Interpolative mapping of mean precipitation in the Baltic countries by using landscape characteristics

Kalle Remm[a], Jaak Jaagus[a], Agrita Briede[b], Egidijus Rimkus[c] and Tiiu Kelviste[a]

[a] Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, 51014 Tartu, Estonia; kalle.remm@ut.ee
[b] Department of Geography, University of Latvia, Alberta St. 10, Riga LV-1010, Latvia
[c] Department of Hydrology and Climatology, Vilnius University, 21/27 Čiurlionio St., Vilnius 03101, Lithuania

Received 23 December 2010, accepted 14 March 2011

Abstract. Maps of the long-term mean precipitation involving local landscape variables were generated for the Baltic countries, and the effectiveness of seven modelling methods was compared. The precipitation data were recorded in 245 meteorological stations in 1966–2005, and 51 location-related explanatory variables were used. The similarity-based reasoning in the Constud software system outperformed other methods according to the validation fit, except for spring. The multivariate adaptive regression splines (MARS) was another effective method on average. The inclusion of landscape variables, compared to reverse distance-weighted interpolation, highlights the effect of uplands, larger water bodies and forested areas. The long-term mean amount of precipitation, calculated as the station average, probably underestimates the real value for Estonia and overestimates it for Lithuania due to the uneven distribution of observation stations.

Key words: precipitation, landscape variables, data mining, Baltic countries.

INTRODUCTION

Long-term meteorological data, precipitation included, are traditionally recorded from permanent observation stations. Such point data, in the sense of spatial analysis, are used to draw conclusions on climate change and to estimate the meteorological and climatological parameters over a given territory – country, catchment area, etc.

Point data measured at meteorological stations have limited representativity for wider regions. Precipitation is characterized by a very high spatial variability due to local convective showers, especially during the warm season. There is a need for spatial averaging of precipitation values for regions of different size (Omstedt et al. 1997; Rutgersson et al. 2001). Knowledge of the amount of precipitation is essential in hydrology for calculating water balance and river runoff (Bergström & Carlsson 1994).

The coverage of the estimated values can be obtained via simple interpolation or using an algorithm that involves characteristics of every location. The involvement of local landscape characteristics usually gives more reliable results than simple interpolation (Daly et al. 1994; Wei et al. 2005; Ninyerola et al. 2006; Sokol & Bližňák 2009; Moral 2010). An overview of predictive statistical models relating mean precipitation to altitude and its derivates has been given by Basist et al. (1994).

Hitherto, maps of the mean precipitation in the Baltic countries were mainly drawn according to visual interpolation and assessments of the distance from the sea coast, elevation and slope exposition. The most often used precipitation map for Latvia was drawn by A. Pastors in 1987. The map was later improved and published by Ziverts (2004). Unfortunately, the data used for this map and the interpolation method are not documented.

The latest published annual precipitation map for Lithuania is presented in Galvonaitė et al. (2007). The map is based on precipitation data of the years 1961–1990 from 75 stations. Only the impact of major relief forms on the amount of precipitation was considered.

In Estonia the first attempt to compose a mean pattern of annual precipitation by using landscape factors was made by Jaagus & Tarand (1988). Four factors for creating a model of the mean precipitation pattern were used – absolute height, windward and leeward parts of uplands, distance to the sea in the southwestern sector and the 3 km wide coastal zone. The model described two thirds of the total spatial variability of annual precipitation at the Estonian stations for 1966–1985. A new version of the maps for annual and monthly

(May, October) precipitation was published, using data from a longer period (1966–1998) (Jaagus 1999). In a previous study on precipitation in the Baltic countries (Jaagus et al. 2010), we interpolated the mean amount of precipitation using kriging interpolation between observation stations. Figures 2 and 3 of that paper depict the mean amount of precipitation in an extremely generalized way, smoothing non-typical results.

Precipitation is characterized by a very high variability in space and time. Its daily values are rather random. General and stable spatial patterns become evident after summing up the daily precipitation into monthly, seasonal and annual values. Variation in the quantity of precipitation is usually observed due to differences in the surrounding landscape. Therefore, the use of precipitation totals recorded over longer periods, such as a season or a year, is much more reasonable when studying relationships between landscape features and precipitation.

The indicator value of landscape characteristics at the stations and surroundings (local landscape variables) was compared in Jaagus et al. (2010), but not applied for map production. We suppose that at least a somewhat more detailed deduction of the spatial distribution of precipitation is possible by involving local landscape variables in interpolative mapping. This investigation is an extension of the above-mentioned study; therefore the observation data and the predictors are mainly the same.

Local landscape variables are involved as explanatory variables in statistical models and as the characteristics of every location where a model is applied for map generation. That is, first, statistical relationships between local landscape variables and precipitation have to be modelled. Then estimated values are calculated from the model at every location (grid cell) of the study area. Many methods have been developed to model statistical relationships. Only methods enabling the prediction of a continuous numerical variable using a large number of categorical and continuous explanatory variables, presuming the relationship is not multi-dimensionally linear or any other simple solution, are applicable in the present task of modelling mean precipitation. The task fits several data mining methods at the boundary of statistical modelling and machine learning. There are no generally accepted rules for how to choose the presumably most effective technique from the expanding diversity of data mining methods. Therefore, methodological comparisons are still urgent.

The aims of this study were: (1) to generate more detailed maps of mean precipitation in the Baltic countries than is possible from pure interpolation, (2) to compare the effectiveness of methods for interpolative modelling, (3) to specify the mean precipitation values over the territory of the Baltic countries.

DATA

Precipitation data

Seasonal (spring – MAM, summer – JJA, autumn – SON, winter – DJF) and annual precipitation data measured at meteorological stations in the years 1966–2005 were obtained from the national weather services of all three Baltic countries (Estonian Meteorological and Hydrological Institute; Latvian Environment, Geology and Meteorology Centre; Lithuanian Hydrometeorological Service). The precipitation data used in this study were collected with the Tretyakov gauge, following the general instructions given by the World Meteorological Organization (WMO). Winter precipitation in the Baltic countries can be liquid as well as solid. Snow and ice were melted before precipitation measurements.

The total number of stations included in this study was 245 (101 from Estonia, 62 from Latvia and 82 from Lithuania). The training data set of 123 stations used to calibrate models for interpolative mapping is the same as described in Jaagus et al. (2010). This was selected following the criterion of complete data coverage; that is the absence of gaps in time series. Single gaps existed only at some stations. The maximum amount of gaps allowed for the training data set was two years or 5% (Jaagus et al. 2010). Precipitation data from additional 122 stations not included in the training set were used as an independent data set to validate the results obtained with different interpolation methods (Fig. 1). The validation data set consists of information from the stations that did not fulfil the criteria for the training data set. The validation set was separated only to compare methods. The predictive maps were calculated, using observations from all 245 stations as training data for the models derived with the most effective methods.

The precipitation data included in the validation data set have observation gaps. The selected stations did not function during the entire study period (1966–2005). The criterion for using the data from each station was 15 years of observations as a minimum. The gaps in the observation series were filled with data from two control stations from the training set. The control stations were selected to be the nearest stations located in opposite directions from the station with observation gaps.

The mean monthly precipitation for the stations with gaps was calculated in the following steps. First, the mean precipitation values of the two control stations were found for every month during 1966–2005. Then,

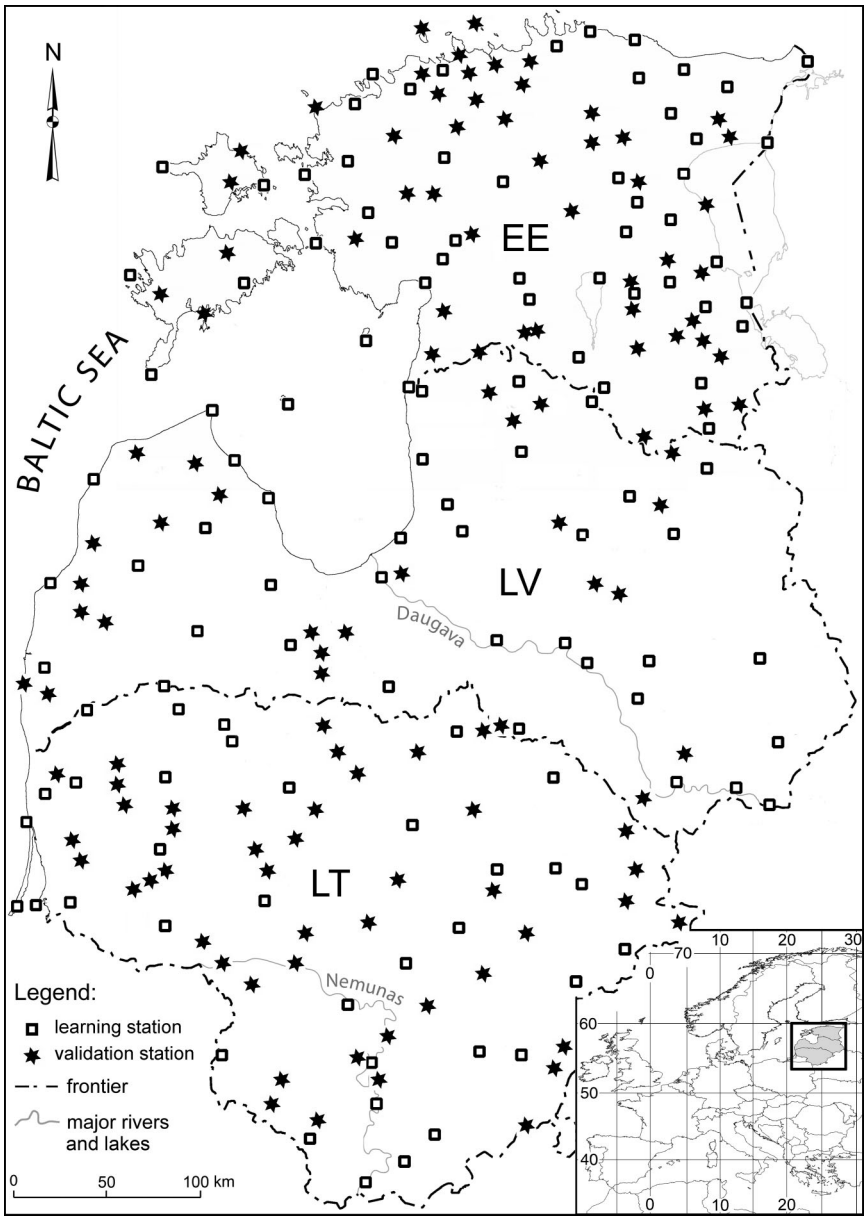

Fig. 1. Location of the study area, with learning and validation stations. EE, Estonia; LV, Latvia; LT, Lithuania.

the ratios of the monthly mean precipitation between the test station and the control stations were calculated for the period of simultaneous observations. Finally, the monthly mean ratios for that period were calculated and multiplied by the monthly mean precipitation values of the control stations for 1966–2005. We assumed that the ratio between monthly precipitation at a test station and at control stations observed during the period of simultaneous observations persisted also during the entire observation period of 1966–2005.

Landscape variables

The local landscape variables used as explanatory variables in the interpolative modelling were the same as in Jaagus et al. (2010): Cartesian coordinates of the Transverse Mercator projection (central meridian 24°E) in the west–east (longitude) and the south–north direction (latitude); 26 variables to characterize land cover diversity and the dominant land cover class; 10 variables to describe land surface elevation; 7 variables to describe

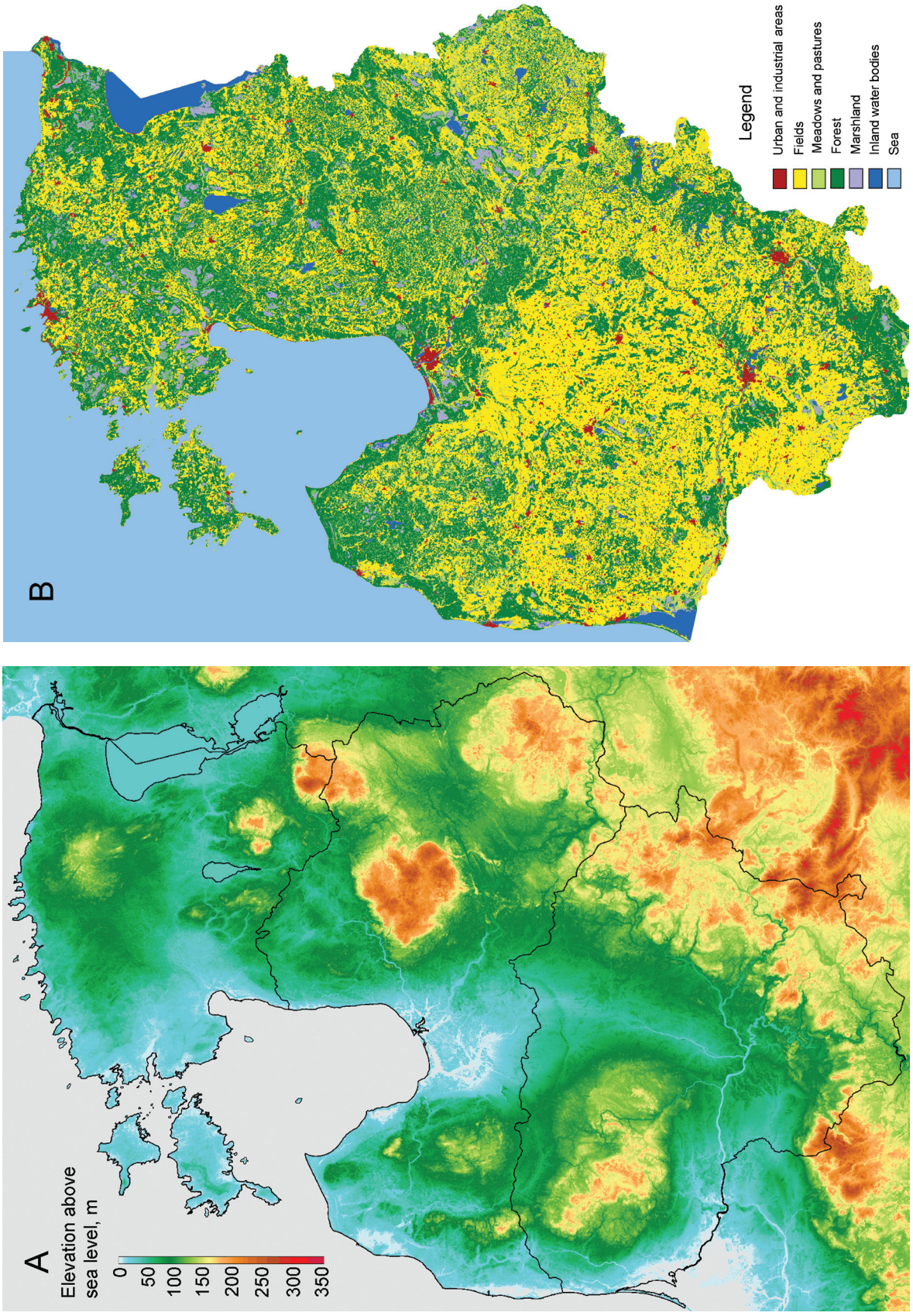

Fig. 2. Data layers for calculating local landscape features: **A**, surface elevation; **B**, land cover.

the distance from the sea coast and the share of water bodies and 6 describe the share of the forested area.

Two data layers were used to derive the landscape features (excluding the coordinates): the Coordination of Information on the Environment (Corine) land cover 2000 database (European Commission) and the global Shuttle Radar Topography Mission (SRTM) surface elevation model (U.S. National Aeronautics and Space Administration) (Fig. 2). Both data layers follow a constant methodology, cover the entire study area and have a suitable level of precision for predictive mapping at a regional scale.

The spatial indices calculated from the data layers included the number of categories, the index of dominance, the share of a particular category, the modal category and the reverse distance weighted modal category in the case of land cover. The mean value, aspect and quotient of variation (QV) were derived from the elevation model. The slope angle was not applied because the SRTM elevation model is not precise in details, and because the slope of the land surface is predominantly close to zero in the Baltic countries, since the differences in elevation are modest (Fig. 2A).

Coordinates in SN and WE directions are needed to describe spatial trends in models. In similarity-based estimations, the spatially closer exemplar stations are more similar regarding predictors SN and WE, and therefore have more impact on the estimated value. In addition, the reverse distance weighted mean precipitation amount of the predictable period, calculated using data from the other stations within a radius of 75 km, was added to the explanatory variables. The radius of 75 km was preferred as the result of comparing the fit of models, which alternatively involved precipitation data from other stations within radii of 30, 50, 75, 100 and 200 km. The application of larger radii would dim local variability of precipitation. Smaller radii would cause missing value problems for some modelling methods, since not all stations in the training data set have a nearest neighbouring station within the given radius. There are six stations missing a neighbour within 50 km in the training data set, but all stations have a nearest neighbour within 75 km (Fig. 3).

The basic data layers, maps of estimated values, similarity maps and models are all available at http://www.geo.ut.ee/Natuurkaart/BalticPrecipitation. Tabular source data are available from the authors.

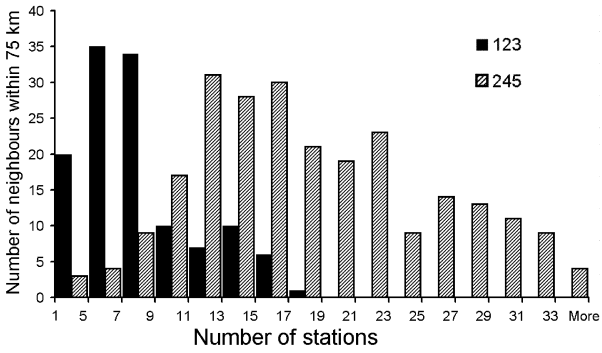

Fig. 3. Frequency of the number of neighbouring stations around observation stations in the learning set (123) and in all data (245).

METHODS

Data transformations and experiment structure

The main stages of the study were: (1) calculating local landscape variables and the reverse distance weighted mean precipitation at stations, (2) comparing the reliability of different models calibrated on the training set of stations, using the independent validation set of stations, (3) calibrating the most effective models according to stage two, using all observation data (training and validation data merged together), (4) generating predictive maps, using the most effective models, (5) deriving the estimated mean values for the countries from calculated values in the predictive maps.

The reverse distance weighted mean precipitation and the spatial indices used as local landscape variables were extracted from raster format data layers at the precise location of the station, within radii of 1, 10 and 20 km, and in the southwest sector of the circular kernel with the LSTATS software (Remm 2005; Tamm et al. 2010; http://www.geo.ut.ee/LSTATS).

A software system for the calculation of spatial indices, machine learning and similarity-based predictions – Constud (Remm & Remm 2008) and the following advanced statistical methods from the Statsoft Statistica 9 Data Miner (DM) package were compared: multivariate adaptive regression splines (MARS), boosted regression trees (BRT), random forest (RF) regression, support vector machine (SVM), *k*-nearest neighbours (KNN) and Statistica artificial neural networks (SANN). All these methods can combine multiple categorical and numeric variables in one analysis. Default values for the initial parameters (the choice of which depends on the method) were used for all methods (Appendix). The default options may not give the best models for all occasions, but are used as a standard to avoid a subjective bias in favour of one method or another when manipulating with the practically infinite number of possible combinations of parameter values. We assumed that software developers have selected default options close to optimal for most occasions.

Most data mining algorithms in Statistica DM are sensitive to the missing values of explanatory variables –

the predicted value is not returned if any of these variables are unknown. Surface elevation in the vicinity of seven stations in the training data is predominantly zero or predominantly flat. The QV calculated from surface values is undetermined on the first, and slope aspect on the second occasion. To include these stations in all comparable DM methods, gaps in the QV and aspect values were replaced with zeros.

Input from and output to binary raster format is integrated to Constud but not to Statistica DM. Therefore, a simple user interface was written in Microsoft Visual Studio 2008 to deploy the DM models to raster format data layers.

The fit of calculated and observed values (objective function) was measured as the relative root mean squared error (relative RMSE) for all methods. The training and validation fit were compared using over-fitting ratio, which was calculated as follows: the relative RMSE in the validation sample divided by the relative RMSE in the training sample. Over-fitting occurs when a model or a learning system describes noise instead of the underlying relationship. In iterative learning, over-fitting starts when further learning may improve the prediction fit according to training data, but not in an independent data set used for validation.

The spatial mean values of precipitation for the three Baltic countries were calculated as: (1) the average measured at observation stations, (2) the reverse distance weighted mean of precipitation at stations within the radius of 75 km calculated for a 1 km grid covering the study area, (3) similarity-based estimations calculated for the 1 km grid using Constud and (4) estimations calculated for the 1 km grid with the best DM method selected according to the validation fit.

Interpolations and predictive models for annual data and for the four seasons were calculated as separate models. Therefore, the sum of estimated precipitation for the seasons is not equal to the estimated annual value of precipitation.

Data mining methods

Data mining (DM) methods compared in this study are implementations of algorithms, the details of which are described in textbooks (e.g. Maimon & Rokach 2005; Witton & Frank 2005; Nisbet et al. 2009) and in publications cited at each method. Constud is a less known software system and therefore is described in more detail below.

Multivariate adaptive regression splines (MARS) is a nonparametric procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables, and can handle both categorical and continuous variables (whether response or predictors). This method partitions the input space into regions, each with its own regression or classification model, and can be considered as a generalization of regression trees where the abrupt binary splits are replaced by smooth basis functions. Overfitting in MARS is restricted by parameters setting limits to the complexity of the model. For more details see Hastie et al. (2009).

Boosting is an iterative procedure used to adaptively change the weights of training data so that the classifier will focus on cases that are hard to classify (Tan et al. 2006). The boosted trees algorithm in Statistica DM generates iterative classification or regression models and assigns weights to the observations according to the accuracy of the prediction. Then the classifier is applied again to the weighted data. As a result, each consecutive classifier is more effective in predicting values for observations that were not well predicted by the previous classifiers. Boosted trees have been successful for presence/absence data, as in the case of modelling species distribution by using climatological data (Elith et al. 2008) and in epidemiology (Remm & Remm 2010).

A random forest (RF) of decision trees consists of a collection of simple tree predictors, which are used to vote for the most popular class, or their responses are averaged to obtain an estimate of the numerically dependent variable. The response of each tree depends on a set of predictor values chosen independently for all trees in the forest as subsets (with replacement) of the predictor values of the original data. Random forest has been used in climatology for the prediction of transitions between weather regimes (Kondrashov et al. 2007).

Support vector machine (SVM) was initially developed as a classification method based on a set of points (support vectors) in the feature space that determines the boundary between different class membership areas. It can also be used as a nonparametric regression technique, where the complexity of the regression curve is controlled via the number of support vectors and not by the dimensionality of the feature space. This method has been used for climate change predictions (Tripathi et al. 2006) and for estimating surface temperature from remote sensing data (Moser & Serpico 2009).

The method *k*-nearest neighbours (KNN) is a similarity-based technique that in Statistica, unlike Constud, does not use iterative training. In KNN predictions are based on a set of prototype examples (exemplars, etalon observations) that are used to estimate values of the predictable variable based on the majority vote (for classification tasks) or averaging (for regression) over a set of *k* nearest prototypes. The KNN method was used for the interpolation of rainfall by Ali (1998).

An artificial neural network (ANN) consists of layers of interconnected nodes with a simple function connecting

inputs, weights and outputs. The weights are given initial settings and then iteratively adjusted according to errors of the predicted values. When the network is executed, the input variable values are placed in the input nodes, after that the hidden and output layer nodes are executed in their sequential order. Each of them calculates its activation value, using the weighted sum of the outputs of the nodes in the preceding layer. When the entire network has been executed, the nodes of the output layer act as the output for the entire network. The ANN method has been used in climatology for estimating evapo-transpiration (Zanetti et al. 2007), for the classification of rainfall variability (Michaelides et al. 2001), for downscaling daily precipitation extremes and variability (Dibike & Coulibaly 2006), for ground rainfall estimation from radar measurements (Liu et al. 2001) and for the reconstruction of precipitation time series (Lucio et al. 2007).

Constud

The software system Constud contains options for locally calculating indices of spatial patterns, iterative fitting (learning) of weights and for calculating similarity-based estimations output to a data table or to a raster map. The learning in Constud includes selection and iterative shifting of weights for features and observations according to the results of the similarity-based estimations in samples of observations and explanatory variables.

Constud estimates are similarity-based like the KNN estimates. Similarity between cases is calculated in Constud as the weighted mean similarity calculated from partial similarities. Every predictor has a weight, which is gradually fitted in iterations, and every predictor returns a partial similarity. In the case of a nominal predictor and matching classes, the partial similarity equals one. If the classes differ and self-similarities of classes are not applied (like in this study), the partial similarity equals zero. When self-similarity tables are used, partial similarity equals the similarity between the classes as set in self-similarity tables.

In the case of a numerical predictor (f), the difference (D) between its values (T_f and E_f) for an exemplar (E) and an observation (T) is calculated according to formula (1):

$$D = \frac{\left|T_f - E_f\right|}{2 w_E w_f}, \quad (1)$$

where w_E is the weight of exemplar E and w_f is the weight of the feature f. If the difference is greater than one, partial similarity is assigned a zero value. Otherwise, partial similarity is calculated by subtracting D from one.

In the case of spatial data, a distance-dependent correction parameter for reducing the effect of spatial autocorrelation is used. This parameter regulates the amount of possible reciprocal prediction between observations at a close distance. The similarity between observations is decreased in proportion to inverse distance between observations. The amount of decrease is regulated by the distance correction value. If the distance between observations remains below the correction value, similarity between them is set to zero. If the distance equals the double correction value, similarity will be decreased by 50%. As a result, the set of exemplars where the observations are spatially more dispersed is preferred (Linder et al. 2010).

Machine learning in Constud includes the automated selection and iterative weighting of features, observations and the sum of similarity sought for a decision. Learning iterations in Constud consist of five stages in a given order: (1) selection of features, (2) weighting of features, (3) selection of exemplars, (4) weighting of exemplars, (5) changing the actuality of features and cases.

The selection of features in Constud involves the inclusion of features in the order of usefulness one by one while gradually decreasing feature weights. The best feature set enters the new weighting iteration during which the feature weights are repeatedly changed one by one to improve the goodness of fit. Optionally, in addition to the change in feature weights, the effect of changing the sum of similarity is tested. The set of changed weights and the amount of similarity that gave the best matching predictions are used in the next iteration.

The best set of feature weights is used for the selection and weighting of cases. Before the selection of exemplars, goodness of fit between estimations and training data using all cases as exemplars is calculated. Thereafter the last observation is removed and the goodness of fit is recalculated. If the goodness of fit has decreased, the removed case will be used as an exemplar, otherwise the case will not be included in the set of exemplars.

In the feature weighting stage three alternative changes of exemplar weight are compared: (1) an increase, (2) weight left unchanged and (3) a decrease in weight. The initial magnitude of the weight change is 0.5; this is divided by 2 in the following iterations.

The actuality value is raised for a feature/exemplar that has been selected for the learning sample and has turned out to be useful (the prediction accuracy has increased). The actuality is decreased when the feature/exemplar was in the sample, but was not necessary for the prediction. Actuality values range between 1 and 200 and affect the sampling of features and exemplars. As the minimal possible actuality value is 1, no feature/case is totally discounted. After approximately ten learning iterations, actualities of features and exemplars are normalized so that the mean actuality of features and exemplars equals 100.

The speed of change of actualities is tuned by the factor of credulousness that has an integer value between 1 and 6. When the learning process is set to be more credulous, the alteration of actualities at every iteration is faster. When a case has been selected as an exemplar, its actuality is increased by the value of case weight multiplied by the factor of credulousness. If the case is in the training sample but appears to be useless as an exemplar, its actuality is decreased by the ratio of the sum of exemplar weights to the number of cases not being used as exemplars. Due to this, the total increase and decrease of actualities is balanced.

The fitting of features results in weights given to the explanatory variables. Similarity in one aspect does not mean similarity in other aspects; that is, not all recorded characteristics of a site have the same indicator value. As a rule, most features are not needed for the prediction of a dependent variable if the dataset contains a large number of interrelated explanatory variables – these features are given a zero weight.

The result of the fitting of cases is also weights. Only observations having a weight above zero are used as exemplars for similarity-based predictions. The number of exemplars used in estimations is fitted during learning in Constud as the sum of similarity sought for a decision. The initial amount of similarity for decisions was set to five, to start from more generalizing predictive sets. The learning process optimized this value for the annual and seasonal data sets to 1.2–4.1.

The final estimate is given in Constud as the average of the values of the dependent variable attributed to the most similar exemplars weighted by the similarity between a new location and exemplar locations. More technological details, the schemes of learning and map generation in Constud are given in earlier publications (Remm & Remm 2008, 2009; Remm et al. 2009) and on the Constud webpage: http://www.geo.ut.ee/CONSTUD.

In addition to the estimated values, the mean level of similarity between observations and exemplars used for the prediction is recorded by Constud. Lower values mark relatively peculiar cases or sites, which are insufficiently represented in the training data. The most similar exemplars used to calculate similarity-based predictions at these locations are actually not very similar to the site. Mean similarity does not represent similarity to the most similar exemplar, but the mean level of similarity between the exemplar sites used to calculate the prediction and that particular site since similarity-based predictions are usually calculated by using more than one exemplar.

RESULTS

Fit of models

Although SVM models gave the most exact estimations in the training data for annual and winter precipitation, Constud models were more reliable according to the average RMSE over annual and seasonal data (Table 1).

Table 1. Relative RMSE (%) of the predictions in the training data and in the independent test data set, and the overfitting ratio. The best validation results separately for the average of all models (annual, DJF, MAM, JJA, SON) and as the best of the DM methods are in bold. See pp. 173 and 175 for abbreviations

	Constud	MARS	BRT	RF	SVM	KNN	SANN
Training							
Annual	4.05	5.26	6.33	7.44	3.45	4.78	4.48
DJF	7.32	9.09	9.53	13.73	7.04	9.15	8.74
MAM	4.60	5.25	10.63	12.08	7.99	4.23	10.42
JJA	2.76	3.59	7.02	6.38	3.05	4.08	2.88
SON	4.07	9.39	7.92	17.66	17.53	18.52	4.38
Average	**4.56**	6.52	8.29	11.46	7.81	8.15	6.18
Validation							
Annual	8.34	**8.73**	9.60	11.02	10.56	12.12	20.52
DJF	11.01	**13.69**	14.73	15.18	14.12	17.55	30.49
MAM	10.72	10.89	17.14	**9.35**	9.54	11.19	35.11
JJA	7.58	7.81	12.05	8.12	**7.73**	8.47	9.45
SON	10.37	**13.00**	13.55	19.29	19.18	19.97	15.97
Average	**9.60**	10.83	13.42	12.59	12.23	13.86	22.31
Overfitting ratio							
Annual	2.06	1.66	1.52	1.48	3.07	2.54	4.58
DJF	1.50	1.51	1.55	1.11	2.00	1.92	3.49
MAM	2.33	2.07	1.61	0.77	1.19	2.64	3.37
JJA	2.74	2.18	1.72	1.27	2.54	2.08	3.28
SON	2.55	1.39	1.71	1.09	1.09	1.08	3.65
Average	2.24	1.76	1.62	1.15	1.98	2.05	3.67

Similarity-based reasoning in Constud outperformed other methods also according to the validation data, except for spring. MARS was the next most effective method according to the prediction fit in the validation data. As the average over all methods, the long-term mean precipitation in spring and summer was easier to model spatially than during autumn and winter when the RMSE was larger.

The RF method was remarkably resistant to overfitting; SANN was the extreme opposite, at least according to these data. Both similarity-based methods, KNN and Constud, have the next highest risk to yield deceptively effective overfitted estimations. On average, models for annual data gave more overfitted results than models estimating seasonal precipitation values.

Maps of estimated precipitation values

The maps of estimated mean precipitation in the Baltic countries were generated by using methods according to the validation fit, employing (1) reverse distance weighted interpolation, (2) Constud as the best of all compared methods on average and (3) the most reliable of DM methods. As a rule, the most effective DM method was MARS, except for in spring when the RF model outperformed all others, and in summer when SVM gave more reliable predictions. As expected, the modelled maps have more spatial details than the interpolation result (Figs 4, 5).

At the general level, all three methods (interpolation, Constud, MARS) result in similar maps for **annual** precipitation, which is highest on the western coast of Latvia and Lithuania (Fig. 4). This can be explained by the windward coastal effect, which is also persistent in western Estonia and in Latvia east of the Gulf of Riga. The area with the lowest precipitation is located east of these belts of higher precipitation on the leeward side from the sea. In the eastern parts of the Baltic countries, topography is the main factor determining mean precipitation. Higher precipitation is observed in the uplands and especially on their windward (western and southwestern) sides. Lower precipitation is typical for leeward sides of uplands and for lowlands.

The differences between the maps in Fig. 4A–C can be seen in the details – Constud and MARS highlight and delineate the effect of the Otepää, Haanja-Alūksne and Vidzeme uplands, while MARS probably overestimates the annual precipitation in Estonia and Latvia (Fig. 4B). The Constud map also indicates a higher level of precipitation in more forested regions than according to simple interpolation. The actual amount of precipitation in the central part of Hiiumaa and Saaremaa islands is not well known – the modelling methods indicate a potentially much greater amount of precipitation than measured at the stations, which are all quite close to the sea.

The mean precipitation over the larger lakes, Peipsi and Võrtsjärv, is unknown. The precipitation values for the lakes are inferred from data obtained at stations located near the coasts of these water bodies. The models take into account the percentage of water bodies in the neighbourhood. Therefore, the precipitation amounts estimated for lakes Peipsi and Võrtsjärv are similar to the estimates for the coastal zone of the sea and for small islands. The uniqueness of Lake Peipsi is also presented on the similarity map (Fig. 4D).

The actual amount of precipitation is also unclear on the southern bank of the Daugava River. There the result calculated with the use of local landscape variables is higher than the precipitation measured at nearby stations.

A significantly higher amount of precipitation close to the sea and a lower amount in the eastern regions is especially characteristic of the **winter** season. Modelled winter precipitation is spatially more detailed than the result of the interpolation. The greatest difference appears on the western and southern coastal regions of the Gulf of Riga, where the modelled values (Fig. 5B, C) are much higher than the interpolated ones (Fig. 5A). The lower actual value could be caused by ice cover, usually occurring in the Gulf of Riga, but absent in the Baltic Proper in winter. Models were not able to include the effect of ice because the distribution of ice cover was not among the explanatory variables.

The spatial variability of precipitation in **spring** is relatively low (Fig. 5D–F) and seems to coincide with air temperature. For example, lower temperatures and lower precipitation are observed in the coastal regions of Estonia. More precipitation is estimated for the more continental southern and eastern parts of the study area, which are the warmest in spring. Constud estimates that highlight high levels of precipitation near the eastern coast of the Gulf of Riga are deduced from the high levels of precipitation recorded at the stations Lagaste (208 mm) and Limbaži (218 mm) during MAM. The area of high precipitation close to the sea is extrapolated northwards up to Pärnu and to a more southern region on the Lithuanian coast, which is not supported by the observations (Fig. 5D, E).

According to validation data, the most reliable modelling method for MAM precipitation was RF, which gave the most conservative estimations – the estimated values are close to the mean value over the entire study area (Fig. 5F). Attempts to use any other method for modelling MAM precipitation were less successful according to the validation data. The only tendency highlighted by the estimates is the higher spring precipitation in eastern Latvia, which is deduced from relatively high

observed precipitation at the stations Dagda (182 mm) and Griškāni (174 mm).

Summer is the season of the highest precipitation in the Baltic countries except in the western coastal region of Lithuania and Latvia, where the maximum precipitation is observed in autumn. Less precipitation is depicted in the coastal zone and higher levels in the hinterland. Estimations for eastern Latvia are relatively low and spatially unstable (Fig. 5G–I). The low mean precipitation measured at Griškāni (154 mm) and Dagda (156 mm) was extrapolated to a wider region because of the low spatial density of observations (Fig. 5G). Modelling methods formally related low values in these two stations to the land cover category pastures, which is the dominant class in this region according to the Corine land cover map.

A somewhat higher mean precipitation indicated in the uplands in summer is visible both on the interpolated and Constud maps (Fig. 5G, H). The differences are mainly in estimates for Lake Peipsi and for the inner areas of Saaremaa and Hiiumaa islands – there should be much more rain in summer according to estimates involving landscape characteristics compared to simple interpolation. The actual mean precipitation level on larger water bodies cannot be directly verified from existing observation data. The low mean precipitation estimate for JJA off the coast of Lake Peipsi is concluded from the similar share of water bodies in the vicinity attributed to the stations on the seacoast.

The pattern of alternative estimations for **autumn** precipitation is quite similar. The area of higher precipitation is located in the western and northwestern parts of the Baltic countries, while the region of lower precipitation is situated in the southeastern and eastern parts (Fig. 5J–L). According to Constud estimates, the high precipitation area in western Lithuania continues along the seacoast in Latvia. The maximum is 10–20 km inland from the coast. Precipitation is somewhat higher in the eastern parts of Saaremaa and Hiiumaa islands than in the western parts.

Similarity maps and prediction residuals

Both prediction residuals and similarity maps calculated with Constud indicate reliability of estimations. Similarity maps show areas that do not have sufficient similar exemplar stations in the training data set. Low-similarity areas either lack stations with similar surrounding landscape or have atypical observation results. Additional stations are needed to represent these low-similarity regions. Absolute similarity cannot be seen anywhere since more than one exemplar was used in predictions.

The similarity map for annual data primarily indicates the uniqueness of large lakes Peipsi and Võrtsjärv (Fig. 4D). The average similarity between the most similar exemplar stations and the estimated location in the central part of Lake Peipsi is about 65–75%, while being mostly between 80% and 90% in other parts of the study area. Eastern Latvia and the diverse landscapes of the Baltic Uplands in eastern Lithuania should be more densely covered by meteorological stations.

The RF model generally overestimates the mean amount of precipitation in northern and western Estonia in spring. Columns of residuals for the Lagaste and Limbaži stations in northern Latvia for MAM stand in one north–south line, exaggerating each other (Fig. 6C, H). The values observed in both stations are notably higher than both RF and Constud estimations and also higher than the values observed in the neighbouring stations. The SVM model tends to underestimate the amount of precipitation in most stations in hilly southern Estonia. The annual, winter and autumn MARS models overestimate precipitation levels in the western and northern sides of the study area, except at stations on small islands and at the westernmost end of the Kõpu Peninsula.

Prediction residuals of Constud estimations have less spatial trends than estimations from the DM models (Fig. 6). Constud estimations have a marked tendency to smooth down extreme values in observations, since the estimated values are calculated as weighted averages from values of more than one exemplar. Although the estimated amount of precipitation in western Lithuania and western Latvia is the highest in the study area in autumn, winter and as the total annual, the observed values in this region are more often higher than lower, compared to the estimated values (Fig. 6A, B, E).

Predictive sets of variables

The predictive sets of explanatory variables selected by learning in Constud contain 2–8 landscape features, SN or WE coordinates (except MAM and JJA) and the reverse distance weighted amount of precipitation at stations within a radius of 75 km during the estimated period (except DJF) (Table 2). The indicator value of a single variable is not the same as in combination with other variables. In Jaagus et al. (2010) we compared variables one by one; here a predictive set of variables is given.

The interpolated amount of precipitation within a radius of 75 km during the estimated period was merely one among other variables used for similarity-based mapping. In most cases it is not even the weightiest one (except MAM, when Constud failed in comparison with RF, SVM and BRT). The best model according to the validation data for MAM was RF, which involved all explanatory variables.

Table 2. Explanatory variables selected by machine learning in Constud for the predictive set of features used for calculating maps. Variables are in descending order according to their weight (*w*)

Annual	*w*	DJF	*w*	MAM	*w*	JJA	*w*	SON	*w*
dw-mode_10sw	1.64	*WE*	1.61	*MAM 75*	1.95	*mode_20*	1.28	*dom_10*	1.68
annual 75	1.34	*SN*	1.50	*dw-mode_20sw*	1.84	*water_10*	1.14	*WE*	1.58
forest_10	1.13	*forest_10*	0.85	*water_20*	1.34	*JJA 75*	0.93	*SON 75*	1.40
elev_10	1.04	*d_coast*	0.03	*dw-mode_1sw*	1.14	*dw-mode_20sw*	0.66	*dw-mode20*	1.03
d_coast	1.00			*water_10*	1.13			*SN*	0.90
WE	0.96			*dw-mode_10sw*	0.40			*water_10*	0.90
mode_10sw	0.96			*mode_20*	0.11			*mode_20*	0.66
water_10	0.87			*forest_10sw*	0.09			*dom_20sw*	0.55
dom_20	0.74							*forest_10*	0.29
forest_10sw	0.30								

Abbreviations: *d_coast*, distance to sea coast; *dom_10*, index of dominance calculated from the coverage of Corine land cover units within 10 km; *dom_20*, index of dominance within 20 km; *dom_20sw*, index of dominance within 20 km in the SW direction; *forest_10*, share of forest within 10 km according to the Corine land cover map; *forest_10sw*, share of forest within 10 km in the SW direction; *water_10*, share of water bodies within 10 km; *water_20*, share of water bodies within 20 km; *elev_10*, mean elevation within 10 km; *mode_20*, Corine land cover mode within 20 km; *mode_10sw*, land cover mode within 10 km in the SW direction; *dw-mode_1sw*, reverse distance weighted land cover mode within 1 km in the SW direction; *dw-mode_10sw*, reverse distance weighted land cover mode within 10 km in the SW direction; *dw-mode_20sw*, reverse distance weighted land cover mode within 20 km in the SW direction; *WE*, west–east Cartesian coordinate; *SN*, south–north Cartesian coordinate; *annual 75*, reverse distance weighted annual precipitation in stations within 75 km; *MAM 75*, reverse distance weighted precipitation in stations within 75 km in March–April–May; *JJA 75*, reverse distance weighted precipitation in stations within 75 km in June–July–August; *SON 75*, reverse distance weighted precipitation in stations within 75 km in September–October–November; DJF, winter; MAM, spring; JJA, summer; SON, autumn.

Mean values of precipitation for the Baltic countries

The spatial mean values of precipitation for the three Baltic countries were calculated with different methods. The annual mean in Lithuania calculated as the average of observations (685 mm) is higher than the estimates from the models (667–681 mm) (Table 3). The difference is pronounced mainly in autumn and is likely caused by the low density of stations in the central part of Lithuania, which is an area with the lowest precipitation and where fields dominate on a flat terrain. The density of stations is higher in the coastal region of Lithuania where the amount of precipitation is the highest in the Baltic countries.

The situation is opposite in Estonia. Regions of lower precipitation are more densely covered by stations, while upland regions and forested areas with higher precipitation have fewer stations. All methods of spatial averaging indicate a higher mean precipitation for Estonia (687–712 mm) than the mean of the stations (681 mm).

The precipitation estimations from different methods are unstable especially in eastern Latvia and on the eastern coast of the Gulf of Riga (Fig. 5). In spite of a high spatial variability of estimates, the annual mean precipitation for Latvia according to the mean of the stations (694 mm), interpolation (690 mm) and Constud (697 mm) is similar. The MARS model yielded a higher estimate (721 mm).

DISCUSSION

According to Galvonaitė et al. (2007), the mean annual precipitation on the basis of the mean of stations in Lithuania is 675 mm, which is nearly identical to the total from the distance weighted interpolation from our data. The mean precipitation in Latvia has previously been estimated to be 703 mm (Ziverts 2004), which is more than the mean of the measurements and interpolations from station data available to us, and more than estimations obtained by using landscape variables and similarity-based reasoning in Constud. A previously published long-term mean annual precipitation value for Estonia is 669 mm (Jaagus 1999), which is notably less than indicated by the results of this study (Table 3).

Recent maps of the mean precipitation in the Baltic countries are basically similar to previous isoline maps (Jaagus & Tarand 1988; Jaagus 1999; Jaagus et al. 2010), but much more detailed. When comparing the

Table 3. Spatial mean precipitation (mm) in Estonia, Latvia and Lithuania according to observations, reverse distance weighted mean (RDW) interpolation, maps calculated in Constud and the DM method according to validation fit. See pp. 173 and 175 for abbreviations

	Estonia 101 stations	Latvia 62 stations	Lithuania 82 stations
Annual			
Mean of stations	681	694	685
RDW	687	690	676
Constud	691	697	667
MARS	712	721	681
DJF			
Mean of stations	137	141	139
RDW	138	140	136
Constud	139	143	137
MARS	142	143	129
MAM			
Mean of stations	117	130	131
RDW	120	133	132
Constud	119	130	128
RF	124	128	128
JJA			
Mean of stations	222	218	223
RDW	224	217	223
Constud	226	224	221
SVM	224	224	225
SON			
Mean of stations	206	205	191
RDW	207	198	185
Constud	206	196	184
MARS	218	209	189

maps of annual mean precipitation in Estonia, created using four landscape factors (Jaagus & Tarand 1988) and by interpolation of station data (Jaagus 1999) with the maps in Fig. 4, the highest similarity is revealed in the case of the Constud results (Fig. 4C). The largest differences between these maps are in the uplands of southern Estonia, in the hinterland of northeastern Estonia and in the forested belt of central Estonia, where the similarity-based estimation indicates higher precipitation.

The largest difference between these and previous maps for Lithuania are related to the Žemaičiai Upland. Traditionally, the dense isohyets are depicted around the highest part of this upland (Galvonaitė et al. 2007). Similarity-based calculations in Constud locate the highest annual precipitation closer to the coast, on the western slope of the upland (Fig. 4A). This result is deduced from the high observed annual values at the exemplar stations: Kartena (838 mm), Vėžaičiai (886 mm), Tubausiai (841 mm), Plateliai (867 mm), Aizpute (849 mm) and Cīrava (864 mm).

The low estimated precipitation 150–200 mm in eastern Latvia in summer (Fig. 5G) is deduced from the low precipitation recorded in the stations Griškāni (154 mm) and Dagda (156 mm). Temnikova (1958) reports a much higher precipitation level for eastern Latvia – about 260 mm; figure 3 B in Jaagus et al. (2010) indicates about 210 mm. According to the present data, these stations belong to a region of continental draught in summer, especially in July and August. A denser network of stations in this region could provide stronger evidence on the spatial and temporal pattern of precipitation, on the reliability of the measurements and on the relationships with topographical variables.

Similarity maps, first of all, highlight the absence of direct data on lakes Peipsi and Võrtsjärv and the scarcity of observations in eastern Latvia in spring and summer. The western slope of the Žemaičiai Upland in Lithuania and coastal areas would need a better coverage in summer. The distance to the nearest coast is up to 20 km in the centre of Lake Peipsi. A water body of this size presumably affects the movement of clouds and the amount of precipitation.

No single model is the best for all data. The results of this study are possibly affected by the modest size of the training sample. Similarity-based predictive sets in Constud tend to be overfitted if the number of training observations is less than 500 (Remm & Remm 2010), but this limit is not strict – overfitting depends also on the number, internal structure and intercorrelation of descriptive features, on the type (boolean, numeric, nominal), variability and predictability of the dependent variable. The extent of overfitting reduces the reliability of most of the compared methods, and therefore the fit and reliability of estimations derived with different methods could also differ if the number of observation stations or the study area were larger.

Another weakness of the methodological comparison in this study, and also in the wider context, lies in the large number of initial parameters and all possible combinations of the parameters that a user can adjust in all of these relatively complicated methods. Therefore, a single comparative study is clearly not enough to prove the superiority of one or another method. The comparison indicates that Constud has some advantages, which may be a result of the more automated process of parameter optimization for the predictive system compared to the DM models in Statsoft Statistica. The development and application of data mining methods in climatology is directed towards the automation of parameter optimization (Moser & Serpico 2009) and to the use of ensemble methods. Ensembles, which search for the best solution

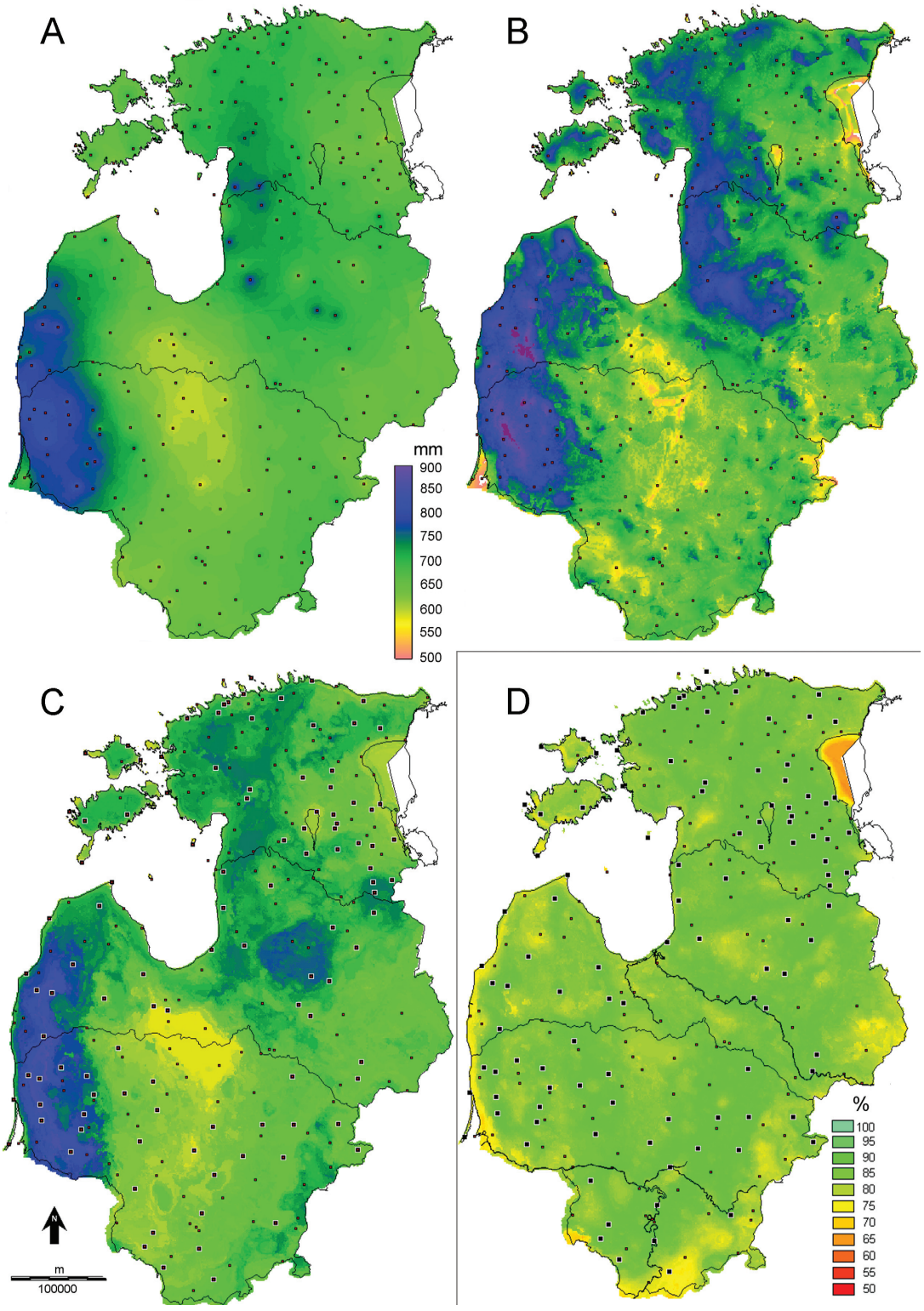

Fig. 4. Estimated annual precipitation in the Baltic countries in 1966–2005 and observation stations: **A**, reverse distance weighted interpolation; **B**, MARS model; **C**, Constud model; **D**, similarity to exemplar stations (larger squares).

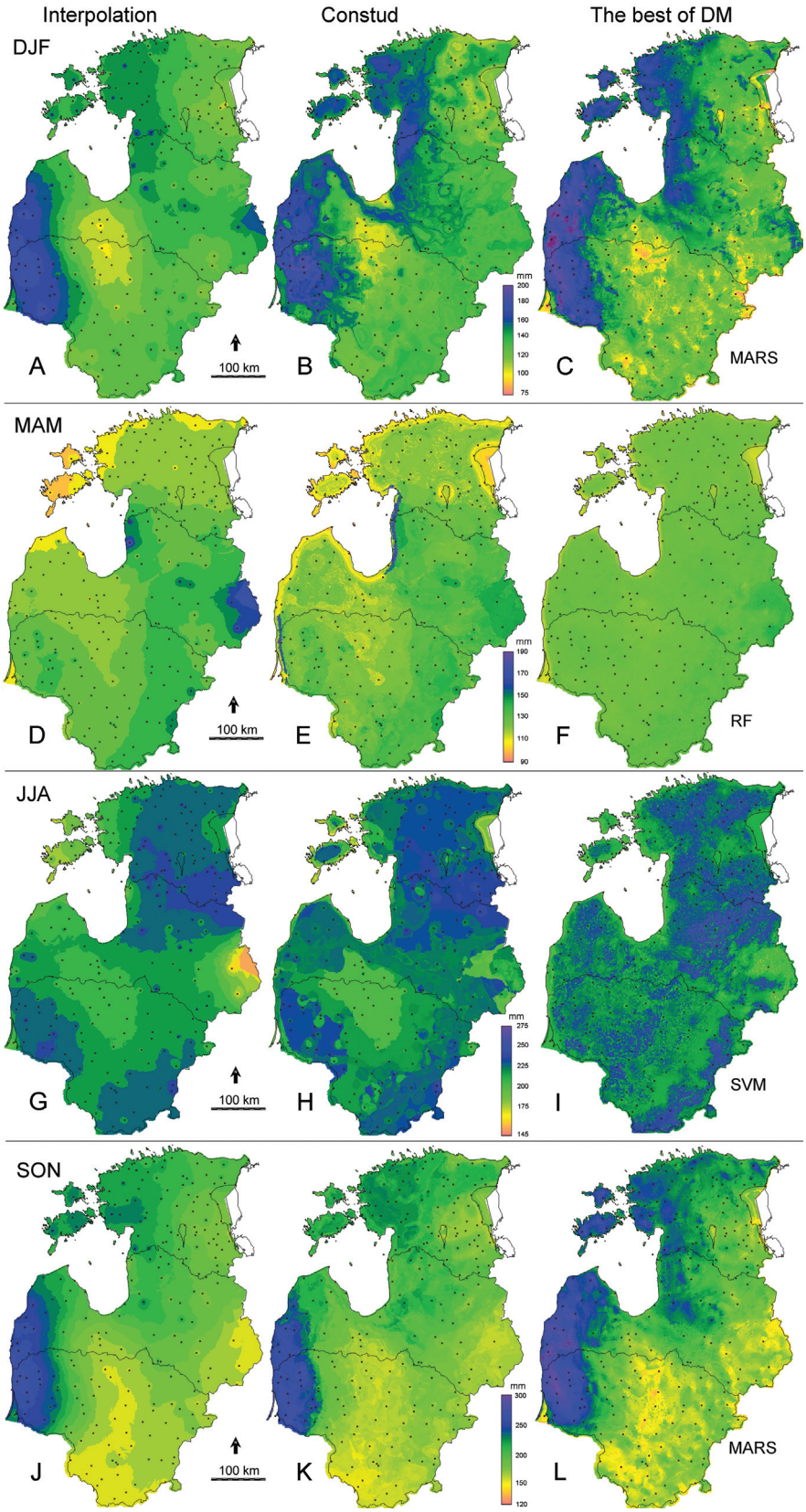

Fig. 5. Estimated seasonal precipitation in the Baltic countries in 1966–2005 and observation stations: **A**, DJF interpolated; **B**, DJF Constud model; **C**, DJF MARS model; **D**, MAM interpolated; **E**, MAM Constud model; **F**, MAM RF model; **G**, JJA interpolated; **H**, JJA Constud model; **I**, JJA SVM model; **J**, SON interpolated; **K**, SON Constud model; **L**, SON MARS model.

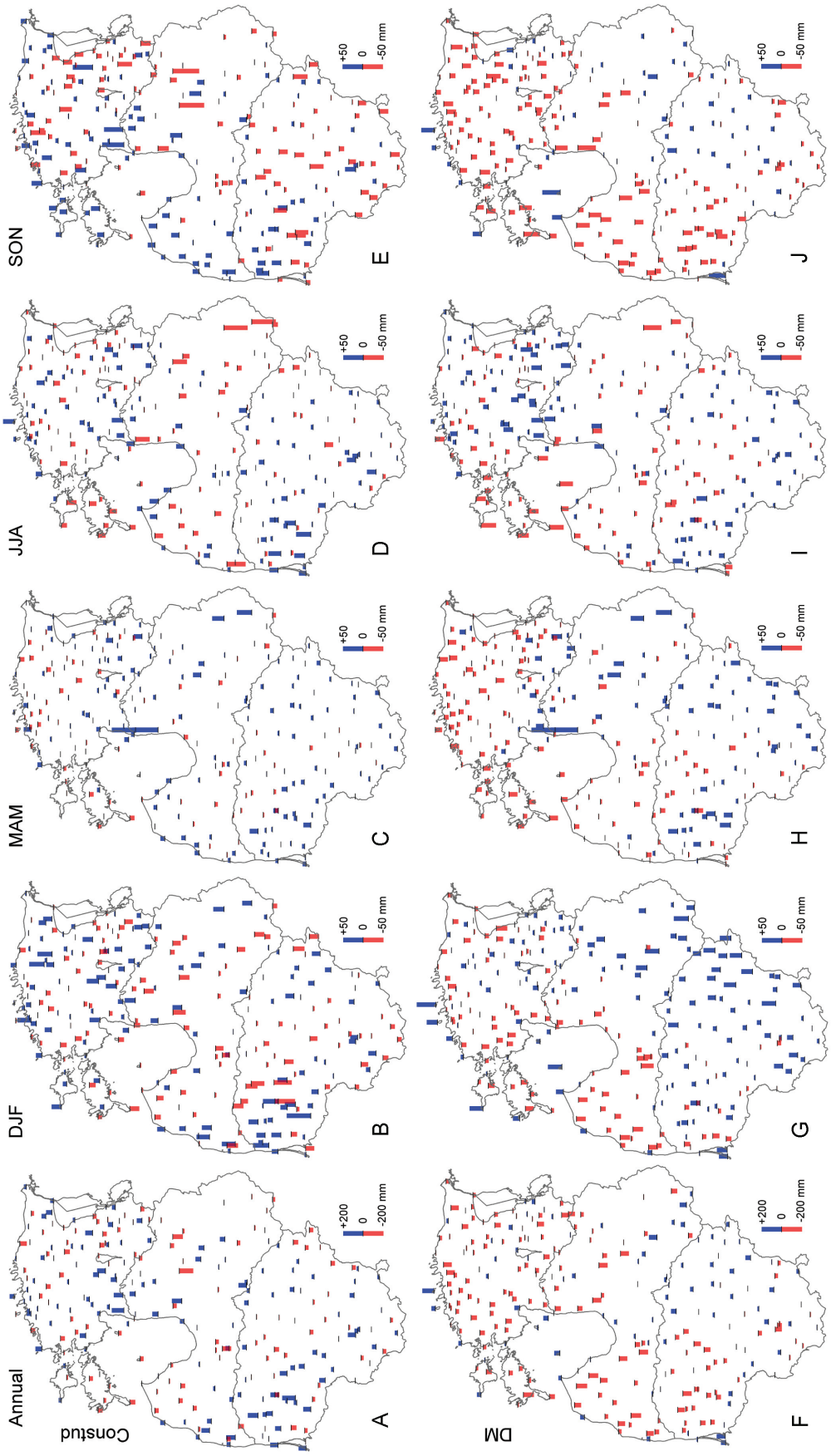

Fig. 6. Prediction residuals of the mean precipitation in the Baltic countries in 1966–2005 at observation stations: **A**, Constud model for annual data; **B**, Constud model for DJF; **C**, Constud MAM; **D**, Constud JJA; **E**, Constud SON; **F**, MARS annual; **G**, MARS DJF; **H**, RF MAM; **I**, SVM JJA; **J**, MARS SON. The height of the column indicates the difference between observations and model-estimation in millimetres: the column is blue and above the location of a station if the observed amount of precipitation is higher than the estimated value; the red column is below the location of a station and indicates that the observed value is lower than model estimation.

from many different models in parallel, have come into use for both short- and long-term weather forecasting. Conceptually, the multi-model approach should be more reliable; however, in reality, the superiority of multi-model predictions is not obvious (Algar et al. 2009; Weigel et al. 2009).

The relative RMSE of precipitation estimations obtained with the best method in this study in validation data was 7.6–11.0%. This is a better result than the simpler regression models applied in some earlier investigations (Daly et al. 1994; Goodale et al. 1998; Ninyerola et al. 2006).

Most modelling methods are sensitive to missing values of explanatory variables. The effect of missing values is not destructive in Constud – similarity is calculated using the existing values. The other reason why some models cannot yield a reasonable prediction is a new category in validation data – a class of a nominal explanatory variable that was not represented in the training data. Here again, Constud is more robust. The main drawbacks of Constud are the complicated and strictly fixed database structure needed within this software system and the time-expense of machine learning. The 2500 learning iterations using 123 observations and 20 explanatory variables (out of 52) took about 15 hours on a 3.4 GHz processor.

Learning in Constud involves random decisions: weights are shifted by a small random value at the beginning of each iteration. In addition, the decision is random in the situation when more than one learning path is equal. As a result, parallel machine-learning processes using the same data and initial parameters in Constud do not yield identical results. There is no guarantee that the predictive set of weights for features and exemplars selected yields the most reliable estimations. Different sets of features can provide approximately equally precise estimations as proven by Remm & Remm (2008), especially if the number of features is large and intercorrelated features are numerous. For example, there are seven features directly describing the amount of water bodies in the vicinity among the landscape variables used in this study. The share of water bodies is indirectly related also to the index of dominance, number of classes and the modal category of land cover.

Universally essential landscape factors in Constud models were features describing the amount of water bodies and the share of forests in the vicinity. The modest role of surface elevation among the landscape features, selected by iterative weighting in Constud compared with other studies (Goodale et al. 1998; Ninyerola et al. 2006), can be related to the relatively small altitude difference in the Baltic countries. The highest land surface point is 318 m above sea level and less than 2% of the area is higher than 200 m above sea level. Though, in Ireland, where the territory above 200 m comprises 13% of the total area, annual precipitation increased with elevation at a rate of 204 mm per 100 m (Goodale et al. 1998). In our models, features, such as the coordinates of a location and distance weighted precipitation values in the neighbourhood, partially replace characteristics of elevation and land cover. Following that, the actual role of elevation might be larger than indicated by the list of predictive features in Table 2.

CONCLUSIONS

The inclusion of local landscape variables enables more detailed and perhaps reliable interpolation than the reverse distance interpolation not including topographical variables.

Similarity-based estimations in the Constud software system were more reliable than data mining methods in the majority of cases, but not always. Automated parameter optimization could improve the output of data mining models. Over-fitting is the most serious threat in using artificial neural networks, but also in the case of similarity-based *k*-nearest neighbours and Constud methods.

The spatial representativeness of the existing network of observation stations could be better; the estimates for eastern Latvia are especially problematic because of the low number of stations. The spatial average precipitation calculated as the mean value of observations probably underestimates the real value for the Estonian territory and overestimates it in Lithuania. There are fewer stations within larger forested areas and uplands than the average density of the observation network in Estonia. The western coastal region of Lithuania has a denser network of stations than central Lithuania. Comparable observations on larger inland water bodies are also needed for reliable full-cover estimations.

Acknowledgements. The investigation was supported by the Estonian Ministry of Education and Research (Project No. 0180049s09). The authors express their gratitude to the Latvian Environment, Geology and Meteorology Centre, Lithuanian Hydrometeorological Service and Estonian Meteorological and Hydrological Institute for kindly providing the precipitation data, to Michael Haagensen for proofreading the text and to two anonymous reviewers for their instructive comments.

APPENDIX

INITIAL PARAMETERS OF THE MODELS, SOME EXPLANATIONS AND COMMENTS

Common parameters

Preclassifier – not used
Type of the predictable variable – numeric
Training sample size – 123
Validation sample size – 122
The number of stations used for calibrating models applied for generating maps – 245

Specific parameters

Constud

Initial value for the sum of similarity – 5 (a relatively large value)
Subsample size in learning – 124 (subsampling while learning is not applied if this parameter is larger than or equal to the training sample size)
Standard deviations of numerical predictors – precalculated from all training data (enhances the speed of learning, has no effect on results if subsampling is not applied while learning; if subsampling is used, the alternative is to calculate the SDs of numerical predictors for every subsample)
Turns of weighting cases – 1 (yields in weights for cases either 0, 0.5, 1.0 or 1.5)
Turns of weighting predictors – 20 (the weights for predictors and the sum of similarity sought for decision are changed in 20 iterations to find out the optimum set of weights)
Fitting the sum of similarity – true (the sum of similarity sought for decision is optimized together with the weights of predictors)
Objective function in learning – RMSE of leave-one-out cross-validation
Number of learning iterations – 2000
Distance-dependent correction – 0.1 m (excludes self-prediction while learning)
Factor of credulousness for cases and predictors – 3 (a medium value)

MARS

Maximum number of basis functions – 21 (determines the maximum complexity of the model)
Degree of interactions – 1 (only first-order interactions between variables are included)
Penalty for adding basis functions – 2 (larger values decrease the number of basis functions actually applied)
Threshold – 0.0005 (prevents overfitting)
Apply pruning – true (controls model complexity)
Memory limit – 30 MB (maximum data size that can be processed)

BRT

Learning rate – 0.1 (weight with which consecutive trees are added into the equation)
Number of additive trees – 200 (the number of trees to be computed in boosting steps)
Subsample proportion – 0.5 (proportion of random observations in learning samples)
Random test data proportion – 0.3 (proportion of random observations in test samples)
Minimum number to stop – 25 (the minimum number of cases in a terminal node that allows further splitting)
Minimum child node size to stop – 1 (the minimum number of cases in a terminal node to apply splitting)
Maximum number of levels – 10 (maximum number of splits)
Maximum number of nodes – 3 (each consecutive tree consists of one root node and two child nodes)
Seed for random number generation – 1 (is used to select the subsamples)
User defined final model – false (the final model is selected automatically)

RF

Number of predictors – 1 (the number of predictors in a simple regression tree)
Number of trees – 200 (the number of simple regression trees to be computed in successive forest building steps)
Subsample proportion – 0.5 (the subsample proportion to be used for drawing the bootstrap learning samples for consecutive steps)
Random test data proportion – 0.3 (the proportion of randomly chosen observations that will serve as a test sample)
Minimum number to stop – 25 (the minimum number of cases in a terminal node that allows further splitting)
Minimum child node size to stop – 1 (the minimum number of cases in a terminal node to apply splitting)
Maximum number of levels – 10 (maximum number of splits)
Maximum number of nodes – 100 (the splitting is stopped if the number of nodes exceeds this number)
Seed for random number generation – 1 (is used to select the subsamples for consecutive trees)
User defined final model – false (the final model is selected automatically)
Cycles to calculate mean error – 10 (specifies a number of cycles over which the error rates are monitored for improvement)
Decrease in training error – 30% (if the rate of improvement drops below this level, training is terminated)

SVM

Subsampling – random
Size of training sample – 75% (the proportion of cases used to form the training sample; the remaining cases are used as the test sample)
Seed for random sampling – 1000 (the random generator seed for sampling of data into train and test subsets)
SVM type – regression type 1 (type of the SVM model)
Capacity – 10
Epsilon – 0.1
Nu – 0.5
Kernel – RBF (Radial Basis Function kernel)
Degree – 3
Gamma – 0.2
Coefficient – 0
Maximum number of iterations – 1000 (the maximum number of iterations that can be applied in training the SVM model)
Stop at accuracy – 0.001 (training stops when the given level of accuracy is reached)
Cache size – 40 MB (limits memory usage)
Shrink data – true (shrinks data for computational efficiency)
Scale inputs – true (linearly scale the inputs within the range 0 to 1)

Scale outputs – true (linearly scale the outputs within the range – 1 to 1)
Apply *v*-fold cross-validation – false

KNN

Subsampling – random
Size of example set – 75% (the proportion of cases used as examples; the remaining valid cases form the test sample)
Seed – 1000 (the random generator seed for dividing data into the example and test sets)
Number of nearest neighbours – 1 (the number of exemplars involved into a prediction)
Distance measure – Euclidean (the metric used for measuring the distance between two points in the input space)
Standardize distances – true (values of predictors are standardized to make their ranges of values comparable)
Use weighted average for predictions – false (makes no difference if the number of nearest neighbours is 1)
Apply *v*-fold cross-validation – false
Restrict memory usage – false

SANN

Subsampling – random
Train sample size – 80% (the proportion of cases used to form the training sample)
Test sample size – 20% (the proportion of cases used to form the test sample)
Validation sample size – 0% (the proportion of cases used to form the validation sample)
Seed for sampling – 1000
Use Multilayer Perception (MLP) – true
MLP min. hidden units – 4 (the minimum complexity of the MLP network)
MLP max. hidden units – 13 (the maximum complexity of the MLP network)
Identity activation function: hidden neurons, identity – true
Identity activation function: hidden neurons, logistic – true
Identity activation function: hidden neurons, tanh – true
Identity activation function: hidden neurons, exp – true
Identity activation function: hidden neurons, sine – false
Identity activation function: output neurons, identity – true
Identity activation function: output neurons, logistic – true
Identity activation function: output neurons, tanh – true
Identity activation function: output neurons, identity – true
Identity activation function: output neurons, exp – true
Identity activation function: output neurons, sine – false

REFERENCES

Algar, A. C., Kharouba, H. M., Young, E. R. & Kerr, J. T. 2009. Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. *Ecography*, **32**, 22–33.

Ali, A. 1998. Nonparametric spatial rainfall characterization using adaptive kernel estimator. *Journal of Geographic Information and Decision Analysis*, **2**, 34–43.

Basist, A., Bell, G. D. & Meentemeyer, V. 1994. Statistical relationships between topography and precipitation patterns. *Journal of Climate*, **7**, 1305–1315.

Bergström, S. & Carlsson, B. 1994. River runoff to the Baltic Sea: 1950–1990. *Ambio*, **23**, 280–287.

Daly, C., Neilson, R. P. & Phillips, D. L. 1994. A statistical topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, **33**, 140–158.

Dibike, Y. B. & Coulibaly, P. 2006. Temporal neural networks for downscaling climate variability and extremes. *Neural Networks Archive*, **19**, 135–144.

Elith, J., Leathwick, J. R. & Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Galvonaitė, A., Misiūnienė, M., Valiukas, D. & Buitkuvienė, M. S. 2007. *Lietuvos klimatas* [*Lithuanian Climate*]. Lietuvos hidrometeorologijos tarnyba, Vilnius, 180 pp. [in Lithuanian].

Goodale, C. L., Aber, J. D. & Ollinger, S. V. 1998. Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model. *Climate Research*, **10**, 35–49.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition*. Springer, 745 pp.

Jaagus, J. 1999. New data about the climate of Estonia. *Publicationes Instituti Geographici Universitatis Tartuensis*, **85**, 28–38 [in Estonian, with English summary].

Jaagus, J. & Tarand, A. 1988. Spatial distribution of precipitation in Estonia. *Yearbook of the Estonian Geographical Society*, **24**, 5–16 [in Estonian, with English summary].

Jaagus, J., Briede, A., Rimkus, E. & Remm, K. 2010. Precipitation pattern in the Baltic countries under the influence of large-scale atmospheric circulation and local landscape factors. *International Journal of Climatology*, **29**, 705–720.

Kondrashov, D., Shen, J., Berk, R., D'Andrea, F. & Ghil, M. 2007. Predicting weather regime transitions in Northern Hemisphere datasets. *Climate Dynamics*, **29**, 535–551.

Linder, M., Jakobson, L. & Absalon, E. 2010. The effect of distance correction factor in case-based predictions of vegetation classes in Karula, Estonia. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science, Hong Kong 26.–28. May 2010* (Guilbert, E., Lees, B. & Leung, Yee, eds), pp. 570–574. International Society of Photogrammetry and Remote Sensing.

Liu, H., Chandrasekar, V. & Xu, G. 2001. An adaptive neural network scheme for radar rainfall estimation from WSR-88D observations. *Journal of Applied Meteorology and Climatology*, **40**, 2038–2050.

Lucio, P. S., Conde, F. C., Cavalcanti, I. F. A., Serrano, A. I., Ramos, A. M. & Cardoso, A. O. 2007. Spatiotemporal monthly rainfall reconstruction via artificial neural network – case study: south of Brazil. *Advances in Geosciences*, **10**, 67–76.

Maimon, O. & Rokach, L. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer, New York, 1383 pp.

Michaelides, S. C., Pattichis, C. S. & Kleovoulou, G. 2001. Classification of rainfall variability by using artificial neural networks. *International Journal of Climatology*, **21**, 1401–1414.

Moral, F. J. 2010. Comparison of different geostatistical approaches to map climate variables: application to precipitation. *International Journal of Climatology*, **30**, 620–631.

Moser, G. & Serpico, S. B. 2009. Automatic parameter optimization for support vector regression for land and sea surface temperature estimation from remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, **47**, 909–921.

Ninyerola, M., Pons, X. & Roure, J. M. 2006. Monthly precipitation mapping of the Iberian Peninsula using spatial interpolation tools implemented in a Geographic Information System. *Theoretical and Applied Climatology*, **89**, 195–209.

Nisbet, R., Elder, I. V. J. & Miner, G. 2009. *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier, Amsterdam, 824 pp.

Omstedt, A., Meuller, L. & Nyberg, L. 1997. Interannual, seasonal and regional variations of precipitation and evaporation over the Baltic Sea. *Ambio*, **26**, 484–492.

Remm, K. 2005. Correlations between forest stand diversity and landscape pattern in Otepää NP, Estonia. *Journal for Nature Conservation*, **13**, 137–145.

Remm, M. & Remm, K. 2008. Case-based estimation of the risk of enterobiasis. *Artificial Intelligence in Medicine*, **43**, 167–177.

Remm, K. & Remm, L. 2009. Similarity-based large-scale distribution mapping of orchids. *Biodiversity and Conservation*, **18**, 1629–1647.

Remm, K. & Remm, M. 2010. Geographical aspects of enterobiasis in Estonia. *Health & Place*, **16**, 291–300.

Remm, K., Linder, M. & Remm, L. 2009. Relative density of finds for assessing similarity-based maps of orchid occurrence. *Ecological Modelling*, **220**, 294–309.

Rutgersson, A., Bumke, K., Clemens, M., Foltescu, V., Lindau, R., Michelson, D. & Omstedt, A. 2001. Precipitation estimates over the Baltic Sea: present state of the art. *Nordic Hydrology*, **32**, 285–314.

Sokol, Z. & Bližňák, V. 2009. Areal distribution and precipitation–altitude relationship of heavy short-term precipitation in the Czech Republic in the warm part of the year. *Atmospheric Research*, **94**, 652–662.

Tamm, T., Remm, K. & Proosa, H. 2010. LSTATS software and its application. In *Proceedings of the Seventh IASTED International Conference: Signal Processing, Pattern Recognition and Applications; Innsbruck, Austria; 17.–19.02.2010* (Zagar, B., Kuijper, A. & Sahbi, H., eds), pp. 317–324. ACTA Press.

Tan, P.-N., Steinbach, M. & Kumar, V. 2006. *Introduction to Data Mining*. Pearson, Boston, 769 pp.

Temnikova, N. S. 1958. *Klimat Latvijskoj SSR* [*Climate of the Latvian SSR*]. Gidrometeoizdat, Riga, 232 pp. [in Russian].

Tripathi, S., Srinivas, V. V. & Nanjundiah, R. S. 2006. Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, **330**, 621–640.

Wei, H., Li, J.-L. & Liang, T.-G. 2005. Study on the estimation of precipitation resources for rainwater harvesting agriculture in semi-arid land of China. *Agric Water Manage*, **71**, 33–45.

Weigel, A. P., Liniger, M. A. & Appenzeller, C. 2009. Seasonal ensemble forecasts: are recalibrated single models better than multimodels? *Monthly Weather Review*, **137**, 1460–1479.

Witton, I. H. & Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Kaufmann, San Francisco; Elsevier, Oxford, 528 pp.

Zanetti, S. S., Sousa, E. F., Oliveira, V. P. S., Almeida, F. T. & Bernardo, S. 2007. Estimating evapotranspiration using artificial neural network and minimum climatological data. *Journal of Irrigation and Drainage Engineering*, **133**, 83–89.

Ziverts, A. 2004. *Hydrology (Introduction and Hydrological Calculations)*. Textbook for students of Agricultural University. LLU, Jelgava, 103 pp. [in Latvian].

Balti riikide pikaajalise keskmise sademete hulga kaardistamine maastiku tunnuste põhjal

Kalle Remm, Jaak Jaagus, Agrita Briede, Egidijus Rimkus ja Tiiu Kelviste

Ilmajaamades mõõdetud sademete hulgad on iseloomulikud vaid suhteliselt väikesele alale jaama ümbruses, seda eriti suvel, kui hoovihma osakaal on suur. Töö eesmärkideks olid: a) maastiku tunnuseid kasutades luua detailsem Eesti, Läti ja Leedu sademete kaart, kui seda on võimalik teha lihtsa interpoleerimisega, b) võrrelda erinevate interpoleerimismeetodite tõhusust sademete kaardi koostamisel ja c) määrata Balti riikide jaoks pindalaliselt keskmine sademete hulk.

Sademete ülepinnaliseks kaardistamiseks kasutati 245 meteoroloogiajaamas aastatel 1966–2005 mõõdetud keskmist sademete hulka ja igat kohta ning selle ümbruse maastikku iseloomustavat 51 kohatunnust. Seoseid kohatunnuste ja sademete hulga vahel modelleeriti 7 erineva andmekaevandamise meetodi (MARS, BRT, KNN, RF, SVM, ANN, Constud) abil. Võrreldes teiste meetoditega (välja arvatud kevadiste sademete hulga hinnangutes), andis sarnasusele tuginev hindamine Constud-i tarkvara abil enamasti usaldusväärsemaid tulemusi. Maastiku tunnustest olid sademete territoriaalse jaotuse kirjeldamisel olulised kõrgustikke, veekogusid ja ümbruse metsasust iseloomustavad tunnused.

Kohatunnuseid arvestav sademete ülepinnaline hinnanguline kaardistamine näitas, et vaatlusjaamade andmete lihtsal riikide kaupa keskmistamisel saadud hinnang tõenäoliselt ülehindab sademete keskmist hulka Leedus ja alahindab seda Eestis. Ülepinnalisel hinnangulisel kaardistamisel saadud tulemused viitavad vaatlusjaamade keskmisest suuremale tihedusele Leedu sademeterikkas lääneosas ja keskmisest väiksemale tihedusele Eesti kõrgustikel ning metsastes piirkondades. Baltimaade sademete keskmise hulga usaldusväärset kaardistamist piiras ka vaatlusjaamade vähesus Ida-Lätis ja nende puudumine suurematel siseveekogudel.