, 2, 3. Where 1 was for areas that probability of RVF risk was high, 2 for lower risk probability and so forth. ArcGis10x environment was used to pre-code the data. Map algebra was then utilized in ArcGis to add the data together. Where most significant data, for instance areas that had 1s, emerged as areas that had high probability risk of RVF. Those areas that had low significant values intersecting made the low probability areas of RVF. The resultant output is the 'friction map 'or probability map of RVF risk.

An RVF risk map was then created by this reclassification of high and low risk. The high and low risk divisions in the area of study thus informed where the focus group discussions would be held. The location to carry out the FGDs was agreed with the help of key informants at the area of study which locations have over time been considered by both government and community as the high and low risk areas.

D. Focus Group Discussions

The FGDs were used in the study as a way of characterizing coping strategies that households used in their livelihoods. In this study, FGD was a discussion of at most 10 local people, the interpreter and moderator. The discussion generally started by way of prayer, introduction of all those in attendance then moderator would set the agenda for discussion. This was then followed by development of seasonal calendar of the past and current climate perceptions to the community. Seasonal calendars helped the FGD participants contextualize occurrences within a calendar year.

After the seasonal calendar was agreed on amongst the FGD members, the discussion went on into a guided discussion for a maximum of an hour. A checklist was used to elicit issues that were pertinent to livelihoods, common livestock economically and their importance, common livestock diseases that have most impact and those that they deem to be climate driven, some of the coping strategies that the community has used, barriers that deter coping, and measures that have been put prior and after a disease outbreak. Issues raised were noted, and pair wise comparison used to determine what was fairer, important, common or of most impact to their livelihoods. Ranking was then used to determine order of priority with the first position given to top of the issues and the last to depict the one with lowest impact or influence. The discussion amongst the FGD members was left to go on until they reached a consensus of what position or items to consider for a topic of discussion. Items that went into pair wise comparison were kept to a maximum of 9 at any time. The discussion was then wrapped up by thanking all those in attendance, commentary of any other issues directly or indirectly related to the agenda, followed by a photo and GPS position of the group being taken.

E. FGD Data Analysis

The booked data from FGDs were then tabulated into Excel. The seasonal calendar was tabulated with the corresponding local names alongside. Climate related diseases from communities perceptions were also tallied to corresponding calendar times. The rest of the issues that came up from the guided conversation were also tabulated with their respective rank positions. The tabulated results gave different insights in terms of priority and action as to how different issues were approached by communities.

IV. RESULTS AND DISCUSSION

Most of Ijara is covered by rainfed croplands (6%), and mosaicked open/ closed grass lands and shrub lands (48%), and broadleaved forest vegetation. The urban and bare areas cover a very small area of the whole district, barely 1% of the district.

The main soil types in Ijara are solenetz, planosols and vertisols and have coverage of approximately 42%, 38% and 17% respectively. The main soil texture types are clay, loam, and sandy. The heavy clayey soil is approximately 60% and 38% of clay loam of the district.

The elevation in Ijara district varies from 0 – 90m above sea level with the elevation gradient rise from south east westwards to the north. Majority of the area is fairly flat with many areas having a maximum of 0 to 15% slope rise covering 72% of the whole district.

The RVF cases were extracted from reports obtained from the Department of Veterinary Services in Kenya that were reported at division level for the period 1912 to 1950. Over 30 of the total 47 districts in this time have reported RVF outbreaks, which is over 63% of all the districts. It is also observed that the number of RVF outbreaks in the last 3 decades have had high impact outbreaks across many divisions with more areas reporting the diseases. The peeks in this period of outbreaks starting from recent are 2006 – 2007, 1997 – 1998, and 1989 – 1990.

TRMM being lowest resolution available at 0.25 degrees for the period December 1997 to July 2013 in netdcf. It can be observed that the easterly part of Ijara is wetter compared to the west with rainfall of over 2mm per day. The data is able to capture the rainfall periods ending January, short rains April – May. The late 2006 early 2007 RVF outbreak coincides with these rainfall periods.

The longest section across Ijara was just over 1degree in distance thus resolving the analysis to this level was a compromise to variability anticipated. Hence as much as rainfall was quite significant (0.04) to the model it was dropped from further analysis in the model. Some landcover type was also significant (0.15) but data available was for the period 2009, and that would not be reflective of the vegetation during the 2006 – 2007 outbreak.

The data thus retained for analysis in the GLM were soil and elevation, see equation 1.

```
Glm ( model;Cases = Soil + Elevation, family=
binomial(link = "logit"))
```

Eq. 1. GLM Model

The model output showed that when the soils; vertisols, solonetz, and luvisols were compared to the other soils they had a higher coefficient to explaining the RVF cases, see Table 1. The solonetz and luvisols had almost the same coefficient of 1.2 while vertisols were lower at 0.66. Areas of elevation between 1000 to 2000m were also compared to areas below 1000 and those above 2000m. The areas below 1000m had a higher coefficient (2.74) to those at a higher altitude (-1.99), see table 1. This result helps confirm that the RVF cases that occurred in these areas could best be explained by these soil types and the elevation.

From the results, solonetz, vertisols and luvisols covered mid-way across Ijara with the 'Other' soil types covering the rest of the district. Areas that had solonetz and were of less than 1000m were areas of highest risk, areas with vertisols then luvisols followed in decreasing risk while areas that had 'Other soils' types in the district were of a lower risk, see Figure 4.

TABLE 1. Bio-physical Risk Factor Coefficients from GLM model

Variable	Levels	Generalised linear model	
		Coefficients (β)	Standard Deviation (SE) of (β)
Soil type	Vertisols	0.66	0.31
	Solonertz	1.19	0.44
	Luvisols	1.21	0.58
	Others	0.00	-
Elevation	0 -1000	2.74	0.38
	>1000-≤2000	0.00	-
	>2000	-1.99	0.48

Four FGDs were successfully held in Ijara District. From the RVF risk map obtained, the locals were consulted as to which locality to carry out the FGDs. It was observed that the locals' perception to areas at high risk and low risk of RVF agreed with our analysis. Two FGDs were then carried out at high risk (Sangailu and Kotile) and at low risk (Hajmahamad and Galmatha; also known as Falmata) divisions, see Figure 2. The FGDs were done with the help of a translator who understood the local language, Somali.

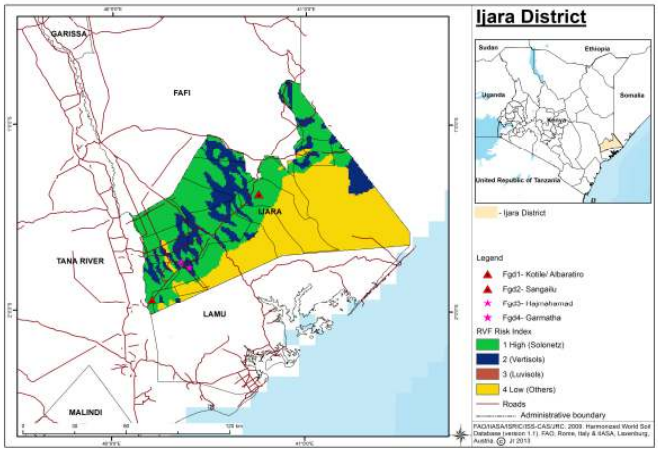


Fig. 2. RVF Risk Map & FGD location- High Risk Areas are halfway northwards of the district. Low risk areas having 'other' soil types are midway south and to the east. The high and low risk locations agreed on by community coincide with Risk Map.

Coping strategies were looked at as those initiatives taken by households prior to a season start so as to reduce risk of getting livestock exposed to some of the livestock diseases. FGD1s mechanisms in order of priority were separation of infected animals, prayer, early treatment of sickly and vaccination, early deworming, and splitting or sharing livestock to friends and family in other areas to reduce impact. FGD2s coping strategies were vaccination, prayer, separation of infected animals, securing nets for the young, early treatment of the sickly animals, taking extra care of livestock left, split any other extra livestock to family and friend for keeping, evacuation of area that disease has broken out, and early communication of diseases to the government veterinary

department. FGD3 separated animals, vaccination, early reporting to government, prayer, moving uninfected animals to safer areas, split livestock, provision of nets to the young. FGD4s mechanisms were to perform early treatment, burn bushes to destroy vector environments, separate animals, prayer, and preparation and separation of money for treatment of livestock.

Barriers that impeded coping of households or a community were identified as those that could undermine the measures taken from being effective and compromises both livestock survival and human health. Lack of requisite knowledge to diagnose diseases was cited as one of the issues impeding coping. FGD1s issues that impeded coping were if the disease was internal and could not be detected early, some disease symptoms are not well known, inaccessibility of medical services to the livestock early enough. FGD2 mentioned that some drugs seem to no longer work or effective when administered to the livestock, there are some new diseases which they have no mastery of the symptoms, trial and error treatment of diseases many animals and inaccessibility of livestock services early enough. FGD3 observed that there seem to be some diseases that have become drug resistant (e.g. Trips), some drugs no longer work when administered, lack of knowledge of new diseases, and lastly inaccessibility of treatment early enough. FGD4 mentioned that at times the spread of the disease is too fast for them to perform timely treatment, late access of drugs pauses many livestock to risk of exposure to disease, and trial and error treatment when the disease is not known.

These are discussions drawn from tabulated results from the FGDs. The statements are informed by rank across the topic discussions across different FGDs which help in further understanding of some of the issues. For example, which of the livestock is most affected by the pertinent issues? Which of the climate driven diseases is common, and which livestock would be most affected? Are some of the measures taken prior and after disease outbreak undermined by the barriers that impede coping? Are some of the coping strategies related to solving the pertinent issues? Are some of the issues common and handled in a similar way by all the groups?

The pertinent issue common to all the groups is livestock diseases, followed by wild animal attacks for FGD3 and FGD4, famine and flooding for FGD1 and FGD2 respectively. The livestock most affected by these issues common to both groups are cattle, sheep and goats. The common livestock disease and of most impact to their livelihoods is different across all groups. All the groups have different opinions of which diseases are closely related to climate patterns.

Some of the issues thought to impede coping could adversely undermine some of the strategies that the communities have put in place. The groups cite cases where the disease occurs internally in the animals without exhibiting external symptoms on the livestock, community lacks requisite knowledge of the disease and it spreads fast, and inaccessibility of some areas due to flooding pose a great challenge to evacuating people, animals and in administering treatment on time.

Prayer was appreciated to be very important in many of the actions and decisions taken by community towards challenges faced. Irrespective of the strategy the communities sought they prayed for the action to be blessed. Similar to the actions taken prior and after the outbreak; the communities' choice to vaccinate livestock, evacuate areas prone to flooding, culling of sickly animals and early reporting to government, is that they hope these actions will be blessed.

From the FGDs it was observed that actions taken by communities at the high risk areas were strategic as they took actions like vaccination, controlled animal movement with awareness to changing climate and disease incidence. Those at low risk areas used reactive, ad hoc coping strategies like treatment of sick animals' as opposed to vaccination. It was thus observed that those communities at the high risk areas would cope better to adverse climate variability and extended disease burden compared to those communities at low risk areas who lack knowledge of some of those strategies.

V. CONCLUSION AND RECOMMENDATION

This study has clearly shown and given a detailed inventory of how GIS techniques and methods can be used to characterize and understand coping strategies exercised by communities at different areas when faced by different climate and diseases. It has been possible to identify bio-physical factors that are quite significant to outbreak of RVF. By geostatistical analysis, a generalized linear model (glm) was used to differentiate the RVF risk index across Ijara district. Soils identified as significant occur in most of the flat areas. When there is extended flooding there would be pools of water suitable for mosquito breeding making these potential hotspots for RVF. These factors are corroborated from other researches that find them quite significant to outbreaks of RVF cases in many low lying areas and are of soil types that have high water retention capacity.

The high and low RVF risk areas were approximately split halfway across the whole district, northwards and southwards respectively. Both FGD and key informant discussions noted that the RVF outbreak always starts in the flood prone areas (for example Sangailu, Albaratiro divisions) then spreads to the rest of the district. Procedures and policies should be made to facilitate containment and support to communities at these places both in access to vaccination and evacuation.

The whole district experiences extended conditions of rainfall events and flooding that are driver conditions for outbreak of RVF as evidenced by the increased rainfall in the 2006 – 2007 period. The climate change profile observed by McSweeney[30] is there would be a likely increase in mean temperatures and rainfall over many areas of Kenya. It being subject to El Nino patterns then more research needs to be done to understand ecology of RVF in this climate outlook. This in effect means that all communities faced by these challenges now and later, those at low risk areas will be most affected as they are unaware of what strategies to use.

It also emerged from the FGD discussions that some of the (re-)actions (like vaccination, deworming, and early treatment of animals) taken by households could be considered as a form of coping strategy. It is there perception that if livestock are

healthy and free of diseases then they stand healthier in the face of an RVF outbreak. How well this works should be investigated so that they can be advised accordingly. From the discussions, and also with key informant, it emerged that more needs to be done to understand climate variability, disease ecology of RVF, community awareness and facilitation (information sharing and empowerment of policy makers).

Other geostatistical models that are able to consider for dynamic factors like rainfall and NDVI (at a low resolution approx. 5km or less) should be explored. This will make it possible to refine the RVF risk map as outbreaks (especially those sporadic) can be factored into the model. Other datasets like livestock and human population, socio-economic data, livelihoods, landcover landuse data should also be factored into the analysis in future. The RVF surveillance, diagnostics and reporting should also be improved so that cases can be reported and controlled early enough below division level so as to reduce extent of future impact of RVF disease.

The study intended to establish whether there was a difference in exercise of strategies by communities in subsequent RVF outbreaks. A test like Fischer's, chi- square would have been used at that point. This could not be done conclusively from the 4 FGDs as it required more discussions be carried in the study area. We were constrained by time. FGDs still stand as crucial points of information to understanding issues and livelihoods of the rural folk. From it a more detailed research, household research could be done so as to further appreciate individual dynamics that are a challenge to coping practices.

REFERENCES

[1] Hussein K, Nelson J. SUSTAINABLE LIVELIHOODS AND LIVELIHOOD DIVERSIFICATION. 1998.

[2] Thornton PK, Boone RB, Galvin K a., BurnSilver SB, Waithaka MM, Kuyiah J, et al. Coping Strategies in Livestock-dependent Households in East and Southern Africa: A Synthesis of Four Case Studies. Hum Ecol 2007;35:461–76.

[3] Rota A. Livestock and pastoralists. Livest Themat Pap Tools Proj Des 2009:1–8.

[4] Kelly PM, Adger WN. THEORY AND PRACTICE IN ASSESSING VULNERABILITY TO. Clim Change 2000;47:325–52.

[5] WHO. Rift Valley Fever t N°207. Fact Sheet N°207 2010.

[6] Anyamba A, Chretien J-P, Small J, Tucker CJ, Formenty PB, Richardson JH, et al. Prediction of a Rift Valley fever outbreak. Proc Natl Acad Sci U S A 2009;106:955–9.

[7] Fontenille D, Traore-Lamizana M, Diallo M, Thonnon J, Digoutte JP, Zeller HG. New vectors of Rift Valley fever in West Africa. Emerg Infect Dis 1998;4:289–93.

[8] Dighe NS, Pattan SR, Bhawar SB, Gaware VM, Hole MB, Waman S, et al. Journal of Chemical and Pharmaceutical Research 2010;2:228–39.

[9] CSIRO, UKaid, DFID. Managing the Risks of Zoonoses - Search Results 2012.

[10] MacMillan S. Economic losses from Rift Valley fever greater than previous documented. ILRI Clippings WordPress.com 2010.

[11] MedicineNet.com. Definition of Rift Valley fever 2003.

[12] HDR. Climate shocks : risk and vulnerability in. 2008.

[13] Yohe G, Tol R. Indicators for Social and Economic Coping Capacity - Moving Toward a Working Definition of Adaptive Capacity Gary Yohe 2001:1–24.

[14] Garba T. SHOCKS, COPING STRATEGIES AND SUBJECTIVE POVERTY: EVIDENCE FROM NIGERIA'S NATIONAL CORE WELFARE INDICATORS QUESTIONNAIRE SURVEY 2006. J Sustain Dev Africa 2011;13:129 – 164.

[15] Ringler C, Hassan RM. Factors Affecting the Choices of Coping Strategies for Climate Extremes The Case of Farmers in the Nile Basin of Ethiopia 2010.

[16] Boko M, Niang I, Nyong A, Vogel C, Githeko A, Medany M, et al. Africa. Africa. Clim. Chang. 2007 Impacts, Adapt. Vulnerability. Contrib. Work. Gr. II to Fourth Assess. Rep. Intergov. Panel Clim. Chang. M.L. Parry, O.F. Canz. J.P. Palutikof, P.J. van der Linden C.E. , 2007, p. 433–67.

[17] Roy BC, Selvarajan S. Vulnerability to Climate Induced Natural Disasters with Special Emphasis on Coping Strategies of the Rural Poor in Coastal Orissa , India. 2002.

[18] Catley A, Alders RG, Wood JL. Participatory epidemiology: Approaches, methods, experiences. Vet J 2012;191:151–60 ST – Participatory epidemiology: Approache.

[19] Wickham H. The Journal. R J 2013;5:1–264.

[20] ESRI. Legal Information | Copyright and Trademarks | Environmental Systems Research Institute, Inc., 2013.

[21] Arino O, Ramos Perez JJ, Kalogirou V, Bontemps S, Defourny P, Van Bogaert E. Global Land Cover Map for 2009 (GlobCover 2009) 2012.

[22] Nachtergaele F, Velthuizen H Van, Verelst L, Batjes N, Dijkshoorn K, Engelen V Van, et al. Harmonized World Soil Database. vol. v1. Rome, Laxenburg: 2008.

[23] Jarvis A, Reuter HI, Nelson A, Guevara E. SRTM 90m Digital Elevation Database v4.1. CGIAR-CSI 2008.

[24] DISC G. Readme for TRMM Product 3B43 (V7). Goddard Earth Sci Data Inf Serv Cent 2013.

[25] TRMM Team TPRT, JAXA JAEA, NASA NA and SA. Tropical Rainfall Measuring Mission (TRMM) Precipitation Radar Algorithm Instruction Manual For Version 6 TRMM Precipitation Radar Team Japan Aerospace Exploration Agency (JAXA) National Aeronautics and Space Administration (NASA) Precipitation Rada. 6th ed. 2005.

[26] NASDA NSDA of J. TRMM Data Users Handbook. 2001.

[27] KUMMEROW C, SIMPSON J, THIELE O, BARNES W, CHANG ATC, STOCKER E, et al. The Status of the Tropical Rainfall Measuring Mission (TRMM) after Two Years in Orbit. J O U R N A L O F A P P L I E D M E T E O R O L O G Y 2000;39:1965–82.

[28] Murithi RM, Munyua P, Ithondeka PM, Macharia JM, Hightower a, Luman ET, et al. Rift Valley fever in Kenya: history of epizootics and identification of vulnerable districts. Epidemiol Infect 2011;139:372–80.

[29] Nelder AJA, Wedderburn RWM. Generalized Linear Models. J R Stat Soc 1972;135:370–84.

[30] McSweeney C, New M, G. L. UNDP Climate Change Country Profiles - Kenya 2012:1–26.

Generating Spatio-Temporal Maximum Entropy Ensembles Using R

Roshan K. Srivastav
Department of Civil and Environmental Engineering
The University of Western Ontario, London, Canada
roshan.srivastav@uwo.ca; roshan1979@gmail.com

Slobodan P. Simonovic
Department of Civil and Environmental Engineering
The University of Western Ontario, London, Canada
simonovic@uwo.ca

Multi-variate ensembles are very practical tools for assessment of uncertainty in the weather data due to changes in future climatic conditions. This paper presents a new multi-variate maximum entropy bootstrap for generating long samples of weather data using R. The model is able to mimic both the temporal and the spatial correlations present in the historical weather data in addition to the other statistical characteristics. The modeling process involves: (i) application of orthogonal transformation to de-correlate the multivariate data; (ii) generation of the samples of decorrelated data using maximum entropy bootstrap; and (iii) inverse orthogonal transformation. The multivariate weather data consists of daily precipitation, mean temperature, minimum temperature, and maximum temperature. An R package is developed to implement the multi-variate maximum entropy bootstrap. The main advantages of using R are: (i) open source; (ii) availability of a number of in-built statistical analysis packages, models and standard statistical tests; and (iii) access to one of the largest collections of user-developed statistical packages that can be easily downloaded, installed and most importantly modified.

Keywords—spatio-temporal, multivariate, entropy, bootstrap, orthogonal transformation

I. INTRODUCTION

Weather generators (WG) play an important role in modeling integrated water resources systems under uncertain future. They can be used for various tasks like (i) filling in the missing hydro meteorological data; (ii) downscaling of global climate model outputs; and (iii) assessing water management decisions under changing conditions. The synthetic replicates obtained from the weather generators capture the statistical characteristics of the observed weather data, which include: (i) basic statistics such as mean and standard deviation; (ii) extremes (minimum and maximum); and (iii) spatial and temporal dependence structure of all the weather variables. In many hydroclimatic studies, the most dominating weather variables that are explicitly affecting the water resources systems are precipitation and extreme temperature (maximum and minimum). These weather variables are collected at various locations and exhibit dependence - called spatial correlation.

Based on the spatial correlation, the weather generator application can be implemented in one of two different ways: (i) single site weather generator where the variables are generated independently at each site without the influence of other sites; and (ii) multi-site weather generator where the variables are generated to include the influence of other sites i.e., spatial correlation.

Research in weather generators offers a variety of tools as in [1] and [2] classified as (i) parametric; (ii) non-parametric and (iii) hybrid or semi-parametric. The parametric models suffer from inherent drawbacks related to reproduction of nonlinearity, generation of negative values and bimodality present in the historical data. Further, the historical data are assumed to be Gaussian in nature, which is not the case in most of the weather datasets. The non-parametric models on the other hand do not make such assumptions, but rather shuffle the data itself, to reproduce the characteristics of the observed data. The major drawbacks of the non-parametric methods are that they cannot reproduce values outside the historically observed range, and preserve temporal correlation. The models which combine the strengths of both parametric and non-parametric approaches are the hybrid or semi-parametric models. In spite of considerable progress, the weather generators proposed in the literature are found to be far from being universally accepted among the researchers and practitioners. The potential reasons may be (i) lack of confidence; (ii) complexity and (iii) computational burden. In general, the weather generators should be reliable, efficient and robust to capture the underlying processes present in the historical data.

In this paper, a new multisite, multivariate Maximum Entropy Bootstrap weather generator (MEBWG) is proposed which preserves both spatial and temporal correlation, in addition to the other statistical characteristics of the daily weather data. The proposed MEBWG model combines (i) maximum entropy bootstrap (MEB) to capture the time-dependent structure and other statistical characteristics of the historical data; and (ii) orthogonal transformations to capture the spatial correlations between the weather variables at multiple sites. The MEBWG tool does not apply (i) any perturbation for smoothing or for generating extremes; (ii) explicit conditioning of variables to preserve spatial or temporal statistics; and (iii) disaggregation from low frequency to higher frequency. Further, the proposed approach is computationally less demanding.

II. MULTI-VARIATE MAXIMUM ENTROPY BOOTSTRAP METHOD

The flowchart of the proposed modeling procedure is shown in Fig 1. The following section presents the multisite multivariate MEB weather generator (MEBWG) model. Let the observed weather variables be represented by $\boldsymbol{W}$ and denoted by

$$W = \left(W_t^{1,1}\, W_t^{2,1} \text{K}\, W_t^{n,1} \text{K}\, W_t^{1,2}\, W_t^{2,2} \text{K}\, W_t^{n,2} \text{K}\, W_t^{1,k}\, W_t^{2,k} \text{K}\, W_t^{n,k}\right) \quad (1)$$

where the superscript 'n' denotes the weather variable at site 'k' and 't' is the index for time. In matrix form $\boldsymbol{W}$ can be expressed as

$$W = \begin{bmatrix} W_1^{1,1} & W_1^{2,1} & \dots & W_1^{n,k} \\ W_2^{1,1} & W_2^{2,1} & \dots & W_2^{n,k} \\ \vdots & \vdots & \dots & \vdots \\ W_t^{1,1} & W_t^{2,1} & \dots & W_t^{n,k} \end{bmatrix} \quad (2)$$

The total number of columns in matrix (2) is equal to N (n times k) and the total number of rows is equal to M (total number of days). The modeling steps of MEBWG: (A) preprocessing; (B) generating replicates; and (C) post processing are as follows.

A. *Preprocessing*

In this stage, the correlated weather variables are uncorrelated by using orthogonal linear transformation. The advantage of using orthogonal transformation is that the characteristics of the overall observed data are assimilated into fewer components in the transformed dataset. The following steps are involved:

Standardize the individual columns of matrix W following

$$w = \left[w^{1,1}\, w^{2,1} \dots w^{n,1} \dots w^{1,2}\, w^{2,2} \dots w^{n,2} \dots w^{1,k}\, w^{2,k} \dots w^{n,k} \right] \quad (3)$$

where $w^{n,k}$ is

$$w^{n,k} = \frac{W_t^{n,k} - \overline{W^{n,k}}}{\sigma_{W^{n,k}}} \quad (4)$$

in which, $\overline{W^{n,k}}$ and $\sigma_{W^{n,k}}$ represent the mean and the standard deviation respectively for each of the weather variables at the kth site (i.e., each column of matrix W Eq(2)).

Apply orthogonal transformation, such that the matrix 'w' (Eq.3) is fully uncorrelated (linear dependence between sites and weather variables). This is achieved by finding out the

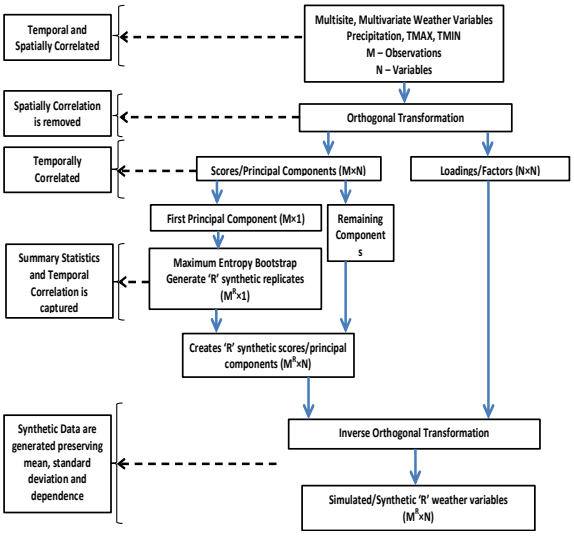


Fig. 1. Flow chart showing the framework for MEBWG weather generator

eigenvalues and eigenvectors of the covariance matrix obtained from the standardized dataset

$$r = Pw \quad (5)$$

where r is a scores matrix which is fully uncorrelated vector components also known as principal components (PC) and has a matrix size same as vector 'w', i.e. $(M \times N)$

$$r = \begin{bmatrix} r_1^{1,1} & r_1^{2,1} & \dots & r_1^{n,k} \\ r_2^{1,1} & r_2^{2,1} & \dots & r_2^{n,k} \\ \vdots & \vdots & \dots & \vdots \\ r_t^{1,1} & r_t^{2,1} & \dots & r_t^{n,k} \end{bmatrix} \quad (6)$$

P is the eigenvectors of the covariance matrix (Ω) and is given by

$$\Omega = COV(w) = \text{E}\Delta\text{E}^T \quad (7)$$

where E is the orthonormal eigenvector matrix or loadings and Δ is the diagonal eigenvalue matrix of size $(N \times N)$. The diagonal eigenvalue matrix (Δ) is arranged in descending order, i.e., the first element of the matrix has the highest eigenvalue then the second and so on.

The replicates in transformed space can be generated either by fitting a generating Maximum Entropy Bootstrap model (MEB) to each component (column) of the score matrix (r), or by selecting significant columns (principal components) and fitting generating model to the selected components. We show in the following steps how to generate replicates in transformed space for one of the principal components using

MEB. The same procedure can be followed to generate replicates for other principal components.

B. Generating Replicates – Maximum Entropy Bootstrap

The maximum entropy bootstrap (MEB) involves: (1) construction of maximum entropy densities which are combination of uniform densities (2) sampling data from the maximum entropy densities and (3) arranging the sampled data according to the historical data ranking. The following seven-step procedure is explains in detail the MEB modeling procedure for generating:

Construction of Entropy Densities (Combination of uniform Densities)

1. The data (first column of the matrix 'r') is sorted in ascending order to create a rank matrix O_t. This will help in ensuring the restructuring of the generated replicates would result in similar pattern as observed in the data.
2. Compute intermediate points (P_t) from the rank matrix, using:

$$P_t = \left(\frac{O_t + O_{t+1}}{2} \right) \quad for\ t = 1, \ldots, M-1 \tag{8}$$

3. To extrapolate beyond the historical extremes (i.e., minimum and maximum values), lower and upper limit of the data are calculated using trimmed mean (T_{trim_mean}) of O_t (step 3)

   Then the lower limit is determined as:

$$P_0 = O_1 - T_{trim\_mean} \tag{9}$$

   and the upper limit is determined as

$$P_M = O_M + T_{trim\_mean} \tag{10}$$

   where O_1 and O_M are the minimum and the maximum value of the data in Step 3.

   This ensures that the replicates generated are beyond the historical extremes. The justification for use of trimmed mean in this study includes: (i) robustness to outliers; and (ii) high computational efficiency for mixed and heavy tailed distributions (very common with some atmospheric variables). It is to be noted that many other methods may be used instead of trimmed means and this is left for future research.
4. The maximum entropy density is constructed, such that the ergodic theorem (mean preserving) is satisfied. The following equations are used to calculate the desired mean:

$$m_1 = 0.75O_1 + 0.25O_2 \tag{11}$$

$$m_k = 0.25O_{k-1} + 0.5O_k + 0.25O_{k+1} \quad \forall\ k = 2, 3, \ldots, t-1 \tag{12}$$

$$m_t = 0.25O_{t-1} + 0.75O_t \tag{13}$$

Sampling from Maximum Entropy Density

5. Uniform random numbers between 0 and 1 are generated and the sample quantiles of the maximum entropy density at those points are obtained and sorted accordingly.

Reordering the Sampled Replicates

6. Using the rank matrix from step 3, the sample quantiles are reordered. This step ensures that the temporal dependence of the historical structure is replicated.
7. Steps 4 to 8 are repeated, till the desired number of replicates of length 'M' are generated

$$r' = \begin{bmatrix} r_{1,1}^1 & r_{1,2}^2 & \cdots & r_{1,rep}^k \\ r_{2,1}^1 & r_{2,2}^2 & \cdots & r_{2,rep}^k \\ \vdots & \vdots & \vdots & \vdots \\ r_{t,1}^1 & r_{t,2}^2 & \cdots & r_{t,rep}^k \end{bmatrix} \qquad \forall rep = 1, \ldots, numrep \tag{14}$$

   where 'numrep' is the total number of replicate series generated

C. Post-Processing

Assuming that the first column of the 'r' matrix (represented as $r_t^{1,1}$ Eq. 6) is considered for generating the replicates in the transformed space (represented as $r_{t,rep}^{1,1}\ \forall rep = 1, \ldots, numrep$, where numrep is the total number of replicates generated), the post-processing continues as:

8. Replace each replicate with the first component in scores matrix to obtain the modified scores matrix. For example to obtain the first set of replicates in transformed space, replace $r_{t,1}^{1,1}$ in place of $r_t^{1,1}$ in matrix 'r' in Eq. 6 as shown below to obtain:

$$r_{rep} = \begin{bmatrix} r_{1,1}^1 & r_1^2 & \cdots & r_1^k \\ r_{2,1}^1 & r_2^2 & \cdots & r_2^k \\ \vdots & \vdots & \vdots & \vdots \\ r_{t,1}^1 & r_t^2 & \cdots & r_t^k \end{bmatrix} \ldots\ldots \begin{bmatrix} r_{1,\mathrm{rep}}^1 & r_1^2 & \cdots & r_1^k \\ r_{2,\mathrm{rep}}^1 & r_2^2 & \cdots & r_2^k \\ \vdots & \vdots & \vdots & \vdots \\ r_{t,\mathrm{rep}}^1 & r_t^2 & \cdots & r_t^k \end{bmatrix} \tag{15}$$

9. Using the modified scores matrix, perform inverse orthogonal transformation to obtain synthetic weather variables at all sites. Note that the entire matrix in Eq. 15 is used to perform inverse transformation, i.e., no dimensionality reduction.
10. Repeat steps 10 and 11, until the number of generated sequences is equal to the total number of replicates to be generated.

11. Perform inverse standardization to obtain the synthetic replicates for all the weather variables at multiple sites.

The replicates obtained from the proposed model are used to calculate summary statistics, dependence structure and climate based indices. Implementation examples and the application of MEBWG to three river basins are shown in [1] and [2].

III. Implementation of Multi-variate Maximum Entropy Bootstrap Method in R

'R' is statistical computing software available under the GNU General Public License and currently developed by R Development Core Team. R has its own programming language built on lexical scoping semantics and S programming language. R software is written in C, Fortran and R and available for various operating systems. R is also based on object-oriented programming which creates wider flexibility for the programmers to work on objects rather than structured programming. Advance R uses can easily link the routines from various other programming languages to improve the computational performance of the program. R programming is used for computational statistics, visualization and data science. More specifically R can be used for data manipulation, statistical modeling and representation multifaceted data with charts and graphs. The advantages of using 'R' are that it is free and open source software. Number of statistical analysis, models and standard statistical tests are in-built into the R programming language. R also has one of the largest collections of user-developed statistical packages, which is easy download, install and most importantly modify.

We developed an R package to implement the multi-variate maximum entropy bootstrap. This package consists of three main components (i) Pre-processing; (ii) Generation of replicates using maximum entropy bootstrap; and (iii) Post-processing. In pre-processing the collinear multivariate data is transformed into linearly independent components using orthogonal transformations. This is achieved by using inbuilt function 'princomp' in R software. We select the first principal component from the transformed variables to generate replicates using maximum entropy bootstrap. Next we use the 'MEboot' R package [3] developed for data generation of uni-variate time series using maximum entropy bootstrap. To generate the replicates beyond the historical maximum the 'MEboot' package requires the trimmed mean percentage as one of the inputs. In order to capture the collinearity between the components the replicates in orthogonal space are inverse transformed using the scores matrix obtain in first step using 'princomp'. The following pseudo R-code shows the implementation of the multi-variate maximum entropy bootstrap model.

```
#####################################
# Load R Package
library(meboot)
#####################################
# Load Multivariate Data
data <- read.csv("data.csv")
#####################################
# Other Inputs
rep <- 100 # Number of replicates
per.trim <- 10 # trimmed mean percentage
#####################################
# Pre-Processing - Orthogonal Transforamation
ortho.trans <- princomp(data)
#####################################
# Generate replicates using MEboot package
comp1.boot <- meboot(ortho.trans[,1], reps = rep, trim = per.trim)
#####################################
# Inverse Transformation
inv.trans <- z.i %*% t(ortho.rep)
#####################################
```

IV. Conclusions

In this paper, a new multisite, multivariate Maximum Entropy Bootstrap (MEBWG) method is presented for generating daily weather variables. The method has the ability to mimic both the spatial and temporal dependence structure, in addition to the other historical statistics. The Maximum Entropy Bootstrap (MEB) is suited to the random generation of non-stationary time series and involves two main steps: (1) random sampling from the empirical cumulative distribution function (ECDF) with endpoints selected to allow limited extrapolation; and (2) reordering of the random series to respect the rank ordering of the original time series (temporal dependence structure). To capture the multi-collinear structure present between the weather variables and between the sites, we combine orthogonal linear transformation with MEB. The implementation of MEBWG is done in R.

References

[1] Srivastav, R. and S.P. Simonovic, "Maximum Entropy Bootstrap based Multi-site, Multivariate Weather Generator", *Climate Dynamics,* 2014a, DOI:10.1007/s00382-014-2157-x, available online http://link.springer.com/article/10.1007/s00382-014-2157-x.

[2] Srivastav, R. and S.P. Simonovic, "Analytical Procedure for the Implementation of Multi-site Multi-season Streamflow Generator using Maximum Entropy Bootstrap", in print *Environmental Modelling & Software,* 2014b.

[3] Vinod, H. D., & Lopez-de-Lacalle, J. "Maximum Entropy Bootstrap for Time Series : The meboot R Package", Journal of Statistical Software, 2009, 29(5).

Potential Use of an Open-Source Software R as a Tool for Performing Climate Change Impact Studies

Abhishek Gaur
Facility for Intelligent Decision Support
Department of Civil Engineering, Western University
London, Ontario, Canada
abhishek.gaur1988@gmail.com

Slobodan P. Simonovic
Facility for Intelligent Decision Support
Department of Civil Engineering, Western University
London, Ontario, Canada

Performing climate change impact analysis requires significant amount of data manipulation and analysis. R programming language has been used extensively as statistics and data analysis tool in the past. Objective of this paper is to highlight the utility of R programming language while performing climate change impact analysis. The paper offers (i) a generalized procedure for the assessment of climate change impacts and (b) description of the role that R programming language and different R packages can play.

Keywords-R programming language; climate change impact analysis.

I. INTRODUCTION

Growing concern about changing climatic conditions across the globe has led to significant amount of research in the field of climate change impact assessment. Research has been done both to assess the impacts as well as on the development of appropriate assessment methodologies. A number of studies have been performed at local (for example [1]), regional (for example [2]) and global (for example [3]) scale which estimate probable changes in climatic and hydrologic conditions in future. Presence of a wide variety of models/methodologies used in several steps of climate change impact studies confirms that none of them is perfect, although their efficacy is gradually improving. As a result steps followed while performing climate change impact studies differ across different studies undertaken in the past. The objective of this paper is: a) to outline the series of steps that can possibly be involved in a typical climate change impact study on flooding frequencies and b) to identify the role that R programing language can play in the implementation of these steps. The vision is to be able to utilize provided guidelines towards framing a uniform methodology for climate change impact studies and implement this methodology using R.

II. CLIMATE CHANGE IMPACT ANALYSIS PROCESS

This section of the paper describes three major steps involved in a typical climate change impact study on flow extremes (as summarized in Figure 1). These steps are: (i) selection of future climate projections, (ii) pre-processing of global climate model (GCM) data and (iii) generation and analysis of future stream flow projections.

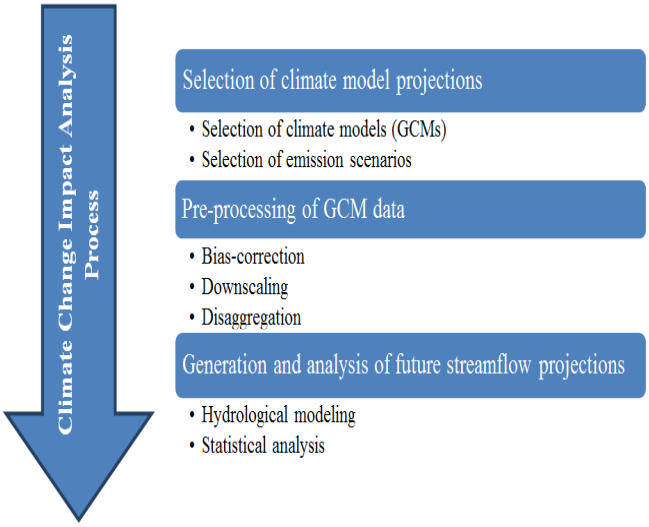


Fig. 1. Steps involved in a typical climate change impact study on flow extremes.

A. *Selection of climate model projections*

A total of twenty three climate models have been identified by the Intergovernmental Panel on Climate Change in their 4th assessment report (AR4) [4] and they have been found to reproduce historical observed climate reasonably well. However, none of the models is able to do it perfectly. It is therefore recommended in AR4 that projections from every model should be considered equally plausible in future and should be included while providing future climate projections. Three Special Report Emission Scenarios (SRES) have been utilized extensively in AR4 and Representative Concentration Pathways (RCPs) are used in detail in upcoming IPCC 5th assessment report (AR5). For generation of flows, climate model data for climate variables precipitation (*ppt*), mean temperature (t_{mean}), maximum temperature (t_{max}) and minimum temperature (t_{min}) is required for baseline (reference) and future timelines. One of the major reasons that hinder usage of all model-scenario combinations is the lack of continuous climate data for these climate variables. In spite of data non-uniformity, large ensembles of climate model data are available for usage. Historical and future data as well as the changes projected by these models differ from each other significantly.

Climate model selection has been performed in the past to fulfil one of the following two objectives: 1) to choose from the range of projections made by the ensemble or 2) to encompass the projections made by the multi-model ensemble. Former approach can be performed by considering averaged, weighted-average or selected projections for analysis. This approach has been extensively used for designing climate-informed water resource systems in the past although it compromises with the total uncertainty associated with GCM projections. Latter approach can be performed by using methods like scatter-plot selection and percentile selection. In the scatter-plot method, GCM-scenario combinations most likely to produce hydro-climatic weather extremes are selected. Cold-dry, cold-wet, hot-dry and hot-wet scenarios in terms of mean changes projected are selected for analysis. In percentile method, selection is made to capture the entire range of changes projected by GCMs. Generally, scenarios corresponding to 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile of the entire range of projections are selected for analysis.

Dependency of these model selection methodologies on the spatial, temporal and distributional scale selected is explored using a method that compares probability density functions of historical observed and model simulated data and allots a skill score measured by the extent of their overlap.

Selection of model-scenario combination can be made a) based on mean changes projected by the multi-model ensemble and b) based on extreme (mean above 99^{th} percentile) changes projected by the ensemble. The implementation of this process shows that model rankings differ in each of experiments performed. This suggests that selected climate projections as well as model performance depends on the spatial, temporal and distributional scale chosen for analysis [1].

B. Pre-processing of model climate data

Climate model datasets are associated with low spatial, temporal and distributional resolution. They have typical horizontal spatial resolutions of 2° x 2° which is close to 220 km x 220 km. This means that physiographic characteristics of such a huge area are approximated into one grid cell of the climate model data. Spatial extents of grid points exceed the catchment scales at which climate change impact studies are typically performed. Climate model datasets are available in yearly, monthly and recently in daily time steps. Climate data obtained from GCMs are associated with some time-independent component of model errors called biases. These biases are evident when simulated climate model data for baseline are compared with historical observed data at the same location. Most hydrologic models require climate variable data at gauging stations in hourly time steps for generating stream flows. To make the raw climate data usable for catchment scale hydrological analysis, methods such as bias-correction, downscaling and temporal disaggregation are employed.

C. Bias correction of gridded GCM climate data

While performing climate change impact studies, bias associated with climate model data can be roughly but safely, defined as the time independent component of model error or the component of model error which remains constant throughout the length of datasets. The purpose of bias correction step is to modify climate model data in a way that the correlation between model and observed data increases. [1] shows that the changes projected by raw and bias-corrected climate data differ significantly. Methods employed to do the bias correction range from those correcting just the means to those correcting entire distributions of climate data. The amount of uncertainty associated with bias-correction approaches has been even found to be comparable to that associated with climate model projections.

D. Statistical downscaling

Downscaling is a method for improving the spatial resolution of GCM output. Two distinct groups of approaches exist for doing so. First group is known as the dynamic downscaling approach in which local physiographic information along with GCM boundary conditions are used to generate higher resolution Regional Climate Model (RCM) datasets. The dynamic downscaling process is highly computationally intensive. Further, dynamically downscaled results are governed significantly by associated GCMs and so, they too are uncertain.

Another less computationally demanding approach for downscaling GCM data is known as statistical downscaling. It is based on the principal that regional data is dependent on large scale climate state as well as local physiographic features. Information regarding large scale climate state is generally extracted from the GCMs while different parametric, semi-parametric and non-parametric methods are employed to transfer this large scale information to local scale. Future projections obtained using different downscaling methodologies can be significantly different from each other.

E. Disaggregation of daily climate model data to hourly timescale

Climate model daily datasets need to be disaggregated to hourly timescale before they can be used for hydrological simulations. Capturing the hourly variability of climate variables accurately is important for simulating daily hydrological response of the catchment. Disaggregation of daily temperature data into hourly timescale is straightforward and simple methods like the cosine formulae have been found effective in doing so. Disaggregation of precipitation is much more complicated than temperature because of the unsystematic variability involved at a sub-daily scale. Methods used currently for disaggregating daily precipitation include a) uniform distribution of daily rainfall across the day; b) stochastic temporal disaggregation of daily data; c) use of detailed information from a nearby station to perform disaggregation; and d) multivariate disaggregation by employing a combination of b) and c).

Detailed analyses by [1] suggest that choice of downscaling method can have a significant effect on future

projections and that the uncertainty associated increases with the number of methods taken into consideration.

F. Hydrological modeling

Selection of hydrological model can be made based on several factors including size of catchment under analysis, availability of climatic and hydrologic data, usage history of hydrologic models within the catchment etc. If the catchment is small (<100 km^2 in area) lumped hydrological models can be used. However, for medium and large catchments (>100 km^2 in area), semi-distributed or distributed models can be used. Further, continuous hydrological modelling is necessary while estimating peak runoff from a catchment. Hydrological model structure is found to be the most significant source of uncertainty in this step. Parameter uncertainty in hydrologic models due to equifinality and robustness requirements under changing climatic conditions are some of the other sources of uncertainty associated with this step. Further, changes in land-use and reservoir release should contribute heavily towards framing future flow peaks. However, uncertainty associated with these two sources has been explored in lesser details than the other sources till now.

G. Statistical analysis

Statistical analysis is performed on time-series of peak flows to develop relationships between flood frequency and flood magnitude. There are two prominent methods of extracting peak flow data from a discharge series, namely Annual Maximum (AM) and Peak Over Threshold (POT) method. In AM method, yearly maximum discharge values while in POT method, discharge events larger than a specified threshold are selected for analysis. Major limitation of AM method is that the values extracted may not be representative of actual peaks in the entire discharge series. Another limitation is that sample size of peak flows obtained from the AM method is small (equal to the number of years of discharge series data) and hence, reliable statistical inferences are hard to be drawn from it. POT method, on the other hand, overcomes these limitations and is extremely useful especially when the available discharge series is short. Selected flow peaks can be fitted using an appropriate distribution and flow quantiles corresponding to chosen return periods can be estimated. Uncertainty in flood frequency analysis methods may arise due to differences in assumption of extreme value statistics, choice of sample, choice of distribution function, choice of parameter estimation method and statistical inference method.

III. USE OF R PACKAGES

Different steps that have been performed and potential role of R packages used in the process presented in the previous section are summarized below:

A. Interpolation of climate model data

Climate model data is provided in the NetCDF format and can be accessed in R using packages like "ncdf". Three dimensional gridded datasets in a NetCDF file are arranged according to latitude (lat), longitude (lon) and time (t). These variables as well as the datasets can be extracted using "get.var.ncdf (filename, variable-name)" function in the "netcdf" package. Extracted datasets can be interpolated at the location of interest using algorithms like inverse distance square method. Data corresponding to four climate model gridpoints surrounding the location of interest are used and interpolation is made based on their distances from the location of interest.

B. Evaluation of climate models

Climate models are evaluated by their ability to reproduce historically observed climate. In R, data corresponding to gauging stations and climate models located within the area of interest can be extracted using packages like "maptools" and skills of different climate models can be established. Formulation of the indices of skill evaluation can be made within R.

C. Selection of future climate model projections

Assessment of the uncertainty is a major objective of any climate change impact analysis study. Uncertainty of future emission scenarios can be assessed using methods like scatter-plot and percentile selection. Percent and absolute changes projected for precipitation and temperature respectively can be calculated and used to select extreme GCM-scenario combinations in R. GCM-scenarios corresponding to "cold-dry", "cold-wet", "hot-dry" and "hot-wet" extreme scenarios can be selected using scatter-plot method while those projecting 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles of changes are selected using percentile method.

D. Bias correction of climate data

Bias in climate model data arises from the fact that climate models are imperfect. It can be corrected using methods correcting data means or those correcting entire distributions of climate model data. Different steps involved while doing so ranging from a) establishment of bias-correction functions b) estimation of parameters involved c) estimation of corrected climate data can be performed in R with ease owing to its statistics and data analysis capabilities.

E. Downscaling of climate data

Downscaling is a technique used to estimate climate data at a local scale. Climate model datasets available at low spatial resolution are used for doing so. Statistical downscaling can be performed in R. Change factors based statistical downscaling is one of the methods that can be used for doing so. Distribution based change factors can be estimated and future scaled climate can be used in weather generators to estimate future climate data at a local scale. Weather generators like KNN-CAD v4 are solely developed in R [5].

Several packages like “timeseries” and “lubridate” have been used in the weather generator to facilitate analysis of climate variable time-series data.

F. Disaggregation of daily climate data

Many hydrologic models require climate data input in hourly time-steps. However even the most recent climate model datasets are available only in daily timesteps. Disaggregation is a process of converting daily scale climate model data into hourly scale. Several methodologies for disaggregating precipitation and temperature data have been recommended [1]. Simple (for instance cosine function method to disaggregate temperature, uniform disaggregation of rainfall) as well as more complicated spatial and temporal statistics based methodologies can be implemented using R interface and R packages like “timeseries”, “sp”, “rgdal” and “maptools”.

G. Hydrologic modelling

Hydrologic modelling in R can be performed for small catchments. Packages like “hydromad” are available which provide users with an ensemble of multiple multiple lumped models. Using more than one hydrologic model is also essential in climate change studies to capture the uncertainty associated with hydrological modeling step. Several other packages like “HydroGOF”, “HydroTSM” are also useful for performing statistical tests on hydrologic data.

H. Flood frequency analysis

Flood frequency analysis is performed to estimate relationship between flood magnitude and return period. Different steps involved while performing flood frequency analysis: a) Selection of extreme flow datasets b) Fitting of appropriate distribution function c) Estimation of distribution function parameters d) Establishment of flood magnitude-return period relationships can be performed by utilising statistical and data analysis capabilities of R. Some of the packages that are useful while doing so are: “Lmoments”, “Lcomco” etc.

IV. CONCLUSION

The paper presents a generalized procedure for the assessment of climate change impacts and identifies the role that R packages can provide. A number of steps in the presented procedure are executed using R and those that are not yet are identified for future development.

REFERENCES

[1] Gaur A, Simonovic SP (2013) Climate Change Impact on Flood Hazard in the Grand River Basin, Ontario, Canada. Water Resources Research Report No. 084. London, Ontario: Facility for Intelligent Decision Support 92 pp.

[2] Dankers R, Feyen L (2009) Flood Hazard in Europe in an Ensemble of Regional Climate Scenarios. Journal of Geophysical Research 114: D16108. doi:10.1029/2008JD011523.

[3] Hirabayashi Y, Mahendran R, Koirala S, Konoshima L, Yamazaki D, Watanabe S, Kim H, Kanae S (2013) Global Flood Risk under Climate Change. Nature Climate Change 3: 816–821. doi:10.1038/nclimate1911.

[4] Intergovernmental Panel on Climate Change (IPCC) (2007) Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA..

[5] King, L., S. Irwin, R. Sarwar, A.I., McLeod and S.P. Simonovic, (2012) “The effects of climate change on extreme precipitation events in the Upper Thames River Basin: A comparison of downscaling approaches”, *Canadian Water Resources Journal,* 37(3):253-274.

An Operational R-Based Interpolation Facility for Climate and Meteo Data

[extended abstact]

Dr. Raymond Sluiter
R&D ICT & Sensor Technology
Royal Netherlands Meteorological Institute (KNMI
De Bilt, The Netherlands
raymond .sluiter@knmi.nl

KNMI is the Dutch national institute for weather, climate research and seismology. It disseminates weather and climate information to the public at large, the government, aviation and the shipping industry. KNMI conducts observations, develops models and performs fundamental research on the climate system. Interpolated datasets are becoming more and more important: environmental changes, such as global warming, force governments to develop policy to counter any adverse effects. For example hydrological models are part of the tool set used to develop and evaluate policy. These models require gridded maps of meteorological parameters such as rainfall, temperature and potential evaporation. KNMI has several ongoing projects to generate such gridded maps using point observations as a basis. The common goal of the research projects is to obtain the optimal interpolation method for each meteorological parameter and to optimize the dissemination of the interpolated maps to the users using innovative web services.

I will present the Climate Atlas project and the datasets needed for the national project "Nationaal Hydrologisch Instrumentarium (National Hydrological Toolbox)". Prior to 1990 many parameters were measured on 5 to 16 weather stations. After 1990 these parameters were measured on about 30 automatic weather stations scattered around the Netherlands. The limited number of measurements poses great challenges for interpolation of 30 year time series as needed by retrospective and future climate studies. Therefore, we tested several methods including inverse distance weighted interpolation, multiple linear regression, thin plate splines and kriging. Quality information of the datasets is provided to the users in the form of cross-validation results and kriging variance. In the project we use R for statistical computing and graphics, including the gstat package. For research and production purposes we developed the GeoSpatial Interpolation Environment (GSIE) which is a shell including the R environment, database connections and OGC web services for visualization.

We provide the datasets through the recently developed KNMI Data Centre (KDC: http://data.knmi.nl). KDC has been built on open standards and proven open source technology, which includes in-house developed software like ADAGUC WMS, the NADC data processing framework and a web portal. KDC includes 1) A user-friendly interface for (meta)data management; 2) Search capabilities using ISO 19115 (INSPIRE) compliant metadata available through a Catalog Services for the Web (CSW) server combined with deep search on geolocation and time; 3) NetCDF-CF files downloadable through FTP and HTTP services; 4) Viewing services based on the OGC Web Mapping Service Standard (WMS).

In the keynote presentation I will present the entire interpolation chain including the latest developments of KNMI's GeoSpatial Interpolation Environment (GSIE), the interpolation techniques, the datasets, data dissemination through KDC, future research, issues and challenges.

Production Of Climate Maps: Operational Issues And Challenges

[extended abstact]

Mojca Dolinar
Climatology Section
Slovenian Environmental Agency
Ljubljana, Slovenia
m.dolinar@gov.si

Key words: Spatialisation, Spatial interpolation, Climatology, Climate variables

I. INTRODUCTION

When calculating climatological variables in regular grids geostatistical methods prove to be very powerful. However, operationally we are faced by many challenges as the consequence of imperfect reality. The basic input in spatialisation procedure are measurements with their measurement error. The spatial density of point measurements is never adequate for our demands and is homogenous neither in space nor in time. The use of auxiliary variables with higher spatial density to resolve the spatial variability of the treated variable can help, but again, the coverage of measurements regarding the auxiliary data domain should be high enough. Another issue regarding measurements is their representativity, which could be inadequate for the task we have to solve. Inclusion of modern, remote sensing measurements with relatively good spatial coverage in spatialisation process is often advantageous but poses another challenge, how to ensure temporal homogeneity of calculated grids. Finally, there are user demands, which are reasonable, but sometimes from the above described or some additional reasons unsolvable for data providers.

II. DATA

Climatological variables in regular grids are calculated from different measurements using more or less sophisticated methods or models. The point in situ measurements are still the basic input data for calculation of grids on regional, local or smaller level. All the measurements have their own measurement error, which is, in most cases, small in comparison to errors and uncertainties aroused from the spatialistaion process. More crucial is the question of spatial measurements density on which depend both: the scale in which the spatial variability could be treated and the final spatial resolution. The lack of point measurements we try to substitute by the inclusion of auxiliary variables with higher spatial density, which are well correlated with and could help to explain the spatial variability of the treated variable. One of the very frequently used auxiliary variables in climatology is elevation, which is especially well correlated with mean temperature. Using it we encounter another problem – the lack of high elevation mountainous stations. Fig. 1 shows a high deficit of measurements on elevations above 1000 m in Slovenia. In such cases there is a high risk of extrapolation instead of interpolation if using auxiliary variables.

Another issue regarding data availability and data density in local scale are country (regional) borders. Usually data across the border are not available or their density is very low. In that case, especially in very odd shaped regions, the grid points close to the border have higher error than inner grids and their value uncertainty is much higher. This difficulty is obvious when we try to merge the grids from different regions and encounter discontinuities on the border (Fig. 2). The discontinuities could be very large when the border is in the complex terrain. One of the possible solutions is inclusion of across border measurements in the spatialisation process, but the discontinuities can not be completely avoided, especially if the variability of the treated climate variable is very high.

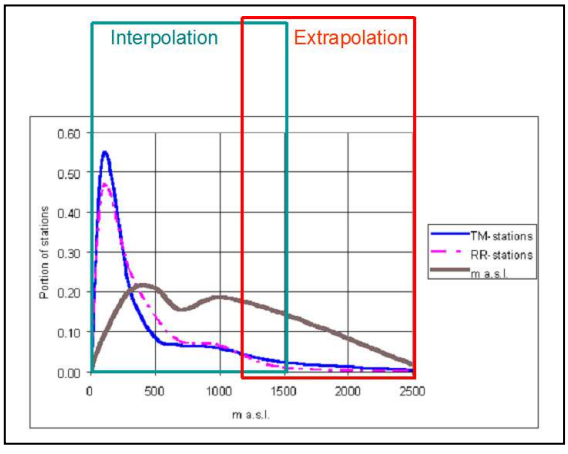


Fig. 1. The relative frequency distribution of Slovenian terrain (brown line) and of the meteorological stations (climatological stations blue line, precipitation stations magenta line) according to the altitude.

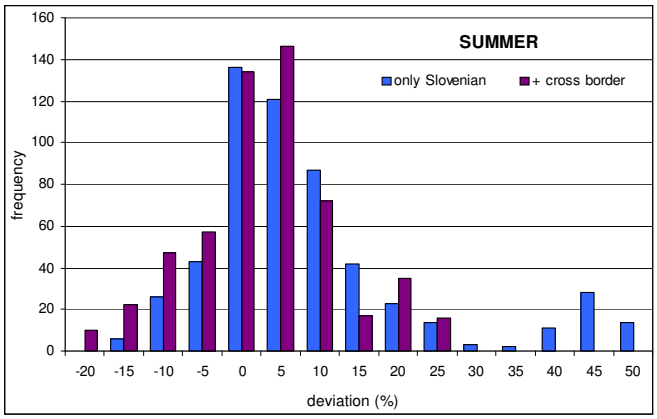


Fig. 2. Frequency distribution of relative differences between Austrian and Slovenian summer precipitation grid (period 1971–2000) in 1 km wide belt on the country border. Slovenian grid is calculated twice: once only with Slovenian measurements and second with additional 28 cross border measurements. It is obvious that additional measurements narrow the distribution towards smaller differences, however, the differences are quite large in both cases. Slovenian precipitation estimations are slightly biased towards higher values than Austrian.

One of the modern challenges regarding data, that we are facing presently or will have to face in near future, is a wide spectra of different (new) measurements (also in situ). The problem is not purely spatialistaion problem but affects spatialisation process a great deal. It rises a question how to assure continuity and temporal homogeneity of grids and at the same time take advantage of additional (but different) information from new technology measurements.

III. AUXILIARY DATA

The selection of auxiliary variable should always base on physical relation between treated and auxiliary variable. Usually the spatial density of auxiliary variable is much higher than the density of treated variable, or at least the spatial correlation structure of treated variable is well known and similar to the spatial correlation structure of the treated variable. As already mentioned, when selecting auxiliary variable, we should be careful that the values range of both variables is similar to avoid extrapolation problem.

Beside the geographical variables (original and derived ones) there are many other variables that can help us to explain spatial variability of the treated variable. The remote sensing measurements of meteorological variables (radar and satellite data) have a great advantage because of their good spatial coverage. However, their absolute values are biased and inhomogenous in space. Combining remote sensing and in situ measurements can improve spatialisation results. Spatial resolution can be increased and small scale structures, which would be overlooked by the in situ measurement network, can be resolved (Fig. 3). High resolution remote sensing data can also be used only to define spatial correlation structure of the treated process (Fig. 4).

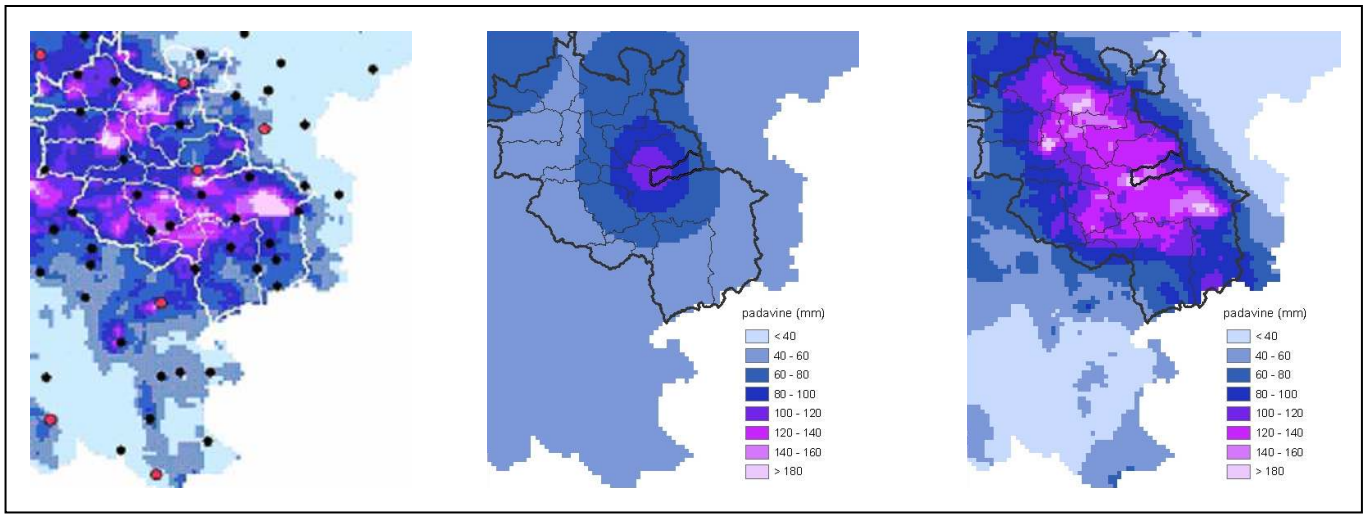


Fig. 3. Radar precipitation measurements in 1 km resolution with indicated in situ precipitation stations (left), Spatial distribution of precipitation in 1 km resolution, using only in situ measurements (middle) and spatial distribution of precipitation in 1 km resolution, using combined in situ and radar measurements (right). It is obvious that with existing in situ measurement network a fine structure of convective precipitation is overlooked. On the other hand, radar measurements show the fine structure of convective cells but the absolute values are not reliable due to attenuation, shading and other errors of radar measurements. Using both kind of measurements in the spatialisation process, we get the fine convective structure in the precipitation field and absolute values of precipitation are in accordance with in situ measurements.

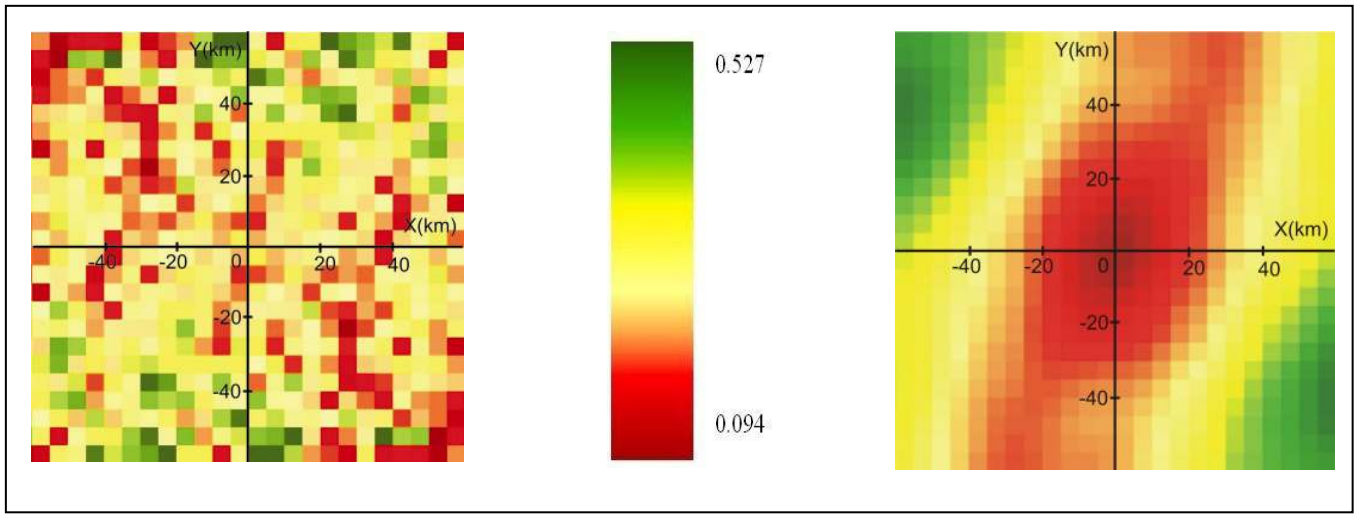


Fig. 4. Empirical variogram map for selected precipitation event. The left one is calculated from in situ measurements, the right one is calculated from the radar measurements. While the strong anisotropy in spatial correlation is evident in the radar variogram map (right) it is imposible to deduce anisotropy from the in situ variogram map (left).

IV. SPATIAL RESOLUTION

The demand for high resolution grids is constant but the final resolution of grid is limited by the spatial density of the input data and by the spatial variability of the treated variable. As already mentioned in the chapter 3, there are possibilities to improve the resolution of the grid. One of them is inclusion of auxiliary variable(s). In some cases we can also improve the final resolution by incorporating the physical knowledge of the spatial realisation of the treated variable. Fig. 5 shows an example of using a simple physical model to improve the final resolution of wind potential grid.

V. USER DEMANDS

With the growing computer power there seem to be no limits for the demanded spatial resolution of grids. Other two frequent user requests are high temporal resolution and (near) real time production of grids. In the effort to meet user demands we are often led to exaggerate and produce grids with very high uncertainty and errors. That is why it is very important to communicate the information of grids uncertainty and their quality to the users. Personal communication (if possible) is proved to be the best, because the uncertainty information, which is simply added to the products, is often ignored by the users.

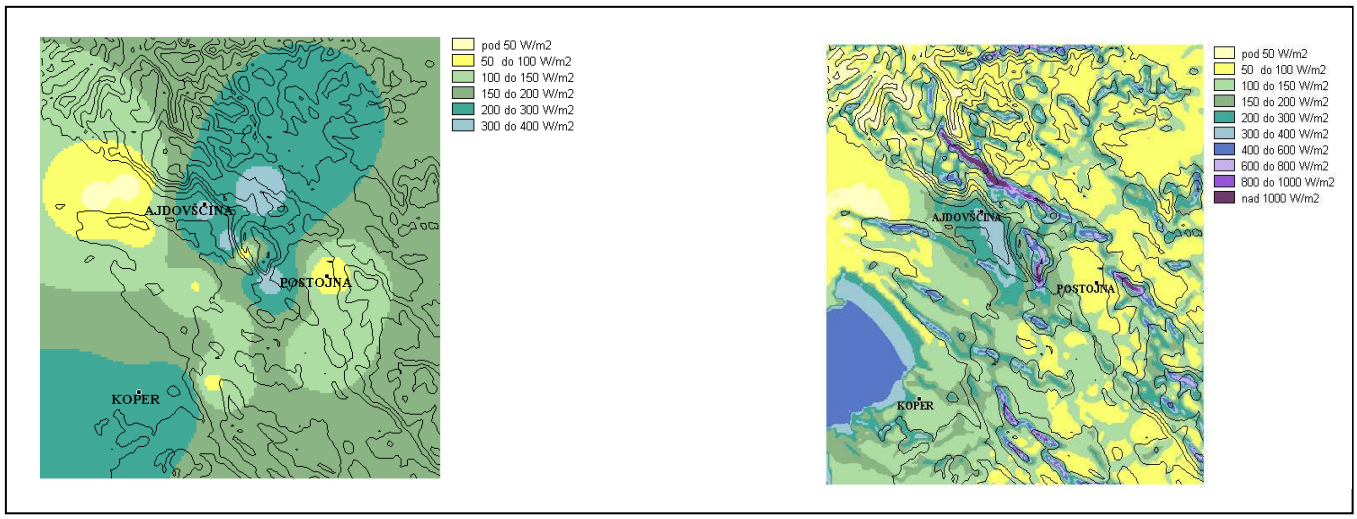


Fig. 5. Wind potential map calculated from wind speed measurements (17 measuring points) using Inverse distance weights interpolation method (left) and combined method approach: mass-consistent physical model, regression, one dimensional mathematical Bora model, local influences (right).

30-Arcsecond Climate Projections for All Global Land Surfaces

[extended abstact]

Thomas M. Mosier
College of Engineering
Oregon State University
Corvallis, Oregon, USA
mosiert@onid.orst.edu

David F. Hill
Civil and Construction Engineering
Oregon State University
Corvallis, Oregon, USA

Kendra V. Sharp
Mechanical, Industrial, and Manufacturing Engineering
Oregon State University
Corvallis, Oregon, USA

***Abstract*— A statistical downscaling method is presented for producing 30-arcsecond monthly time-series of precipitation, mean temperature, minimum temperature, and maximum temperature from GCM data. These downscaled data can be produced efficiently for any global land area, requiring the user only to specify their region of interest. To highlight one use of these data, a metric called potential snowfall is calculated for the Karakorum and Hindu Kush region from 2020 through 2100. These data show that potential snowfall is decreasing over this period and that this change is very spatially heterogeneous. Therefore, using these downscaled data offer a readily apparent advantage compared to using the relatively low spatial resolution GCM output directly. These tools are open source and freely available at GlobalCliamteData.org.**

Keywords*— *downscaling; climate data; monthly precipitation; monthly temperature; snowfall

I. INTRODUCTION

High-spatial resolution time-series of climate data are critical for many hydrological and earth science studies, yet these data are unavailable for most global land areas. Previous work by Mosier et al. [1] developed a downscaling package written in Matlab for producing 30-arcsecond monthly time-series hindcasts of monthly precipitation, mean temperature, minimum temperature, and maximum temperature. These data are useful for studying climate over the twentieth century. The present work expands the downscaling package for use with any GCM projection run associated with the Coupled Model Intercomparison Project Phase 5 (CMIP5) [2]. GCM projection data. Therefore, the updated downscaling package can be used to produce globally available monthly projections at 30-arcseconds for the twenty-first century. These downscaled data have the same temporal features as the original GCM data but include finer scale orographic effects not present in the original GCM. As a demonstration of the new downscaling package's utility, it is used to model potential snowfall for the Karakorum and Hindu Kush (KHK) mountain region of Central Asia (shown in Fig. 1).

II. METHODOLOGY

The current expansion of the downscaling package by Mosier et al. [1] includes an algorithm for downscalig GCM projection data. The two basic components of the GCM downscaling algorithm are to bias correct the GCM data and then to spatially downscale it using a Delta method [3].

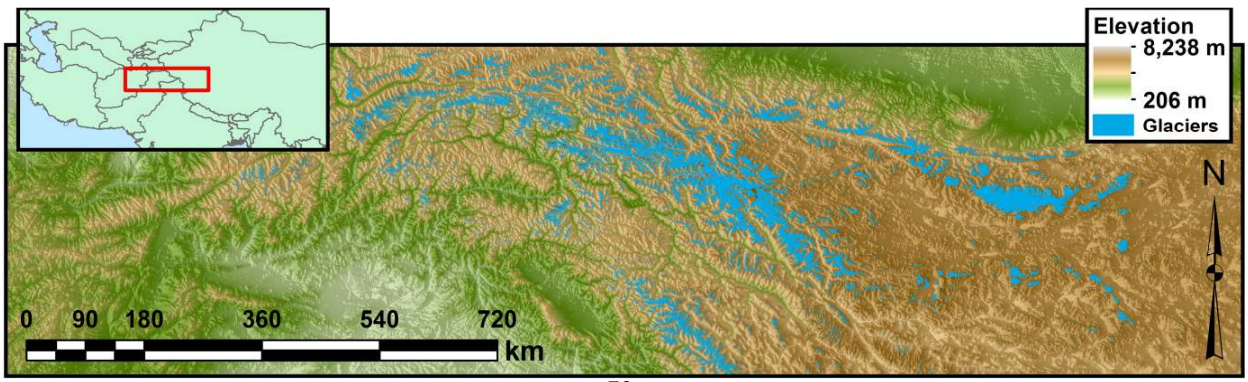


Fig. 1 - The study area for this research includes the Karakorum and Hindu Kush mountains. Elevation data are from GTOPO30 [4], glacier extents are from the Randolph Glacier Inventory [5], and country boundaries are from the Global Administrative Areas database [6].

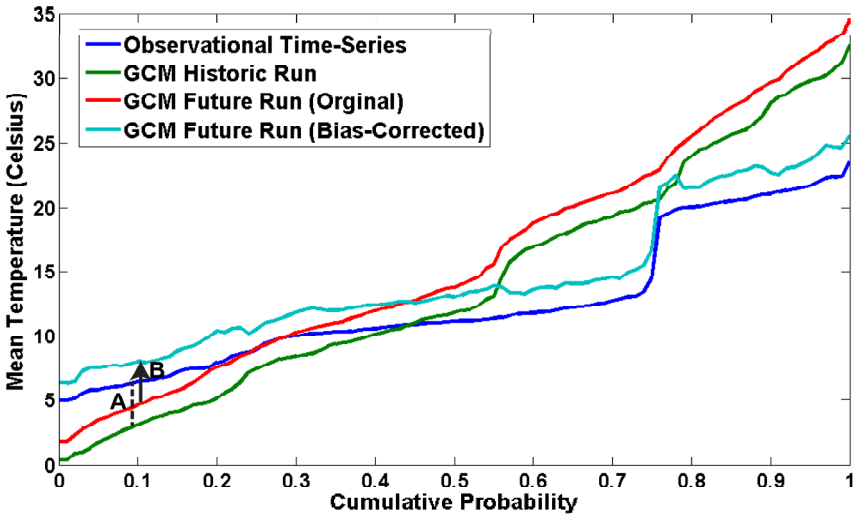


Fig. 2 – Cumulative distribution functions (CDFs) of mean temperature gridded products used in bias correction for January at the grid cell centered at (68.7 °W, 33° N). The CDFs displayed correspond to: CRU monthly time-series from 1950-2000 (denoted *observational time-series*), GISS historic run for 1950-2000 (denoted *GCM historic run*), GISS GCM projection run for RCP 4.5 from 2020-2010 (denoted *GCM Projection Run (original)*), and the bias-corrected version of the GISS GCM projection run for 2020-2100 (denoted *GCM Projection Run (bias-corrected)*). The dotted line, *A*, refers to the transfer function at the 0.1 CDF value and the arrow, *B*, refers to the application of the transfer function at the 0.1 CDF value.

The form of bias correction used is quantile mapping (an example is shown in Fig. 2). The first step of the quantile mapping procedure is to calculate cumulative distribution functions (CDFs) for each gridded dataset used and at all grid locations. Each CDF is composed of the time-series elements over a range of years, for a 3-by-3 cell window centered on the cell being bias-corrected. These CDFs are first produced for the GCM historic run and Climate Research Unit (CRU) [7] gridded observational datasets at all cells in the region being downscaled for the years 1950-2000. A transfer function is then calculated from the GCM CDF to the CRU CDF at each percentile of the CDFs (an example value of the transfer function for the cumulative probability of 0.1 is denoted by the dotted line, A, in Fig. 2). The resulting transfer function is then applied to a projection run from the same GCM (example for cumulative probability of 0.1 denoted by the arrow, B, in Fig. 2), creating a bias-corrected GCM projection time-series.

The downscaled data are produced both with and without the bias correction step in order to quantify the uncertainty in the bias-correction step. There are several other sources of uncertainty related to the current downscaling method. For example, as is shown in Mosier et al. [1], errors in the hindcast output from the downscaling package are strongly correlated to errors in WorldClim [8], the reference climatology dataset used in the downscaling package. Additionally, WorldClim is a representation of climatic conditions from 1950-2000. These precise high-spatial resolution relationships (e.g. relating to lapse rates) are expected to change in the future due to larger-scale changes in the climate, such as changes in the lapse rate feedback [9]. Despite these assumptions and uncertainties, the 30-arcsecond downscaled projected GCM data produced using the downscaling package add useful information about the spatial distribution of climatic conditions within each GCM cell. For example, the downscaled monthly precipitation and mean temperature GCM data are useful in studying projected changes in potential snowfall over the twenty-first century.

III. EXAMPLE APPLICATION

As an example of these downscaled data, an ensemble of downscaled data are produced for a region encompassing the Karakorum and Hindu Kush (KHK) mountains (Fig. 1), which is a region where seasonal snowpack and glacier mass balance contribute significantly to seasonal stream flow [10]. Ensemble members include the MPI-ESM-MR (developed by the Max Plank Institute, GISS-E2-H (developed by the Goddard Institute for Space Studies), and BCC-CSM (developed by the Beijing Climate Center), which are all members of CMIP5 [2]. Monthly precipitation and mean temperature data from these GCMs are downscaled for the 4.5 and 8.5 representative concentration pathways (RCPs) [11]. Monthly precipitation and mean temperature data from these GCMs are downscaled for 2020-2100 and processed to form a metric which the authors refer to as potential snowfall. Potential snowfall is defined as the incident precipitation in cells where the temperature is at or below 0 °C. A result of this analysis is that the annual potential snowfall for the ensemble mean (i.e. all three GCMs and both RCPs) decreases by approximately 17% from 2020 to 2100 and 20% between 2100 relative to the 1950-2000 hindcast climatology.

The changes in potential snowfall for the KHK region are very spatially heterogeneous. These spatial gradients are not present if the GCM data are compared at the original resolution. Assessing potential snowfall using only the

original resolution GCM data, these high spatial resolution changes would not have been detectable. Knowing the spatial distribution is of snowfall is important because the energy balance of snow varies substantially by location, which changes the length of snow storage and timing of its release [12].

Acknowledgment

This research has been supported by the National Science Foundation (award 1137272), the Glumac Faculty Fellowship, and the Oil Spill Recovery Institute.

References

[1] Mosier, Thomas M., David F. Hill, and Kendra V. Sharp. "30-Arcsecond monthly climate surfaces with global land coverage." *International Journal of Climatology*(2013).

[2] Taylor, Karl E., Ronald J. Stouffer, and Gerald A. Meehl. "An Overview of CMIP5 and the Experiment Design." *Bulletin of the American Meteorological Society* 93, no. 4 (2012).

[3] Fowler, H. J., S. Blenkinsop, and C. Tebaldi. "Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling." *International Journal of Climatology* 27, no. 12 (2007): 1547-1578.

[4] DAAC, LP. "Global 30 Arc-Second Elevation Data Set GTOPO30. Land Process Distributed Active Archive Center." (2004).

[5] Arendt, A., Bolch, T., Cogley, J. G., Gardner, A., Hagen, J. O., Hock, R., et al. (2012). Randolph Glacier Inventory [v2. 0]: A Dataset of Global Glacier Outlines. Global Land Ice Measurements from Space, Boulder Colorado, USA.*Digital Media*.

[6] Hijmans, R. "GADM database of Global Administrative Areas. Version 2. University of Berkeley, CA, US, and the International Rice Research Institute, Los Baños, the Philippines." (2011).

[7] Harris, I., Jones, P. D., Osborn, T. J., & Lister, D. H. "Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset." *International Journal of Climatology* (2013).

[8] Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. "Very high resolution interpolated climate surfaces for global land areas." *International journal of climatology* 25, no. 15 (2005): 1965-1978.

[9] Andrews, Timothy, and Piers M. Forster. "CO2 forcing induces semi-direct effects with consequences for climate feedback interpretations." *Geophysical Research Letters* 35, no. 4 (2008).

[10] Winiger, M., G. H. Y., M. Gumpert, and H. Yamout. "Karakorum–Hindukush–western Himalaya: assessing high-altitude water resources." *Hydrological Processes* 19, no. 12 (2005): 2329-2338.

[11] Stocker, T. F., D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley. "Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change." (2013).

[12] DeWalle, David R., and Albert Rango. *Principles of snow hydrology*. Cambridge: Cambridge University Press, 2008.

Building the Semantic Web for Earth Observations

Martina Baučić
Faculty of civil engineering, architecture and geodesy
University of Split
Split, Croatia
martina.baucic@gradst.hr

Damir Medak
Faculty of geodesy
University of Zagreb
Zagreb, Croatia
damir.medak@geof.hr

***Abstract*—The growth in number of Earth sensors and increase in data volumes have raised a problem of observations integration, data analysis and reasoning over the integrated data. The two initiatives are building an interoperable environment for Earth observations: the Sensor Web Enablement and the Semantic Sensor Web. The standards for web services and observation encodings are resolving syntactic interoperability between sensors. The Semantic Web standards are enriching observations with description of data semantics and thus improving data integration. The paper demonstrates the building of Semantic Web for Earth observation data. It explains development of meteorological data ontology and provides an example of transforming meteorological data into Resource Description Framework data model. Although at the very beginning, the current implementations have proved the Semantic Web as an emerging technology for Earth observations integration and web computational modelling.**

Keywords—Earth observations; Sensor Web; Semantic Web; meteorology; ontology; RDF data model

I. INTRODUCTION

Key emerging trends in Earth observations include the growing number of sensors, growth in data volume, real-time processing, distribution via Web, crowdsourcing etc. There is a need for integration of Earth observation data coming from various sensors that will allow analysis and reasoning over integrated data. These trends can be found in reports such as the Report on future trends in geospatial information management by UN Committee of Experts on Global Geospatial Information Management [1].

The work presented here discusses current efforts in building the Semantic Web as an interoperable environment for Earth observations. The Open Geospatial Consortium (OGC) initiative called Sensor Web Enablement (SWE) has defined open standards for exploiting Web connected sensors. The standards include encodings for describing sensors and sensor observations, and interface definitions for web services. The syntactic interoperability is achieved by adoption of these standards, but semantics of observations remain ambiguous. The Semantic Sensor Web initiative by World Wide Web Consortium (W3C) extends SWE standards with spatial, temporal, and thematic description of observations by ontologies. There are three main reference ontologies for building Earth observations ontologies. W3C Semantic Sensor Network (SSN) ontology models sensor devices, systems and processes; W3C Time ontology models temporal concepts such as instants, intervals, durations etc.; OGC GeoSPARQL standard models geospatial objects and their topological and geometrical properties. To build domain ontology, such as ontology for meteorological data, one should define basic concepts in the domain and relations among them. To enable integration of data from various domains, domain ontology concepts should be linked to concepts in reference upper ontologies.

Implementation of the Semantic Web technologies for Earth observations is at the very beginning. There are projects such as European research project TELEIOS that builds Virtual Earth Observatories [2], or National Aeronautics and Space Administration project that builds Semantic Web for computational modeling of the impacts of changing climate [3]. Looking at scope of the projects and organizations that implement them, the Semantic Web will be emerging technology not only in integration of Earth observations but also in web computational modelling of geospatial and temporal data.

The remainder of this paper is organized as follows. Section 2 briefly describes the OGC Sensor Web Enablement initiative. Section 3 provides main information about the Semantic Sensor Web initiative by W3C. Section 4 explains the development of meteorological data ontology. It also provides an example of transforming meteorological data into Resource Description Framework (RDF) data model. Finally, we present the conclusions.

II. SENSOR WEB ENABLEMENT INITIATIVE

There are millions of sensors, in and around the Earth, collecting massive amounts of data. Sensors could be from a barometer at fixed location to hyper-spectral sensor on board of a satellite. Each sensor observes a certain condition (wind, pressure, etc.) in a particular place and time. This spatio-temporal information is stored on the sensor or directly sent to server, but having its own data format and software for processing, and its own semantics. Overwhelming number of observations must be processed and explained, and thus we need interoperability between the heterogeneous sensor data and applications.

The OGC SWE initiative is developing the global standards to enable discovery, exchange, processing of observations and controlling of sensor systems via the Web. The standards include encodings for describing sensors and observations, and

interface definitions for Web services. The built and prototyped SWE standards include the following [4]:

- Observations & Measurements Schema
- Observations and Measurements XML encoding
- Sensor Model Language
- Sensor Observations Service
- Sensor Planning Service
- SWE Common Data Model
- SWE Services Common
- PUCK Protocol Standard
- Sensor Alert Service
- Web Notification Services

The SWE enables interoperability between sensors, models and decision support systems as shown on Fig.1. It is a middleware layer that provides description and discovery of sensor assets and capabilities, access to data, tasking of sensors, and subscription to alerts. The goal of SWE is a distributed sensing system in which information is globally shared and used by all networked clients. Some current SWE implementation efforts are listed in [4]. We will mention the organization 52North that provides a complete set of SWE services under General Public License.

However, while the syntactic interoperability is achieved by adoption of the SWE standards, the semantics of observations remain ambiguous. Also, the SWE standards do not provide a basis for reasoning that can ease development of advanced applications for discovery and retrieval of sensor data.

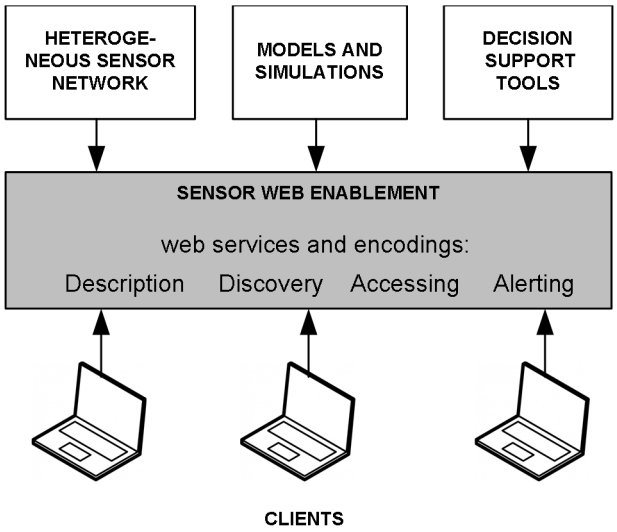


Fig. 1. SWE framework

III. SEMANTIC SENSOR WEB INITIATIVE

The Semantic Web is an extension of the Web facilitating users to find, share, and combine information more easily. It is a vision of "Web of data" that can be readily interpreted by machines, instead of today "Web of documents" that can be read by people. Semantics, or meaning, of information on the Web is formally defined by ontologies. The Semantic Web stack builds on the W3C standards: Resource Description Framework (RDF), SPARQL Protocol and RDF Query Language (SPARQL), Ontology Web Language (OWL), Extensible Markup Language (XML), and Uniform Resource Identifier (URI). These technologies provide machine-readable descriptions of the content of Web documents and reasoning algorithms for automated information search.

To improve semantic interoperability and integration of sensor data, the SWE initiative is enriched with Semantic Web technologies. The Semantic Sensor Web initiative by W3C extends SWE standards with spatial, temporal, and thematic description of observations by ontologies. These ontologies allow integration, classification and reasoning over the sensors data and observations.

The Semantic Sensor Networks Community Group is developing Semantic Sensor Network (SSN) ontology which models sensor devices, systems, processes, and observations [5]. The SSN ontology is domain independent and it merges sensor-focused, observation-focused and system-focused views. It is aligned with the DOLCE Ultra Lite (DUL) upper ontology to facilitate reuse and interoperability. The SNN is a formal OWL Description Logic ontology available as single OWL file [6]. It consists of 41 classes and 39 properties. Fig. 2 shows a small part of SSN ontology with a central concept: *Sensor*, as the broadest concept of any entity capable of sensing. The nine SNN classes (shown on Fig. 2 as white ovals) are connected by properties. The property *subClassOf* (shown on Fig. 2 as arrow with no filled head) means: e.g. any member of *Device* class is a member of *System* class, and of *Physical object* class. The arrows with filled heads show various properties, but their names are omitted from Fig. 2 due to figure size limits. E.g. *observesOnly* is the property linking *Sensor* with *Property*. Some classes are linked with the upper ontology classes of DUL ontology (shown on Fig. 2 as grey ovals). E.g. *Property* class is subclass of *Quality* class.

There are concepts not described with the SSN: e.g. units of measurements, locations, features and property hierarchies. The idea is that knowledge engineers of particular domain include domain feature ontology, location and units ontology by linking them to SSN ontology. E.g. SSN ontology is combined with NASA SWEET (Semantic Web for Earth and Environmental Terminology) ontology modelling the Earth observed properties.

Although recently published, the SNN ontology is already being used in several projects. The examples and uses of the SSN ontology are given in [7]. Some of them are: SENSEI and SPITFIRE projects in the EU's Seventh Framework Programme; the projects of the Kno.e.sis Centre at the Wright State University; the projects of the 52North organization and the SemsorGrid4Env project. Linked Sensor Data and Linked Observation Data are projects of the Kno.e.sis Centre. The

projects RDF datasets contain description of circa 20.000 weather stations and hurricane observations in the USA since 2002. The datasets are part of the Linked Open Data. These projects have shown that the use of SNN ontology is enabling integration of sensor data with other data and applications relying on Semantic Web technologies like RDF and SPARQL.

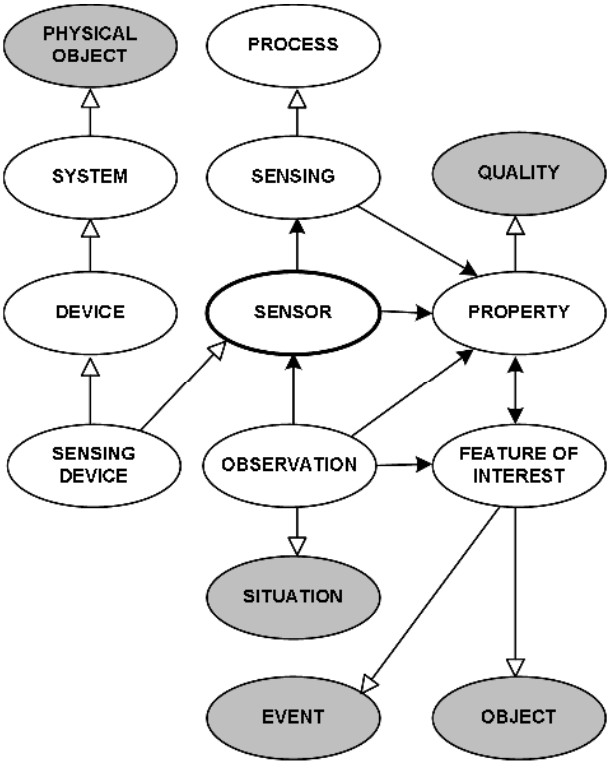


Fig. 2. Part of SSN ontology aligned with the DOLCE Ultra Lite ontology classes (colored in grey)

IV. Development of Meteorological Data Ontology and RDF Database

Ontology represents knowledge of a domain as a hierarchy of concepts (also called classes), their properties (also called attributes) and relationships. Ontology languages are used to construct ontologies. The current W3C standards are: OWL, a formal language based on description logics; RDF; and RDF Schema. Domain ontology represents concepts of a particular domain. Upper ontology represents concepts applicable across a range of domain ontologies (e.g. SUMO or DOLCE ontology). To enable integration of data from various domains, the domain ontology concepts should be linked to concepts in the reference upper ontologies. In addition to taxonomic hierarchies of classes and properties, the ontology can state axioms constraining the possible interpretations and describe the logical inferences that can be drawn from asserted data.

Several methodologies are guiding experts in the process of ontology building, but there are two main steps (Fig.3). In first step, an expert models the domain knowledge: define basic concepts and relations among them, and define axioms and rules for data interpretation and reasoning. The second step is to link the domain concepts with concepts in reference upper ontologies. One should consider the reuse of the already developed ontological resources.

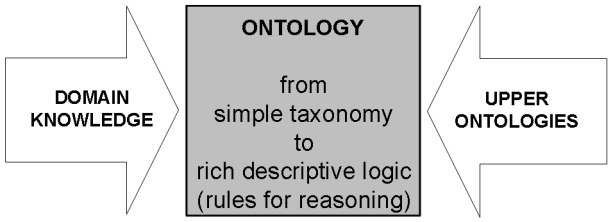


Fig. 3. Ontology building process

The Croatian Meteorological and Hydrological Service is publishing meteorological data in XML files. Fig. 4 shows an excerpt from the XML file. Each file contains 8 meteorological observations from 38 weather stations for a particular date and hour.

```
<?xml version="1.0" encoding="UTF-8"?>
- <Hrvatska>
  - <DatumTermin>
      <Datum>24.05.2014</Datum>
      <Termin>18</Termin>
    </DatumTermin>
  + <Grad>
  - <Grad>
      <GradIme>Crikvenica</GradIme>
    - <Podatci>
        <Temp>22</Temp>
        <Vlaga>68</Vlaga>
        <Tlak>1014.9</Tlak>
        <TlakTend>-</TlakTend>
        <VjetarSmjer>E</VjetarSmjer>
        <VjetarBrzina>04</VjetarBrzina>
        <Vrijeme>-</Vrijeme>
        <VrijemeZnak>-</VrijemeZnak>
      </Podatci>
    </Grad>
  - <Grad>
      <GradIme>Gorice-Nova Gradiška</GradIme>
    - <Podatci>
        <Temp>25</Temp>
```

Fig. 4. Excerpt from XML with meteorological data

Our attempt aims to facilitate the use of meteorological data by adding semantic description and offering as RDF data.

We started with by searching existing ontological resources. We have considered the W3C SSN ontology, W3C Time ontology and OGC GeoSPARQL ontology as the reference ontologies. Fig. 5 shows the links between the main concepts in the three reference ontologies. *Observation* is a subclass of *Temporal entity*, and thus it has its beginning and end. *Observation* and *Feature of Interest* are subclasses of *Feature*, and thus they have their geometries. By linking SSN ontology to GeoSPARQL ontology, the sensor concepts may have complex descriptions of their geospatial characteristics such as types of geometry, coordinate reference systems and topological relations. The Geography Markup Language

(GML) and well-known text (WKT) standards are used for geospatial data encoding.

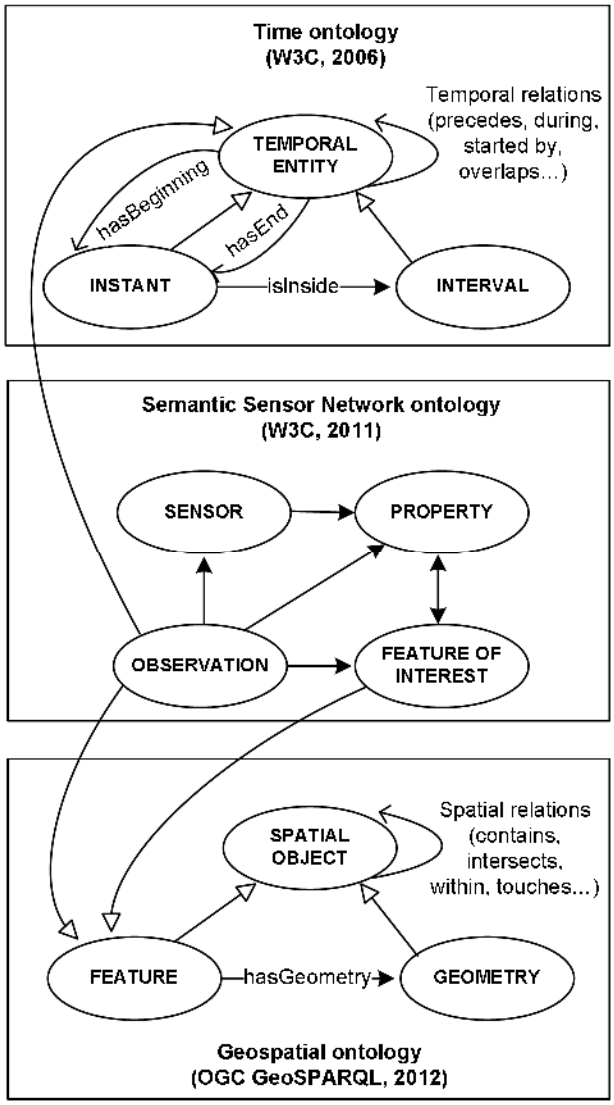


Fig. 5. Links between the main concepts of the three reference ontologies

By extending W3C SSN ontology, we have defined basic concepts and their relationships for the meteorological observations stored in XML file. Fig. 6 shows some meteorological classes, their relationships and links to W3C SSN, OGC GeoSPARQL and W3C Time ontology.

New defined meteorological classes and their relationships can be encoded in a TBox part of knowledge base. The TBox contains ontological schema describing terminology and data semantics. The definition of new class *TemperatureObservation* and its relationship with *Observation* class is written in OWL language with Turtle RDF serialization as follows:

dhmz:TemperatureSensorOutput rdf:type owl:Class.

dhmz:TemperatureObservation rdf:subclass ssn:Observation.

The prefixes *dhmz*, *rdf*, *owl* and *ssn* in the above statements are URI abbreviations (e.g. *rdf* is abbreviation of *www.w3.org/1999/02/22-rdf-syntax-ns#*). A URI provides a global identification for a Web resource.

A key feature of OWL is its ability to describe class by restricting the values allowed for certain properties. It allows us to make inferences about members of a class. The description of class *TemperatureSensorOutput* by restriction is as follows:

dhmz:TemperatureSensorOutput rdf:type owl:Class;

rdfs.subClassOf

[rdf:type owl:Restriction;

owl:onProperty ssn:hasValue;

owl:allValuesFrom dhmz:TemperatureValue].

The above statements will classify all instances as members of class *TemperatureSensorOutput* for which all values of the property *hasValue* come from class *TemperatureValue*.

Having classes and properties written in the TBox, we can encode meteorological observations in an Abox part of knowledge base. The ABox contains asserted instances. For example, air temperature of 22°C, measured at the weather station Crikvenica at 18 o'clock on May 24, 2014. This observation is encoded as follows:

TemperatureObservation_Cr_2405201418 rdf:type dhmz:TemperatureObservation;

dhmz:observationResult dhmz:TemperatureSensorOutput_1234;

geo:hasGeometry dhmz:geo_WS_Crikvenica.

dhmz:TemperatureSensorOutput_1234 rdf:type dhmz:TemperatureSensorOutput;

ssn:hasValue dhmz:TemperatureValue_1;

dhmz:hasTime dhmz:TemperatureDateTime_1.

dhmz:TemperatureValue_1 ssn:hasQuantityValue "22"^^qudt:DegreeCelsius.

dhmz:TemperatureDateTime_1 time:inXSDDateTime "2014-05-24T18:00:00"^^xsd:dateTime.

In order to add geospatial location to the above observation, the weather station Crikvenica is defined as a point with coordinates and a coordinate reference system:

dhmz:geo_WS_Crikvenica rdf:type geo:Point;

geo:asWKT "<http://www.opengis.net/def/crs/EPSG/0/3765> POINT(35787.4 5005291.0)"^^geo:wktLiteral.

The TBox and ABox use the same RDF data model and the same OWL encoding language. The data and their description (semantics) are stored together and can be queried together by SPARQL. Moreover, sensors data from other sources can be converted to RDF, merged into one federated RDF database, and queried together with their semantics.

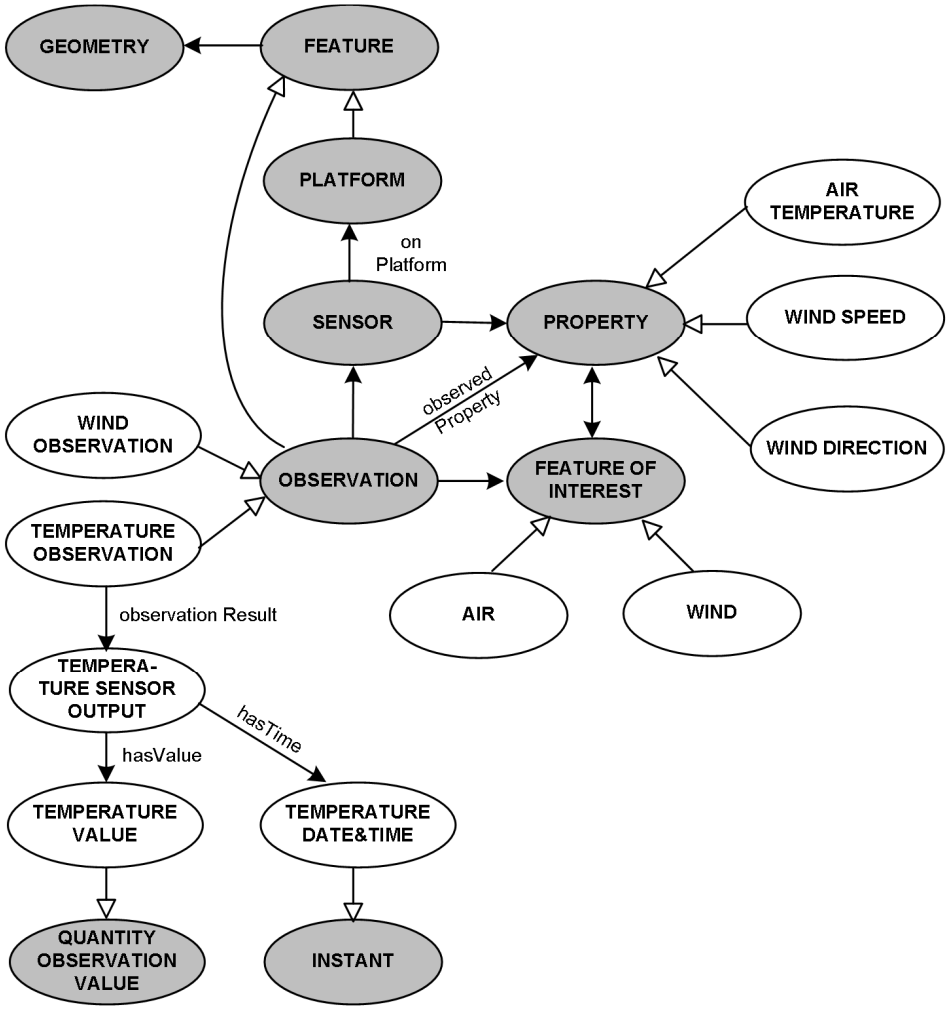


Fig. 6. Some meteorological classes (white ovals), their relationships and links with SSN, GeoSPARQL and Time ontology classes (grey ovals)

An example of SPARQL query over federated RDF database is presented bellow. The query shows which weather stations (labeled with *?ws*) are within which national parks (labeled with *?np*).

```
SELECT ?ws ?np
WHERE {
?ws rdf:type dhmz:Wather_station;
    geo:hasGeometry ?geo_ws.
?np rdf:type hrnp:National_park;
    geo:hasGeometry ?geo_np.
?geo_ws geo:sfWithin ?geo_np.
}
```

In the previous example, the federated RDF database is merged from two imaginary RDF databases: *dhmz* (could be the RDF database of The Croatian Meteorological and Hydrological Service) and *hrnp* (could be the RDF database of Croatian Registry of National Protection Areas). GeoSPARQL property *sfWithin* defines topological relations between weather stations and national parks. The example clearly demonstrates the power of RDF data model in integrating data which could be used for the integration of Earth observations.

V. CONCLUSIONS

Our attempt was to explain and demonstrate the building of Semantic Web for Earth observation data. The new technologies are emerging and able to integrate, process, and explain overwhelming number of observations. The current efforts encompass two initiatives: the OGC Sensor Web Enablement and the W3C Semantic Sensor Web. Recently developed standards are already successfully implemented throughout many projects and it seems the Semantic Web technologies will take a significant place in integration of sensor data and applications.

In this paper we have briefly described Semantic Web concepts and we have demonstrated a domain ontology development combining thematic, spatial and temporal ontologies. The meteorological data available in XML files

published by the Croatian Meteorological and Hydrological Service is converted into RDF data model. Enriched with semantics, the meteorological data can effectively be used with data from other sources. The example of SPARQL query demonstrates the integration of data from two RDF databases and use of the GeoSPARQL topological relation property *stWithin*.

Instead of commonly used W3C Basic Geo Vocabulary standard which can only describe points with latitude, longitude, and altitude in the WGS84 coordinate reference system, we have used GeoSPARQL standard which provides complete semantics of geospatial objects and their spatial relations.

In our future work we will explore qualitative spatial reasoning over the integrated Earth observation data by building more complex OWL models.

REFERENCES

[1] UN Committee of Experts on Global Geospatial Information Management, Future Trends in Geospatial Information Management: The Five to Ten Year Vision, J. Carpenter and J. Snell, Eds. UN, July 2013.

[2] Project TELEIOS - Virtual Observatory Infrastructure for Earth Observation Data. (*http://www.earthobservatory.eu*)

[3] NASA Earth, Life and Semantic Web (ELSeWeb) project. (*https://earthdata.nasa.gov/our-community/community-data-system-programs/access-projects/earth-life-and-semantic-web-elseweb*)

[4] Open Geospatial Consortium, OGC® Sensor Web Enablement: Overview And High Level Architecture, C. Reed, M. Botts, G. Percivall, and J. Davidson, Eds. 2013.

[5] W3C Incubator Group, Semantic Sensor Network XG Final Report, L. Lefort, C. Henson, and K. Taylor, Eds. 2011.

[6] W3C Incubator Group, Semantic Sensor Network ontology OWL file, 2011. (*www.w3.org/2005/Incubator/ssn/ssnx/ssn.owl*)

[7] M. Compton, P. Barnaghi, L. Bermudez, et al., "The SSN ontology of the W3C semantic sensor network incubator group" in Journal of Web Semantics, vol. 17. Elsevier, 2012, pp 25-32.

Spatio-Temporal Interpolation of Soil Moisture, Temperature and Salinity (in 2D+T And 3D+T) Using Automated Sensor Networks

[abstract]

C.K. Gasch, T. Hengl, D. Joseph Brown, B. Graeler
ISRIC - World Soil Information, Wageningen, Netherlands

***Abstract*— Comprehension of dynamic soil properties at the field scale requires measurements with high spatial and temporal resolution. Sensor networks provide frequent in situ measurements of soil properties at fixed locations, providing data in 2- or 3-dimensions and through time (3D+T). Spatio-temporal interpolation of 3D+T point data produces continuous estimates (maps) that can then be used for prediction at unsampled locations, as input for process models, and can simply aid in visualization of properties through space and time. Regression-kriging with 2D+T data has successfully been implemented for a daily air temperature dataset using terrain and temporal imagery as covariates. In this paper, we extend that approach to develop models for mapping soil moisture, temperature, and salinity using regression-kriging on 3D+T data. Currently, the field of geostatistics lacks an analytical framework for modeling 3D+T data, so our long term objective is to develop robust 3D+T models for mapping soil data that has been collected with high spatial and temporal resolution.**

For this analysis, we use the Cook Farm dataset, which includes hourly measurements of soil volumetric water content, temperature, and electrical conductivity at 42 points and five depths, collected over five years. Cook Farm is a 37 hectare experimental farm in the northwest of the United States; it is host to variable soils, cropping systems, microclimates, and landscape positions. The dataset also includes a digital elevation model, topographic wetness index, soil unit description map, daily meteorological data, and annual satellite imagery (for 2011-2013). The sensor data, combined with the spatial and temporal covariates, provide an ideal dataset for developing 3D+T models. The presentation will include preliminary results and address main implementation strategies.

Influence of the Different Precipitation Interpolation Methods on Jadar River Discharge

[abstract]

Marija Ivković, Aleksandra Kržič
Hydrometeorological Service of Serbia, Belgrade, Serbia

Albrecht Weerts
Institut Deltares, Delft, Netherlands

***Abstract*— The frequent occurrence of heavy rainfall in Serbia during June and October causes flash floods in many areas resulting withsignificant economic losses. Consequently, themost partof the annual local community budget is diverted to the reconstruction of the disasters caused by floods every year. In RHMSS, through the DRIHM project, aHBV model, distributed hydrologic model is in the process of assessment as a tool for early warning of the flash floods. This model is a part of the OpenStreams, anopen source distributed hydrologic modeling environment. Here we use OpenStreams HBV-96 setup embedded in OpenDA and based on the open source python-pcraster library. Dynamic input data for the model are gridded time series of the observed precipitation, mean temperatures and potential evaporation. Representation of the catchment characteristics is given through the digital elevation map, land cover as well as soil map and the river network is calculated from the topography map. Precipitation well-distributed in time and space is vital part of the input data preparation, especially when there are not sufficiently long time series of measured precipitation data available. In this study we are examining sensibility of the HBV model to different interpolation methods for precipitation distribution on Jadar river basin. The basin is selected on the basis of its position and exposure to southwestern circulation bringing convective cloudiness and significant precipitation amounts. Several interpolation methods were chosen from Python extension NumPy. For every method gridded precipitations were used as input to HBV, and time series of discharge are produced. Given the fact that hourly precipitation data are available from only two main meteorological stations, the verification of the interpolated precipitation is done by comparison with daily accumulated amounts from precipitation stations. The dissimilarities are also presented through the difference between measured and simulated discharge.**

The Development of Common Gridded Climate Database Through Regional Cooperation- CARPATCLIM

[abstract]

Dragan Mihić
Hydrometeorological Service of Serbia, Belgrade, Serbia

***Abstract*— The main aim of the CARPATCLIM (Climate of the Carpathian Region) project was to make gridded climatological database and Digital Climate Atlas of the larger Carpathian Region as one of the final products, freely available. The data are for the period between 1961-2010, on 0.1^{o} spatial and daily temporal resolution. A dozen of essential meteorological variables, together with variety of climate indices are presented with downloading possibility. The homogenization, harmonization and interpolation of data were carried out through regional cooperation, overcoming some common problems with the use of unified methods. MASH (Multiple Analysis of Series for Homogenization) software was used in each of the participating countries for quality control, homogenization and harmonization of observed climatological data, while MISH (Meteorological Interpolation based on Surface Homogenized Database) was applied for spatial interpolation. The Digital Climate Atlas was developed as a Web GIS application based on modern web standards. Additionally, one very important product is the Metadata Catalog, designed as a searching tool of metadata database by various parameters. It contains all the metadata developed within the project, together with the original observations metadata. All the work was performed with the objective to contribute to availability and accessibility of both climate data and metadata, in order to be used for studies of regional climate variability and regional climate change as well as studies in applied climatology..**

Web GIS for the Carpathian Region Climate Analysis

[abstract]

Igor Antolović, Vladan Mihajlović, Dejan Rančić
Faculty of Electronic Engineering, Niš, Serbia

Dragan Mihić, Vladimir Djurdjević
Hydrometeorological Service of Serbia, Belgrade, Serbia

Abstract— **Acquisition and analysis of meteorological data is crucial for prediction of climate changes, drought and flood. Concerning that this dataset contains a geospatial component and covers long periods of time it requires a specialized GIS (Geographic Information System) for suitable processing and visualization. Also in order to make these data available to a wider scientific community it is recommended to provide a Web based GIS solution (Geoportal ARSO 2013, Czech Republic Climate Atlas 2013). The CARPATCLIM (Climate of the Carpathian Region) Web GIS represents the main entry point for accessing, visualization and analysis of all relevant metrological data acquired on the CARPATCLIM project. The main goal of this project was to create a gridded climatologic database for the Carpathian region in a daily temporal resolution for the period 1961-2010 by using 0.1° spatial resolution. The dataset includes fourteen essential meteorological variables (temperature, precipitation, pressure, global radiation, wind speed and direction, etc.) and variety of climate indices (Standardized Precipitation Index, Palfai and Reconnaissance Drought Index, Palmer Drought Severity Index etc.). The basic functionality of CARPATCLIM Web GIS is to display a particular variable for a particular day as a raster colored grid or as a set of isocontours, along with the automatically calculated information about minimal and maximal values inside the displayed grid. The grid is underlaid with a raster relief map and a vector map of country borders. Advanced features include visualization of averaged grid values for a custom defined time period. Time period can be defined on a monthly, yearly basis or for a particular day, month over a specified year span. All calculations perform in real time due to an efficient database organization resulting in fast data retrieval. The results can be downloaded as an image or in standard raster data format, which is suitable for further processing.**

Predicting Daily Air Temperatures by Support Vector Machines Regression

[full paper]

Miloš Marjanović
Faculty of Mining and Geology
University of Belgrade
Belgrade, Serbia
milos.marjanovic@rgf.bg.ac.rs

***Abstract* — This paper demonstrates an attempt to apply Support Vector Machines regression (SVMr) on meteorological data. The objective was to predict daily air temperatures for continental part of Europe for 1.1.2011. Supplied training set was based on temperature records of 357 weather stations throughout Europe and additional attributes (extracted from appropriate grid maps), comprising of MODIS satellite image for 1.1.2011, elevation, coastline distance and surface insolation. The SVMr algorithm was then optimized by means of 10-fold Cross Validation and obtained parameters were used to learn a regression rule and expand it to the entire continent. The chosen data resolution of 0.036 arc degrees (approximately 4 km) was sufficient to give rough temperature estimation on the specified date. A separate validation set of another 200 weather stations throughout Europe was supplied to evaluate the modeling performance. Resulting RMSE of about ±2°C proved that there is a good potential of applying SVMr or similar Machine Learning based techniques for extrapolating measured temperature data.**

Keywords — air temperature; Support Vector Machines; regression; MODIS; Europe

I. INTRODUCTION

Air temperature maps, alongside with precipitation and cloud cover maps, are probably the most common weather forecast products used for informing the public on a daily bases. They are also used in larger scales in agriculture, industry, traffic (especially air traffic), energy consumption analysis, natural hazards, tourism etc. Short-term (several hours) to mid-term (several days) temperature forecasts are usually of the utmost interest, and numerous researchers have shown how these could be successfully predicted by various numerical deterministic, geostatistical and soft computing techniques [5, 8]. The latter appeared as an alternative to conventional meteorological and climatic models and have turned particularly applicable in cases with multiple meteorological parameters and multiple weather scenarios which are more difficult to cope with by deterministic (finite-element-based or finite-difference-based, such as HIRLAM) models [8]. This is a reasonable and logical trend from another aspect – synoptic meteorological measurements are becoming more available. There are numerous Earth Observation systems nowadays that provide meteorological data with higher temporal and spatial resolutions on global scales. On the other hand, there are local or site-specific observatories specialized in collecting weather conditions at points of interests, with unprecedented accuracy and resolution.

Soft computing techniques for predicting air temperature or other meteorological variables, such as solar radiation, precipitation, wind speed etc, are usually based on neural network computing, including Multi-layer perceptrons, Radial Basis Functions, regression networks and similar techniques [6]. Some researchers have even compared different techniques and highlighted importance of synoptic information and selection of particular scenarios [2, 9]. However, novel approaches are indicating that regression Support Vector Machines (SVMr) technique is more accurate and more reliable for prediction of various meteorological parameters [3, 4, 10, 11].

This paper is an example of such SVMr application and focuses on spatial prediction (spatial interpolation) of peak air temperature for a single day for continental part of Europe (Fig. 1), based on publicly available open data (Table 1). One of the foci is also experimenting with possibilities to include some less-common, non-meteorological spatial variables that could be well correlated with air temperatures, such as various morphometric parameters, land cover type, etc.

II. DATA AND PREPROCESSING

Typical predictor variables for large scale prediction of air temperature include max/min temperature, precipitation, relative humidity, sunshine duration, cloud cover, air pressure and so forth [1]. Synoptic situations arising from large scale circulation patterns (e.g. North Atlantic Oscillation, based on Hess-Brezowsky classification), are also suggested for long-term temperature prediction. These patterns can significantly affect the weather conditions in Central Europe throughout the year [7, 11], but this was not the case in presented research mostly because the prediction was confined to a single day. Instead, some other predictors have been considered.

TABLE I. GENERAL DATASET INFORMATION

Data	Description	Resolution	Source
air temperature	total of weather stations throughout Europe (dependent variable)	point data	GSOD and ECA&D datasets for 1.1.2011, at http://www.ncdc.noaa.gov/ and http://www.ecad.eu/
elevation	Digital terrain model from Shuttle Radar Topography Mission resampled to a coarser resolution (predictor)	100m/4km[a]	SRTM 30+ and ETOPO DEM available at http://worldgrids.org
wetness index	relation of the catchment area a and slope b $\ln(a/\tan b)$, expressing the water retention potential (predictor)	4km	derived from elevation
surface insolation	total amount of solar radiation energy received on a given surface area during a given time (predictor)	4km	derived from elevation and Sun position calculator available at http://pveducation.org/
coastline distance	euclidean distance from Europe's coastline (predictor)	4km	calculated from continents.shp available at http://www.arcgis.com/
MODIS image	LST MOD11A2 Level 2 product, with thermal bands 31 (10.78–11.28μm) and 32 (11.77–12.27μm), resampled (predictor)	1km/4km[a]	Earth Observing System Data and Information System EOSDIS http://reverb.echo.nasa.gov/reverb/
land cover	simplified CORINE Level 1 classes: urban, agricultural, forested bare, wetland and water bodies cover types, resampled	100m/4km[a]	CORINE Land Cover, European Environment Agency http://www.eea.europa.eu
latitude	geographical coordinate in degrees for each grid cell (predictor)	4km	calculated from 4km grid converted to points
longitude	geographical coordinate in degrees for each grid cell (predictor)	4km	calculated from 4km grid converted to points

[a] Original/final resolution.

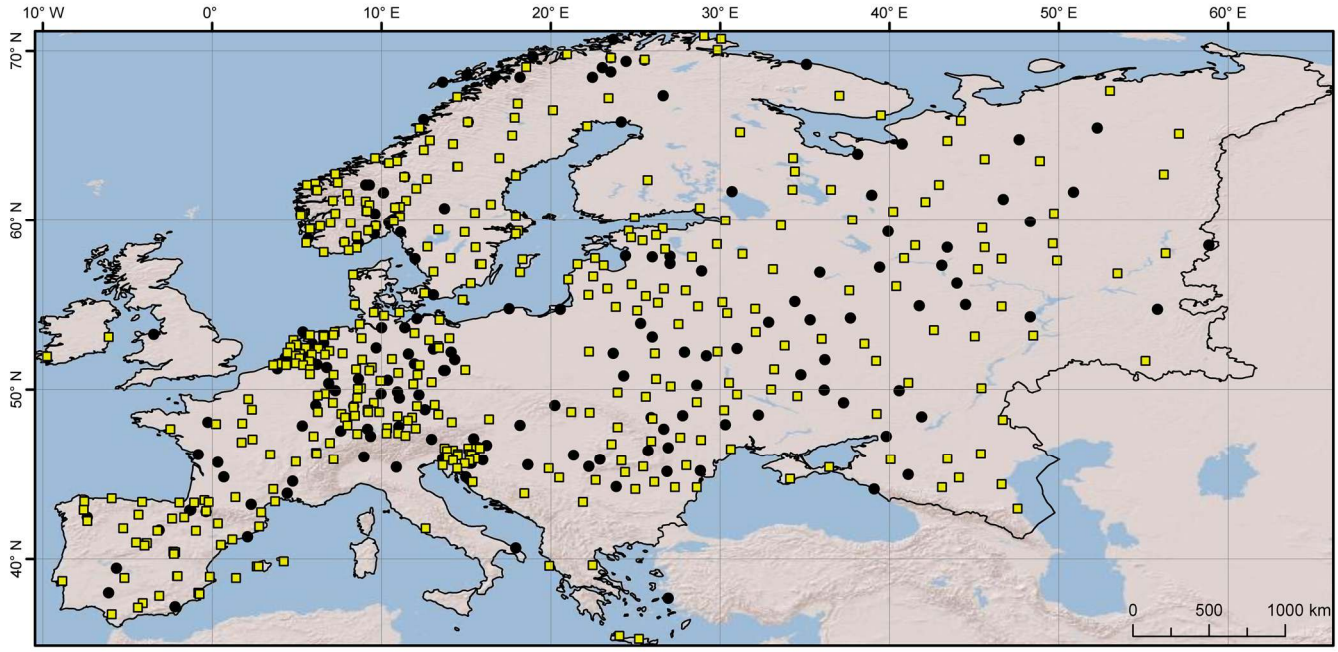


Fig. 1. Location of considered weather stations (yellow squares represent stations used fot training and black dots represent stations used for validation)

The dataset included the following predictors: ground elevation, topographic wetness index, surface insolation, coastline distance, MODIS satellite image, land cover and lat/long coordinates (Table 1).

It has been reported [4, 5] that DTM derivatives are usually well correlated with air temperature or its gradient. In addition to commonly familiar temperature dependence on elevation, temperature can depend on the type of the surface material and surface morphology or roughness. In this respect, wetness index and land cover types are included as predictors to simulate the influence of different material type and different water content, respectively, while surface morphology is modeled by insolation or surface solar reflectance. Insolation depicts distribution of sunlight and shadow, i.e. depicts differences in exposure to the sunlight. It is also well known that large water bodies impose gradual inland shifts in air temperature, which is why the distance from the coastlines is included as a predictor. Another well known fact is the difference in temperature on different locations on the Earth's surface due to its geometry and incident angles of the Sun. This was the underlying reason for including the lat/long coordinates, as they allocate position on the (spherically approximated) Earth's surface. Finally, MODIS satellite image from TERRA satellite, MOD11A2 product in particular, was included. It records the Earth's surface thermal emissivity, which directly influences the near surface air temperature.

The weather station data were obtained by combining two publically available repositories. The first included Global Surface Summary of Day (GSOD) dataset, archived at NOAA's National Climatic Data Center (NCDC) under the code NCDC DSI-9618, while the second included data from European Climate Assessment and Dataset (ECA&D) project. For the purpose of this research only maximum daily temperatures for 1.1.2011 were used (out of dozens of available meteorological parameters). The data from GSOD and ECA&D have been merged and cleaned, giving the final dataset, which was reduced to the total of 544 stations, wherein 376 stations were used for training protocol and 168 for validation (Fig. 1). Elevation data and its derivatives were obtained from Shuttle Radar Topography Mission (SRTM) at 100m resolution. Enclosed free access MODIS image, i.e. MOD11A2 Land Surface Temperature product with 1km resolution, was obtained from NASA's Earth Observing System Data and Information System (EOSDIS). Publicly available 100m CORINE land cover map issued by European Environment Agency, was simplified to slightly modified Level 1 CORINE classification.

All predictors were processed as raster grids and all were resampled to 4km resolution by Nearest Neighbor interpolation (original values of continual data are preserved). Preprocessing further included normalization of continual ordinal grids and splitting of nominal land cover grid into dummy variables (binary grids of each land cover class). Operative predictor dataset included 13 grids in total. Final preprocessing included converting grids to points, as the most suitable format for available software solution. SAGA GIS 2.0.8 was used for (pre)processing, evaluating and visualizing, while the SVMr experiments were placed in WEKA 3.7.

III. PREDICTION AND EVALUATION METHOD

The modeling method was based on the Support Vector Machine regression problem [4]. It implied learning the regression function $y=f(\mathbf{x})$ at known training instances containing a reasonably small number of $V(\mathbf{x},y)$ vectors containing $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2),..., (\boldsymbol{x}_n, y_n)\}$ (where $\boldsymbol{x}_i$ represents predictors and y_i temperature values in this particular case) and expanding it to all remaining n instances given the consistent predictor dataset. The regression function is chosen as the best separating function within a given family of functions F. Particular version of used SVMr algorithm was epsilon-SVMr, with Radial Basis Function (RBF) kernel function implementation, which projects the original feature space to high-dimensional feature space. Basic epsilon-SVM regression implies linear form $f(\mathbf{x})=\mathbf{w}\mathbf{x}_i+b$, where term $\mathbf{w}$ can be considered as the function's sloping and b as its offset. To choose an optimal $f(\mathbf{x})$ among Fs is to minimize a general risk function $R(f)$, (1).

$$R(f)=0.5\|\mathbf{w}\|^2+c\Sigma(|y_i-f(\mathbf{x}_i)|\varepsilon) \quad (1)$$

The loss function $|y-f(\mathbf{x})|\varepsilon$ is controlled by the parameter ε, while c is the cost constant that trades-off between the margin width $(2/\|\mathbf{w}\|)$ and the number of misfits [12]. The optimization thus implies:

$$\min(0.5\|\mathbf{w}\|^2+c\Sigma(\xi_i+\xi_i^*)), \quad (2)$$

subject to:

$$y_i-f(\mathbf{x}_i)\leq \varepsilon+\xi_i,\ i=1,...,n \quad (3)$$

$$-y_i+f(\mathbf{x}_i)\leq \varepsilon+\xi_i^*,\ i=1,...,n \quad (4)$$

$$\xi_i+\xi_i^*\geq 0,\ i=1,...,n \quad (5)$$

where ξ_i are misfit slack variables. The problem comes down to introducing the Lagrangean multipliers and maximizing the term:

$$\max(-0.5\Sigma(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\varepsilon\Sigma(\alpha_i^*+\alpha_i)+\Sigma y_i(\alpha_j^*-\alpha_j) \quad (6)$$

subject to:

$$\Sigma(\alpha_j^*-\alpha_j)=0;\ 0\leq\alpha_j,\alpha_j^*\leq c. \quad (7)$$

The optimized regression function is now:

$$f(\mathbf{x})=(\alpha_j^*-\alpha_j)(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+b \quad (8)$$

Replacing the dot product in (8) with (9), i.e. applying the RBF kernel function, final regression function is obtained (10) [4]. The entire procedure requires introducing of three parameters: ε, c and γ.

$$K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)=\exp(-\gamma\cdot\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2) \quad (9)$$

$$f(\mathbf{x})=(\alpha_j^*-\alpha_j)\exp(-\gamma\cdot\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)+b \quad (10)$$

For evaluation of the resulting models (Fig. 2,3), i.e. for their performance measurement a simple Root Mean Square measure was used in the form:

$$RMSE=((\Sigma(y_i-y_i')/y_i)/k)^{0.5} \quad (11)$$

wherein y_i is a temperature value measured at validating weather station and y_i' is predicted temperature value, while k represents the number of validating instances (k=168 in this case). $\alpha + \beta = \chi.$ (1) (1)

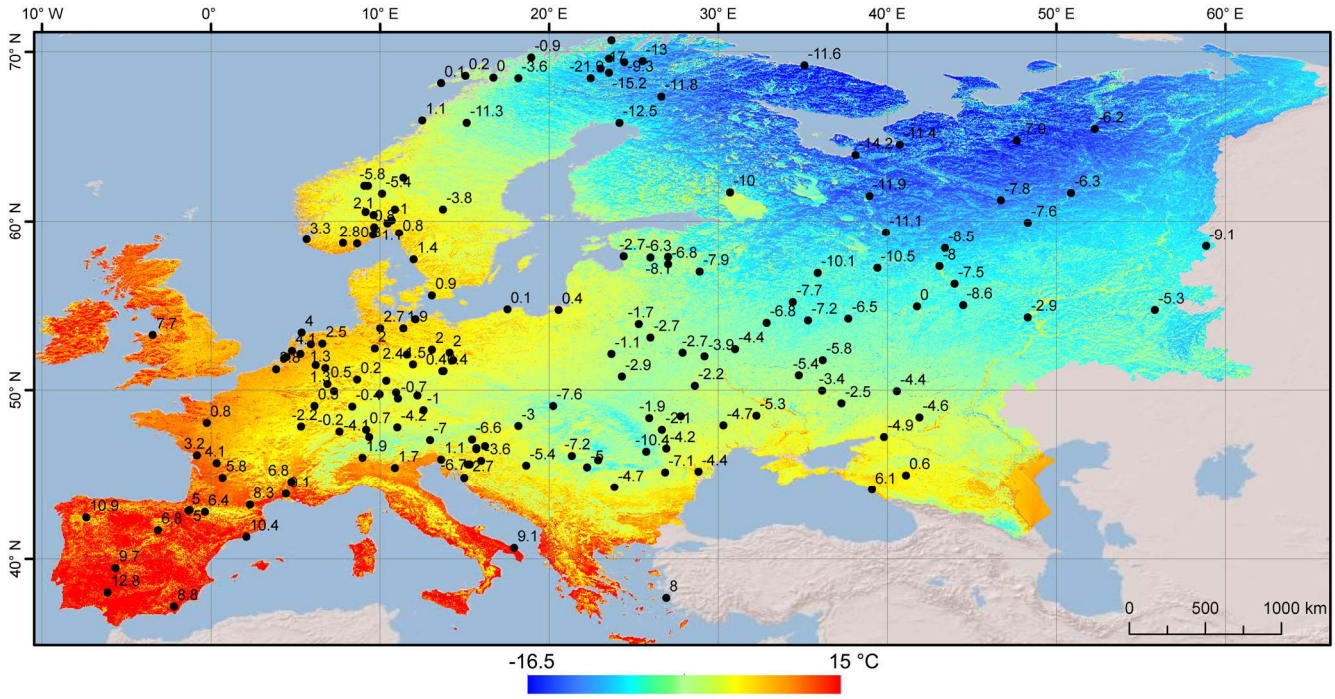


Fig. 2. Temperature prediction map using a complete set of predictors for 1.1.2011, with temperatures of validating weather stations (dots) given in °C

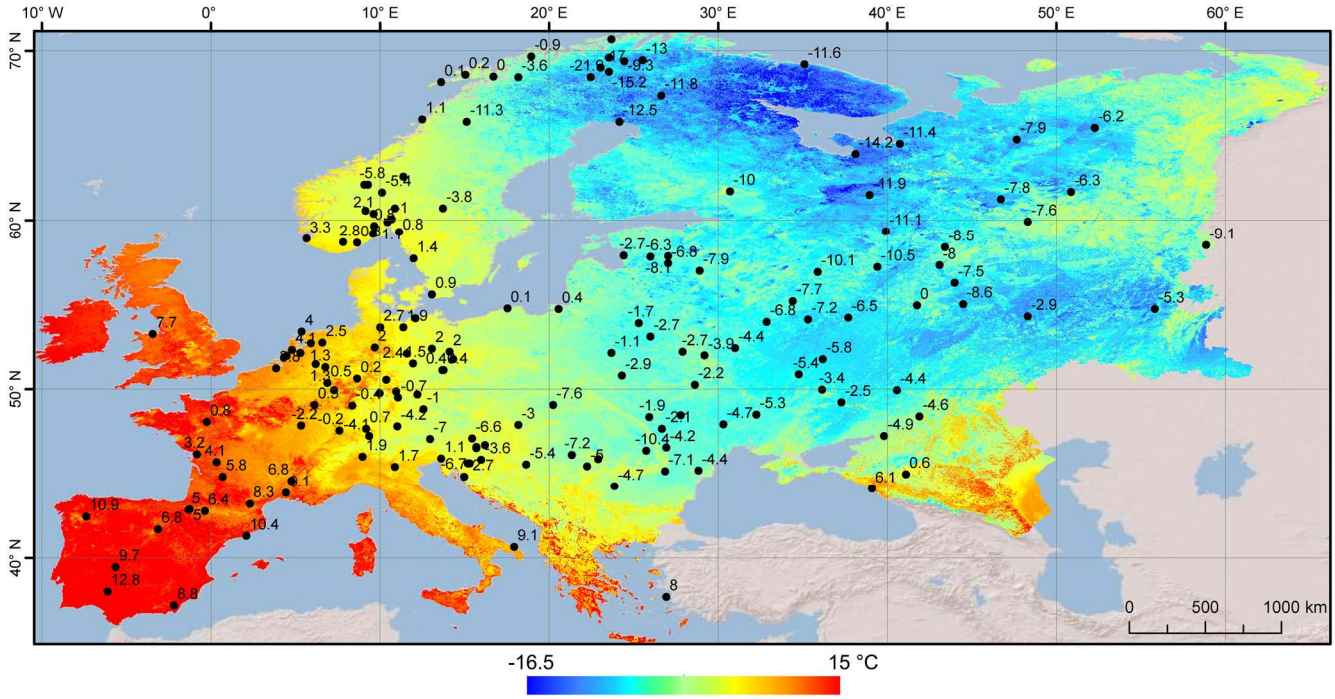


Fig. 3. Temperature prediction map using filtered set of predictors for 1.1.2011, with temperatures of validating weather stations (dots) given in °C

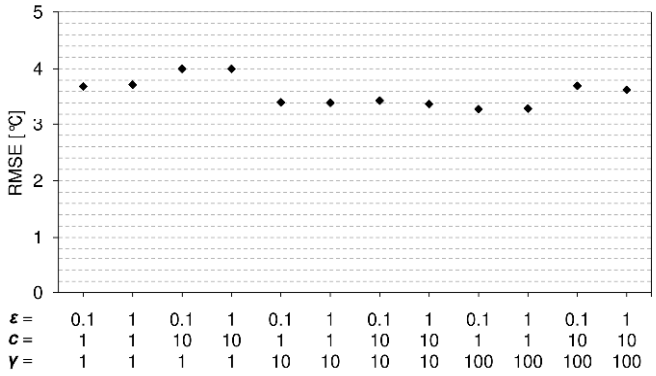


Fig. 4. Optimization of SVMr parameters for given ε, c, γ combinations

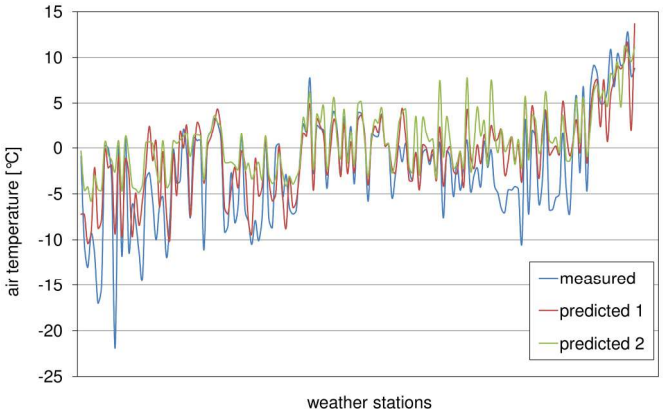


Fig. 5. Comparison of measured and predicted temperatures at 168 validation weather stations

IV. Results and Discussion

The SVMr experiment was designed in the following fashion. From 544 available stations (containing the measured temperatures and values of 13 predictors) 376 training instances were randomly sampled, which makes about one third of the total, which is usually recommendable proportion [2]. Remaining third with 168 stations was reserved for validation.

Initially it was necessary to optimize ε, c, γ parameters, which was performed by means of trial-and-error testing for selected combinations in 10-fold Cross-Validation protocol (Fig. 4). The selection has been narrowed on the basis of the best Root Mean Square Error result. The final parameter values ε=0.5, c=50, γ=2 were obtained after this fine tuning.

The first experiment included the original dataset, with all 13 predictors. The resulting model (Fig. 2) shows relatively good performance with RMSE of only ±2.096°C, which is concurrent to the other reported techniques, for instance geostatistical [5].

The second experiment involved the same procedure, but the number of parameters was reduced by implementing an attribute selection filter. Correlation Feature Selection (CFS) measure was used in combination with Best First search method. According to CFS the selected attributes are well correlated with dependent variable (temperature in this case) while they remain uncorrelated to other predictors. From original 13 predictors, the dataset was reduced to only 4 including: longitude, MODIS image, land cover = *Forested Area* and land cover = *Low Vegetation Area*. Even reduced to such extent, this dataset have resulted in fair model, with RMSE of ±2.553°C. This result (Fig. 3) questions whether the additional spatial data, such as DTM derivatives are necessary, and also emphasizes the importance of the MODIS image and land cover, which have the most apparent significance for predictions.

Comparison of obtained predictions from these two models and trends of their residuals (Fig. 5) suggest that the second model tends to slightly overestimate measured temperatures more than the first model. It is therefore arguable weather the attribute selection should be accepted in such cases.

It can be speculated that various meteorological effects are yet to be included to improve the model. It is usually some remote phenomena that can govern the temperature trends, as previously indicated. However, it seems rather difficult to quantify these effects and include them in the dataset. It is also quite challenging to make long-term predictions with such unsteady data.

V. Conclusions

In this paper SVMr algorithm was applied in prediction of the peak daily air temperature for a single day (Fig. 2,3). The main experiment was concentrated on implementing of various predictors, including some less common spatial data. Another aspect was regarding possibilities of reducing the number of predictors to obtain equally or even better performing model. It is confirmed that MODIS satellite image and specific land cover classes are most influential on the air temperature prediction. With relatively low RMSE values in both cases (±2–2.5°C), it could be inferred that the SVMr regression is an efficient tool for predicting such meteorological measurements for short-term periods. It provides concurrent results to more commonly used interpolation methods.

Research could be relatively easily extended by including predictions of other typical meteorological parameters in short-term domain, i.e. for a single day or for an average in a few days period. Further research can be directed towards mid-term temperature prediction by using appropriate temporal data, but also involving remote or local phenomena and synoptic situations, which is suspected to be of a greater challenge. The latter particularly regards complex meteorological phenomena with highly variable values in short time span.

The research also showed that precise meteorological maps can be obtained from publically available data (Remote Sensing and GIS datasets) and open source software.

Acknowledgments

This research was supported by the by the Ministry of science and technological development of Republic of Serbia (project TR36009).

REFERENCES

[1] M. Begert, E .Zenklusen, C. Häberli, C. Appenzeller and L. Klok, "An automated procedure to detect discontinuities; performance assessment and application to a large European climate data set", Meteorologische Zeitschrift, vol. 17/5, pp. 663-672, 2008.

[2] R.F. Chevalier, G. Hoogenboom, R.W. McClendon and J.A. Paz, "Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks", Neural Comput. & Applic., vol. 20, pp. 151-159, 2011.

[3] B.B. Ekici, "A least squares support vector machine model for prediction of the next day solar insolation for effective use of PV systems", Measurement, vol. 50, pp. 255-262, 2014.

[4] M. Kanevski, A. Pozdnoukhov and V. Timonin, Machine Learning for Spatial Environmental Data: Theory, Applications and Software. Lausanne, EPFL Press, 2009.

[5] M. Kilibarda, T. Hengl, G.B.M. Heuvelink, B. Gräler, E. Pebesma, M. Perčec Tadić and B. Bajat, "Spatio-temporal interpolation of daily temperatures for global land areas at 1 kmresolution", Journal of Geophysical Research: Atmospheres, vol. 119, pp. 2294- 2313, 2014.

[6] A. Koca, H.F. Oztop, Y. Varol and G.O. Koca, "Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey", Expert Systems with Applications, vol. 38,pp. 8756-8762, 2011.

[7] J. Kyselý and R. Huth "Changes in atmospheric circulation over Europe detected by objective and subjective methods", Theoretical and Applied Climatology, vol. 85, pp. 19-36, 2006.

[8] B. Navascués, J. Calvo, G. Morales, C. Santos, A. Callado, A. Cansado, J. Cuxart, M. Díez, P. del Río, P. Escribà, O. García-Colombo, J.A. García-Moya, C. Geijo, E. Gutiérrez, M. Hortal, I. Martínez, B. Orfila, J.A. Parodi, E. Rodríguez, J. Sánchez-Arriola, I. Santos-Atienza and J. Simarro, "Long-term verification of HIRLAM and ECMWF forecasts over Southern Europe History and perspectives of Numerical Weather Prediction at AEMET", Atmospheric Research, vol. 125-126, pp. 20-33, 2013.

[9] E.G. Ortiz-García, S. Salcedo-Sanz, C. Casanova-Mateo, A. Paniagua Tineo and J.A. Portilla-Figueras, "Accurate local very short-term temperature prediction based on synoptic situation Support Vector Regression banks", Atmospheric Research, vol. 107, pp. 1-8, 2012.

[10] E.G. Ortiz-García, S. Salcedo-Sanz and C. Casanova-Mateo, "Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data", Atmospheric Research, vol. 139, pp. 128-136, 2014.

[11] A. Paniagua-Tineo, S. Salcedo-Sanz, C. Casanova-Mateo, E.G. Ortiz-García, M.A. Cony and E. Hernández-Martín, "Prediction of daily maximum temperature using a support vector regression algorithm", Renewable Energy, vol. 36, pp. 3054-3060, 2011.

[12] Y. Radhika and M. Shashi, "Atmospheric Temperature Prediction using Support Vector Machines", International Journal of Computer Theory and Engineering, vol. 1/1, pp. 1793-8201, 2009.

Spatial Dependency Modeling of Daily Mean Temperature Time Series using Spatial R-vine Models

[extended abstract]

Tobias Michael Erhardt, Claudia Czado and Ulf Schepsmeier
Zentrum Mathematik
Technische Universität München
München, Germany
tobias.erhardt@tum.de

***Abstract*—Classical geostatistical approaches for the modeling of spatial dependencies assume a Gaussian dependency structure. This assumption may simplify the modeling process, however it is not always met, when we face real world problems. We are going to introduce a new parametric R-vine copula based modeling approach, which is able to capture non-Gaussian spatial dependencies. R-vine copulas are a class of flexible multi-dimensional probability distributions, composed out of bivariate copulas. For each of these bivariate building blocks we can choose among a variety of different dependency structures (copula families), which are well understood and easy to compute. The proposed "spatial R-vine model" combines the flexibility of vine copulas with the geostatistical idea of modeling spatial dependencies by means of the distances between the variable locations. To illustrate the model development process we consider daily mean temperature time series taken from 54 monitoring stations across Germany.**

Keywords—daily mean temperature; spatial R-vine model; spatial statistics; vine copulas

I. INTRODUCTION

The following is a summary of yet unpublished work [1]. Precise model formulations and results can be found there.

The literature proposes different approaches for the modeling of spatial dependencies (see for instance [2], [3] and [4]). Most of these approaches assume and model Gaussian dependencies. However, multivariate Gaussian distributions are not suitable to model arbitrary data, since they are symmetric and unable to model extreme dependency. Thus, we apply so-called vine copula models (see [5], [6], [7], [8] and reference therein), designed to handle these shortcomings, i.e. to allow for flexible non-Gaussian dependency structures.

We developed a new spatial dependency model relying on a reparametrization of an R-vine copula model, exploiting the relation between the vine copula parameters and the available spatial information. Different model specifications based on distances and elevation differences can be envisioned. Maximum likelihood estimation and a model based prediction method are available. Our approach is developed, based on daily mean temperature time series originating from 54 monitoring stations across Germany, collected over the period 01/01/2010-12/31/2012 by the German Meteorological Service (Deutscher Wetterdienst).

Reference [9] is a further spatio-temporal modeling approach using vine copulas. It combines copulas and geostatistical methods for the modeling of spatial dependencies in the first vine copula trees. On the contrary, our R-vine based approach models the spatial dependencies in all vine copula trees and allows including spatial covariates other than distance.

II. DEPENDENCY MODELING USING VINE COPULAS

Copulas in general are d-dimensional distribution functions living on the unit hypercube $[0,1]^d$. The respective marginal distributions are uniform distributions on [0,1]. A copula C is a tie between a multivariate distribution function F and its marginals $(F_1,\ldots,F_d)$ that captures all dependency information (see [10]). It holds $F(\mathbf{y})=C(F_1(y^1),\ldots,F_d(y^d))$, where $\mathbf{y}=(y^1,\ldots,y^d)'$ is the realization of a random vector $\mathbf{Y}$.

Regular vine (R-vine) copulas are multivariate copulas composed out of bivariate copulas only, which are well understood and easy to compute (see [11], [12] and [13]). The needed bivariate building blocks generated by conditioning, are identified through a set of nested trees (in a graph theoretical sense), which is called R-vine tree structure. Here the bivariate copulas are also called pair-copulas. They can be selected among the many different, parametric, bivariate copula types, which feature different kinds of dependency structures.

In the case of the data set addressed above (d=54) we are interested in the 54-dimensional copula representing the spatial dependencies between the temperatures of the 54 monitoring stations. Marginal distributions corresponding to all 54 stations, which need to capture seasonality effects and temporal dependencies of the time series, are modeled. For further investigations, an ordinary R-vine is fitted to the data using the Dißmann algorithm [13]. This includes the sequential selection

of an R-vine tree structure and adequate, parametric copula families for the bivariate building blocks.

III. SPATIAL R-VINE MODEL

For high dimensions, ordinary R-vine copula models become computationally impracticable, because the amount of parameters increases quadratic. Our new approach, the spatial R-vine copula model (SV) allows to reduce the needed number of parameters significantly, by exploiting the available spatial information.

For the R-vine copula, let us consider only parametric pair-copulas with at most two parameters. For the first parameters of these bivariate copulas, there are copula-family-dependent functional relationships to the well-known rank correlation measure Kendall's tau, which measures the association of two variables. Since we expect comparatively strong spatial association between nearby monitoring stations, we investigated a possible relationship between the Kendall's tau estimates occurring in the R-vine copula fitted to the temperature data and respective station distances and elevation differences. We detected a strong linear relation of the Fisher-z transformed Kendall's tau estimates on the logarithmized station distances and some linear relation on the elevation differences. The only two-parametric copula family we considered was the bivariate Student-t copula, which was chosen for a large share of the modeled pairs of the R-vine copula fitted to the data. In this case, the second copula parameter corresponds to the degrees of freedom of the Student-t copula. We discovered a quadratic trend of the logarithmized estimated degrees of freedom with respect to the R-vine tree number.

We use our previous findings to parametrize all pair-copulas of each R-vine copula tree jointly. In the case of the data set under consideration (d=54), the first R-vine tree models bivariate dependencies between 53 station pairs. This means that the original R-vine requires 53 first copula parameters for this tree. Exploiting for instance the above-detected relationship of these parameters on the station distance, all of these parameters can be replaced by a copula-family dependent transformation of a linear term involving an intercept parameter and a further parameter for the logarithmized station distance. Thus, 53 parameters could be replaced by only two parameters, by using the dependency information captured by the station distances.

The parameters in the higher R-vine trees are replaced in a similar fashion. Due to the nested structure of the R-vine trees, the pairs modeled in higher trees depend on the structure of the lower trees. Thus, one could additionally consider covariates (e.g. distances, elevation differences) for the regression formula of the copula parameters, which are indicated by the conditioning on the previous R-vine trees.

In contrast, the second copula parameters are modeled jointly for all R-vine trees. They are replaced by the exponential of a quadratic polynomial of the tree number, i.e. all occurring second copula parameters are replaced by only three parameters.

Since the model is specified through nesting, error propagation is minimized when strongest dependencies are modeled first (see the Dißmann algorithm [13]). For that reason it is common in practice to consider so called truncated R-vines. This means that after a certain R-vine tree, all pair-copulas of the higher trees are considered to be independence copulas, which reflect conditional independencies and do not involve any parameter. This allows to reduce the needed number of parameters further, but also leads to a slight, maybe negligible decrease in model accuracy compared to the full R-vine copula.

To summarize, a spatial R-vine copula is a truncated R-vine copula where the parameters are regressed on available spatial covariates capturing spatial dependency information.

In order to estimate the parameters of a spatial R-vine copula model we apply maximum-likelihood estimation. The likelihood that has to be maximized is a product of the parametric, family-dependent copula densities occurring in the truncated R-vine. The original pair-copula parameters are determined through common regression parameters corresponding to spatial distances and elevation differences, as described above.

For the purpose of prediction from a spatial R-vine model at an unobserved location, the modeled spatial R-vine has to be extended by one dimension. This is achieved by adding a further variable as a leaf to the first R-vine tree and adjusting the higher order trees accordingly. The new tree edges are selected such that the respective Kendall's tau estimates, estimated based on the tree-dependent model specification chosen above, are maximized. For the new pair copulas, families are selected based on the number of occurrences of each family in the R-vine copula. The parameters are calculated using the above model specification and based on the spatial information indicated by the (relative) location from which to predict. This results in a (d+1)-dimensional copula distribution which allows prediction of the variable corresponding to the unobserved location, given the variables composing the original data set. Hence, the prediction uses the available information for all locations of the data set to predict at a new location.

IV. CONCLUSIONS

Investigations of an R-vine copula fitted to temperature time series observed at 54 locations across Germany led to a new spatial dependency model, the spatial R-vine model. An extensive analysis of the relationship between the Kendall's tau estimates occurring in the R-vine copula and the respective distances and elevation differences propose tree-wise model specifications for the first pair-copula parameters. This allows to reduce the necessary number of parameters immensely. It shows that the station distances are able to explain the spatial dependencies to a large extent. The described model specification for the second pair-copula parameters reduces the number of parameters further. Utilization of different, non-Gaussian copulas as bivariate building blocks of an R-vine copula distinguishes our spatial R-vine model from classical Gaussian approaches to model spatial dependencies.

For the investigated temperature data set, our modeling approach led to a distinct reduction in the number of parameters. In the original (truncated) R-vine copula 733 parameters were needed. This number was reduced to 41 for the selected spatial R-vine model, which is also mirrored in the computation time of the parameter estimation for both models. The joint maximum likelihood estimation procedure took about 3.7 days for the R-vine copula. In comparison, 18 hours were needed in the case of the spatial R-vine model.

In [1] we compare our spatial R-vine model to a classical Gaussian approach of spatial dependency modeling. Prediction results are compared and evaluated based on a validation data set. For the purpose of comparison continuous ranked probability scores (see [14]) were calculated. These scores yield reasonable accuracy of our predictions. Examination of the scores over time show a time variation of the relative prediction performance of both models. Overall consideration of the scores and a comparison in terms of AIC and BIC yield preference of our spatial R-vine model.

Our modeling approach can also be applied to other types of meteorological data like precipitation, wind speed, cloud cover or air pressure. Also an extension of our modeling approach by including further or different covariates is possible.

REFERENCES

[1] T. M. Erhardt, C. Czado, and U. Schepsmeier, "R-vine models for spatial time series with an application to daily mean temperature," unpublished, (preprint available at http://arxiv.org/abs/1403.3500).

[2] K. Stahl, R. Moore, J. Floyer, M. Asplin and I. McKendry, "Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density," Agric. For. Meteorol., vol. 139, pp. 224-236, October 2006.

[3] J. Šaltytė-Benth, F. E. Benth and P. Jalinskas, "A spatial-temporal model for temperature with seasonal variance," J. Appl. Stat., vol. 34, pp. 823-841, September 2007.

[4] X. Hu, I. Steinsland, D. Simpson, S. Martino and H. Rue, "Spatial modelling of temperature and humidity using systems of stochastic partial differential equations," unpublished, (preprint available at http://arxiv.org/abs/1307.1402).

[5] C. Czado, "Pair-copula constructions of multivariate copulas," in Copula Theory and Its Applications, Lecture Notes in Statistics, vol. 198, P. Jarowski, F. Durante, W. K. Härdle and T. Rychlik, Eds. Berlin: Springer, 2010, pp. 93-109.

[6] C. Czado, E. C. Brechmann and L. Gruber, "Selection of vine copulas," in Copulae in Mathematical and Quantitative Finance, Lecture Notes in Statistics, vol. 213, P. Jarowski, F. Durante and W. K. Härdle, Eds. Berlin: Springer, 2013, pp. 17-37.

[7] D. Kurowicka and R. Cooke, Uncertainty Analysis with High Dimensional Dependence Modelling, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd, 2006.

[8] D. Kurowicka and H. Joe, Dependence Modeling: Vine Copula Handbook, Singapore: World Scientific, 2011.

[9] B. Gräler and E. Pebesma, "Modelling dependence in space and time with vine copulas," Presented at: Geostats 2012, Oslo, Norway, 11-15 June 2012.

[10] A. Sklar, "Fonctions de répartition à n dimensions et leures marges," in Publications de l'Institut de Statistique de L'Université de Paris, vol. 8, pp. 229-231, Institut Henri Poincaré, 1959.

[11] K. Aas, C. Czado, A. Frigessi and H. Bakken, "Pair-copula constructions of multiple dependence," Insur. Math. Econ., vol. 44, pp. 182-198, 2009.

[12] E. C. Brechmann and U. Schepsmeier, "Modeling dependence with C- and D-vine copulas: The R package CDVine," J. Stat. Softw., vol. 52, pp. 1-27, January 2013.

[13] J. Dißmann, E. C. Brechmann, C. Czado and D. Kurowicka, "Selecting and estimating regular vine copulae and application to financial returns," Comput. Stat. Data. An., vol. 59, pp. 52-69, March 2013.

[14] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," J. Am. Stat. Assoc., vol. 102, pp. 359-378, March 2007.

The first author likes to thank the TUM Graduate School's Graduate Center International Graduate School of Science and Engineering (IGSSE) for support.

Performance of Neural Network for Estimating Rainfall over Mato Grosso State, Brazil

[full paper]

Nadja Gomes Machado[1,2], Thiago Meirelles Ventura[2], Victor Hugo de Morais Danelichen[2], Marcelo Sacardi Biudes[2]
[1]Laboratório de Biologia da Conservação, Instituto Federal de Mato Grosso, Cuiabá, Brazil. nadja.machado@blv.ifmt.edu.br
[2]Programa de Pós-graduação em Física Ambiental, Universidade Federal de Mato Grosso, Cuiabá, Brazil

Abstract — **Rainfall is the key element in regional water balance, and has direct influence over economic activity. There is increasing interest of climatic and meteorological data in large spatial and temporal scale. Likewise, computational methods and techniques for climatic and meteorological estimates in large areas with small dataset are growing. Thus, we evaluated neural network performance for rainfall estimates over Mato Grosso State located in the Brazilian Midwest region. The technique of neural network allows the algorithm identifies patterns in the data series, allowing estimates to make new datasets. In this preliminary study, a dataset obtained from 12 meteorological stations was used to train the neural network and then it was run to perform estimates, which allowed comparing with TRMM satellite estimates. We chose TRMM satellite estimates because it has estimated appropriately the annual accumulated rainfall in the Brazilian Midwest region. In general, there was an overestimation of total rainfall estimates by neural network of 21.9% in January and 26,219% in September for the year 2010. The higher overestimated rainfall values in January occurred in Pantanal (39.5%) and Amazon forest (25.4%) than in Cerrado (14%); while the higher overestimated values in September occurred in Cerrado (31,225%) and Amazon forest (25,645%) than in Pantanal (1,424%). The rainfall estimates by neural network had better performance in January (wet season) than in September (dry season) which means that neural network was weak to predict lack of rainfall probably due to use just latitude and longitude as auxiliary variables. The better performance of rainfall estimates by neural network was in the Brazilian Savanna in January than in Amazon forest and Pantanal. Bad estimates of rainfall using neural network in Mato Grosso state were due to (i) a short temporal dataset, (ii) few stations with poor spatial variability, (iii) few auxiliary variables to build neural network. The next step will be to analyze the rainfall and other climatic estimates for the whole year for several years on the Midwest region of Brazil by neural network including other auxiliary variables besides latitude and longitude or by other computational frameworks developed by DailyMeteo group.**

Keywords —spatio-temporal dynamics, satellite observations, artificial intelligence, precipitation, Cerrado.

I. INTRODUCTION

The rainfall is the most influential meteorological element [1] with a direct effect on the water balance and an indirect effect on relative humidity, air and soil temperature, which affect plant growth and human development [2]. The annually amount and distribution of rainfall determine the natural vegetation type and agricultural exploitation mode in a region [3]. This characterization allows better planning of agricultural practices, soil conservation structures (contour lines and terraces), constructions (channel drains and dams), and weather forecasts [4].

Economy based on agribusiness involving production chain of agricultural and cattle raising can be directly influenced by the excess or lack of rainfall causing partial and total losses such as in the economy of Mato Grosso state in Brazil [1]. These effects on agribusiness have caused an increasing demand about climatic and meteorological data in large spatial and temporal scales. Analyses of pluviometric regime demand long data series [5] what can be obtained by remote sensing [6] and computational techniques.

Remote sensing and computational techniques are advantageous because they allow the monitoring on a regional scale the energy partitioning, carbon and water cycle with low operating costs and greater data acquisition [7], [8], [9]. It has become powerful tools for obtaining information from natural resources management such as water, soil and vegetation [10]. Several methodologies have been proposed for rainfall estimation using satellite images such as TRMM (Tropical Rainfall Measuring Mission satellite) and neural network.

TRMM as a remote sensing technique with the specific purpose of measuring rainfall in the tropics [11] is been used to investigate pluviometric regime dynamics for many purposes such as: assessing the spatiotemporal dynamics of two sub-regions of the Pantanal [12], studying daily variability of rainfall in the Amazon basin [13], evaluate the vegetation response in northeastern Brazil [14], identifying warm season in urban regions of some cities in USA [15], and investigate flooding causes in a city in south Brazil [16]. Rainfall estimated by TRMM was validated over the Ethiopian highlands [17] and the Midwest region of Brazil [18].

A neural network model is a mathematical construct whose architecture is essentially analogous to the human brain, which is represented by the network topology and pattern of connections between the nodes, its method of determining the connection weights, and the activation functions that it employs [19], [20], [21].

Nevertheless, the interest in artificial neural networks (ANNs) is nowadays increasing because of their high potential for complex, non-linear and time-varying input–output mapping [19]. Recently, artificial neural networks have been applied in meteorological and agro ecological modeling and applications [22]. Kumar et al. [23] applied neural networks for estimation of daily evapotranspiration and compared the performance neural networks with Penman–Monteith method. Most of the applications reported in literature concern estimation, prediction and classification problems [20].

Therefore, computational techniques and remote sensing for climatic and meteorological estimates in large areas with small dataset are growing around the world. Thus, the objective of this paper was to evaluate neural network performance for rainfall estimates over Mato Grosso State, Brazil.

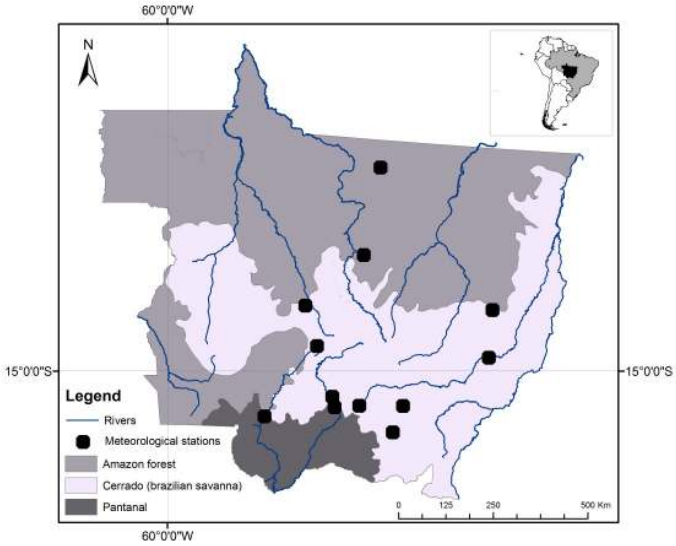


Fig. 1. *Location of Mato Grosso state, Brazil.*

II. MATERIAL AND METHODS

A. Study area

Mato Grosso is one of the Brazilian states, the third largest by area, located in the western part (latitude from 7° to 18° S and longitude from 50° to 62° W) of the country (Figure 1). A state with a flat landscape, alternating plateaus and plain areas, which presents three different ecosystems: Amazon Rainforest, Brazilian Savanna (Cerrado) and the Pantanal (wetland) [24]. The climate is classified as Aw, according to Köppen [25], with a dry season from May to September and a wet season from October to April [18], [26]. The annual temperature average ranges from 23°C to 26.8°C and the annual rainfall average ranges from 1,200 to 2,000 mm [25]. Mato Grosso contributes to form three basins: Paraguay (176,800.60 km^2); Amazon (592,382.54 km^2) and Tocantins (132,237.56 km^2) [27].

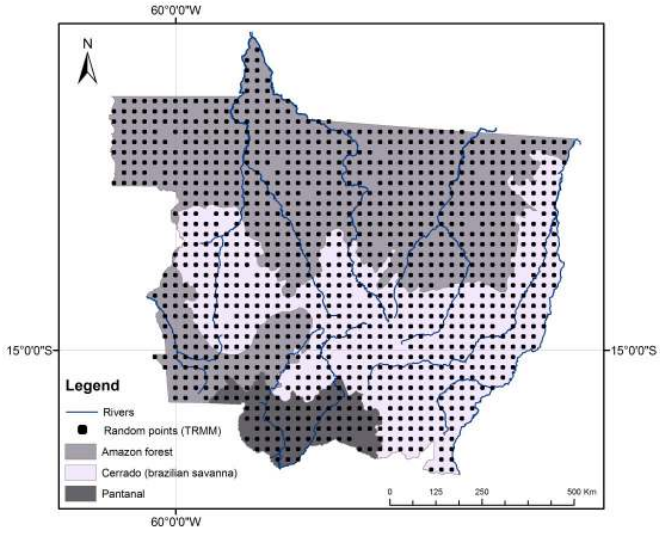


Fig. 2. *Location of random points (TRMM) in Mato Grosso state, Brazil.*

B. Meteorological data

Rainfall data were obtained from 12 meteorological stations (Figure 1 and Table 1) provided by the 'Instituto de Controle de Espaço Aéreo' (ICEA) of 'Comando da Força Aérea' available on the website [http://clima.icea.gov.br/clima/] and TRMM satellite provided by Distributed Active Archive System (DAAC) available on the website [http://disc2.nascom.nasa.gov/Giovanni/tovas/TRMM].There were 02 meteorological stations in Amazon forest, 08 in Cerrado and 02 in Pantanal. The pixel size of TRMM is 25 km². We used data from 3B43 V6 products. We chose TRMM satellite estimates because it has estimated appropriately the annual accumulated rainfall in the Brazilian Midwest region [18].

Table 1 – Characterization of meteorological stations in Mato Grosso state, Brazil. y = longitude. x = latitude. PN = Pantanal. CE = Cerrado. AF = Amazon forest.

Stations	**Biome**	**y**	**x**	**Altitude (m)**
Cáceres	PN	-57.68	-16.05	118
Canarana	CE	-52.27	-13.47	430
Cuiabá	CE	-56.1	-15.61	145
Diamantino	CE	-56.45	-14.4	286
Gleba Celeste	AF	-55.29	-12.28	415
Matupá	AF	-54.91	-10.25	285
Nova Xavantina	CE	-52.35	-14.7	316
Padre Ricardo Remetter	PN	-56.06	-15.78	140
Poxoréu	CE	-54.38	-15.83	450
Rondonópolis	CE	-54.56	-16.45	284
São José do Rio Claro	CE	-56.71	-13.43	350
São Vicente	CE	-55.41	-15.81	800

C. *Artificial neural network*

The dataset of 12 meteorological stations were used to train the neural network and then was run to perform estimates which allowed comparing with TRMM satellite estimates. The input dataset contained 360 ground measurements of daily accumulated rainfall for January (wet season) and 372 September (dry season) for year 2010. Rainfall estimates were performed by neural network as a function of latitude and longitude. We generated 1158 random points and extracted rainfall values from TRMM dataset to compare with the neural network estimates (Figure 2).

D. *Statistical analysis*

The evaluation of rainfall estimates from neural network data in relation to TRMM data was performed by these statistical indices: accuracy of Willmott index "d" (eq. 1), root mean square error "RMSE" (eq. 2), mean absolute error "MAE" (eq. 3), and Spearman's Rank correlation "r" (eq. 4).

The accuracy is related to the distance of the estimated values from those observed. Mathematically, this approximation is widely applied to the comparison between models [28]. Their values range from the value of 0, representing no agreement, to value of 1 representing perfect agreement.

$$d = 1 - \left[\sum (P_i - O_i)^2 / \sum (|P_i - O| + |O_i - O|)^2\right] \quad (1)$$

where Pi is the estimated value, O_i the value observed and O the average of observed values.

The RMSE indicates how the model fails to estimate the variability in the measurements around the mean and measures the change in the estimated values around the measured values [29]. The lowest threshold of RMSE is 0, which means there is complete adhesion between the TRMM estimates and measurements.

$$EQM = \sqrt{\frac{\sum (P_i - O_i)^2}{n}} \quad (2)$$

The MAE indicates the mean absolute distance (deviation) of values estimated from the values measured. The MAE and RMSE values should be close to zero [29].

$$EMA = \sum \frac{|P_i - O_i|}{n} \quad (3)$$

Spearman's Rank correlation coefficient is used to identify and test the strength of a relationship between two sets of data [30].

$$r = 1 - \frac{6 \sum d^2}{n\,(n^2 - 1)} \quad (4)$$

where di = xi – yi is the difference in the ranks given to the two variable values for each item of data.

III. RESULTS AND DISCUSSION

Rainfall ground measurements were higher in January than in September (Table 2), following the seasonal trends of the region [18]. There was a geographical pattern of rainfall from higher to lower values from North to South in January (Figure 3) in Mato Grosso state [18], [31]. However, neural network estimates captured a small amount of the rainfall pattern over Mato Grosso state (Figure 4).

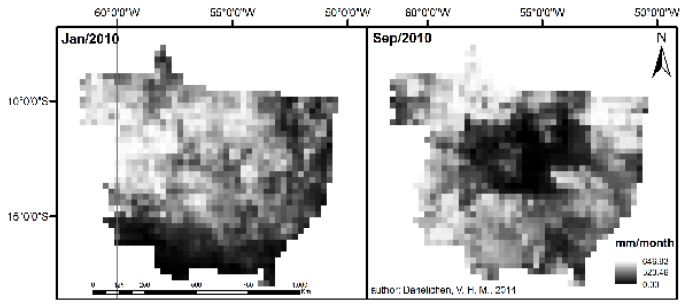


Fig. 3. *Rainfall map using TRMM estimates in Mato Grosso state, Brazil.*

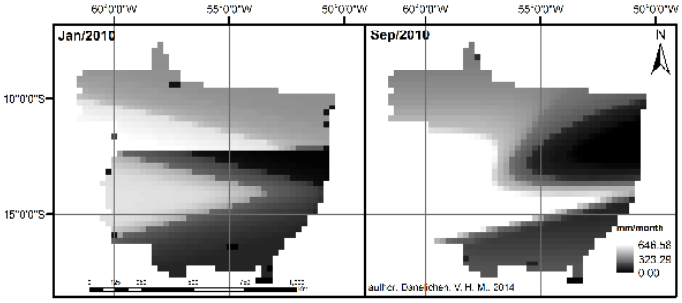


Fig. 4. *Rainfall map using neural network estimates in Mato Grosso state, Brazil.*

In general, there was an overestimation of total rainfall estimates by neural network of 21.9% in January and 26,219% in September for the year 2010 (Table 2). The higher overestimated rainfall values in January occurred in Pantanal (39.5%) and Amazon forest (25.4%) than in Cerrado (14%); while the higher overestimated values in September occurred in Cerrado (31,225%) and Amazon forest (25,645%) than in Pantanal (1,424%).

Table 2 – Spatial and temporal variability of rainfall estimates by meteorological stations, TRMM and neural network in Mato Grosso state, Brazil. MS = Meteorological Stations. NN = Neural Network. AF = Amazon forest. CE = Cerrado. PN = Pantanal.

	MS (mm)		**TRMM (mm)**		**NN (mm)**	
	Jan	**Sep**	**Jan**	**Sep**	**Jan**	**Sep**
Total	345.33	11.45	368.25	20.14	436.85	386.51
AF	376.15	36.35	398.70	24.92	482.28	401.64
CE	346.90	5.92	354.01	13.97	401.58	378.41
PN	308.25	8.65	214.23	24.92	286.97	401.64

The rainfall estimates by neural network had better performance in January (wet season) than in September (dry season) (Table 3). The neural network was weak to predict lack of rainfall probably due to use just latitude and longitude as auxiliary variables. The better performance of rainfall estimates by neural network was in the Brazilian Savanna in January than in Amazon forest and Pantanal.

Table 3 – Statistical performance of neural network for estimating rainfall in Mato Grosso state, Brazil. total = values including all biomes. AF = Amazon forest. CE = Cerrado. PN = Pantanal.

	RMSE	**d**	**MAE**	**r**
Jan_total	124.47	0.67	103.70	0.55
Set_total	404.68	0.06	369.55	0.10
Jan_AF	141.19	0.47	120.97	0.20
Set_AF	418.22	0.08	379.58	0.16
Jan_CE	102.65	0.74	85.16	0.68
Set_CE	401.63	0.03	368.60	-0.12
Jan_PN	97.03	0.32	75.92	-0.16
Set_PN	418.22	0.08	379.58	0.16

Neural network has obtained good estimates of meteorological and climatological variables in several studies. Hijmans et al. [32] obtained good estimates of monthly total precipitation, monthly mean, minimum and maximum temperature using latitude, longitude, and elevation as auxiliary variables in neural networks. Dibike and Coulibaly [19] also obtained good estimates using temporal neural network as method for downscaling both daily precipitation as well as daily maximum and minimum temperature series. Antonic et al. [33] obtained good results for modelling seven climatic variables using neural network with elevation, latitude, longitude, month and time series of respective climatic variable observed at two weather stations as auxiliary variables.

Despite the great popularity of the neural network models in many fields, Hsieh and Tang [34] showed three obstacles to adapting the neural network method to meteorology and oceanography, especially in large-scale with low-frequency studies due to (a) nonlinear instability with short data records and (b) large spatial data fields.

IV. CONSIDERATION

In general, we obtained bad rainfall estimates using neural network in Mato Grosso state due to (i) a short temporal dataset, (ii) few stations with poor spatial variability, and (iii) few auxiliary variables to build neural network.

The next step will be to analyze the rainfall and other climatic estimates for the whole year for several years on the Midwest region of Brazil by neural network including other auxiliary variables besides latitude and longitude or by other computational frameworks developed by DailyMeteo group.

ACKNOWLEDGMENT

N.G.M. acknowledges a grant from CAPES (9768/13-0). T.M.V. acknowledges a grant from CAPES (14277/13-1). V.H.M.D. acknowledges a grant from CAPES. M.S.B. acknowledges a grant from CAPES (9750/13-4).

REFERENCES

[1] R. Dallacort, J.A. Martins, H.M. Inoue, P.S.L. Freitas, and A.J. Coletti. "Distribuição das chuvas no município de Tangará da Serra, médio norte do Estado de Mato Grosso, Brasil". Acta Scientiarum Agronomy, v.33, n.2, pp.193-200, 2011.

[2] J.W.M.C. Santos. "Ritmo Climático e Sustentabilidade sócio-ambiental da agricultura comercial da soja no Sudeste de Mato Grosso". Revista do Departamento de Geografia, v.1, pp.1-20, 2005.

[3] G.A. Buriol, V. Estefanel, A.C. Chagas, and T.D. Eberhard. "Clima e vegetação natural do Estado do Rio Grande do Sul segundo o diagrama climático de Walter e Lieth". Ciência Florestal, v.17, n.2, 91-100, 2007.

[4] M.G.P. Bazzano, F.L.F. Eltz, and E.A. Cassol. "Erosividade, coeficiente de chuva, padrões e período de retorno das chuvas de Quaraí, RS". Revista Brasileira de Ciência do Solo, v.31, n.5, pp.1205-1217, 2007.

[5] F.F.N. Marcuzzo, N.L. Oliveira, F.R.P. Filho, and T.G. Faria. "Chuvas na região Centro-Oeste e no estado do Tocantins: Análise histórica e Tendência futura". Boletim Geográfico, v.30, n.1, pp.19-30, 2012.

[6] R.N. Nóbrega. Modelagem de Impactos do Desmatamento nos Recursos Hídricos da Bacia do Rio Jamari (Ro) utilizando dados de Superfície e do TRMM. Campina Grande, PB: UFCG. 2008. 238p. Tese de Doutorado. Universidade Federal de Campina Grande. 2008.

[7] D. Courault, B. Sguin, and A. Olioso. "Review on estimation of evapotranspiration from remote sensing data: from empirical to numerical modeling approaches". Irrigation and Drainage System, v.19, pp.223-249, 2005.

[8] R. Allen, A. Irmak, R. Trezza, J.M.H. Hendrickx, W. Bastiaanssen, and J. Kjaersgaard. "Satellite-based ET estimation in agriculture using SEBAL and METRIC". Hydrological Processes, v.25, pp.4011-4027, 2011.

[9] Q. Mu, M. Zhao, and S.W. Running. "Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sensing of Environment". Numerical Terradynamic Simulation Group, Department of Ecosystem and Conservation Sciences, The University of Montana, Missoula, MT 59812, USA, 2011.

[10] C.C. Braga, F.R. Soares, F.R.C. Dantas, and L.F.P. Barbieri. "Determinação do albedo e índice de área foliar usando o sensor TM / LANDSAT 5". Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 935-942.

[11] B. Collischonn, D. Allasia, W. Collischonn, and C.E.M. Tucci. "Desempenho do satélite TRMM na estimativa de precipitação sobre a bacia do Paraguai superior". Revista Brasileira de Cartografia, v.59, n.01, pp. 93-99, 2007.

[12] M. Adami, R.M. Freitas, C.R. Padovani, Y.E. Shimabukuro, and A.M. Moreira. "Estudo da dinâmica espaço-temporal do bioma Pantanal por meio de imagens MODIS". Pesquisa Agropecuária Brasileira, v.43, n.10, pp.1371-1378, 2008.

[13] C.M.S. Silva, S.R. Freitas, and R. Gielow. "Ciclo diário da precipitação estimada através de um radar banda S e pelo algoritmo 3B42_V6 do projeto TRMM durante a estação chuvosa de 1999 no sudoeste da Amazônia". Revista Brasileira de Meteorologia, v.26, n.1, pp.95-108, 2011.

[14] E. Arai, M. Adami, R.M. Freitas, Y.E. Shimabukuro, V.B. Rao, and M.A. Moreira. "Análise de Séries Temporais MODIS e TRMM nas áreas de caatinga, cerrado e floresta". Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 5081-5088.

[15] J.M. Shepherd, H. Pierce, and A.J. Negri. "Rainfall Modification by Major Urban Areas: Observations from Spaceborne Rain Radar on the TRMM Satellite". American Meteorological Society. Journal of Applied Meteorology, v.41, pp.689-701, 2002.

[16] E. Collischon. "Climatologia e Gestão do espaço urbano o caso de uma cidade pequena". Mercator, v.9, n.1, pp.53-70, 2010.

[17] T. Dinku, P. Ceccato, E.K. Grover-Kopec, M. Lemma, S.J. Connor, and C.F. Ropelewski. "Validation of satellite rainfall products over East Africa's complex topography". International Journal of Remote Sensing, v.28, n.7, pp.1503-1526, 2007.

[18] V.H.M. Danelichen, N.G. Machado, M.C. Souza, M.S. Biudes. "TRMM Satellite Performance in Estimating Rainfall over the Midwest Region of Brazil". Revista Brasileira de Climatologia, v.12, pp.22-31, 2013.

[19] Y.B. Dibike, and P. Coulibaly. "Temporal neural networks for downscaling climate variability and extremes". Neural Networks, v.19, pp.135–144, 2006.

[20] P.S. Shirgure. "Evaporation modeling with artificial neural network: A review". Scientific Journal of Review, v.2, n.2, pp.73-84, 2013.

[21] S. Haykin. Neural Networks: A Comprehensive Foundation. 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.

[22] G. Hoogenboom. "Contribution of agrometeorology to the simulation of crop production and its applications". Agricultural and Forest Meteorology, v.103, n.1-2, pp. 137-157, 2000.

[23] M. Kumar, N.S. Raghuwanshi, R. Singh, W.W. Wallender, and W.O. Pruitt. "Estimating Evapotranspiration using Artificial Neural Network". Journal of Irrigation and Drainage Engineering, v.28, pp.224-233, 2002.

[24] C.P. Costa, C. Nunes da Cunha, and S.C. Costa. "Caracterização da flora e estrutura do estrato arbustivo-arbóreo de um cerrado no Pantanal de Poconé, MT". Biota Neotrópica, v.10, n.3, pp. 61-73, 2010.

[25] A.P. Souza, L.L. Mota, T. Zamadei, C.C. Martim, F.T. Almeida, and J. Paulino. "Classificação climática e balanço hídrico climatológico no estado de Mato Grosso". Nativa, v.01, n.01, pp.34-43, 2013.

[26] N.G. Machado, M.S.Biudes, V.H.M. Danelichen, M.C. Souza, M.C.J.A. Nogueira, and J.S. Nogueira. "Climate characterization of Cuiabá, Mato Grosso state, Brazil" unpublished.

[27] Mato Grosso. Caracterização hidrográfica do Estado de Mato Grosso. Relatório Preliminar – Versão pra discussão interna. Cuiabá-MT: PRODEAGRO, 1995.

[28] C.J. Willmott, S.G. Ckleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'donnell, and C.M. Rowe. "Statistics for the evaluation and comparison of models". Journal of Geophysical Research v.90, n.C5, pp.8995-9005, 1985.

[29] C.J. Willmott, and K. Matssura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". Climate Research, v.30, pp.79-92, 2005.

[30] J.H.Zar. Biostatistical analysis, 5th ed., Upper Saddle River, NJ. Pearson, 2010.

[31] M.S. Biudes, N.G. Machado, P.H.Z. Arruda, G.A.R. Neves, F.A. Lobo, C.M.U. Neale, G.L. Vourlitis, and J.S. Nogueira, "Patterns of Energy Exchange for Brazilian Tropical Ecosystems in Mato Grosso, Brazil" unpublished.

[32] R.J. Hijmans, S.E. Cameron, J.L. Parra, P.G. Jones, A. Jarvis. "Very high resolution interpolated climate surfaces for global land areas". International Journal of Climatology, v. 25, pp. 1965–1978, 2005.

[33] O. Antonic, J. Krizan, A. Marki, and D. Bukovec. Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks. Ecological Modelling, v.138, pp.255–263, 2001.

[34] W.W. Hsieh, and B. Tang. "Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography". Bulletin of the American Meteorological Society, v.79, n. 9, pp. 1855-1870, 1998.

Assessing Uncertainties in Rainfall Maps from Cellular Communication Networks

[abstract]

Manuel F. Rios-Gaona, Aart Overeem, Remko Uijlenhoet
Hydrology and Quantitative Water Management group, department of Environmental Sciences
Wageningen University - WU
Wageningen, The Netherlands
manuel.riosgaona@wur.nl

Hidde Leijnse, Aart Overeem
Research and Development, Weather Service
Royal Netherlands Meteorological Institute - KNMI
De Bilt, The Netherlands

***Abstract*—Commercial cellular communication networks have been recently used in country-wide rainfall-map retrievals. Rainfall is the main source of attenuation in the electromagnetic signals that travel from one telephone tower to another, across the network. If the received power is measured at one end of a microwave link, the path-averaged rainfall intensity can be retrieved. The use of microwave link networks is a step further in the run for accurate rainfall estimates, given the large amount of information they can potentially collect.**

The aim of this work is to identify and quantify the sources of uncertainty present in rainfall maps retrieved from commercial microwave link networks. We used rainfall estimates, from microwave link data, to interpolate rainfall maps for the entire land surface of The Netherlands. These interpolated rainfall maps were compared to gauge-adjusted actual rainfall fields considered as ground-truth; thus, the uncertainties could be quantified. The Ordinary-Kriging (OK) was the methodology to interpolate the rainfall maps.

We based our uncertainty analysis in four features that mainly define the applicability of commercial microwave link networks in rainfall-map retrievals: the spatial density of the network; the time availability of the microwave link data (or attenuation measurements); the interpolation methodology; and the intrinsic aspects proper of the measurement process (for instance, wet antenna effect, sampling interval of the measurements, wet/dry period classification, drop size distribution, multi-path propagation). The results showed that it is more important to have continuous registries of microwave link data than a microwave link network with higher degree of spatial density.

Keywords—rainfall maps; commercial cellular networks; microwave links; uncertainty; ordinary kriging

Mapping Average Annual Precipitation in Serbia (1961–1990) by Using Machine Learning Techniques

[full paper]

Nenad Višnjevac, Miloš Kovačević, Branislav Bajat
Faculty of Civil Engineering
University of Belgrade
Belgrade, Serbia
nvisnjevac@grf.bg.ac.rs

***Abstract*— Precipitation data are measured at certain places, often quite distanced. Values between the places can be predicted using geostatistical tools like kriging, for which the interpolated values are modeled by a Gaussian process. In this paper we discuss an alternative method, i.e. using Machine learning techniques (Multilayer perceptron, linear regression and Support Vector Regression) to predict precipitation values. For the case study, average annual values of precipitation for a 30-year period in Serbia, three models were created based on attributes related to location data, spatial pattern, distance from nearby sea and data of nearest points (gauges). The best results were obtained by using Support Vector Regression and Linear Regression, indicating the linear nature of the problem. Finally, we have compared the output from kriging prediction and the output obtained by using Machine Learning techniques.**

Keywords—Precipitation, Machine Learning Techniques, Linear Regression, Support Vector Regression

I. INTRODUCTION

Precipitation data are important to many problems in hydrologic analysis and designs. For example the ability of obtaining high resolution estimates of flow accumulations or floods depends on accurate and high resolution precipitation data. The accurate estimation of the spatial distribution of precipitation data requires a very dense network of measured places (gauges), resulting in high installation and operational costs. Hence, it is necessary to estimate the precipitation at unrecorded locations from values at measured sites.

A number of methods have been proposed for the interpolation of precipitation data. The simplest approach consists of assigning to the unrecorded location the precipitation value of the closest measured site [1]. This method creates around each measured location a polygon of influence with the boundaries at a distance halfway between rain gauge pairs, i.e. Thiessen polygons. Another method which can be used estimates values as a weighted average of surrounding values, the weights being reciprocal to the square distances from the unsampled location [2]. Like the Thiessen polygon method, the inverse square distance (IDW) technique is simple method and it does not allow considering factors such as topography, which can improve the accuracy of estimation. Interpolators such as those above are commonly included in GIS packages and have been applied to the interpolation of precipitation data from point-based station records.

The second group of methods consists of algorithms that combine precipitation data with digital elevation model (DEM) like linear regression and geostatistical methods such as simple kriging with varying local mean (SKlm) and kriging with external drift (KED) and collocated ordinary cokriging. All geostatistical algorithms provide better predictions than the methods that ignore the pattern of spatial dependency (Thiessen polygons and IDW) [3]. Bajat et al. [3] used regression kriging for mapping average annual precipitation in Serbia. Obtained results show that the prediction of average annual precipitation by regression kriging is a robust technique. Geostatistical tools are used not only for spatial prediction of climatologic variables but also for mapping detected trends in climatologic changes during certain time intervals [4] [5].

There is no doubt that geostatistical methods such as kriging have been useful for climate surface interpolation (including precipitation data) especially when elevation data is included in the interpolation [6]. However, lately some other prediction methods, like machine learning techniques, have been introduced in spatial prediction and modeling of climate variables [3] [7]. For example, Bryan and Adams [7] used resilient backpropagation artificial neural networks (ANN) to interpolate mean annual precipitation and surface temperature for China. Target variables were represented on the basis of nonlinear relationships of latitude, longitude, and terrain elevation. Also Hong [8] used machine learning techniques for rainfall forecasting.

The procedure of modeling average annual precipitation over a 30-year period by machine learning techniques is presented in this paper. For mapping average annual precipitation three machine learning techniques were used: Multilayer Perceptrons (MP) as an artificial neural network [11], Multiple Linear Regression (MLR) [9] [12] and Support Vector Regression (SVR) [10] [13]. Besides the usual data such as precipitation and location of rain gauge stations (northing, easting, and altitude), which are nowadays

customary deliverable, we calculated 19 more attributes describing each measured place. Finally, we compared the results of machine learning tools with the kriging result obtained by Bajat et al. [3].

The structure of the paper is as follows: "Materials and methods" section contains a brief description of theoretical foundations of machine learning techniques. In addition to this section, the description of used rain gauge meteorological data is given. "Results and comments" section provides details about explanatory data analyses and spatial data modeling and mapping by machine learning techniques. "Conclusions" section concludes the paper.

II. Materials And Methods

A. Machine learning techniques for precipitation prediction

The main objective of the research was to examine the usage of machine learning (ML) techniques to build a predictive model for average annual precipitation in Serbia during the period of 1961 – 1990. Using available measurements from a network of precipitation gauge stations, the model should be capable to predict the average annual precipitation in any desirable location. The problem is formulated as follows: given a location $\mathbf{x}=<x_1,x_2,\ldots x_n> \in R^n$ one tries to find a function f which maps a location into related precipitation: f: $\mathbf{x}\rightarrow p$, where $p\in R$. Each x_i represents a real valued attribute describing the associated location such the altitude or the distance from the nearby sea. ML techniques learn the mapping f by using a set of locations in which all attributes and precipitation are known in advance. This set is referred as a *training set* and each member of the

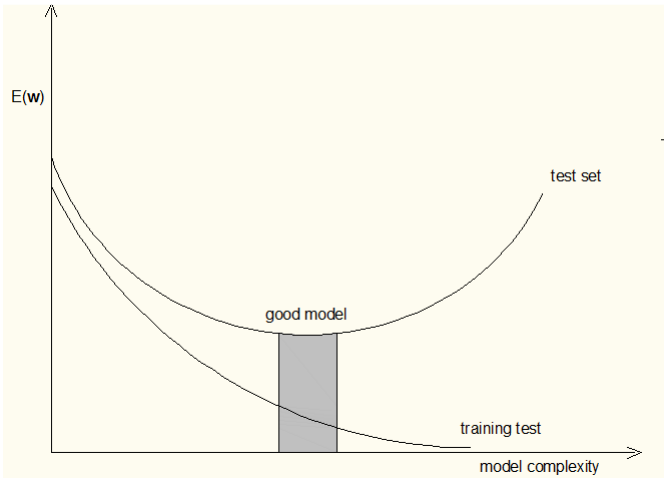


Fig. 1. Finding a good model

set is a pair $<\mathbf{x}_i, p_i>$, i=1, 2, … N, where N is the number of elements (gauge stations) in a training set. To train a predictive model f from the training set one uses techniques such as MP [11] or SVR [10][13].

ML techniques use different approaches to find the appropriate $f = f(\mathbf{x}, \mathbf{w})$ where $\mathbf{w} = <w_1, w_2, \ldots w_k> \in R^k$ represent model parameters. All techniques share similar ideas of finding the best model parameters $\mathbf{w}^*$ that decrease the values of error function $E(\mathbf{w}) = E(f(\mathbf{x}) - p_{true})$ on a training set while retaining the simplicity of the model. If one tries to minimize the error function on a training set this could lead to *overfitting* when a model perfectly predicts training values but fails to predict unknown values on a separate data (Fig. 1). In general, best models were built when one uses sufficient, quality data in which the examples (locations) are represented with carefully selected attributes that convey enough information for the underlying prediction task.

B. Study Area

The study area occupies the territory of Serbia located at the crossroads between Central and Southern Europe. It comprises around 18 % of Balkan Peninsula, covering an area of 88,361 km^2. Serbia is a country of diverse topography (Fig. 2). The northern part of the country is completely located

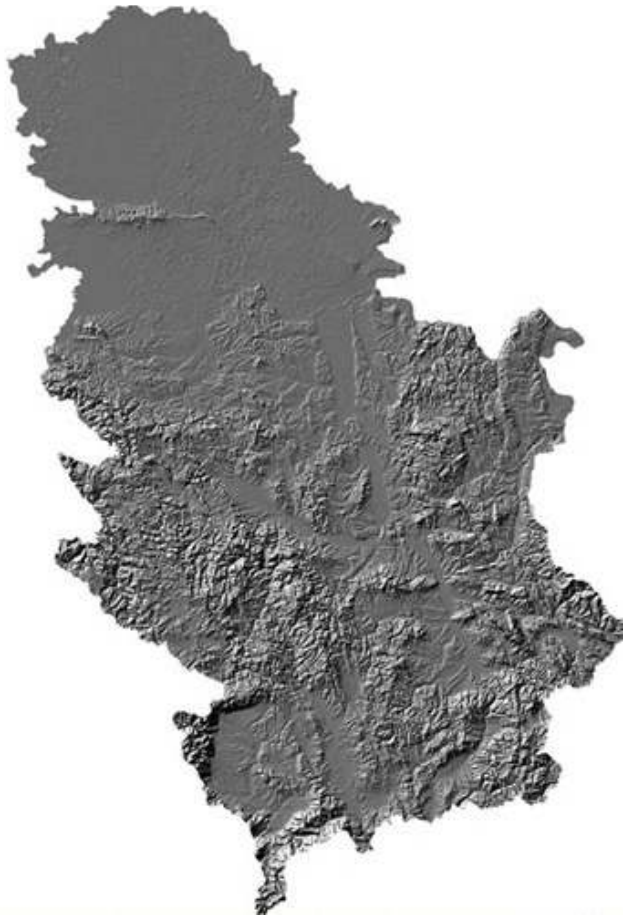


Fig. 2. DEM of Serbia

within the Pannonian Plain. Four mountain systems cross the country: the Dinaric Alps (in the western part of the country), the Carpathian Mountains (along the eastern part of Serbia), Balkan Mountain (eastern border of the country), and the Rhodope Mountains (on the southeast). The country varies in altitude from 29 m near the border between Serbia, Romania,

and Bulgaria to 2,656 m on Prokletije Mountain (on the south).

There are three main types of climate in Serbia: continental, moderate-continental, and modified Mediterranean climate. Geographic location and topography are key factors in formation of different types of climate. The north of the country is characterized by typical continental climate with air masses coming mainly from northwest Europe [14]. The south and southwest of the country is subjected to Mediterranean influences. The amount of precipitation increases with altitude from north to south. The northern parts receive less than 600 mm of annual precipitation; towards the south, it rises to 1,000 mm annually, while some mountains summits in the southwest receive over 1,000 mm per year. Most part of the country has the continental precipitation regime with maximum precipitation in June or May and minimum in February or October. Due to Mediterranean influence, southwestern Serbia has the Mediterranean precipitation regime, with maximum during winter months and minimum in August.

C. Precipitation Dataset

Two separate datasets were used in this study. The first set is related to rain gauge stations, their spatial coordinates (northing, easting, and altitudes) and associated average annual precipitation values. The second set represents publicly available digital elevation model used to improve the basic

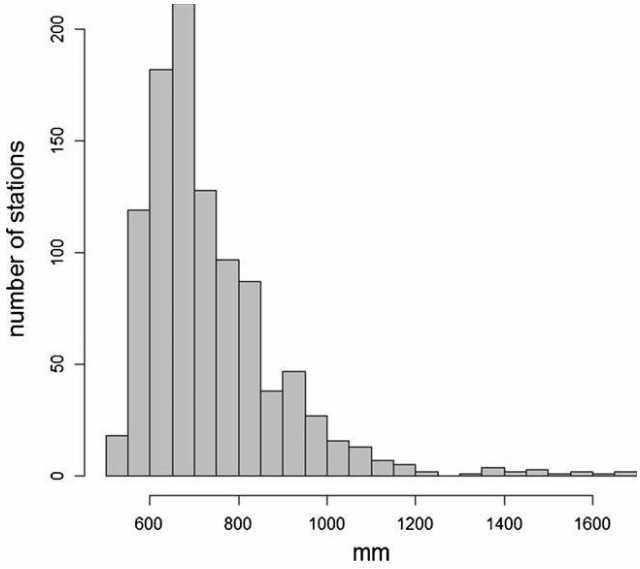


Fig. 3. The histogram of average annual precipitation in Serbia (1961 - 1990)

data representation of the first set.

The first dataset included mean annual precipitation data from 1,014 meteorological stations in Serbia provided by the Hydro-Meteorological Service of Serbia. The data come from 31 main stations, 99 climatological stations, and 885 precipitation stations. Since almost all weather stations in Serbia were used in this research, the precipitation data are very well spatially distributed. At altitudes 0–500m (62% of the whole area), there are 66.1% of all weather stations. From 500 to 1,000m (27% of the whole area), there are 23.4% of all stations. At altitudes higher than 1,000m (11.2% of the whole area), there are 10.9% of the stations. The data covers the period of 1961–1990 and are quality controlled in terms of correction of misprints and relocation of the stations. The histogram of average annual precipitation (Fig. 3) points to a skewed distribution with mean value of 739mm.

For the purpose of this study, we used DEM of 1km resolution produced from GDEM global model (Version 1) of Earth's surface, published in the year of 2009 (http://asterweb.jpl.nasa.gov/gdem-wist.asp).

D. Data representation

In this research each of 1014 locations is represented as a real vector $\mathbf{x}_i$ with 22 coordinates (attributes) given in TABLE I. and an associated precipitation p_i.

TABLE I. USED ATTRIBUTES

Attribute	*Description*	*Group*
x_1	Easting	1
x_2	Northing	
x_3	Altitude	
x_4	Distance to high mountains	2
x_5	Distance to medium-sized mountains	
x_6	Distance to low mountains	
x_7	Direction toward high mountains	
x_8	Direction toward medium-sized mountains	
x_9	Direction toward low mountains	
x_{10}	Distance from the Adriatic Sea	3
x_{11}, x_{12}, x_{13}	Distance to 1st, 2nd and 3rd Nearest Point	4
x_{14}, x_{15}, x_{16}	Direction toward 1st, 2nd and 3rd Nearest Point	
x_{17}, x_{18}, x_{19}	Altitude of 1st, 2nd and 3rd Nearest Point	
x_{20}, x_{21}, x_{22}	Precipitation at 1st, 2nd and 3rd Nearest Point	

First group of attributes represent spatial information of locations. Since combining precipitation data with digital elevation model (DEM) gives better prediction than using just precipitation data [3], altitude was added into this group.

Smith [15] gives a comprehensive review on the complex subject of orographic rain. On the windward side, forced lifting of approaching air masses causes the release of rainfall and an increase in precipitation with elevation [16]. Depending on the mountain size and the efficiency of the release processes, precipitation will decrease on the leeward side. Mountains may also facilitate the formation of convective rainfall [15]. Therefore, in this research we used six attributes which are based on a topography pattern, e.g.

mountains. Three types of mountains were extruded from DEM: high mountains – over 2000 m, *medium*-sized mountains – from 1000 m to 2000 m and low mountains – from 500 m to 1000 m. Based on these data, attributes in group 2 were calculated by using GIS tools.

According to water cycle, which describes the continuous movement of water, water moves from one aggregate state to another, such as from oceans and seas to the atmosphere, and then it falls to Earth's surface as precipitation. Therefore, we used Euclidian distance from the Adriatic Sea as an attribute (group 3).

Geostatistical methods such as regression kriging are based on spatial dependency. Within the kriging algorithm variogram function is used for describing the spatial dependence degree of a spatial random field or stochastic process. Bajat et al. [3] created variogram function for precipitation data in Serbia and it clearly shows that if a point pair is closer the dependence degree is higher. To include this behavior into our mapping model we calculated 12 more attributes based on precipitation and location data of three nearest points (group 4).

E. Evaluation of model performance

In order to determine which ML technique (MP, LR, and SVR) produces the best prediction model it is crucial to apply the correct evaluation protocol. For that purpose five train-test splits were randomly created using 1014 gauge stations. In each train-test split locations were distributed uniformly over the case study area with 811 locations in the train and 203 locations in the test part. After building a model on a train part its performance was evaluated on a test part using standard measures: correlation coefficient, absolute mean error and root mean squared error. Final evaluation was performed after averaging the results for each train-test split.

When building a model in each split one must find the appropriate parameters for each applied technique (i.e. C for SVR or number of neurons in a hidden layer for MP). The number of parameter combinations for each technique is infinite. Therefore, we selected few reasonable parameter combinations for each technique and found the best combination (model) using the 5-fold cross validation [9] on each training set.

III. RESULTS AND COMMENTS

After building models with optimal parameters and applying prediction at gauges in test sets the results are presented in TABLE II.

TABLE II. SUMMARY STATISTIC OF PREDICTION ERRORS

Machine Learning Technique	***Correlation***	***MAE***	***RMSE***
Multilayer Perceptron (MP)	0.87	59.76	85.33
Linear Regression (MLR)	0.89	49.04	74.39
Support Vector Reg. (SVR)	0.88	50.41	75.23

MLR and SVR produced similar results, while MP appeared to be the inferior model.

Fig. 4 shows residual values between input and modeled precipitation at gauge test stations. The spatial pattern of residuals indicates their random distribution. Compared to DEM, it is notable that residuals are highly dependent on

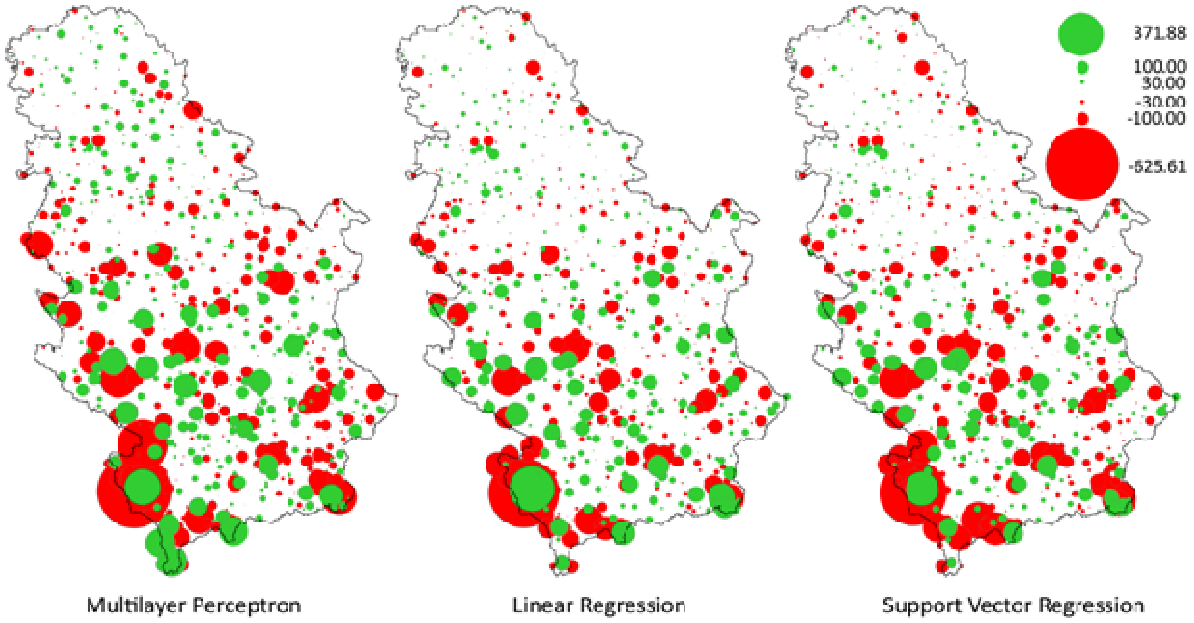


Fig. 4. Residuals on test gauges

altitude. Each of the three trained models causes higher

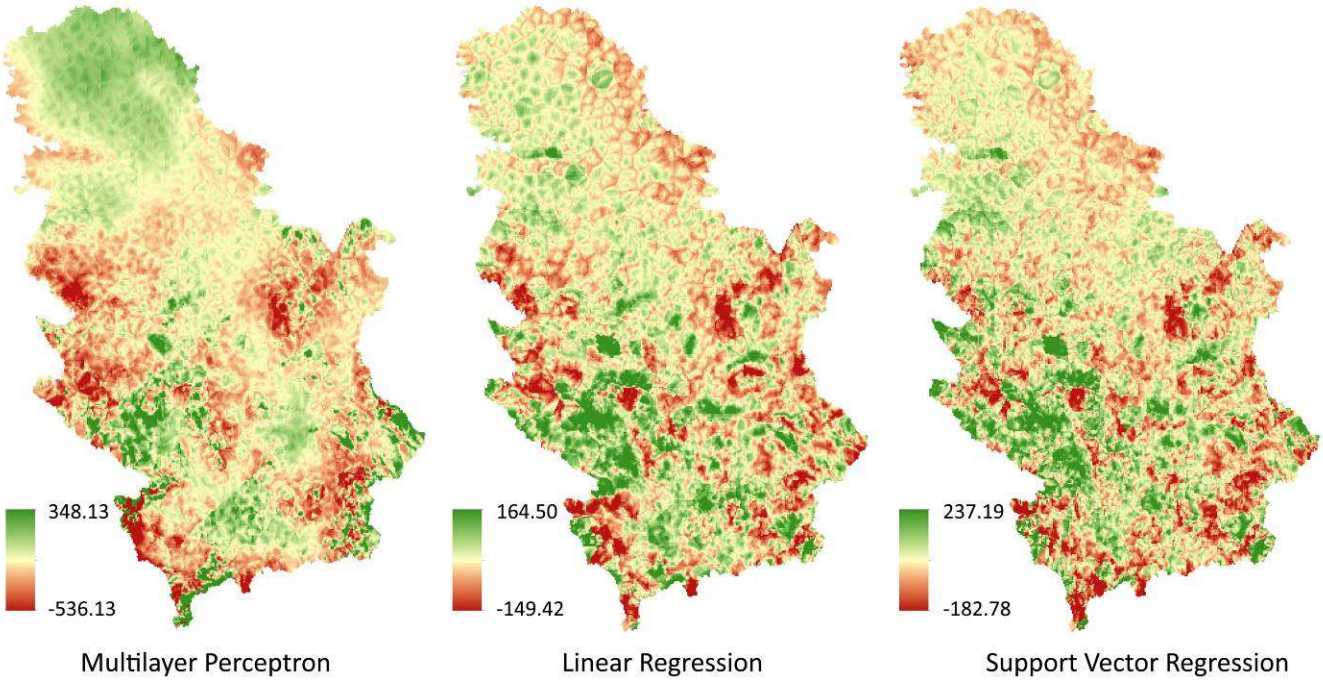


Fig. 5. Differences between trained models and regression kriging

negative residuals (underestimate precipitation value - red bubbles) than positive residuals (overestimate - green bubbles). As expected, the highest residual values are located in Prokletije mountain area closed to the Albanian border. Also, residuals are higher in south and west Serbia than in north part of the country. It is evident that all three models make similar errors in the same parts of the country.

Based on trained models three maps of the mean annual precipitation for Serbia were produced and compared to the reference map obtained by kriging regression (study by Bajat et al. [3]) which is assumed to be more accurate. The smallest difference, compared to the reference map, was produced by SVR (TABLE III.) while MLR produced slightly bigger difference. According to the mean value of differences and standard deviation MP had the biggest difference.

TABLE III. STASTISTICS OF DIFFERENCES TO KRIGING

Statistic measure	*Multilayer Perceptron*	Linear Regression	Support Vector Reg.
Abs Max	536.13	164.50	237.19
Mean	-29.05	6.67	-0.28
Standard dev.	43.97	28.37	26.31

Fig. 5 shows spatial distribution of differences between kriging and trained models. Both SVR and MLR models exhibited similar behavior in predicting precipitation while MP produced slightly different map of differences, especially in the northern part of the country. For all three models the biggest differences are found in the southern part of the country, while the northern part contains smaller and more uniform differences. In addition, all trained models formed rough surface of precipitation with sharp transition between

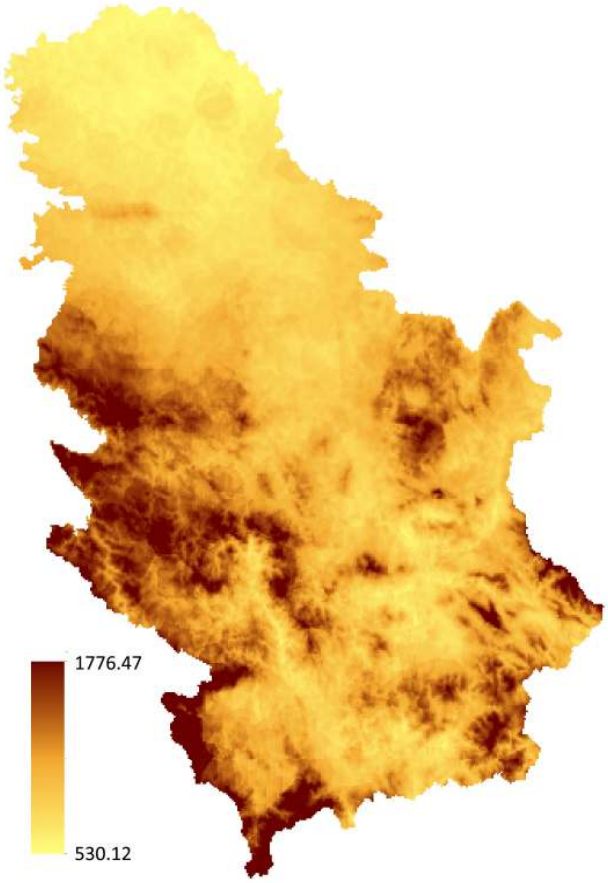


Fig. 6. Average annual precipitation map obtained by Support Vector

values of near sites, while kriging [3] created smooth surface, which is more natural and appropriate to the underlying problem.

Despite the similar performance on test gauges of SVR and MLR models (TABLE II.) SVR exhibited the smallest difference compared to kriging (TABLE III.). Therefore, Fig. 6 presents the map of mean annual precipitation for the period 1961–1990 obtained by SVR model.

The map shows variation in geographic distribution of precipitation. The wettest area is in the mountains at the southwest (Prokletije Mountain) and in the southern parts (Šar Mountain) of the country with average precipitation exceeding 1 300 mm. The driest part is the northern part of Serbia which extends into the Pannonian Plain with less than 600 mm a year. Eastern parts of the country are drier than the western parts. Relative uniformity of the precipitation in the northern and central parts of the country reflects less complex topography. The inconsistency of precipitation in the western, southern, and eastern part is a result of the influence of complex topography.

IV. CONCLUSIONS

The main objective of this work is to generate the map of average annual precipitation in Serbia for the period of 1961–1990 by using machine learning techniques. Based on location data, spatial pattern, distance from nearby sea and data of nearest gauges four groups of attributes were calculated.

Multiple Linear Regression and Support Vector Regression models provided similar results, both on test gauges and compared to kriging, while the results obtained by Multilayer Perceptron model were inferior. Our findings suggest the linear nature of the precipitation model in Serbia for the explained data representation.

The obtained results would be even better with incorporating more additional attributes related to precipitation (i.e. direction of the dominant wind). In addition, including additional gauges and topographic pattern from surrounding countries would probably improve the results and reduce the impact of "edge effects". The produced map confirmed the high influence of regional topography on average precipitation and described noticeable spatial patterns of precipitation values.

ACKNOWLEDGMENT

This study was supported by the Serbian Ministry of Education and Science, under grant No. III 47014.

REFERENCES

[1] P. Goovaerts, "Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall" Journal of Hydrology 228, pp. 113-129, 2000

[2] P.B. Bedient, W.C. Huber, "Hydrology and Floodplain Analysis. 2nd ed." Addison-Wesley, p. 25, 1992

[3] B. Bajat et al., "Mapping average annual precipitation in Serbia (1961–1990) by using regression kriging" Theoretical and Applied Climatology, vol. 112, pp. 1-12, 2012

[4] S. del Rio, L. Herrero, R. Fraile, A. Penas, "Spatial distribution of recent rainfall trends in Spain (1961–2006)" Int J Climatol 31, pp. 656-667, 2011

[5] F.J. Acero, M.C. Gallego, J.A. Garcia, "Multi-day rainfall trends over the Iberian Peninsula" Theoretical and Applied Climatology, vol. 108, pp. 411-423, 2012

[6] M.F. Hutchinson, "Interpolation of Rainfall Data with Thin Plate Smoothing Splines - Part II: Analysis of Topographic Dependence" Journal of Geographic Information and Decision Analysis, vol. 2, pp. 152 -167, 1998

[7] B.A. Bryan, J.M. Adams, "Three-dimensional neurointerpolation of annual mean precipitation and temperature surfaces for China" Geographical Analysis, vol. 34, pp. 93 - 111, April 2002

[8] W. Hong, "Rainfall forecasting by technological machine learning models" Applied Mathematics and Computation, vol. 200, pp. 41 - 57, June 2002

[9] I.H. Witten, E. Frank, M.A. Hall, "Data Mining - Practical Machine Learning Tools and Techniques" Elsevier Inc., 2011

[10] T.M. Mitchell, "Machine Learning" McGraw-Hill Science/Engineering/Math, 1997

[11] B. Yegnanarayana, "Artificial Neural Networks" PHI Learning Pvt. Ltd., 2009

[12] A.E. Hoerl, R.W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems" Technometrics, vol. 12, pp. 69 - 82, 1970

[13] A.J. Smola, B. Scholkopf, "A tutorial on support vector regression" Statistics and Computing, vol. 14, pp. 199 - 222, 2004

[14] M. Unkašević, Đ. Radinović, "Statistical analysis of daily maximum and monthly precipitation at Belgrade" Theoretical and Applied Climatology, vol. 66, pp. 241-249, 2000

[15] RB. Smith, "The influence of mountains on the atmosphere" Advances in Geophysics, vol. 21, pp. 87-229, 1979

[16] B. Johansson, "The influence of wind and topography on precipitation distribution in Sweden: statistical analysis and modeling" International Journal of Climatology, vol. 23, pp. 1523-1535, 2003

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

551.5(082)(0.034.2)

DAILYMETEO.ORG/2014 Conference (2014 ; Belgrade)
Proceedings of DailyMeteo.org/2014 Conference, Belgrade, Serbia 26-27 June 2014. [Elektronski izvor] : abstracts, extended abstracts and full papers / [edited by Branislav Bajat, Milan Kilibarda]. - Belgrade : Faculty of Civil Engineering, 2014 (Beograd : Dosije studio). - 1 elektronski optički disk (CD-ROM) ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovne strane dokumenta. - Tiraž 50. - Bibliografija uz pojedine radove.

ISBN 978-86-7518-169-9

a) Климатологија - Зборници
COBISS.SR-ID 208092940