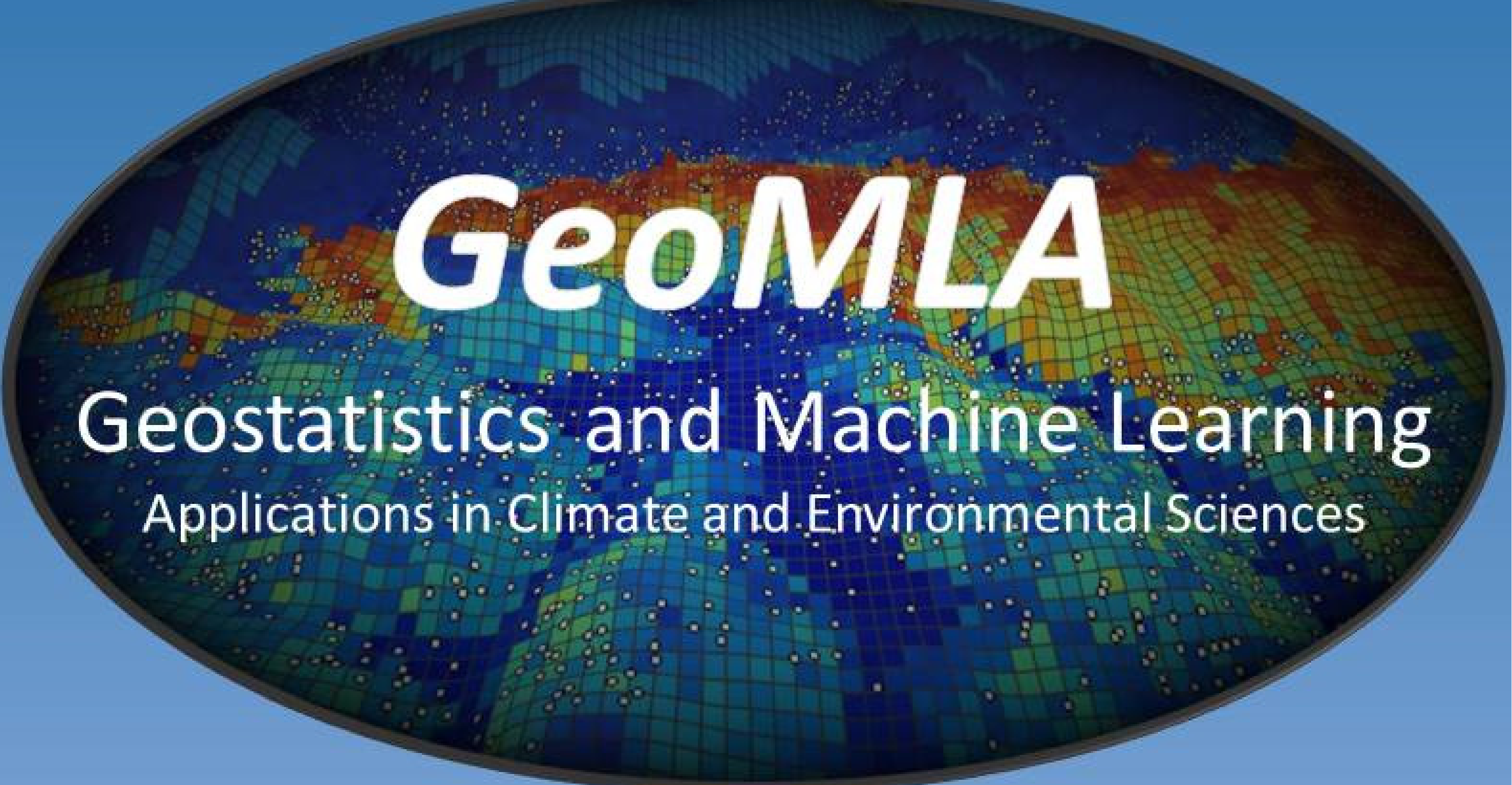
osgl.grf.bg.ac.rs

geomla.org

GeoMLA 2016

Proceedings of



GeoMLA

Geostatistics and Machine Learning

Applications in Climate and Environmental Sciences

Belgrade, Serbia 21-24 June 2016

DHMZ

REPUBLIC OF SLOVENIA
MINISTRY OF THE ENVIRONMENT AND SPATIAL PLANNING
SLOVENIAN ENVIRONMENT AGENCY

Univerzitet u Beogradu
Geografski fakultet

GeoMLA 2016
Abstracts, extended abstracts and full papers
of the GeoMLA 2016 Conference
Belgrade, Serbia, 23-24 June 2016

Edited by
Milan Kilibarda
Jelena Luković

Publisher
Faculty of Civil Engineering, University of Belgrade

For the publisher
Branko Božić

Printed by
Dosije studio, Belgrade
www.dosije.rs

Circulation
50 copies

ISBN 978-86-7518-190-3

©2016 Faculty of Civil Engineering, University of Belgrade
All copyrights reserved. Reprinting and photocopying prohibited.

Contents

Advanced Analysis of Environmental Data Using Machine Learning

[key note lecture abstract]

Mikhail Kanevski
IDYST, University of Lausanne, Switzerland
Mikhail.Kanevski@unil.ch

Abstract—**The application of Machine Learning Algorithms for the analysis, modelling and visualization of environmental data is presented within the framework of an advanced and consistent generic methodology. The main attention is paid to the following topics: monitoring networks analysis and optimization using active learning; exploratory data analysis with machine learning algorithms; recognition of spatially structured patterns; relevant features selection and dimensionality reduction; visualization of high dimensional data and visual data mining; analysis and justification of the results. Simulated data - for better understanding of the ideas, and real data (topo-climatic, natural hazards, pollution) case studies are considered. In conclusion, new challenges of machine learning applications in environmental data science are discussed.**

Automated global soil mapping: discovering spatial soil patterns using machine learning

[key note lecture]

Tomislav Hengl
ISRIC - World Soil Information
Wageningen, The Netherlands
tom.hengl@isric.org

Time series analysis of Sentinel-1 backscatter data on a high performance computing platform

[key note lecture]

Wolfgang Wagner
Department of Geodesy and Geoinformation (GEO)
TU Wien, Austria
Wolfgang.Wagner@geo.tuwien.ac.at

Stjepan Mohorovičić, an early advocate of Milutin Milanković's theory of climate change

[key note lecture abstract]

Mirko Orlić
Andrija Mohorovičić Geophysical Institute, Faculty of Science
University of Zagreb, Croatia
morlic@gfz.hr

Abstract— **Four publications authored by Stjepan Mohorovičić between the years 1921 and 1947, in which he favorably discussed Milutin Milanković's theory of climate change, are commented upon. Moreover, it is shown that M. Milanković cited S. Mohorovičić in his "Canon of insolation and ice-age problem" published in 1941. The life and work of S. Mohorovičić are briefly described. It is pointed out that he made important contributions to mathematics, physics, geophysics and astronomy. In particular, his theoretical prediction of the existence of a bound positron-electron system, which has been subsequently detected by experimentalists and which is today called positronium, is mentioned. It is therefore concluded that S. Mohorovičić was an excellent scientist, who was able to recognize the importance of M. Milanković's theory immediately after it had been first published in 1920 and who was willing to popularize the theory at a time when it was often subjected to criticism.**

Exploring relationship between global temperatures and global sea levels

[key note lecture abstract]

Mirko Orlić
Andrija Mohorovičić Geophysical Institute, Faculty of Science
University of Zagreb, Croatia
morlic@gfz.hr

***Abstract*— Global sea level has been rising over the last hundred or so years. The rise is expected to continue, and even to accelerate, in the future due to the global climate change. Some recent storm surge events, particularly those related to hurricanes, have revealed how sensitive the coastal population is to flooding. The future sea level rise will obviously exacerbate the sensitivity. The presentation starts with a brief overview of three different methods that are used to determine sea level change: reconstruction based on geological indicators, measurement with tide gauges and measurement with satellite altimeters. Sea level variability over hundreds of thousands of years and global sea level rise amounting to approximately 17 cm over the past century is described. It is pointed out that the recent rise depended on an increase of the ocean volume due to the absorption of heat and on an increase of the ocean mass caused by the melting of glaciers and ice sheets. It is also shown that the Mediterranean sea level rose more slowly for a while and that an acceleration was observed recently. Additionally, extreme flooding events in various places around the world are mentioned, in particular those in New Orleans (2005), Myanmar (2008), New York (2012) and Philippines (2013). The forecast of storm surges is briefly commented upon. Flooding events in the Adriatic are described as well; it is pointed out that the Bakar tide gauge – the oldest one in Croatia – recorded four highest sea levels after the year 2008.**

When modeling mean sea level, one usually starts with the comparison of past sea levels obtained by a method to the observed values and then proceeds by applying the method to determine future sea levels. The method utilized may be either process-based or semi-empirical. The presentation proceeds with an overview of three variants of the semi-empirical method being used in analysis and projection of not only sea levels but also sea level trends. The variants differ in assuming that the response of sea level to the temperature forcing is equilibrium, inertial or a combination of the two. All variants enable temperatures, sea levels and/or sea level trends to be successfully regressed, albeit with controlling parameters that differ among the cases. The related response times vary considerably, with a realistic value (ca. 50 years) obtained only if both equilibrium and inertial dynamics are taken into account. It is stressed that a comparison of computed sea levels to those measured over the last century showed that the best agreement is provided by the purely inertial variant of the semi-empirical method. On the other hand, a comparison of computed sea level trends to the corresponding values determined from available measurements pointed to the equilibrium-cum-inertial variant of the semi-empirical method as the most successful one. Sea levels projected by using the three variants were found to be similar through the middle of the 21st century but also to radically diverge by the end of the 23rd century. Sea level trends considerably differed throughout the projection interval. It is therefore pointed out that one has to be careful while calibrating a method on the data having a specific spectral content and then using the method to prepare projections under the forcing having different spectral characteristics.

Observation based gridded climate data in Norway

[key note lecture]

Ole Einar Tveito
Norwegian Meteorological Institute
Oslo, Norway
ole.einar.tveito@met.no

Feature Extraction for Rasters Using Autoencoders

[extended abstract]

Miloš Manić, Mladen Nikolić
Department of Computer Science
Faculty of Mathematics, University of Belgrade
Belgrade, Serbia
nikolic@matf.bg.ac.rs

***Abstract*—Machine learning models are usually trained assuming that instances of interest are represented by vectors of numerical features, often crafted by domain experts. As an alternative to expert crafted features, neural networks were used to automatically extract features from raw data in variety of tasks. Autoencoders are a specific kind of neural networks which extract features in an unsupervised manner by learning a compressed representation of raw data in their hidden layers, such that satisfactory reconstruction of the input data is possible from that representation. We used autoencoders to learn such features for raster data. Moreover, we show how autoencoder training can be performed faster by initializing autoencoder for higher resolution rasters using weights of autoencoder trained for lower resolution rasters. We performed evaluation on daily temperature maps. We evaluated autoencoders with number of hidden units (i.e. extracted features) much smaller than the number of inputs and outputs, meaning that the compression ratio is high. We show high correlation between distances computed over extracted features and distances computed over original rasters, meaning that autoencoder based feature extraction can facilitate use of various distance based machine learning or information retrieval tasks on rasters.**

Keywords—feature extraction; rasters; autoencoders

I. Introduction

Feature extraction is one of the major tasks of machine learning, since the availability of informative features is of great importance for any application of machine learning methods. For a specific problem, features are often crafted by domain experts. An alternative approach is to extract features automatically from the raw data using methods designed for that task [1]. Artificial neural networks are characteristic for seamlessly merging feature extraction with learning, features being learned by hidden units of the network. A specific kind of neural networks which extract features in an unsupervised manner, and thus without relation to specific learning task, are autoencoders [2,3,4,5].

The volume of raster data related to environmental sciences is rapidly growing, constantly generating the need for improving various aspects of automated processing, search, and storage methods. Machine learning has already proven to be a valuable tool in environmental data processing [6,7,8]. As already noted, one of the key ingredients for successful application of machine learning methods are good quality features. Relatively small set of such features could also be used to speed up the search for similar rasters in a raster database. Namely, considering the usual size of rasters, finding a raster in a database similar to the given one may be computationally expensive task if the rasters are to be compared on pixel level. However, if a raster can be approximately represented by small number of numerical features, that task becomes computationally less demanding.

The contributions of our work are the following:

- We use autoencoders to extract small number of informative features from temperature rasters for Serbia and evaluate the quality of reconstruction of rasters from those features.
- We demonstrate that distances computed over these features are highly correlated with distances computed over original rasters, meaning that they can be used to facilitate use of various distance based machine learning or information retrieval tasks on rasters.
- We propose a procedure that speeds up autoencoder training, which relies on training smaller autoencoders on lower resolution rasters and using their weights to initialize the training of autoencoders for higher resolution rasters.

II. Autoencoders

One of the most successful machine learning models are *artificial neural networks* (ANNs) [9,10]. ANNs usually consist of connected units which perform nonlinear transformation over their multiple inputs first by computing a linear combination of the inputs and then applying a nonlinear *activation function* to that value. Linear combination is determined by a set of *weights*. Training ANNs consists of determining the set of weights for all units of an ANN in order to achieve desired behavior of the ANN. ANNs come in various flavours, one of the simpler and most often used being *layered feed-forward neural networks*. Such ANNs are characterized by arrangement of units in layers so that each unit in each layer takes only outputs of units of the previous layer as its inputs. The last layer is called *output layer* and other layers are called *hidden layers*. Inputs themselves are often called *input layer* even though they do not consist of units.

Autoencoder is a specific kind of unsupervised layered feed-forward neural network [2,3,4,5]. The number of its inputs and the number of its output units is equal. It is trained to copy its inputs to outputs, apparently a useless task. However, autoencoder has at least one hidden layer and it is exactly those layers which are important. Autoencoder consists of an *encoder* and a *decoder*. In its simplest form, the encoder consists of one layer which computes latent representation of the input data and the decoder consists of one layer which reconstructs the input data from that latent representation. In case that the encoder and the decoder consist of more than one layer, they make a *deep autoencoder*.

More formaly, an autoencoder implements two functions, encoder $f: R^N \to R^M$and decoder $g: R^M \to R^N$such that $g(f(x)) \approx x$. One of the usual choices for encoder and decoder are

$$\mathrm{h}=\mathrm{f}(x)=\sigma(\mathrm{Wx}+\mathrm{b})$$

$$g(x)=\sigma(\mathrm{W'h}+\mathrm{b'})$$

where W,W',b,b'are weights to be learned from the data and σ is a function:

$$\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}}$$

which maps real numbers to interval [0,1]. The training is performed by choosing the weights in order to minimize some loss function $L(\mathrm{x},\mathrm{g}(f(x)))$, the usual choice being:

$$L(\mathrm{x,y}) = (x-y)^2$$

Most important applications of autoencoders are feature extraction and dimensionality reduction. Feature extraction is performed by encoder which yields a vector of new features h. Considering that the original data can be approximately reconstructed from these features by decoder, it follows that these features contain most of the information present in the original data. Since the number of extracted features $M$is usually much smaller than the number of inputs N, they can be understood as a kind of compression or low dimensional representation of the original data. The decoder can be understood as nonlinear low dimensional parametrization of high dimensional data manifold. However, sometimes autoencoders are used to learn high dimensional representations of the data and then it holds M>N. In such cases, model contains large number of weights and its flexibility is controlled by additional regularizations.

III. Training Procedure

We propose a procedure for autoencoder training, which relies on training smaller autoencoders on lower resolution rasters and using their weights to initialize the training of autoencoders for higher resolution rasters. For simplicity, we focus on the simplest form of autoencoders, but the extension of the procedure to deep autoencoders can be formulated in an analogous manner.

Let Nbe the number of inputs of the autoencoder and $M$the number of hidden units. Let $N_0$and M_0be integers and let $N_{\mathrm{i}+1}=\mathrm{n}_i N_i$and $M_{\mathrm{i}+1}=\mathrm{m}_i M_i$for $1 \leqslant i \leqslant L$where $n_i$and $m_i$are sequences of integers such that N_L=Nand M_L=M. Denote by □ Kronecker product of matrices defined as:

$$\mathrm{A}\otimes\mathrm{B}=\begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{\mathrm{mn}}B \end{bmatrix}$$

Denote by $1_{p\times q}$a matrix of dimensions $p \times q$such that its all entries are 1. The training procedure is as follows:

1. Set i=0.

2. Initialize W_i,W_i',b_i,$\mathrm{b}_i'$to random matrices of appropriate dimensions.

3. Train autoencoder with $M_i$hidden units on rasters with $N_i$pixels with initial parameter values W_i,W_i',b_i,$\mathrm{b}_i'$and update the parameters by values obtained during training.

4. Set

$$W_{\mathrm{i}+1} = \frac{W_i \otimes 1_{m_i\times n_i}}{n_i} \quad \mathrm{W'}_{\mathrm{i}+1} = \frac{\mathrm{W'}_i \otimes 1_{n_i\times m_i}}{m_i}$$

$$b_{\mathrm{i}+1}=\mathrm{b}_i \otimes 1_{m_i\times 1} \quad \mathrm{b'}_{\mathrm{i}+1}=\mathrm{b'}_i \otimes 1_{n_i\times 1}$$

5. If i<L, set i=i+1and go to step 3. Otherwise, stop.

IV. Experimental Evaluation

We evaluate the quality of autoencoder based feature extraction and the ability of extracted features to represent rasters with distance based techniques. Also, we evaluate the effectiveness of proposed training procedure.

A. *Data*

In the experiments we use temperature map rasters representing maximal daily temperature on the territory of Serbia, for each day during 10 year period of 2006-2015. We use rasters in different resolutions. We will express resolution by the area a pixel represents (e.g. 2km × 2km). Data from years 2006-2014 are used for training and data from year 2015 are used for testing. We use three datasets of rasters of resolution 8km × 8km, 4km × 4km, and 2km × 2km. In each dataset, each raster consists of of 1973, 7888, and 31645 pixels respectively. First two are auxiliary ones used for the training procedure we propose and the last one is intended for evaluation.

B. *Experiments*

As a first step in evaluation of autoencoder based feature extraction, we train autoencoder, using both proposed training procedure and ordinary training approach. The proposed training procedure uses rasters of resolution 8km × 8km, 4km × 4km, and 2km × 2km. The ordinary training approach is applied directly on rasters of resolution 2km × 2km. In both cases, the final autoencoder has 31645 inputs and outputs and the number of units in the hidden layer is chosen to be 12. We did not try different number of units in the hidden layer, since this choice performed well and we lacked computational resources in this stage of our research. For comparison, we also use the mean value of the training set as the most naïve prediction method. The experiments were performed on laptop computer with Intel Core i7 processors with 4 cores working at 2.1GHz and 6 GB RAM. For autoencoder training, we used TensorFlow library [11].

Regarding training efficiency and quality, we report training time, number of epochs used in the training (number of training iterations), and error obtained on the test set and compression ratio. Error is computed as an average of absolute differences between true and reconstructed pixel values which are expressed in percentages of temperature range of the data (from -9.3°C to 39.4°C). Compression ratio is computed by comparing the size of the compressed dataset and the size of the original dataset. Compressed dataset consists of parameters of the decoder and feature vectors of all rasters in the dataset.

Regarding the usefulness of extracted features in context of distance based machine learning methods, we report two statistics. The first one is the Pearson's correlation coefficient of Euclidean distances between all pairs of original rasters and Euclidean distances between their 12 dimensional feature vectors. The second one is computed as follows. For each raster in the test set, we find its nearest neighbor with respect to Euclidean distance over original rasters. Then, for each raster we sort whole test set according to the Euclidean distance over 12 dimensional feature vectors to that given raster, its nearest neighbor is found in that sequence, and its rank (index) is recorded. We report the average of those ranks.

The results of our experiments are given in Table I. They show that there is a significant decrease in training time thanks to our training procedure. It yields a model with several times lower test error in much less time, compared to ordinary training. The compression ratio is very high. The correlation coefficient is also very high, indicating that the distances over feature vectors could be used instead of distances over rasters. The average rank indicates that if the rasters were retrieved from a database, based on extracted features, truly most similar raster would be among 5 or 6 top ranked rasters.

TABLE I.

	8 × 8 → 4 × 4 → 2 × 2	2 × 2	Mean
Training time (s)	7643	37539	-
Epoch count	3000+2000+1000	8000	-
Test error	0.65%	2.15%	17.42%
Compression ratio	99.64%	99.64%	-
Correlation coefficient	0.95	0.93	-
Average rank	5.68	8.43	-

V. CONCLUSIONS AND FUTURE WORK

In this paper we showed how one can perform automated feature extraction from rasters. We used autoencoders, a kind of neural network which has been used for similar tasks in other domains. For daily temperature data, we demonstrated that the number of extracted features can be very small compared to the number of pixels in the raster and, still, the rasters can be reconstructed from them with very small average error. That means that a small number of informative features approximately parametrizes the high dimensional raster space of temperature data. We performed experiments which demonstrated that distances between vectors of extracted features approximately correspond to distances between original rasters, meaning that numerous distance based machine learning, data mining, and information retrieval methods can be used with these features, instead of the original rasters, thus relieving the computational burden. The applications may include classification or search tasks.

Regarding future work, a more thorough evaluation, which requires greater computational resources, is to be performed. The number of units in the hidden layer should be chosen by cross-validation instead of relying on vague intuition. Deep autoencoder approach should be evaluated, too. Details of the training procedure should be considered more thoroughly in order to provide maximal training speed up. Evaluation on 1km × 1kmrasters is intended. Regarding applications, we intend to implement a fast raster search engine based on autoencoder based feature extraction method.

ACKNOWLEDGMENT

The authors thank Branislav Bajat, Aleksandar Sekulić, and Milan Kilibarda from Faculty of Civil Engineering at the University of Belgrade, for kindly providing the data needed for this research and useful discussions and advice.

REFERENCES

[1] I Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Eds., “Feature extraction, foundations and applications”, Springer, 2006.

[2] D. Rumelhart, G. Hinton, R. Williams, “Learning representations by backpropagation errors”, Nature, vol. 323, pp. 533-536, 1986.

[3] M. Ranzato, F. Huang, Y. LeCun, “Unsupervised learning of invariant feature hierarchies with applications to object recognition”, Proceedings of Computer Vision and Pattern Recognition Conference, 2007.

[4] G. Hinton, S. Osindero, Y. Teh, “A fast learning algorithm for deep belief nets”, Neural Computation, vol 18, pp. 1527-1554, 2006.

[5] G. Hinton, R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, Science, vol. 313, 2006.

[6] M. Kanevski, Ed., “Advanced mapping of environmental data”, Wiley, 2008.

[7] M. Kanevski, V. Timonin, A. Pozdnukhov, “Machine learning for spatial environmental data: theory, applications, and software”, CRC Press, 2009.

[8] W. Hsieh, “Machine learning methods in the environmental sciences”, Cambridge University Press, 2009.

[9] M. Hassoun, “Fundamentals of Artificial Neural Networks”, MIT Press, 2003.

[10] I. Good, Y. Bengio, A. Courville, “Deep learning”, unpublished.

[11] M. Abadi et al, TensorFlow: large-scale machine learning on heterogeneous systems, unpublished.

Kriging with machine learning covariantes in environmental sciences: A hybrid approach

[full paper]

Velibor Ilić[a], Jovan Tadić[b], Aleksandra Imširagić[c]

[a] RT-RK Institute for Computer
Based Systems,
21000 Novi Sad, Serbia
velibor.ilic@rt-rk.com

[b] Carnegie Institution for Science,
Department of Global Ecology,
Stanford, CA 94305, USA
jtadic@stanford.edu

[c] University of Novi Sad, Faculty of
Technical Sciences,
Department of Environmental
Engineering,
21 000 Novi Sad, Serbia
leapsv@gmail.com

Abstract - **Kriging is probably the most frequently used method in spatial interpolations in environmental sciences. Also, this method can be used as a model of innovative risk assessment of air quality and climate changes. It is generally accepted that inclusion of auxiliary variables improves the accuracy of the kriged values. Auxiliary information could be incorporated into kriging scheme in two ways: using co-kriging, or kriging with external drift. The fact that kriging is a linear technique could represent a disadvantage in cases where more complex spatial auto-correlation structure is encountered. Machine learning methods (and neural networks among them) could handle well the nonlinearity in the data, but being a non-exact interpolator they are not guaranteed to provide unbiased estimates, nor estimates with minimized variance. In this study we applied a combined (or hybrid) method on CO_2 mixing ratio spatial distribution pattern. Kriging was used as a primary interpolation method and neural network predictions were used as covariates through kriging with external drift interpolation framework. There are two main conclusions from this study: (a) neural networks cannot compete with geostatistical tools specifically developed for geospatial analysis, (b) incorporation of neural networks outputs as covariates in kriging schemes, can improve the overall accuracy despite the poorer separate neural network results.**

Keywords—Kriging with external drift, Universal kriging, Airborne measurements, estimation CO2 mixing ratio, Neural networks, Machine learning, Ensemble, environmental engineering

I. Introduction

Kriging is probably the most frequently used method in spatial interpolations in environmental sciences. It represents an exact (exactly reproduces observations) geostatistical interpolation method in which the estimated value is expressed as a linear combination of known/measured values, after modeling the spatial covariance structure of the data, usually through the process known as variography or variogram analysis[1]. The estimation of the interpolation uncertainties is obtained as a kriging by-product. In kriging, it is easy to make use of auxiliary information, as long as a correlation exists between the interpolated quantity and auxiliary variable(s). It is generally accepted that inclusion of auxiliary variables improves the accuracy of the predictions[2].

Auxiliary information could be incorporated into kriging scheme in two ways, using co-kriging[1], or kriging with external drift[3]. In kriging with external drift it is necessary to have values of the auxiliary variables available at the estimation locations, while in co-kriging the spatial correlation structure is modeled in a similar way as the auto-correlation structure of the primary variable – usually through variography. The fact that kriging is a linear geostatistical technique could represent a disadvantage in cases where more complex spatial auto-correlation structure is encountered. On the other hand, machine learning methods (and neural networks among them) could handle well the nonlinearity in the data, but being a non-exact interpolator they are not guaranteed to provide unbiased estimates, nor estimates with optimal variance. Artificial neural networks have been used as an alternative method for spatial interpolation [4,5,6,7]. Combining geostatistical (linear, unbiased and best) interpolators with machine learning methods have some promising properties. It should preserve the convenient properties of the kriging, while including effects of the non-linearity in the data through machine learning component of the hybrid method. How to combine them? In majority of the existing studies machine learning was used as a primary method[8] and some sort of kriging was used to interpolate residuals on the estimates produced by a machine learning technique. The method yielded promising results, and in a review study of the various interpolation methods used in environmental sciences it was shown that a combination of Random forest (machine learning method) together with ordinary kriged residuals produced 30% more accurate predictions that any other method[9]. There are at least two serious objections towards such an approach. First, it is hard to compare interpolation performance using different test cases, and second, the weaknesses of the machine learning methods (e.g. biasedness) are preserved.

In this study we try a different approach, built on the work by Tadić et al.(2015)[4]. We use kriging as a primary interpolation method, but we use machine learning predictions as covariates through kriging with external drift interpolation scheme. This alternative approach showed some promising properties[4], but the evidence was inconclusive.

Notice that in such an approach primary advantages of kriging are preserved. We applied the new method to the set of *in situ* airborne measurement collected by NASA over a cluster of refineries near San Ardo, CA. The performance of the new method, and universal kriging and neural networks applied separately in reproduction of the observed values, is evaluated using leave-one-out cross validation technique.

II. Measurements

In situ airborne CO_2 data were collected in a flight around a cluster of refineries near San Ardo, CA (35º57'N- 120º52'W), on 02/21/2013. The aircraft and instrumentation have been described in detail elsewhere[10] and will be only briefly discussed here. Measurements were taken by a tactical fighter aircraft (Alpha Jet, based at NASA Ames Research Center), equipped with a Picarro 2301-m cavity ring-down instrument for CO_2, CH_4 and H_2O measurements, GPS and inertial navigation systems that provide position information (latitude, longitude, altitude).

III. Kriging with external drift

Before we describe details of the applied technique, we point out to the difference between kriging with external drift and universal kriging. Universal kriging is the kriging technique used for data with a significant spatial trend where only coordinate and functions of the coordinates are used as auxiliary variables, while in kriging with external drift the choice is more flexible – any auxiliary variable could be used. The inclusion of spatial explanatory variables accounts for the possible nonstationarity of the random CO_2 mixing ratio field[11]. In this study we used both functions of the coordinates and machine learning outputs as auxiliary variables and thus classified the method into kriging with external drift category. Height and height squared were used as covariates, mimicking the approach from[4]. Before the kriging was performed, the coordinates were converted from WGS84 coordinate system (lat/lon) into a Universal Transverse Mercator (UTM) coordinate system, to allow for computation of distances and angles using Euclidean geometry[8].

IV. Using neural networks for CO2 mixing ratio estimation

For estimation CO_2 mixing-ratio in this study is used back-propagation neural network with following configuration, 8 nodes at input layer, 9 nodes at hidden layer, and one node at output layer. Input for neural network represent array with 8 values that contains absolute values (longitude, latitude, and altitude), CO_2 mixing-ratio at one of referent point, distance between current point and referent point and relative values (longitude, latitude, and altitude) between current point (*CurrentPoint[X]*) and referent points (*RefPoint[Y]*). Referent points are extracted from a flight path log file as every *n* point. Training set is generated as relation between current points and each of referent points, see figure 1.

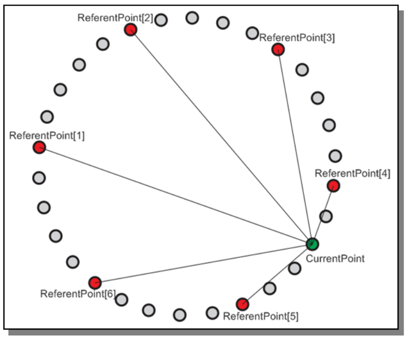


Fig. 1. Creating training set using referent points

The inputs of the ANNs were calculated using the values of the estimation location and values of one of the referent points. Input values of neural network are calculated using the following formulas:

Input1 = CurrentPoint[X].Latitude
Input2 = CurrentPoint[X].Longitude
Input3 = CurrentPoint[X].Altitude
Input4 = RefPoint[Y].CO2mixingRatio
Input5 = Distance(RefPoint[Y], CurrentPoint[X])
Input6 = (RefPoint[Y].Latitude - CurrentPoint[X].Latitude)
Input7 = (RefPoint[Y].Longitude - CurrentPoint[X].Longitude)
Input8 = (RefPoint[Y].Altitude - CurrentPoint[X].Altitude)

Output values of neural network are the concentrations of CO_2. All values in the neural networks training set should be re-scaled into the 0-1 range.

File with the measurement data contains 2749 measured points. Every 25th point from that file is used as referent point, in this way is obtained 110 referent points. Calculating relations between referent points and other measured points is created training set of 290290 pairs. For training neural network and estimations was used a PC-based software "ANN V2.5"[12] When an ANN estimates the CO_2 mixing-ratio at a certain point, it actually estimates mixing ratios at that point related to each of 110 referent points, the mean of which represents the final result, see figure 2.

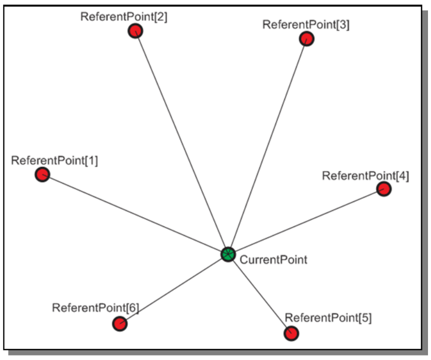


Fig. 2. Estimation CO_2 mixing-ratio

In this study, an ensemble of 25 ANNS was used in interpolations. To calculate CO_2 mixing-ratio at estimation location (long/lat/alt) it is required to determine the 8 input values using same procedure and referent points as it is used for generating training set. Those 8 values are sent to the inputs nodes of all of 25 neural networks in ensemble. The final result is calculated as average value of output results of those networks, figure 3.

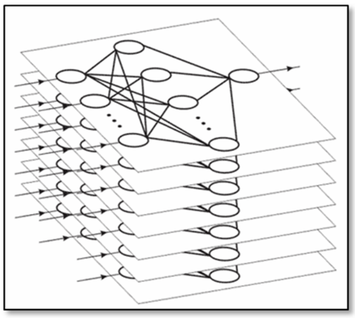


Fig. 3. Ensemble of neural networks

V. Cross-validation

Cross-validation is a widely used method for assessing prediction accuracy of statistical models[13]· In this study, we used the leave-one-out cross-validation to assess the prediction accuracy of all three methods. A detailed description of the method can be referred to Cressie (1993)[14]. The cross-validation is carried out by first removing one observation and then making its prediction using the remaining dataset. This process is repeated for each observation using all three methods. As a measure of the performance the mean absolute error (MAE) was used. The MAE was found to be 0.0841, 0.4568 and 0.0834 ppm, for universal kriging, ensemble neural networks and hybrid method, respectively.

VI. Discussion and conclusions

This study was built up on the work by Tadić et al., 2015[4], and provided further information about the potential of hybridizing machine learning with geostatistical methods, in this case universal kriging. There are two main conclusions from this study: (a) neural networks, being a widespread estimation tool used not only in geospatial analysis, cannot compete with geostatistical tools specifically developed for geospatial analysis if input information is limited to coordinates and/or functions of coordinates. Significantly higher MAE compared to kriging (4.5 times higher) undoubtedly confirm such a conclusion, and (b) incorporation of machine learning outputs as covariates in kriging schemes, even in cases where machine learning performs significantly worse than kriging, does not adversely affect overall prediction, but opposite, we observed (though small) improvement in the overall accuracy despite the poorer separate neural network results.

Apart from two obvious conclusions, we believe that incorporating machine learning predictions as covariates could have additional benefit. Namely, the ancillary information used as inputs into neural networks could widely vary in nature, and could incorporate both numerical and categorical variables. Thus, machine learning represents a way to combine and incorporate disparate information sources and subsequently include them into kriging through covariates.

The hybridization technique used in this paper is not the only one available and future studies should reveal the true benefit of using machine learning as a primary technique followed by kriging of the residuals, compared to using kriging with external drift (where machine learning outputs are used as covariates).

On Figure 4 we show the predictions of all three methods on the numerical grid fitted into the flight trajectory. The plot is intended to demonstrate the degree up to which methods are capable of providing spatially resolved fine-scale estimates. The plot shows that inclusion of the machine learning (in this case neural network) outputs as covariates significantly changes the spatial patterns of the predictions.

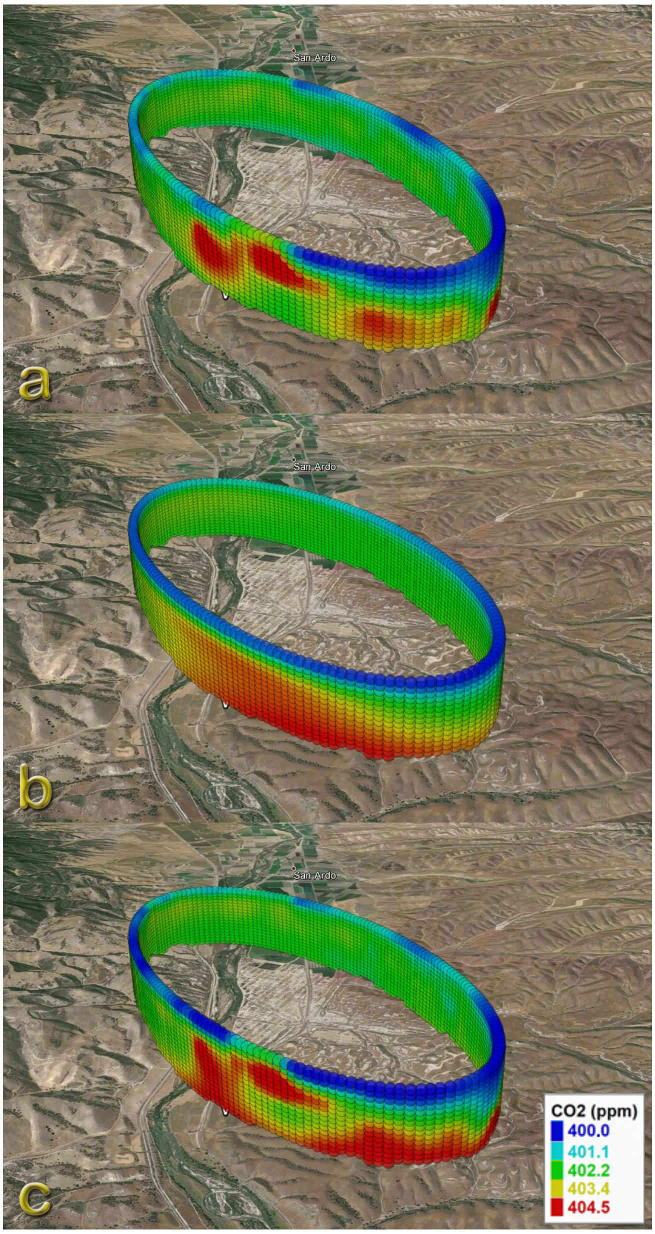


Fig. 4. Estimates on arbitrary projection grid fitted into flight trajectory using (a) Universal kriging, (b) Neural networks, and (c) Kriging with external drift using neural network outputs as covariates.

Method with neural networks and kriging can be used as a model of innovative risk assessment of air quality and climate changes [15,16]. Also it can be used for monitoring development which is important aspect in preservation and improvement of air quality. Based on monitoring results can be undertaken preventive measures in segments significant for protecting air from pollution: informing the public and giving recommendations for actions in situations of air pollution, monitoring and evaluation of trends of polluting substances, modeling (dispersion and prognosis of polluters concentration), evaluation of exposure of population and ecosystem, identification of sources of pollution, reviewing effects of undertaken measures to level of air pollution[15].

Acknowledgment

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number: TR32031.

References

[1] Chiles, J., and P. Delfiner (1999), Geostatistics: Modeling Spatial Uncertainty, 695 pp., John Wiley, Hoboken, N. J.

[2] Hengl, T., 2007. A Practical Guide to Geostatistical Mapping of Environmental Variables. Office for Official Publication of the European Communities, Luxembourg, p. 143.

[3] Wackernagel, H., 1998. Multivariate Geostatistics: An Introduction With Applications, 2nd ed. Springer, Berlin.

[4] Tadić J. M., Ilić V., Biraud S., (2015), "Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments", Atmospheric Environment, Volume 111, ISSN: 1352-2310, June 2015, pp 28–38, Elsevier Ltd, http://www.sciencedirect.com/science/article/pii/S1352231015002873

[5] Liu S., Zhang Y., Ma P., Lu B., Su H., A Novel Spatial Interpolation Method Based on the Integrated RBF Neural Network, Procedia Environmental Sciences, Volume 10, Part A, 2011, Pages 568-575, ISSN 1878-0296, http://dx.doi.org/10.1016/j.proenv.2011.09.092. (http://www.sciencedirect.com/science/article/pii/S1878029611002878)

[6] Gumus K., Sen A., 2013. Comparison of spatial interpolation methods and multi-layer neural networks for different point distributions on a digital elevation model. Geodetski vestnik, 57(3), pp.523-543.

[7] Sen A., Gümüsay M.U., Kavas A., Bulucu U.. Programming an Artificial neural network tool for spatial interpolation in GIS-A case study for indoor radio wave propagation of WLAN. Sensors. 2008 Sep 25;8(9):5996-6014.

[8] Li, J., Heap, A., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists No. Record 2008/23. Geoscience Australia, Canberra.

[9] Li, J., Heap, A., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, Vol. 26(12), 1647-1659.

[10] Tadić, J. M., M. Loewenstein, C. Frankenberg, A. Butz, M. Roby, L.T. Iraci, E.L. Yates, W. Gore, and A. Kuze (2014), A comparison of in-situ aircraft measurements of carbon dioxide and methane to GOSAT data measured over Railroad Valley playa, Nevada, USA. IEEE Trans. Geosci. Remote Sens., 52(12), doi: 10.1109/TGRS.2014.2318201.

[11] Chilès JP, Delfiner P. Wiley Series in Probability and Statistics. Geostatistics: Modeling Spatial Uncertainty, Second Edition. 2012:705-14.

[12] Ilić, V., 2000. Force learn algorithm e training neural networks with patterns which have highest errors. In: Seminar on Neural Network Applications in Electrical Engineering "NEUREL 2000", Belgrade, Sponsored by IEEE Signal Processing Society, pp. 46e48.

[13] Rivoirard, J. Two key parameters when choosing the kriging neighborhood. Mathematical geology, 1987, 23 19(8), 851-856. DOI: 10.1007/bf00893020.

[14] Cressie, N. Statistics for spatial data. John Wiley & Sons: USA, 1993; pp. 29-210.

[15] Imsiragić, A. (2013). Air Quality testing in the territory of the City of Belgrade . Master of Sciences, Novi Sad, Faculty of Technical Sciences

[16] Atkinson, P. M., and Lloyd, C. D., 1998: Mapping precipitation in Switzerland with ordinary and indicator kriging. J. Geogr. Inform. Decis. Anal., 2, 65 76

Regression trees for modelling water demand in Sevilla city (Spain)

[extended abstract]

Víctor Rodríguez-Galiano
Departamento de Geografía Física y Análisis Geográfico Regional
Facultad de Geografía e Historia. Universidad de Sevilla.
Sevilla, Spain
vrgaliano@us.es

María C. Villarín-Clavería
Departamento de Geografía Humana
Facultad de Geografía e Historia. Universidad de Sevilla.
Sevilla, Spain
mvillarin@us.es

***Abstract*— Domestic water demand emerged as an important focus for scientific research since the 1950s. More recently, other variables associated to water demand, such as territorial, demographic, environmental or technological factors have been incorporated to improve planning and supply management.**

This work shows the application of a machine learning method to the modelling of water demand for the first time. Regression trees, a multivariate, spatially non-stationary and non-linear machine learning approach, was used to build a predictive model of water demand for the city of Sevilla. The RT model was fitted to the relationship between the annual water demand and numerous explanatory variables related to socio-demographics and urban aspects. RTs allowed estimation of water demand with an error of 22 L and determination of the main driving variables. This research, thus, shows an alternative to the hitherto applied cluster and linear regression approaches for modelling water demand and paves the way for a new set of further scientific investigations based on machine learning methods.

Keywords—water demand, cenus tract, machine learning,

I. Introduction

Water resources management has emerged as an important focus for scientific research in the last decades. Water is a scarce resource of paramount importance for many aspects of our life and the environment. Population growth, rapid industrialization, and expanding and intensifying food production are exerting a significant pressure on the available water resources [1]. The European Environmental Agency reported that water resources were under increasing pressure as a consequence of higher demands of good quality water for many uses [2]. Besides this human pressure being made on water resources, recent studies suggest that climate change will lead to significant changes on freshwater systems.

Water demand modelling has been accessed through multivariate regression methods such as multivariate linear regression, analysis of hierarchical segmentation (CHAID method) and nominal multivariable logistic analysis; previously complemented in some cases with factor and cluster analysis [3-6]. The intricacy of water consumption modelling has required the inclusion of more precise and larger explanatory variables referred to socio- demographic and buildings-urban characteristics in the last years. In the context of multi-proxy approaches that include a high number of variables there is a need for the application of new generation computational tools to assist in extracting as much information as possible from the rapidly growing volumes of digital data. This is the case of the present research, related to a considerably large socio-demographic and buildings-urban dataset retrieved from census track.

II. METHODS

Domestic water demand modelling is a complicated task that might be driven by many different drivers, which are not necessarily the same within a given study area. Therefore, assembling a single global model for predicting water demand of an entire city cannot be very realistic and hopelessly confusing. Additionally, the drivers or predictive variables for water demand may interact in complicated, non-linear ways, which can undermine the potential of ordinary statistical techniques. An alternative approach to classical multivariate regression is to sub-divide, or partition, the space into smaller regions, where the interactions are more manageable. Regression trees appears as an alternative to traditional regression (global single predictive models) allowing for multiple non-linear regressions using recursive partitioning.

The choice of using trees (regression trees in our case) algorithms is usually associated to their simplicity and interpretability, to their low computational cost and to the possibility of being graphically represented. Hence, the main benefit of using a hierarchical tree structure to perform regression is that the tree structure can be viewed as a "white box", which in comparison to other machine learning techniques is easier to interpret for understanding the relations between the dependent and independent variables.

The socio-demographic and urbanisation variables (explanatory variables) and the water demand values (target variable) were combined together into a set of input feature vector. At each census tract 1 values from each socio-demographic and urbanisation variable were combined to form a vector. These vectors formed the input to Regression Tree [7] and Random Forest algorithms [8]. The water demand was used as target values for the induction of the models. Data processing for the induction of the MLA consisted of two stages: (i) training and parameterisation of the algorithms; and (ii) accuracy assessment. All of the MLA models were created using the R 3.2.3 (R-Project) free software. Within this environment, “e1071 library” was used for inducting both RT and RF.

III. RESULTS

A. Regression Tree (RT) Model.

The regression model with the lowest RMSE (22.60 L/y) was created by using a minimum value of 36 cases in every terminal node and a cost complexity factor of 0.001. This model was considered robust given the complexity of data. From a set of fifteen socio-demographic and urbanisation variables (Table 1), the obtained RT considered only five (HS, ABSA, RD, P<15 and AAP) as the most relevant for establishing differences between samples. Changes in the selected variables will be interpreted as reflecting the characteristics of groups with different water demands. A summary of obtained first and second best ranked decision parameters (Table 2) was consistent with significant relations previously identified for the database under scope (Table 3), with the exception of HS, whose importance was underestimated when using Pearson correlation coefficient. HS is the variable that provided the best solution for the analysed database (Table 2 and Fig. 1 node 1). RD as the second best ranked variable (Table 3, and Fig. 1 node 3) is also in agreement with the obtained correlation between HS and RD (Table 3). AAP was also very important as a primary split variable and especially as a surrogate variable.

TABLE 1. EXPLANATORY VARIABLES

Name	Units
Domestic water consumption (DC)	L/inhab./day
Population under 15 (P<15)	%
Population between 15 and 34 (P1534)	%
Population between 35 and 64 (P3564)	%
Population over 65 (P>65)	%
Youth index (YI)	%
Aging index (AI)	%
Foreigners (FRG)	%
Average age of population (AAP)	years
Average cadastral value (ACV)	€
Average built surface area (ABSA)	m^2
Weighted average height (WAH)	number of floors
Average gross density (AGD)	inhab./m^2
Average net density (AND)	inhab./m^2 (constructed)
Household size (HS)	inhab./household
Residential density (RD)	(inhab./household) ×100m^2

TABLE 2. VARIABLE IMPORTANCE IN THE RT MODEL

	HS	**RD**	**AAP**	**YI**	**P>65**
	20	14	13	11	10
ABSA	**P1534**	**P<15**	**ACV**	**AI**	**AND**
8	6	6	5	5	1

*A variable may appear in the tree many times, either as a primary or a surrogate variable. An overall measure of variable importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness * (adjusted agreement) for all splits in which it was a surrogate. (Table footnote)*

TABLE 3. CORRELATION MATRIX (PEARSON CORRELATION COEFFICIENT)

DC				
P<15	**P1534**	**P3564**	**P>65**	**AAP**
-0.398**	-0.207**	0.04	0.360**	0.453**
DC				
FRG	**YI**	**AI**	**ACV**	**ACV**
0.93*	-0.287**	0.369**	0.132**	0.420**
DC				
WAH	**AGD**	**AND**	**HS**	**RD**
0.197**	0.02	-0.308**	-0.329**	-0.479**

It should be noted that interpretability of the model plays an important role in this research and in many fields of social sciences. The goal was not to achieve the final results provided by the obtained regression tree (i.e., water demand estimation) but to better understand the synergies between the remaining variables. Since the application of the present method allowed a better overview of the spatial distribution of water demand throughout the city, this feature will be presented (Fig. 1), together with the decision rules that precedes it. The obtained regression tree is shown in Fig. 4, where each socio-economic variable is accompanied by the respective decision threshold value. The average estimated water demand is indicated at each terminal node, as well as to which city area the estimated values belong to. It is worth of mention that the distribution of samples along the decision tree is not arbitrary, but organized either by variables representing different characteristics of the city and the socio-urban complexity of it.

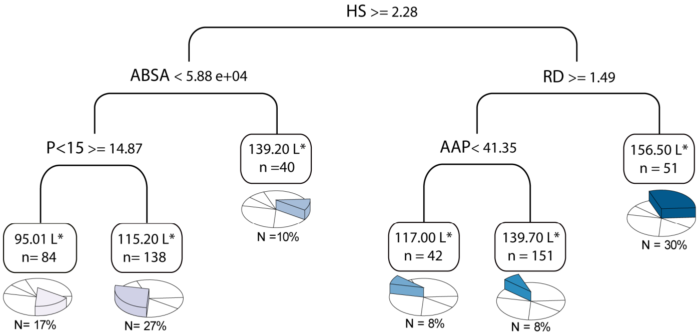


Fig. 1. Decision tree obtained for the census tract. Each explanatory variable is accompanied by the respective threshold value and a pie diagram indicates the number of samples included (N). Random Forest (RF) Model. Water demand values are given in L/inhab/day.

B. Random Forest (RF) Model.

The main drivers of water demand in Sevilla city were identified through the application of a feature selection procedure (see methods). Figure 2 shows the RMSE in the prediction of different models after removing the least important predictor. RMSE error values ranged from 18.89 L/y to 26.91 L/y. Figure 3 shows the pseudo-R^2 of the models as well as the relative importance of each explanative variable. RF water demand models explained a percentage of the variance up to 56%. Regarding the relative importance of the drivers, the same ranking in importance was observed within the different models, which reflected the stability in the RF importance estimation, and a high reliability of the results. To interpret the main socio-demographic drivers of the spatial variation in water, simplified model with reduced number of predictors was selected. The model was composed of 6 predictors (pseudo-R^2=0.54 and RMSE of 18.96 L/y). As in the case of RT, Our results suggest that spatial variation in the water demand of Sevilla city is driven mainly by HS and RD, assigning it much greater importance than to the rest of variables. Therefore, water consumption is Sevilla city is mainly associated to the population size. Also AAP, ACV, WAH and P<15 were of a significant importance in the model. Additionally, a linear regression between predicted values from RF and observed water demand produced R^2 values equal to 0.55 and RMSE value of 13.55 L/y (Figure 4). However, the lower and higher water demand values were over and underestimated, respectively.

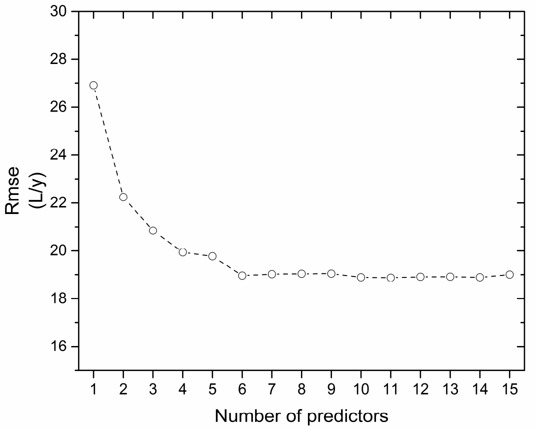


Fig. 2. Root mean square error (RMSE) of the models fitted as a result of the feature selection approach.

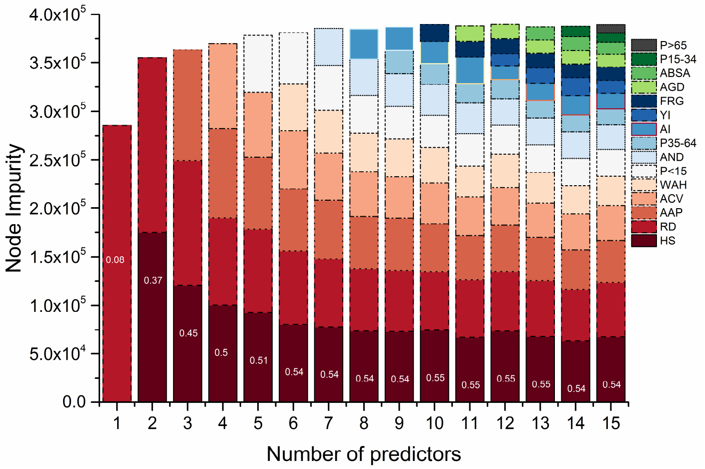


Fig. 3. Relative importance of each independent variable in predicting water demand in Sevilla city. Different models derived from the feature selection approach are represented in each column. Numbers given over each column represent the coefficient determination of each model

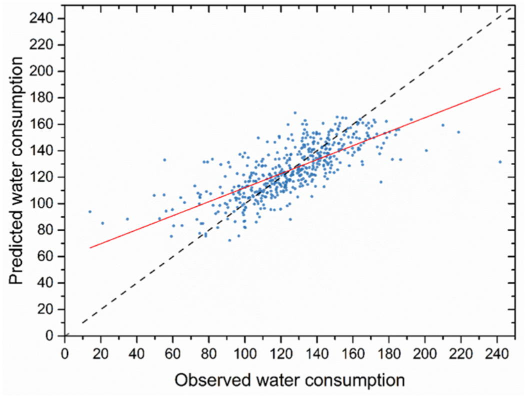


Fig. 4. Scatterplots between observed water demand and the predictions calculated using a selection of predictors (see Figure 2 and Figure 3). The dashed lines represent an exact 1:1 relationship (expected fitting), the solid lines show a linear regression of these data. The explained variances (percentage R^2) and RMSE values are 55% and 13.55 L/y.

Acknowledgment

The first author is a Juan de la Cierva Grant holder. The authors are grateful for the financial support given by the European Commission under the Seventh Framework Programme and the Spanish MINECO (project BIA2013-43462-P). We acknowledge, Andalusian Statistics and Cartography Institute (IECA), General Directorate of Cadastre and Sevilla Metropolitan Drinking and Waste Water Company (EMASESA) for proving the water demand data.

References

[1] E. C. Corcoran, N. E., B. R. *et al.*, *Sick Water?. The Central Role of Wastewater Management in Sustainable Development*, 2010.

[2] EEA, "Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy," *Official Journal of the European Parliament,* vol. L327, no. September 1996, pp. 1-82, 2000.

[3] H. E. Campbell, E. H. Larson, R. M. Johnson *et al.*, *Some best bets in residential water conservation: results of a multivariate regression analysis. City of Phoenix, 1990-1996. Final Report*, Arizona, USA, 1999.

[4] E. Domene, and D. Saurí, "Urbanisation and Water Consumption: Influencing Factors in the Metropolitan Region of Barcelona," *Urban Studies,* vol. 43, no. 9, pp. 1605-1623, August 1, 2006, 2006.

[5] M. Loh, and P. Coghlan, *Domestic water use study in Perth, Western Australia 1998-2001*, Perth, Australia, 2003.

[6] P. W. Mayer, W. B. DeOreo, E. M. Opitz *et al.*, *Residential End Uses of Water*, USA, 1999.

[7] L. Breiman, *Classification and regression trees*: Chapman & Hall/CRC, 1984.

[8] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

Machine learning as a generic framework for spatial and spatiotemporal prediction

[abstract]

Tomislav Hengl
ISRIC - World Soil Information
Wageningen, The Netherlands
tom.hengl@isric.org

***Abstract*—A generic framework for spatial and spatiotemporal predictions based on Machine learning (quantile regression forests and Gradient Boosting Tree) is described and illustrated with some data examples from the literature: spatial prediction of Zinc concretions in soil (with and without covariates), spatial prediction of soil organic carbon in 3D, and spatiotemporal prediction of daily temperatures. To account for spatial autocorrelation i.e. spatially dependent variation, buffer distances to groups of values in response space are used in addition to remote sensing (RS) and DEM-based covariates. This allows for a full extension of MLA tree-based models to spatial problems where spatial dependence needs to be included in the model building. The proposed method is compared to standard geostatistical techniques such as ordinary kriging, regression-kriging (based on a Generalized Linear Model) and spatiotemporal regression-kriging. The cross-validation (accuracy assessment) results show promising potential for machine learning. The methodology shows especial usefulness for sorting the predictor variables based on importance, for visualizing complex non-linear relationships, and for highlighting possible outliers and blunders in the input data.**

Random Forest for modelling anomalies in the land surface phenology of the European forest

[extended abstract]

Victor Rodriguez-Galiano
Departamento de Geografía Física y Análisis Geográfico Regional
Facultad de Geografía e Historia. Universidad de Sevilla.
Sevilla, Spain
vrgaliano@us.es

Jose A. Caparros-Santiago
Departamento de Geografía Física y Análisis Geográfico Regional
Facultad de Geografía e Historia. Universidad de Sevilla.
Sevilla, Spain

***Abstract*— Vegetation phenological events, such as the timing of leaf onset or leaf senescence, vary between years depending on climatic conditions. This research reveals new insights into the weather drivers of interannual variation in land surface phenology (LSP) across the entire European forest, while at the same time establishes a new conceptual framework for predictive modelling of LSP. Specifically, the Random Forest method, a multivariate, spatially non-stationary and non-linear machine learning approach, was introduced for phenological modelling across very large areas and across multiple years simultaneously: the typical case for satellite-observed LSP. The RF model was fitted to the relation between LSP interannual variation and numerous climate predictor variables computed at biologically-relevant rather than human-imposed temporal scales. In addition, the legacy effect of an advanced or delayed spring on autumn phenology was explored. The RF models explained 81% and 62% of the variance in the spring and autumn LSP interannual variation, with relative errors of 10% and 20%, respectively: a level of precision that has until now been unobtainable at the continental scale. Multivariate linear regression models explained only 36% and 25%, respectively. It also allowed identification of the main drivers of the interannual variation in LSP through its estimation of variable importance. This research, thus, shows an alternative to the hitherto applied linear regression approaches for modelling LSP and paves the way for further scientific investigation based on machine learning methods.**

***Keywords*— modelling; anomalies; land surface phenology; climate variability; MERIS**

I. Introduction

Land Surface phenology (LSP), the study of the timing of recurring cycles of vegetation growth using time series of satellite sensor derived vegetation indices, is a valuable tool for the monitoring of vegetation at global or continental scales and allows evaluating the impacts of climate change. Vegetation phenological events, such as the timing of leaf onset or leaf senescence, vary between years depending on climatic conditions.

Modelling efforts to characterize LSP have generally relied on functions (usually linear) of meteorological drivers, such as average temperature and precipitation, growing degree days (GDD), light and temperature, minimum temperature, photoperiod, vapour pressure deficit, or minimum relative humidity. However, there is lack of understanding on number of important aspects, such us the multivariate influence of meteorological variables (temperature, precipitation, solar radiation) driving phenology, or the effect of additional drivers in the modelling of autumnal phenophases [1]. However, the modelling of interannual variation in LSP considering its potentially complicated relationship with climate in a multidimensional feature space (i.e. high number of multivariate weather drivers) might not be possible using traditional linear regression models [2]. In this sense, phenological modelling may benefit from machine learning techniques such as the Random Forest (RF) method [3], reducing uncertainties and bias [4]. RFs have the potential to identify and model the complex non-linear relationships between phenology and climate, being able to handle a large number of predictors and determine their importance in explaining phenology.

II. Data

Three sources of data were used for this research: i) Satellite sensor derived temporal composites of MERIS Terrestrial Chlorophyll Index (MTCI), ii) temperature and precipitation data from the European Climate Assessment and Data project (http: //www.ecad.eu) and iii) surface radiation daylight (DAL; w/m^2) data and surface incoming shortwave (SIS; w/m^2) radiation data from the Climate Monitoring Satellite Application Facilities (http://www.cmsaf.eu). We used weekly composites of MTCI data at 1 km spatial resolution from 2002 to 2012. This dataset was supplied by the European Space Agency and processed by Airbus Defence and Space. Daily temperature (mean, minimum and maximum) and daily precipitation data were derived from the European Climate Assessment & Dataset time-series (version 10.0) with spatial resolution of 0.25° ×0.25°, covering the period from 2002 to 2011. Both radiation datasets, DAL and SIS were derived from Meteosat satellite sensors at a spatial resolution of 0.05° x0.05° covering the same period as temperature and precipitation datasets.

III. METHODS

A. LSP computation

The time-series of MERIS MTCI data was used to estimate both the onset of greenness (OG) and end of senescence (EOS) from 2003 to 2011. The yearly values of OG and EOS were estimated for each image pixel of the study area using the methodology described in [5]. Z-score values during the study period were used as a proxy to measure interannual variation in the LSP parameters. The z-score values for a given year were defined as the difference from the multi-year mean, normalized by the standard deviation across years. To match the spatial resolution of the weather predictors, the LSP z-score values for each year were resampled to a spatial resolution of 0.25°×0.25° by calculating the median of all the LSP z-score values within this area after excluding the areas with fewer than 50 LSP estimates and the non-forest pixels according to the Globcover2005 and Globcover2009.

B. Computation of weather predictors

A suite of weather predictors were computed for each 0.25 ×0.25° grid cell associated with the occurrence of positive or negative z-score values in LSP (see Table 1).The different weather predictors were computed based on the 30 and 90 days previous to the day of the year (DOY) of the z-score values in OG and EOS The chilling requirements for spring modelling and freeze predictors were an exception, as the period for its computation starts 90 days prior to the OG. Relative differences between each predictor and its multi-year average for the same period were computed to capture the inter-annual variability in climate variables at the pixel level for every predictor and to facilitate the modelling of climate-driven variation in phenology (Table 1).

TABLE I. PREDICTORS USED IN THE MODELING.

OG anomalies	EOS anomalies
Averages (M)	
Maximum temperature (TX)**	Maximum temperature (TX)**
Minimum temperature (TN)**	Minimum temperature (TN)**
Average temperature (TG)**	Average temperature (TG)**
Precipitation (PP)**	Precipitation (PP)**
Surface incoming shortwave radiation (SIS)**	Surface incoming shortwave radiation (SIS)**
Surface radiation daylight (DAL)**	Surface radiation daylight (DAL)**
Cumulates (C)	
Growing Degree Days (0º C) (GDD)**	Growing Degree Days (0º C) (GDD)**
Growing Degree Days (5º C) (GDD)**	Growing Degree Days (5º C) (GDD)**
Chilling requirements (CHIL)*	Chilling requirements (CHIL)**
Precipitation (PP)**	Precipitation (PP)**
Surface incoming shortwave radiation (SIS)**	Surface incoming shortwave radiation (SIS)**
Surface radiation daylight (DAL)**	Surface radiation daylight (DAL)**
Date of specific events	
First freeze (FF)*	First freeze (FF)*
Last freeze (LF)*	OG z-score value (OGA)

* predicted over a period of 90 days. ** predicted over a period of the 30 and 90 days previous to the date of the z-score value.

C. Modelling interannual variation in LSP

The Random Forest (RF) method was applied to phenological modelling across very large areas and across multiple years simultaneously. The RF model was fitted to the relation between LSP interannual variation and numerous climate predictor variables. The locations with z-score in LSP greater than 1 (positive and negative) were selected to build a RF predictive model on OG and EOS. Z-score values of OG or EOS for each year were combined together with the different weather predictors. The z-score values in OG were assessed as an extra predictor to evaluate the legacy effect of an advanced or delayed spring in the modelling of EOS.

The values of these variables at the selected years and locations (spatiotemporal model) were combined into a set of input feature vectors (3900 feature vectors for the spring model and 3124 for autumn) as an input to the RF algorithm. These feature vectors were divided equally into two subsets, one for the training of the models (inbag) and one as an additional test to the one internally computed by RF (out of bag; oob) to evaluate performance. RF models composed of 2000 trees were grown using different subsets of predictors, varying the number of random predictors from 1 to 9. The Random Forest method within the package implemented in the R statistical software was used to build the different models.

D. Feature selection

A feature selection approach, based on the ability of the RF to assess the relative importance of the predictors, was used to identify the minimum number of drivers which can better explain spring or autumn interannual variation in phenology. In order to reduce the number of drivers the least important predictor was removed iteratively at different steps. Then, a 5-fold cross-validation was applied to obtain a stable estimate of the error of the model built after predictor deletions. Finally, the model with a better trade-off between number of predictors and error was chosen as the basis for interpreting the likely drivers of interannual variation in phenology.

IV. RESULTS AND DISCUSSION

Numerous models were built on the basis of different predictor combinations considering different temporal windows prior to the spring and autumn phenological events. Although, we did not carry out an exhaustive analysis of the optimum GDD parametrization, our results showed a systematic pattern in spring models, presenting slightly larger pseudo-R^2 for models which used 0^0 C as a threshold for the computation of GDD (rather than 5^0 C). Regarding, the length of the temporal windows for weather function computation, spring models using 30 and 90 days for the computation of averaged and cumulative functions were more accurate, whereas for autumn models with 90 day-averaged predictors outperformed the rest.

The main drivers of interannual variation in LSP were identified through the application of a feature selection procedure. Spring models were more accurate than autumn, with median relative error values of 10% to 27% (12 to 1 predictor), versus 26% to 60% of autumn (14 to 1 predictor). Fig. 1 shows the pseudo-R^2 of the models as well as the relative importance of each predictor. Spring models (explained a percentage of the variance up to 81% (Fig. 1a), whereas autumn explained up to 61% (Fig. 1b). Cook, Smith and Mann [6], using a modelled based on GDD only, explained 63% on the variance of onset date for mixed and boreal forest.

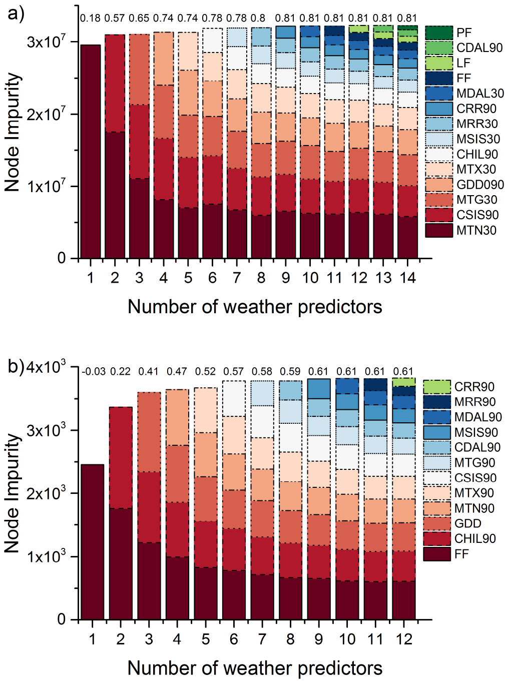


Fig. 1. Relative importance of each independent variable in predicting LSP interannual variation in Europe. Different models derived from the feature selection approach are represented in each column. Numbers given over each column represent the coefficient determination of each model. Plots at the top and bottom represent the spring (a) and autumn interannual variation in LSP (b), respectively.

Fig. 2 shows the relative error in the prediction of different models after removing the least important predictor. Regarding the relative importance of the drivers, the same ranking in importance was observed within the different models of each phenophase, which reflected the stability in the RF importance estimation, and a high reliability of the results. To interpret the main weather drivers of the interannual variation in phenology, simplified models with reduced number of predictors were selected for spring and autumn, respectively. The spring model was composed of 6 predictors (pseudo-R^2=0.77 and median relative error of 10%) and the autumn model of 5 predictors (pseudo-R^2=0.59 and median relative error of 28%). Our results suggest that interannual variation in the onset on greenness (LSP) of temperate forest species are driven mainly by the daily temperature of the 30 days prior to onset (but not necessarily the GDD), with the most important driver being the minimum temperature. Photoperiod was also important, the most accurate empirical prediction was obtained by a combined temperature-radiation forcing, integrating the SIS of the previous 90 days. For senescence, temperature was suggested to be more important than photoperiod in controlling the senescence process, with the most important drivers being the date of the first freeze and the accumulation of chilling temperatures. However, we did not observe a legacy effect of a much earlier or later spring onset on the date of senescence.

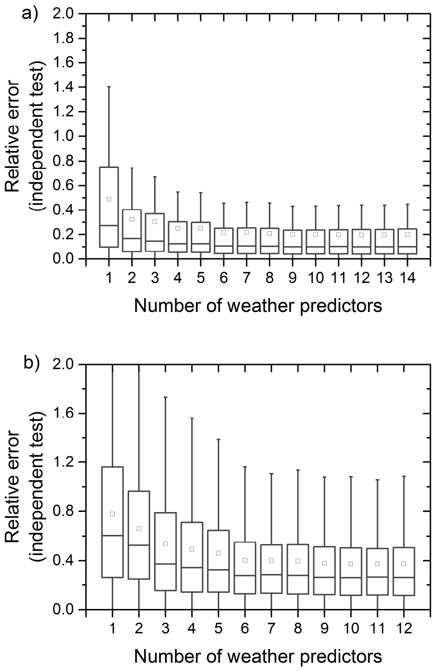


Fig. 2. Relative error of the models fitted as a result of the feature selection approach. Median (interior horizontal line), mean (interior square), 1% and 99% quantiles (edge of boxes), range (extremes).

Acknowledgment

The author is a Juan de la Cierva grant holder. The author is grateful for the financial support given by the Spanish MINECO (project BIA2013-43462-P). We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu). Surface radiation data were obtained from EUMETSAT's Satellite Application Facility on Climate Monitoring (CM SAF).

References

[1] J. T. Morisette, A. D. Richardson, A. K. Knapp *et al.*, "Tracking the rhythm of the seasons in the face of global change: phenological research in the 21st century," *Front. Ecol. Environ.*, vol. 7, no. 5, pp. 253-260, 2009/06/01, 2008.

[2] K. M. de Beurs, and G. M. Henebry, "Land surface phenology and temperature variation in the International Geosphere-Biosphere Program high-latitude transects," *Glob. Change Biol.*, vol. 11, no. 5, pp. 779-790, 2005.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.

[4] M. F. Zhao, C. H. Peng, W. H. Xiang *et al.*, "Plant phenological modeling and its application in global climate change research: overview and future challenges," *Environ. Rev.*, vol. 21, no. 1, pp. 1-14, 2013.

[5] J. Dash, C. Jeganathan, and P. M. Atkinson, "The use of MERIS Terrestrial Chlorophyll Index to study spatio-temporal variation in vegetation phenology over India," *Remote Sens. Environ.*, vol. 114, no. 7, pp. 1388-1402, 2010.

[6] B. I. Cook, T. M. Smith, and M. E. Mann, "The North Atlantic Oscillation and regional phenology prediction over Europe," *Glob. Change Biol.*, vol. 11, no. 6, pp. 919-926, 2005.

Statistical modeling of phenological phases in Poland based on coupling satellite derived products and gridded meteorological data

[extended abstract]

Bartosz Czernecki
Department of Climatology
Adam Mickiewicz University in Poznań
61-680 Poznań, Poland
nwp@amu.edu.pl

Katarzyna Jabłońska
Department of Air Pollution Modelling,
Institute of Meteorology and Water
Management - National Research
Institute,
01-673 Warsaw, Poland
katarzyna.jablonska@imgw.pl

Jakub Nowosad
Institute of Geoecology and
Geoinformation
Adam Mickiewicz University in Poznań
61-680 Poznań, Poland
nowosad@amu.edu.pl

Abstract— **Changes in timing of phenological phases are important proxies in contemporary climate research. Phenological data could be also used in the reconstruction of long-time temperature time-series. The aim of the study was to create and evaluate different statistical models for reconstructing and predicting the selected phenological phase. Quality-controlled dataset of Syringa vulgaris and Aesculus hippocastanum phenophases from the years 2007-2014 was used. For each plant species models were build using the most commonly applied regression-based and random forest methods. Three types of data sources were applied as predictors: (i) satellite derived products, (ii) preprocessed gridded meteorological data, and (iii) spatial features (longitude, latitude, altitude) of the monitoring sites. The obtained results showed potential for coupling meteorological derived indices with remote sensing products in terms of phenological (late spring) modelling. It was also shown that choosing a specific set of predictors and applying a robust preprocessing procedures is more affecting final results than applying a statistical model.**

Keywords— climate change; plant phenology; phenology modeling; phenological reconstruction; statistical modeling; machine learning

I. Introduction

Phenology of the plants is mainly influenced by photoperiod and temperature. Previous studies showed that global warming determine the advance of phenological events [1-5]. Therefore, changes in timing of phenological phases are important proxies in contemporary climate research. Additionally, phenological data could be used in the reconstruction of long-time temperature time-series [6-9].

In Poland, the oldest discovered local records of phenological observations were found in the 15th and 16th centuries [10]. Disorganized modern phenological observations began in the late 19th in the Polish territories. Currently, phenological observations in Poland are carried out independently by different institutions, such as botanical gardens, agricultural and forestry departments. The phenological data in each institution were collected using different methodologies and data formats [11]. Additionally, the Polish Institute of Meteorology and Water Management established nationwide phenological monitoring after World War II. However, the network was abandoned between 1993 and 2005 and newly established network in 2006 was moved into a new locations. Keeping the aforementioned in mind the main aim of the study was to create and evaluate different statistical models for reconstructing and predicting day of year of selected phenological phases occurrence. Authors also decided to evaluate possibilities of using a wide-range of statistical modeling techniques to create synthetic archive dataset using only free of charge data remote sensing and meteorological data as predictors. Therefore it was also possible to (1) distinguish the amount of information provided by both sources of data, (2) define whether they are unrelated and contain possible sources of not overlapping information, (3) and thus may (or may not) robustly contribute in phenological research, especially in terms of phenological modeling.

II. Materials and methods

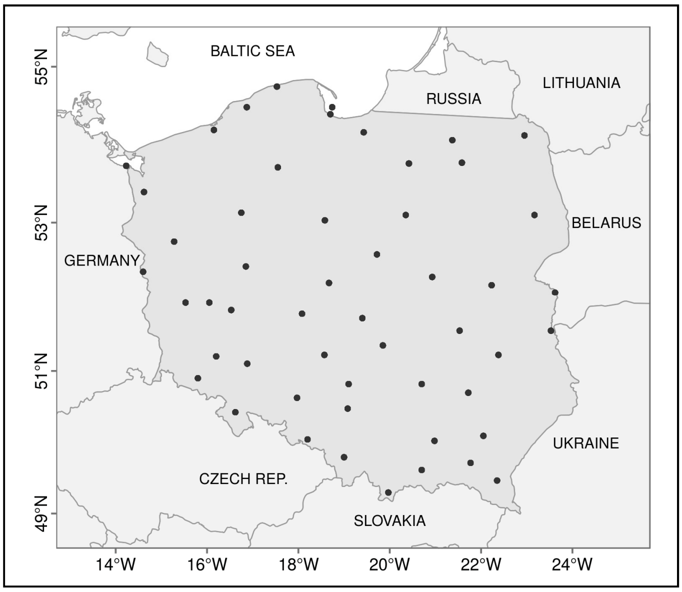


Fig. 1. The location of the sites used for the study

A. Phenological data

Study period covers the years 2007-2014 and contains only quality-controlled dataset of *Syringa vulgaris* and *Aesculus hippocastanum* flowering dates on 52 stations in Poland (Fig. 1). Phenological data used in this study originates from the observational network run by the Institute of Meteorology and Water Management - National Research Institute (IMGW-PIB). The phenological network follow the BBCH (abbr. from German: "Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie") methodology, which was akin to most of European countries with similar growth stages of plant species.

B. Predictor variables

Three types of data sources were used as predictors: (i) satellite derived products, (ii) preprocessed gridded meteorological data, and (iii) spatial features (longitude, latitude, altitude) of the monitoring sites. Moderate-Resolution Imaging Spectroradiometer (MODIS) level-3 vegetation products were used for detecting onset dates of particular phenophases. Following indices were used: Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Leaf Area Index (LAI), and Fraction of Photosynthetically Active Radiation (fPAR). Additionally, Interactive Multisensor Snow and Ice Mapping System (IMS) products were chosen to detect occurrence of snow cover. Due to highly noisy data, authors decided to take into account pixel reliability information. Besides satellite derived products (NDVI, EVI, FPAR, LAI, Snow cover), a wide group of observational data and agrometeorological indices derived from the European Climate Assessment & Dataset (ECA&D) were used as a potential predictors, such as, cumulative growing degree days (GDD), cumulative growing precipitation days (GPD), average monthly temperatures for each month over the previous year for each site.

C. Model development

A few commonly applied statistical methods, including (multiple) linear regression (lm), linear regression with stepwise selection (lmAIC), generalized linear model (glm), generalized linear model with stepwise feature selection (glmAIC) and random forest (RF), were tested and evaluated against the onset dates of phenophases. All the calculations were carried out using R [12] and R packages [13-16].

This study split the potential predictors into four sub-groups that might be applied for the needs of statistical modeling: (i) consisting only of meteorological-derived variables and locations' features, (ii) MODIS-derived predictors, (iii) all available variables pre-processed with the use of Boruta algorithm that finds all-relevant features (14), (iv) and all available variables without any pre-selection.

Repeated *k-fold* cross validation was used to avoid overfitting and to estimate the accuracy of the models. This method randomly divides the data into *k segments*. The model is trained on *k-1 segments*, and the held out segment is used to evaluate the model. The overall performance is obtained by averaging the *k estimates* of performance [18]. The whole procedure was repeated twice.

Model's performances were characterized using the coefficient of determination (R2), root-mean-square error (RMSE), mean absolute error (MAE). An R2 value is the squared correlation coefficient between the observed and predicted values. RMSE is the difference between predicted values and observed values. MAE is the mean absolute error. Additionally, standard deviations (SD) of the model's performance statistics were calculated.

III. Results

Created statistical models showed substantial impact of selected predictors on final results. The chosen set of predictors showed very similar results in terms of calculated models' performance statistics in every of tested regression based algorithms (Fig. 2). The impact of selected predictors was smaller on random forest models than on regression models. However, the obtained pattern was similar in every of analysed models, i.e. the best fitted models were preprocessed using Boruta algorithm (Fig. 2). The AIC stepwise screening applied for linear models hardly influences the obtained results and in authors' opinion this solution do not redress computational time that is required while applying this procedure.

The models based on meteorological characteristics were better fitted to observational time-series than remote sensing-based models. Even though, conjunction of both data sources showed high potential in improving model's accuracy (arround 1 day) in predicting late spring phenology phases.

In general, the constructed models based only on coupling phenophases with meteorological indices accounted for about 80% of variance in *Syringa vulgaris* and *Aesculus hippocastanum* flowering dates, while in case of applying remote sensing data and preprocessing with Boruta algorithm

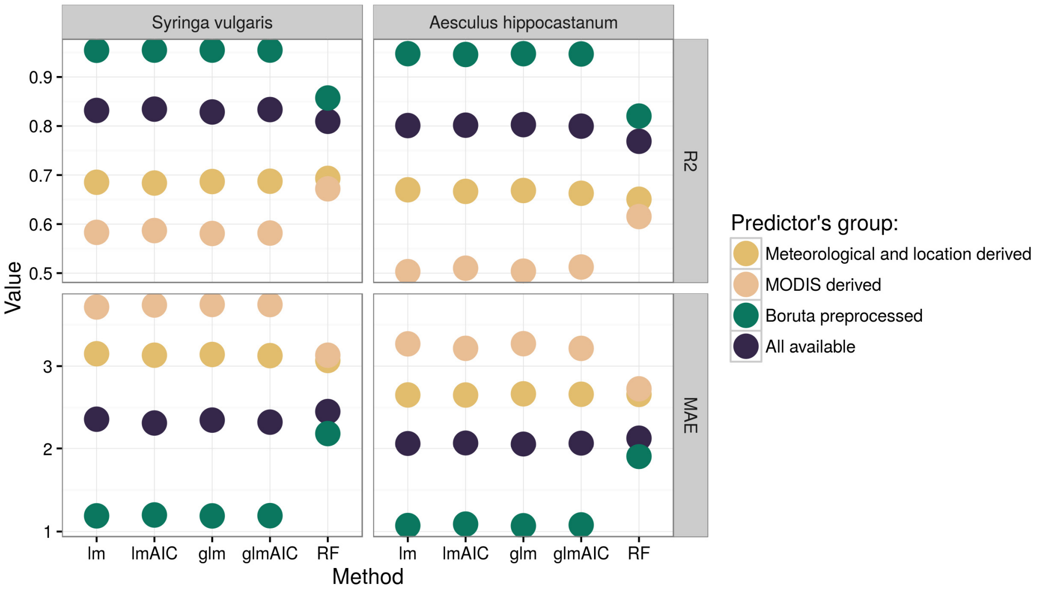


Fig. 2. A few commonly applied statistical methods, including (multiple) linear regression (lm), linear regression with The location of the sites used for the study

this value increased to over 95%. This suggests clear needs of applying preprocessing procedures while using numerous set of predictors. Additionally, expanded validation strategy to avoid model overfitting should be conducted.

IV. Conclusions and Future Work

The obtained results show good potential of using statistical models in filling the temporal and spatial gaps in data, as well as, for forecasting selected phenological phases. However, there are some clear limitations of applying modern satellite observation in plant phenology modeling. For instance, calculated contribution of each variable showed small influence of satellite derived products. These data may contain noisy information and thus were omitted while applying preprocessing procedure. Therefore, most of the created phenology models are primarily based on agrometeorological indices with only slightly improvements while using satellite derived products.

Acknowledgment

This work has been supported by the Polish Ministry of Science and Higher Education under grant no. DEC-2012/05/N/ST10/03703. We would like to acknowledge also the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu).

References

[1] Bradley NL, Leopold CA, Ross J, Huffaker W. Phenological changes reflect climate change in Wisconsin. Proc Natl Acad Sci U S A. 1999;96:9701–4.

[2] Root T, Price J, Hall K, Schneider S. Fingerprints of global warming on wild animals and plants. Nature. 2003;421(6918):57–60.

[3] Menzel A, Sparks TH, Estrella N, Koch E, Aasa A, Ahas R, et al. European phenological response to climate change matches the warming pattern. Glob Chang Biol. 2006;12(10):1969–76.

[4] Parmesan CN. Ecological and evolutionary responses to recent climate change. Annu Rev Ecol Evol Syst. 2006;37:636–7.

[5] Cleland EE, Chuine I, Menzel A, Mooney HA, Schwartz MD. Shifting plant phenology in response to global change. Trends Ecol Evol. 2007;22(7):357–65.

[6] Aono Y, Kazui K. Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. Int J Climatol. 2008;28(7):905–14.

[7] Schleip C, Rutishauser T, Luterbacher J, Menzel A. Time series modeling and central European temperature impact assessment of phenological records over the last 250 years. J Geophys Res Biogeosciences. 2008;113(4):1–13.

[8] Bradley RS. Paleoclimatology: Reconstructing Climates of the Quaternary. Elsevier Science; 2013.

[9] Zheng J, Hua Z, Liu Y, Hao Z. Temperature changes derived from phenological and natural evidence in South Central China from 1850 to 2008. Clim Past. 2015;11(11):1553–61.

[10] Obrębska-Starklowa B. About the phytophenological research in Galicia in the nineteenth century on the the background of the development of phenology network in Europe, Geophys Rev.1993;3-4:289-295

[11] Jabłońska K, Rapiejko P. Using the results of a nationwide phenological network to examine the impact of changes in phenology of plant species on the concentration of plant pollen in the air. Acta Agrobot. 2010;63(2):69–74.

[12] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016.

[13] Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth. New York: Springer; 2002.

[14] Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. J Stat Softw. 2010;36(11):1–13.

[15] Wright MN. ranger: A Fast Implementation of Random Forests. R package version 0.3.0. 2015.

[16] Kuhn M. caret: Classification and Regression Training. R package version 6.0-64. 2016.

[17] Keatley MR, Hudson IL. Phenological Research. Hudson IL, Keatley MR, editors. Phenological research methods for environmental and climate change analysis. Dordrecht: Springer Netherlands; 2010. 525 p.

[18] Kuhn M, Johnson K. Applied predictive modeling. Springer New York; 2013.

Mapping of the East Vardar Zone ophiolite using remote sensing and geophysical data

[extended abstract]

Dragana Đurić
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia
dragana.petrovic@rgf.bg.ac.rs

Abstract— **The central part of the Balkan Peninsula has very complex tectonic settings. Tethys was closed during the upper Mesozoic, and ophiolites in the west and ophiolites of Vardar zone represent the remains of the former ocean. The easternmost part of Vardar zone is different from all other ophiolite on the Balkan Peninsula, and the position of this ophiolite is still the subject of discussion among researchers. The main goal of this study was to determine the position of the East Vardar Zone ophiolite, based on remote sensing and geophysical data in Serbia, Macedonia and Greece. Boundaries of EVZ were determined, as well as a precise position of ophiolite exposures. The western border of the East Vardar Zone ophiolite was clearly expressed, while the eastern border was uncertainly established primarily due to a very heterogeneous Serbo-macedonian Massif. Based on the analysis of transformed data it was established that ophiolites are elongated in the direction of NNW-SSE and that represent a unique body.**

Keywords—remote sensing; geophysical potential field; ophiolite; East Vardar Zone

I. Introduction

Vardar Zone (VZ), as one of several, subparallel and NNW–SSE stretching belts that left behind Mesozoic convergence and subsequent collision, is settled between two continental units, Drina-Ivanjica Unit on the east and Serbo-macedonian Massif (SMM) on the west. VZ is generally divided into three subzones: West Vardar Zone (WVZ), Central Vardar Zone (CVZ) and East Vardar Zone (EVZ) [1]. East Vardar zone is located in central part of Serbia. On the north it extends to Romanian Apuseni Mountains up to Transilvanian depression. On the south VZ extends to Macedonia and Greece, named Peonias Zone. Further south VZ extends to Izmir-Ankara Zone in Turkey. This complicated geotectonic relationship was one of the main reasons of disagreement related to position of EVZ. Generally EVZ represents small, narrow belt on the easternmost part of VZ. Present tectonic contact with adjacent Units is characterized by steep normal faults, reactivated in Miocene [2]. There were many disagreements between researchers related to spatial position of EVZ [3] [4] [5]. Also, there were no precise locations of boundaries EVZ with adjacent units. The aim of this study was to use satellite images and geophysical data (potential field data) with different scale in order to determine boundaries of EVZ with adjacent units. By defining and interpreting the main structural features of these ophiolite complexes, an attempt was made to refine and improves earlier views about the position of EVZ. The obtained results suggested that EVZ represent single belt that is not interrupted on or near the surface.

Main part of East Vardar zone ophiolites is built up of gabbro-dolerites, dolerites, dolerite dykes and rare basaltic pillow lavas. Occasionally, at the margins of EVZ there are small and isolated mainly serpentinised harzburgites. All ophiolitic rocks are associated with intermediate and acid calcalkaline granitic rocks [6]. In Serbia and Macedonia, ophiolites are overlain by Tithonian limestone [7]. The biggest exposure of EVZ are located in Serbia (near Kuršumlija and Kragujevac) as bodies elongated 20 km and wide few km oriented NNW-SSE. In Macedonia EVZ ophiolites are represent by NW-SE elongated bodies, long 50 km and wide 25 km. This ophiolite include ophiolite complex of Demir Kapija and Gevgelija. Further south, ophiolite complex is present in Greece as body long 20 km intruded by Fanos (Furka) granite and separated into two parts.

II. Methodology

A. Preprocessing of satellite images

Satellite images used in this paper were Landsat 7 ETM+ satellite images which were placed in the WRS (World Reference System) at positions 183-187 / 28-32. Bands of this mission use wavelengths from 0.45 to 12.50 μm. Preprocessing eliminate the atmospheric effects on transmission of electromagnetic energy through the atmosphere from Earth's surface to the sensor, as well as noise. Panchromatic sharpening was applied by Brovey transformation. New resolution of images was 15m.

B. Processing of satellite images

Processing of satellite images includes transformation of each single band in order to obtain basis that allows the best possible analysis and interpretation. Transformation includes change of wavelength of pixels.

The Normalized difference vegetation index (NDVI) uses the visible and near-infrared bands of the electromagnetic spectrum to analyze vegetation in investigated area. Morphological filter which is used in the paper is "erosion". In this way, the regional linear structures are featured [8] [9]. Mosaic of 13 processed images was created.

Principal component analysis (PCA) includes spectral values of each band. The usefulness of the method is correlation within a set of variables and reduction of dimensionality of space, in order to use as small a number of variables to explain more variables. PCA1 channel was replaced by a panchromatic channel, which was adjusted to fit PCA1, in order to not dissipate the spectral information. Then, the reverse operation was performed. Multispectral data were automatically resampled and their resolution was raised using the "nearest neighbor". The process of automatic classification was carried out on the images, which were previously passed through the PCA, in order to separate several classes of interest. EVZ ophiolite was allocated by automatic classification as class1, as most marked class. As a separate unit, in this way, area under water and space with increased vegetation was allocated (class 4). Sedimentary covers were allocated as class 3. Class 2 comprise of granite rocks, while class 5 cover all other rocks that appear in the study area.

C. *Qualitative analysis of geophysical data*

Qualitative analysis of geophysical data includes gravity and geomagnetic data (satellite, areal and terrestrial), application of mathematical transformation on these data in order to define rupture assembly and body position / causes anomalies in the horizontal plane.

Standard geostatistical methods were used for determination of spatial distribution of the rock properties present on the ground. Gravity data were obtained by detailed gravimetric surveys in Serbia and Macedonia realized between 1952 and 1984 by the Geophysical Institute [10]. Gravimetric data, which cover the area of northern Greece, included digitized map of Bouguer anomaly [11] and they were used for the analysis of gravity data in the Northern Greece. Geomagnetic data include data of surveys in Serbia and Macedonia of Z component. These data were reduced to the epoch 1960.0. The geomagnetic data, also, comprised airborne data which were obtained by digitization of total field anomaly map of the Earth's field, whereas the grid was created with 500 m spacing points. Normal magnetic field was calculated according to the formula of Geomagnetic Institute (Grocka) for the epoch 1960.0. Analysis of global data from the CHAMP satellite comprise positioning a zone of interest, i.e. positioning the dominant anomaly values at the regional level. These data served as the basis for local research.

Above the gravimetric data, respectively, over the Bouguer anomaly map (calculated with 2.67 t/m3), following mathematical transformation were used: Hanning filter, vertical derivation of gravitational acceleration and normalized standard deviation. Above geomagnetic data we applied the following methods of mathematical transformation: reduction to pol and tilt angle (TDR).

III. Results

A. *PCA and Automatic classification*

Automatic classification, which is applied over the PCA channels, has given good results at this form of mapping. Ophiolites are clearly visible and separated from other units (Fig. 1). The results are in agreement with Basic Geological Map, but boundaries are more precisely located (Fig 1).

B. *Results of geophysical data analysis*

Medium density of gravimetric measurement is 1.4 point per km^2. The anomalies are calculated with a mean density of 2.67 t / m^3. Bouguer anomaly values range from -70 to 30 mg (SI system: 1 mgal = 10^{-5} m / s^2). Anomalies are oriented in a NNW-SSE direction and indicate the presence of large linear structures, which generally follow the provision of EVZ and restrict it from both sides. Abrupt changes of the anomaly values suggest that the contact between adjacent units is vertical or sub vertical. Hanning filter was applied on Bouguer anomaly map. On this map anomalies in the area of IVZ are emphasized and boundary between EVZ and SMM in the east and the western boundary EVZ with Kopaonik unit (KU) are better defined. EVZ is characterized by high values of anomalies, while the neighboring tectonic units are clearly separated by abrupt changes and low values. In the area of Greece the digitized map of Bouguer anomaly [12] was used. This map was digitized with a grid of 250 m. In this area, values of Bouguer anomaly range from -22 to 55 mgal. Dominant anomalies are elongated in the direction of N-S and NNW-SSE. There are three characteristic regions: the western part with anomaly range of -22 to 20 mgal, the central part, which is NW-SE elongated and covers the southeastern part of the research area (EVZ, i.e. ophiolites Evzoni and Skra) with a range of anomalies from 25 to 55 mgal and the north-eastern part (SMM or the Circum Rhodope) with a range of anomalies from 0 to 25 mgal. On geomagnetic map there are three areas elongated in the direction of NW-SE with a high positive values of anomalies. The first area is positioned in the northwestern part of the study area and can be linked to the West Vardar ophiolites. Dominant anomalies range from -10 to 450 nT, while the range of values of the anomaly zone in the southern part of Macedonia is from 100 to 450 nT. The reason for this different values of anlomalies was in using a different type of data (areal and terrestrial measurments in northern and southern part). The second area allocated in the central part of the study area, with a range of anomalies from -30 to 350 nT, corresponds to the IVZ (Fig 1). Further south, this area is characterized by high values of anomalies and represents the impact of diabase. This anomaly could be associated with granodiorite on the eastern border [5]. Abrupt changes of anomaly values indicate tectonic boundaries. The third zone with high value anomalies occupies the eastern part and correlates with SMM. In this area, the anomalies have a value of -35 to 50 nT. On the territory of Macedonia and Greece anomalies appear in the range of 100 to 350 nT indicating the presence of granitoid rocks. On TDR map position of the causes of anomalies in the area of EVZ was determined, which agrees with the position of which can be seen on maps Bouguer anomalies. On the TDR map it is evident that the boundaries (Fig.1) between adjacent units are clear and sharp. The position of the boundaries (Fig. 1/black polyline) of causes of anomaly corresponds to the provision of isoline with the value zero [13].

Interpreted ruptures suggest that the entire area were controlled by steep normal faults NNW-SSE to S-J oriented [3] [14].

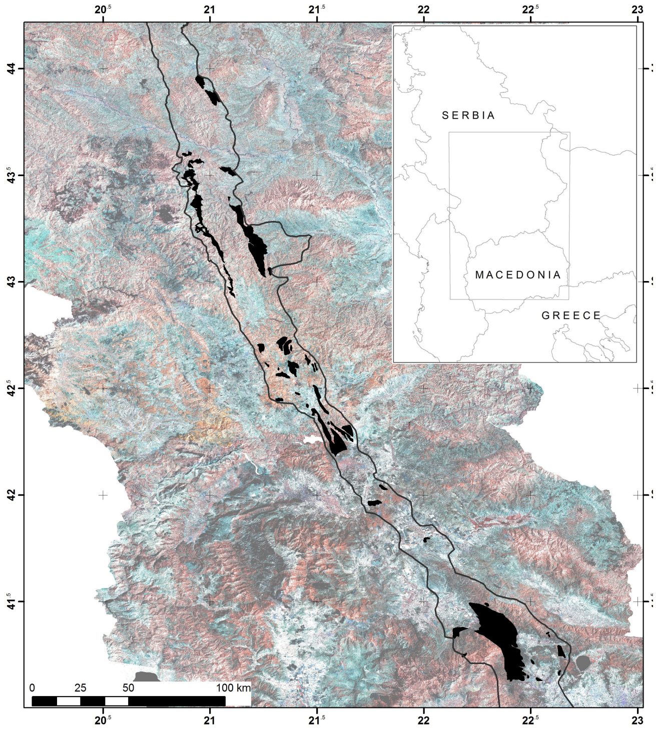


Fig.1. Landsat satellite imagery mosaic (bands: 4, 5, 7, 8); position of EVZ ophiolite determined by PCA and automatic classification (black polygons); boundary of EVZ determined by geophysical data (black polyline)

IV. Conclusion

Based on the results it is concluded that the EVZ has clear and sharp boundary with westward units (Kopaonik unit and WVZ) and this boundary is interpreted as sub vertical tectonic boundary. Based on remote sensing and geophysical data can be concluded that the boundary between IVZ and SMM is diffuse. It cannot be clearly distinguished as a unique lineament either is assigned an exact position. Overall, the horizontal distribution of ophiolite IVZ is bigger in the south (southern Macedonia and Greece) than in the northern part of the investigated area (Serbia and northern Macedonia). Mapped EVZ ophiolite covers area of about 920 km^2 (black polygons on Fig.1). Area of the central part of EVZ determined by remote sensing and geophysical data was estimated to about 9100 km^2 in its central part (black polyline on Fig. 1). Determination of boundaries between EVZ and adjacent unit could be the starting point for all further research in this area.

Acknowledgment

This study was supported by Serbian Ministry of Education, Science and Technological Development, project no. 176016. D. Đ. thanks to Society of geophysical exploration.

References

[1] Robertson, A. H. F., Trivić, B., Djerić, N., Bucur, I., 2013. Tectonic development of the Vardar ocean and its margins: Evidence from the Republic of Macedonia and Greek Macedonia. Tectonophysics 595, 25-54, doi: 10.1016/j.tecto.2012.07.022.

[2] Brown, S. A. M. and Robertson, A. H. F., 2004. Evidence for Neotethys rooted within theVardar suture zone from the Voras Massif, northernmost Greece.Tectonophysics 381, 143-173.

[3] Karamata, S., 2006. The geological development of the Balkan Peninsula related to the approach, collision and compresion of Gondwanan and Euroasian units. in: Eds Robertson, A. H. F. & Mountrakis, D.: Tectonic Development of the Eastern Mediterranean Region. Gseological Society, London, Special Publication, 260, 155-178.

[4] Schmid, M.S., Bernoulli, D., Fügenschuh, B., Matenco, L., Schefer, S., Schuster, R., Tischler, M., Ustaszewski, K., 2008. The Alps-Carpathians-Dinarides-connection: a correlation of tectonic units. Swiss Journal of Geosciences, doi: 10.1007/s00015-008 1247.

[5] Petrović, D., Cvetkov, V., Vasiljević, I., Cvetković, V., 2015: A new geophysical model of the Serbian part of the East Vardar ophiolite: Implications for its geodynamic evolution, Journal of Geodynamics, v.90, p. 1-13; DOI:10.1016/j.jog.2015.07.003.

[6] Resimić-Šarić, K., Cvetković, V., Balogh, K., Koroneos, A., 2006. *Main characteristics of ophiolitic complexes within the eastern branch of the Vardar zone composite terrane in Serbia*. International Symposium "Mesozoic ophiolite belts of northern part of the Balkan Penninsula". Belgrade-Banja Luka, 112-115.

[7] Dimitrijević, M. D., 1997. *Geology of Yugoslavia*, Belgrade: Barex, Special Publications.

[8] Kresić, N., 1995. Remote Sensing of Tectonic Fabric Controlling Goundwater Flow in Dinaric Karst. Rem. Sens. Environ., 53(2): 85-90.

[9] Mohanty, C., Baral, D. J. and Malik, J. N, 2004. *Use of Satellite Data for Tectonic Interpretation, NW Himalaya*, Journal of Indian Society of Remote Sensing, Vol. 32, No. 3, pp 241-247.

[10] Bilibajkić, P., Mladenović, M., Mujagić, S., Rimac, I., 1979. *Explanation for the Gravity Map of SFR Yugoslavia – Bouguer Anomalies – 1:500 000,* Federal Geol. Inst., Belgrade.

[11] Savvaidis, A. S., Tsokas, G. N., Papazachos, C. B., Kondopoulou, D., 2000. A geophysical study of the ophiolite complex and sedimentary basins in the northwest part of the Chalkidiki Peninsula (N. Greece), Surveys in Geophysics, 21(5):567-595. DOI: 10.1023/A:1006723025183.

[12] Tassis, G. A., Grigoriadis, V. N., Tziavos, I. N., Tsokas, G. N., Papazach os, C. B., Vasiljević, I., 2013. *A new Bouguer gravity anomaly field for the Adriatic Sea and its application for the study of the crustal and upper mantle structure*, Journal of Geodynamics, Volume 66, p. 38-52.

[13] Cooper, G. R. J., Cowan, D. R. D., 2008. Edge enhancement of potential-field data using normalized statistics. Geophysics 73, 3, H1-H4, http://dx.doi.org/10.1190/1.2837309.

[14] Robertson, A.H.F., Karamata, S., Šarić, K., 2009. Overview of ophiolites and related units in the Late Palaeozoic–Early Cenozoic magmatic and tectonic development of Tethys in the northern part of the Balkan region. Lithos, ISSN: 0024-4937.

Sensitivity of vegetation indices derived from Sentinel-2 data to change in biophysical characteristics

[extended abstract]

Dragutin Protić, Stefan Milutinović, Ognjen Antonijević, Aleksandar Sekulić, Milan Kilibarda

Department of geodesy and geoinformatics
Faculty of civil engineering, University of Belgrade
Belgrade, Serbia
protic@grf.bg.ac.rs

Abstract — **Optimal spectral, spatial and temporal characteristics of Sentinel-2 data makes it attractive for applications in agriculture and monitoring of crop status and health in particular. Crop growth is increasing of certain biophysical characteristics of crops such as biomass and leaf area index (LAI). A number of vegetation indices (VIs) have been designed and used for monitoring biophysical parameters of vegetation, however a common solution does not exist due to the fact that the sensitivity of VIs to biophysical parameters depends on factors like crop architecture and plant growth stage. In this study, several VIs that are documented in the literature to be good estimators of biophysical parameters of crops were calculated for 2 winter crops: wheat and barley. The VI raster layers were generated on days when Sentinel-2 data is available during the crop growth phase (January - June), for a number of parcels where in situ data is available. Response of the VIs to temporal progression of the crops is statistically analyzed to gain a deeper understanding on usability of various Sentinel-2 based VIs in monitoring of winter crops growth. The conclusions contribute to developing methodology for generating crop health related information.**

Keywords — Sentinel-2, vegetation indices, biophysical paramteres, crops

I. Introduction

The new Copernicus Sentinel-2 mission with two planned satellites is going to provide valuable information for monitoring of crops. Vegetation Indices (VIs) are empirical parameters derived from spectral bands commonly used to describe health status and development of vegetation. VIs are often "translated" into biophysical parameters using transfer functions to ensure information on vegetation conditions with physical meaning.

A number of VIs has been designed based on the spectral characteristics of green vegetation which exhibits low reflection in the red portion of the spectrum and strong reflection in the near-infrared range (NIR) [1]. VIs have different sensitivity to crop parameters in various phonological stages [3]. Recent studies showed that red-edge spectral range is highly significant in estimating a number of biophysical parameters [2].

For this study, multitemporal Sentinel-2 data from January to June was obtained for several agricultural fields under two arable crops: wheat and barley.

This research is conducted within the scope of the APOLLO project – Advisory platform for small farms based on Earth Observation (H2020).

II. Materials and Methods

A. Satellite data

Sentinel-2 data for tile 34TDQ was collected from ESA Sentinels Sci Hub for 6 dates: 1-1-2016, 18-3-2016, 7-4-2016, 27-4-2016, 27-5-2016 and 30-5-2016. The Level-1c product was corrected for atmospheric and topographic effects using sen2cor processor to generate surface reflectance Level-2A product (Bottom of Atmoshpere -BOA).

B. In situ data

Field data collection campaign was organized 28th and 29th of May to coincide with the Sentinel-2 data acquisition that occurred 27th and 30th of May. There were 29 sampling units of which 11 under winter wheat and 5 under barley. The data collected include: crop phenology, soil type, soil wetness estimation, canopy height, leaf angle estimation. The data were estimated for an average plant on the sampling unit. In addition, LAI and biomass (after process of drying) are measured for the selected average plants and extrapolated (knowing exact number of plants per m^2) to 1 m^2 area.

C. *Experimental work*

To study the behavior of Sentinel-2 data and the derived VIs during crop development period or winter wheat and barley, several experiments have been conducted.

First, average BOA spectral signatures in VNIR region of wheat and barley crops on the selected fields were generated to track changes in spectral reflectance on crop level in different Sentinel-2 bands related to changes in biophysical properties (crop development).

Second, sensitivity of several VIs, namely NDVI, Clred-edge, IRECI and NIR/R to crop canopy development is estimated by comparison of normalized VI values derived from average BOA data on the selected fields.

Third, influence of soil background on each of the above mentioned VIs was described comparing simulated values of VIs calculated for different soil-vegetation ratios for dark and light soil respectively.

Fourth, best fitting linear regression function was found to serve as a transfer function between LAI and VIs.

III. Results and Discussion

A. *Temporal changes of spatial signaturesof crops*

Spectral signatures of winter wheat and barley across VNIR Sentinel-2 bands are presented in Fig. 1 and Fig. 2 respectively. In both cases the differences in reflectance in Bands 7, 8 and 8A are very low. It is also clear that spectral signatures of wheat and barley is almost identical until 7th of April image when reflectance in NIR-Red Egde bands for barley are in their peak and start to decline, while for wheat the reflectance continue rising until 27th of April.

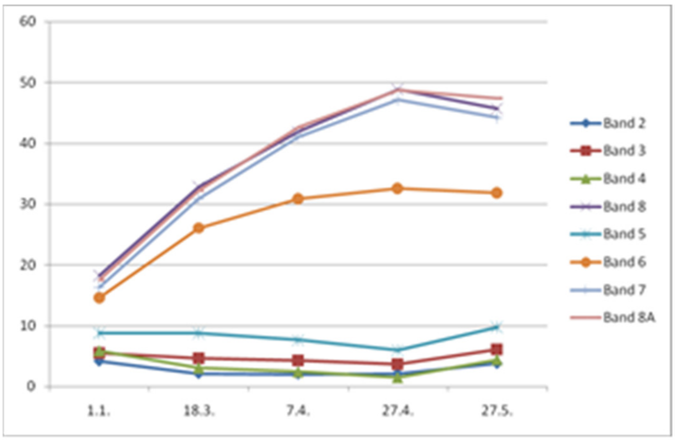


Fig. 1. Temporal changes of BOA reflectances in VNIR S-2 bands for winter wheat

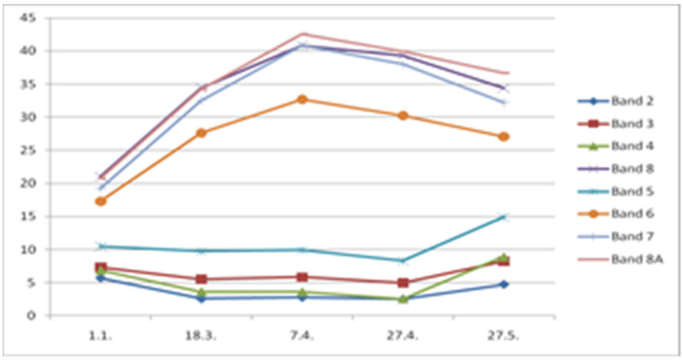


Fig. 2. Temporal changes of BOA reflectances in VNIR S-2 bands for barley

B. *Sensitivity of VIs to crop canopy development*

Fig. 3 and Fig. 4 present temporal changes in four VIs (normalized to the range 0-1 in order to be comparable) during growth period for winter wheat and barley respectively. In both cases, the most popular vegetation index NDVI becomes saturated already on 18th of March. The other VIs behave similarly in the case of wheat, while in the case of barley, NIR/R shows the most uniform sensitivity during the crop development period.

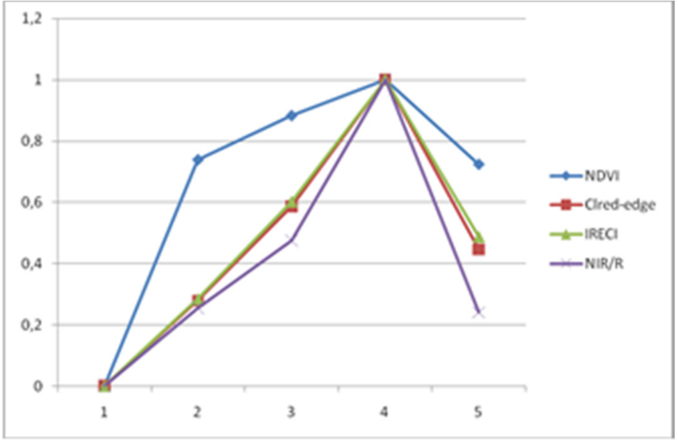


Fig. 3. Temporal changes of normalized VIs for winter wheat

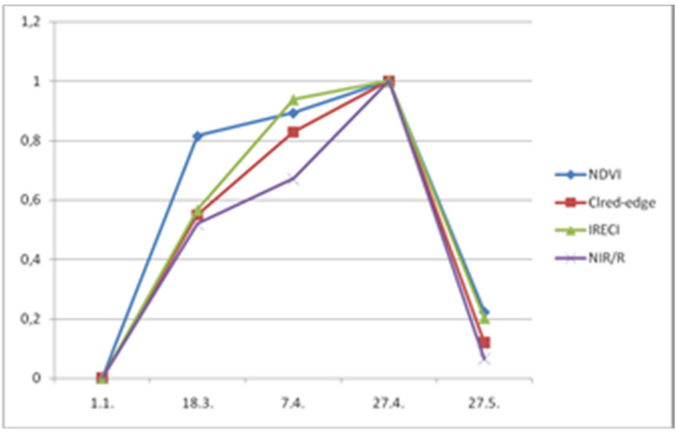


Fig. 4. Temporal changes of normalized VIs for barley

C. *Influence of soil background*

Differences between VIs of crops with dark (wet) and light (dry) soil background for various crop-soil ratios are presented for four VIs in Fig. 5-8. The smallest influence is observed for IRECI index.

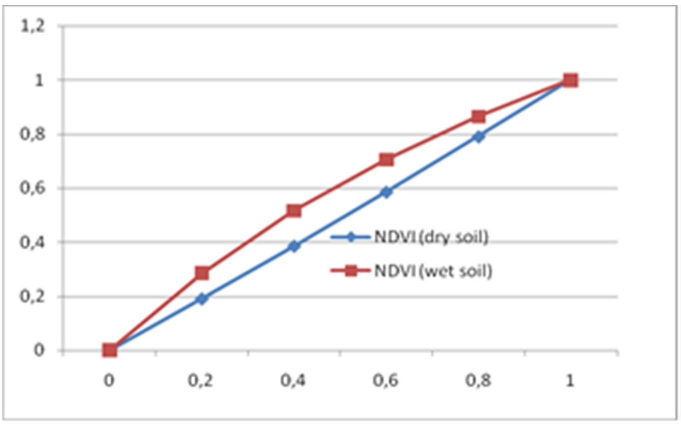


Fig. 5. NDVI values from S-2 bands for different crop-soil ratios for dry and wet soils

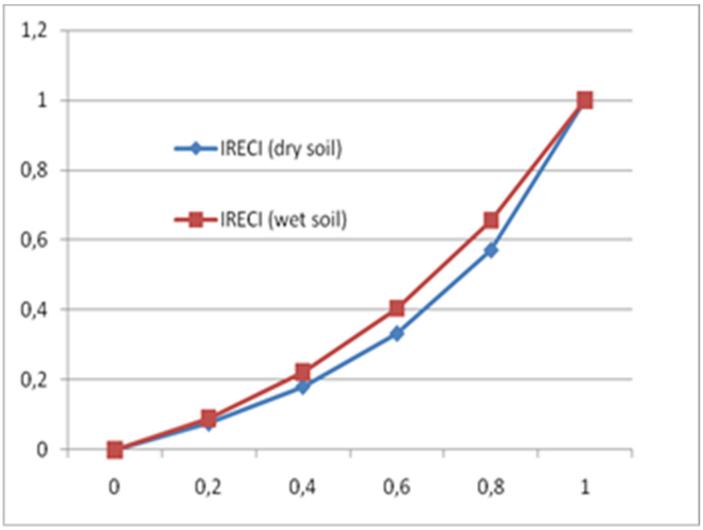


Fig. 6. IRECI values from S-2 bands for different crop-soil ratios for dry and wet soils

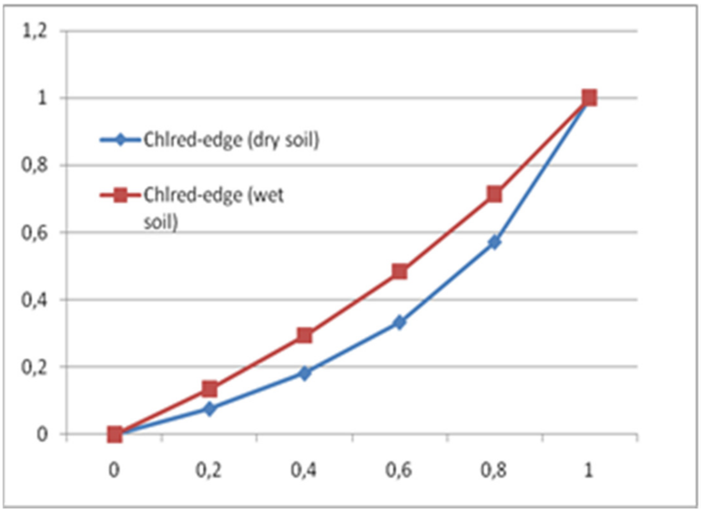


Fig. 7. IRECI values from S-2 bands for different crop-soil ratios for dry and wet soils

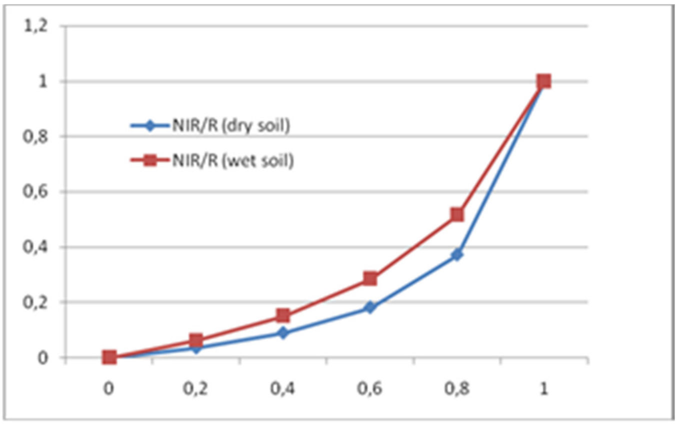


Fig. 8. NIR/R values from S-2 bands for different crop-soil ratios for dry and wet soils

D. *Estimation of LAI from VIs*

Linear regression was performed to derive linear transfer functions between LAI and VIs. The best performing VI was Clred-egde with R^2=0.77 (Fig. 9). For IRECI, R^2=0.68. These results are lower then presented in the reference studies [4], [5]. This could be due to the noise present in Sentinel-2 data and *in situ* data.

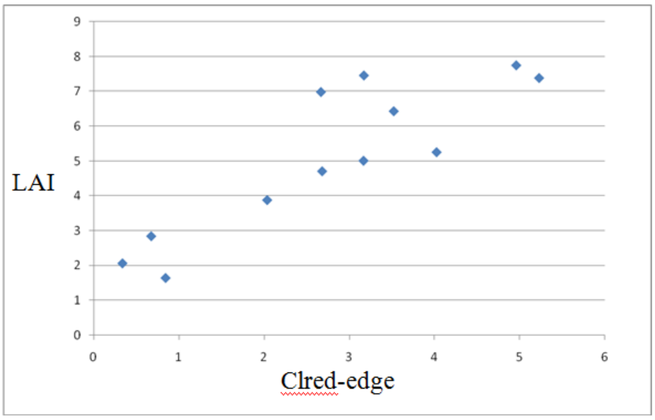


Fig. 9. Relationship between LAI and Clred-edge

IV. CONCLUSIONS

Reviewing the previous studies and analyzing the results of the conducted experiments lead to the following conclusions:

1. Models for operational generation of biophysical parameters from VIs should be crop specific.

2. To define operational models, testing of the transfer functions proposed by the literature and fine calibration using significant amount of *in situ* data could be a solution.

3. It is important to minimize the noise of both Sentinel-2 and in situ data in calibration phase and to be able to detect Sentinel-2 data noise in operational phase.

4. Collecting high quality in situ data is essential for calibration of operational models, but is also the biggest challenge.

ACKNOWLEDGMENT

The paper represents the result of the research carried out on the project TR36035, financed by the Ministry of Education and Science of the Republic of Serbia and the project APOLLO that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 687412.

REFERENCES

[1] A. A. Gitelson, “Wide Dynamic Range Vegetation Index for Remote Quantificatin of Biophysical Characteristics of Vegetation”, J. Plant Physiology, vol. 161, pp. 165-173, 2004

[2] J.G.P.W. Clevers, A.A. Gitelson, “Remote estimatin of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3”, International Journal of Applied Earth Observation and Geoinformation, vol. 23, pp. 344-351, 2013

[3] J. L. Hatfield and J.H. Prueger, “Value of Using Different Vegetative Indices to Quantify Agricultural Crop Characteristics at Different Growth Stages uner Varying Management Practices”, Remote Sensing, vol. 2, pp. 562-578, 2010

[4] J. Delegido, J. Verrelst, L. Alonso, J. Moreno, "Evaluation of Sentinel-2 Red-Edge Bands for Empirical Estimation of Green LAI and Chlorophyll Content", Sensors, vol. 11, pp. 7063-7081, 2011

[5] A.L. Nguy-Robertson, Y. Peng, A.A. Gitelson, T.J. Arkebauer, A. Pimstein, I. Herrmann, A. Karnieli, "Estimating green LAI in four crops: Potential of determining optimal spectral bands for universal algorithm", Agricultural and Forest Meteorology, vol. 192-193, pp. 140-148, 2014

Recurrent Neuronal Network tailored for Weather Radar Nowcasting

[extended abstract]

A. Scheidegger [1,2]

[1] Urban Water Management
Eawag
8600 Dübendorf, Switzerland
andreas.scheidegger@eawag.ch

[2] Institute of Hydraulic Engineering
University of Stuttgart
Stuttgart, 70569, Germany
andras.bardossy@iws.uni-stuttgart.de

***Abstract*—A non-standard recurrent neuronal network for short-term weather radar predictions is presented. A special layer allows warp the input images to mimic advection. The aim is to combine the flexibility of machine learning methods with the robustness and efficiency of conceptual models.**

***Keywords*—machine learning, spatial transformer, deep learning, convolution**

I. Introduction

Weather radars provide images of the rain field in high temporal (few minutes) and spatial (0.1 to 10 km) resolution. This high resolution and the large coverage often out-weights relatively large measurement errors and biases in comparison to rain gauges [1]. For many applications not only is the currently observed rain field of interest, but also predictions of its development in the near future (such short-term predictions are often called *nowcasting*). Potential applications of such predictions are disaster warning, traffic coordination, or the control of urban sewer systems.

Various methods have been proposed for rain field nowcasting. Most are either based on the identification, tracking and extrapolation of single rain cells [2], or on the estimation of a velocity field by comparing consecutive radar images and translation of the last observed image [3,4]. From a statistical perspective radar nowcasting is similar to predicting the next steps of a very high dimensional time series. For such predictive tasks data driven machine learning techniques can outperform more physically based models if a sufficiently large data set is available. It is therefore not surprising that machine learning methods, in particular artificial neuronal networks (ANN), have been applied in various studies for radar nowcasting—with varying degree of success [5,6,7].

Recently, the machine learning and artificial intelligence community has made astonishing progress with ANNs, often rebranded as *deep learning*. Shi et al. [7] applied some of the popular deep learning techniques for radar nowcasting. They achieved promising results with a combination of LSTM and convolutional layers (see the Method Section for an explanation of this terms) .

The nowcasting method presented in the following is also based on state-of-the-art machine learning techniques. Instead of relying on a generic standard model structures as commonly done in machine learning, the model structure was carefully designed to mimic traditional velocity field based approaches. This results in a more efficient, easier to train model.

II. Methods

A. Model structure

The model is based on a recurrent artificial neuronal network (RNN) [8]. RNNs use the output of one time step as additional input for the next time step. Therefore RNNs are a natural choice to model dynamic systems. The RNN is implemented as long short-term memory (LSTM) to better capture potential long-term dependencies [9]. While standard feed-forward ANNs have been used to predict radar images [6], RNNs are conceptually more satisfying and have the advantage that they learn which information to store in the memory and for how long.

Further, the model consists of four different types of layers (functions with learnable parameters) described below: a) fully-connected layers, b) convolution layers, c) a spatial transformer, and d) deconvolution layers. Fig. 1 shows the computational flow between the different types of layers. The last observed image P_t contributes via two paths to the prediction of the next time step: directly via the spatial transformer (to model advection), and indirectly via the internal state H_t for intensity corrections.

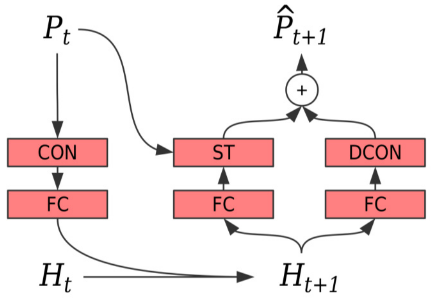


Fig. 1. Simplified computational flow from the rain field P_t to the prediction at the next time step. The internal state is denoted as H_t, CON stands for convolution, FC for fully connected, and DCON for deconvolution layers. ST is the spatial transformer.

a) Fully-connected layers

Traditional ANNs consist of one or more fully-connected layers (FC) where an output is calculated as the weighted sum of every input. Often a non-linear function is additionally applied to the outputs.

Because FC layers can approximate almost any function, their task is to learn dependencies from the data that cannot be clearly described. The mapping from the internal state to the parameters of the spatial transformer is an example of such an unknown dependency. As non-linearity the so called Leaky Rectified Linear Unit function, *max(x, x/5),* is used.

b) Convolutional layer

In recent years the field of computer vision was enormously successful applying ANNs to image classification tasks. The key elements to this are convolutional layers that transform an image (or in general any matrix) into another image with different properties. This is done by moving local filters over the complete image (mathematically a convolution). Depending on the parameters of a filter, the resulting image emphasizes different properties of the original image. The idea is that the ANN learns the parameters of these filters during calibration. This enables the model to extract the most informative features of an image for the task at hand.

Twenty different convolution filters are applied to each radar image with the aim of producing good features for the following FC layers.

c) Spatial transformer

The central part of the model is the spatial transformer (ST) [10]. Originally, STs have been developed to improve image classification: they give a classification model the possibility to crop or correct the perspective of an input image. The main contribution of [10] was to find an image transformation for which all derivatives can be calculated allowing gradient based training algorithms to be applied.

In the current study a ST is used to model advection by translating the last radar image. Translating alone is not enough, because the advection may differ for different regions. Therefore the ST can also "warp" an image, i.e. stretch it locally in different directions. This is achieved with an approach similar to thin-plate warping [11]. The parameters of the ST are provided by a FC layer.

d) Deconvolution layer

Deconvolution layers are similar to convolution layers. Instead of applying many filter to one input image, different filters are applied to different input images which are then averaged to obtain a single output image. This is helpful in reconstruction of an image.

Here the deconvolution layer is used to construct a "correction image" that is added to the transposed image. This gives the model the possibility to change intensities, e.g. to correct for the influence of a mountain range.

B. Training

The parameters of all layers are estimated jointly (trained or learned in machine learning speak) by minimizing a loss function. The square of the difference between the true and predicted rain intensity averaged over all pixels was used as loss function.

Because the model is based on carefully designed layers, the derivative of every parameter with respect to the loss function can be computed. This enables the use of efficient

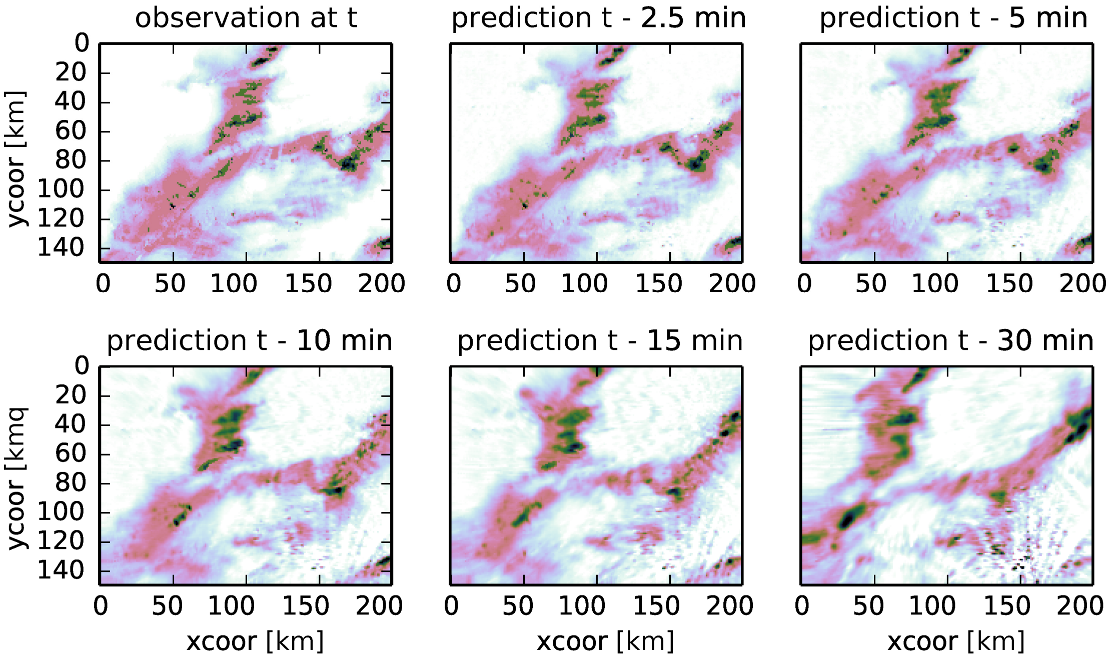


Fig. 2. Rain field predictions for time point *t* made with different forecast horizons. The true observed image is shown top left. This data have not been used to train the model.

stochastic gradient descent algorithms. In addition, *dropout* was applied to all FC layers to avoid over-fitting [12].

C. Implementation

The method is implemented with *chainer* [13] a Python based deep learning framework that can be extended easily with new functionalities such as the ST. At the same time common building blocks of ANNs are provided and computations on GPUs are greatly facilitated.

III. Results

The model was trained and validated on the operational weather radar product PRECIP-SV (1km spatial, 150sec temporal resolution) of Meteo Swiss [14]. The product covers the whole of Switzerland, however, to reduce computational time during development, the model was only applied to a subregion with an area of 150x200 km^2. Training was performed with Hinton's RMSprob algorithm on data of one month (ca. 17'000 images) for 25 epochs (i.e. the algorithm iterated 25 times over every image), although ten epochs would have been sufficient. Using a middle class graphic card for computation, the training time for one epoch with one month of data takes approximately three minutes.

Preliminary results for data that have not been used for training are presented in Fig. 2: the true and the predicted rain field for the same point in time are shown for different forecast horizons. The model captures the position and intensities quite well, however, predictions with a longer forecast horizons loose some details. Note that some radar artifacts are also predicted. While not exactly desirable, it shows that the model is capable of learning small local features from data.

IV. Discussion and Outlook

The presented results and model structure are preliminary. Systematic tuning and validation on larger data sets and different conditions are needed to find the optimal configuration and to quantify the expected prediction error.

Data driven models are known to perform well if the input lies within the range covered by training data but poorly if extrapolation is needed. The proposed approach mitigates this problem by using a specifically designed model structure to provide the model some "guidance" and at the same time give the freedom to learn patters from data. For example if the input contains a rain event of much larger intensity as the training data, it only needs to be able to detect advection correctly and the spatial transformer preserves the high intensities.

There is potential to improve the model in different directions. One of the main advantages of constructing a model as ANN is that the parameters can be learned iteratively: after every new image the parameters are slightly adjusted. Instead of training the model once and then keeping the parameters constant, the parameters value could be updated in real time. In operation the model could be re-calibrate every time a new image becomes available. This would ensure that the model always uses the optimal parameters for the present situation.

Another promising approach is to include additional input information. For example it seem plausible that wind predictions of numerical weather models may help or that the time of day contains some information about the occurrence of convective rain events. A data driven approach has the advantage, that such information can be included without the need of defining an exact relationship. During calibration the model learns how to weight the different inputs.

The presented method could also be applied to the results of data assimilation methods and not only to pure radar products. Such assimilation methods combine the information of different types of rain sensors, for example radar and rain gauges measurements [15,16].

The availability of efficient methods for parameter estimation is the main advantage of ANNs and hence also the presented approach. The resent advances of machine learning research have made it much easier to tailor the model structure of an ANN to a particular problem. Being able to use machine learning techniques to learn patterns from data and, at the same time, using a non-standard model structure that reflects system understanding offers a promising approach to now-casting tasks. This method may also be useful of other than radar products.

References

[1] Rinehart, R. E. (2004) Radar for Meteorologists, Columbia, MO, Rinehart.

[2] Wilson, J. W., Crook, N. A., Mueller, C. K., Sun, J., and Dixon, M. (1998) Nowcasting Thunderstorms: A Status Report. Bulletin of the American Meteorological Society, 79(10), 2079–2099.

[3] Bowler, N. E., Pierce, C. E., and Seed, A. (2004) Development of a precipitation nowcasting algorithm based upon optical flow techniques. Journal of Hydrology, 288(1), 74–91.

[4] Peura, M. and Hohti, H. (2004) "Motion vectors in weather radar images" in Proceedings of the 7th International Winds Workshop, Helsinki, Finland.

[5] Chow, T. W. S. and Cho, S. Y. (1997) Development of a recurrent Sigma-Pi neural network rainfall forecasting system in Hong Kong. Neural Computing & Applications, 5(2), 66–75.

[6] French, M. N., Krajewski, W. F., and Cuykendall, R. R. (1992) Rainfall forecasting in space and time using a neural network. Journal of hydrology, 137(1), 1–31.

[7] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W. (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv preprint arXiv:1506.04214.

[8] Jaeger, H. (2002) Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach, GMD-Forschungszentrum Informationstechnik.

[9] Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. Neural computation, 9(8), 1735–1780.

[10] Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015) Spatial Transformer Networks. arXiv:1506.02025 [cs].

[11] Glasbey, C. A. and Mardia, K. V. (1998) A review of image-warping methods. Journal of applied statistics, 25(2), 155–171.

[12] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958.

[13] Tokui, S., Oono, K., Hido, S. and Clayton, J., Chainer: a Next-Generation Open Source Framework for Deep Learning, Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), (2015)

[14] Germann, U., Galli, G., Boscacci, M., and Bolliger, M. (2006) Radar precipitation measurement in a mountainous region. Quarterly Journal of the Royal Meteorological Society, 132(618), 1669–1692.

[15] Sideris, I.V., Gabella, M., Erdin, R., Germann, U., 2014. Real-time radar–rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. Q.J.R. Meteorol. Soc. 140, 1097–1111.

[16] Scheidegger, A., Rieckermann, J., 2014. Bayesian assimilation of rainfall sensors with fundamentally different integration characteristics. International Symposium Weather Radar and Hydrology, Washington, DC.

Observed NDVI change in poplar plantation based on LANDSAT 8 images

[abstract]

Dejan Stojanović, Bratislav Matović, Zoran Novčić, Saša Orlović
Institute of Lowland Forestry and Environment
University of Novi Sad, Serbia
dejan.stojanovic@uns.ac.rs

Abstract—**The normalized difference vegetation index (NDVI) is often used tool for assessing changes in vegetation. It is mostly used in agriculture and forestry. Some of the most important applications in forestry are assessment of the forest health status, monitoring of pest attacks, observing of forest fires occurrence, etc. LANDSAT 8 satellite images provide good framework for evaluation of NDVI across the Earth . Specifically, spectral bands red (640 – 670 nm) and near infrared (850-880 nm) are used for such evaluations. Similar approach has been applied in this research. NDVI index change between summer 2013 and summer 2015 for the area of the experimental estate of the Institute of Lowland Forestry and Environment near Novi Sad has been evaluated. Data about forestry management measures within the observed period were collected and compared with observed image data. Acquired results confirmed that NDVI can be effectively used for monitoring of changes in poplar plantations.**

Total electron content prediction using machine learning techniques

[full paper]

Miljana Todorović Drakul[1], Mileva Samardžić Petrović, Sanja Grekulović[1], Oleg Odalović[1], Dragan Blagojević[1]
[1]Department of geodesy and geoinformatics, Faculty of Civil Engineering
University of Belgrade, Serbia
mtodorovic@grf.bg.ac.rs

Abstract— **Total electron content (TEC) is one of the basic parameters that are determined in order to investigate ionospheric phenomena and as such it has practical applications in various fields like satellite positioning and navigation as well as in other Earth Observation Systems (EOS). Since TEC values vary both spatially and temporally, they are dependent on many factors, primarily from the Sun's solar activity and Earth's geomagnetic activity. Accurate modeling and prediction of TEC values represents a complex nonlinear problem and it's solving is critical for improvement of Global Navigation Satellite Systems (GNSS) measuring accuracy. The representation and prediction of various dynamic spatial phenomena is often achieved using different types of machine learning (ML) techniques. The capability of ML to learn and derive patterns for examined phenomena based on data, has made these techniques very popular in many geosciences, particularly in geodesy and geophysics. This paper presents creation and comparison of results for TEC prediction models acquired using different machine learning methods. TEC values used in this study were obtained using dual-frequency GNSS observations collected on stations that are integral part of permanent stations network of Republic of Serbia.**

Keywords—TEC; machine learning; ionosphere; GNSS

I. Introduction

Global Navigation Satellite System (GNSS) i.e. Global Positioning System (GPS) techniques have been significantly improved over the past two decades, however several sources of errors still remain, which may limit accuracy, practical operation and performances of precise positioning. Ionosphere is the major source of errors for the GNSS positioning [1]. In this region, ionizing radiation from the Sun causes the existence of electrons, in the quantities that influence radio-waves propagation [2]. The number of electrons intercepted by the electro-magnetic waves traveling through ionosphere is known as the Total Electron Content - TEC. It represents an integral of electron density per unit of volume, along the signal path between the satellite and the GNSS receiver. It is noted in TECU units, with 1 TECU being 10^{16} electrons per square meter of cylindrical cross-section.

Ionosphere is a very dynamic environment, and the electron density may significantly vary per minute at the given location, which leads to temporal and spatial variations in the Total Electron Content. The most significant changes occur due to the geomagnetic activity and Earth's revolution around the Sun in the period of the equinox and solstice [3]. Indicators of these changes are primarily Solar flux (SF), Sunspot number (SSN), Index of geomagnetic activity (Ap index). Furthermore, the largest signal delay caused by the ionospheric influence occurs in the 10-14h local time interval [4]. The testing is conducted for years 2013, 2014 and 2015 that are at the peak of 11 year cycle of solar activity. In order to model and predict previously mentioned extreme TEC values, ML techniques were used.

Machine learning (ML) techniques are empiric modeling approaches that have the capability to extract information and reveal patterns by exploring unknown relations between input and output variables (dependent and independent continual and categorical variables). ML techniques include a great many techniques with different kinds of learning algorithms. However, two most commonly used ML techniques in order to model/predict TEC values are Neural Networks (NN) and Support Vector Machines (SVM) [5,6,7,8]. Therefore, the main objective of this research study is to examine the capability of those ML techniques to model and predict extreme TEC values. For that purpose, we separately examined and analyzed the capability of SVM and NN to model both spatial – temporal and spatial variation of TEC values.

II. Methods

A. TEC based on GPS observation

Ionosphere delay is nearly proportional to the Total Electron Content along the signal path and inverse proportional to the frequency squared. This dispersion property of ionosphere provides for dual frequency GNSS receivers to compensate for the errors of ionosphere delay and measure the *TEC*.

To compensate ionospheric delay, dual-frequency GPS receivers use L1 (1575.42 MHz) and L2 (1227.60 MHz) frequencies. Delay, Δt=t2-t1, measurement between L1 and L2 frequencies is used to calculate TEC along the signal path:

$$\Delta t = \left(\frac{40.3}{c}\right)\cdot\frac{\mathrm{TEC}}{\left(\frac{1}{f_2^2}-\frac{1}{f_1^2}\right)} \tag{1}$$

where *c* is the speed of light in open space. Considering that using pseudo-ranges provides absolute TEC, while using phase differences improves the accuracy, GPS data provides for the efficient method of estimating TEC values with greater spatial and temporal coverage [9, 10]. Having that the frequencies used by the GPS system are sufficiently high, the signals are minimally influenced by ionospheric absorption and Earth's magnetic field, both in short and in long-term changes in the ionosphere structure. Here, the values of slanted TEC were obtained as the sum of slanted TECs*,* hardware satellite delay b_S and hardware receiver delay b_R. Thus, vertical TEC (VTEC) may be expressed as follows:

$$\mathrm{VTEC} = \frac{(\mathrm{STEC} + b_S + b_R)}{S(e)} \tag{2}$$

where STEC is slanted TEC, e is elevation angle of satellites in degree, *S*(e) is the slant factor against the zenith angle *z* at the Ionospheric Pierce Point (IPP) and VTEC is vertical TEC in the IPP point. The slant factor, *S*(e) (or the mapping function) is defined as [11]:

$$S(e) = \frac{1}{\cos(z)} = \left(1 - \frac{R_e \times \cos(e)}{R_e + h_i}\right)^{-0.5} \tag{3}$$

where *R*e is the average Earth's radius in km, and *h*i is the (effective) height of ionosphere over the Earth's surface.

B. TEC based on ML techniques

In this research we used well known SVM and Radial Basis Function (RBF) [12] as kernel function, and for NN we used Multi-layer Perceptron (MLP) [13], with softplus activation function. The efficient application of most ML techniques including SVM and NN techniques requires selection of optimal combination of function parameters. Therefore, in order to use those algorithms appropriately it was necessary to find optimal combination of two parameters; γ of the RBF and penalty C for TEC SVM based models and the number of neurons in a hidden layer for TEC NN based models. Furthermore, the attributes with little or irrelevant information can often confuse the algorithm learning process and lead to wrong conclusions, for that reason it is very important to perform the attributes selection. The main goal of attribute selection is to choose a subset of informative attributes by eliminating those with little or no information relevant for the TEC values [14]. For that purpose, we used Correlation-based Feature Subset (CFS) [15] selection method. The CFS method automatically determines a subset of *m* relevant attributes (*m* < *n*, *n* is number of all considering attributes), i.e. attribute that are highly correlated with the TEC but uncorrelated with each other.

III. Data

Data for three base stations have been taken over from the archive, in the form of 30-second RINEX files (Fig.1). Examined stations Kanjiža (S_K), Novi Pazar (S_{NP}) and Šabac (S_S), belong to the permanent GNSS network of the Republic of Serbia under the name AGROS (Active Geodetic Reference Base of Serbia). Collected data contain five-days observations from March, June, September and December of 2013, 2014 and 2015. Temporal series of TEC measurements have been obtained using (2). GPS TEC Analysis software, developed at Boston University, was used for processing [16]. Phase and code values on both frequencies have been used to eliminate clock and tropospheric effect errors, in order to calculate relative values of slanted TEC [17]. Afterwards, absolute TEC values have been obtained by removing hardware delays, i.e. differential code discrepancies between the satellite (produced by the Data Centre of the Bern University, Switzerland) and the receiver (obtained by minimizing TEC value between 2:00 AM and 6:00 AM - local time) [16]. Trigonometric single-layer mapping function (3) was used to convert TEC to VTEC at AGROS stations and at the IPP point at the altitude of 350 km. Elevation angle was limited to the value of 20 degree to decrease a potential effect of multiple signal reflection during the tests. Data sampling was done at 30 seconds.

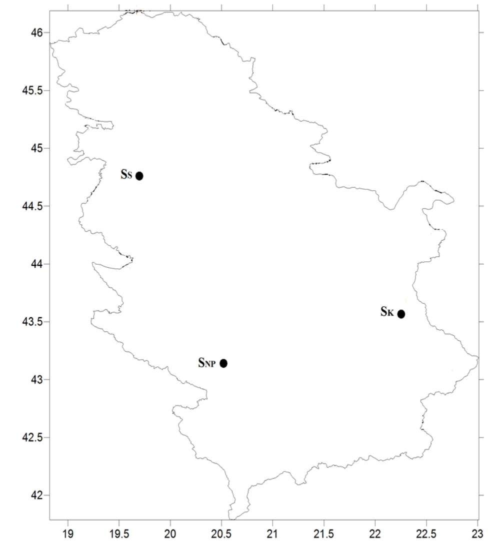


Fig. 1. Distribution of AGROS stations, S_K, S_S and S_{NP}, within the territory of Serbia.

Given that ionosphere influence on GPS observations is largest in the period 10-14 h UT, values of TEC are calculated and averaged for this time interval. Fig. 2 shows TEC values for all three stations in the studied time interval. It can be concluded that TEC values for the same time intervals and for all stations vary in the 1 to 7 TECU range. Changes in TEC values during different seasons are significant and vary from 10 to 55 TECU, where the lowest TEC variations are recorded in June, and the highest in Match period for all considered years.

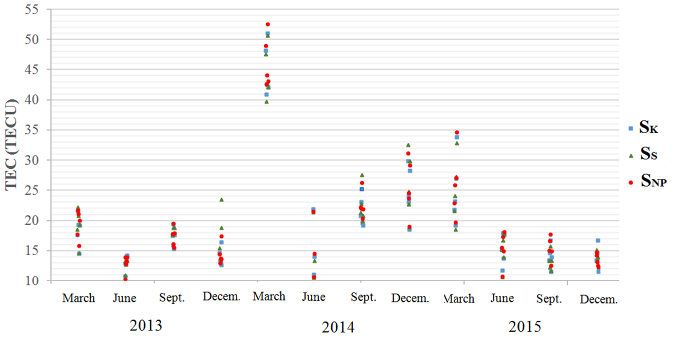


Fig. 2. Distribution of TEC values for examined time intervals based on observation from three stations, S_K, S_S and S_{NP}.

Given that the SF, SSN and Ap index are factors that influence ionospheric variations, they are included into the model as attributes. Their values were downloaded from NASA's Space Physics Data Facility (http://omniweb.gsfc.nasa.gov/form/dx1.html), for all 12 periods of interest (Fig. 3).

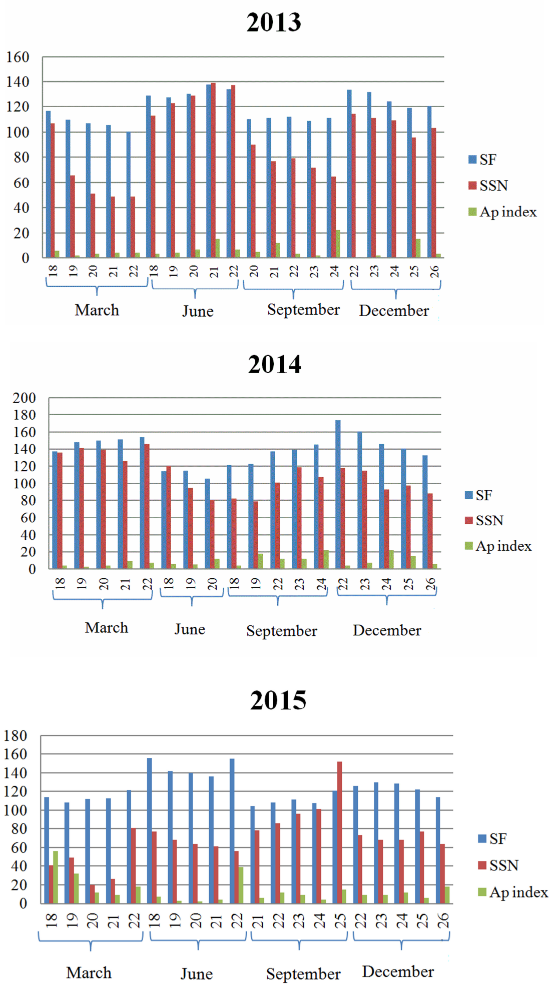


Fig. 3. Distribution of SF, SSN and Ap index values for examined time intervals at 2013, 2014, and 2015 years.

IV. Result and Discussion

In order to create and validate spatial – temporal ML TEC models, two independent datasets were created based on the values of TEC, Solar flux (SF), Sunspot number (SSN), Index of geomagnetic activity (Ap index), and geographic coordinates (lat, long, h). Furthermore, in order to indicate the time intervals in which TEC value was obtained in regards to winter and summer solstice and autumnal and vernal equinox, additional attribute was defined, labeled as Month. Each dataset contains TEC values for all three stations (S_K, S_{NP} and S_S), represented as a row vector of attributes a_{it}, i=1,…,7, including TEC values as target attribute. The training dataset, label as TR_1, contained the time points from years 2013 and 2014 and the test dataset, label TE_1, contained the time points from year 2015, where the actual TEC (TEC based on GPS observation) at year 2015 was used for comparison with the predicted TEC values in the year 2015.

Furthermore, in order to create and validate spatial TEC ML models, two independent datasets were created based on the values of TEC, SF, SSN, Ap index, lat, long, h, and Month, where each dataset is represented as a row vector of attributes a_{it}. i=1,…,7, including TEC values as target attribute. The first dataset used for training, labeled as TR_2, contained data form station S_K, S_{NP} for all investigated time points and second data used for a validation, labeled as TE_2, contained data form station S_S, where the actual TEC at all time points was used for comparison with modeled TEC values at station S_S.

After the creation of datasets, attribute selection was performed on training dataset TR_1 and TR_2. The CFS automatically determines a subset of *m* relevant attributes, which are highly correlated with the TEC values and uncorrelated with each other. Given that both sets contain the same attributes, CFS method selected the same subset of four attributes: Month, SF, lat and long for the both training sets TR_1 and TR_1. All TEC models were therefore built and validated on training and test datasets that contain only selected attributes, labeled as $TE1_{CFS}$, $TR_{1CFS,}$ TR_{2CFS} and TE_{2CFS}, respectively. The implementation of CFS and ML was performed using Weka software [18]. The SOMreg algorithm was used to implement the SVM for regression algorithm and MLPreg algorithm to implement Neural Network. Using training datasets and 10 fold cross-validation, optimal combination of parameters was found for both ML techniques and for both types of models. Minimum, maximum and mean error, standard deviation (St.Dev.) and root mean square error (RMSE) are used as spatial and spatial – temporal ML TEC models quality controls measure and, accordingly, the best performing models are present in Table 1. Histograms of the errors for those models are present in Fig. 3.

TABLE I. BEST PERFORMING MODELS FOR TE$_{1CFS}$ AND TE$_{2CFS}$ DATASETS

Datasets/ML techniques	Quality controls measure				
	Min	*Max*	*Mean*	*St.Dev*	*RMSE*
TE$_{1CFS}$/SVM	-7.72	9.03	1.19	3.84	4.02
TE$_{1CFS}$/NN	-10.52	8.27	0.03	4.10	4.10
TE$_{2CFS}$/SVM	-7.72	2.23	-0.36	1.54	1.58
TE$_{2CFS}$/NN	-4.72	3.72	0.15	1.69	1.70

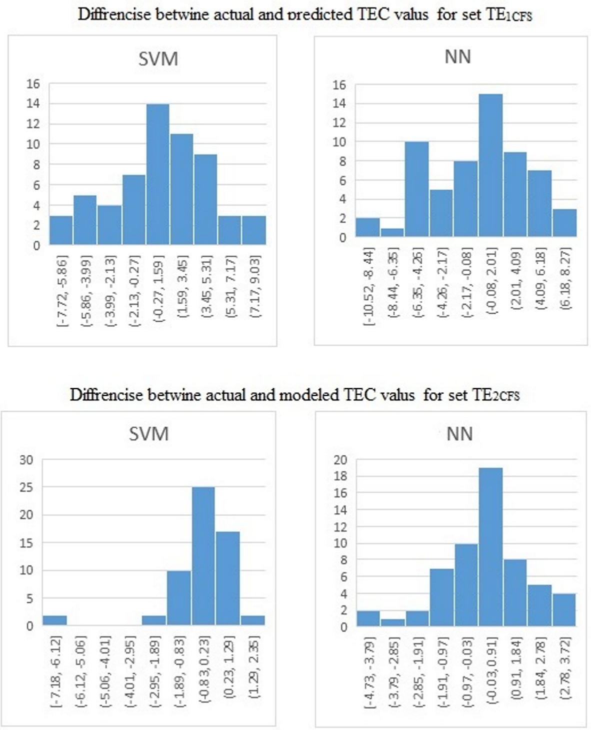


Fig. 4. Histograms of the differences between actual and modeled TEC values for best performing SVM and NN techniques for both type of models, spatial -temporal TEC and spatial TEC models.

Based on the results it can be concluded that both ML techniques define trend of TEC values and its variations through space more efficiently then through space and time. Those obtained results were expected due to the existence of larger variations of TEC values through time in comparison to the variations through space. Considering the range of actual TEC, from 10 to 53 TECU, given the relatively small time interval and small samples, and considering that the objective was to predict the extreme values of TEC, the obtained results for differences range, mean values and standard deviation (Table. 1) are acceptable. The obtained standard deviation and RMSE are equal or less than 1.7 TECU and mean values of differences between actual and modeled TEC values are -0.36 (with range from -7.72 to 2.23) and 0.15 (with range from -4.72 to 3.72), for spatial TEC SVM and NN based model, respectively. Furthermore, considering that differences between actual TEC values per station vary from 1 to 10 TECU at same time point, it can be concluding that the obtained results of spatial TEC are also satisfactory. When analyzing the distribution of actual and spatially modeled TEC differences (Fig.3), it can be noted that results obtained from spatial TEC NN based model are closest to the normal distribution, which indicates that NN have slightly better performance comparing to SVM. Furthermore, it is evident that results obtained from spatial TEC SVM based model contained grouped outliers.

V. CONCLUSION

This research study examined the possibility of two most commonly used ML techniques, SVM and NN, in order to model extreme values of TEC. For that purpose, we used GPS observation from three AGROS stations, distributed within the territory of Serbia. The largest variation of TEC follows the changes in solar activity (11-year cycle of Sun activity, winter and summer solstice and autumnal and vernal equinox). Therefore, the average values of TEC between 10 and 12 UT, for five days, for each season and for three years of interest (2013, 2014 and 2015) were calculated from GPS observations from each station and used as samples for modeling. The two types of models were examined, the spatial-temporal and just spatial. Therefore, two pairs of appropriate independent training and test datasets were created for each type of model.

The used factors (attributes) of TEC changes were Solar flux, Sunspot number, Index of geomagnetic activity, geographic coordinate of stations and one additional attribute (Month) generated in order to represent the time intervals. In order to remove uninformative attributes, CFS attribute selection method was performed and two new pairs of training and test datasets that contain only four attributes (Solar flux, Month, latitude and longitude) were created.

After performing attribute selection by CFS method and finding appropriate and optimal combination of parameters for both ML techniques and types of models, the best performing models were analyzed.

Generally, the differences between the results obtained by SVM and NN models are small, thus indicating that both techniques are capable to adequately predict and spatially model extreme TEC values. Based on the results it can be concluded that both ML techniques define trend of TEC values and its variations through space more efficiently then through space and time. Since that it can be expected that the model can be improved using larger number of samples and time intervals, in future work our attention will be dedicated to extending the samples.

ACKNOWLEDGMENT

The support of this study was made through Ministry of Education, Science and Technological Development of the Republic of Serbia (TR36020) awarded to first, third, fourth and fifth authors.

REFERENCES

[1] B. Hofmann-Wellenhof, H. Lichtenegger and J. Collins, “Global Positioning System, Theory and Practice,” 4th edition, Springer-Verlag, Berlin, Heidelberg, New York, 1992, pp 389.

[2] A. Kleusberg and P.J. Teunissen, “GPS for Geodesy,” Springer-Verlag Berlin Heidelberg, 10.1007/BFb0117676, 1996.

[3] G. Wautelet and R. Warnant, “Climatological study of ionospheric irregularities over the European mid-latitude sector with GPS,”. J. Geodesy, 88(3), 2014, pp. 223-240.

[4] Z.R. Radzi, M. Abdullah, A.M. Hasbi, J.S. Mandeep, S.A. Bahari, “Seasonal variation of Total Electron Content at equatorial station, Langkawi, Malaysia,” Proceeding of the 2013 IEEE International Conference on Space Science and Communication (IconSpace), 1-3 July 2013, Melaka, Malaysia pp

[5] V. Barrile, M. Cacciola, F.C. Morabito and M. Versaci, "TEC measurements through GPS and artificial intelligence." Journal of electromagnetic waves and applications 20, no. 9, 2006, pp. 1211-1220.

[6] Z. Huang and H. Yuan, "Research on regional ionospheric TEC modeling using RBF neural network." Science China Technological Sciences 57, no. 6, 2014, pp. 1198-1205.

[7] J.B. Habarulema and L.A. McKinnell, "Investigating the performance of neural network backpropagation algorithms for TEC estimations using South African GPS data." In Annales Geophysicae, Copernicus GmbH, vol. 30, no. 5, 2012, pp. 857-866.

[8] D. Okoh, O. Owolabi, C. Ekechukwu, O. Folarin, G. Arhiwo, J. Agbo, S. Bolaji and B. Rabiu, "A regional GNSS-VTEC model over Nigeria using neural networks: A novel approach,". Geodesy and Geodynamics, 2016, pp. 19-31

[9] K. Davies, G.K. Hartmann, “Studying the ionosphere with the global positioning system,” Radio Sci, 1997, 32(4), pp. 1695–1703.

[10] K. Igarashi, M. Nakamura, P. Wilkinson, J. Wu, A. Pavelyev, and J. Wickert. "Global sounding of sporadic E layers by the GPS/MET radio occultation experiment." J. Atmos. Sol-Terr. Phy., 2001, 63(18), pp. 1973-1980.

[11] R. Langley, M. Fedrizzi, E. Paula, M. Santos and A. Komjathy, “Mapping the low latitude ionosphere with GPS,” GPS World, 2002, pp. 13 ,41 –46.

[12] S. Abe, “Support Vector Machines for pattern classification,” Springer, London, 2010, pp. 471.

[13] D. Rumelhart, G. Hinton and R. Williams, “Learning internal representations by error propagation,” In D. E. Rumelhart, & J. L. McClelland (Eds.), Parallel distributed processing: explorations in the microstructures of cognition Cambridge: MIT Press 1, 1986, pp. 318–362.

[14] Y.S. Kim, W.N. Street and F. Menczer, “Feature selection in data mining,” In: Wang, J. Data mining: opportunities and challenges. Idea Group Inc., 2003, pp. 80-105.

Extensions of 3D trend models of soil variables

[abstract]

Milutin Pejović
Department of geodesy and geoinformatics
Faculty of civil engineering, University of Belgrade
Belgrade, Serbia
mpejovic@grf.bg.ac.rs

Abstract— **From a geostatistical point of view, trend in soil data is systematic non-random part of variation that can be represented by function of spatial coordinates or some other environmental variables. Considering the fact that the all environmental variables are in fact surface related, 3D trend model of soil variables is mainly represented as sum of lateral (2D) and depth-wise components. Lateral component relates environmental variables (spatial covariates) to the modeled soil properties while the vertical component models vertical variation as linear or non-linear (spline) function of soil depth. In some cases, this form could be too restrictive, because the effects of spatial covariates were not allowed to vary with depth as well as the effects of vertical components terms cannot vary in 2D space. One way to overcome this issue is to extend such linear model by allowing interactions between the effects of lateral and depth-wise components. By doing this, different soil depth layer can be represented by different trend. However, in case of large number of covariates, allowing the interactions between spatial covariates and depth increase the number of potentially useful predictors by twice. Problem increases with the presence of categorical variables, which require to be adequately coded prior to model fitting. Determining which predictors, including the interaction terms, should be included in a model is becoming the crucial issue at this point. Problem can be more complicated, one can insist to honor the hierarchy principle which implies that an interaction term can only be included in the model if one or both variables a statistically important. In this paper, we tried to examine whether the extension of existing linear 3D trend models based on interaction of spatial covariates and soil depth can improve the predictive capabilities of model and in what extent. In addition, we propose one predictive approach to extend existing linear 3D trend model allowing an interactions between spatial covariates and depth. Presented modeling approach is based on shrinkage regression method Lasso. Lasso enables to perform variable selection and model optimization at the same time on very effective way. Hierarchy principle was also enabled to be honored. Proposed methodology was tested on soil profile observations of Soil Organic Matter (SOM), pH and Arsenic (As) concentration sampled on the 10x20 km area in central Serbia. In order to provide reliable accuracy assessment model performance measures (R2 and RMSE) were calculated based the nested 5-fold cross-validation procedure. Folds were stratified taking into account the 3D distribution of observation as well as the whole range of response variables. Final model was selected based on whole data set within the 5-fold cross validation procedure. Obtained results show that taking interaction into account can improve the predictive capabilities of trend model up to 20% in terms of R2 and slightly less in terms of RMSE. As expected, the greatest improvement was achieved with variables that have strong decreasing trend with depth (SOM and As).**

Gridded monthly temperature fields for Croatia for the 1981–2010 period

[abstract]

Melita Perčec Tadić
Meteorological and Hydrological Service of Croatia
Zagreb, Croatia
melita.percec.tadic@cirus.dhz.hr

Abstract— **At least six reasons why gridded data are so important in meteorology, climatology and other research fields are described in [1]. Among those the two are especially interesting to us:**

1) such interpolated data sets allow best estimates of meteorological variables at locations away from observing stations, thereby allowing studies of local climate in data-sparse regions,

2) validation of Regional Climate Models (RCMs) that generally represent area averaged rather than point processes is most appropriate with interpolated observed data for present climate since such comparison assumes that the observations and model are indicative of processes at the same spatial scale.

Hence, there is a high motivation to derive sets of gridded climate data of different temporal and spatial scales for Croatia. Moreover, the intention is to outperform in accuracy, and provide the fields in higher spatial resolution, than available similar European projects. There are several important gridded data sets derived from observations only: 1) the gridded E-OBS of mean daily temperature and precipitation, 2) the EURO4M-APGD Alpine daily precipitation set [2] of 5 km resolution for 1971–2008, 3) GPCC-FD monthly precipitation data set (1901–2013) on coarser 0.5°~56 km resolution and 4) the CRU TS v. 3.23 dataset of temperature, precipitation, air pressure and water vapour (1901-2014).

Most of the previously mentioned data sets are developed with some variant of regression on climatic factors and interpolation of the anomalies. We will investigate the spatial and temporal scale of those products compared to a newly derived gridded monthly temperature fields for Croatia for the 1981–2010 period. Those grids will be derived with geostatistical methods using regression on climatic factors and interpolation of the anomalies also. These methods belong to the parametric statistical learning methods since they make an assumption about the shape of function that connects predictors and observed variable. This class of methods are useful when it is important to make an inferential statistics about the data. Some other interpolation and mapping approach are of growing interest, namely machine learning [3] and bayesian spatial and ST modelling [4],[5]. Compared to parametric statistical learning methods, those belong to non-parametric methods. Those methods do not make explicit assumptions about the functional form of the underlined process. Instead they seek an estimate that gets as close to the data points as possible without being too rough or wiggly [6]. They need a large number of observations to get an accurate estimate, while it is important not to over-fit the model.

Selected nonparametric methods will be tested to derive the monthly temperature grids and the results will be cross-validated to estimate the accuracy of the proposed methods.

References

[1] Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New (2008), A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res., 113(D20), D20119, doi:10.1029/2008JD010201.

[2] Isotta, F. A. et al. (2014), The climate of daily precipitation in the Alps: Development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data, Int. J. Climatol., 34(5), 1657–1675, doi:10.1002/joc.3794.

[3] Appelhans, T., E. Mwangomo, D. R. Hardy, A. Hemp, and T. Nauss (2015), Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt . Kilimanjaro , Tanzania, Spat. Stat., 14, 91–113, doi:10.1016/j.spasta.2015.05.008.

[4] Blangiardo, M., and M. Cameletti (2015), Spatial and spatio-temporal bayesian models with R-INLA, Wiley.

[5] Ingebrigtsen, R., F. Lindgren, I. Steinsland, and S. Martino (2015), Estimation of a non-stationary model for annual precipitation in southern Norway using replicates of the spatial field, Spat. Stat., 14(C), 338–364.

[6] James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), An Introduction to Statistical Learning with applications in R, Springer Texts in Statistics, Springer New York, New York, NY.

High resolution daily temperatures for Serbia (1960-2015)

[extended abstract]

Aleksandar Sekulić, Milan Kilibarda, Branislav Bajat

Department of geodesy and geoinformatics
Faculty of civil engineering, University of Belgrade
Belgrade, Serbia
asekulic@grf.bg.ac.rs

Abstract — **Publicly available global meteorological data sets, like GSOD, ECA&A, GHCN-daily, etc. are really popular for climate phenomena modeling nowadays. The question is whether these data sets could be used at local scales. Interpolation of maximum and minimum daily temperatures for Serbia has been obtained for the period 1960 – 2015 at spatial resolution of 1km. Temperature prediction is performed based on a global spatio-temporal regression–kriging model. The main aim of this study was to produce the simplest and fastest methodology for temperature interpolation for local areas using a global model. DEM, TWI and the geometrical temperature trend are covariates used in the regression model and the GHCN-daily station measurements are used for kriging. Even though there are only a few stations in Serbia, cross-validation and RMSE results show that the global model could be very simple and fast solution for local prediction of temperatures.**

Keywords — Geostatistics, spatio-temporal regression-kriging, temperature, GHCN-daily

I. Introduction

Spatio-temporal geostatistics has made a breakthrough in the past decade with theoretical concepts [1] and various examples of applications have been provided [2][3][4][5][6]. Most meteorological parameters, including temperature, vary both in space and time and these observations are correlated in space and time. The fitting of spatio-temporal models and making predictions using spatio-temporal covariates (regression-kriging) implies more than just the smoothing of station data.

There are a lot of publicly available global meteorological data sets from ground-based stations, such as Global Surface Summary of Day (GSOD), European Climate Assessment & Dataset (ECA&A), Global Historical Climate Network – daily (GHCN-daily), etc. which could be combined with open source software providing free solution for everyone [7].

The spatio-temporal regression-kriging model based on GSDO and ECA&A datasets is presented by Kilibarda et al. [8]. Covariates used in this model are: Digital Elevation Model (DEM), SAGA Wetness Index (TWI), geometrical temperature trend (GTT) and MODIS Land Surface Temperature (LST) 8 day images. The results of cross-validation RMSE for this model are 2-3°C for both maximum and minimum temperatures with R-square approximately 95%.

The question is whether the above mentioned methodology and GHCN-daily data set could be used for the prediction of daily maximum and minimum temperatures at a spatial resolution of 1km at local scales.

II. Materials and Methods

A. Covariates

The following three covariates are used in order to compute the linear trend model for maximum and minimum temperatures:

- DEM – DEMSRE3, Global Relief Model based on SRTM 30+ and ETOPO DEM at 1/120 arcdeegres (WorldGrids.org)
- TWI – TWISRE3, SAGA Topographic Wetness Index derived using the DEMSRE3
- GTT – expresses temperature dependency of latitude (φ) and the day of the year (day):

$$t_{geom} = a \cdot \cos \varphi - b \cdot (1 - \cos \theta) \cdot \sin |\varphi| \qquad (1)$$

where a and b are constants, different for maximum ($a = 37$, $b = 15.4$) and minimum ($a = 24.2$, $b = 15.7$) temperatures and θ is:

$$\theta = (\mathrm{day} - 18)\frac{2\pi}{365} + 2^{1-\mathrm{sgn}(\varphi)}\pi \qquad (2)$$

GTT is a very important covariate because it explains about 70-75% of maximum and minimum temperature variations. MODIS LST 8 day images need a lot of preparing in order to be used as covariates. Closing gaps because of the missing pixels and disaggregation in time through the use of splines for each pixel in order to get daily images is a time consuming processes. These are the reasons why MODIS LST images haven't been taken into account.

The GHCN-daily meteorological data set has been used in spatio-temporal kriging to compute the best unbiased linear prediction.

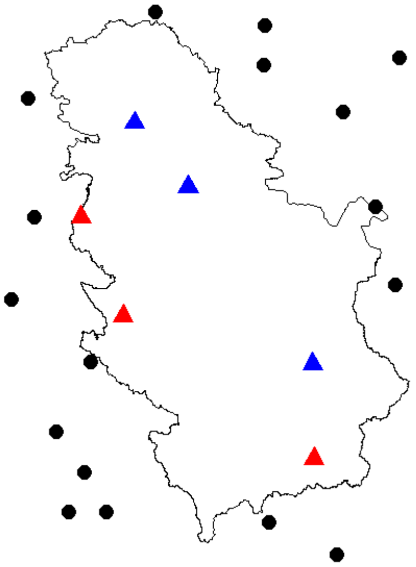


Fig. 1. GHCN stations distribution near Serbia

All of the regression coefficients, spatio-temporal variograms and other functionalities like computation of GTT, managing data and making the prediction are implemented in the R [9] Meteo package [11].

B. Experiment

In order to estimate maximum and minimum temperatures in Serbia, measurements from the twenty nearest GHCN stations and one day before the observed day are used for each 1km pixel. Positions of the commonly used stations are presented in Fig. 1 (black circles and blue triangles). The difficult circumstances were that there are only three stations in Serbia: Novi Sad, Belgrade and Niš (Fig. 1, blue triangles).

III. Results and Discussion

All images of maximum and minimum daily temperatures for Serbia for the period between 1960 – 2015 at a spatial resolution of 1km are available on OSGL Belgrade Rasdaman server (osgl.grf.bg.ac.rs/rasdaman/ows) and are free to download [11].

In Fig. 2, maximum (up) and minimum (down) temperatures for dates 15 January 2014 (left) and 15 July 2014 (right) are given. It's quite obvious how big an impact those three points have on temperature prediction, which is manifested by yellow line in Fig. 2, in Serbia. Only bottom right picture in Fig 2. doesn't show this effect because there are no observations for minimum temperatures on 15 July 2014 for these three stations.

A. Cross-validation

Aggregated results of cross-validation for the three stations in Serbia show that RMSE ranges from 1˚C to 3˚C (TABLE I.).

TABLE I. Aggregated RMSE for the three stations in Serbia

Stations	Belgrade	Novi Sad	Niš
MAX	1.908462	1.547734	2.406954
MIN	2.924081	1.388202	1.879265

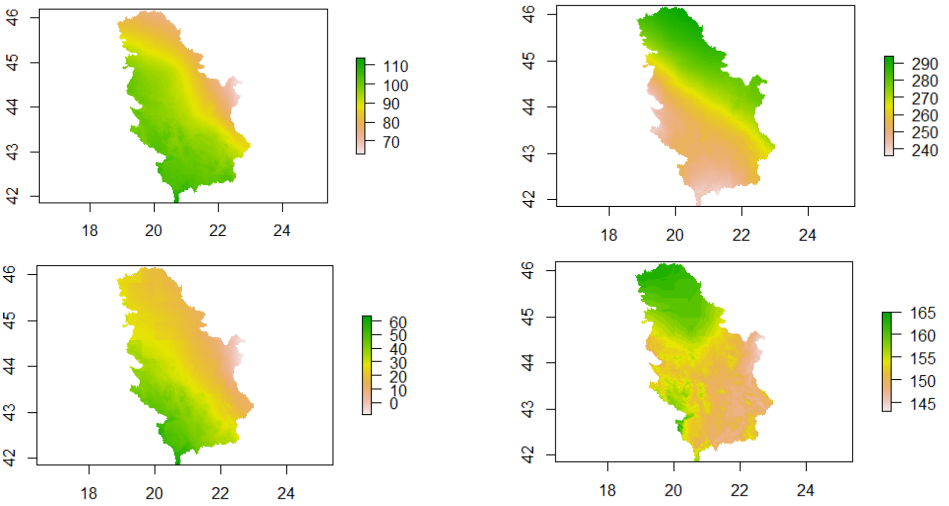


Fig. 2. Maximum (up) and minimum (down) temperatures for dates 15 January 2014 (left) and 15 July 2014 (right)

B. Testing

Three new stations, for the year 2014, which were never used, neither for modeling nor prediction, were used for testing of the maximum and minimum temperature model. These three stations are Zlatibor, Vranje and Loznica (Fig. 1, red triangles). Aggregated results of the difference between real measurements and predicted values for these stations show that RMSE ranges again from 1˚C to 3˚C (TABLE II. and TABLE III.), except for the Zlatibor station. This confirms our assessment that for areas on higher altitude the temperature model can't give an accurate prediction.

TABLE II. DIFFERENCE BETWEEN REAL MEASUREMENTS AND PREDICTED VALUES - MAXIMUM TEMPERATURES

Stations	Zlatibor	Vranje	Loznica
Altitude	1028m	432m	121m
RMSE	4.57	1.67	2.33
Min	-9.7	-4.7	-4.4
1st Qu.	-5.3	-0.9	0.5
Mean	-3.69	0.02	1.43
Median	-4.0	0.1	1.5
3rd Qu.	-2.3	1.0	2.6
Max	17.5	6.5	7.0

TABLE III. DIFFERENCE BETWEEN REAL MEASUREMENTS AND PREDICTED VALUES - MINIMUM TEMPERATURES

Stations	Zlatibor	Vranje	Loznica
Altitude	1028m	432m	121m
RMSE	2.87	1.78	2.14
Min	-12.1	-5.2	-2.5
1st Qu.	-3.5	-1.6	0.6
Mean	-2.09	-0.44	1.46
Median	-2.2	-0.5	1.5
3rd Qu.	-0.8	0.6	2.4
Max	4.8	5.3	15.7

It must be emphasized that although we did not use outlier detection, predictions are still usable for areas on lower altitudes. Because of that, the maximum and minimum differences are quite large (TABLE II. and TABLE III.). e.g. the maximum difference for minimum temperatures for the Loznica station is 15,7 (TABLE III.). This is obviously an outlier because the minimum temperature for this day was higher than the maximum temperature. If detection of outliers were performed, RMSE would be lower.

IV. CONCLUSIONS

Analyzing all of the results of the conducted experiment, a conclusion could be drawn that the global model for minimum and maximum temperatures can be used at local scales for areas on lower altitude. Even though there are few meteorological stations and MODIS LST wasn't used for prediction of maximum and minimum temperatures, this model gives good results. Using of MODIS LST and more stations in Serbia would improve the accuracy of prediction. The observed model, regarding the lower accuracy with higher altitude, is especially usable in agriculture. Future work will be to improve the model for higher altitudes. There are twenty more stations in Serbia for year 2014 which could be used for testing.

ACKNOWLEDGMENT

The paper represents the result of the research carried out on the project III47014 and project TR36035, financed by the Ministry of Education and Science of the Republic of Serbia.

REFERENCES

[1] Cressie N., Wikle C.K., "Statistics for spatio-temporal data", Wiley, 2011

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

[3] Heuvelink G. B. M., Griffith D. A., "Spatio-temporal geostatistics for geography: A case study of radiation monitoring across parts of Germany", Geographical Analysis, vol. 42, pp 161-179, 2010

[4] Heuvelink G. B. M., Griffith D. A., Hengl T., Melles S. J., "Sampling Design Optimization for Space-Time Kriging", John Wiley & Sons, Ltd., pp. 207-230, 2012

[5] Gräler B., Gerharz L., Pebesma E. J., "Spatio-temporal analysis and interpolation of PM10 measurements in Europe", ETC/ACM Technical Paper, 10, 2011

[6] Hengl T., Heuvelink G. B. M., Prečec Tadić M., Pebesma E. J., "Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images", Theoretical and Applied Climatology, vol. 107, pp. 265-277, 2012

[7] Kilibarda M., Prečec Tadić M., Hengl T., Luković J., Bajat B., "Publicly available global meteorological data sets: sources, representation, and usability for spatio-temporal analysis", International Journal of Climatology, 2013

[8] Kilibarda M., Hengl T., Heuvelink G., Graeler B., Pebesma E., Prečec Tadić M., Bajat B., "Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution", Journal of Geographical Research: Atmospheres, vol. 119, pp. 2294-2313, 2014

[9] R Core Team. „R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2015

[10] R *Meteo* package, Spatio-temporal analysis and mapping of meteorological observations, https://cran.r-project.org/web/packages/meteo/index.html, 2015

[11] OSGL, Laboratory for development of the open source technologies, http://osgl.grf.bg.ac.rs/

Monthly gridded datasets for temperature and precipitation over Slovenia

[full paper]

Mojca Dolinar
Department for meteorological applications
Slovenian Environment Agency
Ljubljana, Slovenia

Abstract —Gridded time series of climatological variables are an essential factor in all environmental studies where climate variability and climate change impact are under consideration. The time series should be homogeneous, since inhomogeneities mask the real climate signal and thus lead to false conclusions of the studies. In the article the methodology for calculation of the gridded dataset of monthly mean temperature and precipitation in 1 km resolution is presented. The mean monthly value of the treated variable (temperature or precipitation) is decomposed into climate normal and monthly anomaly signal. Spatialisation into 1 km grid is preformed separately for the individual signal. For spatial interpolation of monthly normals all available data (including inhomogenous and incomplete climate data series) were used. On the other hand, only homogenised climate series were used for calculation of gridded monthly anomalies in order to insure temporal homogeneity of the resulting monthly gridded time series. Cross validation of interpolation models showed promising results. However, the results for precipitation grids were slightly better than those for temperature due to a denser network of homogenous precipitation time series.

Keywords —Mean monthly temperature, Monthly precipitation sum, gridded data, spatial interpolation, kriging, homogenised time series, Slovenia

I. Introduction

Gridded climatological data, from normals to monthly or daily datasets, have become increasingly important in the recent past in a variety of different professions like agriculture engineering, ecology, hydrology, etc. [1], [2]. These datasets are expected to provide realistic representation of spatial variability of climate variables for the areas with poor station density such as mountain areas as well [3]. In the same time, for the climate variability and climate change studies, the gridded datasets should be homogeneous in time to avoid misleading results which could be the consequence of artificial signals in the datasets. At the beginning of the previous century the gridded normals (period 1971-2000) for Slovenia were constructed [4]. The calculation of gridded monthly datasets for main climatological variables proved to be a challenge mainly due to high variability in station density during the last decades and considerable inhomogenities in station datasets, which both made the calculation of homogeneous gridded datasets difficult. In 2009 a homogenisation project was launched in Slovenian Environment Agency (ARSO) in the framework of which all major climatic variables datasets for the period of 1961–2011 were homogenised on monthly basis [5], [6]. The homogenised monthly datasets for selected high-quality stations enabled the generation of homogeneous monthly gridded datasets for a range of climatic variables over Slovenia. In the following article the methodology for generation the homogeneous monthly gridded dataset of temperature and precipitation is presented. It takes into account all high–quality measurements to describe high spatial variability of temperature and precipitation. However, when generating monthly datasets, only homogenous station datasets were used.

II. Data

Throughout the history ARSO has been managing two basic types of meteorological stations, defined by their organisation and t extent of measurements [7]:

- Climatological stations with observations at 7a.m., 2 p.m. and 9 p.m. local time
- Precipitation stations with observations at 7a.m.,

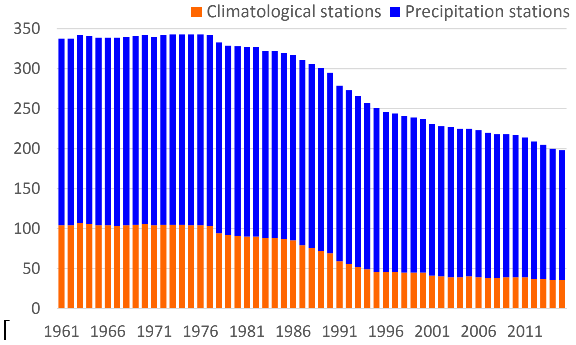


Fig. 1. Temporal variability of number of stations measuring temperature (Climatological stations – orange) and precipitation (Climatological and Precipitation stations – orange and blue) from 1961 to 2014.

Temperature is measured on climatological stations only, while precipitation is measured on both climatological and precipitation stations. Although temperature and precipitation have been continuously measured on Slovenian territory from the year 1850 on, the measurements have been systematically digitised from 1961. For the period before 1961 only selected time series are in digital form. Throughout the history the station network was subject to many changes, resulting in very high variability of station density on a year to year basis (Figure 1). There was a significant decrease in number of both station types in 1975 and the decreasing trend continues into present day.

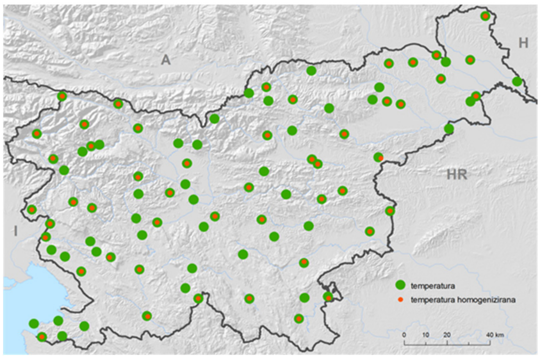


Fig. 2. Spatial distribution of stations with temperature datasets. Larger green circle indicates the stations, which temperature datasets were used for spatial interpolation of monthly normals. Smaller orange circle indicate stations with high quality temperature datasets which were homogenised. These datasets were used for spatial interpolation of monthly temperature anomalies.

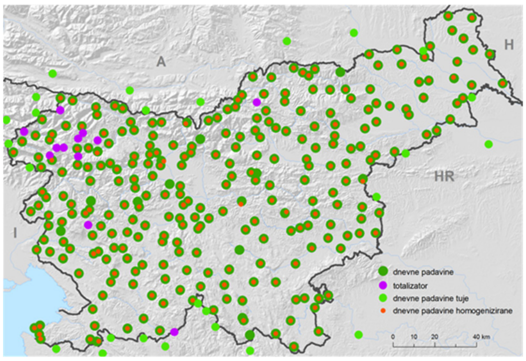


Fig. 3. Spatial distribution of stations with precipitation datasets. Larger dark green circles indicate Slovenian and light green circles indicate foreign (Italian, Austrian and Croatian) stations from which precipitation datasets were used for spatial interpolation of monthly normals. Purple circles indicate locations of totalisers. Smaller orange circles indicate stations with high quality precipitation datasets, which were homogenised. These datasets were used for spatial interpolation of monthly precipitation anomalies.

The highest number of station measurements in digital form was limited to the period of 1961-1975, when there have been more than 100 stations carrying out temperature measurements and close to 350 stations performing precipitation measurements (Figure 1). Slowly decreasing in number from 1975 on, there have only been 75 stations performing temperature measurements and 298 taking precipitation measurements at the end of 1990. This was the reason that 30-year reference period 1961-1990 was selected for calculation of gridded monthly normals. For spatial interpolation of monthly normals it is important to have as much measurements as possible to describe the high spatial variability of variable (temperature or precipitation).That is the reason that all datasets with more than 15 years of high quality measurements were used for calculation of gridded monthly normals. Missing data (on monthly basis) were calculated using estimated correlation with closest representative station for the period with no missing data. For gridding procedure of precipitation normals data from totalisers were used as well. With those measurements high precipitation variability in mountain region could be assessed, all the while few other measurements were available in that region (Figure 4). The seasonal amount of precipitation (from September until September next year) measured by totalisers was allocated to each individual month taking into account the monthly precipitation course of neighbouring precipitation stations. Additionally, precipitation measurements from near-border stations from Italy, Austria and Croatia were available for the reference period 1961-1990. Finally, there were 89 station datasets available for the spatial interpolation of monthly temperature normals (Figure 2) and 319 station datasets available for interpolation of monthly precipitation normals (Figure 3).

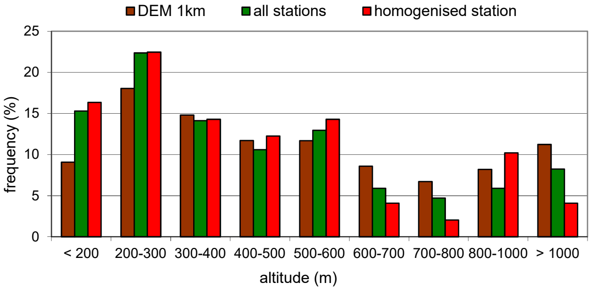


Fig. 4. The relative frequency distribution of stations with temperature measurements (green column – all stations, orange column – stations with homogenised datasets) according to the altitude in comparison with relative frequency distribution of Slovenian terrain in 1 km resolution (brown column).

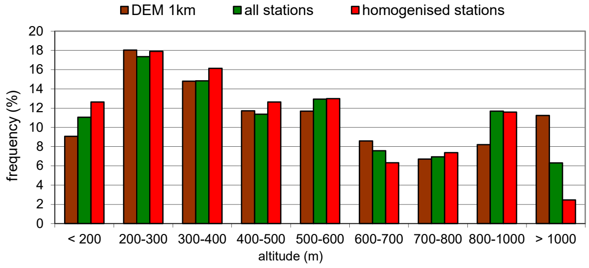


Fig. 5. The relative frequency distribution of stations with precipitation measurements (green column – all stations, orange column – stations with homogenised datasets) according to the altitude in comparison with relative frequency distribution of Slovenian terrain in 1 km resolution (brown column).

Gridded monthly anomalies of both temperature and precipitation were based on homogenised monthly datasets. In the process of homogenistaion the artificial signals (caused by changes in the surroundings of the stations, changes of measuring equipment, station replacements, etc.) were removed from monthly datasets (period 1961-2011) [5], [6]. For the period after 2011, the datasets were completed simply by adding monthly measurements, while in the process of homogenization the datasets were adapted according to the last period [5]. In the homogenization process, the quality of some datasets was found to be inadequate for homogenization. Therefore, they were removed from the homogenisation process and from any further analysis. Thus, the spatial density of available homogenized datasets was much smaller than the density of stations used in the first step (calculation of monthly normals). For gridding procedure of temperature anomalies there were 49 stations available. Spatial distribution of those stations is represented in Figure 2 and their distribution in relation to their elevation on Figure 3. For spatial interpolation of monthly precipitation anomalies, much more station datasets were available (267). Their spatial distribution is represented in Figure 3 and their distribution in relation to their elevation in Figure 5.

III. Methods

The procedure for calculation of monthly grids was the same for both temperature and precipitation. The gridding process for an individual variable differs in the quantity of input datasets only, which was much larger for precipitation than that of temperature.

The methodology is based on the decomposition of monthly value into two signals:

- 30-years monthly normal value
- Monthly anomaly,

Spatial interpolation in a regular grid is performed for each signal separately and final monthly grids are the composition of both signals. The gridding procedure is schematically described on Figure 6.

A. *Spatial interpolation of monthly normals*

To describe as much of very high spatial variability of monthly normals as possible, dense station datasets are needed along with adequate distribution by altitude. In this stage of gridding process as much input data as available was used. By definition, normal value is the signal, averaged over time. By averaging, minor errors and temporal inhomogenities are filtered out from the dataset. To increase the input data density, for spatial interpolation of monthly normals lower quality datasets and datasets with minor inhomogenities were used as well.

For spatial interpolation of monthly normals Regression-Kriging was applied, explained by [8] and [9]. Regression kriging uses correlation with multiple environmental predictors (explanatory variables) through regression and spatial autocorrelation of the targeted variable through kriging. It is frequently used in climatology [10], [11], [12].

Explanatory variables in regression kriging were chosen for each variable and month separately, according to multivariat regression analysis results. The statistically significant explanatory variables were selected among geographical variables: longitude, latitude, altitude and their second-degree polynomial terms. Regression residuals were interpolated using ordinary kriging method [13], [14], [15].The uncertainty of the spatial interpolation predictions was evaluated using leave-one-out cross-validation (LOOCV) approach [13], [11]. Detailed explanation of geostatistical methods (regression kriging) and validation methodology (cross-validation) is described in [13] and [14].

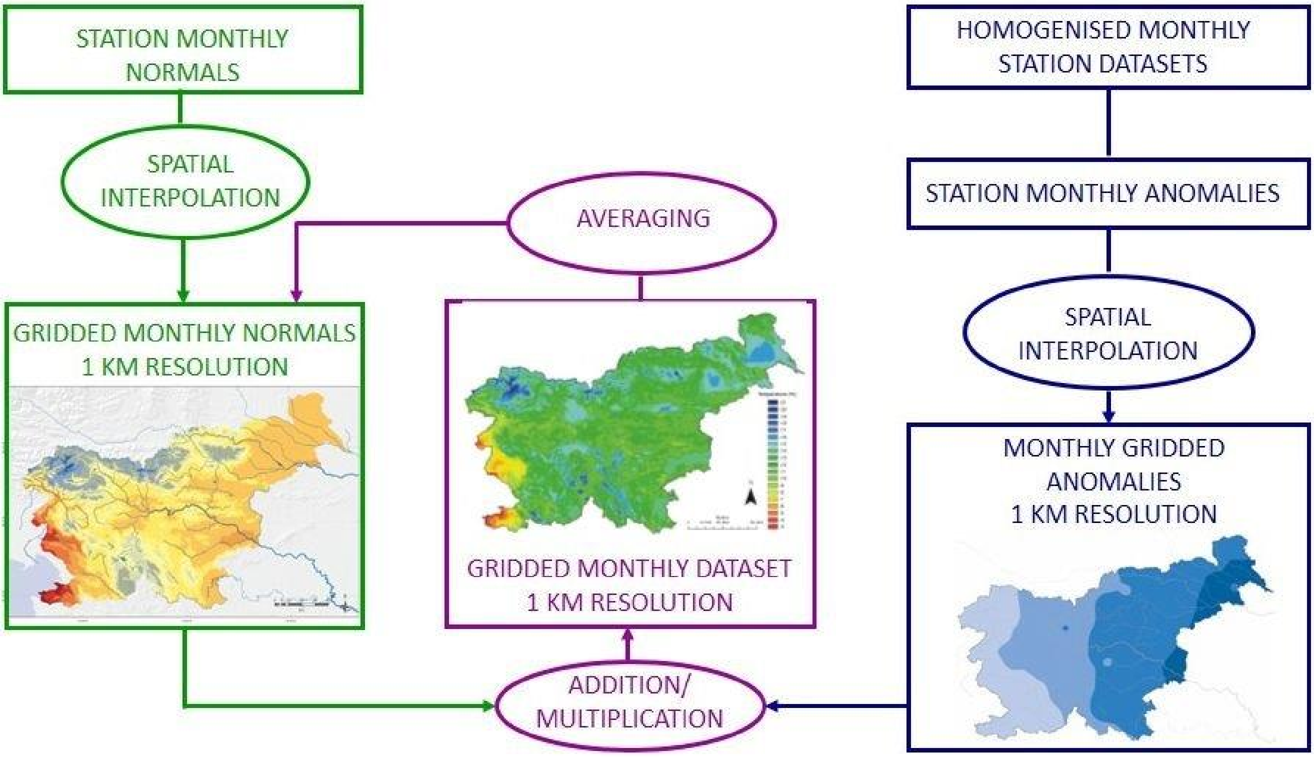

STATION MONTHLY NORMALS
SPATIAL INTERPOLATION
AVERAGING
HOMOGENISED MONTHLY STATION DATASETS
STATION MONTHLY ANOMALIES
SPATIAL INTERPOLATION
GRIDDED MONTHLY NORMALS 1 KM RESOLUTION
GRIDDED MONTHLY DATASET 1 KM RESOLUTION
MONTHLY GRIDDED ANOMALIES 1 KM RESOLUTION
ADDITION/ MULTIPLICATION

Fig. 6. Computational framework for production of monthly grided datasets for temperature and precipitation.

B. *Spatial interpolation of monthly anomalies*

Monthly temperature and precipitation anomalies are more spatially coherent than the absolute values (Jones and Trewin 2000). Therefore the spatial density of monthly station datasets could be smaller than the data density for spatial interpolation of absolute monthly values or monthly normal values. On the other hand, in order to provide temporal homogeneity of monthly grids, the input datasets for anomaly grids calculation should be homogenous in time. Therefore, only homogenous station datasets were used in this stage of gridding procedure.

For spatial interpolation of monthly anomalies the same methodology (regression kriging) as of that for spatial interpolation of monthly normals was applied. As the basic characteristic of monthly anomailies is their high temporal variability, the interpolation model could not be constant throughout the months and years. It should be developed out of the spatial characteristics for each single month separately. Out of an apriory set of explanatory variables (longitude, latitude, altitude and their second-degree polynomial terms) the statistically significant set of variables was chosen according to multivariate regression analysis results. Regression residuals were interpolated using ordinary kriging method [13], [14], [15]. The uncertainty of the spatial interpolation predictions was evaluated using leave-one-out cross-validation approach [11] and [13].

C. *Calcualtion of monthly absolute values*

After spatial interpolation of monthly normals and monthly anomalies, the two signals were added up to get absolute values for every single month. Monthly values were than averaged through the period 1961-1990 to get a new set of monthly normals, which slightly differed from the first normals, calculated in the first step, due to different set of station datasets used in anomaly calculation. With these updated normals the composition of signals was repeated to get final grids of monthly absolute values.

IV. Results

A. *Monthly temperature normals*

Validation results of spatial interpolation of monthly temperature normals are presented in Table 1. High percentage of spatial variability for all months is explained by explanatory variables in the deterministic part of the interpolation model. Coefficient of determination (R^2) in regression analysis was the lowest in February (79 %). For all other months R^2 was higher, with the highest values in summer months (91 and 92 %). In all cases altitude was one of the explanatory variables which statistically significant explained the spatial variability of monthly temperature normal. The correlation coefficient between predicted and measured values on station level in LOOCV procedure were higher than 0.89 (January) with highest values 0.98 in May.

TABLE I. Validation results of spatial interpolation of monthly temperature normals (period 1961-1990). For every month statistically significant explanatory variables are presented in addition to coefficient of determination (R2) of regression analysis and the correlation coefficient (R) between predicted and measured values on station level in LOOCV procedure.

Month	Regression analysis		Cross validation
	Explanatory variables	***R^2 (%)***	***R***
Jan	x ,x^2, y^2, z	81	0.89
Feb	x, x^2, xy, z	79	0.90
Mar	x, y ,x^2, y^2, z	84	0.92
Apr	x, y ,x^2, y^2,xy, z	88	0.95
May	x, y ,x^2, y^2,xy, z	90	0.98
Jun	x, y ,x^2, y^2,xy, z	92	0.96
Jul	x, y ,x^2, y^2,xy, z	91	0.97
Aug	x, y ,x^2, y^2,xy, z	92	0.96
Sep	x, x^2, y^2,xy, z	90	0.95
Oct	x, y ,x^2, y^2, z	86	0.94
Nov	x, x^2, y^2,xy, z	88	0.90
Dec	x, x^2, y^2, z	83	0.91

B. *Monthly precipitation normals*

Validation results of spatial interpolation of monthly precipitation normals are presented in Table 2. The portion of spatial variability explained by explanatory variables in the deterministic part of the interpolation model is a bit lower than in the case of temperature, yet it is still high. The lowest R^2 was in February (59 %), same as in the case of the temperature. The highest value of R^2 was 69 % (April). Altitude, longitude and latitude and/ their second-degree polynomial terms were among statistically significant explanatory variables in every month. For the colder part of the year relative height above sea level of the nearest geographical barrier in the NE direction (zNE) was an important explanatory variable. The majority of precipitation in the colder season is associated with the orographic effect, when moist air inflow from SW hits the Alpine-Dinaric barrier, which crosses Slovenia from NW to NS. As expected, the cross-validation results for precipitation are worse than those for temperature, which could be attributed to much higher precipitation variability. The lowest correlation coefficient between predicted and measured values is in the case of August, when very high precipitation variability is associated with the convective type.

TABLE II. VALIDATION RESULTS OF SPATIAL INTERPOLATION OF MONTHLY PRECIPITATION NORMALS (PERIOD 1961-1990). FOR EVERY MONTH STATISTICALLY SIGNIFICANT EXPLANATORY VARIABLES ARE PRESENTED IN ADDITION TO THE COEFFICIENT OF DETERMINATION (R2) OF REGRESSION ANALYSIS AND THE CORRELATION COEFFICIENT (R) BETWEEN PREDICTED AND MEASURED VALUES ON STATION LEVEL IN LOOCV PROCEDURE.

Month	Regression analysis		Cross validation
	Explanatory variables	***R^2 (%)***	***R***
Jan	x , y, x^2, zNE, z	68	0.89
Feb	x, y, x^2, xy, z	59	0.82
Mar	x, y ,x^2, xy, zNE, z	60	0.85
Apr	x, y ,x^2, y^2,xy, z	69	0.91
May	x, y ,x^2, y^2,xy, z	68	0.90
Jun	x, y ,x^2, y^2,xy, z	64	0.86
Jul	y ,x^2, y^2,xy, z	68	0.84
Aug	x, y ,x^2, y^2,xy, z	60	0.78
Sep	x, x^2, zNE, z	65	0.91
Oct	x, y ,x^2, xy, zNE z	67	0.91
Nov	x, y, x^2, y^2,xy, zNE, z	65	0.85
Dec	x, y, x^2, xy, zNE, z	61	0.92

C. *Monthly temperature anomalies*

Validation results of spatial interpolation of monthly temperature anomalies are presented on Figures 7 and 8. The variability of cross-validation results is very high and could be attributed to high temporal variability of anomalies and relatively small number of station datasets (49). In majority of cases a substantial portion of spatial variability of temperature anomalies could be explained by explanatory variables. However, in March and June there were cases, where none of explanatory variables statistically significant explained the spatial variability (Figure 7). The correlation coefficients between predicted and measured values on station level in LOOCV procedure also presented high variability (Figure 8). In all months except January the lowest detected correlation coefficient was 0.20. On average correlation coefficients were higher than 0.50, except for April (0.39) and June (0.40).

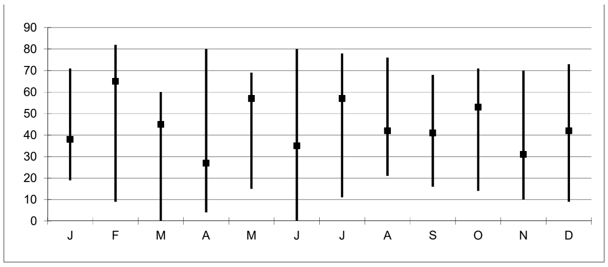


Fig. 7. The distribution of the coefficient of determination (R^2) of the deterministic part of interpolation model for monthly temperature anomalies. The distribution is presented on monthly level and represents the results for 54 interpolation models (from 1961 to 2014). Square sign indicates the median of all 54 R^2 for the selected month, while the straight line links both extreme values of R^2.

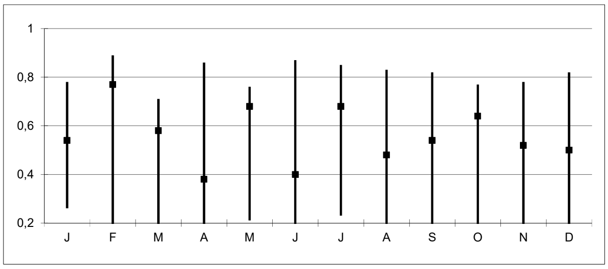


Fig. 8. The distribution of the correlation coefficient (R) between predicted and measured values on station level in LOOCV procedure for monthly temperature anomalies. The distribution is presented on monthly level and represents the results for 54 interpolation models (from 1961 to 2014). Square sign indicates the median of all 54 R for the selected month, while the straight line links both extreme values of R.

D. *Monthly precipitation anomalies*

Validation results of spatial interpolation of precipitation anomalies are presented on Figures 9 and 10. The variability of cross-validation results is high, however, mostly due to higher input data density, the results are relatively better than for temperature anomalies. In the colder part of the year the correlation between precipitation anomalies and explanatory variables is higher (Figure 9), as well as correlation between measured and predicted values (Figure 10). On the contrary, in the warmer half of the year, especially in summer months, the results are more variable. In some cases there was no correlation between precipitation anomalies and explanatory variables whatsoever and on average, the portion of explained spatial variance by explanatory in the warm months was lower than in the cold months (Figure 9). The correlation coefficients between predicted and measured values on station level in LOOCV procedure were much higher and less variable than in the case of temperature (Figure 10). In all except summer months mean correlation coefficient was above 0.80. Lower correlation coefficient as well as lower correlation with explanatory variables in summer months could be attributed to higher variability of convective precipitation, which prevails in the summertime.

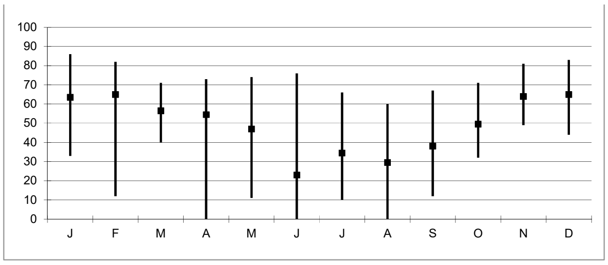


Fig. 9. The distribution of the coefficient of determination (R^2) of the deterministic part of interpolation model for monthly precipitation anomalies. The distribution is presented on monthly level and represents the results for 54 interpolation models (from 1961 to 2014). Square sign indicates the median of all 54 R^2 for the selected month, while the straight line links both extreme values of R^2.

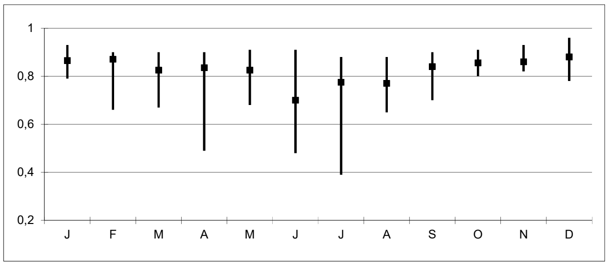


Fig. 10. The distribution of the correlation coefficient (R) between predicted and measured values on station level in LOOCV procedure for monthly precipitation anomalies. The distribution is presented on monthly level and represents the results for 54 interpolation models (from 1961 to 2014). Square sign indicates the median of all 54 R for the selected month, while the straight line links both extreme values of R.

V. Conclusions

With the described procedure, gridded datasets of monthly temperature and precipitation were generated from 1961 on with 1 km spatial resolution. In addition, as a parallel product of the procedure, monthly anomalies grids were generated for that same period. The grids are updated on monthly basis with a one-month lag due to data collection and the data quality control procedure.

With the approach described in the paper, all available data were used to describe very high spatial variability of both variables. On the other hand, with signal separation and different datasets used for spatial interpolation of each individual signal, the gridded dataset temporal homogeneity was preserved, which is the greatest advantage of the generated gridded dataset of temperature.

References

[1] C. Daly, W.P. Gibson, G.H. Taylor, G.L. Johnson, P. Pasteris, »A knowledge-based approach to the statistical mapping of climate«, Clim. Res. vol. 22, pp. 99-113, 2002

[2] C. Daley, »Guidelines for assessing the suitability of spatial climate data sets«, Int. J. Climatol., vol. 26, 707–721, 2006

[3] C. Daly, M. Halbleib, J.I. Smith, W.P. Gibson, M.K. Doggett, G.H. Taylor, J. Curtis, P.P. Pasteris, »Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States«, Int. J. Climatol. vol. 28, pp. 2031–2064, 2008

[4] Dolinar M., ed., Climate of Slovenia (1971-2000), Slovenian Environment Agency, Slovenia,.2006

[5] G. Vertačnik, Podnebna spremenljivost Slovenije v obdobju 1961–2011: 2 Kontrola in homogenizacija podatkov, Agencija RS za okolje, Slovenija, 2016

[6] G. Vertačnik, M. Dolinar, R. Bertalanič, M. Klančar, D. Dvoršek, M. Nadbath, »Ensemble homogenization of Slovenian monthly air temperature series«, Int. J. Climatol., vol. 35 (13), pp. 4015–4026, 2015

[7] M. Nadbath, Podnebna spremenljivost Slovenije: Meteorološka opazovanja I, Agencija Republike Slovenije za okolje, Slovenija, 2015

[8] E. Pebesma, The role of external variables and GIS databases in geostatistical analysis, T GIS vol. 10 (4), pp. 615-632, 2006

[9] T. Hengel, A Practical guide to Geostatistical Mapping of Environmental Variables, JRC Scientific and Tehnical Report, Luxemburg, 2007

[10] H. Wackernagel, Multivariate geostatistics: an introduction with applications, Springer, New York, 2003

[11] O.E. Tveito (ed), The use of geographic information systems in climatology and meteorology : COST action 719. R. Beratalanič, Z. Bihari, H. Dobesch, M. Dolinar, Spatialisation of climatological and meteorological information with the support of GIS (Working Group 2). Luksemburg: Office for official publ. of the EC, pp. 36-163, 2007

[12] M. Perčec-Tadić, "Gridded Croatian climatology for 1961-1990", Theor. Appl. Climatol., vol. 102, pp. 87-103, 2010

[13] N. A. C. Cressie,, Statistics for Spatial Data. John Wiley &Sons, USA, 1993.

[14] Isaaks, E. H. in Srivastava R. M. 1989: An Introduction to Applied Geostatistics. Oxford University Press, USA

[15] R. Daley, Atmospheric Data Analysis. Cambridge University Press, Canada, 1991.

[16] D.A. Jones, B. Trewin, "The spatial structure of monthly temperature anomalies over Australia", Aust. Met. Mag. vol 49, 2061-276, 2000

Combining observations with model data for improved high resolution temperature interpolation

[full paper]

M. Dirksen, MSc
KNMI
De Bilt, The Nederlands
marieke.dirksen@knmi.nl

Dr. R. Sluiter
KNMI
De Bilt, The Netherlands
raymond.sluiter@knmi.nl

Abstract— **High-resolution meteorological data is beneficial for end users in the fields of safety, environment and agriculture. The goal of this research is to obtain high-resolution temperature climatology based on integrated in-situ observations and Numerical Weather Prediction (NWP) models. We investigated high resolution temperature interpolation for the Netherlands based on 33 observation points. As the number of observations is rather limited and they are not optimally distributed, interpolating is a difficult task. Large scale temperature patterns in the Netherlands are influenced by a.o. the sea. In summer temperatures near the coast are generally lower compared to inland values, while in winter the opposite is true. Previous investigations included this sea effect by using Universal Kriging, in combination with the distance to the sea as ancillary data. However, the distance to the sea is a time independent variable. The interpolation can be further improved in space and time by including the effects of large water bodies (e.g. IJsselmeer and Westerschelde), wind (direction and speed) and the temperature gradient between sea and land which are included already in NWP models. By combining the temperature observations with NWP model data from RACMO and HARMONIE an improved high-resolution temperature interpolation is obtained. The model output was combined in high-resolution temperature grids (1x1km). Several interpolation techniques, e.g. Universal Kriging with ancillary datasets, Ordinary Kriging, Thin Plate Spline and Inverse Distance Weighting, were compared. Universal Kriging interpolation with NWP model data as ancillary data has improved previous methods both statistically and spatially and performed better than solely NWP model data. For long term averages R^2 values significantly improved from below 0.50 up to 0.75. Small scale features like lakes and cities can be recognized.**

Keywords— High Resolution, Interpolation, Temperature observation, Kriging, NWP model, HARMONIE, RACMO

I. Introduction

An estimate of the temperature in between observations allows studies in regions without observations. By averaging interpolated data over an area, area's with low and high measurement densities can be compared. Also, climate variability and climate change studies often require regular spaced data. Another important application of interpolated data is the validation of models [1].

Large scale temperature patterns in the Netherlands are influenced by the sea. A pressure gradient, induced by differential heating between the land and sea causes the formation of a circulation cell. This has a major impact on the meteorological conditions near the sea. The resulting sea breeze turns, on the northern hemisphere, towards the right, e.g. the Coriolis acceleration. The see breeze is a.o. influenced by the sea temperature, land temperature, wind direction and wind speed [2]. Up to now, the distance to the shoreline is used as a (simple) explanatory variable for temperature gradients [3]. Interpolating with the distance to the sea as ancillary data is thus a simplification of the complicated sea breeze process. The studies of [4], [5] and [6] show that also other parameters, such as height, can be important for the spatial temperature patterns. A way to include all these variables is the use of Numerical Weather Prediction (NWP) models. The model outcome is expected to improve the temperature interpolation. This research uses NWP model output from RACMO (12km grid size) and HARMONIE (2.5km grid size). Output from a 20 year period, from 2 January 1995 until 29 July 2014, was used. The goal of this research is to obtain high-resolution temperature climatology based on integrating in-situ observations and NWP models. To reach the research goal various verification methods for minimum, mean and maximum temperature for different time scales, using different ancillary datasets, are used. The verification methods include statistical analysis of cross validation results, differences between observations and models and a comparison between variogram fits.

II. Data and Methods

A. Datasets

For the interpolation and validation a dataset of almost 20 years, from 2 January 1995 until 29 July 2014, was used. Various time steps (day, month, year, 20 years and 20 year months) were compared. Also a 30 years period with 20 years of ancillary data was examined. As input data mean, maximum and minimum temperature observations were used, either with ancillary data (distance to the sea, RACMO or HARMONIE) or without ancillary data. For the HARMONIE data only the mean temperature is available. Several interpolation techniques were compared: Ordinary Kriging, Universal Kriging, Thin Plate Spline (TPS) and Inverse Distance Weighted (IDW) interpolation. The interpolation techniques and model input will be validated by means of cross validation and point differences.

B. Temperature observations

The temperature data used for interpolation originates from the 'Klimaat Informatie Systeem' (KIS). The database includes 33 sampling locations, since 1990 (Fig. 1). The sampling locations are representative and suitable for interpolation of temperature data. The distribution of the sampling locations is slightly better for the west coast, were the population density is highest [3]. Air temperature measurements at an altitude of 1.5m were used. For the long term interpolation of 20 years the stations Berkhout and Ell were not considered. Both Berkhout and Ell started measuring in 1999.

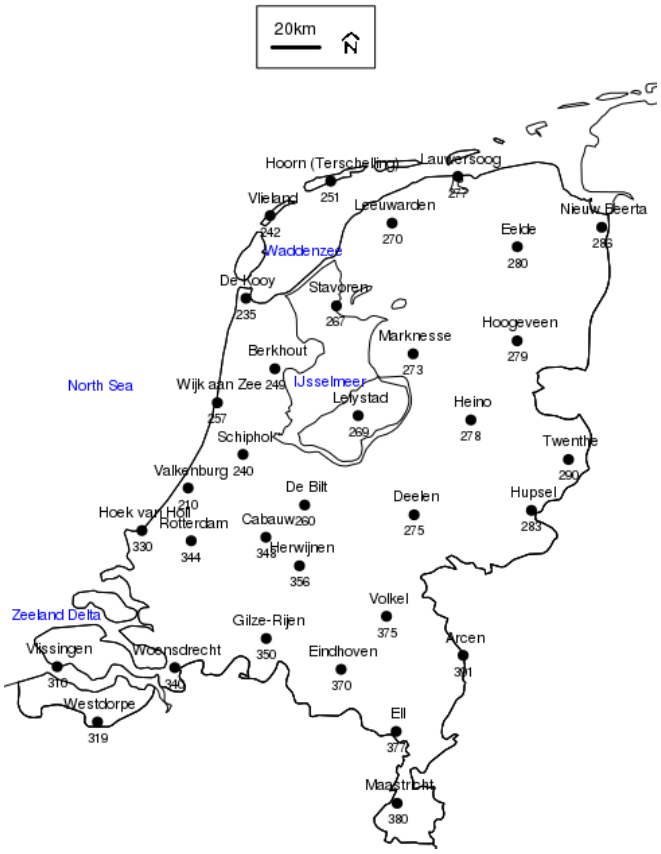


Fig. 1. Overview of the names and numbers of the KNMI measurement stations in the Netherlands. Large water bodies and the sea are indicated in blue

C. Ancillary data

Large scale temperature patterns can be explained with the distance to the sea. Two previous researches used his variable during interpolation [3,7]. The distance to the sea is a digitalized virtual shoreline with values >0, enabling logarithmic transformations. According to [3] a logarithmic distance to the sea algorithm preforms best. The distance to the sea and logarithmic distance to the sea will be compared.

The model output from [8] was used. RACMO calculates a 36 hour forecast from which only the last 24 hours are used. This is due to the adaptation time for the smaller resolution. The grid size of the model is 12 km. The NWP model output, stored in NetCDF files, was used as ancillary data for the interpolation. The RACMO run from 2 January 1995 until 29 July 2014, with minimum, mean and maximum daily temperatures was used. Details on the model can be found in [9].

The model output from [10] was used for this research. HARMONIE has a 7 hour forecast with a spin up of 1 hour. The grid size of the model is 2.5 km. The HARMONIE run stores hourly NetCDF files. The hourly NetCDF files are combined into daily NetCDF files. A HARMONIE run with mean temperature from 2 January 1995 until 31 December 2013 was used.

The model output from RACMO and HARMONIE, was converted to grid files using an online server [11,12]. The projection of the NetCDF files was corrected, also the coordinate system, bounding box coordinates and grid size were adjusted. The resampling method for the projection and grid size is nearest neighbour. The Dutch RD-coordinate system was used (with EPSG projection number: 28992). The bounding box coordinates were set to the boundaries of the Netherlands. The grid size for both NWP models is adjusted to 1x1km.

D. Validation

Different validation methods were used to validate the models (point difference) and determine the best interpolation method (cross validation and variogram fits). Several variogram models were compared: spherical, exponential and Gaussian. RACMO and HARMONIE temperature grids were compared with the observations. For the interpolation with the grids the temperature patterns is of importance, thus not the absolute values; though absolute values are important for model validation. For large datasets Leaf One Out Cross Validation is recommended, e.g. one data point is left out and predicted. From the difference between predictions and observations statistics can be calculated. For the quality of the fitted trend R^2 was calculated, the actual performance was calculated using the root mean square error standard deviation (RMSEsd). Also the root mean square error (RMSE) and scaled mean error (MEmean) were calculated, according to [7].

E. Interpolation methods

The focus of this research is Kriging interpolation. Several settings for Kriging can be used: with or without ancillary data and using different variogram models. Ordinary Kriging interpolates without ancillary data. Universal Kriging interpolates with ancillary data: distance to the sea, logarithmic distance to the sea and, RACMO and HARMONIE temperature grids. Besides Kriging also TPS and IDW were examined. A detailed explanation of the interpolation methods can be found in [3,7].

III. RESULTS

A. NWP models compared with measurements

The temperature observations were on average 0.26℃ higher than the predicted temperature by RACMO. Not all areas are underestimated: the western Waddenzee area is slightly over predicted. It can be concluded that over a time period of 20 years the mean temperature is accurately predicted and representative. HARMONIE predicted temperature values are more accurate. On average observations were 0.09°C lower than predicted values.

B. *Statistical analysis*

TABLE 1 shows the statistical analysis for the 20 year period. For this time period most of the interpolation methods performed good, except for the spherical variogram model. For this time period the statistical differences between the currently used interpolation (TPS) and interpolations with NWP model data are small.

For the minimum and maximum temperature only ancillary data from RACMO is available. For the 20 year period minimum and maximum temperature larger differences between TPS and KEDexpR are found. For the minimum temperature interpolation R^2 values improved from below 0.4 up to 0.66 and RMSE values improves from 0.6 to 0.01. For the maximum temperature R^2 values improved from 0.78 up to 0.84 and RMSE values improves from 0.12 to 0.03. Similar results were found for the 30 year period minimum and maximum temperature.

C. *Interpolated temperature maps*

Interpolated temperature maps were compared for the different time periods: day, month, year, 20 years, 20 year months and 30 years.

Results are shown in Fig. 2. For 20 year period interpolation Universal Kriging with an exponential model is chosen, based on statistics and variogram fits. Differences are found depending on the ancillary data, i.e. logarithmic distance to the sea, temperature grids from RACMO and temperature grids from HARMONIE. A spatial temperature pattern with increasing temperature from north-east to south-west is found when using the logarithmic distance to the sea. The spatial patterns resulting from an interpolation with temperature grids from RACMO show this same trend, but several differences can be observed. Here, lower temperatures in the southeast (Maastricht) and in the central part of the Netherlands, higher temperatures are observed in the southwest (Zeeland delta) and around the lake IJsselmeer. Large spatial differences are found for interpolations with temperature grids from HARMONIE and the other interpolations. This temperature pattern includes cities, which can be recognized by an elevated temperature; also different vegetation types are represented.

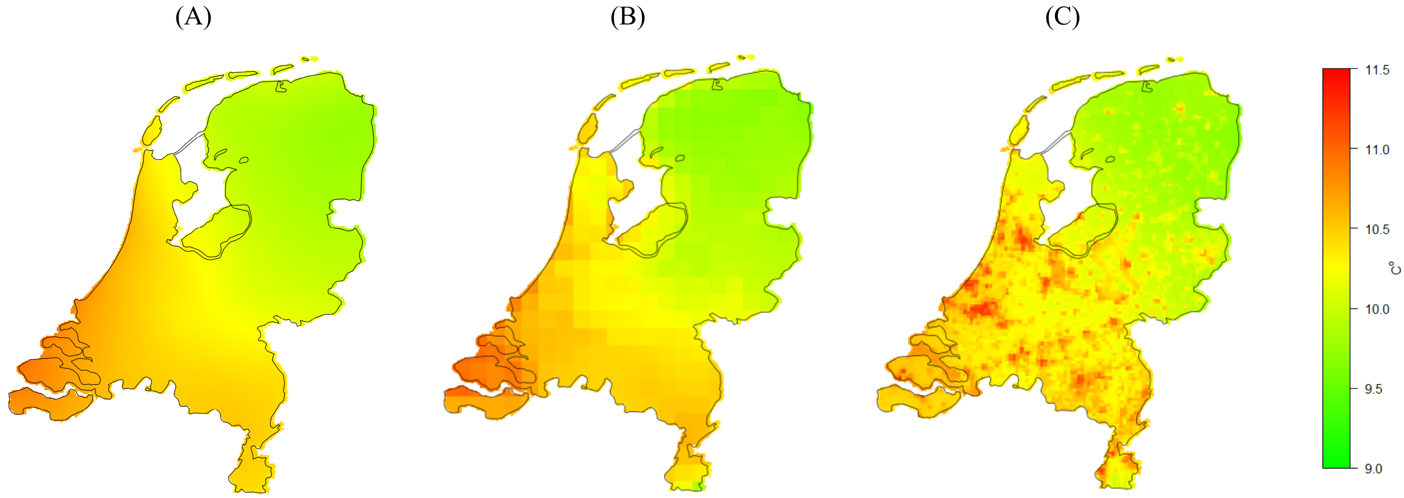


Fig. 2. Overview of the Universal Kriging interpolation with exponential variogram model (A) LOG distance to the sea as ancillary data (B) RACMO temperature grid as ancillary data (C) HARMONIE temperature grid as ancillary data. The areas with large cities, such as Amsterdam and Rotterdam, show higher temperatures.

TABLE I. EXAMPLE OF THE STATISTICAL ANALYSIS OF THE 20 YEAR PERIOD.

Interpolation method	Statistical Analysis			
	R^2	***RMSE***	***RMSEsd***	***MEmean***
Okexp[a.]	0.746	0.004	0.010	0.000
Oksph[b]	0.013	0.070	0.191	0.001
IDW	0.763	0.001	0.002	0.000
TPS	0.805	0.009	0.025	0.000
KEDexpS[c]	0.725	0.001	0.002	0.000
KEDsphS[d]	-0.020	0.064	0.174	0.001
KEDgauS[e]	0.781	0.011	0.031	0.000
KEDexpLS[f]	0.755	0.012	0.032	0.000
KEDsphLS[g]	0.015	0.079	0.215	0.001
KEDgauLS[h]	0.784	0.023	0.064	0.000
KEDexpR[i]	0.787	0.037	0.100	0.001
KEDsphR[j]	0.695	0.042	0.115	0.001
KEDgauR[k]	0.790	0.035	0.097	0.001
KEDexpH[l]	0.787	0.150	0.239	0.002
KEDsphH[m]	0.078	0.187	0.298	0.002
KEDgauH[n]	0.778	0.095	0.151	0.001

[a.] Ordinary Kriging with exponential variogram model. [b.] Ordinary Kriging with spherical variogram model. [c.] Universal Kriging with exponential variogram model and distance to the sea. [d.] Universal Kriging with spherical variogram model and distance to the sea. [e.] Universal Kriging with Gaussian variogram model and distance to the sea. [f.] Universal Kriging with exponential variogram model and logarithmic distance to the sea. [g.] Universal Kriging with spherical variogram model and logarithmic distance to the sea. [h.] Universal Kriging with Gaussian variogram model and logarithmic distance to the sea. [i.] Universal Kriging with exponential variogram model and RACMO as ancillary data. [j.] Universal Kriging with spherical variogram model and RACMO as ancillary data. [k.] Universal Kriging with Gaussian variogram model and RACMO as ancillary data. [l.] Universal Kriging with exponential variogram model and HARMONIE as ancillary data. [m.] Universal Kriging with spherical variogram model and HARMONIE as ancillary data. [n.] Universal Kriging with Gaussian variogram model and HARMONIE as ancillary data.

IV. DISCUSSION

The station of Vlissingen, in the southwest of the Netherlands, is surrounded by water bodies and is not representative for the inlands of Zeeland islands. The measurements of the station Rotterdam are influenced by the city, the city effect is mainly noticeable with southwestern winds. The station of Stavoren is located next to the IJsselmeer, a large inland water body. As not all stations are representative for the surrounding area the statistical values should not be interpreted as absolute values, but as an indication of the performance of the interpolation methods.

In general, it can be observed that when the time period decreases R^2 decreases and RMSE increases. Daily values of R^2 and RMSE for TPS are 0.6 and 0.17. For KEDexpR daily values of R^2 and RMSE are respectively 0.65 and 0.09. KEDexpH has R^2 and RMSE of 0.65 and 0.07. Not only based on the statistical analysis the interpolation has improved. The daily temperature pattern is better captured by the model grids.

Mainly for shorter time periods and extreme temperatures the interpolation methods with RACMO and HARMONIE differ. The input of the RACMO model influences the temperature grids and thereby also the interpolations. The land/sea mask grid is too large for the Dutch islands. As a consequence the cross validation of the observation point Hoorn (Terschelling) is more than 0.5°C higher than observed. For other coastal areas the predictions are better. The temperature differences between the temperature grids from HARMONIE and the observations are smaller.

V. CONCLUSION

The goal of this research was to obtain high-resolution temperature climatology based on integrated in-situ observations and NWP models. Observations and NWP model data were analysed from 2 January 1995 until 29 July 2014. Spatial patterns from interpolations with mean temperature grids from HARMONIE provide a more detailed pattern. Spatially small scale features like lakes and cities can be recognized. Main differences between interpolation methods are found in the coastal regions and elevated areas. The ancillary data of RACMO and HARMONIE has improved the temperature interpolation. When we compare the two NWP model inputs, interpolations with HARMONIE show a more detailed spatial pattern. Especially for shorter time periods interpolations with ancillary data from HARMONIE improved the interpolation. Based on these results it can be concluded that NWP model data improves the interpolation of high resolution temperature.

ACKNOWLEDGMENT

The authors would like to thank Steven de Jong (Utrecht University) for the constructive criticism and feedback.

REFERENCES

[1] Haylock M.R., Hofstra N., Klein Tank A.M.G., Klok E.J., Jones P.S., New M. (2008) – A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. J. Geophys. Res. 113, D20119.

[2] Gahmberg M., Savijarvi H., Leskinen M. (2009) – The influence of synoptic scale flow on sea breeze induced surface winds and calm zones. Tellus, 62A, 209-217.

[3] Salet F.W.J. (2009) - Het interpoleren van temperatuurgegevens. Master Thesis, WUR. Available online on: http://bibliotheek.knmi.nl/stageverslagen/stageverslag_Salet.pdf

[4] Di Piazza A., Lo Conti F., Viola F., Eccel E., Noto L.V. (2015) – Comparative Analysis of Spatial Interpolation Methods in the Mediterranean Area: Application to Temperature in Sicily. Water 2015, 7, 1866-1888.

[5] Vincente-Serrano S.M., Saz-Sánchez M.A., Guadrat J.M. (2003) – Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. Climate Research Vol. 24: 161-180.

[6] Aalto J., Pirinen P., Heikkinen J., Venäläien A. (2013) – Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models. Theor. Appl. Climatol. 112: 99-111.

[7] Hiemstra P., Sluiter R. (2011) - Interpolation of Makkink evaporation in the Netherlands. De Bilt, TR-327. Available online on: http://bibliotheek.knmi.nl/knmipubTR/TR327.pdf

[8] Van Meijgaard E., Van Ulft L.H., Lenderink G., De Roode S.R., Wipfler L., Broers R., Timmermans R.M.A. (2012) – Refinement and application of a regional atmospheric model for climate scenario calculations of Western Europe. KvR 054/12, ISBN/EAN 978-90-8815-046-3.

[9] Van Meijgaard E., Van Ulft L.H., Van den Berg, Bosveld F.C., Van den Hurk B.J.J.M., Lenderink G., Seibesma A.P. (2008) - The KNMI regional atmospheric climate model RACMO version 2.1. TR-302. Available online on: http://bibliotheek.knmi.nl/knmipubTR/TR302.pdf

[10] Van den Brink H., Baas P., Burgers G. (2013) - Towards an approved model set-up for HARMONIE Contribution to WP 1 of the SBW-HB Wind modelling project.

[11] Climate4Impact (2015) - dev.climate4impact.eu/impactportal/

[12] Déandries C., Pagé C., Branconnot P., Bärring L., Bucchignani E., Som de Cerff W. et al. (2014) – Towards a dedicated impact portal to bridge the gap between the impact an climate communities: Lessons from use cases. Climatic change, V. 125, I. 3, p. 333-347. DOI: 10.1007/s10584-014-1139-7.

Annual precipitation data proccessing and interpolation for the weather stations of Western Ukraine

[full paper]

Alexander Mkrtchian
Department of Geography
Lviv national Ivan Franko university
Lviv, Ukraine
alemkrt@gmail.com

Abstract— **There is a significant applied demand on accurate and precise spatial data on precipitation values distributions. While there are many methods allowing to spatially interpolate point data of meteorological records, the methods based on geostatisics gain momentum for the last decade. The task of climatic data interpolation for less-developed countries is additionally challenged by sparse networks of observation points and discontinuities in data series. The effort has been made to interpolate data on average annual precipitation sums gained through summarizing records of 50 weather stations located in Western Ukraine. Daily data have been downloaded from GHCN database, then preprocessed and summarized in R to obtain average annual precipitation sums for each station. As a first option, data were interpolated by ordinary kriging, using a set of functions from the R gstat package. A second option involved considering the relationships of precipitation with a set of DEM–derived terrain morphometric parameter through multiple regression model. The estimation by leave-one-out cross validation revealed that the second method produced much better accuracy, accounting for more than 90% of initial data variance. It was further revealed that the regression model residuals show no spatial autocorrelation, probably the result of quite large distances between data points.**

Keywords—precipitation; ordinary kriging; multiple regression; Ukraine; R; gstat

I. INTRODUCTION

There is a significant demand for accurate and reliable spatially distributed data on climatic conditions, e.g. for the purposes of landuse and conservation planning, the management of transportation, agriculture, tourism, emergency services and other activities. While climatic maps are a customary component of commonly released thematic atlases, the methods of their creation are often vague and not formally defined, thus the accuracy and reliability of their information is not known. The scales of such climatic maps are usually rather small (e.g. 1:8000000 for the maps in the National atlas of Ukraine [13]), their lack of spatial detail and unknown data accuracy limiting their applied value.

The modern tools for geospatial and statistical data analyses coupled with the availability of digital data sources provide possibilities to develop and apply formal and objective methods of mapping the climatic characteristics. As the primary source of climatic data is the data records of weather stations, the primary formal task connected with the creation of climatic maps is that of the spatial interpolation of point data. The modeling of spatial fields through data interpolation is a common task of geostatistics. The simple geostatistical interpolation procedures like ordinary kriging based on the analysis of the spatial autocorrelation structure of the variable can be refined by considering the relationships of the variable of interest with other exhaustively-sampled explanatory variables available in the area of interest. In the case of climatic characteristics, these variables may correspond to the terrain morphometric parameters that influence the climate through their impact on energy balance and air masses movements.

The precipitation amount is one of the most practically meaningful climatic elements that directly influences the water resources and hydrologic processes, the conditions for agriculture and a number of other activities. The modeling and mapping of its spatial distribution have obvious applied value.

It is only quite recently that the geostatistical methods became commonly applied for the interpolation of climatic and specifically precipitation data. Phillips et al., performing interpolation of precipitation values in mountainous terrain in western Oregon, found that methods which take into account precipitation-elevation relationships, like detrended kriging and cokriging, provide better accuracy and precision compared with ordinary kriging [14]. Yet they notice that precipitation-elevation relationships may weaken at larger scales for areas of complex geography.

In the study of Diodato and Ceccarelli [2] the mapping of the mean annual and monthly precipitation from rainfall observations in a region of southern Italy has been performed using several methods, revealing that linear regression and ordinary cokriging has produced better results compared to the inverse distance interpolation while the best results (indicated by cross-validation) were produced by the multivariate geostatistical methods that utilized elevation data as an auxiliary explanatory variable [2]. In another study concerned with the mapping of monthly precipitation in Great Britain from sparse point data it was shown that kriging with an external drift (informed by elevation data) provided more accurate estimates (judging by the cross-validation RMSE) than either ordinary kriging or deterministic moving window regression [9]. Goovaerts, comparing the performance of the different interpolation techniques for the interpolation of monthly and annual precipitation data for 36 stations located in Southern Portugal, revealed that general linear regression of rainfall versus elevation gave much better predictions than methods which ignore elevation information (like inverse square distance method and ordinary kriging) while the best results were obtained with methods that take into account elevation data while performing geostatistical interpolation [3].

The description of the general principles and theoretical foundations of geostatistical mapping, the software tools for its realization together with some practical examples can be found in [4,5].

Concerning regional efforts at creating detailed spatially distributed climatic datasets, CARPATCLIM project should be mentioned that has been financed by European Commission and carried out by a consortium of institutions from nine countries with Hungarian Meteorological Service as the leading organization. Precipitation sums were among a set of examined variables that after quality-checking and homogenization were interpolated into 10-km resolution grids [15]. The interpolation was implemented with MISH software, that applies AURELHY method developed in 1980-th at the French Meteorological Service. The latter is based on regression kriging in which explanatory variables (predictors) are the principal components derived from a set of elevation differences calculated in local neighborhood, supposed to reflect the orography variability [1]. A similar large effort has been the creation of a high-resolution (1*1 km) monthly temperature dataset for the Greater Alpine region using multilinear regression techniques and regionalization based on data for 1961-1990 period [6]. In comparison with the former case, the predictors used here are justifiable by theoretical considerations of mountain climatology.

The aim of present research has been the creation of accurate and reliable map of annual precipitation distribution with well-documented and reproducible methodology and using open data sources and software.

II. MATHERIALS AND METHODS

A. Study area

The area of interest encompasses the western part of Ukraine with the total area of ~156 thousand sq. km (Fig. 1). It includes the total extent of 8 first rank administrative regions of Ukraine and the partial extent of another two ones. Ukraine is characterized by the relatively low density of weather stations with often discontinuous observation series, presenting additional challenges. The area of interest, while being characterized by diverse terrain conditions (lowland, upland and mountainous) doesn't have marked features of latitudinal zonality characteristic of more eastern parts of Ukraine. 50 weather stations where precipitation is regularly measured on standardized gauges are located inside the bounds of study area.

B. Data acquisition and preprocessing

The precipitation data have been acquired from open-access Global Historical Climatology Network (GHCN) database. The preprocessing of data involved several steps aimed at making data temporary comparable and homogeneous and at summarizing data for the further analyses. The downloaded data consisted of separate data files for each of 50 weather stations, with headers and some redundant columns. R code has been written to remove these and to merge separate data files into one based on common "DATE" column, with columns that correspond to separate stations. Columns were appropriately renamed, and "nodata" codes were adjusted to "NA" R standard. Than the "DATE" column has been split into separate "year", "month", and "day" parts.

The obtained data frame contained 33512 daily precipitation observations spanning from 1924 to 2011. No station however possessed the uninterrupted observation sequence for this time span. Data source contained plenty of "NA" values and periods of present and missing data values for different stations didn't match. To solve the issue, the decision was made to include into analysis only those observation dates for which the data were available for all of the 50 locations. Additionally, the time span was restricted to 1960–1990, when global climatic conditions were changing relatively slowly prior to rush in temperatures observed starting from 1990-th.

The total of 3432 daily observations have thus been selected. Monthly sums and averages were calculated using the combination of *by* and *sapply* R functions. As different months had different numbers of selected observations, this should be accounted for by the appropriate weighting when summarizing monthly data for obtaining annual values. Thus monthly sums were divided by the number of observations pertaining to appropriate month, then multiplied by the number of days in certain month, and then summed up to get annual precipitation values.

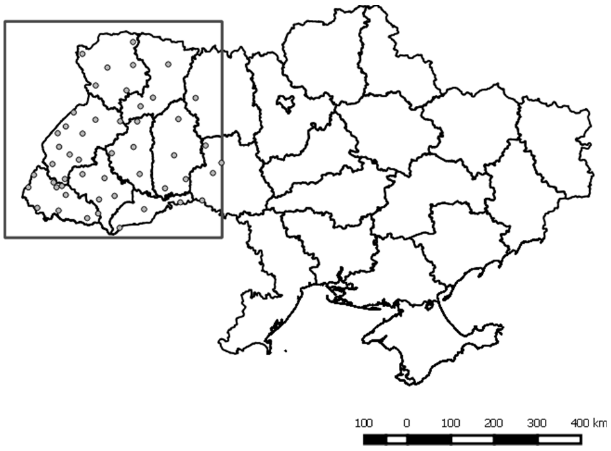


Fig. 1. The location of study area.

The locations of weather stations were mapped onto the shapefile that was imported into R spatial points data frame using rgdal package, and the previously calculated annual precipitation sums were then merged into it.

To prepare the data on terrain morphometric parameters, the 4 SRTM Version 4.1 DEM tiles were downloaded (tiles 41_02, 41_03, 42_02, 42_03) and then merged into one raster [7]. It has then been reprojected into UTM 35N coordinate system (common with station locations shapefile) and resampled to 720 m resolution (this is justified by the low density of stations separated by tens of kilometers). A set of terrain morphometric parameters has been derived from DEM using focal (neighborhood) operators. These regard terrain roughness factor, calculated as a variance of elevation values inside a circular moving window, and aspect factor, calculated as differences in mean elevation values between two opposite circular sectors. Is was hypothesized that increased terrain roughness could positively correlate with precipitation due to increased air flow turbulence that promotes vapor condensation, while the aspect influences precipitation values through well-known rain shadow and orographic precipitation effects. Each of these factors can be calculated on moving windows of different sizes, thus capturing the effects of different-scaled processes, and for the aspect factor different values of angles defining opposite circular sectors and corresponding cardinal points can be specified.

In our previous studies, different factors and scales combinations were tested for the strengths of their relationships with the precipitation sums for two separate years (1961, 1970), with the subsample of weather stations used in present study [10-12]. The terrain roughness factor was thus selected with 7.2 km moving window, while three different versions of aspect factor appeared to be independently significant, namely for the 36 km moving window – the aspect factors NW/SE and W/E, and for the 50.4 km moving window – the aspect factor NW/SE.

The mentioned tasks of DEM data preprocessing and the derivation of morphometric parameters have been accomplished using appropriate SAGA GIS modules and gdal tools under QGIS.

C. Data spatial interpolation

The first part of the analysis was aimed at the interpolation of data using ordinary kriging. This method ignores the terrain morphometric parameters and interpolates data exploiting only the spatial distribution of precipitation values at data points.

The R gstat package has been used for the task, that contains essential commands to perform all the necessary analysis steps. To produce a sample (experimental) semivariogram, *variogram()* function is used, that accepts the cutoff distance and the width of subsequent distance intervals (bins) as parameters. This variogram can be visually inspected with *plot()* command to access its parameters: nugget, sill, and range. The *vgm()* function is then used to specify a theoretical variogram, which requires specifying the values of variogram parameters and the selection of model type. Then *fit.variogram()* function is used to adjust the specified variogram parameters to better fit the data. Lastly, the theoretical variogram model is used as a parameter to *krige()* function that carries out the interpolation. The latter function also requires parameters indicating the interpolated data and the locations of prediction values. The product of *krige()* function is a spatial points data frame containing fields of predicted values (with .pred extension) and of estimated prediction variance (with .var extension). The spatial points data frame object can be converted to raster object of either predicted values or prediction variance (*rasterize* function of raster package) and then to customary raster format like GeoTIFF.

Another very handy gstat function is *krige.cv()* that performs cross-validation for kriging. The leave-one-out cross validation (LOOCV) sequentially visits every data point and predicts the value at that location by leaving out the observed value when performing interpolation, summarizing validation results for every data point. This produces the most reliable accuracy measure.

Another interpolation method applied was a multiple regression of precipitation values on terrain attributes mentioned above. To perform multiple linear regression on spatial data, *krige()* function has been applied with the appropriate input formula and parameter model=NONE. The interpolation methods also entail the conversions between different data structures (rasters, data frames, spatial points data frames) carried out by appropriate R functions. E.g., rasters have to be combined into single spatial points data frame with its data columns corresponding to separate terrain morphometric parameters.

III. RESULTS

In case of ordinary kriging of original precipitation data, a well-pronounced empirical variogram has been produced, with following parameter values (after adjustment with *fit.variogram()*): nugget 30000 mm^2, partial sill 66000 mm^2, range 468 km (Fig. 2). The available theoretical variogram models have been tested by cross-validation, revealing that the best accuracy is obtained by exponential model, following by spherical and Gaussian ones.

The interpolated map, while reflecting general spatial structure of data, lacks spatial detail and produces doubtful results for areas sufficiently remote from nearest data points.

The model of multiple regression of precipitation values on terrain attributes shows pretty good fit (adjusted R-squared 0.9313, F-statistic 133.9 on 5 and 44 DF, p-value of model <2.2e-16). Each of the terrain parameters used as predictors appears to be statistically significant with $p < 0.01$ (Table 1). It appears that the terrain parameter having the strongest impact on precipitation is not elevation ($t = 4.8$) but terrain roughness ($t = 8.3$) which is consistent with our former findings [10-12].

The validity of regression model has been assessed by examining the distribution of residual values, e.g. their normality being checked by Shapiro-Wilk test (*shapiro.test()* R function). Its obtained values ($W = 0.986$, p-value = 0.8125) suggest that regression residuals distribution is not distinguishable from normal.

TABLE I. EXPLANATORY TERRAIN CHARACTERISTICS AND RESPECTIVE REGRESSION MODEL PARAMETERS

Terrain characteristic	**Moving window, km**	**Regression model parameters**		
		Coeff.	t value	Pr(>|t|)
Terrain elevation	-	0.3	4.83	1.7e-05
Terrain roughness	7.2	2.1	8.3	1.5e-10
Aspect factor NW/SE	36	-1.2	-2.75	0.0086
Aspect factor NW/SE	50.4	1.68	4.31	9e-05
Aspect factor W/E	36	0.16	2.81	0.0074

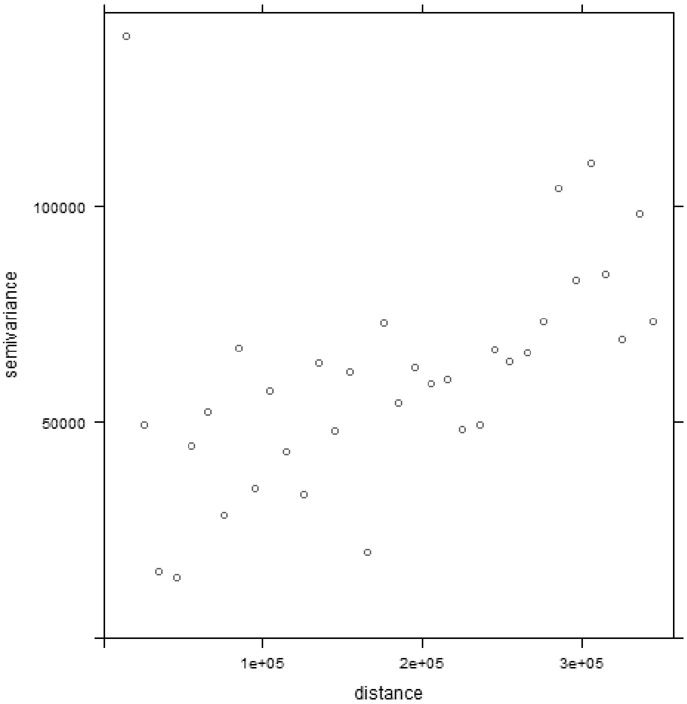


Fig. 2. Empirical variogram of precipitation data

The precipitation map produced by regression model is characterized by much better spatial detail and has good visual appearance (Fig. 4). It is supplemented by another map showing the spatial distribution of the estimated error of the modeled precipitation map (Fig. 5).

The best test of relative accuracy of the two methods is given by cross-validation (Tab. 2). It appears that 36,2% of initial variance of values has been retained after the interpolation by ordinary kriging, while the regression model retains only 9.65% of initial data variability.

IV. DISCUSSION AND CONCLUSIONS

The research demonstrates the opportunities the modern geostatistical methods provide for objective and reproducible analysis and spatial interpolation of point data on summarized weather records. Nowadays decent results can be obtained using open access data and software. The interpolation of precipitation values for 50 weather stations located in the western part of Ukraine has shown that multiple regression of precipitation values on terrain morphometric attributes produced much better results compared with ordinary kriging. Further, the residuals of regression model showed no spatial autocorrelation, rendering unjustified in this case the more sophisticated regression-kriging method.

Some modeling options like the justified number of explanatory variables have been restricted because of the small number of data points. The spatial detail and accuracy of interpolated result can be significantly increased by the inclusion of the records of rain gauges, yet the problem arises of the their correct and precise georeferencing. Another commonly acknowledged problem is the undersampling of high elevation and complex terrain locations, making them underrepresented in factor space.

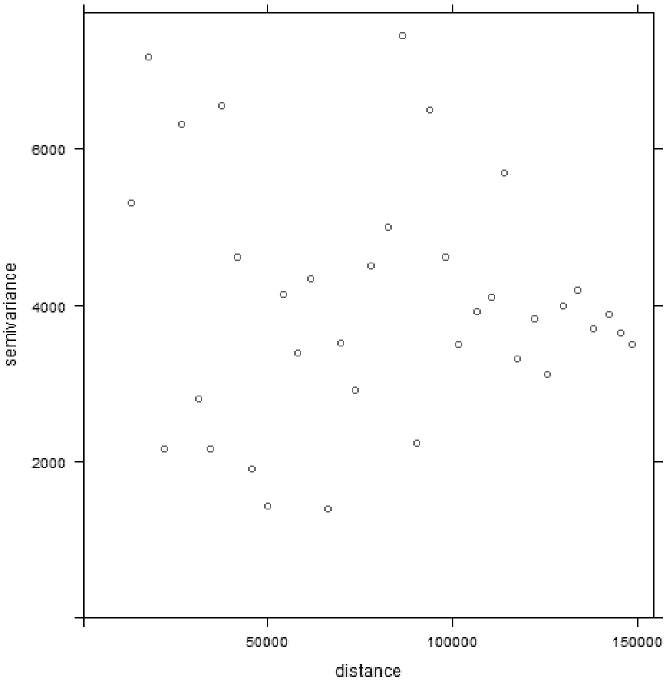


Fig. 3. Empirical variogram of regression model residuals

TABLE II. THE INITIAL AND RESIDUAL VARIANCE OF PRECIPITATION VALUES AFTER APPLICATION OF TWO INTERPOLATION METHODS, ESTIMATED BY LEAVE-ONE-OUT CROSS VALIDATION (MM)

Initial	**After ordinary kriging**	**After regression modeling**
63096	22856 (36,2%)	6087(9,6%)

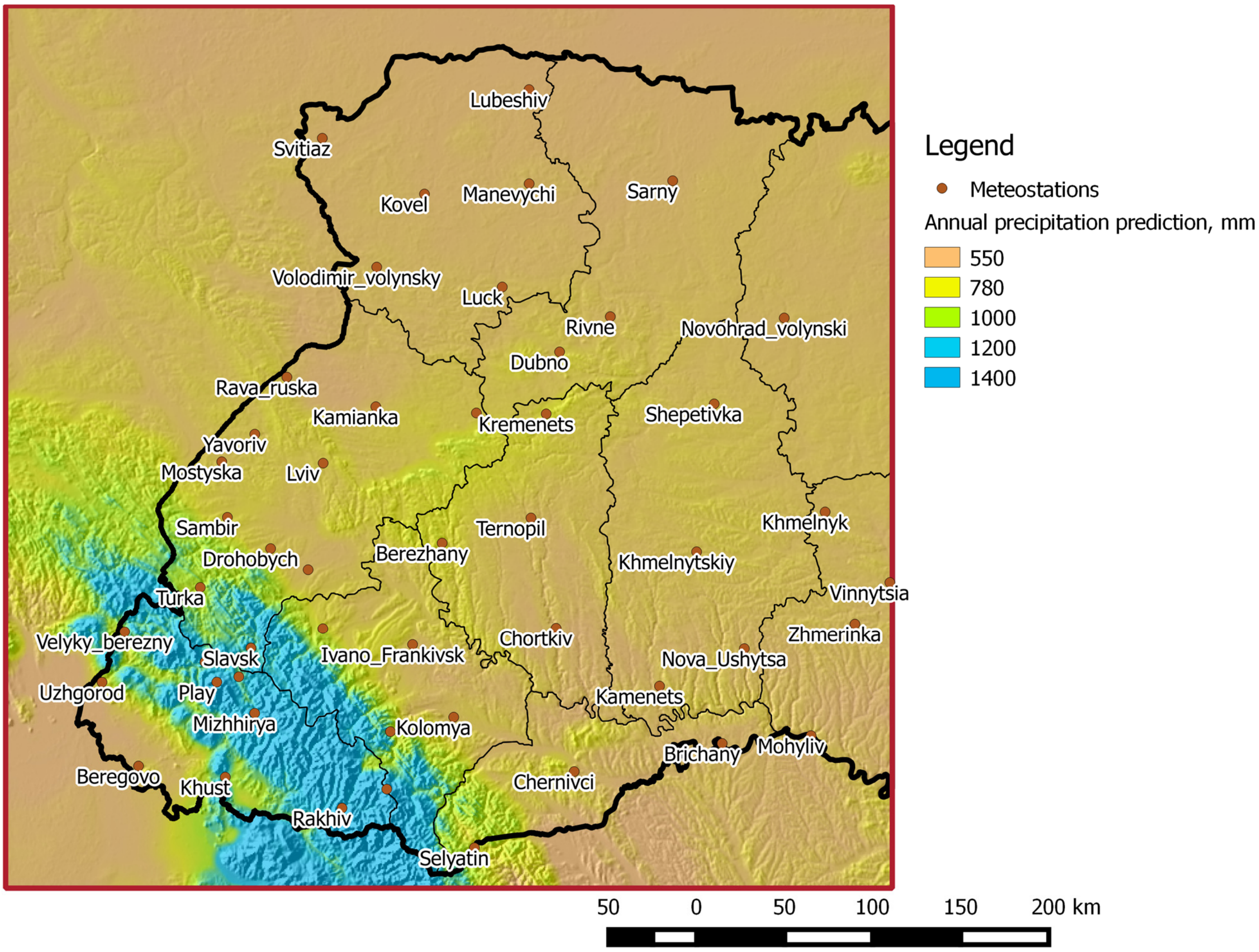


Fig. 4. Precipitation map produced by regression model

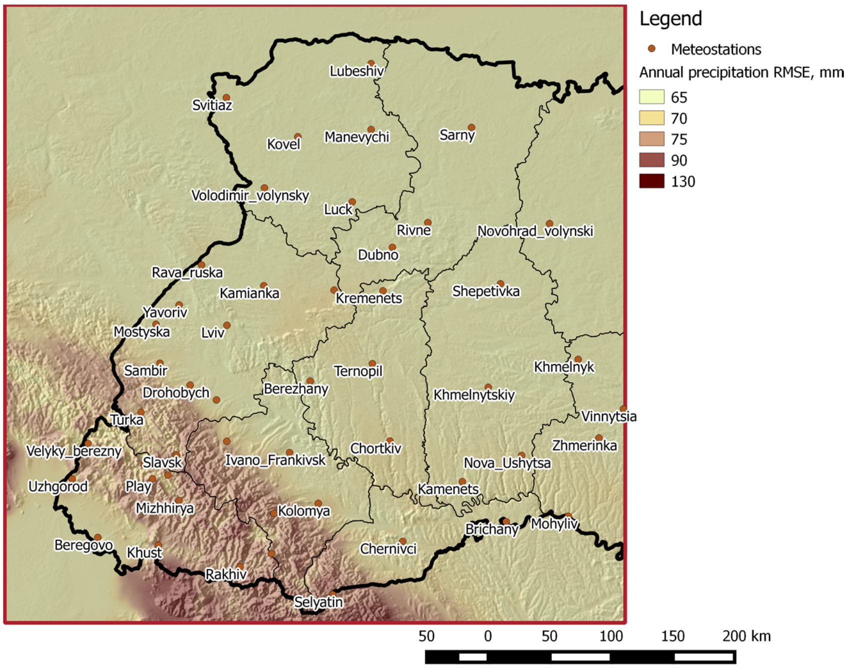


Fig. 5. Estimated RMSE of precipitation map produced by regression model

Better sampling will allow to apply more sophisticated models, e.g., with larger number of explanatory terrain characteristics. Some land cover parameters derived from satellite imagery might be candidates for additional explanatory variables. Denser spatial sampling could make feasible more sophisticated geostatistical methods like regression-kriging with more detailed analysis options, like the consideration of spatial trends and variogram anisotropy.

The resulting map of annual precipitation is reproducible due to formal and transparent method of its creation. It is also accompanied by the map showing the spatial distribution of its estimated accuracy (Fig. 5).

The accuracy of interpolated results were compared with those of CARPATCLIM by calculating the REP parameter (see [15]) for a common subset of stations (those located in Ukrainian Carpathians) by leave-one-out cross-validation. The obtained value of 0.69 is slightly better than 0.66 of CARPACLIM precipitation prediction grid. While our modeling result cannot match the latter with its daily temporal detail and scores of predicted climatic variables, it has finer resolution (1 vs. 10 km) and no less accuracy, even while ignoring data records of rain gauges.

REFERENCES

[1] Bénichou P, Lebreton O., “Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques,” La Météorologie, vol 19, 1987, pp. 23–34.

[2] Diodato N., Ceccarelli M., “Interpolation processes using multivariate geostatistics for mapping of climatological precipitation mean in the Sannio Mountains (southern Italy),” Earth Surface Processes and Landforms, vol. 30, iss. 3, 2005, pp. 259-268.

[3] Goovaerts P., “Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall,” Journal of Hydrology, vol. 228, 2000, pp. 113–129.

[4] Hengl T., “A practical guide to geostatistical mapping”. University of Amsterdam, 2009, 291 p.

[5] Hengl T, Heuvelink G.B.M., Rossiter D.G,. “About regression-kriging: From equations to case studies, “ Computers & Geosciences, vol. 33, 2007, pp. 1301-1315.

[6] Hiebl J., Auer I., Bohm R., Schoner W., Maugeri M., Lentini G., Spinoni J., Brunetti M., Nanni T., Tadic M.P., Bihari Z., Dolinar M., Muller-Westermeier G., “A high-resolution 1961-1990 monthly temperature climatology for the greater Alpine region”, Meteorologische Zeitschrift, vol. 18, no 5., 2009, pp. 507-530.

[7] Jarvis A., Reuter H.I., Nelson A., Guevara E., “Hole-filled seamless SRTM data v4,” International Centre for Tropical Agriculture (CIAT), 2008, available from http://srtm.csi.cgiar.org.

[8] Klein Tank A.M.G. and coauthors, “Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment,” Int. J. of Climatol., vol. 22, 2002, pp. 1441-1453.

[9] Lloyd C.D., “Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain”, Journal of Hydrology, vol. 308 (1-4), 2005, pp. 128–150.

[10] Mkrtchian A., Shuber P., “Geostatistical interpolation of meteodata observations,” Proceedings of international conf. “Geographical science and education in Russia: history and current state”, Saint Petersburg, 2010, pp. 847-855 (in Russian).

[11] Mkrtchian A., Shuber P., “Interpolation of meteorologic obserbations on precipitation and other climatic variables by the regression kriging method”, Visnyk of Lviv university. Ser. geogr., vol.. 42, 2013, pp. 258–264 (in Ukrainian).

[12] Mkrtchian A., Shuber P., “Methods of geospational modeling and mapping of climatic characteristics from the meteorologic stations records,” Visnyk of Lviv university. Ser. geogr., vol. 39, 2011, pp. 245-253 (in Ukrainian).

[13] National atlas of Ukraine. Kyiv: DNVP “Cartographiya”, 2007.

[14] Phillips D.L., Dolph J., Marks D., “A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain,” Agric. For. Meteorol., vol. 58, 1992, pp. 119-141.

[15] Spinoni J. and the CARPATCLIM project team (39 authors), “Climate of the Carpathian Region in 1961-2010: Climatologies and Trends of Ten Variables “, International Journal of Climatology, vol. 35, iss. 7, 2015, pp. 1322-1341.

Spatial interpolation of daily snow depth over Romania

[full paper]

Alexandru Dumitrescu & Marius-Victor Birsan
Department of Climatology
Meteo Romania (National Meteorological Administration)
Bucharest, Romania
marius.birsan@gmail.com

Abstract— **Snow cover has major effects on surface albedo and energy balance, and represents a major storage of water. The snowpack strongly influences the overlying air, the underlying ground and the atmosphere downstream. Snow cover duration influences the growing season of the vegetation at high altitudes. This study presents the spatial interpolation procedure from snow depth measurements at weather stations implying the following stages: (1) Spatial interpolation at 1 km by 1 km resolution of the mean multiannual values (2005-2015) corresponding to each month, computed from the data extracted from the climatological database; (2) Computation of the daily deviations against the multiannual monthly mean for every day and year over 2005-2015 and their spatial interpolation; (3) Spatio-temporal datasets were obtained through merging the two surfaces obtained in stages 1 and 2. The anomalies were considered to be the ratio between the daily snow depth values and the climatology. The spatial variability of the data used in the first stage was accounted for through the use of a series of predictors derived from the digital elevation model DEM and CORINE Land Cover product. To plot the maps with the climatological normals (multiannual means), the Regression-Kriging (RK) spatial interpolation method was used. In order to choose the optimum method applied in spatializing deviations, four interpolation methods were tested using a cross validation procedure: Multiquadratic, Ordinary Kriging (separated and pooled variograms) and 3d Kriging.**

Keywords—snow pack; Kriging; multiquadratic; cross-validation; Romania.

I. Introduction

The realization of high-quality climatic data is essential for realistically assessing the impacts of climate variability and change on a region [1,2]. Gridded data are useful for evaluating the performance of regional climate models, and they serve as input data for spatially distributed agrometeorological and hydrological models [3,4].

Snow cover is a climatic parameter occurring exclusively in the cold season in Romania, being strongly conditioned by air temperature and precipitation type, strongly affecting the surface albedo, the energy balance, the water resources and the hydrological regime [5-7]. In this study we propose a methodology for constructing a gridded dataset over Romania using as target variable the daily snow depth values, measured in cold season (December – March) from 2005 to 2015. Long-term climatic changes over Romania is well documented in various recent papers [8-11]. Climatic changes in Romania since 1961 show increasing temperatures in all seasons except autumn [12], a decrease in snow depth [13] and in wind speed [14], an increase in rain shower frequency [15,16].

The spatial interpolation procedure implied completing the two stages below:

(1) Spatial interpolation at 1000×1000m spatial resolution of the mean multiannual values (2005-2015) corresponding to each cold season month, computed from data extracted from the climatological database;

(2) Computation of daily deviations against the multiannual monthly mean for each day and year from the same period, and combining the maps representing the deviations with the climatic maps.

The spatio-temporal datasets were obtained through merging the two surfaces obtained in stages 1 and 2. The anomalies were considered to be the ratios between the daily values and the climatology.

II. Data and methods

A. Data

The main datasets used in this work consist in daily snow depth values recorded at the meteorological stations, in cold season between December 2005 to March 2015. In this study all weather stations with full records were used (159 stations). Also the auxiliary data listed further, derived from the Digital Elevation Model (DEM) were taken for interpolation the multiannual values: altitude, mean altitude in a 20-km radius, latitude, distance to the Black Sea and distance to the Adriatic Sea.

B. Methods

To interpolate the maps with the climatological normals (multiannual means), the Regression-Kriging (RK) spatial interpolation method was used. To choose the optimum method applied in gridding the deviations, four interpolation methods were tested through the cross validation procedure: Multiquadratic (MQ), Ordinary Kriging with separate (sepOK) and pooled semivariograms (pvOK) and 3D Kriging (K3d).

RK is a multivariate method that can take for computation one or more variables with a spatially continuous distribution (digital elevation model, satellite images, etc.) It results from summing the surface determined through the least squares method (applied to multiple regression) and the surface obtained through spatially interpolating the regression residuals, using the Kriging method. With this method, the first step consists in statistically validating the deterministic model, in the sense of verifying the intensity of the relationships between predictors and the dependent variable. Choosing the best regression method could be performed through the stepwise regression procedure. In the case of RK method, the matrix of the multiple regression grid points represents the large scale variability of the analysed parameter, function of the explanatory variables, interpolated residuals constituting the local peculiarities of the target variable, modeled with the help of the semivariogram [17]:

$$\hat{Z}(s_0) = \sum_{i=1}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{N} \lambda_i \cdot e(s_i) \qquad (1)$$

where $\hat{\beta}_k$ are the coefficients of the regression model, q_k is the value of the predictor in the point localised through the s_0 coordinates for which a new value is estimated and λ_i are the weighting coefficients of the residuals of $e(s_i)$ regression with s_i coordinates. Regression coefficients can be obtained either through the simple method of the least squares or through applying the generalised regression model.

MQ belongs to the class of exact interpolation methods called Radial Base Functions (RBF), which resembles very much to the Kriging family class only differing through that it does not benefit from the contribution of the data spatial structure analysis through the semivariogram. Johnston et al. 2001 defines the general form of this category of interpolators as follow:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \omega_i \, \phi(\|s_i - s_0\|) + \omega_{n+1} \qquad (2)$$

where $\phi(r)$ is the radial base function, $r = \|s_i - s_0\|$ is the radial distance between the point for which a new s_0 value is computed and the points with s_i measured values and ω symbolises the weights to be estimated.

The value of the weight of each point used in interpolation results after solving a system of equations using the matrix computation of the type:

$$\begin{pmatrix} \phi & 1' \\ 1 & 0 \end{pmatrix} \begin{pmatrix} w \\ \omega_{n+1} \end{pmatrix} = \begin{pmatrix} z \\ 0 \end{pmatrix} \qquad (3)$$

with ϕ being the matrix of the distance between the points with known values to which a radial base function is applied; z denotes the vector with the distances between the location chosen for estimation and the points with measurements, to which the same radial base function is applied; w are the estimated weights and ω_{n+1} are the residuals.

MQ radial function is given by the relationship

$$\phi(r) = (r^2 + \sigma^2)^{-1/2} \qquad (4)$$

The smoothing parameter σ can be chosen through computing the minimum sum of squared errors resulted from the application of the cross validation procedure or directly by the user.

OK computes the weights on the basis of the functions that also take for computation the spatial configuration of data [18]. The first step in the interpolation through the OK method is the analysis of the spatial interdependence of the dataset, performed through constructing the semivariogram of the sampled points [19]:

$$\hat{\lambda}(\overline{h}_j) = \frac{1}{2N_j} \sum_{i=1}^{N_j} (Z(s_i) - Z(s_i + h))^2 \qquad (5)$$

where N_j is a set of pairs of locations separated by distance h and $\overline{h}$ = the average of the distances between the N_j distinct pairs.

The assessment in a new location is based on regression against local neighborhood data of the surrounding data points, weighted according to the spatial covariance values [20]

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i), \sum_{i=1}^{N} \lambda_i = 1 \qquad (6)$$

OK weighting functions take for computation both distance and the geographical arrangement of data. The value of the weights of each point used in interpolation results from solving a system of equations through a matrix calculus of the type:

$$\begin{pmatrix} C & 1' \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ m \end{pmatrix} = \begin{pmatrix} c \\ 1 \end{pmatrix} \qquad (7)$$

matrix C representing the covariance's between the points with known values, vector c being made up of the covariances between the points with known values and the point with unknown value, λ = vector of the Kriging weighting coefficients and m = Lagrange multiplier utilized in minimizing errors through the relationship:

$$\sigma^2 = \sum_{i=1}^{N} \lambda_i c + m \qquad (8)$$

In this work two versions of OK method were investigated: (1) with daily estimation of the variograms (separated fitted daily semivariograms – sepOK), and (2) with pooled semivariograms (one single variogram is constructed relying on all data, treating each day as a copy of the same spatial dependence structure – pvOK) [21].

3dK is a three-dimensional extension of the two-dimensional Kriging method, which considers time to be the third orthogonal dimension. The predictions from the space–time cube are based only on one semivariogram model for the period of analysis, while the classical Kriging interpolation models require one semivariogram per time unit [22]. Because good results of this method are achieved when an isotropic covariance model is used, the time dimension must be rescaled in order to align to the spatial directions [23].

C. Validation

To choose the optimum method for interpolating the deviations, the leave-one-out cross validation was applied. This implies the elimination one by one of the values from the set of observed values and determining the value of the point excluded on the basis of the other observed data. The difference between the P estimated data and the O measured ones represents the ε experimental value:

$$\varepsilon_i = P(s_i) - O(s_i) \tag{9}$$

Quantification of differences between estimations and observed data was performed with the help of the error measurement indicators:

– mean error (ME) represents the means of the differences between estimated and measured values respectively:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(Ps_i - Os_i) \tag{10}$$

– mean absolute error (MAE) represents the means of the absolute differences between estimated and measured values respectively:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(Ps_i - Os_i)| \tag{11}$$

– root mean square error (RMSE) is sensitive to the presence of large errors, the squaring process attributing the residuals disproportionate weights:

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(Ps_i - Os_i)^2\right)^{\frac{1}{2}} \tag{12}$$

The box-plot and Taylor-type diagram were also used in the quantitative analysis of results yielded by the four interpolation methods applied in interpolating the ratios [24].

III. Results

A. Climatological maps

In this stage (achieving gridded climatology), there were used as main data the mean multiannual monthly data (1 December 2005 – 31 March 2015) for the parameter of interest. Maps representing the climatological normals were obtained with the RK method.

Due to the existence of the collinearity effect, the predictors derived from the DEM were subjected to the filtering process through the principal component analysis. Filtering the predictors through the principal component analysis (PCA) is performed through transforming the initial variables into a new set of variables, uncorrelated and of a smaller size. The new data set thus obtained contains most part of the original dataset variability (figure 1).

Figure 2 depicts the explained variance of the five principal components obtained through processing the DEM's derived predictors. It can be seen that the first three components explain the main characteristics regarding the spatial variability, representing the strongest configurations in explaining the variance present in the predictor fields (almost 95% of the explained variance), hence only those were taken for computation of the climatological maps.

Prior to applying the RK method, there were identified the statistical relationships between the snow depth values and the auxiliary variables (PCA predictors) for each month. Through applying the retrograde type stepwise regression there can be selected for each case (month) taken apart the statistically significant predictors (table 1). Analyzing the frequency distributions it was noted that the target variables have a positive skewed distribution, hence they were transformed to a close normal distribution by applying the natural logarithm

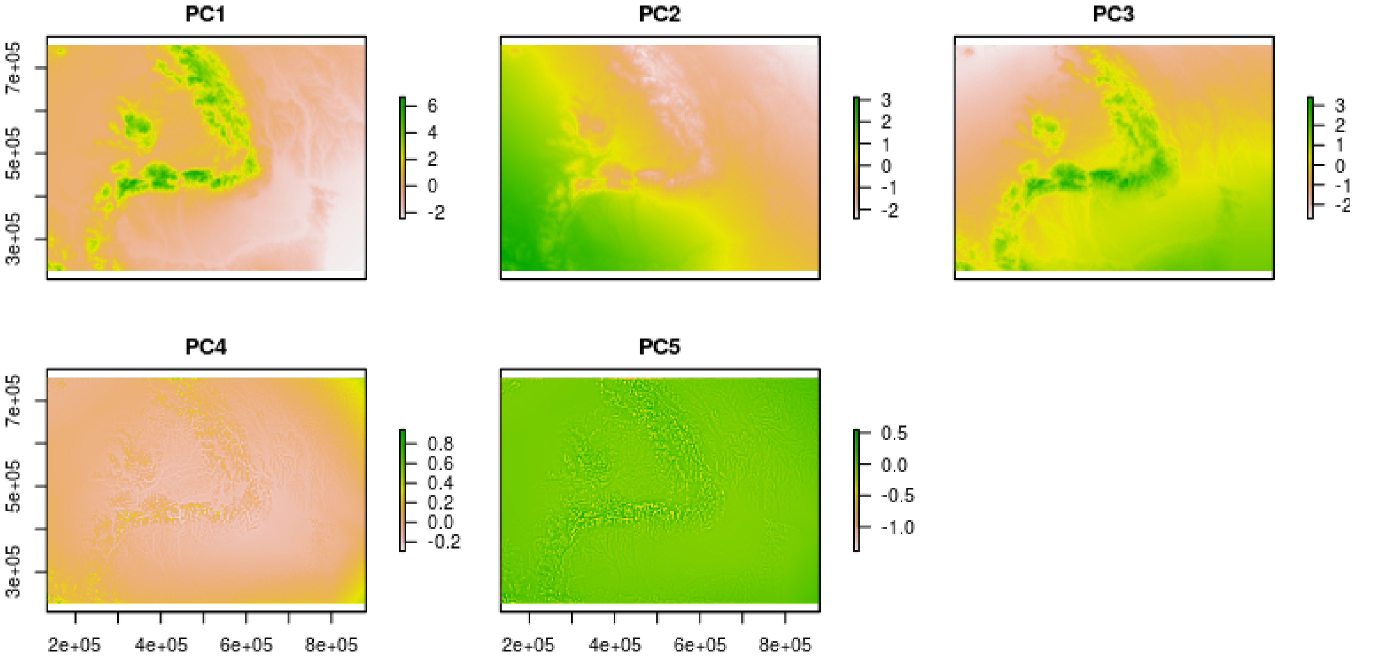


Fig..1. PCA transformed predictors.

function. The log1p() function from R language was used, which can also be applied when the data series contain values of zero. The estimations were back-transformed to real values with a help of expm1() function (https://stat.ethz.ch/R-manual/R-devel/library/base/html/Log.html).

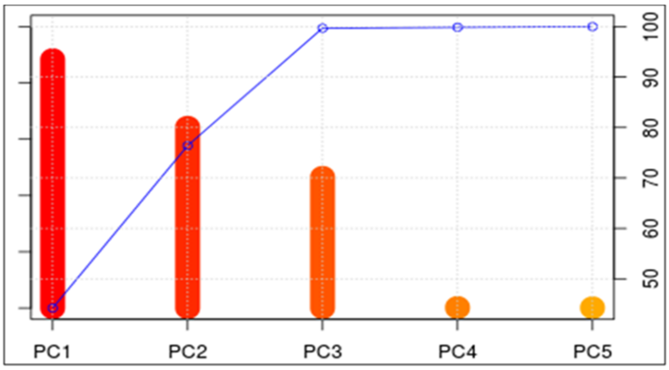


Fig.2. Variance explained by the principal components (PCA) computed from the set of predictors obtained from the numerical altimetric model.

The predictive power of the regression models varies from month to month, with smallest R-squared value in January, and the largest value in March. For all months more than 70% of the spatial variability of snowpack depth being explained by the predictors. From analyzing the maps constructed with RK method (figure 3), it can be seen that the highest values are recorded in the closing month of the cold season, being generated by persisting below zero temperatures at high altitudes, which favors constant accretion of snow.

TABLE I. SNOWPACK DEPTH: PREDICTORS SELECTED THROUGH USING THE STEPWISE REGRESSION AND R^2 DETERMINATION COEFFICIENTS

Month	*Predictors*	*R^2*
log1p(Dec)	PC1 + PC2 + PC3	0.749
log1p(Jan)	PC1 + PC2 + PC3	0.733
log1p(Feb)	PC1 + I(PC1^2) + PC3	0.736
log1p(Mar)	PC1 + I(PC1^2) + PC2 + PC3	0.844

B. *Daily gridded dataset*

In this stage, a number of interpolation methods were tested so as to choose the optimum interpolation method of daily anomalies: Multiquadratic (MQ), Ordinary Kriging - separate (sepOK) and pooled variograms (pvOK)- and 3D Kriging (K3d). For the sepOK method the semivariograms were automatically estimated through the use of the automap R package [23].

Since there are regions where the mean multiannual snowpack depth is equal to zero, at the stations located on lower altitudes, a 1 cm value was added to the multiannual means prior to computing the daily anomalies.

The cross validation procedure was applied to the anomalies computed over the period 1 Dec 2014 – 31 Mar 2015. According to figure 4, estimations performed with the four methods are very similar, a difference being apparent with

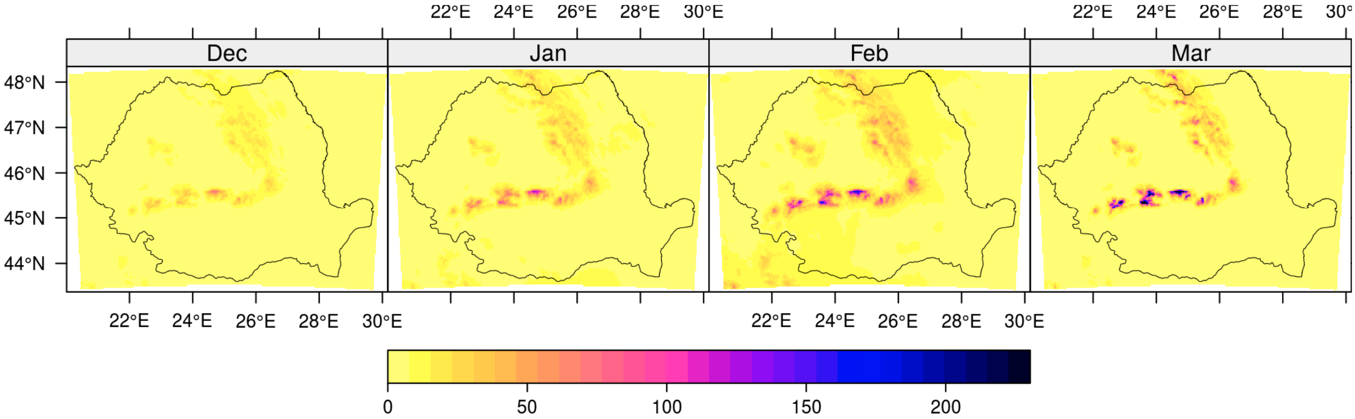


Fig.3. Mean snowpack depth (cm) Dec 2005 – Mar 2015.

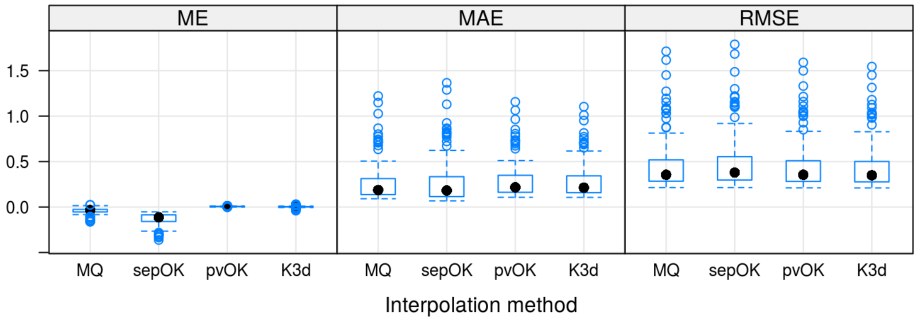


Fig.4. Mean Box-plot type diagram of daily anomalies (Dec 2014 – Mar 2015), ME (left), MAE (middle) and RMSE (right) computed through using the original datasets against those estimated through the cross validation procedure using MQ, sepOK , pvOK and K3d interpolation methods.

the help of RMSE indicators, and ME that points out the superior estimations performed with K3d method.

The Taylor diagrams (figure 5) confirm that the best estimates are provided by K3d method, regardless the month analyzed. pvOK obtains comparable results, with nearly the same computed values for Pearson's correlation coefficient and slightly larger standard deviation values. sepOK has the poorest accuracy in terms of the three computed indicators. Note that all methods underestimate the variability of the observed data, the poorest performance being computed for the March month, when the snow pack depth value are greater than 0 only in the mountain regions.

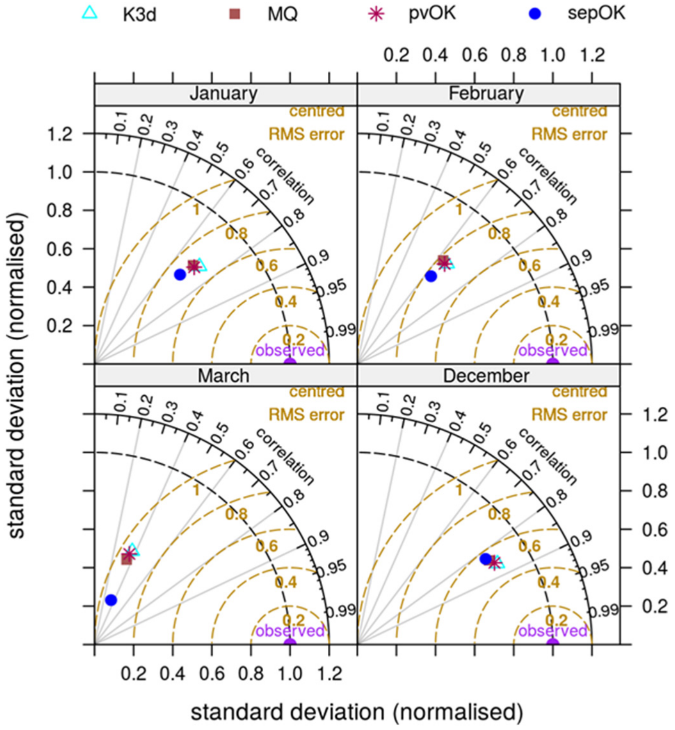


Fig.5. Taylor-type diagram of daily deviations obtained through the cross validation procedure for the four interpolation methods (MQ, sepOK , pvOK and K3d).

Due to the good results in interpolating ratios and to the fewer steps required for producing the maps, K3d method was chosen to generate daily anomaly maps. The final daily snowpack depth maps were generated by multiplying the ratio maps with those representing the monthly climatology.

Using gridded daily data regarding the snowpack depth, constructed with the help oh the K3d method, the monthly maximum snowpack depth was computed in every grid point (figure 6). The highest values of this parameter correspond to the high mountain areas (more than 200 cm starting from January), persisting till March due to the negative mean temperatures. A considerable snowpack (deeper than 50 cm) can also be found in the extra-Carpathian areas as a consequence of the blizzard episodes specific to January and February.

IV. Conclusions

Using an interpolation procedure that implies completion of a number of stages, there were obtained the gridded datasets for the snowpack depth values. Those were constructed at 1000×1000m spatial resolution, using meteorological records from 2005 to 2015, only for the months December, January, February and March.

In the first stage, the monthly climatology maps were constructed based on a multivariate geostatistical model based on the RK method. This can take for computation in the process of spatialization one or more variables with a continuous spatial distribution (DEM's derived predictors). In order to choose the optimum combination of potential predictors, the stepwise regression was used, selecting in the interpolation process only those predictors which are statistically significant for each analyzed case (month).

The second stage implied spatial interpolation of the daily deviations against the monthly normals. In view to choose the optimum interpolation method, four spatialization methods were tested. Through using the cross validation procedure and computing some error indicators, conclusion was drawn that the best estimates are obtained through the K3d, hence this method was applied in interpolating the anomalies. Through combining the maps with the daily anomalies with those rendering the monthly normals, daily snow depth maps were constructed.

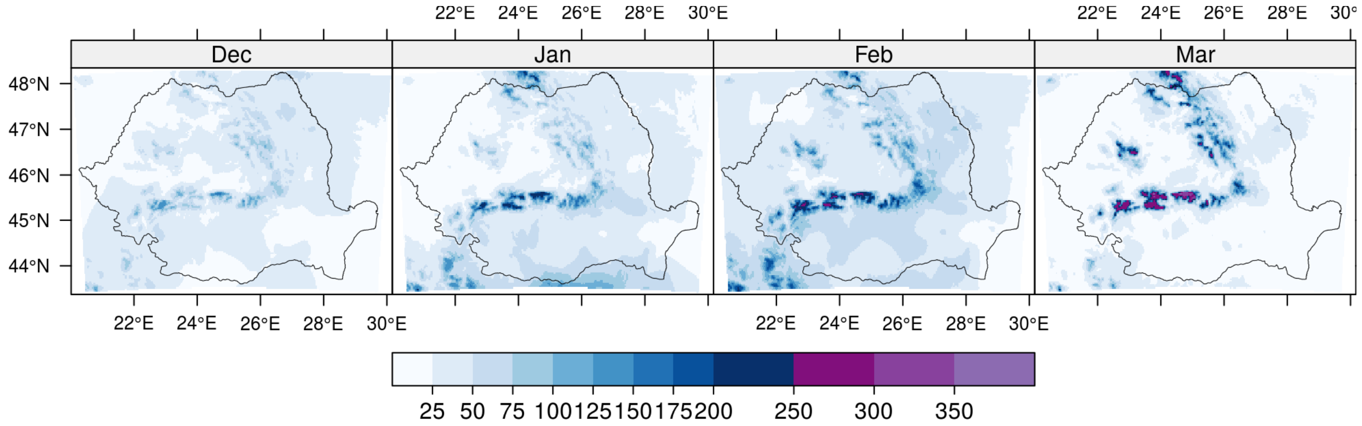


Fig.6. Maximum snowpack depth (2005-2015).

Using the gridded daily data, other parameters may be computed: number of days with snowpack, the first and last day with snowpack, the maximum snowpack depth, etc.

Maps obtained within this stage supply an overall picture of the analyzed variables, however with a precision directly influenced by the scale at which those were performed, by the spatial estimation errors specific to the geostatistical methods and by the density of the measurement points (weather stations operated by the National Meteorological Administration). In certain areas, where peculiar climatic conditions are specific and where no meteorological measurements are performed, it is recommended to achieve detailed studies regarding the spatio-temporal variability of the parameters of interest stressing the spatio-temporal local development character of the meteorological phenomena.

Acknowledgment

This study was realized within the framework of the project SNOWBALL funded by the European Economic Area Financial Mechanism (EEA Grants), contract number EEA-JRP-RO-NO-2013-1-0285.

References

[1] A. Dumitrescu, and M.V. Birsan, "ROCADA: a gridded daily climatic dataset over Romania (1961–2013) for nine meteorological variables," Nat. Hazards vol. 78(2): pp. 1045-1063, 2015. DOI: 10.1007/s11069-015-1757-z

[2] A. Dumitrescu, M.V. Birsan, and A. Manea, "Spatio-temporal interpolation of sub-daily (6-hour) precipitation over Romania for the period 1975-2010," Int. J. Climatol. vol. 36(3), pp. 1331-1343, 2016. DOI: 10.1002/joc.4427

[3] O.E. Tveito, M. Wegehenkel, F. van der Wel, and H. Dobesch, "The use of geographic information systems in climatology and meteorology," COST Action 719 Final Report, 2006.

[4] M.V. Birsan, "Application of a distributed physically-based hydrological model on the upper river basin of Somesul Mare (Northern Romania)," Rom. Rep. Phys. vol. 65(4), pp. 1469-1478, 2013.

[5] M.V. Birsan, "Trends in Monthly Natural Streamflow in Romania and Linkages to Atmospheric Circulation in the North Atlantic," Water Resour. Manage. vol. 29(9), pp. 3305-3313, 2015. DOI: 10.1007/s11269-015-0999-6

[6] M.V. Birsan, L. Zaharia, V. Chendes, and E. Branescu, "Recent trends in streamflow in Romania (1976–2005)," Rom. Rep. Phys. vol. 64(1), pp. 275-280, 2012.

[7] M.V. Birsan, L. Zaharia, V. Chendes, and E. Branescu,"Seasonal trends in Romanian streamflow," Hydrol. Process. vol. 28(15), pp. 4496-4505, 2014. DOI: 10.1002/hyp.9961

[8] S. Cheval, M.V. Birsan, and A. Dumitrescu, "Climate variability in the Carpathian Mountains Region over 1961–2010," Global Planet. Change vol. 118, pp. 85-96, 2014. DOI: 10.1016/j.gloplacha.2014.04.005

[9] S. Cheval, A. Busuioc, A. Dumitrescu, and M.V. Birsan, "Spatiotemporal variability of meteorological drought in Romania using the standardized precipitation index (SPI)," Clim. Res. vol. 60, pp. 235-248, 2014. DOI: 10.3354/cr01245

[10] A. Dobrinescu, A. Busuioc, M.V. Birsan, and A. Dumitrescu, "Changes in thermal discomfort indices in Romania and responsible large-scale mechanisms," Clim. Res. vol. 64(3), pp. 213-226, 2015. DOI: 10.3354/cr01312

[11] L. Marin, M.V. Birsan, R. Bojariu, A. Dumitrescu, D.M. Micu, and A. Manea, "An overview of annual climatic changes in Romania: trends in air temperature, precipitation, sunshine hours, cloud cover, relative humidity and wind speed during the 1961–2013 period," Carpath. J. Earth Env. vol. 9(4), pp. 253–258, 2014.

[12] A. Dumitrescu, R. Bojariu, M.V. Birsan, L. Marin, and A. Manea, "Recent climatic changes in Romania from observational data (1961-2013)," Theor. Appl. Climatol. vol. 122(1-2), pp. 111-119, 2015. DOI: 10.1007/s00704-014-1290-0

[13] M.V. Birsan, and A. Dumitrescu, "Snow variability in Romania in connection to large-scale atmospheric circulation," Int. J. Climatol. vol. 34, pp. 134-144, 2014. DOI: 10.1002/joc.3671

[14] M.V. Birsan, L. Marin, and A. Dumitrescu, "Seasonal changes in wind speed in Romania," Rom. Rep. Phys. vol. 65(4), pp. 1479-1484, 2013.

[15] A. Busuioc, M.V. Birsan, D. Carbunaru, M. Baciu, and A. Orzan, "Changes in the large-scale thermodynamic instability and connection with rain shower frequency over Romania. Verification of the Clausius–Clapeyron scaling," Int. J. Climatol. vol. 36(4), pp, 2015-2034, 2016. DOI: 10.1002/joc.4477

[16] A. Manea, M.V. Birsan, G. Tudorache, and F. Cărbunaru, "Changes in the type of precipitation and associated cloud types in Eastern Romania (1961-2008)," Atmos. Res. vol. 169, pp. 357-365, 2016. DOI: 10.1016/j.atmosres.2015.10.020

[17] T. Hengl, G.B.M. Heuvelink, and D.G. Rossiter, "About Regression-Kriging: From Equations to Case Studies," Comput. Geosci. vol. 33(10). pp. 1301-1315, 2007.

[18] E.H. Isaaks, and R.M. Srivastava, "An Introduction to Applied Geostatistics," 1989. Oxford University Press.

[19] E.J. Pebesma, "Multivariable Geostatistics in S: The Gstat Package." Comput. Geosci. vol. 30 (7), pp. 683-691, 2004.

[20] K. Johnston, J.M. Ver Hoef, K. Krivoruchko, and N. Lucas, "Using ArcGIS Geostatistical Analyst.", 2001. ESRI Press, Redlands.

[21] B. Gräler, M. Rehr, L. Gerharz, and E. Pebesma, "Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009," Technical Paper 2011/10, The European TopiCentre on Air Pollution and Climate Change Mitigation (ETC/ACM), 2013.

[22] E. Pebesma, "Spatio-temporal geostatistics using gstat," 2013. http://cran.r-project.org/web/packages/gstat/vignettes/st.pdf.

[23] P.H. Hiemstra, E.J. Pebesma, C.L.W. Twenhöfel, G.B.M. Heuvelink, "Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network," Comput. Geosci. vol. 35(8), pp. 1711-1721, 2009. DOI: 10.1016/j.cageo.2008.10.011

[24] K.E. Taylor, "Summarizing Multiple Aspects of Model Performance in a Single Diagram," J. Geophys. Res. vol. 106(D7), pp. 7183–7192, 2001. DOI:10.1029/2000JD900719.

Evaluation of phenology parameters as proxi for drought measurements

[extended abstract]

Boudewijn van Leeuwen, Zsuzsanna Ládanyi
Department of Physical Geography and Geoinformatics,
University of Szeged
Szeged, Hungary
leeuwen@geo.u-szeged.hu, ladanyi@geo.u-szeged.hu

***Abstract*—Drought is a large problem in many parts of the world. It is challenge to quantify, monitor and predict its severity and the resulting impacts. This research aims to quantify the relationship between changes in phenology and productivity parameters and the development of drought using medium resolution satellite data. Phenology curves are calculated and the area below them is used as an indicator of vegetation productivity. The development of this parameter in time is evaluated as a proxy of measurements of drought. The strength of the relationship between drought and vegetation productivity varies between vegetation types.**

Keywords—Phenology and productivity parameters; drought; MODIS

I. Introduction

Drought is a phenomenon that is difficult to measure or quantify, but one of its main indicators is the stress it causes to vegetation. Long term time series allow for the analysis of the relationship between plant stress and climate change [1]. Our research aims to quantify the relationship between changes in phenology and productivity of vegetation and the development of drought. A strong relationship would allow for the use of the phenology parameter as a proxy for the measurement of drought. To estimate the development of vegetation productivity, medium resolution MODIS vegetation satellite data were preprocessed and interpolated to data curves. Drought was quantified by calculation of the Pálfai drought index (PAI). This research is a continuation of preliminary work presented in [2] .

II. Methodology

A. MODIS data

Moderate resolution imaging spectroradiometer (MODIS) satellite data of the period 2000 until 2015 were used to create the time series data. These data have a spatial resolution of 250 meter and is created by calculating daily normalized difference (NDVI) and enhanced (EVI) vegetation indices [3]. Over a period of 16 days, the highest values are stored in the final MOD16Q1 product. These data are geometrically and atmospherically corrected. The EVI data were used for further processing since they are less sensitive to atmospheric disturbances [4]. Spatial subsets from the data were created from the Illancs study area near the Danube in the south of Hungary. The surface is mainly covered by blown sand. The groundwater level in the area has decreased severely during the last 50 years. Plotting the values of individual pixels against time and interpolating between the dates results in a curve with many spikes and outliers (Fig. 1).

B. Filtering of outliers

The spikes and outliers in the data set were filtered in two steps. All raw values higher than 0.8 were removed because, they are considered unrealistically high. Then, the seasonal trend decomposition (STL) method was applied which is a procedure that extracts the trend, the seasonal behavior and the remaining signal from the time series [6]. The remaining signal is considered noise and was removed from the data set.

C. Curve fitting

The cleaned data set showed the trend and seasonality of the vegetation growth, but still incorporated many small fluctuations due to varying imaging conditions like atmospheric disturbance, shadow and lighting conditions. To exclude these from the phenology parameters, a least squares fitting method originally developed by Savitzky-Golay was applied [7].

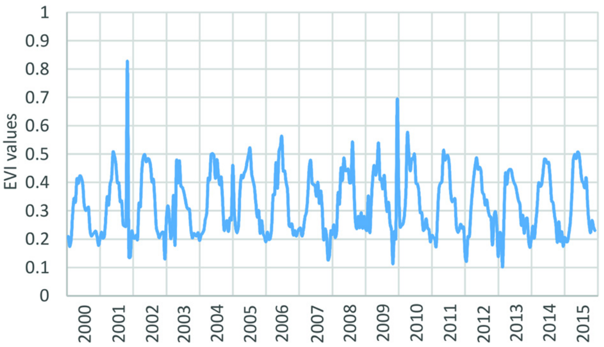


Fig. 1. Raw EVI values of one locust forest pixel

D. Vegetation productivity calculation

The seasonal small integral (S-integral) is defined as the integral of the area (s) under the graph with the season start (a) and end (b) as boundaries and is a dimensionless number larger than zero (Fig. 2). It is an indicator of the vegetation productivity. The season start and end are defined as a certain percentage of the amplitude of the curve [8][9]. The S-integral was calculated for every year in the data set.

III. Results

A. *Pálfai drought index*

The PAI values in the period 2000 until 2015 varied between 3.63 and 15.00 (Fig. 3). The average PAI was 7.7. In 10 years the PAI was larger than 6, indicating a drought year, while 2010 was the most humid year of the century (over 1000 mm precipitation). From the 10 drought years 3, 5 and 2 years were extremely heavy, medium and moderate drought years, respectively. These observations endorse that the study area is strongly affected by climate extremities.

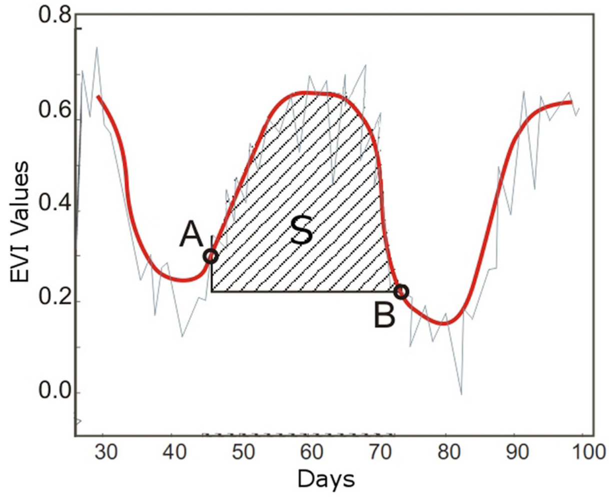


Fig. 2. Phenology curve with start of season (A) and of season (B), and phenology production (S)

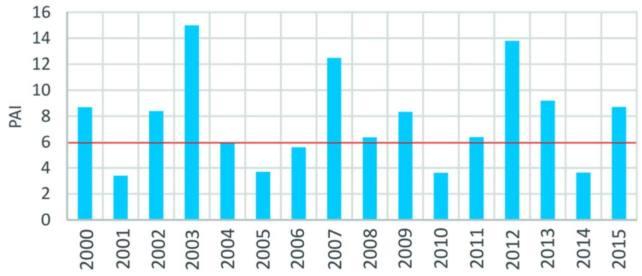


Fig. 3. PAI values for Kiskunhalas between 2000 and 2015

B. *Vegetation productivity*

The vegetation productivity for the 8 test areas with two different types of forests showed a high variation (Fig. 4). The deviation of the S-integral showed higher differences for the locust forests compared to pine. Obviously, this is due to its different foliage and leaf structure. Negative anomalies can be observed in 6 from the 16 years, and the highest peaks occurred in 2000, 2003, 2012 and 2013. Positive anomalies occurred in 9 from the 16 years 16, and the highest peaks were found in 2002 and 2015.

C. *Comparison between Pálfai drought index and vegetation productivity*

The observed pattern of the vegetation productivity deviations from the average showed a strong agreement with the PAI index in case of locust forests (Fig. 5). The extremely heavy drought years resulted in the highest decrease in vegetation productivity. The positive anomalies of extremely humid years (e.g. 2010) were not so spectacular. These observations endorse that the physical geographical background of the area, which is exposed to high water scarcity due to the lowering groundwater table (on the highly elevated areas, a decrease of up to 10 m compared to the 1970s has been observed), and by the genetic soil type, that consists of sandy soils, which are characterized by high infiltration capacity. The water scarcity causes significant decrease of productivity, however, the extreme surplus did not result in extreme positive peaks in productivity. Vegetation is highly dependent on the precipitation and temperature conditions in this area, which was well reflected in the determined relationship with the drought index. The quantification of the relationship between the factors did not result in significant determination coefficients, the R^2 varied between 0.259 and 0.424 for the 4 sample locust forests (Fig. 6). This is due to the fact that the previous years have also impact on the following years. A humid year can result in more balanced conditions in a following drought year (e.g. 2001-2002), and the damages of a drought year can impact the tree growth in a subsequent average years (e.g. 2012-2013). Furthermore, local conditions can also have an impact on vegetation growth.

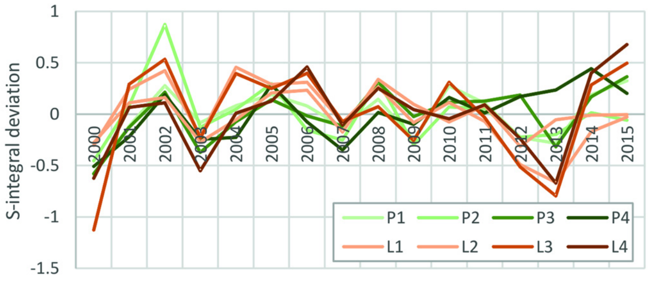


Fig. 4. Deviation of S-integral from the average between 2000 and 2015) for locust (L) and pine (P) forests

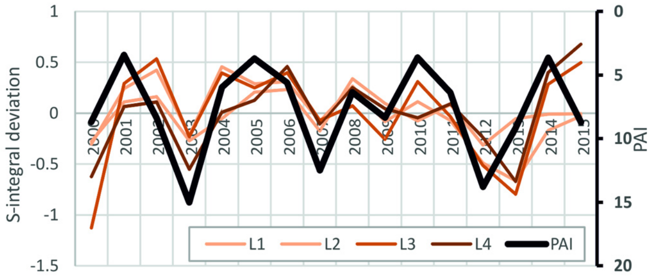


Fig. 5. Comparison of the productivity parameter devations for locust with the PAI values

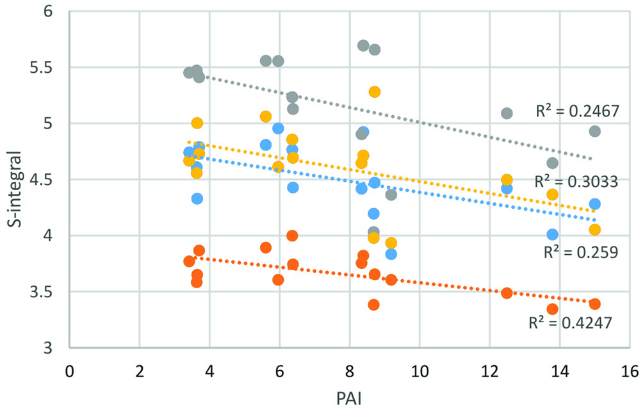


Fig. 6. The relationship between the vegetation productivity (S-integral) of locust forests and the PAI index values

IV. Conclusions and Discussion

A relationship between the vegetation productivity and the PAI drought index could be identified in this research, however, the result shows that apart from temperature and precipitation other factors play important role in the process (e.g. ecological attributes, local influencing factors). Further analysis of phenology parameters is planned in the future to improve knowledge about the behavior and patterns of response of the vegetation to the phenomenon on vegetation in regional scale in the Carpathian Basin.

References

[1] W.W. Hargrove, J.P. Spruce, G.E. Gasser and F.M. Hoffman, “Toward a national early warning system for forest disturbances using remotely sensed canopy phenology”, Photogrammetric Engineering – Remote Sensing, vol 75, 2009, pp. 1150-1156.

[2] B. van Leeuwen, Zs. Ladányi and D. Bátori, “Medium resolution satellite data based estimation of phenology and productivity parameters for drought monitoring”, Carpathian Journal of Earth And Environmental Sciences, (in press).

[3] A. Huete, C. Justice and W. van Leeuwen,”MODIS vegetation index (MOD13) algorithm theoretical basis document”, Greenbelt: NASA Goddard Space Flight Centre, 1999, http://modarch. gsfc. nasa. gov/MODIS/LAND/# vegetation-indices.

[4] A. Huete, K. Didan, T. Miura, E.P. Rodriguez, X. Gao and L.G. Ferreira, “Overview of the radiometric and biophysical performance of the MODIS vegetation indices”, Remote sensing of environment, vol. 83(1), 2002, pp.195-213

[5] I. Pálfai and Á. Herceg, “Droughtness of Hungary and Balkan Peninsula”, Riscuri si Catastrofe An X 9/2., 2011, pp 145–154.

[6] R.B. Cleveland, W.S. Cleveland, J.E. McRae and I. Terpenning, “STL: A seasonal-trend decomposition procedure based on loess” Journal of Official Statistics, vol 6(1), 1990, pp.3-73.

[7] A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures", Analytical chemistry, vol 36(8), 1964, pp 1627-1639.

[8] E. Ivits, M. Cherlet, G. Tóth, S. Sommer, W. Mehl, J. Vogt, F. Micale,” Combining satellite derived phenology with climate data for climate change impact assessment”, Global and Planetary Change, vol 88, 2012, pp. 85-97.

[9] L. Eklundh and P. Jönsson, “TIMESAT 3.2 with parallel processing, Software manual”, 2015, p. 88, http://web.nateko.lu.se/timesat/docs/TIMESAT32_software_manual.pdf

Comparison of Three Interpolation Schemes to Generate Daily and Monthly Gridded Precipitation Analyses by the Global Precipitation Climatology Centre (GPCC) and its Application to Produce Global Analyses

[extended abstract]

Ziese, M.; Schneider, U.; Meyer-Christoffer, A.; Rustemeier, E.; Finger, P.; Schamm, K.; Becker, A.
Deutscher Wetterdienst (DWD)
Global Precipitation Climatology Centre
Offenbach am Main, Germany
gpcc@dwd.de

Abstract— **The Global Precipitation Climatology Centre (GPCC) collects in-situ precipitation observations and provides gridded analyses with different timeliness. An overview of the data basis and quality control is given. A comparison of three interpolation schemes was done to assess their performance. Furthermore a description of the gridded precipitation analyses from the GPCC is given.**

Keywords—precipitation; interpolation; comparison; global

I. Introduction

The Global Precipitation Climatology Centre (GPCC) collects and controls quality of daily and monthly in-situ precipitation measurements to produce gridded analyses, as mandated by the World Meteorological Organization (WMO) of the United Nations (UN). It was established in 1989 and is operated since then by Deutscher Wetterdienst (DWD), the national meteorological service of Germany. The GPCC is part of the Global Precipitation Climatology Project (GPCP) and contributes to WMO World Climate Research Project (WCRP) and Global Energy and Water Cycle Exchanges Project (GEWEX).

GPCC has a very strict data policy: no station observations or metadata are distributed to third parties, as the GPCC does not hold the copy right of these data. But the gridded analyses are available free from charge at ftp://ftp-anon.dwd.de/pub/data/gpcc/html/download_gate.html. GPCC's data sets are DOI-referenced to guaranty a long lasting availability and easy citation.

II. Data base and quality control

Data supplies from national meteorological and hydrological services are the backbone of GPCC's data base. In addition, global as well as regional data collections are integrated, e.g. Global Historical Climatology Network (GHCN), United Nations Food and Agriculture Organization (FAO), Climate Research Unit of the University of East Anglia (CRU) or European Climate Assessment and Dataset (ECA&D). SYNOP and CLIMAT reports transmitted via WMO Global Telecommunication System (GTS) and monthly totals calculated at the Climate Prediction Centre (CPC) of the National Oceanic and Aeronautic Administration (NOAA) are the basis for near-real time products.

The data are stored in a relational data bank with source specific slots and quality flags. This allows a comparison of the different sources to detect and correct errors. Quality codes are applied to track changes with a roll-back option to the original value.

Several automated, semi-automated and manual quality control procedures are applied before and during the loading into the data bank. This includes automated consistency and threshold checks during the decoding of SYNOP reports as well as tests against station and grid based background statistics. Metadata are checked twice – before and during import into the data bank.

Typical errors detected are factor-10 and conversion errors of the observations, observations assigned to a wrong station, errors in the meta data, observations shifted by one day, one month, one year or other periods, wrong coded missing values and gaps filled with long term means or data from other stations. Station names differ between data sources due to different transcriptions from the original language or replace of language specific characters in ASCII-coded files.

III. Comparison of interpolation schemes

Gridded precipitation analyses are generated by means of interpolation by the GPCC. For monthly totals a modified version of the SPHEREMAP scheme [1] is applied, since 2008

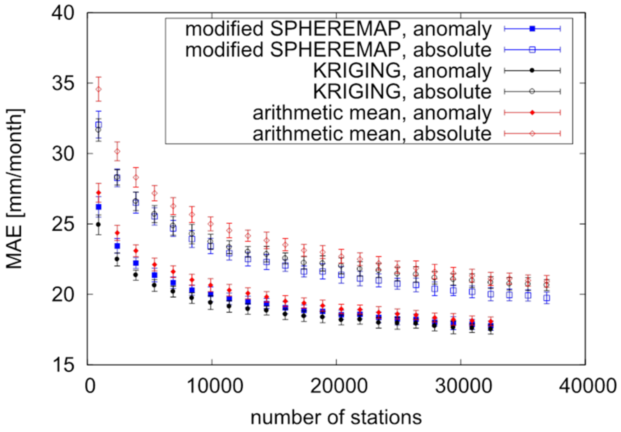


Fig. 1. Comparison of the performance of interpolation schemes. Open symbols: interpolation of totals, filled symbols: interpolation of anomalies.

as interpolation of anomalies [2]. To generate daily gridded fields, ordinary block kriging is utilized [3] as an anomaly technique.

A comparison of three interpolation schemes was done to assess their performance: a modified SPHEREMAP, ordinary block kriging and arithmetic mean for the interpolation of totals and anomalies. Cross-validation techniques were utilized to evaluate the skill of the schemes. As the task of the GPCC is the production of global analyses, no regional assessments were done.

For monthly totals 45.000 global distributed stations were applied. These were divided into 150 collectives with 300 stations. The number of input collective was varied as the amount of reference stations was kept at 4800 stations to assess the impact of the station density. For each station density 50 comparison runs were done with changing reference and input stations. As depicted in Fig. 1, the skill of all tested schemes gets better with increasing station density. The difference between the interpolation of totals or anomalies is larger than the variance between the interpolations schemes. If the station density exceeds a certain limit, the performance of the different interpolation schemes is comparable!

For the assessment of interpolations of daily precipitation a slightly different approach was applied [4]. All daily observations for one year (2008) were taken. A leave-one-out technique was utilized to compute the mean squared error (MSE) for each Köppen-Geiger climate zone. Once again, totals and relative anomalies with respect to the monthly total were chosen. Contrary to the monthly assessment no clear outperformance of any interpolation scheme - whether anomalies or totals - was found. In a detailed view ordinary block kriging of anomalies was slightly better than the other choices.

IV. GRIDDED PRECIPITATION ANALYSES

The GPCC produces several near-real time as well as non-real time precipitation analyses optimized for different applications.

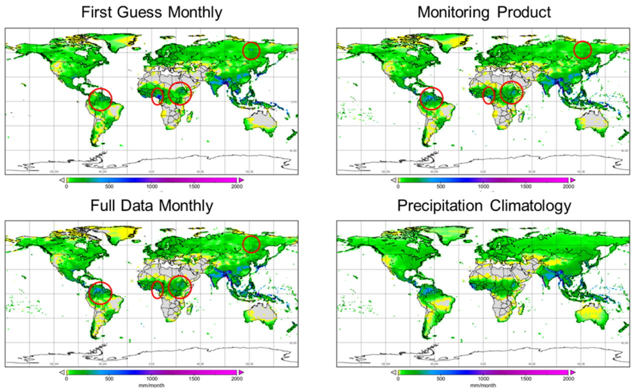


Fig. 2. Comparison of gridded products. Red circles highlight differences between the products originating in quality and station density issues.

Near-real time analyses are the First Guess Daily, First Guess Monthly and Monitoring Product. The First Guess products are based on SYNOP reports from roughly 7.200 stations and an automated quality control. These analyses are released three to five days after the end of each month. CLIMAT reports and monthly totals calculated at the CPC are additional data used for the Monitoring Product, increasing the number of stations to roughly 8.000. The quality control is enhanced by a visual inspection of automated marked questionable values. Due to the additional data and quality control effort, the Monitoring Products is released two months after the observation. In addition to the First Guess products also a correction factor depending on the weather conditions and the fraction of liquid/mixed/solid precipitation is provided.

Non-real time products are the Climatology, Full Data Daily, Full Data Monthly and HOMPRA-Europe. The Climatology are gridded long term means focusing on the period 1951-2000. If this period is not available, means from one of the alternative periods 1931-1960, 1951-1980, 1961-1990, 1971-2000 or 1981-2010 are taken. Up to ten years of missing data are allowed in the above mentioned periods. If no one of these periods is available, but at least ten years of consecutive data, this period is taken. At least, ten years of available data are used. Therefore 75.000 stations are utilized to compute the Climatology. This Climatology is also the background field for the anomaly interpolation applied at GPCC.

Full Data Daily and Full Data Monthly are analyses based on the maximal available number of stations for each month, up to 52.000 and 30.000 stations with monthly and daily values in the best covered month. Full Data Monthly covers the time period from 1901 to 2013, whereas Full Data daily covers 1988 to 2013.

HOMPRA-Europe is based on homogenized monthly precipitation totals from 5.500 stations in Europe with at least 90% coverage for the years 1951 to 2005. The break detection and correction was done by an automated scheme similar to PRODIGE [5].

A comparison of First Guess Monthly, Monitoring Product, Full Data Monthly and Climatology is shown in Fig. 2 and depicts the effect of station density and quality control effort on the spatial details and reduction of obvious errors.

V. ADDITIONAL PRODUCTS

For drought monitoring purposes the GPCC provides the GPCC-drought index (GPCC-DI) [6], which is a combination of the SPI [7] and SPEI [8] to achieve a global coverage. It is based on the First Guess Daily products and gridded temperature analysis from CPC [9]. Several aggregation periods are calculated. Due to the availability of data, this product is calculated at the tenth of the following month.

The Interpolation Test Dataset was developed to provide the users community a possibility to test the GPCC interpolation procedure against others. It is based on freely available GHCN data. The used station data are provided together with gridded analyses of these stations using the GPCC gridding technique.

ACKNOWLEDGMENT

We thank all data providers for contributing their data for GPCC analyses. It are only these data contributions that in total have put GPCC into the position to provide the global precipitation analyses described, and we are looking forward to their further contributions, which are crucial to maintain and enhance GPCC's level of products in terms of scope and quality.

REFERENCES

[1] Willmott, C.J., Rowe, C.M. and Philpot, W.D.: Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring, The American Carthographer,1985, Vol. 12, Is. 1, 5-16.

[2] Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U. and Ziese, M.: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901-present, Earth System Science Data, 2013, Vol. 5, Is. 1, 71-99.

[3] Krige, D.G.: Two-dimensional weighted moving average trend surfaces for ore valuation, Proceedings of the Symposium on Mathematical Statistics and Computer Applications in Ore Valuation, 1966, 13-38.

[4] Schamm, K., Ziese, M., Becker, A., Finger, P., Meyer-Christoffer, A., Schneider, U., Schröder, M. and Stender, P.: Global gridded precipitation over land: a description of the new GPCC First Guess Daily product, Earth System Science Data, 2014, Vol. 6, Is. 1, 49-60.

[5] Mestre, O.: Correcting climate series using ANOVA technique, Proceedings of the Fourth Seminar for Homogenization and Quality Control in Climatological Databases, 2004, 93–96.

[6] Ziese, M.; Schneider, U.; Meyer-Christoffer, A.; Schamm, K.; Vido, J.; Finger, P.; Bissolli, P.; Pietzsch, S.; Becker, A. :The GPCC Drought Index - a new, combined and gridded global drought index, Earth System Science Data, 2014, Vol. 6, Is. 2, 285-295

[7] McKee, T.B., Doesken, N.J. and Kleist, J.: The Relationship of Drought Frequency and Duration to Time Scales, Eighth Conference on Applied Climatology, 1993.

[8] Vicente-Serrano, S. M., Begueria, S. and Lopez-Moreno, Juan I.: A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index, Journal of Climate, 2010, Vol. 23, Is. 7, 1696-1718.

[9] Fan, Y. and van den Dool, H.: A global monthly land surface air temperature analysis for 1948 –present, Journal of Geophysical Research, 2008, Vol. 113, D01103.

New 1981–2010 climatological normals for Croatia and comparison to previous 1961–1990 and 1971–2000 normals

[full paper]

Irena Nimac and Melita Perčec Tadić
Meteorological and Hydrological Service of Croatia
Zagreb, Croatia
irena.nimac@cirus.dhz.hr

Abstract— **Mean values of climate parameters (climatological normals) provide an insight into the climate characteristics of the region. Comparison of climate parameters for different 30-year periods can gain an insight into the stability of climate conditions of some area or their variability may be an indication of climate change. In this paper, daily precipitation amount and daily temperature data from 20 meteorological stations in Croatia are used to calculate climatological normals for three 30-year periods (1961–1990, 1971–2000 and 1981–2010). Although Croatia is relatively small country, large topographic variety, openness towards Pannonian Plain and position along the Adriatic Sea define different regions, so the selection of stations is in accordance with that. Spatial distribution of annual and seasonal climatological normals of temperature and precipitation amounts is shown for each period. Relative changes of temperature and precipitation amounts between three 30-year periods are calculated and presented at annual and seasonal scales. Important temperature and precipitation indices, like number of cold or warm days and number of days above some precipitation threshold are discussed in the climate change context.**

Keywords— climatological normals; temperature; precipitation; indices; climate change.

I. Introduction

Climate is defined through the mean or the variability of weather conditions expressed by mean values, extremes and variability of climate variables in a longer period over some area. As defined in [1], climatological normal (CLINO) is the average value of climatological data computed for relatively long period, at least 30 years, while normals calculated for following consecutive 30-year periods, 1901–1930, 1931–1960, etc are called climatological standard normals.

In calculating climatological normal for a station, data record for 30 year period should meet certain requirements like homogeneity and completeness. Regarding homogeneity; changes in location, instruments or observation procedures that would influence result are unwanted, while regarding completeness; records should be complete with no missing values since normals calculated from incomplete datasets can be biased. In case of missing data it is also mentioned [2] that normals should be calculated only when values are available for at least 80% of the years of record with no more than three consecutive missing years.

A 30-year period is used as long enough to filter out short-term interannual fluctuations or anomalies, but sufficiently short to be able to show long term climatic trends. Besides being a measure to which recent or current observations can be compared, climatological normals also serve as estimate of conditions most likely to be experienced at given location [3]. Heaving in mind those two purposes, it is obvious that when normals are used as a reference it is convenient to have standard period that is not changing frequently. On the opposite, when estimating the most expected value for the climatic element that is experiencing a trend, like e.g. temperature during recent decades, predictive accuracy is improved by updating the averages frequently.

Many studies have shown that 30 year is not optimal averaging period for normals when used for estimation of the most probable value of the climatological element. Optimal period for temperatures is often shorter, while for precipitation is greater than 30 years [3]. In global climate change research [4] where shifts in some teleconnection indexes and shifts in Köppen climates were analyzed, the optimal averaging period of at least 15 years is statistically suggested as representative for climatological mapping. Research was based on CRU monthly mean temperature and precipitation data set on 0.5° spatial resolution. Following that results, the 25-years moving averages were applied to the observed and projected monthly temperature and precipitation data from four emission scenarios producing 176 Köppen-Geiger climatic maps [5] used for estimating the effect of climate change on patterns of climate classes.

Last few decades climate change is important topic, not only among atmospheric scientists, since climate change affects human health, agriculture, industry and tourism. Further on, it is not easy to assess its influence or to validate and compare the different climate models outputs. It has been demonstrated that climate classes can be a used to verify the

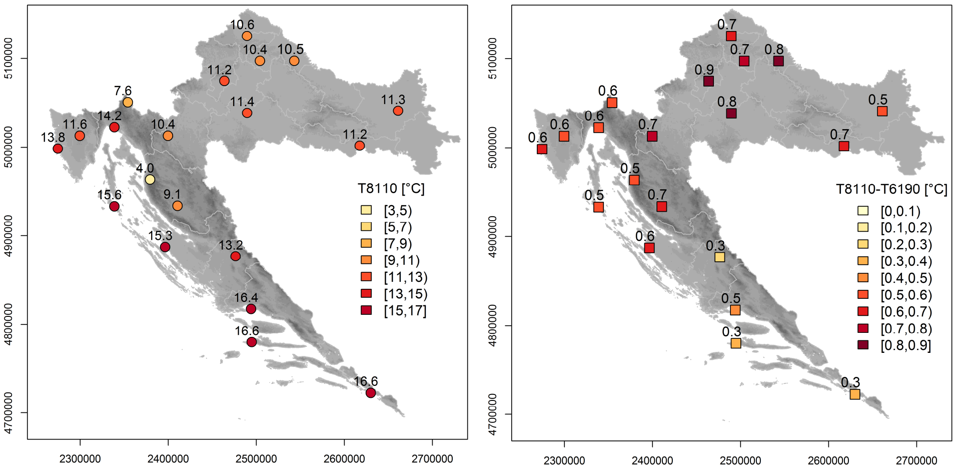


Fig. 1. Temperature climatological normals for latest period 1981–2010 (left) and differences from the standard 1961–1990 period (right). Statisticaly significant differences at 95 % confidence level are marked with sqares.

output of GCM [6] or to serve for model intercomparison to investigate a climate change in Europe [7].

II. Data and Methods

A. Area of interest

Climate in Croatia is determined by its geographical position and medium and large atmospheric circulation systems that does not affect equally all parts of the country [8]. Main climate modifiers are Adriatic and Mediterranean Sea, orography of Dinarids with its shape, altitude and position towards the prevailing flow and openness of the northeastern parts towards Pannonian Plain. Therefore, three main climatic regions in Croatia are continental, mountain and maritime. Mainland of Croatia has moderate continental climate with variety of weather situations and frequent and intensive changes during the year, as it is in a circulation area of mid-latitudes. At higher altitudes climate is mountain and differs from wider area by its lower air temperatures and heavy snowfall. Coastal area is also in mid-latitudes circulation zone, but in summer this area comes under the influence of the subtropical zone as Azorean anticyclone prevents cold air outbreaks to the Adriatic.

According to Köppen climate classification, defined by mean annual temperature course and precipitation amount, most of the Croatia has warm temperate rainy climate with temperature of the coldest month higher than -3°C and lower than 18°C (symbol *C*) [8]. Only highest mountainous areas (>1200 m asl) have a snow-forest climate with average temperature of the coldest month below or equal to -3°C (symbol *D*). Continental part has *Cfwbx"* climate with average temperature of the warmest month of the year below 22°C (war summer, *b*), no extremely dry months during year *(f)* while month with the lowest precipitation amount is in the cold part of the year (*w*) and two maxima in the annual precipitation course (*x"*). Lower mountainous areas have climate class *Cfsbx"* with the highest monthly precipitation in the cold part of the year (*s*), while highest areas have *Dfsbx"* climate. In accordance with mentioned diversity of climate conditions, selection of the stations used in this study is representative.

B. Data and methods

Data used in this study are daily data from 20 meteorological stations which are part of meteorological network of Meteorological and Hydrological Service of Croatia (DHMZ). Selected stations have time series long enough to cover observed 50-year period. Within activities of DHMZ, regular processing and control of the data is carried out. Daily mean temperature is calculated from measurements at 7 am, 2 pm and 9 pm local time, while precipitation daily amounts are measured at 7 am. Few stations had some missing data which were interpolated according to the records of the nearby representative station.

In this work focus is on the analysis of changes in climatological normals of temperature and precipitation amount for three different 30-year periods: 1961–1990 (6190), 1971–2000 (7100) and 1981–2010 (8110). Statistical significance of the temperature change between latest and first observed period will be tested with two sample Student's *t*-test for differences of mean, while for precipitation significance will be tested with Wilcoxon rank-sum test which is nonparametric [9]. Due to 10-year overlapping between those two periods, one of the assumptions that require independence is not fully satisfied. Beside changes on the annual scale, we will examine the seasonal changes and changes in some temperature and precipitation indices, such as number of cold and warm days and number of days above certain precipitation threshold.

Besides statistical measures applied on the temporal series, the long-term temperature and precipitation series are the basis for determining of Köppen-Geiger climate classes. This has been done in R statistical framework [10] using the ClimClass library [11].

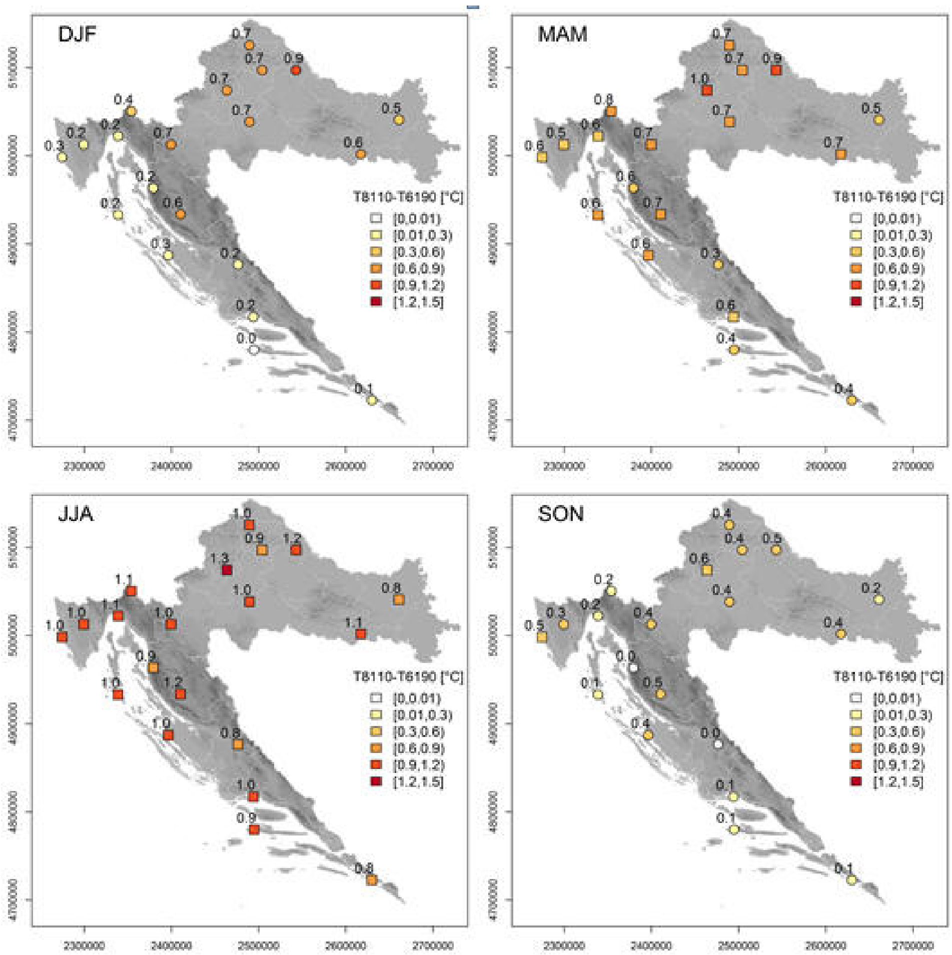


Fig. 2. Seasonal changes in temperature climatological normals between latest 1981–2010 and standard 1961–1990 period for winter (DJF), spring (MAM), summer (JJA) and autumn (SON). Statisticaly significant differences at 95 % confidence level are marked with sqares.

III. RESULTS

A. Temperature

The factors that mostly influence air temperature are the ground surface by either warming or cooling the air, as well as heat radiation of the air itself. Therefore, beside general atmospheric circulation and geographical latitude, spatio-temporal characteristics of air temperature in Croatia are mainly influenced by land-sea distribution because of the difference in heat accumulation, and by elevation. Hence, the highest average 30-year values of temperature, up to 16.6°C, occur at southern coastal area where sea is warming the air in the winter, while the lowest temperatures (4°C) are in the mountains because of temperature decrease with height (Fig. 1 left).

Comparison of mean temperature for latest normal period (T8110) to the standard climatological period 1961–1990 (T6190) shows significant increase in mean temperature at all observed stations (Fig. 1 right). Changes are larger in continental mainland, while lower differences occur at the coast, especially at southern part. Such result is in agreement with stronger heating of the land, while response on the coast is weaker due to more inert changes of the sea temperature. Comparing average temperatures differences between two consecutive periods (7100-6190 and 8110-7100), it is obtained that warming during the latter period was stronger (Tab. 1).

TABLE I. AVERAGE DIFFERENCES IN MEAN ANNUAL AND SEASONAL TEMPERATURE, MEAN MAXIMAL AND MINIMAL TEMPERATURE AND AVERAGE CHANGES IN NUMBER OF DAYS WITH TEMPERATURE ABOVE OR BELOW CERTAIN TRESHOLD

	Comparing periods		
	7100-6190	***8110-7100***	***8110-6190***
ΔT_ANN [°C]	0.2	0.4	0.6
ΔT_DJF [°C]	0.5	0.0	0.4
ΔT_MAM [°C]	0.2	0.4	0.6
ΔT_JJA [°C]	0.4	0.6	1.0
ΔT_SON [°C]	-0.1	0.4	0.3
Δtmax [°C]	0.3	0.4	0.7
Δtmin [°C]	0.2	0.4	0.6
Tmin ≤ -10°C	-1.3	0.3	-1.0
Tmax < 0°C	-1.9	0.0	-1.8
Tmin < 0°C	-1.2	-1.9	-3.1
Tmax ≥ 25°C	3.6	6.4	10.0
Tmax ≥ 30°C	2.8	5.9	8.7
Tmin ≥ 20°C	1.8	3.5	5.2

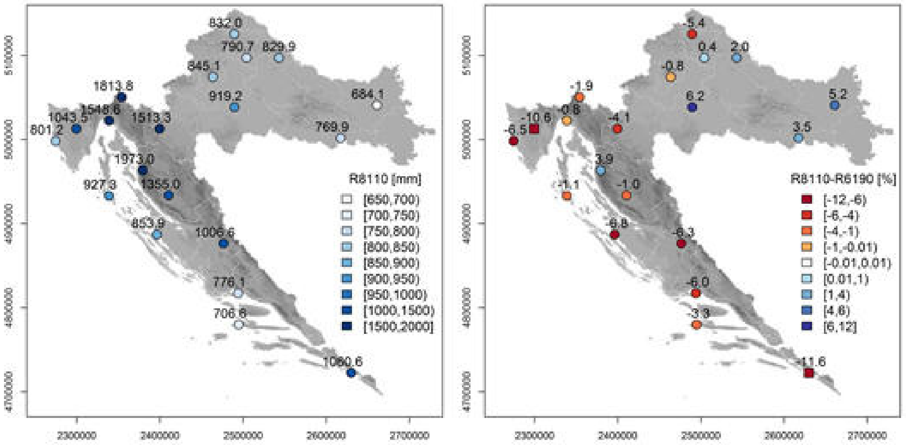


Fig. 3. Precipitation climatological normals for latest period 1981–2010 (left) and relative differences comparing to standard 1961–1990 period (right). Statisticaly significant differences at 95 % confidence level are marked with sqares.

Such result is in accordance with [12] where decadal temperature trends in period 1961–2010 from 41daily temperature and 137 daily precipitation series were examined. The increase in mean annual temperature was shown significant at all stations, with stronger warming in continental area than along the coast.

Looking at seasonal changes of T8110 comparing to T6190, the temperature increased in all seasons and at almost all stations (Fig. 2). Temperature change is the largest in summer, around 1°C, and significant at all stations. Spring temperature increase is also significant at most of the stations, but weaker than in summer, from 0.4–1.0°C. Such increase is result of successive warming, stronger in later periods (Tab. 1).

In colder part of the year, autumn and winter, increase is stronger inland but those changes are not statistically significant. In [12], seasonal analysis of decadal trend showed that in summer and spring warming is significant at most of the stations, while winter increase is significant only in central inland and at few stations of coastal hinterland.

Another indicator of warming is increase in mean maximal and minimal air temperature (Tab. 1), larger between two more recent 30-year periods, 8110-7100. Spatial distribution of changes (not shown) suggests larger amplitudes in inland and mountains, and lower at coastal region, which is also in accordance with results from [12].

Some of temperature indices were also examined such as number of tropical nights (T_{min}≥20°C), warm days (T_{max}≥25°C), hot days (T_{max}≥30°C), icy days (T_{min}<0°C), frosty days (T_{max}<0°C) and days with minimal temperature below or equal -10°C. Due to larger occurrence of tropical nights and hot days along the coast, and icy and frosty days in mountainous and continental area, amplitudes of changes follow that pattern. Analysis of changes of different temperature indices between latest and standard 30-year period shows increase in warm and decrease in cold indices (Tab. 1). Comparison of consecutive periods shows that for warm indices such positive changes are result of successive increase, larger between latest periods, while for some cold indices slight increase occurred at few stations comparing last two 30-year periods.

B. Precipitation

Precipitation in Croatia is mostly result of passing cyclones and related atmospheric fronts which does not affect equally all parts of country [8]. Occurrence, amount and location of precipitation depend of humidity of the air mass, intensity and direction of the air current, and of the vertical component of its movement. Formation of the clouds and development of precipitation can be significantly intensified by some local factors such as distance from the sea and orography. While orography of the Dinarids represents an obstacle for maritime air masses moving toward mainland and for continental masses moving toward the coast, at the same time it can enhance convection and formation of clouds and intensified precipitation. In [13], they analyzed daily precipitation amount of wider Alpine region with purpose of making high resolution maps (5 km). Large area of Croatia was included in analysis, except most southern Dalmatia and eastern part of Slavonia.

TABLE II. AVERAGE RELATIVE DIFFERENCES IN ANNUAL AND SEASONAL PRECIPITATION AMOUNT AND AVERAGE CHANGES IN NUMBER OF DAYS WITH DAILY PRECIPITATION AMOUNT ABOVE CERTAIN TRESHOLD

	Comparing periods		
	7100-6190	***8110-7100***	***8110-6190***
ΔR_ANN [%]	-2.1	-0.2	-2.2
ΔR_DJF [%]	-6.2	3.1	-3.2
ΔR_MAM[%]	-4.8	-1.7	-6.4
ΔR_JJA [%]	-5.2	-3.3	-8.2
ΔR_SON [%]	6.0	0.6	6.7
Rd ≥ 0.1 mm	-3.3	-1.6	-5.0
Rd ≥ 1 mm	-3.0	-1.4	-4.3
Rd ≥ 5 mm	-2.0	-0.7	-2.6
Rd ≥ 10 mm	-1.0	-0.1	-1.1
Rd ≥ 20 mm	-0.4	0.5	0.1
Rd ≥ 50 mm	0.1	0.0	0.1

The largest mean precipitation amount for days with precipitation is estimated for the peaks of Dinarids and larger

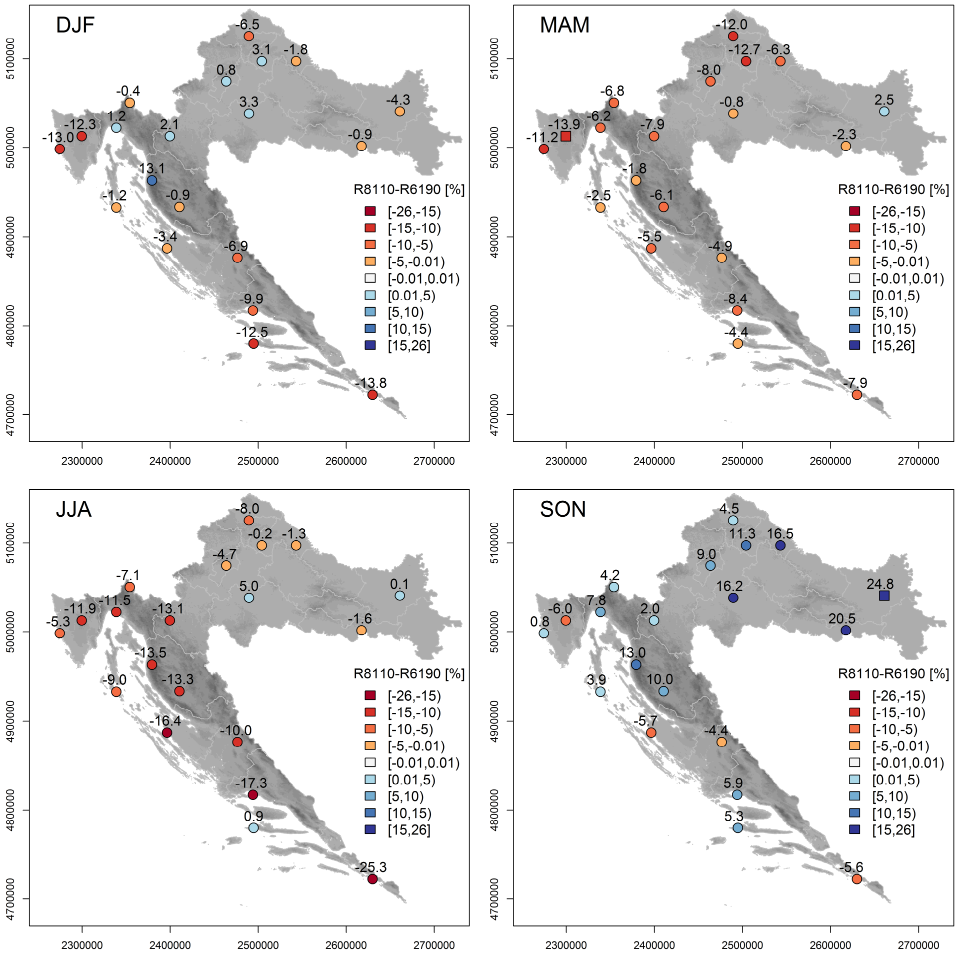


Fig. 4. Seasonal changes in precipitation climatological normals between latest 1981–2010 and standard 1961–1990 period for winter (DJF), spring (MAM), summer (JJA) and autumn (SON). Statisticaly significant differences at 95 % confidence level are marked with sqares.

along the coast compared to inland. Accordingly, the highest values of precipitation amount are at the high mountain stations along the Dinarids with an average 30-year amount over 1500 mm at the selected stations (Fig. 3). Even though it is not possible to appreciate the full spatial variability of precipitation from the 20 selected stations, by keeping in mind the precipitation maps in [8] it is noticeable that precipitation values in the mainland are decreasing eastward as moist air from south-west loses humidity on the way, while those air masses from north-east direction are dry. Looking at coastal area, the distinguished is the most southern station Dubrovnik with the highest precipitation amount due to impact of topography and flow of moist air from the sea with the southern wind. Analysis of changes between precipitation for latest period (R8110) and the first one (R6190), suggests precipitation increase in eastern continental area and mountains and decrease in rest of the area, significant only at most southern station Dubrovnik and in central Istria (Fig. 3). Same results were shown in [14] where they analyzed long-term decadal trends in regional time series for period 1961–2010 from 132 stations. They obtained positive trend in eastern mainland and negative trends in other regions, significant only in mountainous region. Comparing consecutive periods shows that average changes are larger between first two periods (Tab. 2) which is result of larger decrease at the coast between those two and stronger continental increase between latest periods (not shown). Annual course of precipitation in Croatia has bimodal shape. For continental area, maximums occur in summer and autumn as a result of stronger ground base warming, which enhances convection, and the advection of cold and moist air from northwest direction [15]. Along the coast, larger amount of precipitation are in colder part of the year, in autumn and winter. Comparison of R8110 and R6190 shows various changes for each season (Fig. 4). In warmer part of the year, spring and summer, the precipitation decreased at most of the stations, but significant is only at central Istria in spring. For both mentioned seasons, decrease between first two periods (7100-6190) is larger and occurs at almost all stations (Tab. 2), while between last two periods increase is obtained at few stations in eastern part of country (not shown). For winter, decrease in precipitation appears along the Adriatic coast and in eastern Croatia, while on few continental and mountain stations the precipitation increased. Changes between successive periods show for winter different pattern (Tab. 2). While comparing first two periods shows decrease in precipitation all over the country, comparison of last two gives increase at

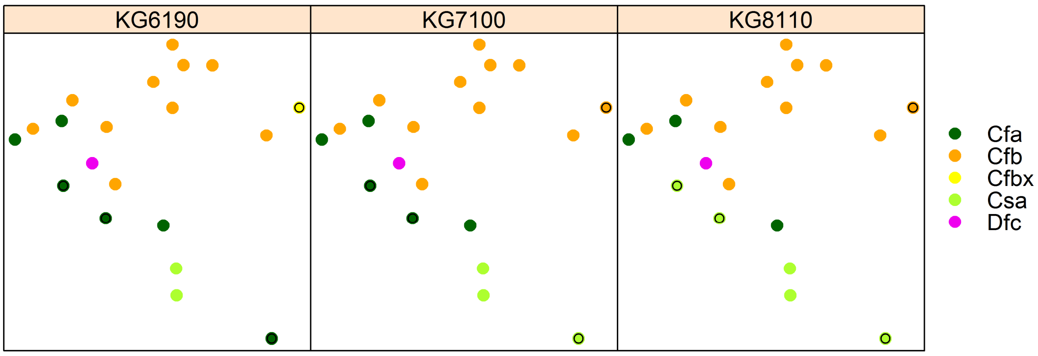


Fig. 5. Köppen-Geiger climate classes for 1961–1990 (KG6190), 1971–2000 (KG7100) and 1981–2010 (KG8110) period. Stations where climate class changed are circled.

relatively all stations. In autumn, increase occurs at most of the stations, with the largest amplitude and only one significant is in most eastern continental part. Analysis of consecutive periods showed successive increase in inland, with larger differences between first two, and decrease in precipitation in the coastal area for latest periods. Such results indicate equalization of summer and autumn maximum in continental area and increase in differences between autumn and winter maxima along the coast. Obtained seasonal changes are in accordance with [8] where they found prevailing decreasing trend in almost all seasons, significant in summer in central hinterland, mountainous and mountainous littoral region. In autumn they found increase in mainland being statistically significant in most eastern parts.

It is interesting to mention that at seven coastal and mountainous stations minimal annual precipitation amount was registered for the same year, 1983. The largest registered daily precipitation amount (352.2 mm) is registered at the station Zadar in 1986.

The highest occurrence of days with precipitation over 0.1 mm, 1 mm and 5 mm is in high mountains, decreasing toward east, and the lowest along the coast (compare with Fig. 7a in [13]). For number of days with stronger precipitation (more than 20 and 50 mm) largest values are also in Dinarids, while the lowest are now in continental part. While in the mainland the days with greater daily precipitation amount are often result of short-term convective precipitation that occurs more often during summer, in the highest areas and along the coast strong precipitation is result of long-term precipitation usually during cold part of the year. Comparing latest and first 30-year period, at almost all stations decrease in number of days with daily precipitation amount above 0.1, 1 and 5 mm is occurred. Such distribution is result of successive decrease with larger amplitudes between first two periods (Tab. 2), except for number of days with precipitation above 5 mm where comparing last two periods increase in mainland occurs. Number of days with precipitation amount above 10 and 20 mm is increased in continental inland and relatively decreased in the coastal area, also as result of successive increase/decrease comparing consecutive periods.

C. Köppen- Geiger climate classes

The presented changes in temperature and precipitation regimes can result in change of Köppen-Geiger climate classes [5]. During the 1971–2000, climate class changed at first at the most southern station Dubrovnik where *Cfa*, warm temperate climate with hot summers, changed from fully humid (*f*) to dry summer precipitation regime, *Csa* (Fig. 5). The most eastern continental station Osijek lost its slightly wettest beginning of the summer compared to the summer's end of the warm temperate fully humid climate, experiencing a change from *Cfbx* to *Cfb*. In the next period 1981–2010 northern Dalmatian island station Mali Lošinj and coastal station Zadar, experienced a change from fully humid Cfa to dry summer precipitation regime, *Csa*, like Dubrovnik already did before.

Not shown in this article, but available from the connected study, is the signal on the Puntijarka mountain near Zagreb, with mean temperature of the coldest month slightly below -3°C, where the increase in temperature resulted in change from snow-forest climate *Dfb* to warm-temperate climate *Cfb*.

IV. CONCLUSION

The analysis of climatological normals of temperature and precipitation for different time periods has given various results depending on region and season. Comparing latest and the standard climatological period significant increase in temperature is observed at all selected stations. Largest warming occurs in continental part while the lowest amplitudes of changes are in southern Adriatic. Seasonal analysis of temperature changes between recent and first 30-year period shows increase for all seasons, significant only in warmer part of the year. While spring and summer increase is result of successive warming, relatively larger between recent periods, autumn and winter increase is moderated with relative temperature decrease between first two, i.e. last two periods. Increase in mean minimal and maximal temperature and warm indices as well as decrease in cold indices confirms warmer climate in Croatia for recent period.

Precipitation response shows dual nature; increase in eastern part of the country and decrease along the coast. Main contribution to eastern continental precipitation increase is in autumn, when most of the selected stations observe positive changes. This will be discussed further in connection to change in climate classes in eastern continental region. Precipitation is increased in winter at highest station Zavižan and at few central inland stations. In spring and summer at almost all stations precipitation amount is decreased, especially summer precipitation in middle and southern Adriatic. This affected also the change in climate classes in this region. Also, decrease in Istria is expressed in all season except autumn. Such seasonal changes indicate increase in differences between two maximums, autumn and winter on the coast and autumn and summer in the inland. Analysis of number of days with precipitation amount above some threshold gave decrease in number of days with daily precipitation amount above 0.1 mm, 1 mm and 5 mm at almost all stations. For number of days with daily precipitation above higher threshold, increase occurred in continental inland, while along the coast precipitation is relatively decreased.

Changes in Köppen-Geiger climate classes happened mainly due to changes in precipitation regime. These changes are consistent with decrease in precipitation, especially along Dalmatian coast during summer, resulting in change from warm temperate climate with hot summers and fully humid precipitation regime to dry summer precipitation regime (change from *Cfa* to *Csa*). This was confirmed on several other stations not included in this study. Further on, there are several eastern continental stations, including Osijek, with characteristic more humid May and July than August and September (*x* letter in *Cfbx* comes from $R_{May} + R_{July} \geq 1.3\ R_{August} + R_{September}$) during 1961–1990 period. Due to increased precipitation in autumn in this region, those stations lost this characteristic wettest beginning of the summer compared to its end and became more similar to the rest of the Croatian stations where these precipitation amounts are more similar in those two periods.

Not shown in this article, but available from the previous study, is the signal on the Puntijarka mountain with mean temperature of the coldest month slightly below -3°C where the increase in temperature resulted in change from snow-forest climate to warm-temperate climate, due to change in temperature regime.

The presented analysis indicates significant changes in climatological normals of temperature and precipitation in Croatia. Spatial dependency of climate response due to different climate factors is noticeable. Köppen-Geiger climate classes, as one of the indices that can describe the climate patterns associated with global circulation [4], showed to be able to recognize the climate change signal. The most prominent change is from fully humid to summer dray warm temperate climate with hot summer in Dalmatia. Noticeable change is losing more humid beginning of the summer compared to the end in eastern continental Croatia. For a comparison with a global picture, the digital Köppen-Geiger world map [16] on climate classification, valid for the second half of the 20th century is available from http://koeppen-geiger.vu-wien.ac.at/.

ACKNOWLEDGMENT

This work has been supported in part by Croatian Science Foundation under the project 2831 (CARE).

REFERENCES

[1] WMO, Technical Regulations, vol. 1, Geneva: WMO-No.49, 1988.

[2] WMO, Guide to Climatological Practices, Third edition, WMO- No.100, Geneva, 2011.

[3] WMO, The Role of Climatological Normals in a Changing Climate, WMO/TD-No. 1377, WCDMP-No. 61, Geneva

[4] K. Fraedrich, F.-W. Gerstengarbe and P.C. Werner, “Climate shifts during the last century“, *Clim. Change*, 50, 405–417, 2001.

[5] F. Rubel, M. Kottek, “Observed and projected climate shifts 1901-2100 depicted by world maps of the Köppen-Geiger climate classification“, *Meteorol. Zeitschrift,* 19, 135–141, 2010.

[6] S. Manabe, J.L. Holloway, “The seasonal variation of the hydrologic cycle as simulated by a global model of the atmosphere”, *J. Geophys. Res.* 80, 1617–1649. doi:10.1029/JC080i012p01617, 1975.

[7] M. De Castro, C. Gallardo, K. Jylha, H. Tuomenvirta, “The use of a climate-type classification for assessing climate change effects in Europe from an ensemble of nine regional climate models”, *Clim. Change* 81, 329–341. doi:10.1007/s10584-006-9224-1, 2007.

[8] K. Zaninović, M. Gajić-Čapka, M. Perčec Tadić et al, Climate atlas of Croatia 1961–1990., 1971–2000., Zagreb: Državni hidrometeorološki zavod, 2008.

[9] D. S. Wilks, Statistical Methods in the Atmospheric Sciences: An Introduction, Academic Press, 1995.

[10] R Core Team. „R: A language and environment for statistical computing“. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2015

[11] E. Eccel, E. Cordano and G. Toller, „ClimClass: Climate Classification According to Several Indices“, R package version 2.0.1. https://CRAN.R-project.org/package=ClimClass, 2015.

[12] M. Gajić-Čapka, K. Zaninović and K. Cindrić, “Climate Chnage Impacts and Adaptation Measures - Observed Climate Change in Croatia,” in *Sixth National Communication of the Republic of Croatia under the United Nation Framework Convention on the Climate Change (UNFCCC)*, Zagreb, Ministry of Environmental and Nature Protection, Physical Planning and Construction, 2014, pp. 100-140.

[13] F.A. Issota et al, ”The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data”, *Int. J. Climatol.,* vol. 34, no. 5, pp. 1657-1675, 2013.

[14] M.. Gajić-Čapka, K. Cindrić and Z. Pasarić, “Trends in precipitation indices in Croatia, 1961–2010,” *Theoretical and Applied Climatology,* vol. 121, no. DOI: 10.1007/s00704-014-1217-9., pp. 167-177, 2014.

[15] E. Lončar and A. Bajić, “The weather types in Croatia”, *Hrvatski meteorološki časopis,* vol. 23, pp. 31-41, 1994.

[16] M. Kottek, J. Grieser, C. Beck, B. Rudolf and F. Rubel, “World Map of the Koppen-Geiger climate classification updated”, *Meteorol. Zeitschrift* 15, 259–263. doi:10.1127/0941-2948/2006/0130, 2006.

Improving seasonal precipitation mapping using GPCC data over the western US

[abstract]

Jelena Luković
Department of geodesy and geoinformatics
Faculty of civil engineering, University of Belgrade
Belgrade, Serbia
jelenalu@yahoo.com

Milan Kilibarda
Department of geodesy and geoinformatics
Faculty of civil engineering, University of Belgrade
Belgrade, Serbia
kili@grf.bg.ac.rs

Branislav Bajat
Department of geodesy and geoinformatics
Faculty of civil engineering, University of Belgrade
Belgrade, Serbia
bajat@grf.bg.ac.rs

***Abstract*—Seasonality is the main characteristics of the western US. In order to estimate participation averages for the period of the 1961-2010 at the seasonal and annual scale, we used Regression Kriging. Cooperative Observer Program (COOP) observations in combination with Global Precipitation Climatology Centre (GPCC) grids for the year 2010 (monthly and annuals) as the predictors were used to obtain the trend model. The five variogram models (four seasonal and one annual) were fitted on regression residuals. The prediction was made at one kilometre spatial resolution over California, Oregon and Washington States. To achieve R-square and RMSE we used live-one-out cross validation procedure and obtained summarised results are: Annual: RMSE=122 mm/year, R2=0.94; Winter: RMSE=282 mm/year, R2=0.89;Spring: RMSE=150 mm/year, R2=0.90; Summer: RMSE=36 mm/year, R2=0.95; * Autumn: RMSE=147 mm/year, R2=0.93; The amount of variation explained by Regression Kriging model is around 90%, and this model is adequate for seasonal gridded estimation of participation averages.**

Earlier onset of spring in Serbia

[extended abstract]

Biljana Basarin, Minučer Mesaroš, Dragoslav Pavić, Tin Lukić
Faculty of Sciences
University of Novi Sad
Novi Sad, Serbia
minucer.mesaros@dgt.uns.ac.rs

Abstract—**This paper describes the analysis of spring onset dates in Serbia for the period 1950-2013, determined by the Ensemble Empirical Mode Decomposition (EEMD) method. The obtained values were visualized using interpolation, highlighting geospatial patterns of spring onset dates. The results show earlier spring onset for all considered meteorological stations, with a trend of increasing statistical significance towards the north-west.**

Keywords—spring onset; Ensemble Empirical Mode Decomposition (EEMD); climate change; geostatistics;

I. Introduction

The timing of the spring season is very important for natural ecosystems and human activities such as agricultural planning. Recent studies determine the earlier onset of spring based on phenological evidence [1]; [2]; [3]. Global warming has been seen as the main cause of the changes in spring onset during the last century [4]. The date of spring in Europe has mainly been correlated to large-scale atmospheric circulation; especially the North Atlantic Oscillation (NAO) for the period mainly from the 1950s to the present day. The timing of the onset of spring is defined by temperatures rising above a certain threshold. The 5ºC air temperature threshold is widely accepted and used for determining climatic spring onset at mid-latitudes [5]; [4]; [6].

In this study the timing of climatic spring onset from the daily temperature records at Serbia during 1950-2013 was determined using the Ensemble Empirical Mode Decomposition (EEMD) method [7]. EEMD was used to isolate the amplitude-frequency modulated annual cycle (MAC) [8] and to determine adaptively the temporally varying trend [9] from a homogenized daily temperature series in Serbia.

II. Material and methods

The study was based on a homogenized daily surface air temperature (SAT) datasets for the territory of Serbia downloaded from European Climate Assessment & Dataset [10]. We used the data from 32 meteorological stations in Serbia. In this study, the spring onset is defined as the date of the first appearance of 5ºC in an adaptively and temporally locally determined low-frequency part of the daily temperature series containing annual cycle and longer timescale component (ALC). The ALC was isolated adaptively and temporally locally using EEMD. The EEMD is a decomposition method that decomposes any complicated data series into small number of amplitude frequency modulated oscillatory components called intrinsic mode functions (IMFs) of different timescales. The steps to obtain ALC by EEMD are as follows:

1. Add a white noise series, with amplitude of 0.3 times the standard deviation of the raw daily data.
2. Decompose the daily data series with added white noise into IMFs using EMD.
3. Repeat step 1 and step 2 for 1000 times.
4. Obtain the (ensemble) means of corresponding IMFs of the decompositions.
5. Combine the sixth to the last ensemble components obtained from step 4 as the final ALC.

For the obtained yearly spring onset dates the average values for the period between 1950 and 2013 were calculated and interpolated using the Radial Basis Function method (spline with tension) in ArcGIS, as shown in fig.1. The values of the annual rate of change in spring onset dates were also interpolated and visualized with the above mentioned method, shown in fig. 2.

A section of five years (1950-1954) of the raw data and its ALC for Novi Sad are displayed in Fig. 3. The adaptively determined ALC fitting to the daily SAT is visually appealing and it has only one intersection with 5ºC line in each spring, thus the onset date of every spring can be uniquely determined. The linear trend of spring onset dates at each station was calculated and the significance of the trend assessed using the Mann-Kendall test. These values were interpolated

III. Results

At all investigated sites the advancing of the spring onset is found. The mean onset date for all investigated sites is at the 74th day (15 March) and most of the onset dates are between the beginning of March and early April. On average, the latest onset is registered at the 100th day (10 April) for Sjenica station while the earliest onset is at the 61st day (2 March) for Belgrade. On the other hand, if individual years were observed it could be seen that the absolute earlier onset was in 2007, when the 5ºC was observed on 1st of January at four investigated stations (Banatski Karlovac, Beograd, Kikinda and

Zrenjanin). During 2007, Serbia experienced one of the strongest heat waves since the meteorological recordings begun [11]. The absolute latest onset was identified for Kopaonik, the station with the highest altitude (1700 m).

Fig. 1 shows the average dates of the spring onset for all investigated stations. The map indicates earlier spring onset in the south-eastern parts of Vojvodina region, while the latest onset can be observed in the mountainous regions of southwestern parts of central Serbia.

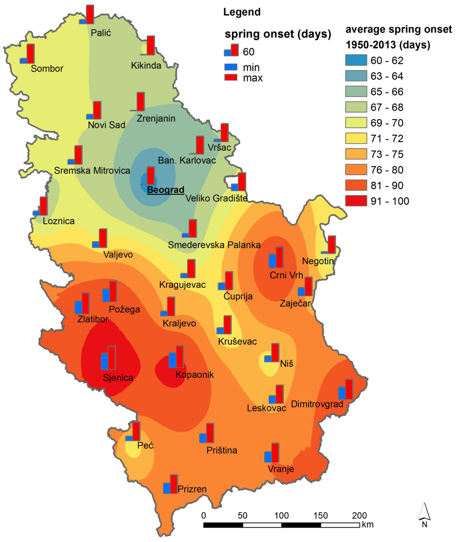


Fig. 1. Average spring onset dates 1950-2013.

The amplitudes between the earliest and latest spring onset date shown with blue and red bars on the map also follow a similar general pattern with lowest amplitude for higher elevation and higher amplitude for highest elevation. An exception from this pattern can be observed in the Southern Morava river valley, where the amplitudes are similar to those in higher elevations.

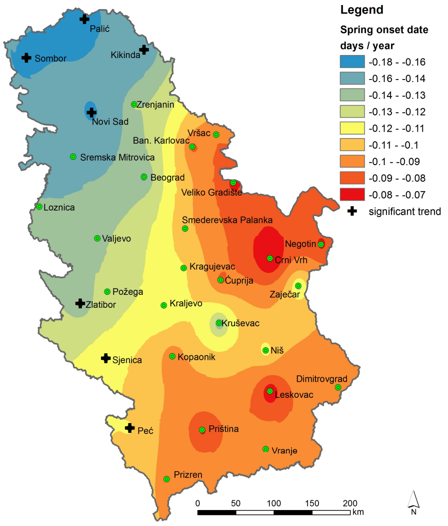


Fig. 2. Annual rate of change in spring onset date (days). Stations showing statistically significant trend are marked with+

The linear trends are significant under Mann-Kendall's test ($p<0.05$) for only seven investigated sites. On the other hand at additional eight sites the linear trends become significant under reduced statistical confidence ($p<0.1$) (Fig. 2).

A. *Example of Novi Sad*

A piece of five years (1950-1954) of the raw temperature data and its ALC are displayed in Fig. 3. The ALC is fitted to the daily temperature data, it has only one intersection with 5ºC line in each spring, thus the onset date of every spring can be uniquely determined.

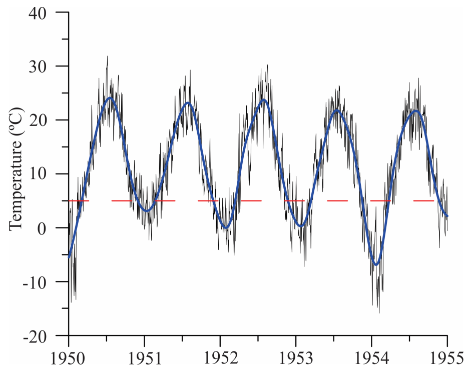


Fig. 3. The raw SAT data (black) at Novi Sad and the sum of its annual cycle and longer timescale component (blue line) for 1950–1955. The red line indicates 5ºC threshold.

The onset of spring for each year at Novi Sad is shown in Fig. 4. The overall linear trend is 0.202 d/yr using EEMD and 0.23 d/yr using the 30-day running mean. Both trends are significant at $p<0.05$ using the Mann-Kendall trend test.

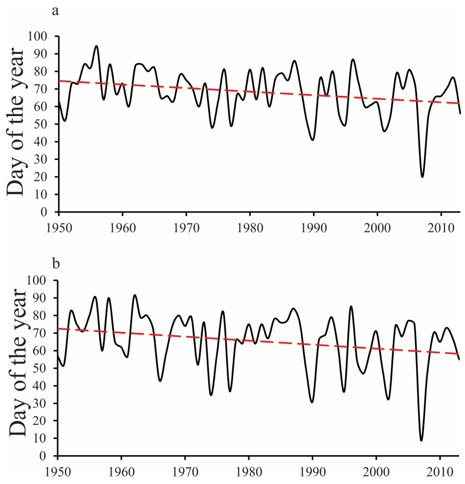


Fig. 4. Spring onset dates associated with the annual cycle component obtained using EEMD and its linear trend line; b. spring onset dates determined using the 30-day-window running mean of daily temperature series.

The results displayed in Fig. 3a show that the year-to-year difference in spring onset dates associated with the modulated annual cycle can be as large as 74 days (20 January for 2007 and 3 April for 1956). The changes in spring onset could be related to those in the phase of the annual cycle and to the warming trend. The combination of the annual cycle and the decadal warming trend for 1956 is compared with those for 2007 in Fig. 5.

The differences between the annual cycles for these two years are quite large. This result implies a significant impact of phase change in the annual cycle on the interannual variability in spring onset. But the earlier spring onset must also be attributable to the warming trend. Also, some studies suggest a strong connection between NAO phase and earlier onset of spring in the second half of the 20th century [4].

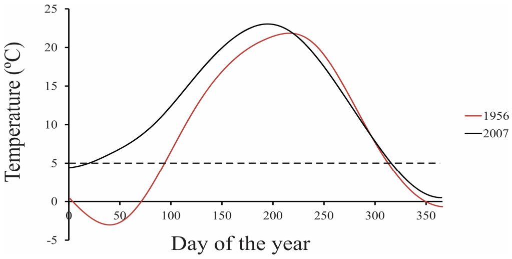


Fig. 5. Spring onset dates for 1956 (red line) and 2007 (black line) the change in the phase of the annual cycle is clearly visible.

IV. Discussion and Conclusion

The spring onset dates are advancing in Serbia which is in good agreement with the studies done worldwide [4];[12]. Spatially, a zonal orientation can be seen in general, i.e. the higher the latitude, the later the onset of spring. The topographic setting plays an important role for spring onset dates, but it is not a predominant factor. An example for this are the meteorological stations in Southern Morava valley, where the observed amplitudes are lower than for the stations with the similar elevation (Fig.1). Possible reasons for this are local factors and conditions that require further investigation.

The results obtained for Novi Sad are consistent with the estimation of [13] who showed that, based on trends in phenological phases in Europe between 1951 and 1996, spring events, such as leaf unfolding, have advanced on average by 6.3 days or –0.21 d/yr. Similarly, [14] emphasized that the most prominent temperature driven changes in plant phenology are an earlier start of spring in the last three to five decades of, on average, 0.25 days/decade, mainly observed in midlatitudes and higher latitudes of the northern hemisphere. Using the modulated annual cycle in Serbia the results of linear trends indicate that the start of the spring is advancing between 0.1 and 0.25 d/yr during the last 63 years. For the Czech Republic [15] found that the prevailing pattern across tree species since around the year 1976, was a consistent advancement of the onset of spring phenophases (leaf unfolding and flowering) and subsequent acceleration of fruit ripening, and a delay of autumn phenophases (leaf coloring and leaf falling).

Similarly to our results many studies suggest that both atmospheric variability and the warming trend play an important role in the earlier onset of climatic spring recent decades. These findings are consistent with the result of similar studies. These studies suggest that the secular trend of earlier onset of spring is dominated by the warming trend, but in recent decades, the NAO is also highly correlated with spring onset at Stockholm [4]. Further studies are needed in order to better understand the advancing trend of spring onset in Serbia

Acknowledgment

This research was supported by Project number 176020 and 43002 of the Serbian Ministry of Education, Science and Technological Development.

References

[1] H.W. Linderholm, Growing season changes in the last century, Agric. For. Meteorol., vol. 137, pp. 1-14, 2006.

[2] M.D. Schwartz, R. Ahas, and A. Aasa, Onset of spring starting earlier across the Northern Hemisphere, Global Change Biology, vol. 12, no. 2, pp. 343-351, 2006.

[3] J. Zheng, Q. Ge, Z. Hao, and W.-C. Wang, Spring phenophases in recent decades over Eastern China and its possible link to climate changes, Climatic Change, vol. 77, pp. 449-462, 2006.

[4] C. Qian, C. Fu, Z. Wu, and Z. Yan, On the secular change of spring onset at Stockholm, Geophys. Res. Lett., vol. 36, L12706, doi: 10.1029/2009GL038617, 2009.

[5] H.W. Linderholm, A. Walther, and D. Chen, Twentieth-century trends in the thermal growing season in the Greater Baltic Area, Climatic Change, vol. 87, pp. 405-419, 2008.

[6] Y. Song, H.W. Linderholm, D. Chen, and A. Walther, Trends of the thermal growing season in China, 1951–2007, Int. J. Climatol., vol. 30, pp. 33-43, 2010.

[7] Z. Wu, and N.E. Huang, Ensemble Empirical Mode Decomposition: A noise-assisted data analysis method, Advances in Adaptive Data Analysis, vol. 1, no. 1, pp. 1-41, 2009.

[8] Z. Wu, E.K. Schneider, B.P. Kirtman, E.S. Sarachik, N.E. Huang, and C.J. Tucker, The modulated annual cycle: An alternative reference frame for climate anomalies, Climate Dyn., vol. 31, pp. 823-841, 2008.

[9] Z. Wu, N.E. Huang, S.R. Long, and C.-K, Peng, On the trend, detrending, and variability of nonlinear and nonstationary time series, Proceedings of the National Academy of Sciences, USA, vol. 104, no. 38, pp. 14889-14894, 2007.

[10] M.R. Haylock, N. Hofstra, A.M.G. Klein Tank, E.J. Klok, P.D. Jones, and M. New, A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, Journal of Geophysical Research (Atmospheres), vol. 113, D20119, doi:10.1029/2008JD10201, 2008.

[11] M. Unkašević and I. Tošić, Seasonal analysis of cold and heat waves in Serbia during the period 1949–2012. Theor Appl Climatol. vol. 120, no. 1, pp. 29-40, 2015.

[12] C. Qian, C. Fu, Z. Wu, and Z. Yan, The Role of Changes in the Annual Cycle in Earlier Onset of Climatic Spring in Northern China. Advances In Atmospheric Sciences, vol. 28, no. 2, pp. 284-296, 2010.

[13] A. Menzel, Trends in phenological phases in Europe between 1951 and 1996. Int J Biometeorol. vol. 44, no. 2, pp. 76-81, 2000.

[14] A. Menzel, N. Estrella, and C. Schleip, Impacts of Climate Variability, Trends and NAO on 20th Century European Plant Phenology. In (ed) Brönnimann S, Luterbacher J, Ewen T, Diaz HF, Stolarski RS, Neu In: Climate Variability and Extremes during the Past 100 Years. Springer Netherlands, 2008.

[15] E. Kolářová, J. Nekovář, and P. Adamík, Long-term temporal changes in central European tree phenology (1946–2010) confirm the recent extension of growing seasons. Int J Biometeorol, DOI 10.1007/s00484-013-0779-z, 2014.

Correlating local palynological record and weather conditions for an individualized pollen alarm

[extended abstract]

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

Mirjana Mitrović Josipović
Ministry of Agriculture and Environmental Protection
Environmental Protection Agency
Belgrade, Serbia
mirjana.mitrovic@sepa.gov.rs

Bojana Božanić
Jantar Group
Belgrade, Serbia
bojanabozanic@gmail.com

Vít Pászto
Palacký University Olomouc,
Faculty of Science
Olomouc, Czech Republic
vit.paszto@gmail.com

Lukaš Marek
University of Canterbury,
Department of Geography
Christchurch, New Zealand
lukas.marek@canterbury.ac.nz

***Abstract*—Pollen is affecting more and more people in Serbia, causing seasonal health disorders. Current health services in Serbia provide a general pollen alarm (www.sepa.gov.rs) which does not involve individualized pollen alarms. On the other hand, the voluntary database (www.pollendiary.com) contains individualized records of daily symptoms of patients from Serbia. The idea of this research is to explore the possibility of individualizing pollen conditions using publicly available and voluntary data, thereby relating weather conditions and pollen concentrations in the air, as well as the individual prognosis of daily symptoms and reactions to pollen. Prognosis is based on SVM regression, wherein several aspects of learning protocol are tested. The latter included examining the influence of learning sample size, attribute selection and meta-regression. Results indicate relatively good potential of using SVM regression in learning individual pollen alarm trends.**

Keywords— pollen; palynology; weather data; predictors; Machine Learning; SVM; regression;

I. Introduction

Pollen is considered the main trigger of respiratory allergies [1]. The number of people suffering from such conditions grows worldwide. The ongoing climate change further accelerates dispersion of allergenic pollens. Allochtonous and invasive plants are particularly threatening, e.g. Common ragweed (*Ambrosia artemisiifolia*) [2]. This is why it is of great importance to improve the knowledge of pollen-induced allergies using better availability of data and multidisciplinary approach.

Similar trends are found in Serbia, where pollen is monitored by governmental agency for environment (www.sepa.gov.rs). Available data show that 8.8% of total population suffered from some kind of allergy excluding allergic asthma in 2013 [3], while in 2006 this percentage was much lower – 5.3% [4]. Other source shows that percentage of patients oversensitive to *Ambrosia* pollen increased from 51% to 83.7% in Serbia through 2000-03 [5].

Principal plant species causing pollinosis are more-or-less the same throughout Europe. Depending on species, and local weather condition, start of the season, peak, duration, and pollen concentrations may vary [6, 7]. All aforementioned phenological characteristics are under the direct influence of global climate change [8]. Several researchers addressed possibilities of predicting pollen concentrations in Europe in respect to weather conditions. They used meteorological data such as temperature [6,9], rainfall/precipitation [10], humidity [11], solar radiation and wind speed and direction [12] or their combination, for correlating pollen dispersion with weather predictors. Only few of the studies used additional data based on patients' symptoms to address pollen allergies [13, 14].

Herein, weather predictors, as well as their autocorrelation indices, were combined with measured pollen concentrations to predict symptoms of pollen allergy of one patient in Belgrade in 2012-14. The principal idea was to predict pollen concentrations from trends of weather data only, which would enable further prediction of personal pollen symptoms of a patient. Thereby, it would be possible to predict symptoms from short-term (e.g. 5-day) weather forecast, without needing any pollen concentration measurement, which is innovative in respect to previous research.

II. Data and Methods

Weather data were collected from publically available archive for 2012-14 from www.hidmet.gov.rs, and included several predictors: averaged air pressure, temperature, relative humidity, wind speed, insolation, overcast and precipitation. It is important to mention that most of these predictors are available in short-term (5-day) weather forecasts at www.hidmet.gov.rs, which opens their potential use in further predictions.

Pollen data included pollen concentrations of 20 plant groups in Belgrade City area (measurement station Zeleno Brdo) for the same interval (2012-14). Eventually only those pollen groups that were correlated with patient symptoms entered final datasets. These included concentrations of: *Ambrosia*, *Betulaceae*, *Cupressacea*, *Moraceae*, *Poaceae* and *Urticaceae*. The data are available for preview at www.sepa.gov.rs, but actual measurements were obtained by the courtesy of www.sepa.gov.rs.

The pollen symptoms data, structured according to www.pollendiary.com standards and covering 2012-14 were voluntarily handed over by one patient.

Support Vector Machine regression (SVMr) was implemented for learning weather-pollen-symptoms trend. Details on SVMr can be found in [15]. Learning experiment included several aspects according to the principal objectives of this research. Apart from raw weather predictors, their autocorrelation indices for appropriate time lags were supplemented. It was expected that these would increase predictive power of the regression algorithm. After filtering and preprocessing the independents and predictors, SVMr algorithm was used in two separate tests. TEST 1 included learning on 2012 data and validating on 2013 data while the TEST 2 used 2012-13 for learning and 2014 for validating. Thereby, learning sample size was explored by comparing regression success between 2013 and 2014. Learning included prediction of trends of particular (target) pollen groups that have been proven correlated with patient's symptoms by using correlation coefficient subset attribute selection. Finally, the learning of symptoms was achieved by as a two-step regression task. Firstly, target pollen groups were predicted in TEST1 and TEST2 experiments only from weather data. Secondly, symptoms in TEST1 and TEST2 variants were predicted from previously predicted pollen concentrations and compared to symptom predictions originating from measured data.

III. Results and Discussion

First learning aspect included learning target pollen group concentrations from weather predictors. Expectedly, TEST2 variants were more successful in catching the trend, but it is generally underachieving to have *RMSE* over 10% which was the case in most target groups. Figure 1. depicts example of *Poaceae* pollen concentration prediction in TEST1 and TEST 2 variant. The trend is relatively well traced but its *RMSE* is still unsettling, and requires further possibilities for improvement.

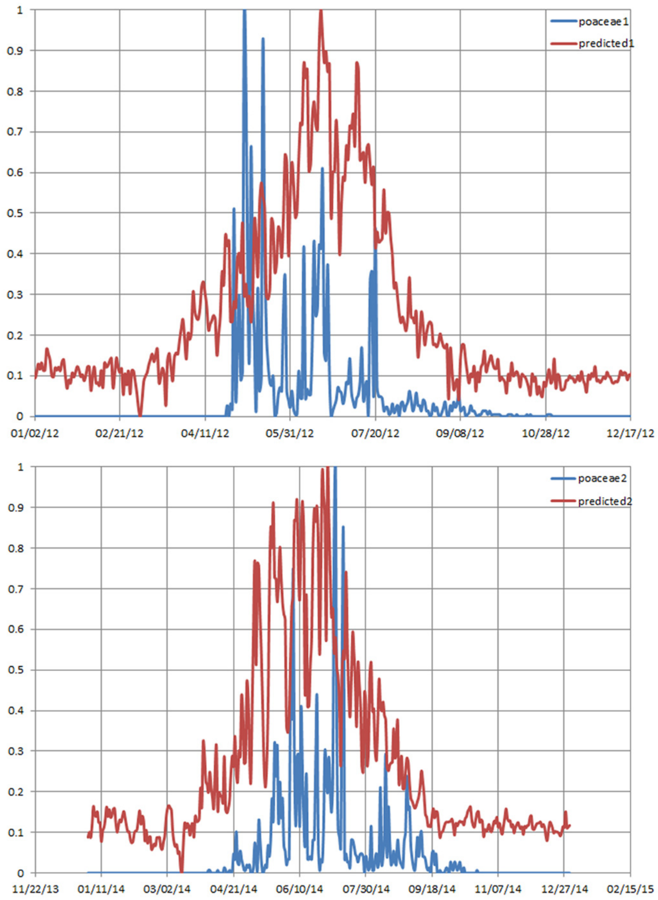


Fig. 1. Prediction of *Poaceae* pollen group from weather predictors in TEST1 (upper) and TEST2 (lower) variant.

In contrast to expectations, results of symptom prediction indicate that the accuracy did not increase by learning longer, as TEST 2 performed poorer than TEST1 (Table 1). The prediction of symptoms generally seems rather random and requires further improvements as the original hypothesis was not achieved, i.e. the symptoms could not been successfully predicted solely from pollen predictions. However, prediction of symptoms was satisfying for both TEST1 and TEST2 variants that were trained on entire dataset (including measured, not predicted pollen data).

TABLE I.

Variant	Symptom prediction *RMSE* in %	
	without measured pollen	***with measured pollen***
TEST1	10.43	6.00
TEST2	39.00	37.03

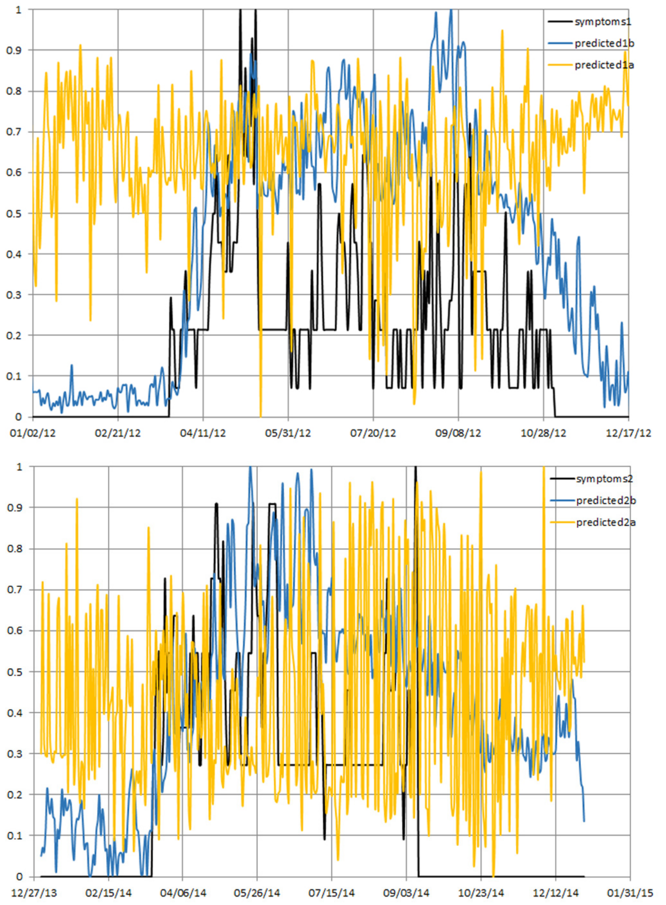


Fig. 2. Symptom prediction in TEST1 and TEST2 variant (upper and lower, respectively). Actual symptoms are depicted as black line, symptoms predicted by using measured pollen are given as blue line, while symptoms predicted by using predicted pollen are given as yellow.

IV. Conclusion

This paper summarizes the attempt of predicting pollen concentrations from weather data and further pollen allergy symptom prediction in the second round of regression based learning. The overall conclusion discourages the hypothesis of bridging symptoms directly to weather predictors but leaves space for improvements in respect to algorithm optimization and data preparation. It has been shown that adding autocorrelation indices of weather predictors as a synthetic supplement to a dataset seems to be a significant improvement, and will be further investigated.

Further research has to be directed towards spatial context of pollen concentrations and predictors distribution over a wider area or on a national level for Serbia. In addition, expected increase of accuracy by learning longer should be further inspected by observing 2015 and 2016 data series.

Acknowledgement

This research is supported by project of Ministry of education, science and technological development of Republic of Serbia TR36009. The authors thank to Environmental Protection Agency of Republic of Serbia for subletting the pollen concentration data of several measurement stations.

References

[1] C. Höflich, G. Balakirski, Z. Hajdu, J. M. Baron, L. Kaiser, K. Czaja, H. F. Merk, S. Gerdsen, U. Strassen, M. Bas, H. Bier, W. Dott, H. Mücke, W. Straff, A. Chakerc, S. Röseler, “Potential health risk of allergenic pollen with climate change associated spreading capacity: Ragweed and olive sensitization in two German federal stated,” Int. J. Hyg. Envir. Heal., vol. 219, pp. 252-260, January 2016.

[2] L.Makra, I. Matyasovszky, G.Tusnády, Y. Wang, Z. Csépe, Z. Bozóki, L.G. Nyúl, J. Erostyák, K. Bodnár, Z. Sümeghy, H. Vogel, A. Pauling, A. Páldy, D. Magyar, G. Mányok, K. Bergmann, M. Bonini, B. Šikoparija, P. Radišić, R. Gehrig, A. Kofol Seliger, B. Stjepanović, V. Rodinkova, A. Prikhodko, A. Maleeva, E. Severova, J. Ščevková, N. Ianovici, R. Peternel, M. Thibaudon, “Biogeographical estimates of allergenic pollen transport over regionalscales: Common ragweed and Szeged, Hungary as a test case,” Agr. Forest Meteorol., vol. 221, pp. 94–110, February 2016.

[3] K. Boričić, M. Vasić, J. Grozdanov, J. Gudelj Rakić, M. Živković Šulović, N. Jaćović Knežević, V. Jovanović, B. Kilibarda, T. Knežević, M. Krstić, D. Miljuš, N. Mickovski Katalina, D. Simić, “Results of the national health survey of the republic of Serbia, 2013, ” The Institute Of Public Health Of Serbia “Dr Milan Jovanović Batut”, ISBN 978-86-7358-062-3, Belgrade, Serbia, 2014.

[4] J. Grozdanov, D. Vuković, M. Krstić, B. Vančevska-Slijepčević, T. Milosavljević, “ National Health Survey Serbia: 2006: key findings, ” Ministry of Health of the Republic of Serbia, ISBN 978-86-83607-34-1, Belgrade, Serbia, 2006.

[5] B. Zvezdin, P. Radišić, M. Kojičić, S. Obradović-Anđelić, D. Jarić, A. Tepavac, L. Vrtunski-More, “Alergijske bolesti respiratornog trakta i polen ambrozije kao njihov uzročni faktor, ” Pneumon, vol. 41, 2004.

[6] T.B.Andersen, “A model to predict the beginning of the pollen season, ” Grana, vol. 30(1), pp. 269-275, 1991.

[7] G. D'Amato, L. Cecchi, S. Bonini, C. Nunes, I. Annesi-Maesano, H. Behrendt, G. Liccardi, T. Popov, P. van Cauwenberge, “Allergenic pollen and pollen allergy in Europe, ” Allergy, vol. 62(9), pp. 976-90 September 2007.

[8] K. M. Shea, R. T. Truckner, R. W. Weber, D. B. Peden, “Climate change and allergic disease, ” J Allergy Clin. Immunol., vol. 122(3), pp. 443-453, September 2008.

[9] I. Chuine and J. Belmonte, “ Improving prophylaxis for pollen allergies: Predicting the time course of the pollen load of the atmosphere of major allergenic plants in France and Spain, ” Grana, vol. 43(2), pp. 65-80, 2004.

[10] H. García-Mozo, J. A. Oteros and C. Galána, “Impact of land cover changes and climate on the main airborne pollen types in Southern Spain, ” Sci. Total. Environ., vol. 548-549, pp. 221-228, April 2016.

[11] I. Kasprzyk, B. Ortyl and A. Dulska-Jeż, “Relationships among weather parameters, airborne pollen and seedcrops of Fagus and Quercus in Poland, ” Agr. Forest Meteorol., vol. 197, pp. 111–122, October 2014.

[12] G. Astray, M. Fernández-González, F.J. Rodríguez-Rajo, D. López, J.C. Mejuto, “Airborne castanea pollen forecasting model for ecological and allergological implementation, ” Sci. Total Environ., vol. 548–549, pp. 110-121, April 2016.

[13] C. Costa, P. Menesatti, M.A. Brighetti, A. Travaglini, V. Rimatori, A. Di Rienzo Businco, S. Pelosi, A. Bianchi, P.M. Matricardi, S. Tripodi, “Pilot study on the short-term prediction of symptoms in children with hay fever monitored with e-Health technology, ” Eur. Ann. Allergy. Clin. Immunol., vol. 46(6), pp. 216-225, November 2014.

[14] D. Voukantsis, U. Berger, F. Tzima, K. Karatzas, S. Jaeger, K. C. Bergmann, “Personalized symptoms forecasting for pollen-induced allergic rhinitis sufferers, ” Int. J. Biometeorol., vol. 59, pp. 889–897, July 2015.

[15] A.J Smola and B. Schölkopf, “A tutorial on support vector regression, ” Stat. Comput., vol. 14(3), pp. 199-222, August 2004.

Modeling the protection of environment regarding climate change in the design and construction of water supply

[full paper]

Ivan Milojković
Institute for Water Resources "Jaroslav Černi"
Belgrade, Serbia
ivan.milojkovic@jcerni.co.rs

Roland Kröpfl
TIROLER ROHRE GMBH
Innsbruck, Austria
roland.kroepfl@trm.at

Nikola Radojlović
GP "GraditeljNS"
Novi Sad, Serbia
n.radojlovic@yahoo.com

Dragan Stefanović
BINVEX-STANDARD d.o.o.
Belgrade Investment and Export Company
Belgrade, Serbia
dragan.stefanovic.binvex@gmail.com

***Abstract*—The good solution for design and building water supply is the construction that is applied of adequate pipe system. Important is to supply with clean water all consumers. The following procedure represents selection of an adequate pipeline which would be stable in terms of floods, landslides and earthquakes. Due to climate change, there was a new requirement that is placed in front of water supply systems. In terms of variable rainfall, possible flooding, the emergence of new landslides is required detailed sensitivity analysis of water supply to these influences. It should not be adverse impact on the environment. In the case of a pipeline it is essential to choose the appropriate type of pipeline that would be resistant to new influences. In the case of main water supply pipeline Ø 800 of Zučka kapija to the settlement Kaluđerica in this paper describes the procedure for selecting the types of pipes which were adopted in the design and with whom he performed this pipeline. Authors used a contemporary MCDS PROMETHEE method in order to more realistically comprehend the conditions of operation and maintenance of water supply systems. The result of selecting the type of water supply pipeline is ductile pipes Ø800.**

Keywords—environment, waterworks, climate change

I. Introduction

Objective function in the following procedure is selection of an adequate pipeline which would be stable in terms of floods, landslides and earthquakes. Due to climate change, there was a new requirement that is placed in front of water supply systems. In terms of variable rainfall, possible flooding, the emergence of new landslides is required detailed sensitivity analysis of water supply to these influences. It should not be adverse impact on the environment. In the case of a pipeline it is essential to choose the appropriate type of pipeline that would be resistant to new influences [3], [6], [7], [8], [10], [11], [12]. In the case of main water supply pipeline Ø 800 of Zučka kapija to the settlement Kaluđerica [5] in this paper describes the procedure for selecting the types of pipes which were adopted in the design and with whom he performed this pipeline. Authors used a contemporary MCDS PROMETHEE method in order to more realistically comprehend the conditions of operation and maintenance of water supply systems. The result of selecting the type of water supply pipeline is ductile pipes Ø800 [2]. The subject of this paper is a section of the projected shaft connecting the regional water supply system Makiš-Mladenovac (km 14 + 374 - code manhole V8), to connect with the previously completed section showed on figure 1.

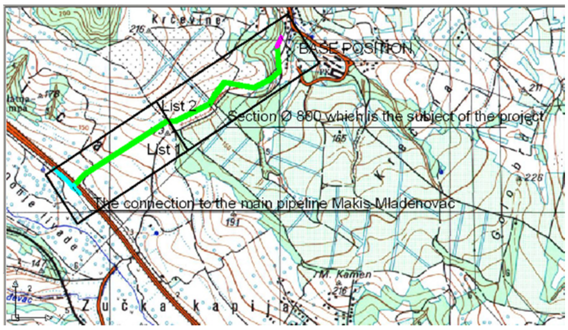

Krčevine
BASE POSITION
List 2
Section Ø 800 which is the subject of the project
List 1
The connection to the main pipeline Makiš-Mladenovac
Zučka kapija

Fig. 1. A section of the projected shaft connecting the regional water supply system

The layout of the pipeline route and the diameter are defined previously in the planning and technical documentation for the city of Belgrade in Serbia. The pipeline is designed and made from ductile pipes. On the route of the pipeline is envisaged to develop two manholes for discharges, one for air vent and one for the sectoral valve. Section length is 1593.84 m.

The route of the pipeline passes through several morphological units. In the first chainages is located in Put down relatively wide, the valley of the river Zavojnička. At the end of the valley the pipeline passing through several meters high embankment of the highway. Then, the route rises by undulating slopes, a distance of about half a mile, passing over a short plateau and down into the valley of a tributary valley Boleèica.

Hypsometrically terrains there are no major differences angles have values from 110 to 170 meters above sea level. Slope terrain, no steep slopes (up to 10 °) made it possible to develop deluvial, to a lesser extent and proluvial processes. Denudation by their accumulation eased the work of the slope. Corrugated surface due to local terrain depressions and projections, indicate that in the geological history happened before slipping lower. Under present conditions, the terrain is stable.

II. Methods

Visual PROMETHEE and GAIA is a multicriteria decision methods (MCDA). PROMETHEE stands for Preference Ranking Organization METHod for the Enrichment of Evaluations. GAIA stands for Graphical Analysis for Interactive Aid. MCDA stands for Multi Criteria Decision Aid. It includes many approaches, models and methods to handle decision or evaluation problems where multiple evaluation criteria have to be taken into account. MCDA methods are designed to assist decision-makers in such a context.

The PROMETHEE and GAIA methods can help engineers to solve many problems [1], [4], [9]. The PROMETHEE Rankings will show engineers the best compromise solutions according to the evaluation criteria, preferences and priorities. The PROMETHEE Sort procedure will help engineers, economists and other to allocate an item to a predefined class. The GAIA visual analysis will help analyst to understand better the decision problem, to see what is possible and what is not, to justify your choices or to acknowledge that some choices cannot be justified. It will also help analyst to explain to other persons why some decisions are better. PROMETHEE V will assist analyst in selecting different options according to constraints such as budget limit, incompatibilities, diversification, etc. The GDSS extensions of PROMETHEE and GAIA are available to help a group of persons to make a decision together. In multicriteria optimization desires are shown criteria and limits opportunities. The task multicriteria decisions are explicitly given until the restrictions implicit in the "admissibility" alternative, because all the alternatives from the set A must satisfy all constraints (g (x) > 0 in classical optimization).

III. EXPERIMENTAL

The projected water supply pipeline is planned and constructed by the ductile pipe with diameter of 800 mm. Operating pressure for the pipeline is designed to 16 bar. In this case, alternative solutions are different types of pipe material, when choosing tubes. With regard to the offer pipes, which is very topical in the market to build stocks observed proposed the following types of pipe material: ductile iron, polyester, high density polyethylene (HDPE), steel and reinforced concrete.

In this case, alternative solutions are different types of pipe material, when choosing tubes. With regard to the offer pipes, which is very topical in the market to build sections observed proposed the following types of pipe material:

a) High density polyethylene (HDPE)

b) Polyester

c) Ductile iron

d) Steel

e) Reinforced concrete

The aforementioned types of pipes with regard to the pipe material (alternatives) are evaluated on the following criteria:

1. Durability of pipes
2. Corrosion and abrasion protection
3. Length tube segments
4. Eligibility from the maintenance of pipelines
5. Eligibility from the standpoint of the design of pipelines
6. Resistance on rainfall
7. Resistance on flood
8. Resistance on landslides

IV. RESULTS AND DISCUSSION

Conditions for the design of external water supply network PUC "Belgrade Waterworks and Sewerage" causes the mode design of water supply network for a given location. The main project was prepared in accordance with the requirements for the design and applicable technical regulations for this type of work.

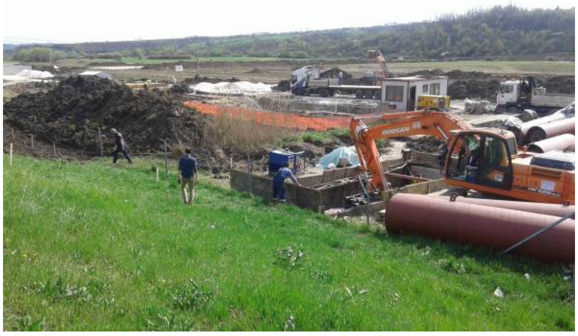


Fig. 2. View of construction site from the point of eligibility of maintenance and durability of pipelines and pipeline landfall resistance

In Figure 2 you can see the site at the beginning of the section that is being built and it was concluded that a good accessibility of the pipeline, which enables its good maintenance. Ductile pipes are highly resistant to mechanical and chemical influences. No special conditions for their storage and manipulation which greatly facilitates the construction. Before and during the construction of the building was carried out a detailed collection of data on the area where the works are performed, as well as on the position and function of the structure in relation to the entire water supply system and other infrastructure systems.

Figure 3 shows good overview of the operator on construction machine that performs very well carry out the excavation. Close to the highway construction site and pipeline places very high demands in terms of resistance to the impacts of the pipeline from landslides. They are very large forces that affect the pipeline and work construction of traffic which is transmitted to land close to the site and pipeline.

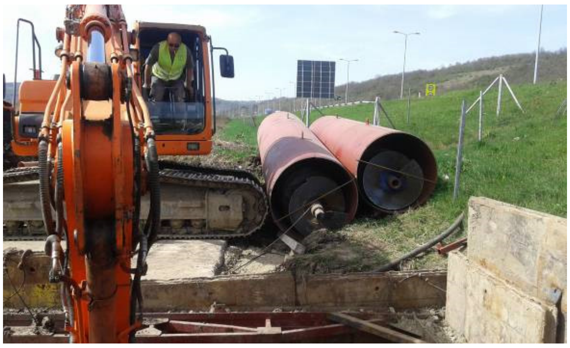


Fig. 3. View of construction site from the point of eligibility of maintenance and durability of pipelines and pipeline landfall resistance

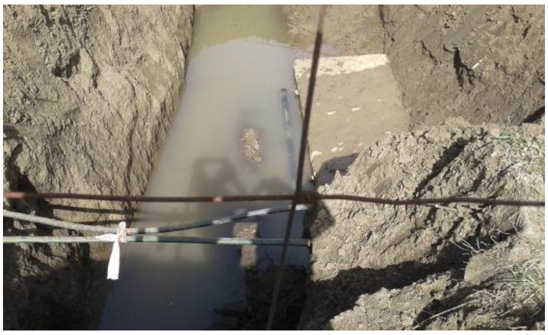


Fig. 4. View of construction site from the point of eligibility of maintenance and durability of pipelines and pipeline landfall resistance

Accuracy embodiment works is 1 cm, the construction works, or 1 mm for assembly work in setting up the pipeline. This is achieved by adequate geodetic and surveying information system that supports the carrying out.

Figure 4 also recorded high levels of underground water which poses major problems for the builders, owners and users of the water system. Durability of pipes comes into play since it is necessary to incorporate a permanent pipe segments to ensure quality functioning of the water system without jeopardizing the surrounding content. In Serbia, water systems and sometimes last more than a hundred years. The lifespan of water supply systems administered prior to all participants in the project as a priority in the application of quality materials and technologies.

Corrosion and abrasion protection are of great importance, especially in water supply systems. It is necessary at all costs to ensure healthy drinking water without any impurities that are harmful to human health. Often corrosion and abrasion products reach the water, and the source of the contamination is usually the water pipes. Therefore, it is necessary to pay special attention to the materials used during construction. In addition maturities of corrosion and abrasion in drinking water, it is very dangerous and the release of these products into the environment water supply system. If you come to the release of toxic substances into the environment, water pipes, are in danger and residents who live near these facilities, animal and plant life as well as employees working on maintenance. The high level of groundwater, especially for long periods, can have a very negative impact on the water supply system facilities.

Figure 5 shows comparative view of the resistance of sewer pipes built of different materials where tests show that the ductile pipes are most resistant to abrasion [2], which was taken into account in the multi-criteria decision making. The tests of resistance of different materials water supply pipes show that the ductile pipes are most resistant to abrasion, which was taken into account in the multi-criteria decision making. These tests were performed according to standard Darmstädter tip shanneling test ac. to DIN EN 295-3 and DIN 19565-1.

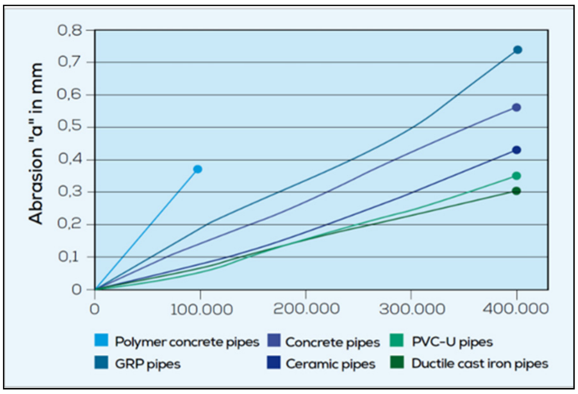


Fig. 5. Comparative display of resistance to abrasion of various tube

Length tube segments have a great importance from the standpoint of construction and maintenance of the facility. If the segments tube segments are longer, all the more easier and faster to build, as well as maintenance of the pipeline. Ductile pipes are manufactured in laying length of 6 meters, which in practice turned out well from the point of installation, replacement and manipulation.

Eligibility from the maintenance of pipelines is very important in water supply systems since they have a very strong great length of operation and maintenance. It is essential to use the most suitable pipeline systems from the point of installation, replacement of damaged pipes, and stability of the pipeline on various external, climatic conditions. Also, water is usually located right next to other infrastructure facilities and maintenance of water supply systems must be fully in accordance with the surrounding infrastructure.

Eligibility from the standpoint of the design of pipelines conditional use of the latest technologies related to all aspects of water supply, especially when the water supply network in question. Planning sewer is critical to anticipate all the conditions in which the object is to operate waterworks. How to build pipeline, so needed to maintain that there is such a pipeline system, which could be easily installed and dismantled. There are different installation technologies and the construction of the pipeline. There is also the possibility of welding the pipeline of different materials. Each method has its advantages and disadvantages, and it should always strive for the best solutions for given design conditions.

Figure 6 shows an advantage when mounting and dismounting of ductile pipe in the application of newly developed mounting compound. This is very important for mining, pipelines when moving to another location, in accordance with the growth of landfill.

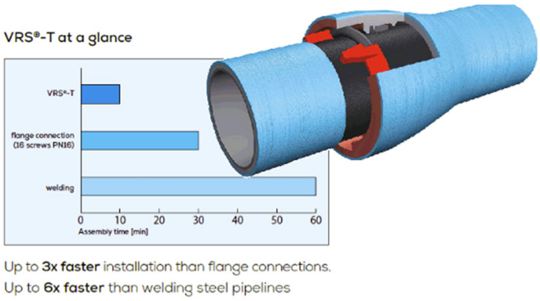


Fig. 6. Benefits during installation of ductile pipes

Resistance on rainfall in recent years has become a priority in Serbia, particularly bearing in mind the most important infrastructure systems. Aqueducts, sewers, roads and other infrastructure must be as resistant to the new climate conditions in Serbia default road emergence of large amounts of rainfall that along the river beds, so and locally. The high level of ground water in this case causes a very high sensitivity to water supply in precipitation since precipitation may occur sudden increases in groundwater levels and hence damage to the pipeline. This case causes the use of the best pipes for the construction of the aqueduct.

Resistance on flood is particularly important since the pipeline runs on relatively flat ground, near the highway and pass under the highway, so that in case of occurrence of local flooding in the area can easily come to the damage of water supply and the highway. Buildings, water supply and highway are of regional and republican character, so it must provide absolute protection against possible floods that may occur in large, unpredictable rainfall. Resistance on landslides and slope stability of land along the highway must be absolutely satisfied, given the importance of water supply and the highway. The technology used must be the best for this type of work. There no way not to compromise the functionality and stability of the considered objects. All terrain data, spatial model of terrain, groundwater, and soil bearing capacity must be taken into account in the design and construction of these facilities. Drilling in under a highway must be carried out in accordance with the weather conditions.

Performing works on the manhole belonging to this part of the Belgrade water supply system had to be adapted to the environmental conditions that are found (Figure 7). Manholes are made watertight, the required dimensions, along the highway.

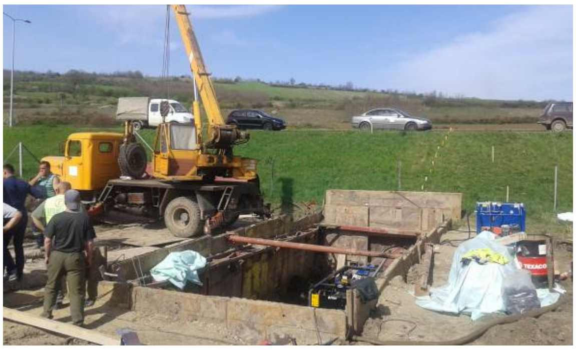


Fig. 7. Benefits during installation of ductile pipes

Engaged part of the field for the most part built Quaternary (the ground), a very small part, about 100 m, limestone (wall). Within the Quaternary alluvial deposits vary, eluviation-talus clay and loess clay. Since all three genetic types of Quaternary grade silty-clay component, they have similar physico-mechanical properties.

In the wider area of the pipeline route at different times and different occasion, registered more active and calmed the occurrence of landslides. The location of the pipeline is now in balance i.e. has a stable surface slopes. As the pipeline raises the most part perpendicular to the contour lines, its construction would not adversely affect, or will worsen, reduce the level of safety of slopes.

Defining geotechnical conditions was based on the total knowledge of this part of the terrain and a small number of field research works.

In terms of the recommended values of the physico-mechanical properties, data reliability is acceptable because in these areas performed relatively numerous laboratory testing field.

Perceiving the one hand, the type and sensitivity of the object, on the other hand engaged level of exploration of the terrain and the reliability of geological-geotechnical data can be considered to have met the conditions for making a satisfactory geotechnical basis for project design and construction of the intended object.

The pipeline of Ø800 Zučka kapija to the village Kaluđerica is connected to the regional water supply system Makiš-Mladenovac which was built in ductile pipe material. Advantages of pipes of ductile pipe material are as follows:

- They have excellent and stable hydraulic characteristics;
- They have excellent mechanical properties;
- Easy to install and do not require a mandatory set of sand in the trench as well as cots and more can be fine-grained materials;
- They are very stable and inert to climatic conditions;
- Suitable for installation in aggressive soils;
- Long-life is extremely reliable and does not require maintenance;

• It does not require cathodic protection, galvanized pipe from the outside before applying the final anti-corrosive coatings;

• In anchored type pipeline no anchor blocks, serving a much larger displacement and settlement of land than any other pipeline.

At the other end of the pipeline R800 yellowish gate to the village Kaluđerica connected to the pipeline already constructed of reinforced polyester. Built sections of reinforced polyester are placed next to the valley roads for Leštane and Kaluderica. Compressibility of the terrain and the case of instability is less pronounced for steel pipelines and ductile piping unbreakable BLS compound relative to polyester pipes. Relief (morphology) steeper terrain, specific for solid incompressible rock mass and stable terrain, have less negative impact on the steel and ductile piping with a TRS-T unbreakable connection, because the impact of flexibility and displacement capacity of the pipeline prominent in polyester pipeline. For these reasons, the polyester pipes need a larger number of anchor blocks.

Given that the projected route of the pipeline with a diameter of 800 mm, from Zučka Kapija to the village Kaluđerica, passes through hilly terrain, which is undulate, conditionally stable ground, an analysis was made for pipe material selection. Taking into account the conducted analysis, geotechnical conditions for construction, security of water supply system's asteroid belt and subsequent maintenance can be accepted pipelines from steel or ductile pipe material. Pipelines of ductile orderly materials provide the highest possible safety in exploitation. The pipeline of ductile pipe material withstands pressures up to 40 bars, has excellent mechanical properties with high resistance to static and dynamic influences. Construction of the due-meter steel pipeline is cheaper than building a pipeline ductile. However a number of advantages, especially if you take into account the durability of the pipe material, cheaper maintenance and redundancy of cathodic protection prefer pipeline of ductile pipe material. Pipelines of ductile pipe material in large-scale use in the reconstruction and construction of water supply network in Belgrade's water supply system anchored to the type of compound TRS-T.

Input data for model in process of evaluating which pipe is good for building pipeline for water supply pipeline Ø 800 of Zučka kapija to the settlement Kaluđerica showed in table I. On the PROMETHEE I Partial Ranking showed on figure 8, the leftmost bar shows the ranking of the actions according to Phi+: c) Ductile iron is on top, followed by a) High density polyethylene (HDPE) , b) Polyester , d) Steel and e) Reinforced concrete

TABLE I. PART OF THE INPUT DATA USED IN THE MODEL

actions	Values criteria functions							
	1.Durability of pipes	2.Corrosion and abrasion protection	3.Length tube segments	4.Eligibility from the maintenance of pipelines	5.Eligibility from the standpoint of the design of pipelines	6.resistance on rainfall	7.resistance on flood	8.resistance on landslides
	point	2 – in 1– middle 0 – out	m	point	point	point	point	point
a) High density polyethylene (HDPE)	9	0	6.0	9	10	8	9	9
b) Polyester	9	0	6.0	9	9	8	8	8
c) Ductile iron	10	2	6.0	10	10	10	10	10
d) Steel	6	2	6.0	5	6	10	9	9
e)Reinforced concrete	7	1	2.0	6	4	10	9	5
extremization	max	max	max	max	max	max	max	max

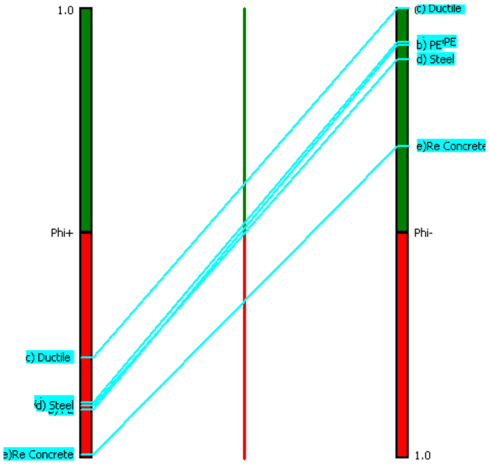


Fig. 8. PROMETHE ranking I

The rightmost bar shows the ranking according to Phi-: c) Ductile iron is still on top, and it is followed by: a) High density polyethylene (HDPE), b) Polyester, d) Steel and e) Reinforced concrete.

We can conclude that:

- c) Ductile iron is preferred more in compare to all the other actions in the PROMETHEE I ranking.
- c) Ductile iron, a) High density polyethylene (HDPE) and b) Polyester are on top.

- All actions are comparable because they have a similar score on Phi+ and on Phi-.
- c) Ductile iron, a) High density polyethylene (HDPE) and b) Polyester are close to each other.
- d) Steel is worse material by score according to Phi, Phi+ and Phi- from c) Ductile iron, a) High density polyethylene (HDPE) and b) Polyester which are close to each other.
- e) Reinforced concrete is worse from all other materials

This is confirmed by the PROMETHEE II complete ranking (figure 9). Three groups of actions appear clearly:

- c) Ductile iron has a higher Phi score, but near score a) High density polyethylene (HDPE) and b) Polyester.
- d) Steel has lower scores from c) Ductile iron, a) High density polyethylene (HDPE) and b) Polyester. They is more average action.
- e) Reinforced concrete has also very close but negative scores. They are at the bottom of the PROMETHEE II ranking.

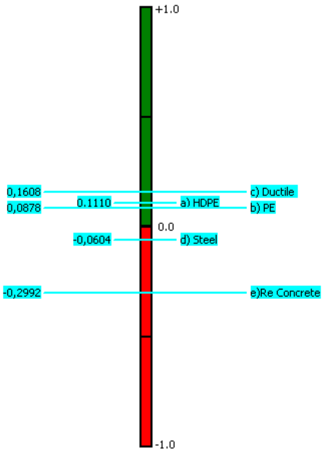


Fig. 9. PROMETHE ranking II

While the PROMETHEE II complete ranking is easier to explain it is also less informative as the differences.

V. CONCLUSION

The result of selecting process the type of water supply pipeline is ductile pipe for main water supply pipeline Ø 800 of Zučka kapija to the settlement Kaluđerica. Ductile iron, high density polyethylene (HDPE) and polyester are on top on the ranking. All pipe materials are comparable because they have a similar score on Phi+ and on Phi- in PROMETHEE method. Authors include in procedure for selection an adequate pipeline, criteria functions: floods, landslides and earthquakes and are get acceptably solution.

ACKNOWLEDGEMENTS

The authors are grateful to the Ministry of Education, Science and Technological Development of Republic of Serbia that the project TR 37014 allow the realization of this research.

REFERENCES

[1] Abu-Taleb, M.F., Mareschal, B. 1995 Water resources planning in the Middle East: Application of the PROMETHEE V multicriteria method. European Journal of Operational Research, 81, 500–511.

[2] Ductile iron pipesystems for drinking water. 2015 TIROLER ROHRE GMBH, Innsbrucker Strasse 51, 6060 Hall in Tirol, Austria.

[3] Jevtic, M., Milojkovic, I., Stojnic, N. 2011 Research of the performance of pulse electrohydrodynamics in blockage removal, Water Science & Technology, 64.1, 102-108.

[4] Kessili, A., Benmamar, S. 2016 Prioritizing sewer rehabilitation projects using AHP-PROMETHEE II ranking method. Water Science & Technology, 73(2), 283–291.

[5] Main project pipeline ø 800 from Zučka kapija to settlements Kaluđerica. 2014 Eko - water project Belgrade, Belgrade.

[6] Milojković, I., Despotović, J., Popin, K. 2015a Model for planning maintenance of sewerage networks based on external inspection. IWA Balkan Young Water Professionals 2015, 10-12 May 2015, Thessaloniki, Greece, Publisher: Hellenic Water Association, 76-80.

[7] Milojković, I., Despotović, J., Karanović, I. 2015b Model for Maintenance of Sewerage System based on Inspection. IWA 7th Eastern European Young Water Professionals Conference, 17-19 September 2015, Belgrade, Serbia, Publisher: IWA - International Water Association, 538-543.

[8] Milojković, I., Pusara, N., Bejatovic, S., Kröpf, R. 2015c Modeling of the pipe material selection for the protection of groundwater mining dumps, X International Symposium on Recycling Technologies and Sustainable Development, 4-7 November 2015, Bor, Serbia, Publisher: University of Belgrade – Technical faculty Bor, Editor: Prof. dr Zoran Markovic, ISBN 978-86-6305-037-2 86-92.

[9] Roozbahani, A., Zahraie, B., Tabesh, M. 2012 PROMETHEE with Precedence Order in the Criteria (PPOC) as a New Group Decision Making Aid: An Application in Urban Water Supply Management. Water Resources Management, 26(12), 3581–3599.

[10] Savić, A.D. 2009 The use of data-driven methodologies for prediction of water and wastewater asset failures, Centre for Water Systems, University of Exeter, North Park Road, Exeter, EX4 4QF, United Kingdom, Chapter published in the Springer book: Risk Management of Water Supply and Sanitation Systems, 181-190.

[11] Ward, B., Savić, A.D. 2012 A multi-objective optimization model for sewer rehabilitation considering critical risk of failure. Water science & Technology, 66.11: 2410-2417.

[12] Ward, B., Kawalec, M., Savić, D. 2014 An optimised total expenditure approach to sewerage management. Proceedings of the institution of Civil Engineers, 1-9.

Hydrological modeling - availability and reliability of the data in real time

[extended abstract]

Marija Ivković
Republic Hydrometeorological Service of Serbia (RHMSS)
marija.ivkovic@hidmet.gov.rs

Julijana Nađ
Republic Hydrometeorological Service of Serbia (RHMSS)
julijana.nadj@hidmet.gov.rs

***Abstract*—Within FP7 DRIHM project, a distributed HBV model, with time step of one hour, was developed according to the specific requirements of the project, parallel with a version to be used operationally for early flash flood warning. Once the necessary model structure and interface had been put in place, the model of the pilot river basin Kolubara was calibrated for issuing early warnings of the extreme hydrological phenomena. Main problems faced in this process were concerned with availability of sufficient real-time and historical precipitation and temperature data. Inadequate number of stations with precipitation data and their spatial distribution motivated a deeper analysis of other available sources of precipitation information. Possibility of using PAC products from the radar located at the Fruška Gora hill and their reliability under extreme weather conditions was examined. Calibration of the HBV model by using available data of the extraordinary hydrological flood event that occurred in May 2014 represented a great challenge. Huge volume of water spilled into the river plains causing water accumulation in the Stubo-Rovni and unexpected inundation of the Kolubara coal mines. In addition, flood wave had broken and flashed away two water level recorders (at the hydrological stations Valjevo and Draževac) and water level data become unreliable during the most important period. The model parameters were calibrated by comparing volume and maximum discharge of the flood hydrograph at the river reaches where the natural flow regime was not significantly disturbed.**

Keywords—wflow_hbv, hydrology, radar, flood

I. Introduction

Complex topography of the flash flood prone areas impose the need to have a fully distributed hydrological models with finer spatial and time resolution. The accuracy of the hydrological model largely depends on the quality of the input data. An analysis of availability and quality of the precipitation data in real time was conducted for an extreme hydrometeorological event in May 2014 when western Balkan was affected by cyclone Ivette. The observational network in Serbia is very sparse and one station covers area of 1,870 km^2 therefore it is necessary to analyze the availability and quality of data in real time, especially in areas where frequent flash flooding occurs. Spatially distributed hourly precipitation accumulations and air temperatures in real time are necessary both for the operational work and calibration and verification of the model. Input data are available from automatic weather stations (AWS) after interpolation and from PAC (Precipitation ACummulation) RAINBOW software product [4] installed on most radars in Serbia. Precipitation remotely observed with radar and precipitation measured on the automatic weather stations was compared with observations from the rain gauges and discussed also with their availability in real time.

II. Hydrological model

Model wflow_hbv is a spatially distributed hydrological model based on the HBV-96 model developed at the Swedish Meteorological and Hydrological Institute (SMHI) [1]. Model transforms rainfall to runoff using PCRaster [5] system for dynamic modeling of distributed systems. Unlike the original version, wflow_hbv use kinematic wave as method for river routing. Three major routines manipulate precipitation, soil moisture and control the volume of the flood wave. Input variables are hourly accumulated precipitation and mean hourly temperatures and potential evapotranspiration. Model is conceptual and 21 model parameters need to be calibrated against discharge [6].

The catchment area for Kolubara River is defined with hydrological profile Draževac on the 250x250m grid with hourly time step. Static maps are prepared using digital elevation maps from SRTM DEM [3] and the vegetation type for the analyzed basins was taken from Corine Land Cover 2000 [2].

III. Precipitation and temperature data in real time

Hourly precipitation and temperature data for Kolubara model are available on AWS Majinović, Štavica, GMS Valjevo and AWS Sopot, Stragari and Velika Ivanča installed on nearby basins. Automatic weather stations are not evenly distributed on the basin and it is important to check if precipitation obtained by interpolation can faithfully represent the distribution of the hourly precipitation amounts. Distance weighting method was the only interpolation method applicable in the situation when the number of AWS is small. Maps were compared with measurements on all available rain gauges (RG) located on the basin or around it. The comparative analysis of rainfall measured on the RG and precipitation registered on AWS showed underestimation of the daily sums on two stations while on the other two stations sums were overestimated (Fig. 3). This implies that either the measured amounts on the rain gauges were not correctly measured or the instruments on AWS were not properly calibrated.

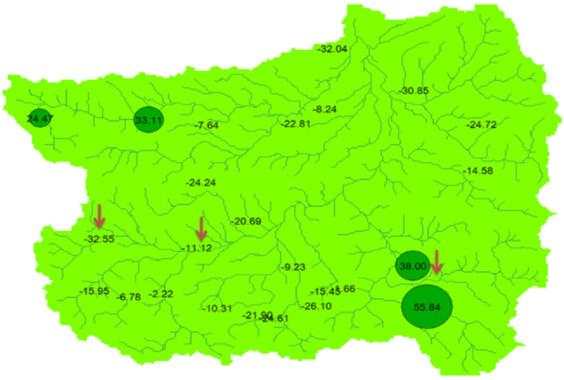


Fig. 3. Deviation of the interpolated precipitation for May 13th and 14th from rain gauge measurements

Another source of distributed rainfall are hourly radar PACs. During the whole year radar on Fruška Gora is operational and PACs are available and can be used in hydrological models. Based on reflectivity precipitation is estimated (QPE) using the formula:

$$Z=200R^{1.6}$$

where Z is radar reflectivity factor and R is the amount of rain (mm). This relationship is based on the Marshall-Palmer distribution of droplet size [7]. This relationship is recommended as basic formula for all regions and type of cloudiness. When two days accumulated PACs were compared with the data measured on RG average underestimation of 46% was noted (Fig. 4). Errors due to incorrect choice of coefficients in the Z-R relationship can be significant in assessing the quantity of rainfall, especially in extreme conditions, when there is the greatest risk of flash flooding. The values of coefficients range from $Z = 31R^{1.71}$ for orographic to $Z = 486R^{1.37}$ for heavy precipitation [8]. In the analyzed period the rain was moderate, occasionally with strong intensity. This fact provides the basis for the assertion that the coefficients of Z-R relationship should have value between $Z = 31R^{1.71}$ and $200R^{1.6}$ i.e. the lower value of reflexivity should give greater amounts of precipitation.

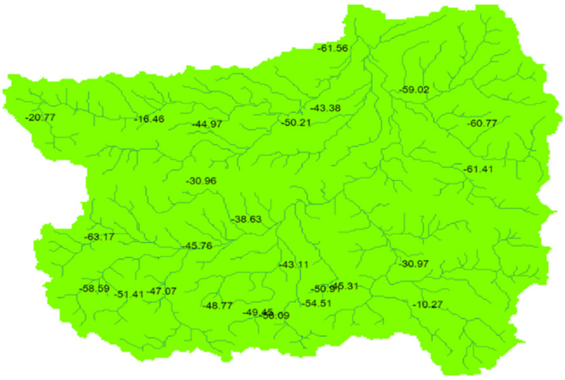


Fig. 4. Deviation of the interpolated precipitation for May 13th and 14th from PACs

By applying the methods of correlation on a sample of 190 members a new relation $Z = 62.0R^{1.6}$ was obtained with correlation coefficient r = 0.5865 (Fig. 5). A period of two days is certainly not sufficient to determine the Z-R relation and this procedure requires a long series of data classified according to the type of cloudiness and shape of relief.

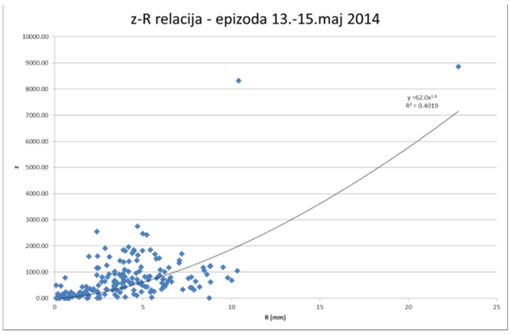


Fig. 5. Z-R relation May 13th to May 15th

Amount of precipitation obtained with the new equation now have a better agreement with the measured values on RG but significant overestimation are still present at the station Koceljeva, Donje Crniljevo, RC Valjevo, Lukavac, Liplje and Štavica. It is notable that the higher amounts of rainfall occur on the locations that are more than 300 meters above sea level. This indicates the potential existence of clutter but such differences could be also the result of the imprecise rain gauge measurements.

If we apply methods [10] to correct rainfall received from radar ($Z = 200R^{1.6}$) with precipitation registered on AWS we expect to have improved rainfall quantities over the basin. To be able to do that it was necessary to resample PACs from 600 x 600m grid to the grid of the hydrological model (Fig. 6).

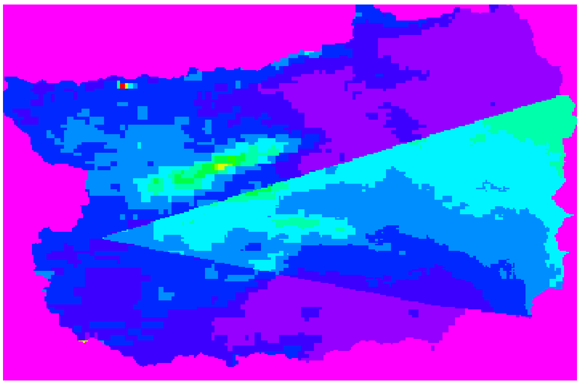


Fig. 6. QPE corrected with precipitation data from AWS

Due to small number of automatic weather stations correction can be made partly on the central and eastern part of the basin. The differences between data registered on RG and corrected ones are significantly reduced. In order to successfully apply correction method it is necessary to have a dense network of AWS.

IV. Hydrology Analysis

All sources of precipitation were used as inputs to wflow_hbv hydrological model [9]. Comparison between obtained volume and shape of the hydrograph was analyzed with the purpose to determine which source of rainfall is reliable in extreme situations. Hydrographs at selected profiles obtained using rainfall registered on automatic stations have quite good agreement with observed hydrograph (**Error! Reference source not found.**). As it was expected hydrographs obtained with precipitation estimated with radar have much smaller volume while corrected rainfall gave much better agreement with observed ones (Fig. **8**).

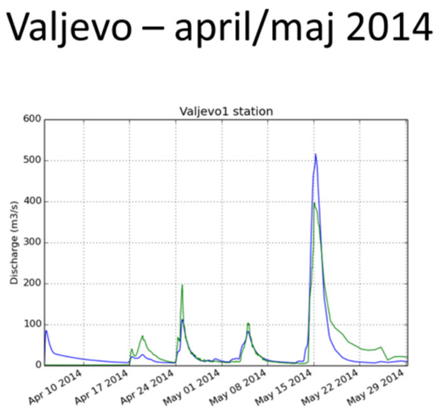


Fig. 7. Simulated (blue) and observed hydrograph (green)

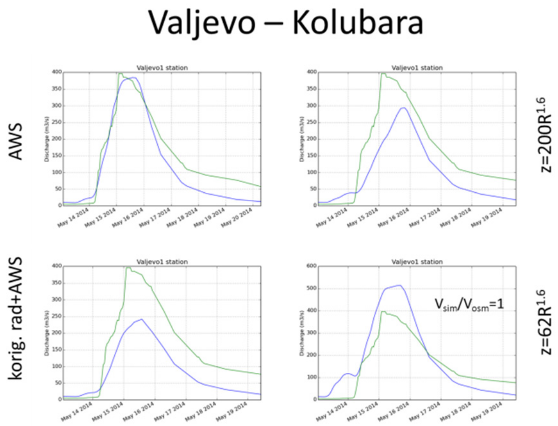


Fig. 8. Hydrograph profile Valjevo, period April-May

V. Conclusion

Hydrological wflow_hbv model requires accumulated precipitation and temperature grids over the basin and those data can be generated from measurements on automatic precipitation stations and as radar's estimate. These sources of rainfall information had not been used as input to any hydrological models in RHMSS, therefore their availability and reliability has not been analyzed.

Automatic weather stations and radar are good and reliable sources of information if they are regularly maintained and calibrated.

References

[1] Bergstrom S., Forsman A., “Development of a conceptual deterministic rainfall-runoff model”, Nordic Hydrology, 4, pp 147-170, 1973.

[2] EEA, CORINE Land Cover Technical Guide – Addendum 2000, Technical report No. 40, European Environment Agency, 2000.

[3] Farr T.G. et al., “The shuttle radar topography mission’, Reviews of Geophysics, 45, p. RG2004, 2007.

[4] GematronikGmb, Rainbow Product Manual,Version 3.4, Document Release 4.2 (2001-05-02)

[5] Karssenberg D., Schmitz O., Salamon P. De Jong, K.Bierkens, M.F.P., “A software framework for construction of process-based stochastic spatio-temporal models and data assimilation”, Environmental Modelling and Software25(2009): 489-502, 2009.

[6] Lawrence D., Haddeland I., Langsholt E., “Calibration of HBV hydrological models using PEST parameter estimation”, Report 01, Oslo, 2009.

[7] Marshall, J.S., R.C., Langille, and W. McK. Palmer, “Measurement of rainfall by radar”, J.Meteor.,4,186-192, 1947.

[8] Radinović Đ., Kostić A. “Radarsko merenje padavina u Srbiji”, studija, Republički hidrometeorološki zavod Srbije, 1997

[9] Schellekens J., Home OpenStreamsdeltares public wiki, http://publicwiki:deltares:nl/display/OpenS/Home, 2012.

[10] Wilson, J. W., “Integration of Radar and Raingage Data for Improved Rainfall Measurement. J. Appl. Meteor., 9, 489-497, 1970.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

551:311]:004(082)(0.034.2)
004.85(082)(0.034.2)

GEOMLA Conference (2016 ; Beograd)
Proceedings of GeoMLA, Geostatistics and Machine Learning, Application in Climate and Environmental Sciences, Belgrade, Serbia 21–24 June 2016. [Elektronski izvor] : [abstracts, extended abstracts and full papers of the GeoMLA 2016 Conference] / [edited by Milan Kilibarda, Jelena Luković]. – Belgrade : Faculty of Civil Engineering, University, 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 978-86-7518-190-3

a) Геостатистика – Примена рачунара – Зборници
b) Машинско учење – Зборници

COBISS.SR-ID 224905740

ISBN 978-86-7518-190-3

9 788675 181903