UNIVERSITY OF BELGRADE

FACULTY OF CIVIL ENGINEERING

Milan S. Kilibarda

AUTOMATED MAPPING OF CLIMATIC VARIABLES USING SPATIO-TEMPORAL GEOSTATISTICAL METHODS

DOCTORAL DISSERTATION

Belgrade, 2013

UNIVERZITET U BEOGRADU

GRAĐEVINSKI FAKULTET

Milan S. Kilibarda

AUTOMATSKO KARTIRANJE KLIMATSKIH VARIJABLI PRIMENOM PROSTORNO-VREMENSKIH GEOSTATISTIČKIH METODA

DOKTORSKA DISERTACIJA

Beograd, 2013

Mentor:

v. prof. dr Branislav Bajat, dipl. inž. geod., Građevinski fakultet, Beograd

Članovi komisije:

prof. dr Vladan Ducić, dipl. geogr., Geofraski fakultet, Beograd

v. prof. dr Ivan Nestorov, dipl. inž. geod., Građevinski fakultet, Beograd

v. prof. dr Ivana Tošić, dipl. met., Fizički fakultet, Beograd

dr Tomislav Hengl, dipl. inž. šum, Građevinski fakultet, Beograd (gostujući prof.) / ISRIC, Wageningen University and Research, the Netherlands

Datum odbrane:

Dedicated to my family.

Acknowledgements

First and foremost, I would like to thank my major professors, Branislav Bajat and Tomislav Hengl. They have kept me motivated throughout my PhD and generously shared their knowledge with me.

Alongside many ideas, theoretical and practical support, dr Tomislav Hengl has introduced me to many important figures in the fields of: spatial statistics, open-source GIS and R. He allowed me to spend two months at ISRIC - World Soil Information, as visiting researcher. There, I met professor Gerard Heuvelink who helped me a lot to understand many important spatio-temporal geostatistical issues.

Melita Perčec Tadić, Jelena Luković, Edzer Pebesma, Graeler Benedikt, Robert Hijmas and Ivana Tošić have given many comments, suggestions and support and they have influenced in the final shape of this thesis.

I would like to thank to developers of the R language for statistical computing. All computations were performed in the R environment, and partly in SAGA-GIS, PostGIS, rgdal open-source GIS applications. Especially, I would like to thank developers: Roger Bivand, Edzer Pebesma, Barry Rowlingson,Tim Keitt, Robert Hijmas, Graeler Benedikt, Alexander Brenning. They have developed many functionalities for spatial analyses in R.

Finally, without organisations providing high quality publicly availing data, this research could not be possible. Three data providers were crucially important for this thesis: National Climatic Data Center (NCDC), European Climate Assessment and Dataset Project (ECA&D), both providing ground station data sets and National Aeronautics and Space Administration (NASA), providing MODIS images.

UNIVERSITY OF BELGRADE

Abstract

Faculty of Civil Engineering
Department of Geodesy and Geoinformatics

Automated Mapping of Climatic Variables Using Spatio-Temporal Geostatistical Methods

Publicly available global meteorological data sets, from ground stations and remote sensing, are used for spatio-temporal interpolation of air temperature data for global land areas. Publicly available data sets were assessed for representation and usability for global spatio-temporal analysis. Three aspects of data quality were considered: (a) representation in the geographical and temporal domains, (b) representation in the feature space (based on the MaxEnt method), and (c) usability i.e. fitness of use for spatio-temporal interpolation (based on cross-validation of spatio-temporal regression-kriging models). The results show that clustering of meteorological stations in the combined data set (GSOD and ECA&D) is significant in both geographical and feature space. Despite the geographical and feature space clustering, preliminary tested global spatio-temporal model using station observations and remote sensing images, shows this method can be used for accurate mapping of daily temperature. Around 9000 stations from merged GSOD and ECA&D daily meteorological data sets were used to build spatio-temporal geostatistical models and predict daily air temperature at ground resolution of 1 km for the global land mass. Predictions were made for the mean, maximum and minimum temperature using spatio-temporal regression-kriging with a time series of MODIS 8 day images, topographic layers (DEM and TWI) and a geometrical temperature trend as covariates. The model and predictions were built for the year 2011 only, but the same methodology can be extended for the whole range of the MODIS LST images (2001–today). The results show that the average accuracy for predicting mean, maximum and minimum daily temperatures is RMSE $= \pm 2$°C for areas densely covered with stations, and between ± 2°C and ± 4°C for areas with lower station density. The lowest prediction accuracy was observed

in highlands ($>$ 1000 m) and in Antarctica with a RMSE around 6°C. Automated mapping framework is developed and implemented as R package meteo. Likewise, package plotGoogleMaps for automated visualisation on the Web, base on Google Maps API is developed.

Key words: spatio-temporal interpolation, spatio-temporal kriging, space-time variogram, linear regression, MaxEnt, daily air temperature, MODIS LST, global model

Scientific area: Geodesy

Scientific sub-area: Geodetic cartography

UDC number: 007:528.9(043.3) ; 551.581(043.3)

UNIVERZITET U BEOGRADU

Rezime

Građevinski fakultet

Odsek za geodeziju i geoinformatiku

Automatsko kartiranje klimatskih varijabli primenom prostorno-vremenskih geostatističih metoda

Javno dostupni meteorološki podaci, kako sa stanica tako i iz daljinske detekcije, korišćeni su za prostorno vremensku interpolaciju temperature vazduha iznad površine Zemlje. Zastupljenost i pogodnost javno dostupnih podataka je ocenjena, kroz tri aspekta kontrole kvaliteta: (a) zastupljenost u geografskom i prostornom domenu, (b) zastupljenost u karaktestičnom prostoru (*feature space*; bazirano na MaxEnt metodi), kao i (c) pogodnost korišćenja podataka za prostorno-vremensku predikciju (na osnovu kros-validacije prostorno-vremnskog regresionog kriginga). Rezultati pokazuju da je kombinovani set podataka (GSOD i ECA&D) značajno klasteriran i u geografskom i u karakterističnom prostoru. Uprkos klasteriranju, preliminarni rezultati globalne interpolacije primenom prostorno-vremenskog regresionog kriginga koristeći merenja sa stanica i snimke daljinske detekcije su pokazali da se tako mogu dobiti precizne globalne karte dnevne temperature. Oko 9000 stanica kombinovanog seta podataka (GSOD i ECA&D) je korišćeno za prostorno-vremensko geostatističo modeliranje i predikciju dnevnih temperatura u rezoluciji 1 km, iznad površine Zemlje. Za predikciju srednjih, minimalnih i maksimalnih temperatura korišćen je regresioni kriging uz pomoćne prediktore: MODIS LST 8-dnevni snimci, topografski lejeri (DEM i TWI) i geometrijski temperaturni trend. Model i predikcija se odnose na 2011 godinu, ali ista metodologija bi se mogla primeniti od 2001 godine do danas (od kada su dostupni MODIS snimci). Rezultati pokazuju da je prosečna tačnost predikcije za srednju, minimalnu i maksimalnu temperaturu vazduha oko ± 2°C za oblasti gusto pokrivene stanicama i između ± 2°C i ± 4°C za oblasti koje su slabo pokrivene stanicama. Najniža tačnost predikcije je dobijena u planinskim predelima i na Antartiku, oko

6°C. R softverski paket, meteo, je razvijen kao resenje za automatsko kartiranje. Razvijen je i paket plotGoogleMaps za automatsku vizuelizaciju na Web-u, koristeći Google Maps API.

Ključne reči: prostorno-vremenska interpolacija, prostorno-vremenski kriging, prostorno-vremenski variogram, linearna regresija, MaxEnt, dnevne temperature vazduha, MODIS LST, globalni model

Naučna oblast: Geodezija

Uža naučna oblast: Geodetska kartografija

UDK broj: 007:528.9(043.3) ; 551.581(043.3)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the main applications of using data from meteorological stations is to produce maps showing spatio-temporal patterns of climatic variables, also referred as mapping climatic variables. The term **mapping**, in this thesis, is considered as an interpolation on regular grids that are also called raster grids, climatic images or surfaces. Chapter 2 describes interpolation methods used in meteorology and climatology; they range from nearest neighbour methods, splines, regression and kriging, to neural networks and machine learning techniques. The most of used methods, and related works of global mapping at daily temporal resolution (see Chapter 2 for literature review on interpolation of climatic variables) uses only spatial interpolation. The reason for this purely spatial modelling of spatio-temporal phenomenons might be that the areas of spatial statistics (and spatial GIS modelling) have been much more developed in contrary to spatio-temporal statistics, able to model processes, essential dynamics. Similar, time series analysis have been developed and used mostly without considering spatial component of time series of observations.

Spatio-temporal geostatistics has made a breakthrough in the past decade with theoretical concepts (Cressie and Wikle, 2011) and various examples of applications have been provided (Gething et al., 2007; Heuvelink and Griffith, 2010; Heuvelink et al., 2012; Gräler et al., 2011; Hengl et al., 2012). An extension from purely spatial statistical models to spatio-temporal models is a logical evolution of the field, especially since we know that most meteorological parameters vary both in space and time and that 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. The insights obtained from the process are much richer and allow one to distinguish sources of variability and to isolate purely temporal, spatial and spatio-temporal components of variability. Moreover, we can predict values using spatio-temporal observations in individual domains such as the spatial domain (e.g. as daily map, pure spatial data),the temporal domain (predict missing values at a certain meteorological station in the form of a time series) and in the spatio-temporal domain as spatio-temporal full grid data as is demonstrated in the spatio-temporal class (data model) developed by Pebesma (2012).

A literature review shows that no group has previously attempted to interpolate daily values of meteorological variables using spatio-temporal regression-kriging with a time-series of remote sensing based covariates, especially at a fine resolution of 1 km. Hengl et al. (2012) describe a framework for space-time regression kriging interpolation of daily temperatures that makes use of a time-series of MODIS images, which are presented as a Croatian case study. *Questions that remain include: 1) Can this methodology now be extended and improved? 2) How can daily maps of climatic elements at ground resolution of 1 km for the global land mass be produced? 3) Can this method be implemented using publicly available global data sets? 4) What remote sensing images and environmental layers can be used to model the trend? 5) Can we make an automated mapping procedure that can be applied to create an archive of global weather patterns at very high resolution for serving daily maps at 1 km and are similar to* `WorldClim.org` *?* (`WorldClim.org` as a climatic repository Hijmans et al. (2005).)

Chapter 3 provides a review of publicly available meteorological data sets form ground stations and remote sensing. In addition, the chapter discusses the results of analysis, points to the possible problems with using this data for climatological mapping and suggests new directions of development in creating daily spatial grids for global land areas using the spatio-temporal regression kriging. International datasets from ground stations used in this thesis include the Global Surface Summary of Day (GSOD) disseminated by the National Climatic Data Center (NCDC) and the data set from the European Climate Assessment and Data sets (ECA&D). This data is intended to be used for free and is unrestricted when used in research, education, and other non-commercial activities.

Predictions of the daily mean, the minimum and maximum air temperatures using spatio-temporal regression-kriging with a time series of MODIS 8 day images, the topographic layers (DEM and TWI) and a geometrical temperature trend as covariates is described in detail within Chapter 4. The methodology, accuracy assessment and prediction is made for year 2011, but the same methodology can be extended for the whole range of the MODIS LST images (2001–today).

Automated mapping is a data driven approach to mapping with little or no human interaction.In this thesis, geostatistical mapping is assumed and used, which mostly requires expert involvement in the mapping procedure. An automated mapping framework for the mapping daily meteorological observations using spatio-temporal regression kriging is developed and implemented in the R environment (R Development Core Team, 2012) as a package called `meteo`. The implemented framework, which also includes some special adaptations for climatic mapping, is described in the Chapter 5. The package source code is available on `https://r-forge.r-project.org/projects/meteo/`.

Chapter 6 presents the spatio-temporal visualisation of meteorological data using plotGoogleMaps (Kilibarda and Bajat, 2012), which is a part of the automated mapping framework, and also highlights isolated solutions for scientific cartographic communication in climatic mapping research (Luković et al., 2013). Interactive web maps related to the results of this thesis were produced using plotGoogleMaps and are available on `http://dailymeteo.org/`.

Chapter 2

Methods for the interpolation of climatic variables

In this chapter the basic principles of the spatial interpolation methods mostly used in climatology and meteorology are presented.

2.1 Introduction

An interpolation method is the process of estimating (assessing) the values of a target variable at any spatial or temporal location where the target variable has not been measured. The mapping of climatic variable is one of the most important tasks for many applications. The term mapping in this study is considered as an interpolation on regular grids that are also called raster grids, climatic images or surfaces.

The interpolation methods used in meteorology and climatology are quite diverse; they range from nearest neighbour methods, splines, regression and kriging, to neural networks and machine learning techniques (Tveito et al., 2006). Hartkamp et al. (1999); Tveito et al. (2006) gives a review of the interpolation methods used in climatic/meteorological mapping (interpolation). There are also various studies reported in literature that describe comparisons of the most commonly used interpolation methods and are given by Price et al. (2000), Jarvis and Stuart (2001), Stahl et al. (2006) and Hofstra et al. (2008).

One of the main applications of using data from meteorological stations is to produce maps showing spatio-temporal patterns of climatic variables. This is not only interesting for predicting approaching events, but also to create an archive of weather patterns (so called *"climate normals"*). Images of weather conditions are commonly produced for different spatial and temporal supports, e.g. ranging from single day to 50 or 100-year time periods at the local, national and global scale. At the global scale, the first monthly images of land surface temperature at 0.5° decimal degrees resolution was produced by Legates and Willmott (1990). They used a collection of data consisting of 24,635 independent terrestrial station records, 2223 oceanic grid-point records and a series of interpolations made using a spherically based interpolation procedure. Leemans and Cramer (1991) generated grids at the same resolution for mean monthly temperature, precipitation and cloudiness using a triangulation network followed by smooth surface fitting. New et al. (1999, 2000) mapped terrestrial climatic variables at 0.5° decimal degrees resolution showing the monthly space-time variability of global land areas excluding Antarctica for the period 1901–2000. Mitchell and Jones (2005) further refined interpolation techniques to produce climatic grids for nine climate variables (temperature, diurnal temperature range, daily minimum and maximum temperatures, precipitation, wet-day frequency, frost-day frequency, vapor pressure, and cloud cover) for the period between 1901–2002, on a monthly temporal scale. Further on, Hijmans et al. (2005) used a thin-plate smoothing spline on a collection of public meteorological data-sets of monthly records to produce global (land mass) climatic images at 1 km resolution for the period from 1960 to 1990. Becker et al. (2012) recently mapped monthly precipitation for the whole world using an empirical interpolation method based on angular distance weighting at resolutions of 0.25°, 0.5°, 1.0° and 2.5° using data from the Global Precipitation Climatology Centre (**GPCC**).

Examples of the most recent applications of interpolation methods on daily observation at regional or global scales are worth listed below. The first global terrestrial gridded data set of daily temperature averages and ranges, and daily precipitation has been developed by Piper and Stewart (1996) for use in terrestrial biospheric modelling. Daily station observations, commencing from the year 1987, have been interpolated to a 1 by 1 degree grid (longitude, latitude) using a nearest neighbours interpolation technique. Frei and Schaer (1998) used an advanced distance weighting scheme commonly adopted for the

analysis of precipitation on a global scale to create a daily precipitation grid at 25 km resolution. The produced maps covering the European Alps were based on a station network with more than 6000 stations within countries of the Alpine region. Global daily predictions of meteorological variables were produced by Kiktev et al. (2003); Alexander et al. (2006), who used an angular distance weighting technique to interpolate extreme daily precipitation and temperature indices onto a 2.5° latitude by 3.75° longitude grid. Caesar et al. (2006) mapped daily maximum and minimum temperature anomalies using the same method and the same output resolutions as Alexander et al. (2006). There are now numerous approaches to produce daily weather images with fine resolutions at regional or local scales. Haylock et al. (2008) produced European-coverage maps of daily mean, minimum and maximum temperature and precipitation at 0.25° and 0.5° resolution using the European Climate Assessment and Dataset Project (**ECA&D**). These maps were generated by first estimating monthly averages, whereby daily anomalies from those averages were interpolated using kriging and added back to monthly estimates (Haylock et al., 2008). Van den Besselaar et al. (2011) mapped sea level pressure for Europe using the same data source and global kriging. Di Luzio et al. (2008) presented a method for mapping daily precipitation and temperature across conterminous USA at 2.5 arc-minutes (around 4 km) for the period of 1960–2001. Their method also combines interpolation (inverse distance weighting) of daily anomalies from respective gridded monthly estimates. In their case, interpolations were generated from the Parameter Elevation Regressions on Independent Slopes Model (PRISM).

The different interpolation methods listed above can be divided into several categories according to the fundamental mathematics they are based on. Classification given by Tveito et al. (2006) is given here:

1. Deterministic methods,
2. Probabilistic methods,
3. Artificial neural networks,
4. Physical methods,
5. Hybrid methods.

In this chapter the basic principles of the methods mostly used in climatology and meteorology are presented. The presentation of the methods will be structured in sections: "Deterministic methods", "Probabilistic methods" and "Artificial neural networks". The described interpolation techniques, in this chapter, have been widely applied in spatial modelling and not just in meteorology/climatology. Two presented methods specially developed for meteorology and climatology are presented in a separate section entitled "Methods specially developed for meteorology and climatology", and hybrid methods are described in the section "Probabilistic methods". Physical methods are not included in the scope of this study.

2.2 Deterministic methods

Deterministic interpolation techniques create surfaces from measured points that are based on either the extent of similarity (inverse distance weighted) or the degree of smoothing (polynomial characteristics). Similar to other interpolation groups, deterministic interpolation techniques can be global or local. Global techniques calculate predictions using the entire set of observations in the domain of interpolation. Local techniques calculate predictions from the measured points within neighbourhoods smaller then the study area. These methods are exact interpolators (splines can be exact but not necessary) and the resulting surface is made to pass through the data values.

The most used deterministic methods in meteorology and climatology are:

1. Nearest neighbours,
2. Triangulation,
3. Inverse distance weighting,
4. Splines and local trend surfaces,
5. Thin plate splines.

2.2.1 Nearest neighbours

Assuming that the area of interpolation is divided into polygons, the nearest neighbour method predicts the value of a variable at a target point that depends on the accompanied polygon. Due to the fact that the area of interpolation can be divided into polygons in an infinite number of ways, the Thiessen (or Dirichlet/Voronoi are also the same methods but with different names) method is used for dividing a domain of interpolation. Specifically, it divides a test area into polygons using lines that are equidistant between pairs of observation stations.

The Thiessen method also has two other requirements when forming polygons. The first requirement is that each polygon contains only one observation point. The second requirement is that any unobserved location from the polygon is closer to its enclosed observation than to any other observation or observations contained within other polygons. So, the area of the one polygon is considered as the area of a unique value of target variable, where by the value is the same as the observation enclosed within the polygon (see Figure 2.1). The predicted image (surface) is similar to a mosaic, depending only on geometrical distribution of observation. For more details about the nearest neighbours method see Ripley (1981); Isaaks and Srivastava (1989); Burrough and Mcdonnell (1998); Li and Heap (2008).

The nearest neighbours method has rarely been used in meteorology and climatology in recent times. However, the method is widely used in hydrology for estimating areal precipitation, but gives poor accuracy in comparison to novel methods.

2.2.2 Triangulation

The triangular irregular network (also known as Triangulation) was developed for digital elevation modelling and is also a geometrical method. The area of interpolation is divided into a network of non-overlapping triangles between observation points (similar to the network of polygons in the nearest neighbours method) but the triangles are empty. Unlike the nearest neighbors method, observation points are vertices of the triangles, see Figure 2.2. Triangle network creation follows Delaunay triangulation principals (for details and variants see Tsai (1993)). Simplistically, the method tends to avoid skinny triangles.

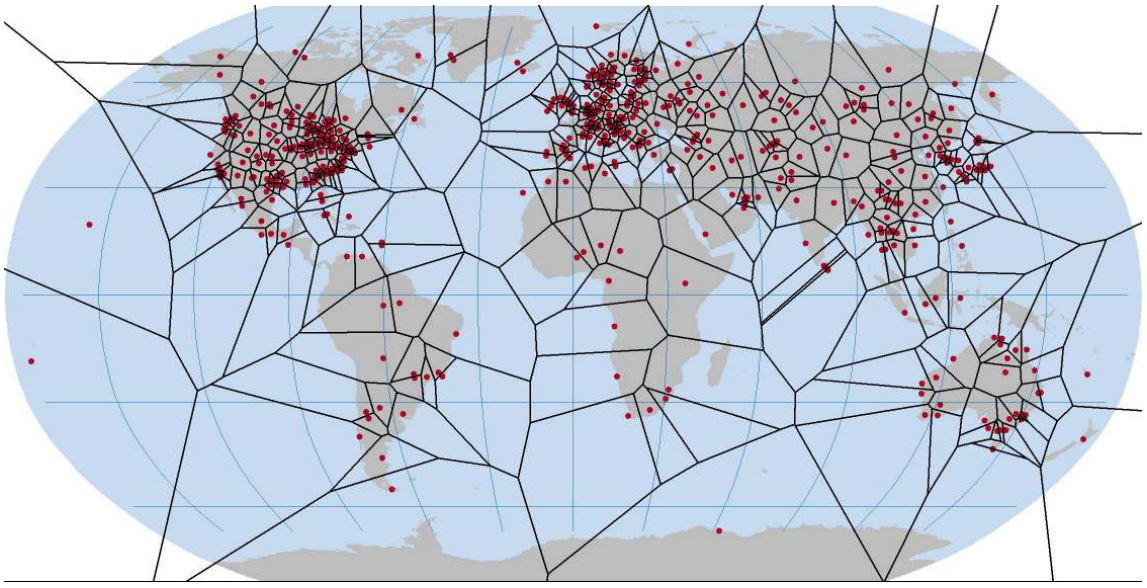

FIGURE 2.1: Sample of meteorological stations and accopanied Thiessen's polygons. Projected in the Robinson projection system.

An added bonus is that each triangle will become part of a 3D surface (e.g. similar as in terrain modelling) if we consider observation values as third coordinates of the triangle vertices. As such, the constructed triangulated surface provides a surface representation of the target variable (estimated values at any location in study area).

The triangulation method is efficient for estimation purposes but, as of late, is rarely used in meteorology and climatology. The main reason for this is that the method is very sensitive to the sampling design. The sampling design should cover all characteristic points of the target phenomenon, e.g.all the points where the phenomenon has local minimum or maximum. The approach is commonly used in geodesy and land surveying where topographic points are well designed and sample sizes are relatively large compared when compared to other disciplines. In contrast, sample sizes for meteorology and climatology are considerably smaller.

2.2.3 Inverse distance weighting method

The inverse distance weighting or inverse distance weighted (IDW) method estimates the values of a variable at an unobserved location using a linear combination of values at

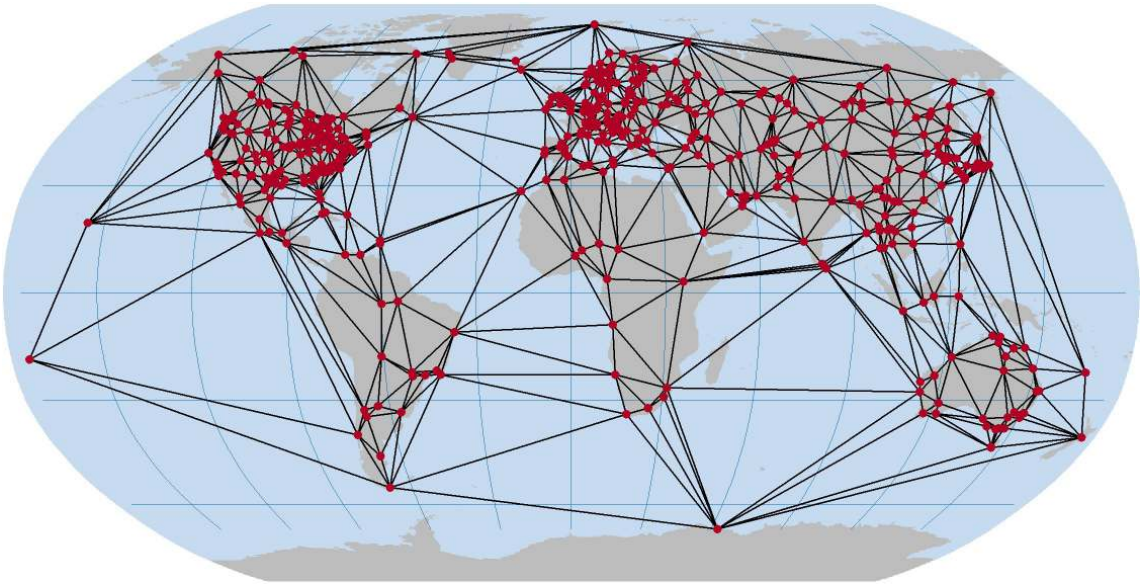

FIGURE 2.2: Sample of meteorological stations and accopanied triangulation polygons. Projected in the Robinson projection system.

sampled points. The weights of this linear combination are proportional to the inverse of the distance between the interpolated and measured points. These weights are then normalised so the sum for all stations is equal to 1 (within the search neighbourhood for local IDW) (Tveito et al., 2006). The formula for the IDW estimation at an unmeasured location $\mathbf{s}_0$ is:

$$\hat{z}(\mathbf{s}_0) = \frac{\sum_{i=1}^{m} \frac{1}{d_i^p} z(\mathbf{s}_i)}{\sum_{i=1}^{m} \frac{1}{d_i^p}}, \tag{2.1}$$

where $z(\mathbf{s}_i)$ the measured value at location $\mathbf{s}_i$, m is number of observations used for prediction (if $m = n$ where n is total number of observation it is global IDW, but often $m < n$, local IDW, the searching criteria is based on the maximum distance to used observations for estimation or maximum numbers of neighbouring observations), p is a power parameter, d_i is distance from estimated location to i-th observation.

Formula 2.1 shows that if the power parameter is higher, the nearby observation has a heavier weight and has more influence on the estimation. This means that higher power

parameters hardly decreases the influence of the observations that are far apart from the prediction location, because weight is defined as powered inverse distance. Therefore, the high power parameter gives the high local influence of closer observations (almost local interpolation) even if global IDW is used, because influence of distant observations is small. IDW is referred to as a "moving average" when p is zero, "linear interpolation" when p is 1, "weighted moving average" when p is not equal to 1 and "inverse distance squared method" when p is two (Li and Heap, 2008). The choice of the parameter p is often arbitrary and gives a biased solution for the result of interpolation (Burrough and Mcdonnell, 1998). However. the p can be chosen on the basis of error measurement, e.g. root mean square error or minimum mean absolute (MAE) error, to optimize the IDW. The calculation of the optimal power parameter that is based on MAE is provided by intamap, R software package (Pebesma et al., 2010).

IDW is frequently applied in climatology and meteorology (Tveito et al., 2006). The weakness of this very fast method is that the direct measurement of uncertainty can not be obtained and that the spatial dependency is only modelled by the inverse distance weights.

2.2.4 Splines and local trend surfaces

Mitas and Mitasova (1999) describes splines as part of variational interpolation methods:

"The variational approach to interpolation and approximation is based on the assumption that the interpolation function should pass through (or close to) the data points and, at the same time, should be as smooth as possible. These two requirements are combined into a single condition of minimising the sum of the deviations from the measured points and the smoothness."

The interpolation function consists of a series of polynomials with each polynomial of degree p. For degree p is 1, 2, or 3, a spline is called linear, quadratic or cubic respectively.

The local trend surface fits a polynomial surface for each predicted point when using only nearby observations. The local influence is ensured by using weighted least squares, whereby the local point is the most influential (Venables et al., 1994).

2.2.5 Thin plate splines

Thin plate splines (TPS), previously known as "laplacian smoothing splines", is a very popular interpolation method in climatic mapping (Hartkamp et al., 1999; Li and Heap, 2008). "The TPS function minimises the surface curvature and imitates a steel sheet forced to pass through the data points: the equilibrium shape of the sheet minimises the bending energy that is closely related to the surface curvature" (Mitas and Mitasova, 1999). In simpler words, the TPS function minimises observation deviations from surface and smoothing. In climatological applications, the smoothing parameter is calculated by minimising the generalised cross validation function (Li and Heap, 2008).

2.3 Probabilistic methods

This group of methods is based on a probabilistic framework in which expected values are of primary importance (Isaaks and Srivastava, 1988). This means that the measured observations represent one of the possible realisations of reality when considering the randomness of observed values. The resulting interpolation produces the expected value along with its associated uncertainty and confidence intervals for the prediction. The following section briefly describes linear regression and geostatistics.

2.3.1 Linear regression

Linear regression explores a possible linear relation (this is stochastic not functional relationship) between the primary variable (interpolated variable, e.g. temperature) and the explanatory variables (e.g. geographical coordinates, elevation, distance to coast line), which are easy to measure (or already known) over the domain of interpolation (Burrough and Mcdonnell, 1998). These explanatory variables are usually referred to as secondary variables, predictors, auxiliary variables, ancillary variables or covariates. Spatial interpolation is often interpolation on a regular grid, so the explanatory variables should also be regular grids that cover the domain of interpolation. In further text below, the explanatory variable is referred to as covariate. Linear regression is often used in geostatistical

applications for de-trending, where the part of variation estimated by regression is called the deterministic part of a target variable phenomenon Z.

The linear regression model is given by:

$$m(\mathbf{s}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{s}), \tag{2.2}$$

where the β_i are unknown regression coefficients, $\mathbf{s}$ any location in domain of interpolation, the f_i covariates that must be exhaustively known over the spatial domain, and p is the number of covariates. Covariate f_0 is taken as unity, resulting in β_0 representing the intercept.

2.3.2 Geostatistics

Webster and Oliver (2007) gives a very interesting history of geostatistics. It is is usually believed that origin of geostatistics were in mining (Krige, 1951), but the first origin was actually in agronomy (Mercer and Hall, 1911, in their article shows yields crop plots and ideas about spatial dependence, correlation range ect.) and the second was in meteorology (Kolmogorov, 1931).

Kolmogorov tried to describe and predict turbulence of the air and weather in his studies. He recognized a spatial correlation phenomenon and modelled it using a 'structure function'. He also tried to apply this function for an optimal and unbiased interpolation method with minimum variance. Kolmogorov's study is published with the name "Interpolated and extrapolated stationary random sequences", in 1942. Inspired by Kolmogorov, Gandin (1963) developed a method for use in synoptic meteorology called optimal interpolation (His research was entitled "Objective analysis of meteorological fields") and the method is very similar to kriging, which was developed at same time by Matheron (1963) in France and is based on Krige's practical studies. Geostatistics includes several methods that use kriging algorithms for estimating continuous variables in space (2D and 3D) and the space-time domain (2D + time).

This section provides a short description of methods published in books written by Isaaks and Srivastava (1989); Cressie (1993); Burrough and Mcdonnell (1998); Webster and Oliver (2007); Hengl (2007); Bivand et al. (2008); Li and Heap (2008).

In geostatistics, the spatial correlation is usually modelled by a variogram. A variogram plots semi-variance as a function of distance. The term semivariogram and variogram are mostly considered as synonymous in geostatistical practices.

For processes modelled using geostatistics, the stationary assumption is considered. The intrinsic stationary assumes that the observation Z can be decomposed into a mean and a residual part (Burrough, 1998; Hengl, 2009):

$$Z(\mathbf{s}) = m + \varepsilon'(\mathbf{s}) + \varepsilon' \tag{2.3}$$

where m is constant mean and $\varepsilon'(\mathbf{s})$ is the spatially correlated stochastic part of variation andε' uncorrelated stochastic component (pure noise) and:

$$\mathbb{E}(Z(\mathbf{s})) = m \tag{2.4}$$

and spatial dependency is defined by the variogram as:

$$\gamma(h) = \frac{1}{2}\mathbb{E}(Z(\mathbf{s}) - Z(\mathbf{s}+h))^2 \tag{2.5}$$

where$h$ is Euclidean distance $|\mathbf{h}|$, $\mathbb{E}$ denotes mathematical expectation and $\mathbf{s}$ is any location in the domain of interpolation.

The variogram model can be understood as measure of the average dissimilarity between data separated in the spatial domain of an interpolation. Typically, we assume that processes occurring spatially close to each other are stronger related than processes occurring farther apart (Tobler 's law).

The sample (experimental) semivariogram $\hat{\gamma}(h)$ can be estimated from a set of observations by calculating the semivariance from observation pairs $z(\mathbf{s}_i)$ and $z(\mathbf{s}_i + h)$, here referred to as separation groups where h can be interpreted as distance intervals (e.g. all

pairs separated from 50-70 km are taken as one group), and every separation group contains $n(h)$ number of pairs.

$$\hat{\gamma}(h) = \frac{1}{2} \cdot \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(\mathbf{s}_i + h) - z(\mathbf{s}_i))^2 \tag{2.6}$$

The sample variogram is used for the fitting of a variogram model. The most used variogram models are: *Nugget*, *Exponential*, *Spherical*, *Gaussian*, *Linear*, and *Power*.

When a variogram model is known (modelled),a spatial covariance function is also implicitly known and kriging interpolation can be performed. An ordinary kriging interpolator (base for many kriging variants) is a linear combination of measured values with weights depending on the spatial correlation between the observations. It is an unbiased interpolator since it aims at minimizing the variance of the errors and the mathematical expectation of the errors is zero.

Kriging covers a range of least-squares methods of spatial prediction.Li and Heap (2008) shortly desribes 22 geostatistical interpolators in their applications and Tveito et al. (2006) remarks several interpolators as important for meteorological/climatic mapping:

1. Ordinary kriging,
2. Universal kriging,
3. Kriging with external drift,
4. Residual kriging,
5. Indicator kriging,
6. Cokriging.

Hengl (2007) shows that both universal kriging, kriging with external drift and regression kriging (residual kriging) are basically the same technique. Bivand et al. (2008) explaine ordinary, universal and kriging with external drift as a special case of universal kriging that depends on trend computation over a domain of interpolation. In the text below, ordinary, regression, indicator and cokriging are briefly presented.

2.3.2.1 Ordinary kriging

Ordinary kriging is by far the most common type of kriging in practice (Webster and Oliver, 2007). An ordinary kriging interpolator is a linear combination of observations that are a weighted sum of nearby observation. The weights depend on the variogram model and the sum of weights is one. These weights are estimated under the condition that kriging variance is minimal.

The ordinary kriging estimator for variable Z at the location $\mathbf{s}_0$ is (Isaaks and Srivastava, 1989; Webster and Oliver, 2007):

$$\hat{Z}(\mathbf{s}_0) = \hat{\boldsymbol{\lambda}}_{\mathbf{0}}^{T} \cdot \begin{bmatrix} Z(\mathbf{s}_1) & \cdots & Z(\mathbf{s}_n) \end{bmatrix}^{T} \tag{2.7}$$

where $\hat{\boldsymbol{\lambda}}_{\mathbf{0}}$ is the estimated vector of weights for the location $\mathbf{s}_0$, n is the number of observation of the variable Z.

The kriging variances and their square roots, the kriging errors, can be mapped similarly and give an idea of the reliability of the maps of estimates. The reliability of ordinary kriging or any kriging interpolator depends on how accurately the variation is represented by the chosen spatial model.

The varinace formula for variable Z at the location $\mathbf{s}_0$ is (Isaaks and Srivastava, 1989; Webster and Oliver, 2007):

$$\sigma^2(\mathbf{s}_0) = \hat{\boldsymbol{\lambda}}_{\mathbf{0}}^{T} \cdot \begin{bmatrix} \gamma(\mathbf{s}_1, \mathbf{s}_0) & \cdots & \gamma(\mathbf{s}_1, \mathbf{s}_0) & 1 \end{bmatrix}^{T} \tag{2.8}$$

The formula for the $\hat{\boldsymbol{\lambda}}_{\mathbf{0}}$ is (Isaaks and Srivastava, 1989; Webster and Oliver, 2007):

$$\begin{bmatrix} \hat{\boldsymbol{\lambda}}_{\mathbf{0}} \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(\mathbf{s}_1, \mathbf{s}_1) & \cdots & \gamma(\mathbf{1}_n, \mathbf{s}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{s}_n, \mathbf{s}_1) & \cdots & \gamma(\mathbf{s}_n, \mathbf{s}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(\mathbf{s}_n, \mathbf{s}_0) \\ \vdots \\ \gamma(\mathbf{s}_n, \mathbf{s}_0) \\ 1 \end{bmatrix} \tag{2.9}$$

the additional parameter μ is a Lagrange multiplier, see details in Isaaks and Srivastava (1989).

Ordinary kriging has become very popular in climatology and meteorology and is often applied as the stochastic component in residual interpolation (Tveito et al., 2006).

2.3.2.2 Regression kriging

Regression kriging uses a spatial multiple regression for de-trending so that the observed phenomenon is decomposed into two parts that are namely 1) the deterministic part (trend) and 2) the residual (regression residuals) stochastic part. Regression kriging assumes that deterministic and stochastic components of spatial variation can be modelled separately. It is mathematically equivalent to the interpolation method variously called "universal kriging" and "kriging with external drift", where auxiliary predictors are used to directly solve the kriging weights (Hengl, 2007).

The regression kriging model for a spatial variable Z at any of the locations$\mathbf{s}$ is:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \varepsilon'(\mathbf{s}) + \varepsilon'' \tag{2.10}$$

where $m(\mathbf{s})$ is the linear regression trend defined by Eq. 2.2, $\varepsilon'(\mathbf{s})$ is the regression residual, stochastic component spatially correlated and ε'' the pure noise component of the observed variable. The sample variogram needs to be calculated from residuals.

The regression kriging estimator for variable Z at the location $\mathbf{s}_0$ is:

$$\hat{z}(\mathbf{s}_0) = \hat{m}(\mathbf{s}_0) + \hat{e}(\mathbf{s}_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot f_k(\mathbf{s}_0) + \sum_{i=1}^{n} \lambda_i \cdot e(\mathbf{s}_i) \tag{2.11}$$

where the $\hat{\beta}_i$ are estimated regression coefficients, the f_i covariates that must be exhaustively known over the spatial domain, and p is the number of covariates. Covariate f_0 is taken as unity, resulting in β_0 representing the intercept. The λ_i are kriging weights determined by the spatial dependence structure and $e(\mathbf{s}_i)$ is the regression residual from an observed location $\mathbf{s}_i$.

The regression coefficients can be estimated using ordinary least squares or Generalized Least Squares (GLS), (Cressie, 1993):

$$\hat{\beta}_{\mathtt{GLS}} = \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{q}\right)^{-\mathbf{1}} \cdot \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{z} \tag{2.12}$$

where $\hat{\beta}_{\mathtt{GLS}}$ is the vector of estimated regression coefficients using GLS, $\mathbf{C}$ is the covariance matrix of the residuals, $\mathbf{q}$ is a matrix of covariate values at the sampling locations and $\mathbf{z}$ is the vector of measured values of the target variable.

The Eq. 2.11 can be rewritten in matrix form and the kriging estimator at the location $\mathbf{s}_0$ is (Christensen, 2001; Hengl, 2009):

$$\hat{z}(\mathbf{s}_0) = \mathbf{q_0^T} \cdot \hat{\beta}_{\mathtt{GLS}} + \boldsymbol{\lambda}_{\mathbf{0}}^{\mathbf{T}} \cdot (\mathbf{z} - \mathbf{q} \cdot \hat{\beta}_{\mathtt{GLS}}) \tag{2.13}$$

where $\hat{\boldsymbol{\lambda}_{\mathbf{0}}}$ is the estimated vector of weights for the location $\mathbf{s}_0$. Prediction variance is defined (Christensen, 2001; Hengl, 2009):

$$\begin{aligned} \hat{\sigma}^2(\mathbf{s}_0) = (C_0 + C_1) - \mathbf{c_0^T} \cdot \mathbf{C^1} \cdot \mathbf{c_0} \\ + \left(\mathbf{q_0} - \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0}\right)^{\mathbf{T}} \cdot \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{q}\right)^{-\mathbf{1}} \cdot \left(\mathbf{q_0} - \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0}\right) \end{aligned} \tag{2.14}$$

where $C_0 + C_1$ is the sill variation and $\mathbf{c}_0$ is the vector of covariances of residuals at the unvisited location, $\mathbf{C}$ is the covariance matrix of the residuals, $\mathbf{q}$ is a matrix of covariate values at the sampling locations, ($\mathbf{q_0}$ is a matrix of covariate values at the unvisited location.

2.3.2.3 Indicator kriging

Typically, indicator kriging is used for mapping binary variables, whereby such variables denote the presence or absence of a phenomenon, e.g. precipitation occurrence.The creation of binary data may be through the use of a threshold for continuous data, e.g. mapping the precipitation higher than the defined threshold. Another example of binary data

creation is the mapping of radiation higher than the allowed limit of radioactivity from the continuous radioactivity observations. Conversion of continuous to binary variable is given:

$$\omega(\mathbf{s}) = \begin{cases} 1 & \text{if } z(\mathbf{s}) < z_c \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

The indicator variable is $\omega(\mathbf{s})$ and is derived from a continuous variable, $z(\mathbf{s})$, which is achieved quite simply by scoring binary values depending on specified threshold z_c.

Converting a continuous variable into an indicator clearly loses much of the information in the original data and it might seem prodigal to transform quantitative data in this way (Webster and Oliver, 2007). This is often the case when many zero data points are in observational data sets and the data distribution is far from a normal distribution; as a result, indicator kriging can then be used for delineation of zero and non-zeros areas. For example, the indicator kriging technique is used first to delineate the raining areas from rain gauge observations and then ordinary kriging or regression kriging is used to determine the rainfall estimates in raining areas.

The results of the indicator kriging are values lying between 0 and 1. Such values are effectively the probability levels given to the data. Zero probability $Z(\mathbf{s}) < z_c$ is defined with 0 and 100% probability with 1.

2.3.2.4 Cokriging

Cokriging estimator beside primary variable uses additional variable(s) that exhibit some correlation with the primary variable. An additional (secondary)variable is known at sampled locations and often on the more discrete point in the domain of interpolation, but limited number of points,not over all domain of interpolation like covariates in regression kriging. Cokriging covariance matrix depends on the variogram and the cross-variogram model. The cross-variogram model shows a correlation of the spatial variability of a target variable with the secondary variable, see Eq 2.16. Cross-correlated information contained in the secondary variable should help reduce the variance of the estimation errors and

the result should not be worse than univariate kriging, which does not account for cross-correlation.

The sample cross-semivariance (or cross-variogram) $\hat{\gamma}_{12}(h)$ for two variables can be estimated from the observations by calculating the semivariance from the observation pairs of z_1 and z_2 , $n(h)$ (as described previously in the methods section), which is the number of point pairs separated with h, distance in Euclidian space.

$$\hat{\gamma}_{12}(h) = \frac{1}{2} \cdot \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z_1(\mathbf{s}_i + h) - z_1(\mathbf{s}_i))(z_2(\mathbf{s}_i + h) - z_2(\mathbf{s}_i)) \tag{2.16}$$

It was shown that cokriging gives better results in comparison to the univariate kriging approach particularly when spatial correlation between secondary variables (covariables) and the variable of interest is high and when the covariables are oversampled with respect to the target variable (Tveito et al., 2006).

2.4 Methods specially developed for meteorology and climatology

2.4.1 PRISM

PRISM is an acronym (Parameter Regression on Independent Slopes Model) for a method developed by Daly et al. (1997). PRISM is a knowledge based method that uses point observations, digital elevation models and other spatial datasets to produce climatic maps based on climate-elevation regressions. A detailed description of the PRISM knowledge based method is given by Daly et al. (2002). PRISM is based on a linear regression function accounting the dominant influence of elevation on the climatic maps. The method interpolates the target variable using a weighted combination of stations data, where weights for a specific location depend on many distinct types of spatial information, e.g. distance (from station to grid point), elevation, cluster, vertical layer (includes a two-layer of atmosphere representation), topographic facet (e.g. rain shadows, windward sides ect.), coastal proximity, and effective terrain weights (expert based).

Tveito et al. (2006) summarise the PRISM advantage in comparison with traditional methods:

"*In a comparison with kriging and detrended interpolation PRISM was shown to be the preferable method in regions with sparse station networks and relatively low precipitation gradients, and very powerful in areas where the station network is unrepresentative for the variation in topography. The more traditional methods showed better results in the areas with a very dense network, where the variability due to the terrain is represented by the stations.*"

2.4.2 AURELHY

AURELHY stands for "Analyse Utilisant le RElief pour l'Hydrométéorologie" (Topography-based analysis for hydrometeorology). In this method, the local topography is used to explain variables by multiple linear regression and regression residuals are interpolated by ordinary kriging. This can be considered as a typical regression kriging method.The AURELHY method was introduced by Meteo France (Benichou and Le Breton, 1987). There are many similar examples of regression kriging like methods that are named differently, e.g. MISH (Meteorological Interpolation based on Surface Homogenized Data Basis) recently developed at the Hungarian Meteorological Service.

2.5 Artificial neural networks

Artificial Neural Networks (ANN) is machine learning technique used for data analysis and modelling in many fields and have been recently applied in spatial prediction. The machine learning algorithms model specific phenomena using an empirical approach for finding the relation of the input and output parameters via a computer program. ANN computing techniques are adaptive and "learning by example" replaces traditional "programming" in solving problems. In traditional programming, the modelling process is coded into the computer program based on the physical or statistical model (stochastic or functional relationship is defined). In contrast, themodel is unknown to the user in the "learning by example" approach. Instead, software that uses a mathematical/statistical

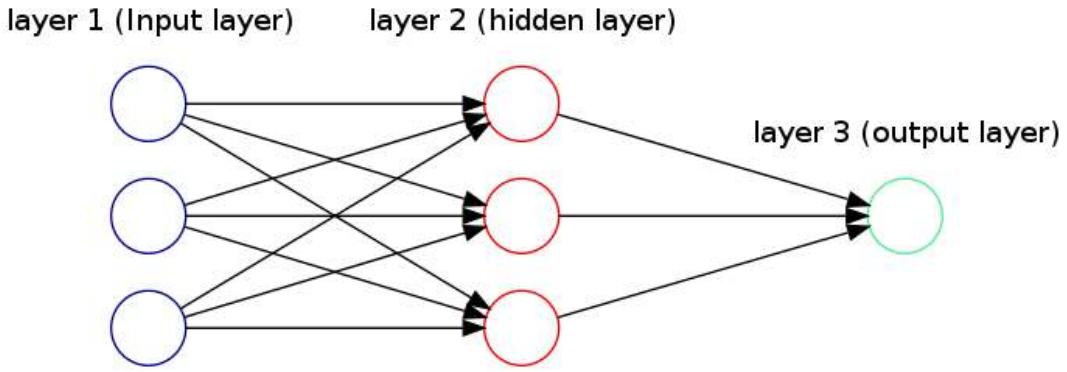

FIGURE 2.3: Simple model of artificial neural networks

model of optimisation synthesizes the relationship between input and output parameters for the user.

The application of machine learning in spatial prediction and analysis is given by Kanevski et al. (2009) in the book " *Learning for Spatial Environmental Data: theory, applications and software*", where many applications in climatic mapping and a short description of ANN are supplied . There are many applications of the ANNs in meteorological/climatic mapping and some of the first studies cited are Hsu et al. (1997); Demyanov et al. (1998); Antonić et al. (2001).

ANN are inspired by the structure of the biological neural network that are composed of a set of numerous interconnected elements that process information (impulses, signals). ANN uses a set of non-linear functions as the so called processing elements (neurons, cells or nodes). The input information (signal, set of known parameters, e.g. longitude, latitude, elevation) are passed to nodes (non-linear functions) and the target variable (e.g. temperature) is modelled as a linear combination of the nodes (non-linear functions). Figure 2.3 shows a simple model of the ANN.

The blue circles from Figure 2.3 represent the input parameters in a similar fashion as regression known over the domain of interpolation (e.g. elevation, longitude, latitude), the red circles represent the nodes (neurons). The nodes are functions defined generally by formula:

$$f(\mathbf{x}, w) = f(\sum_{i=1}^{K} \omega_i x^i + b) \tag{2.17}$$

where x^i are some elements (some predictors) of the input vector $\mathbf{x}$, ω_i are weights and b is bias passed to the function f, so called transfer function. Example of transfer function is hyperbolic tangent:

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.18}$$

Therefore, the resulting circle (the green one depicted in Figure 2.3) is the prediction representing the linear combination of processing components (nodes). Thus, the weights from the Eq. 2.17 are essential parts of the ANN technique since they are fitted (tuned) by the ANN algorithm to match certain criteria that is based on reducing error from predicted and observed values.

The most frequently used neural networks consist of multi-layers, which contain several hidden layers of neurons that are fully connected. The step by step procedure for using a multilayer neural network in spatial interpolation is given by Kanevski et al. (2009), who strongly recommend using a combination of geostatistical and ANN approaches.

Chapter 3

Publicly available global meteorological data sets and preliminary spatio-temporal analyses[1]

The chapter reviews publicly available global meteorological data sets from ground-based stations and/or remote sensing systems, prepared and maintained by national and international organizations: the National Aeronautics and Space Administration (NASA), Global Precipitation Climatology Centre (GPCC), European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), and United States National Oceanic and Atmospheric Administration (NOAA) with focus on available data. A merge of the Global Surface Summary of Day (GSOD) and the European Climate Assessment & Dataset (ECA&D) consisting of 10,695 stations for the year 2011 were assessed for representation and usability for global spatio-temporal analysis. Three aspects of data quality were considered: (a) representation in the geographical and temporal domains, (b) representation in the feature space (based on the MaxEnt method), and (c) usability i.e. fitness of use for spatio-temporal interpolation (based on cross-validation of spatio-temporal regression-kriging models). The results show that clustering of meteorological stations in the combined data set is significant in both geographical and feature space. Most of the

[1] Based on article: Kilibarda M, Perčec Tadić M, Hengl T, Luković , Branislav B (2013?) Publicly available global meteorological data sets: sources, representation, and usability for spatio-temporal analysis. Under review in *International Journal of Climatology*

distribution of the stations (84%) can be explained by population density and accessibility maps, also with elevation, showing that higher elevations areas are less covered with stations, as spared populated and inaccessible areas. Although a spatio-temporal regression-kriging model of mean daily temperature on 8-day MODIS LST images produced average global accuracy of 2–3°C, prediction for polar areas and mountain chains was 2 times worse than for areas densely covered with meteorological stations. Despite the geographical and feature space clustering, the presented global spatio-temporal model using station observations and remote sensing images can be used for production of global mean daily air temperature images at very high resolution.

3.1 Introduction

Publicly available meteorological and/or climate data are also one of the foundations of democracy. Combine open access data with the open source software tools and everyone can build his/her opinion about global change. The same way anyone is now able to zoom into the Ikonos and QuickBird images available via the Google Engine (Butler, 2006), and witness deforestation or massive land degradation (possibly not reported anywhere yet!), anyone should also be able to plot meteorological variables per station, per region, for any given selection, and identify possible changes and trends. Fortunately, *"open-access climate science is becoming easier than ever"* (Kleiner, 2011). There are now multiple data portals where anyone can download original meteorological data in a variety of formats. Some of the major U.S. and global open meteorological data sources have been reviewed by Yang et al. (2010). The most popular sources of freely available satellite data for agro-climatic application have been described by Toulios and Stancalie (2011). Becker et al. (2012) provide a review of global precipitation data sets and their limitations for global change analysis. Recently, several data portals have been launched to support free exchange of climatic and meteorological data sets. The University Corporation for Atmospheric Research (UCAR) Climate Data Guide[2], for example, is a repository for the climate community supporting a wide range of observational data sets and their appropriate use in analyzes and climate model evaluation. The data comprises ground and satellite observations and re-analyzes and model simulations with links to sources of data. Royal

[2] `https://climatedataguide.ucar.edu`

Netherlands Meteorological Institute (KNMI) Climate Explorer portal[3] is another source of data and tools for climate research. In the framework of the Berkley Earth Project[4], monthly temperature observations from 16 sources, including the National Climatic Data Center (**NCDC**) data, were used to produce average, minimum and maximum temperature anomaly grids.

It is for example well know that meteorological stations are often allocated to represent provide higher density information for areas of high population density. Consequently, mountains and uninhabited areas are often miss-represented in the national observation networks. See for example maps of station location used by Vose et al. (1992); Peterson and Vose (1997); Klein Tank et al. (2002) and/or Lawrimore et al. (2011). In addition to the problem of station clustering, most spatial prediction methods do not consider varying uncertainty in input data and its effects on the final outputs. Most of the climate grids that can be find on the data portals listed above have no uncertainty measures attached, or the uncertainty is not spatially assessed. This is obviously a problem because the data quality is an important aspect for decision making. The scale of global or regional temperature change is often very fine (e.g. 0.2–0.5°C) and high uncertainty can lead to misinterpretation of produced patterns (Rohde et al., 2012).

The national budgets for weather monitoring in the developing countries are of course very limited. Consequently, the representation of stations globally does not necessarily reflect complexity of climate. This is a well know paradox in ecology that we probably know the least about the areas of the highest ecological and climatic complexity (Chapman, 2005). Thus, consistent and harmonized (unbiased) grids are needed for global analysis and climatic planning. But can we produce such harmonized grids using the ground and remote sensing data we have at the moment at all?

In this paper we look at the general usability of global meteorological data sets coming from ground-based stations and/or remote sensing systems which are of interest for spatio-temporal analysis. The first part of the paper provides a review of the available meteorological data sets. The second part shows the results of analysis, points the possible

[3] `http://climexp.knmi.nl`
[4] `http://www.berkeleyearth.org`

problems with using this data for climatological mapping and suggests direction of development of daily spatial grids for global land areas using the spatio-temporal regression kriging.

We focus on three important aspects of publicly available data: (a) *content*, (b) *representation*, and (c) *suitability for spatio-temporal interpolation*. For this purpose we have merged GSOD and ECA&D station observation, as clean and consistent meteorological (point), and then run a number of standard spatial analysis operations to determine what could be possible problems with using the global data sets for spatio-temporal interpolation and time-series analysis. The spatio-temporal regression kriging model on MODIS LST 8 day images is made just for mean daily temperature records for a year 2011.

3.2 Measurements at ground stations

Climate research relies heavily on the records from instruments at these near-surface weather stations (Peterson and Vose, 1997), as the most accurate and reliable measures of weather. Although, ground station measurements are the most accurate and the most reliable records of the weather at near surface, they are the only one available records of spatial and temporal variability of climatic variables to 1960, when the first weather remote sensing mission had been lunched by NASA.

Flowing sections describes publicly available data sets at global or near global coverage.

3.2.1 NCDC's Global Surface Summary of Day (GSOD)

The Global Surface Summary of Day (**GSOD**) data set is produced and archived at the NOAA's National Climatic Data Center (NCDC) under the code `NCDC DSI-9618`. The input data used in building GSOD are the Integrated Surface Data (`NCDC DSI-3505`), which includes global hourly data obtained from the US Federal Climate Complex (FCC) consisting of about 27,000 stations.

GSOD (Fig. 3.1) is certainly the most consistent and probably still the largest publicly available international station data sets. It contains daily measurements for a list of 11

meteorological parameters of several climatic variables (since 1929): temperature, humidity, pressure, wind, precipitation (liquid and solid) and phenomena:

1. mean, minimum and maximum temperature (precision of 0.1°F),
2. mean dew point (0.1°F),
3. mean atmospheric pressure and mean sea level pressure (0.1 mb),
4. mean visibility (0.1 miles),
5. mean wind speed, maximum sustained wind speed and maximum wind gust (0.1 knots),
6. precipitation amount (0.01 in),
7. snow depth (0.1 in),
8. and an indicator (class) for occurrence of fog, rain or drizzle, snow or ice pellets, hail, thunder, and tornado/funnel cloud.

This data set is continuously being updated so that the latest daily summary data are normally available 1–2 days after the date-time of the observations used in the daily summaries. The data summaries provided in the GSOD are based on data exchanged under the World Meteorological Organization (**WMO**) World Weather Watch Program, following the WMO Resolution 40 (WMO-No. 837, 1996). This allows WMO member countries to place restrictions on the use or re-export of data for commercial purposes outside of the receiving country. The GSOD data is intended for free and unrestricted use in research, education, and other non-commercial activities.

3.2.2 NCDC's Global Historical Climate Network Dataset

The previously mentioned GSOD together with more than 20 other sources are part of the world's largest collection of daily climatological data that is the **GHCN** (Global Historical Climatology Network)-Daily database (GHCN-D). It contains historical data on daily temperature, precipitation and snow over the global land areas and it is updated daily where possible. One or more of the 40 meteorological elements (maximum/minimum

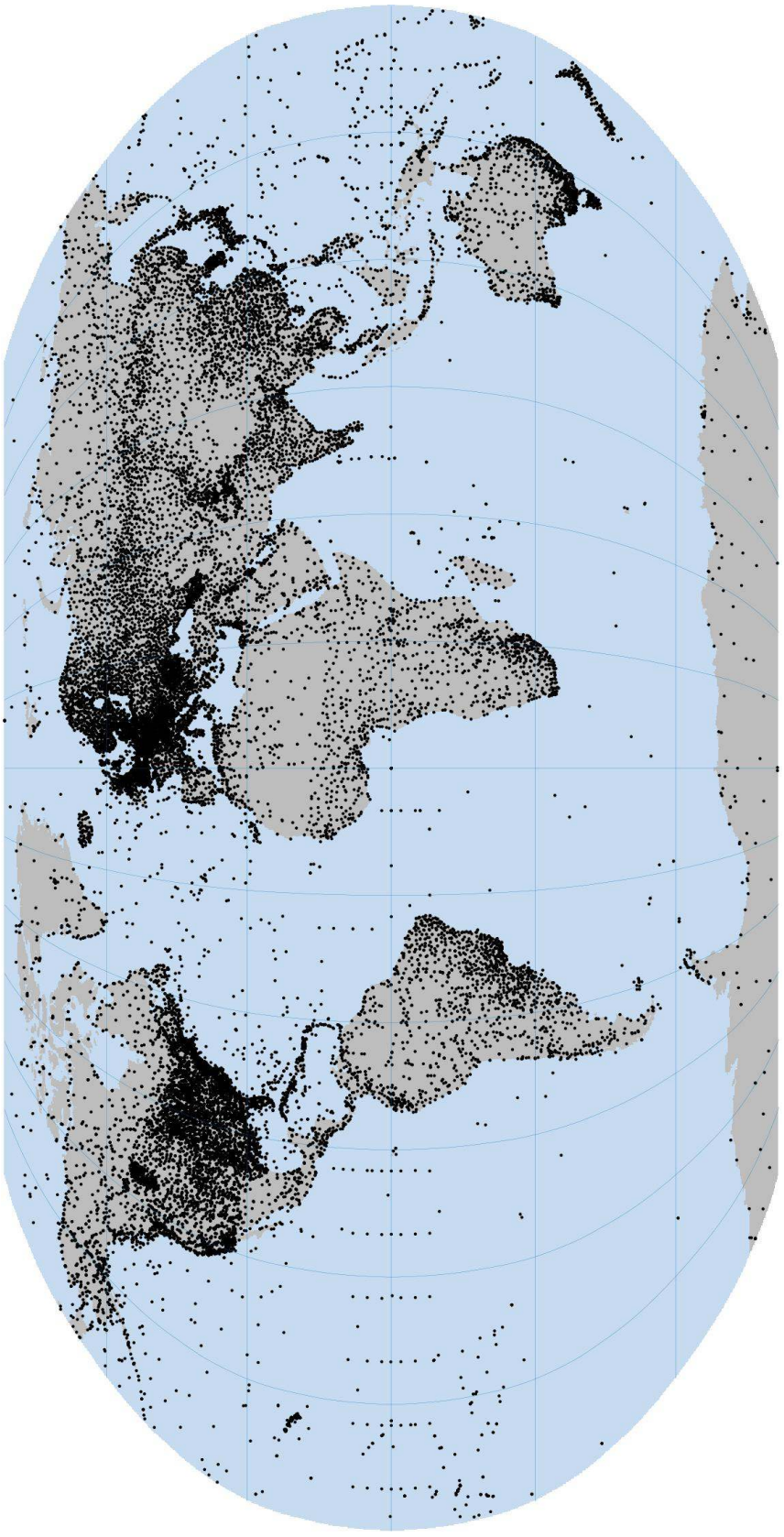

FIGURE 3.1: International network of the meteorological stations (around 27,000 stations shown). GSOD (Global Surface Summary of Day) stations (9000) are a subset of the total network of stations for which harmonized data is available. Projected in the Robinson projection system.

temperature, precipitation, snowfall, snow depth, snow water equivalent, wind maximum, cloudiness, etc.) are collected on more than 80,000 stations in 180 countries and territories (Menne et al., 2012). GHCN-D is especially useful for monitoring the frequency and magnitude of extremes due to high temporal resolution.

The GHCN-Monthly (GHCN-M) temperature data set was first developed in the early 1990s (Vose et al., 1992). A second version was released in 1997 following extensive efforts to increase the number of stations and length of the data record (Peterson and Vose, 1997). GHCN-M version 3 released in 2011 focused on four areas (Lawrimore et al., 2011):

(*a*) *consolidating duplicate station records*;

(*b*) *improving station coverage, especially during the 1990s and 2000s*;

(*c*) *enhancing quality control, and*

(*d*) *applying a new bias correction methodology that does not require use of a composite reference series*.

The version 3 currently contains monthly mean temperature, monthly maximum temperature and monthly minimum temperature. The station network for the time being, is the same as GHCN-M version 2 (NCDS, 2012). GHCN-M version 2 has data for precipitation (20,590 stations, at precision of 0.1 mm), mean temperature (7280 stations, at precision of 0.01°C; Fig. 3.2), and minimum and maximum temperature (4966 stations, at precision of 0.01°C).

The GHCN-M has geographical station information such as latitude, longitude, elevation, station name, etc., and also extended metadata information, such as surrounding vegetation and similar[5].

3.2.3 European Climate Assessment & Dataset

The largest European publicly available meteorological data set is the European Climate Assessment and Dataset Project (**ECA&D**). The idea of ECA&D project was to provide

[5] `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v3/README`

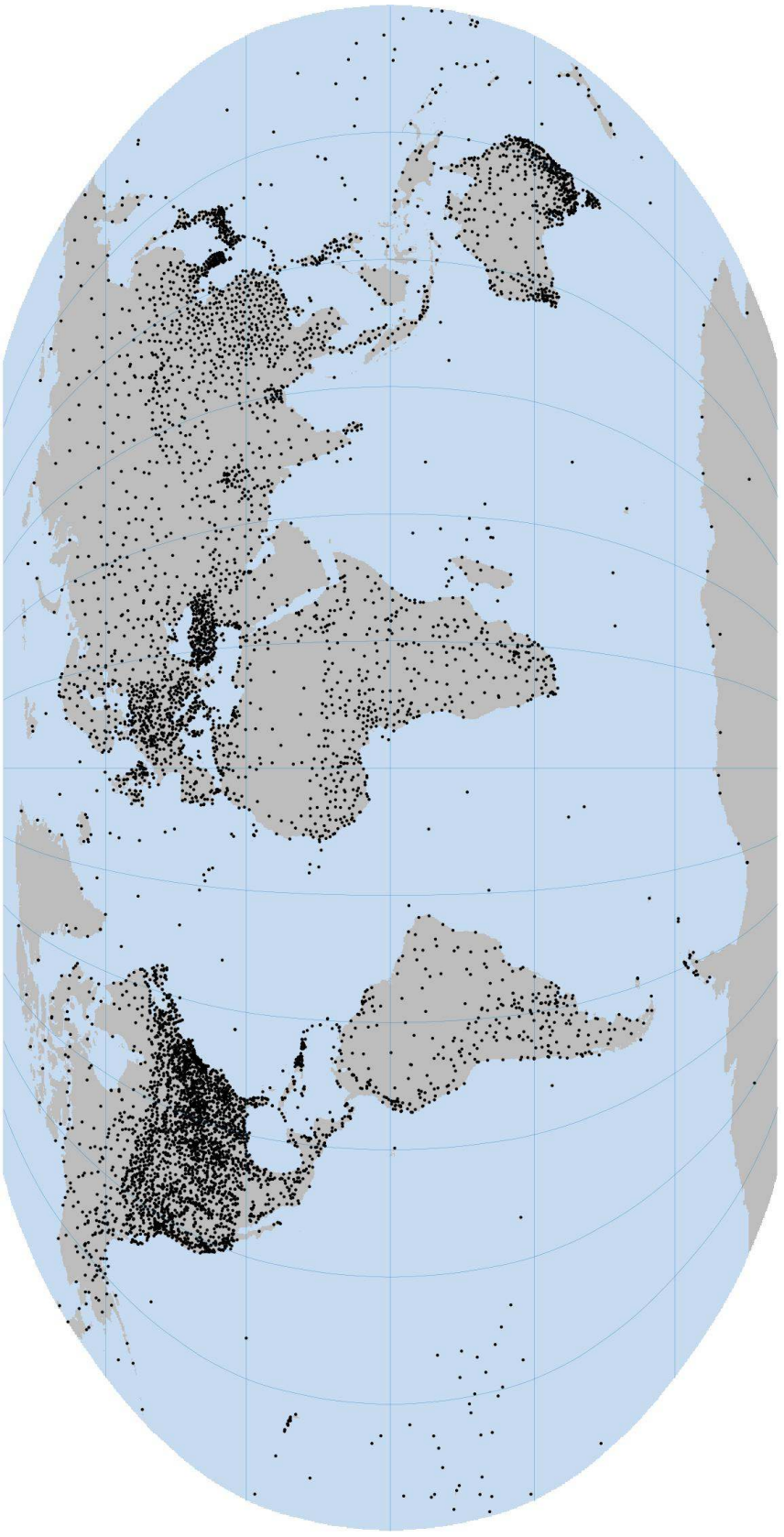

FIGURE 3.2: Locations of the meteorological stations (ca 7280) included in the GHCN-M (Global Historical Climatology Network-Monthly) mean temperature data set.

uniform analysis methodology to daily observational series from 62 European countries and 6596 European and Mediterranean meteorological stations (Klein Tank et al., 2002). Fig. 3.3 shows the geographical distribution of stations with daily time series, throughout Europe and the Mediterranean (around 2700 stations). The number of observed daily climatic elements varies in geographic and time domain. This data set contains measurements for the following meteorological variables and their parameters:

1. minimum, mean and maximum temperature (resolution of 0.1°C),
2. humidity (1 %),
3. mean sea level pressure (0.1 hPa),
4. mean wind speed (0.1 m/s), wind direction (degrees), maximum wind gust (0.1 m/s),
5. precipitation amount (0.1 mm),
6. snow depth (1 cm),
7. cloud cover (oktas),
8. sunshine duration (0.1 hours).

Only a portion of this data can be downloaded and used freely for non-commercial research. On the other hand, licensed daily data are used together with the publicly available daily data to calculate derived value-added products, such as indices of extremes or daily maps of gridded data (E-OBS) available from the ECA&D project web page[6].

3.2.4 Aviation Routine Weather Report (METAR)

Another station data set of interest for global analysis is distributed as the international standard code format for hourly surface weather observations — **METAR**. METAR roughly translates from French as *Aviation Routine Weather Report* and is predominantly used by pilots in fulfillment of a part of a preflight weather briefing, and by meteorologists. Typical METAR report contains data for temperature, dew point, wind speed and direction,

[6] `http://eca.knmi.nl/documents/ECAD_datapolicy_v5.pdf`

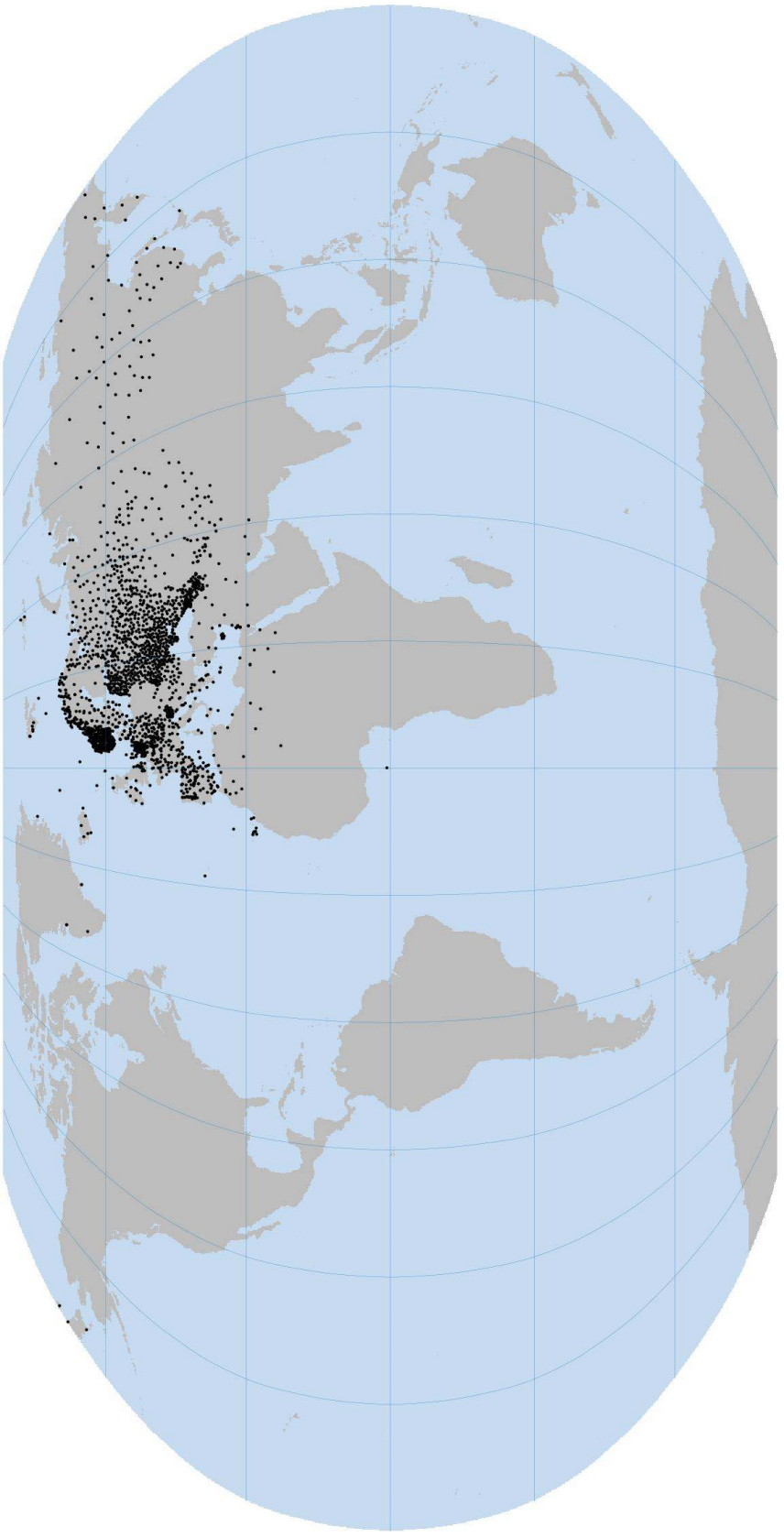

FIGURE 3.3: Locations of the meteorological stations (ca 2700) included in the ECA&D (European Climate Assessment and Dataset) project.

precipitation, cloud cover and heights, visibility, and barometric pressure. Current data, original and decoded, for individual stations are available from the US National Weather Service (**NWS**) FTP site[7].

Historical data of METAR reports are not available from official FTP site. Current observations in KML format[8] are available from the NWS web site showing also the largest density of observations in US and Europe.

3.2.5 Climatic Research Unit (CRU) land station temperature database

In addition to the meteorological data at point support, a number of data sets is available also at block supports. For example the Climatic Research Unit (**CRU**), University of East Anglia, UK derived several gridded monthly temperature products covering the land land/or sea regions on the 5° by 5° grid. These are referred to as the CRU-TEM data sets: CRUTEM3 and the new CRUTEM4 (from March 2012) land air temperature anomalies on a 5° by 5° grid (Jones et al., 2012), and their (variance) adjusted versions CRUTEM3v and CRUTEM4v. CRU and UK Met Office Hadley Centre have also produced combined land and marine (sea surface) temperature anomalies on a 5° by 5° grid (HADCRUT3) and associated variance adjusted versions of HadCRUT3, HadCRUTv3.

The station data used to generate those gridded fields are available from the CRU website[9]. For example, CRUTEM4 underlying data set contains 5583 monthly station temperature time series some extending back to 1850. They cover the global land area, with relatively rough positional precision of the station. The location of stations are available at 0.1 degrees precision in geographic coordinates, that is around 10 km at middle geographic latitudes. HadCRUT and CRUTEM station data had assigned codes to each station, giving the principal source for each series (Brohan et al., 2006).

[7] `ftp://tgftp.nws.noaa.gov`
[8] `http://www.srh.noaa.gov/gis/kml/metar/metarlink.kml`
[9] `http://www.cru.uea.ac.uk/cru/data/temperature/`

3.2.6 FAOCLIM 2.0

Another valuable source of historical meteorological observations is the FAOCLIM 2.0 global climate database i.e. CD ROM (Environment and Natural Resources, 2001). FAOCLIM 2.0 contains monthly data for weather stations across the world. This station database contains monthly data for around 28,000 stations, precipitation data for 27,372 stations, mean temperature data for 20,825 stations, and minimum and maximum temperature data for 11,543 stations (Hijmans et al., 2005). FAOCLIM 2.0 also contains both long-term averages (1961–1990) and monthly time series for precipitation and temperature.

3.3 Publicly available remote sensing data

The first weather Television and Infrared Observation Satellite (**TIROS**-1) was launched on 1960 by NASA. It was in operation for just 78 days, but it sent back thousands of pictures of cloud patterns. It proved the theory that satellites could effectively survey global weather from space. TIROS was followed by nine more test satellites launched between 1960 and 1965 (TIROS X) to provide routine, daily weather observations. This polar-orbiting satellite was providing images of clouds across parts of the globe that may be compared with coincident meteorological observations. Since then, the development of satellite systems has advanced significantly, with different satellite platforms and instruments operating on board (Kidd, 2001).

Satellite images are now routinely used in climate studies due to its availability, spatial coverage and multi-decade length of the series (Struzik et al., 2011). They also allow determination of different climatic parameters at different scales. Some data are freely available trough the Internet while the others require registration. Many research agree that meteorological satellites are the key to weather forecast, and analysis of climate.

Satellite images often need to be calibrated using the ground data which can be tricky since even small systematic differences can lead to wrong conclusions. For example, Santer et al. (2000) found an apparent difference between surface estimated warming of 0.2°C per decade since 1979 and the much smaller temperature trend in the lower troposphere

estimated from satellites and radiosondes. According to them these differences mainly come from data quality problems in either the surface and/or satellite and radiosonde data. Further on, the difference may be due to the effects of natural variability and/or external forcing. A third reason for the observed difference in the temperature trends is the discrepancy in spatial coverage between surface and satellite data. According to Thorne et al. (2005), the choice of data set in meteorological studies can even change the sign of upper-air temperature trends relative to those reported at the surface. Many papers have been published dealing with analysis of ground based and satellite based data (Proedrou et al., 1997; Feidas et al., 2004). The focus is mainly put on assessing the bias of remote sensing based measurements of ground conditions, and on methods that can be used to minimize the bias (Smith et al., 2006).

There is a close cooperation between two US agencies, NASA and NOAA in making the land, oceans and atmosphere visible from space and sometimes their roles is hard to distinguish. While NASA is responsible mainly for satellite pre-launching and launching faze, the NOAA is taking the responsibility for the satellite operation, processing, distribution and archiving of the data. Convention for naming the NASA/NOAA satellites is to set the name and letter before the launch, e.g. last launched meteorological satellite was **GOES**-P, and after it reached its proper orbit the it was renamed GOES-15.

It can be said that all meteorological remote sensing system are often specialized in monitoring specific meteorological features at some working spatio-temporal scale. For practical reasons, distinction between two groups of Remote Sensed (RS) systems is made: RS systems focused on monitoring surface temperatures and RS systems focused on monitoring precipitation (Table 3.1).

3.3.1 The National Oceanic and Atmospheric Administration (NOAA)

NOAA's Satellite and Information Service (**NESDIS**) is leading provider of meteorological satellite data. This USA agency operates meteorological satellites, processes and distributes climate data managing one of the world's largest climate data archive at the National Climatic Data Centre (**NCDC**). NOAA's satellites monitor (Ohring et al., 1989; Ohring and Booth, 1995):

- *Atmospheric temperatures*,
- *Oceanic temperatures*,
- *Greenhouse gases*,
- *Ozone Sea ice, glaciers, snow cover*,
- *Sea level Ocean acidification*,
- *Extreme temperatures and floods/droughts*.

NOAA's satellites provide consistent, long-term observations, 24-hours-a-day, 7-days-a-week, at a basic resolution of 1 km.

The NOAA series of satellites began weather monitoring in 1970. These series of satellites were known as Polar Orbiting Environmental Satellites (**POES**). Currently, NOAA-15 through NOAA-18 serve as the stand-by satellites and NOAA-19 as the operational one performing the morning and afternoon global coverage, orbiting the Earth every 6 hours with a spatial resolution 1.1×1.1 km. The POES among other instruments include the Advanced Very High Resolution Radiometer (**AVHRR**) instrument and the Advanced TIROS Operational Vertical Sounder (**ATOVS**) suite.

From 2012 the Suomi NPP became fully operational new satellite as bridge to the forthcoming series of the advanced Joint Polar-Orbiting Satellite System (**JPSS**) whose first satellite JPSS-1 is planned for launching in 2016. Currently two Geostationary Operational Environmental Satellites (GOES-13 and GOES-15) circle the Earth at the same speed as Earth's rotation. This allows them to hover continuously over one position on the surface. GOES-13 (or GOES EAST) monitors North and South America and most of the Atlantic Ocean, while GOES-15 (or GOES WEST) monitors part of North America and the Pacific Ocean basin. The era of these satellites began with launching of Synchronous Meteorological Satellites-1 (SMS-1) on 1974. However, this program officially started on 1975 in cooperation between NOAA and NASA by launching GOES-1. This series operates until today and will be improved by new generation of GOES-R satellites scheduled for launching in 2015 (Davis, 2007).

TABLE 3.1: Common sources of meteorological RS imagery with near to global coverage.

Provider	*URL*	*Satellite / mission*	*Spatial coverage*	*Temporal coverage*	*Availability*
NOAA	`http://www.ncdc.noaa.gov/oa/satellite.html`	NOAA / GEOS	Global	Since 1979	Free registration required
UAHMSU	`http://discover.itsc.uah.edu/amsutemps/amsutemps.html`	NOAA	Global	Since 1979	Free registration required
EUMETSAT	`http://www.eumetsat.int`	MFG, MSG, METOP, JASON-2	Europe	Since 2002	Free registration required
NASA	`http://modis.gsfc.nasa.gov`	MODIS Terra, Aqua	Global	Since 2000	Free

3.3.2 The National Space Science and Technology Centre (NSSTC)

The National Space Science and Technology Centre (**NSSTC**) is a mission conducting and researching in order to support NASA mission. In cooperation with The University of Alabama in Huntsville (UAH) produces temperature data set for the lower and mid-troposphere and the lower stratosphere that merge data from the nine Micro Sounding Units (MSUs) and two Advanced Micro Sounding Units (AMSUs).

The data are obtained from Microwave Sounding Units (MSUs) on the NOAA's TIROS-N (polar orbiting) satellites, which relate the intensity or brightness of microwaves emitted by oxygen molecules in the atmosphere to temperature. Images and data for the download are available via the NSSTC web site. Spatial coverage of the MSU data set is nearly global while temporal coverage is limited, as the MSU data set is in existence since 1979 (Christy et al., 2000).

3.3.3 European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT)

EUMETSAT is the European operational satellite agency for monitoring weather, climate and the environment. It operates a system of meteorological satellites monitoring the atmosphere and delivering weather satellite data on a daily basis 365 days a year. EUMETSAT is offering a list of atmosphere products available from their geostationary MSG satellites, polar orbiting METOP and low orbiting satellites. EUMETSAT Polar System program (EPS) is European contribution to a joint European-US satellite system called the Initial Joint Polar-Orbiting Operational Satellite System (IJPS). This is an agreement between EUMETSAT and the NOAA on providing instruments for each other's satellites, exchange all data in real time, and assist each other with backup services. Other partners are the European Space Agency (ESA) and the *Centre National d'Etudes Spatiales* (CNES) of France.

3.3.4 National Aeronautics and Space Administration (NASA)

The Moderate Resolution Imaging Spectroradiometer (**MODIS**) images of the Terra and Aqua Earth Observing System (EOS) platforms provide the possibility for retrieving atmospheric, oceanographic including biological variables using different techniques. There are many data products from MODIS observations describing land (temperature, land cover), oceans (sea surface temperature, optical thickness) and atmosphere (water vapor, cloud product, atmospheric profiles) that can be used for studies at different scales, local to global. In meteorological studies are very often used Land Surface Temperature (LST) data and images obtained from MODIS thermal bands and distributed by Land Processes Distributed Active Archive Center (LP DAAC FTP) of USGS (US Geological Survey).

Additionally to this MODIS Level 2 or higher level data, there are also the MODIS Level 1 data that are distributed trough the LAADS portal hosted at Goddard Space Flight Center. The MODIS LST data are available on daily basis and have spatial resolutions of 1×1 km (Coll et al., 2009). The accuracy of MODIS LST is 1°K. However, some validations reported accuracy better then 1°K in clear sky conditions within the temperature range from -10°C to 50°C (Yoo et al., 2011). MODIS LST data and/or images are one of the mostly used and best documented publicly available RS products in the world.

3.3.5 NASA/Goddard's Space flight Center Laboratory for Atmosphere

Precipitation satellites are not able to estimate ground conditions as accurately as e.g. land surface temperature sensors (Mendelsohn et al., 2007). This can be overcome through combining the satellite with the rain gauge data. The international authority that gathers both sources of precipitation data is the Global Precipitation Climatology Project (GPCP), established at the Laboratory for Atmospheres at the NASA Goddard Space Flight Center. The aim of this project is merging the precipitation data taking advantage of the each data type. Rain Gauge data contributing to this project are available from German Weather Service's project GPCC. Satellite precipitation estimates project are computed from geostationary satellites (GOES — United States, Meteosat — Europe, GMS — Japan), and polar-orbiting satellites (NOAA).

Set of precipitation estimates by Geostationary Satellite Precipitation Data Center (**GPSPDC**) is another standard resource of the RS data. Data from NASA Aqua and TIROS are also included in GPCP. Regarding temporal coverage daily precipitation data are available since 1996, while monthly series are available since 1979 in mmday. Precipitation data with a global spatial coverage is currently available only for coarse spatial resolutions of 1°×1° resolution for daily data and 2.5°×2.5° for monthly data. Monthly and daily data are freely available in a FORTRAN binary format with software for reading from NOAA's National Climate Data Center (NCDC), and from the German Weather Service[10]. An overview of the remote sensing system used to map precipitation is given by Prigent (2010).

3.4 Environmental layers

The environmental layers used for this thesis were provided from WorldGrids.org[11] portal. WorldGrids.org is component of GSIF (Global Soil Information Facilities) funded and maintained by ISRIC (International Soil Reference and Information Center). The portal serves continues raster layers, for example, DEM, MODIS Terra products, various climatic, land cover and geological layers. Layers are thematically grouped in sections:

- *Climatic and meteorological images*,
- *DEM–derived parameters*,
- *MODIS products*,
- *Land cover and land use*,
- *Urbanization and Lights at night images*,
- *Biodiversity and human impact maps*,
- *Land, vegetation and water masks*,
- *Harmonized World Soil Database images*,

[10] `ftp://ftp.dwd.de/pub/data/`
[11] `http://www.worldgrids.org/doku.php`

- *Various layers*.

The environmental grids from the data repository are available at various resolutions from 1 to 20 km. The short description of the most commonly used layers for this research is presented in further sections.

3.4.1 Global relief model (DEMSRE)

The environmental layers used for this research were provided from WorldGrids.org portal. WorldGrids.org is component of GSIF (Global Soil Information Facilities) funded and maintained by ISRIC (International Soil Reference and Information Center). The portal serves continues raster layers, for example, DEM, MODIS Terra products, various climatic, land cover and geological layers. Global relief model at 1km resolution was derived as combination of SRTM 30+ and ETOPO DEM at 1/120 arcdeegres resolution. The model is based on SRTM 30+ and ETOPO DEM, publicly available data sets.

Shuttle Radar Topography Mission (SRTM) is an international project managed by National Geospatial-Intelligence Agency (NGA), National Aeronautics and Space Administration (NASA), National Imagery and Mapping Agency (NIMA) and Italian and German space agency (Deutsche Zentrum fur Luft und Raumfahrt - DLR). SRTM for the first time provides a near global high quality DEM at resolution levels of 1 and 3 arc sec, covered land mass between 60°N and 57°S. The horizontal spacing is 1 arc sec; the elevation value is given in meters. WGS84 is used as horizontal and vertical datum. This means that ellipsoidal heights are provided. The DEM accuracy requirements are Âś16 m absolute and Âś6 m relative vertical accuracy (Rabus et al., 2003). For detail description of DEM produced by SRTM see Rabus et al. (2003); Farr et al. (2007). SRTM 30+ is a near-global digital elevation model (DEM) comprising a combination of SRTM data and U.S. Geological Survey's GTOPO30 data set.

ETOPO is a 1 arc-minute global relief model of Earth's surface that integrates land topography and ocean bathymetry. It was built from numerous global and regional data sets covering complete global topographic and bathymetric coverage. The detail description is provided in publication Amante and Eakins (2009).

DEMSRE comes with an accompanying processing script providing detailed instruction for layer reproduction. DEMSRE is a main source of many geomorphometric layers on the portal, such as slope, potential incoming solar radiation, topographic wetness index, etc. Figure 3.4 shows DEMSRE global Earth coverage.

3.4.2 SAGA Wetness Index (TWISRE)

The topographic wetness index (TWI), which combines local upslope contributing area and slope, is commonly used to quantify topographic control on hydrological processes (Sorensen et al., 2006), but also can be used as an indicator of cold air accumulation (Bader and Ruijten, 2008). Methods of computing this index differ primarily in the way the upslope contributing area is calculated. TWI is defined in the equation 3.1 :

$$TWI = \ln\left(\frac{A}{\tan(\beta)}\right) \tag{3.1}$$

where A (m^2) being the contributing area, and $\tan(\beta)$ being the slope.

SAGA GIS documentation contains description of SAGA Wetness Index:

"The 'SAGA Wetness Index' is, as the name says, similar to the 'Topographic Wetness Index' (TWI), but it is based on a modified catchment area calculation ('Modified Catchment Area'), which does not think of the flow as very thin film. As result it predicts for cells situated in valley floors with a small vertical distance to a channel a more realistic, higher potential soil moisture compared to the standard TWI calculation. "

The process of computing the SAGA Wetness Index for global land areas is very time consuming, even if a strong PC configuration is provided the computation takes several days. To achieve this, it is necessary to tile the DEM at continental level, re-project to equal area projection, compute 'SAGA Wetness Index' for each tile and build a mosaic for the global land mass (Figure 3.5). Majority of grid analyses and computation for this research was done in Sinusoidal projection. The result of distortion analysis and pixels omission process, in context of mapping global image data, suggests use of Sinusoidal projection in comparison to other equal area projections (Seong et al., 2002). The script

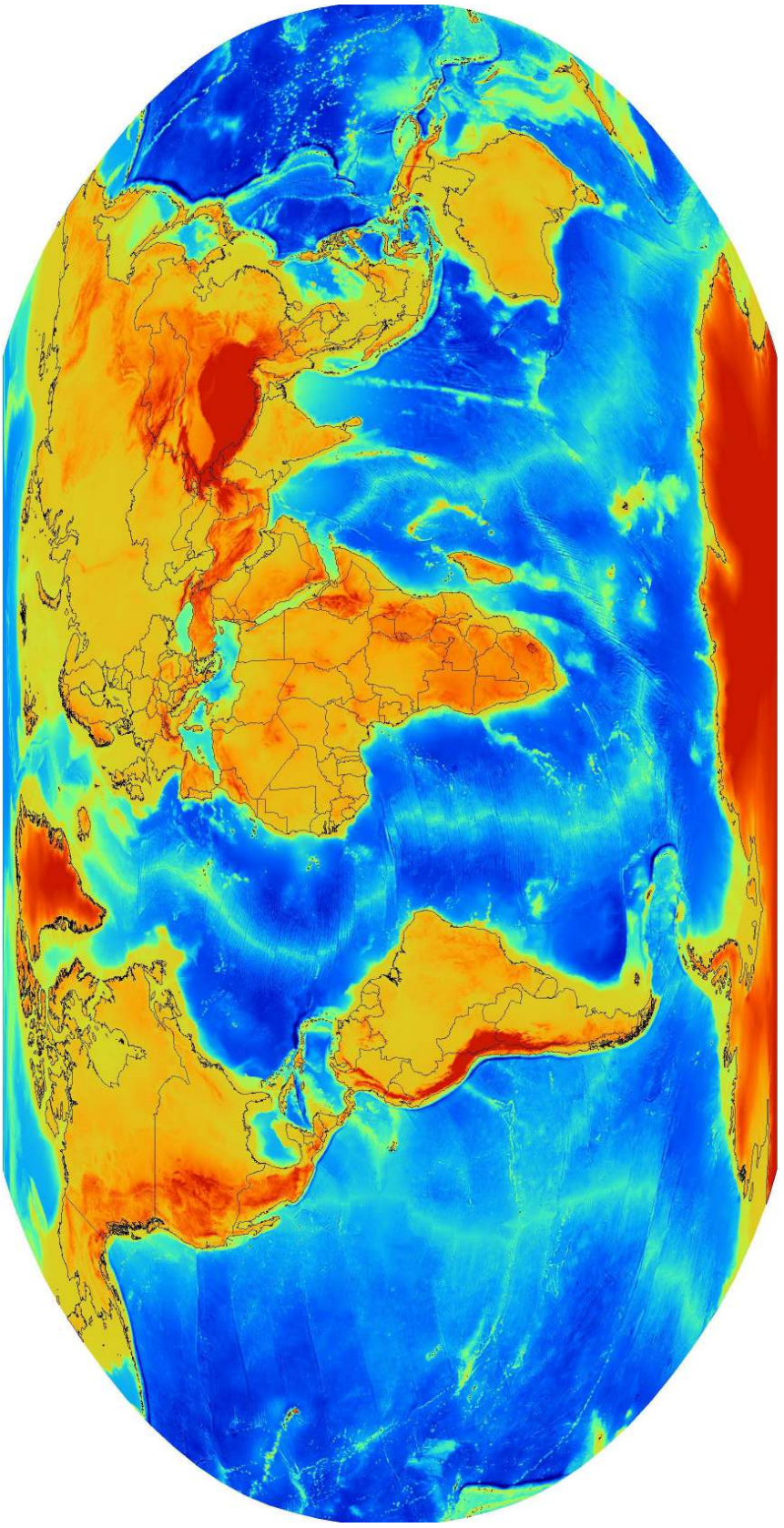

FIGURE 3.4: Global relief model (DEMSRE), full global covarage.

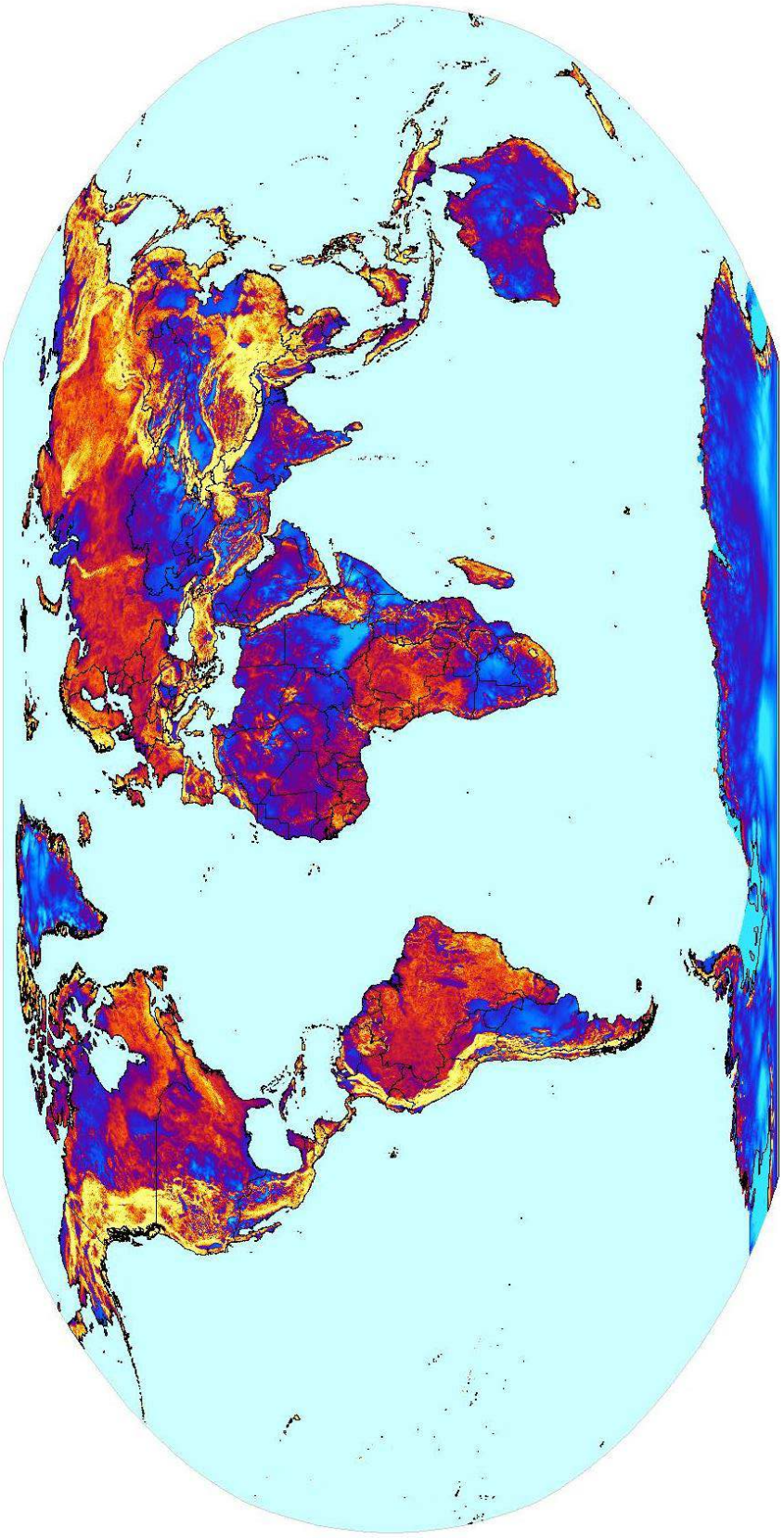

FIGURE 3.5: SAGA Wetness Index, global land mass covarage.

for layer production produced by Tomislav Hengl and Milan Kilibarda is available on the WorldGrids.org.

3.4.3 Potential incoming solar radiation derived in SAGA GIS (INMSRE)

Potential incoming solar radiation is topo-climatic parameter which depends on few factors, it is not just DEM derivative. Topo-climatology is the part of climatology which deals with impacts of land surface (i.e. topography) on climate (Boehner and Antonic, 2009).

Potential incoming solar variability depends on (Boehner and Antonic, 2009) :

"There are three major causes of spatial variability of radiation at the land surface: (1) orientation of the Earth relative to the sun, (2) clouds and other atmospheric inhomogeneities and (3) topography. The first cause influences latitudinal gradient and seasons. The second cause is associated with local weather and climate. The third cause such as spatial variability in elevation, slope, aspect and shadowing, can create very strong local gradients in solar radiation.

One of the first articles about GIS modelled solar radiation is written by Dubayah and Rich (1995) as initial proposal of using GIS for computing solar radiation, the formulas and illustration are provided in publications Hofierka and Suri (2002); Boehner and Antonic (2009).

Author calculated potential incoming solar radiation for 8-days intervals in SAGA GIS, thus computing is very time consuming and for global land areas takes more than 20 processing days. The processing steps were in general:

- *re-projection of initial DEM (DEMSRE) in equal area Sinusoidal projection,*
- *creation of tiles across land mask, around 500 tiles,*
- *calculation of potential incoming solar radiation for 8 day period,*
- *mosaicking results.*

Finally, result of this computation contains 46 images at 1 km resolution. The script for computing is provided on WorldGrids.org, but only available at the moment on the portal is annual average (Figure 3.6) and standard deviation grid at resolution 1 km, or smaller.

3.4.4 Distance from the sea coast line(DICSRE)

Distance from the sea coast line (`DICGSH`) was derived using the Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS) (Wessel and Smith, 1996). The coastline vector map was rasterized to the Robinson projection with three central meridians at -120, 0 and 120 degrees, and then for each tile metric distance from coast line has been derived, and then merged to create a complete and consistent distance to the coast line map in kilometers.

3.5 Methods

After identifying major publicly available sets and following the preliminary inspection of their temporal and geographical coverage and data formats, three clean and consistent data sets have been merged for purpose of testing: GSOD and ECA&D (station data) and time-series of MODIS Land Surface Temperature images. We run a number of standard spatial analysis operations in the R environment for statistical computing (R Development Core Team, 2012) to determine possible problems with using this data for spatio-temporal interpolation and time-series analysis.

We focused on three important aspects of data quality:

(*a*) *representation of station data in geographical space and temporal coverage* — assessed using kernel density analysis;

(*b*) *representation of station data in feature space* — assessed using the MaxEnt analysis;

(*c*) *suitability of data for spatio-temporal interpolation* — assessed using cross-validation of spatio-temporal prediction models;

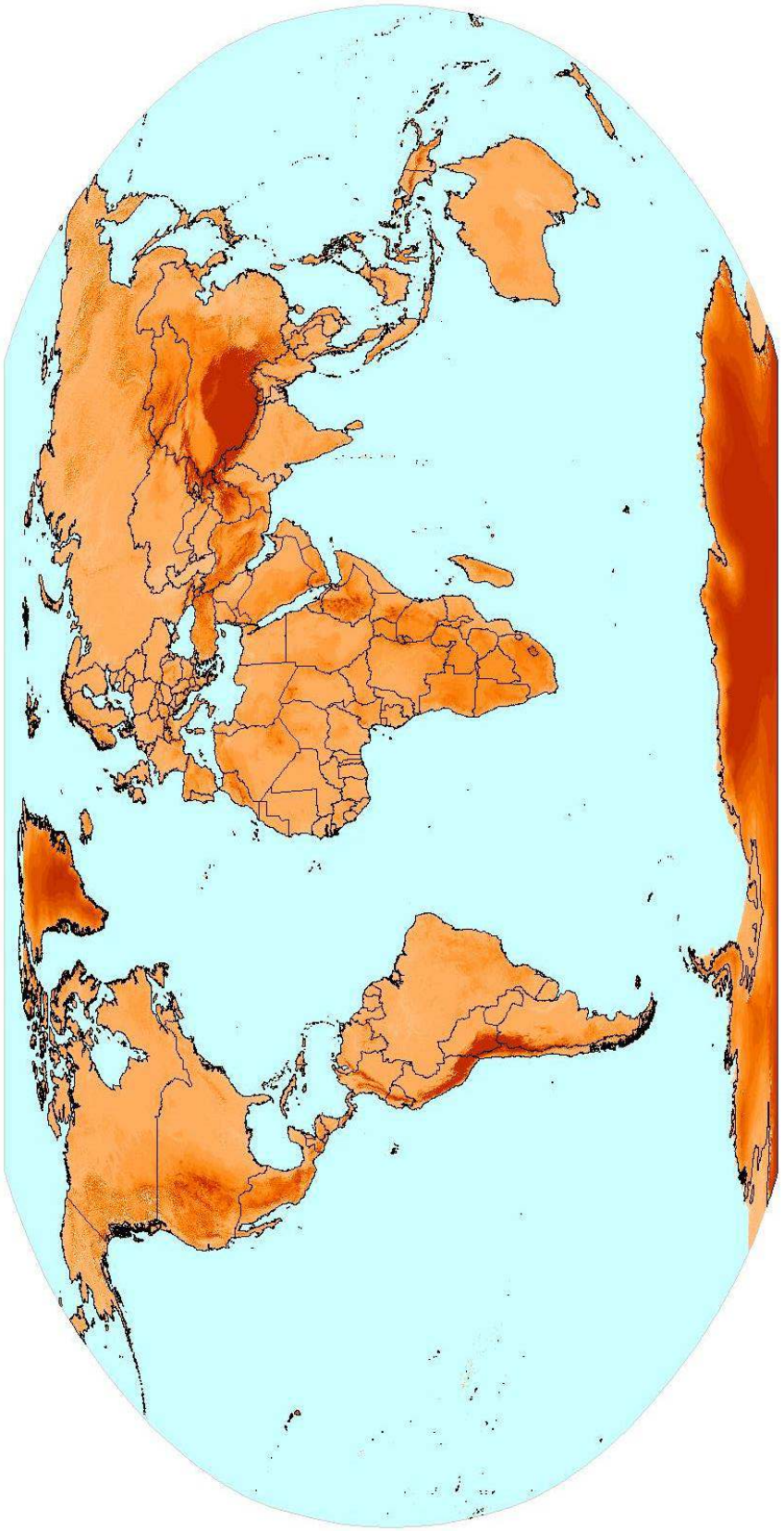

FIGURE 3.6: Potential incoming solar radiation derived in SAGA GIS (INMSRE), annual average, global land mass covarage.

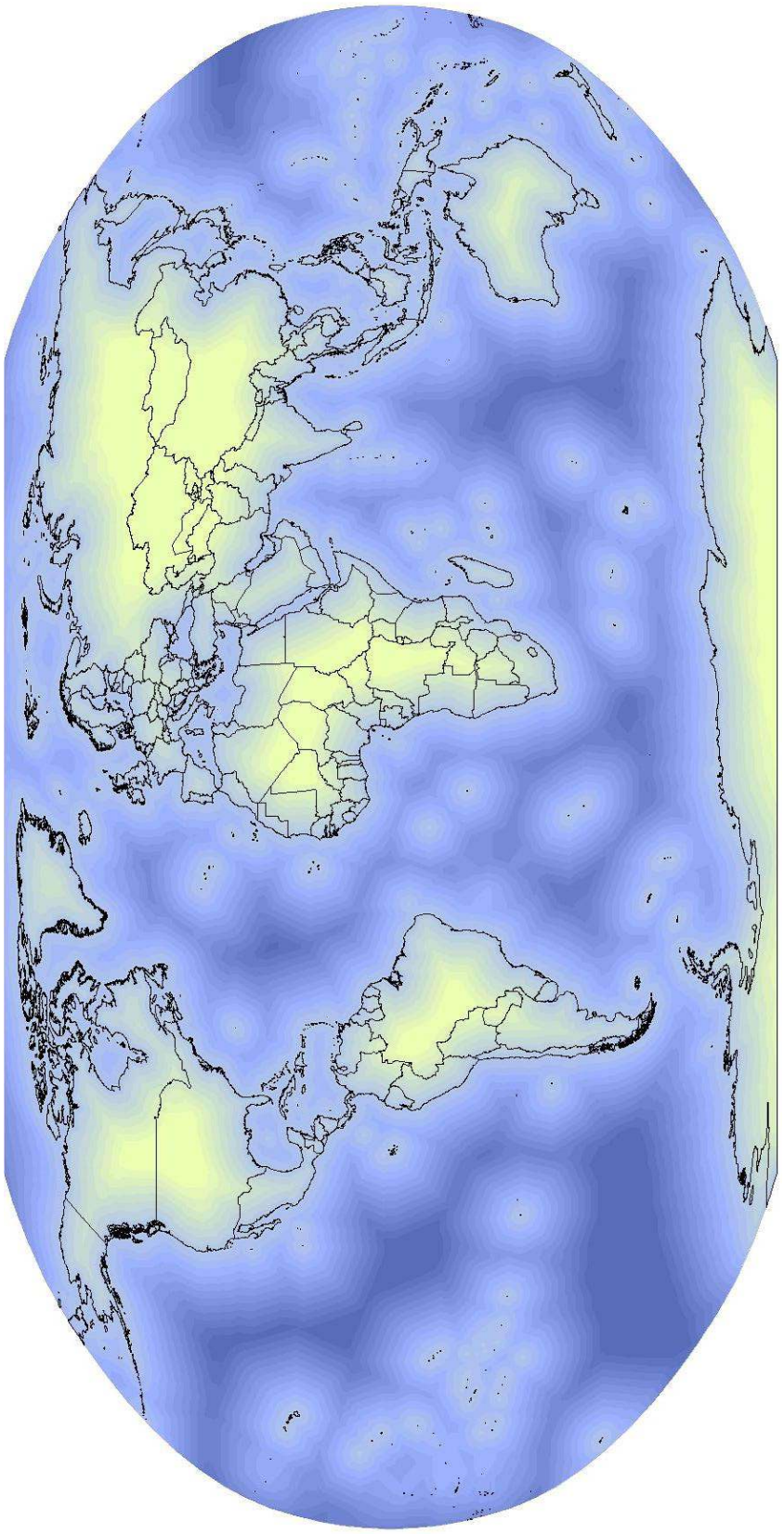

FIGURE 3.7: Distance from the sea coast line(DICSRE).

3.5.1 Representation of station data in geographical space and temporal coverage

To assess temporal coverage (*(a)*) we plot change in number of stations and then try to explain differences and trends. To assess geographical representation we produced kernel density maps using the *Gaussian* smoothing kernel as implemented in the spatstat package for R (Baddeley and Turner, 2006). Kernel density estimation of point process intensity requires an optimal bandwidth selection. To derive the optimal bandwidth we used algorithm based on cross validation as described in Diggle (2003, pp.115–118). The obtained bandwidth was 70 km for the temperature and 40 km for the precipitation stations. To provide comparable maps of density for two meteorological variables, we further derived and plotted only the relative density maps with values in the range $[0, 1]$.

3.5.2 Representation of station data in feature space

To asses the feature space representation i.e. the sampling bias (*(b)*), we use the MaxEnt analysis (Phillips et al., 2006). MaxEnt uses maximum entropy model, purposely developed to explain distribution of animal or plant species as a result of environmental conditions (Phillips et al., 2006), to derive a probability of occurrence of point patterns. Moreover, this technique could be successfully used to explain distribution of people in certain areas as well as to assess spatio-temporal dynamics of human populations as a function of environmental predictors (Bajat et al., 2011). MaxEnt is consider to be one of the most robust methods to assess the feature space coverage (Elith et al., 2011). In this paper we use it to assess the sampling preference i.e. feature space representations of meteorological station network. As *'occurrences'* we used records of mean temperature for year 2011 and precipitation for the same year respectively (GSOD and ECA data sets). MaxEnt generate the receiver operating curve (ROC) and the area under the curve (AUC) which allows for evaluation of the model performance between occurrence of climate locations and their absence. The area under the curve (AUC) values can be interpreted as indicating the probability that, when a presence site and an absence site are drawn at random from the population, the first will have a higher predicted value than the second.

Models with an AUC value above 0.75 are considered as *significant* (for more details refer to Elith et al. (2006)).

As environmental predictors for the MaxEnt analysis we used the following five publicly available environmental layers:

- DICGSH — distance from the sea coast line,
- GLCESA — land cover classes based on the MERIS FR images (factor-type variable),
- DEMSRE — global Relief Model based on SRTM 30+ and ETOPO DEM,
- PDMGPW — gridded Population of the World, version 3 (GPWv3),
- GACGEM — world accessibility map,

Distance from the sea coast line (DICGSH) was derived using the Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS) (Wessel and Smith, 1996). The coastline vector map was rasterized to the Robinson projection with three central meridians at -120, 0 and 120 degrees, and then for each tile metric distance from coast line has been derived, and then merged to create a complete and consistent distance to the coast line map in kilometers.

The environmental grids listed above are available for download via the WorldGrids.org (http://www.worldgrids.org/doku.php) data repository at various resolutions from 1 to 20 km. In this case study we used the grids with spatial resolution of 0.05 degrees i.e. 5 km when projected in the Robinson projection system (http://spatialreference.org/ref/esri/54030/).

3.5.3 Suitability of data for spatio-temporal interpolation

For assessing usability of point and RS data (8-day MODIS images) for spatio-temporal modeling (*(c)*), we first fit a spatio-temporal regression-kriging model for mean daily temperatures, then run cross-validation to detect possible outliers and critical areas. Within the regression-kriging framework, the regression and residual kriging parts are dealt with

separately, the regression part is used as trend surface and residuals are fit with a global space-time sum-metric variogram model. The interpolation surface is the sum of trend surface (regression part) and the residuals surface (spatio-temporal kriging prediction).

As auxiliary predictor to model the spatio-temporal trend we used a time series of 8-day MODIS day-time LST images. Original MODIS LST were converted to degrees Celsius from Kelvins. Missing pixels in the original LST maps were replaced using SAGA GIS function *"Close Gaps"*. Close Gaps function uses splines as robust method for filling the gaps in areas with sparse or irregularly spaced data points (Neteler, 2010), the missing pixels filtering from 8-day MODIS mosaics were performed because the predictor need to be known over the domain of interpolation, i.e. over the land mass area. The temporal dissaggregation from 8-day images to daily images is done by applying spline in temporal domain, by splining 8-day MODIS LST pixels to provide daily values. Thus, we provided daily values of predictor for global land mass area at 1 km spatial resolution. The linear regression for 2011 year were applied on the MODIS spline images. Figure 3.8 shows the plot of observation against the 8-day MODIS spline values.

The spatio-temporal regression-kriging method is described in detail in Hengl et al. (2012) and Heuvelink et al. (2012), they used sum-metric variogram structure for temperature modeling, as we in this paper. The sum-metric space-time variogram were fitted via the space-time kriging framework from `gstat` package (Pebesma, 2004), the package that is also capable of working with space-time class data (Pebesma, 2012). The sum-metric covariance structure assumes that the component of variation can be explained by pure spatial variability (space variogram), the second part with pure time component (temporal variogram) and the third part is explained by joint variogram (spatio-temporal interaction), where the time component is multiplied by geometric anysotropy coefficient. Finally, fitted variogram surface is described by 10 parameters. The prepared spatio-temporal point data set contained around 3 billion measurements of daily average temperature from 8713 stations for the selected year 2011.

The cross-validation technique was performed by using *"leave-one-station-out"* procedure. This takes neighborhood stations to predict the all values in temporal domain (365 days in this case for each day in 2011) without using any observation from cross-validated station. Cross-validation predictions are than compared with actual observed values to

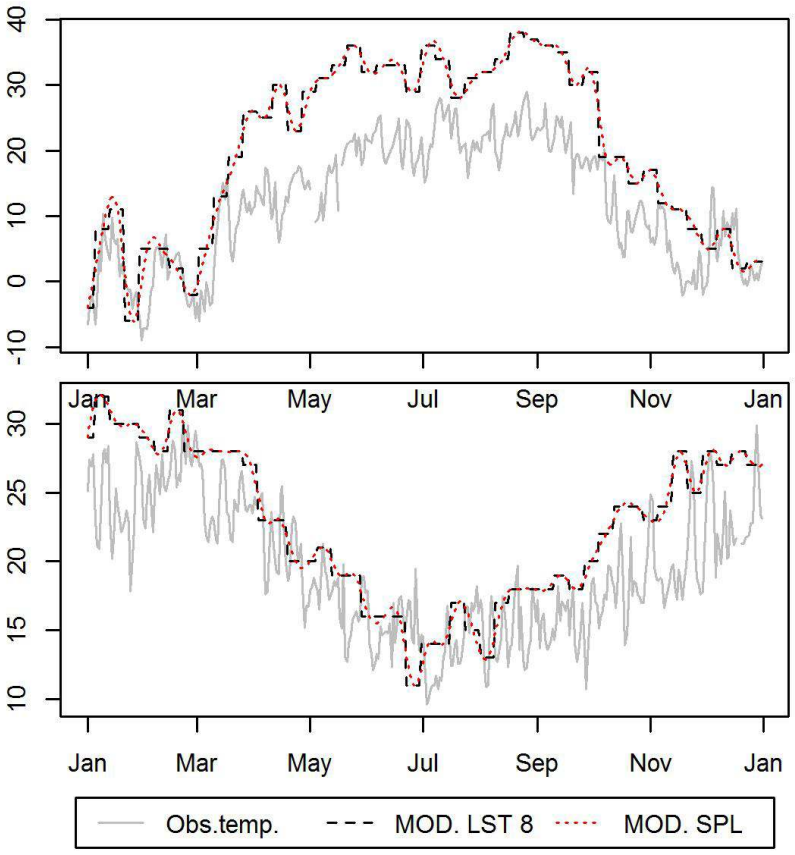

FIGURE 3.8: The station Novi Sad, Serbia ($\lambda = 19.850, \phi = 45.333$) (above), gray solid line: mean daily temperature observation in 2011; black dashed line: 8-day MODIS LST values; red dashed line: MODIS LST spline (red); Station Swanbourne, Western Australia ($\lambda = 115.767, \phi = -31.950$) (below).

derive the root mean square error (RSME). The RMSE was derived for station $\times$ day (Figure 3.19) and for each day of the year globally (Figure 3.17).

Before model fitting, station data needed to be cleaned-up for duplicates and inconsistencies. All assumed duplicate stations (closer than 2 km) were removed from the analysis. Also, stations that have obvious artifacts (usually not visible in the original data) were iteratively removed by comparing the cross-validation versus the observed values. For example, 27 GSOD stations with large difference between cross validation and actual observed values for Canada are shown in Figure 3.9. These are assumed to be gross errors and were removed from final analysis.

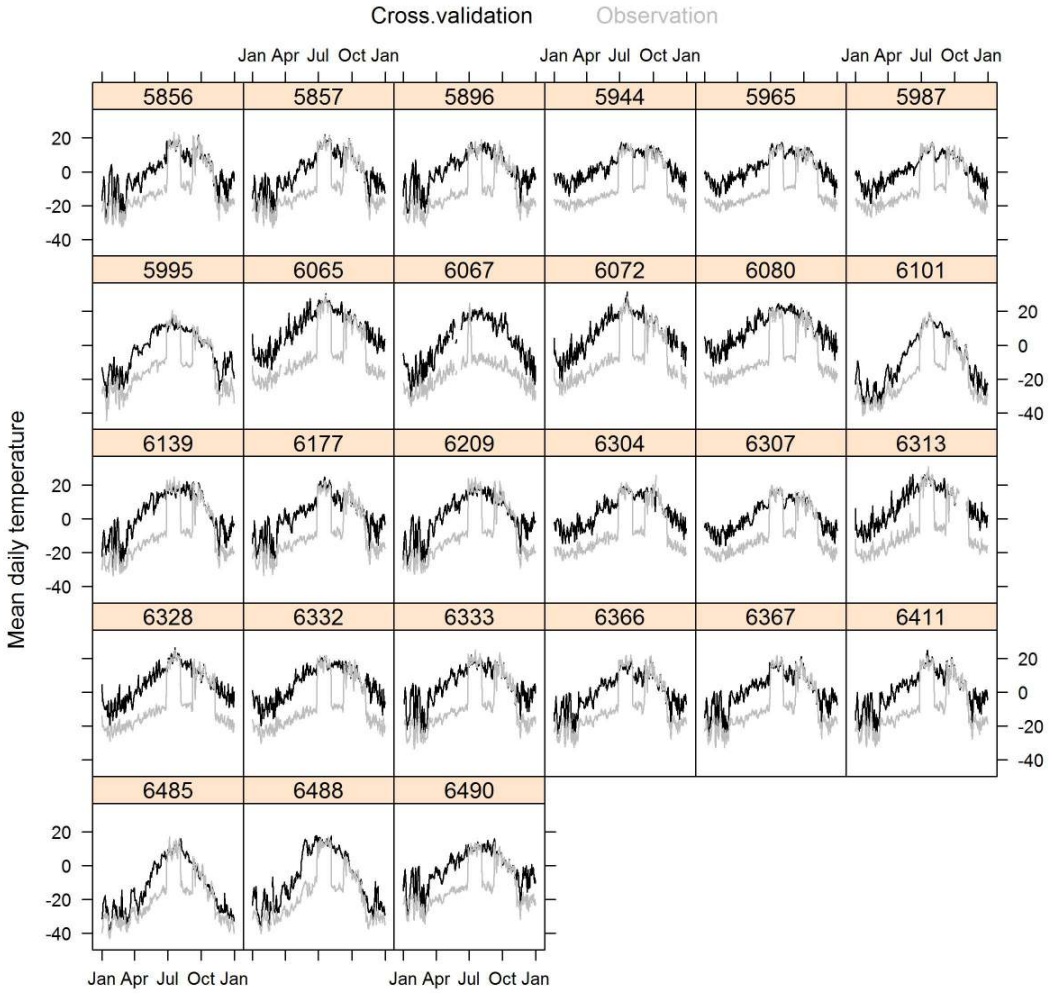

FIGURE 3.9: Outliers and inconsistencies detected for station data from Canada. The observed mean daily temperature (grey) and cross-validation prediction of temperature (black line) in °C. Heading numbers refer to internal identifier of stations.

3.6 Results

3.6.1 Temporal coverage

Change in annual number of available stations is illustrated separately for daily (Figure 3.10, upper) and monthly (Figure 3.10, lower) data sets. As plots indicate, there are several major differences in these sets. ECA&D data set comprises some historical data back to 1800s while GSOD starts in 1890s. Maximum number of stations in GSOD is

more than three times larger than in ECA&D due to GSOD's global geographical coverages compared to ECA&D's European one. ECA&D number of station reached steady plateau from 1960s through 1990s when number of station started to drop down. GSOD number of stations per year had strong increase from 1940s to 2000 but then the number also declines.

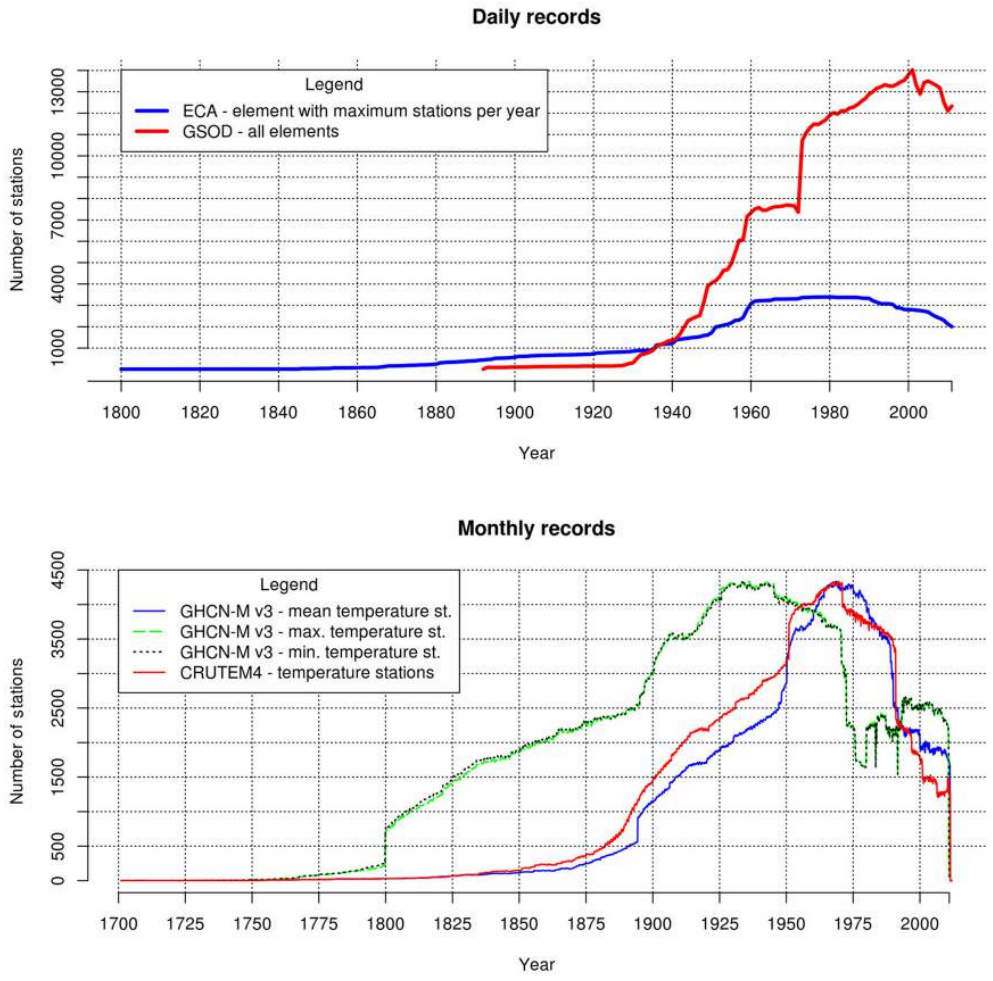

FIGURE 3.10: Number of stations per year with daily records in ECA&D (European Climate Assessment & Dataset) and GSOD (Global Surface Summary of Day) data sets (above). Number of stations per year with monthly records in GHCN-M V3 (Global Historical Climatology Network-Monthly) and CRUTEM4 (Climatic Research Unit land stations) (below).

The reasons for decline in number of stations could be many. The development of computers and telecommunications are mainly responsible for the rise in number of stations

while reasons for the reduction in number are harder to explain. Ground measurements are expensive to maintain so at one moment it was suspected that automatic meteorological stations and remote sensed data (radars and satellites) could replace ground measurements. This technological shift is still on-going, so that the plots show a slight drop as there are less records for the last decade. Also the number of stations available in all global data sets is dependent on national data exchange policies, so sometimes just part of the national data are available. This is also subjected to change. For example ECA&D has been expanded in 2013 with Finnish and a number of Russian stations.

The number of stations with monthly temperature records per year are provided for GHCN-M v3 and CRUTEM4 data-sets. For selected meteorological element these numbers for global area (Figure 3.10, lower) are comparable with number of stations in European ECA&D data set even though the overall number of stations in GHCN is larger. This reflects in lower average global spatial density of GHCN set compared to European ECA&D (except in eastern US and some other smaller regions).

3.6.2 Geographical coverage

Geographic coverage analysed with Gaussian smoothing kernel is shown in Figure 3.11. The regions with average and higher than average station density values are coloured orange to red (Europe, North America, South and East Asia, coastal part of South America, Africa and Australia). Regions with lowest relative density and mostly zero density pixels are coloured in white (< 0.1 or 10%). The later are the areas where spatio-temporal prediction models would possibly result in data extrapolation.

The results of the neighborhood analysis show that for 25% of global land areas the distance to nearest station is >200 km. Further analysis of station frequency counts for 100 km×100 km blocks (about one decimal degree cells) shows that 72% of land areas contains no stations, 15.5% of land area has only one station per 10,000 km^2, while only 12.5% of land areas is covered with more then one station per 10,000 km^2.

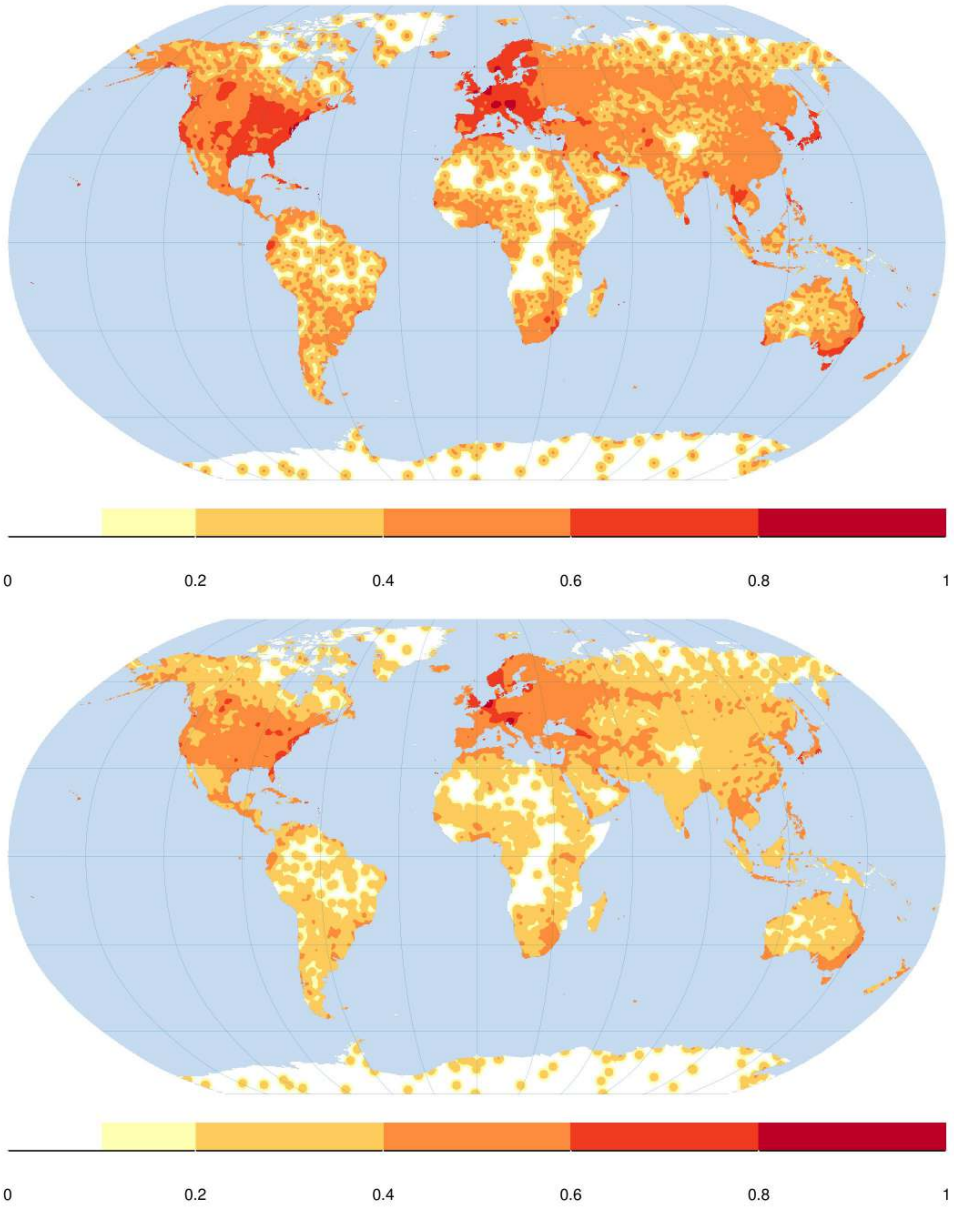

FIGURE 3.11: Relative density of stations for 2011: (above) estimated for the ECA&D (European Climate Assessment & Dataset) and GSOD (Global Surface Summary of Day) daily temperature data set, (below) estimated for the GSOD and ECA&D daily precipitation data set. Bandwidth used to derive kernel density: H=70 km.

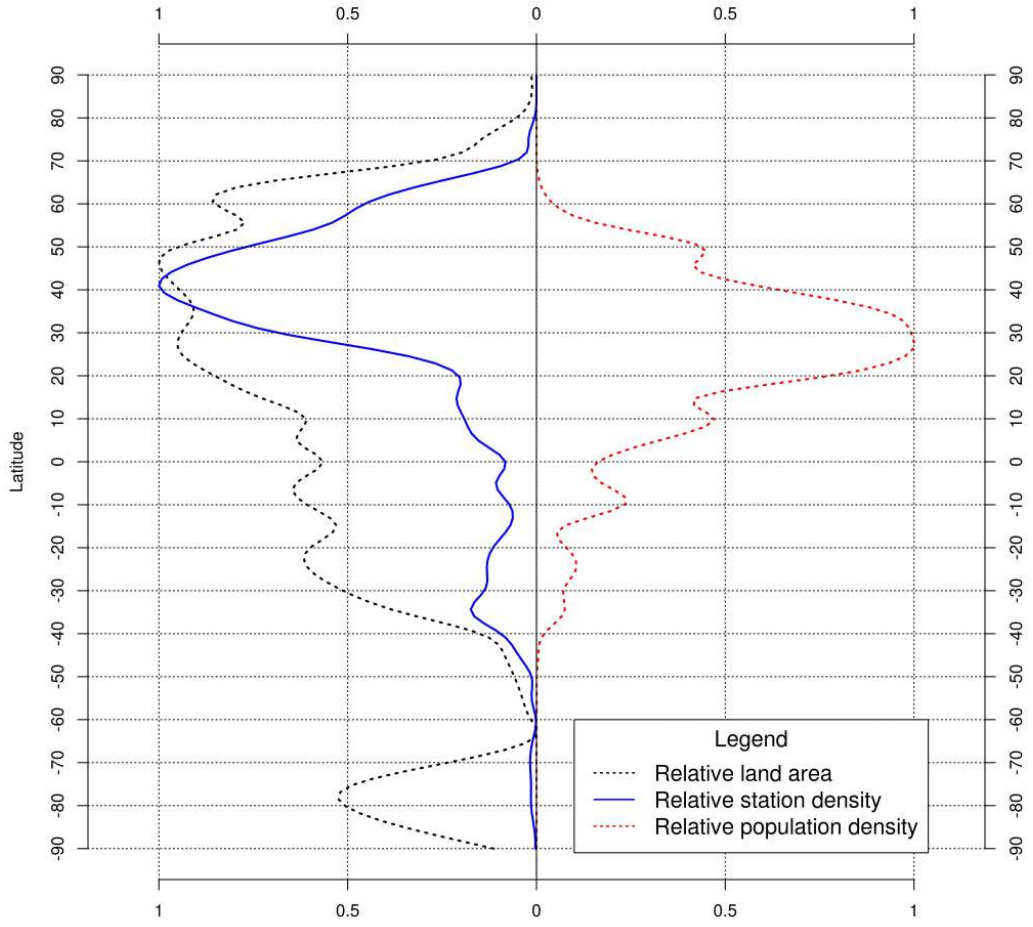

FIGURE 3.12: Relative station density compared to relative population density and land areas arrangement depending on latitude. Density values are in the range $[0, 1]$ for the all showed elements.

3.6.3 Feature space coverage

The results of cross-validation using 75% of records for training and 25% of records for validation in MaxEnt show that distribution of GSOD and ECA sets are fairly controlled by selected environmental layers (AUC$=0.84$). This means that distribution of station locations could be explained with these covariates with an accuracy of 84%. The current distribution of stations locations is strongly controlled by all continuous predictor

layers (DICGSH, DEMSRE, PDMGPW, GACGEM). Even if each of this predictors is used separately, they explain about 80% of distribution individually. Obviously, DICGSH, DEMSRE, PDMGPW, GACGEM are highly cross-correlated, and highly correlated with the density of meteorological stations. The environmental layer that decreases the gain the most when omitted from the prediction is GLCESA (land cover map), which therefore appears to have the most information that is not present in the other layers.

Figure 3.13 indicates that there is indeed a clear sampling preference for meteorological monitoring. As expected, most of Europe is densely covered and represented with meteorological stations (except Alps region), so are the South and East Asia and coasts of North and South America. On the other hand, the areas of Central Asia (Himalaya Range), Andes Range and Amazon Basin in South America, Central Australia, north coasts of Asia and North America, Central Australia and Sahara desert in Africa, as well as all of Greenland, Antarctic and Arctic poles suffer from a significant misrepresentation in the feature space.

The station clustering is also illustrated using histogram plots in Figure 3.14. This shows that there is a higher grouping of stations along the coast lines. Note also in Figure 3.14 (upper) that some stations are located further from the coast line, which is not an artifact, but the consequence of the resolution of the covariate layers (5 km). Many meteorological stations are located on small islands (with area $<$25 square kilometer) not mapped in the DICGSH map. From Figure 3.14 (lower) it is also obvious that large mountain chains are globally under-represented in the meteorological networks.

3.6.4 Spatio-temporal models for temperature

The results of regression modeling show that time series of 8-day MODIS LST images explain 81% of the variability in mean daily temperature values for the year 2011. MODIS LST images are significant estimators of the daily temperatures but with somewhat lower precision than indicated by e.g. Wan et al. (2004): the correlation plot in Figure 3.15 indicates an average precision of $\pm$5.3°C. Note also that we use the 8-day averages of MODIS LST images, hence somewhat higher error can be expected.

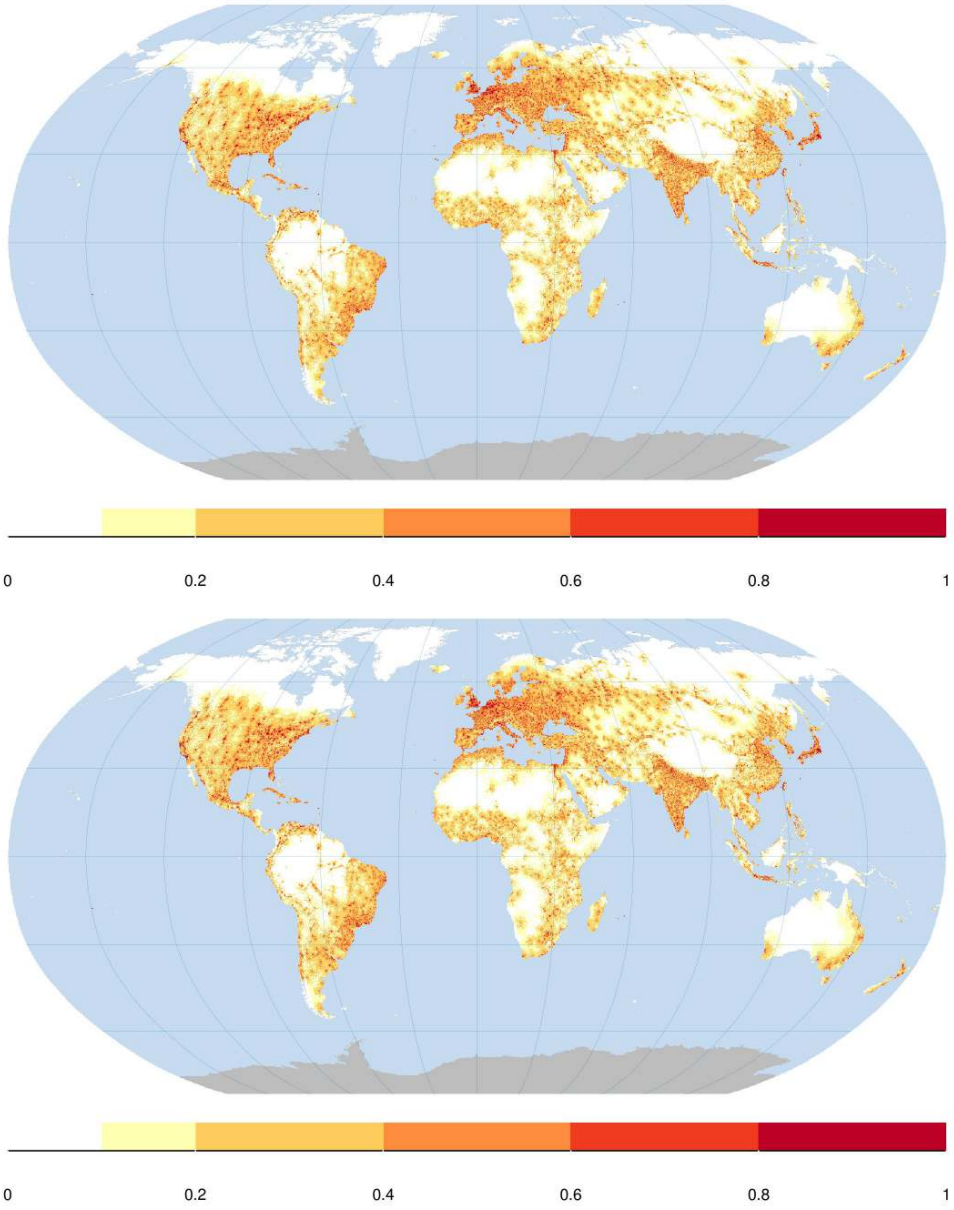

FIGURE 3.13: Sampling bias in feature space derived using the MaxEnt software and standard covariates (distance from the sea coast line, land cover classes, elevation map, population map, world accessibility map): (above) probability of station occurrence derived for observed temperature data sets (European Climate Assessment & Dataset and Global Surface Summary of Day; ECA&D and GSOD), (below) probability of station occurrence derived for observed precipitation data sets (ECA&D and GSOD). White colored areas indicate extrapolation areas. Spatial resolution of the maps is 5 km.

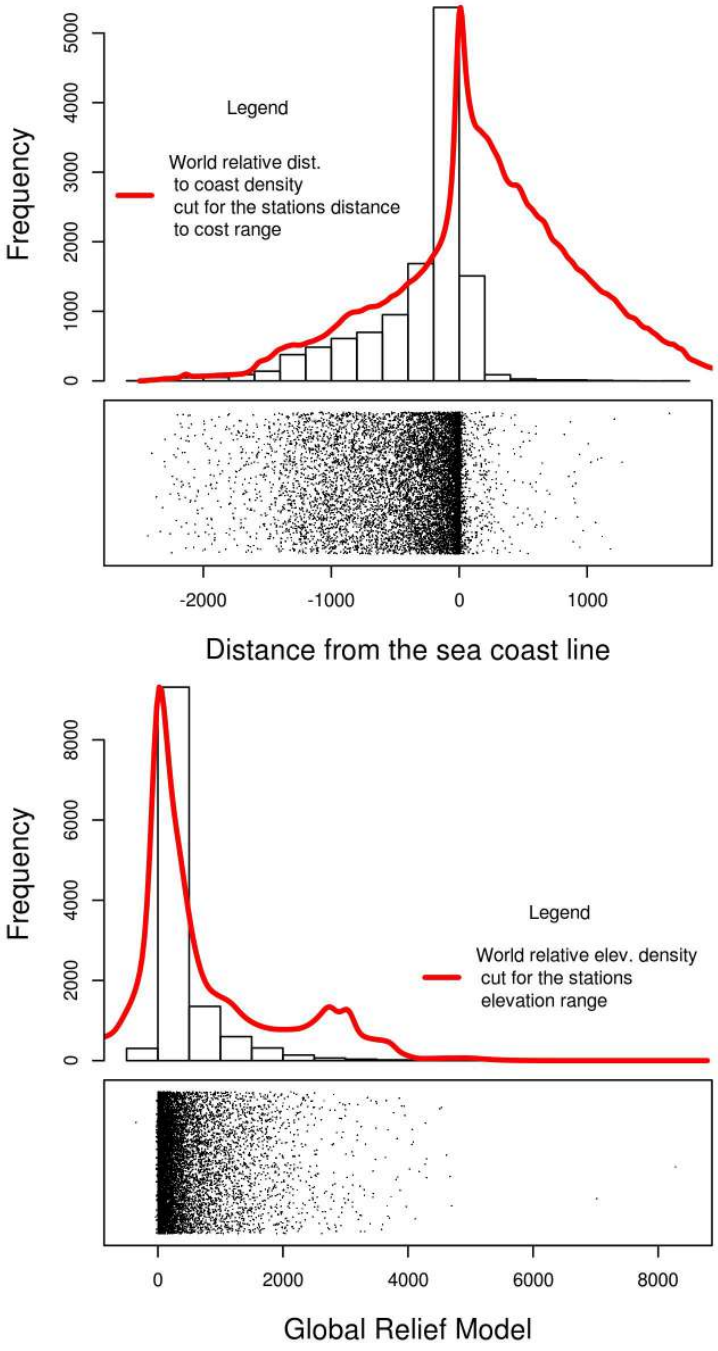

FIGURE 3.14: Station clustering for observed temperature data sets (European Climate Assessment & Dataset and Global Surface Summary of Day) visualized in feature space (distance to the coast line and elevation). The histograms were derived by overlaying stations and environmental layers. The points below the two histograms show actual meteorological stations. The red lines shows global relative density distribution of distance from coast line and elevation.

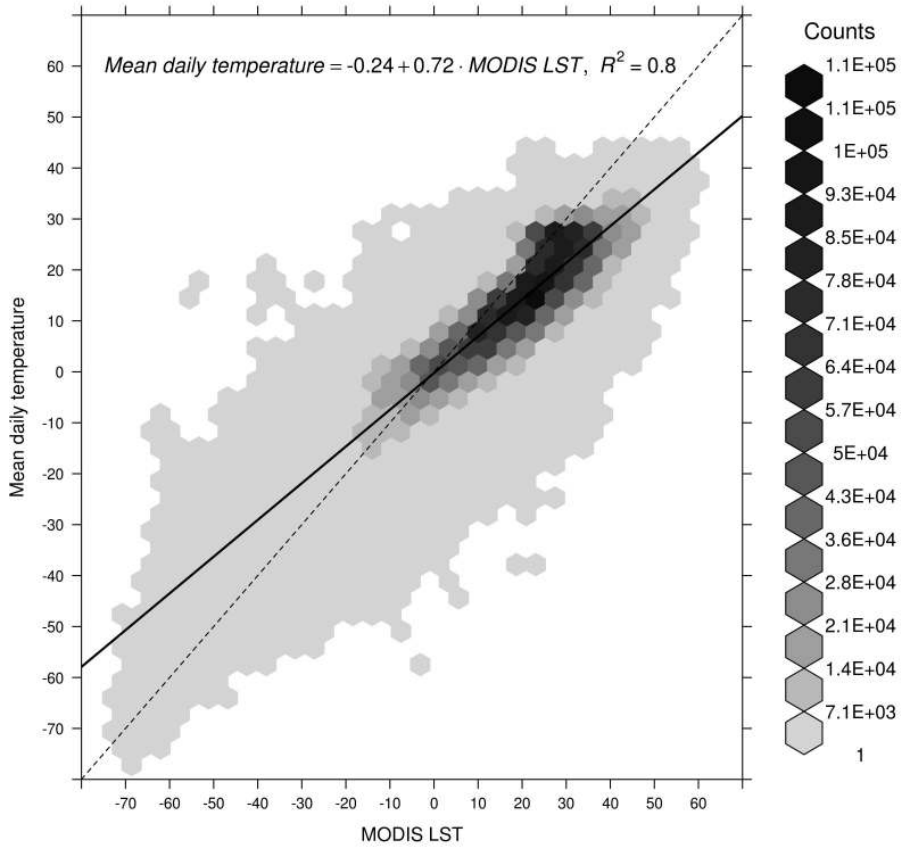

FIGURE 3.15: Scatter plot showing the general relationship between daily temperature and 8-day MODIS LST images. The fitted regression line and the 1:1 line (dotted).

Figure 3.16 shows the space-time variogram for residuals fitted using automated fitting in `gstat`. The pure spatial component of the space-time variogram, showing the residual correlation at any time separation between two spatial point exists, pure spatial spherical variogram has the nugget of 3.22°C^2, the sill of the 18.28°C^2 is reached with the range more than 6000 km. Thus, if it is known just spatial distance between two observations without knowing time separation the part of total covariance is known. The pure time component of variogram structure is zero, implying that if it is just known time separation between two points even part of total covariance is not known, implying that all temporal variability is explained by space-time interaction components. The joint variogram, representing space-time interaction is modeled by spherical function with the parameters: nugget 1.8°C^2, the sill 8.35°C^2 and range 2349 km. The geometric anisotropy parameter

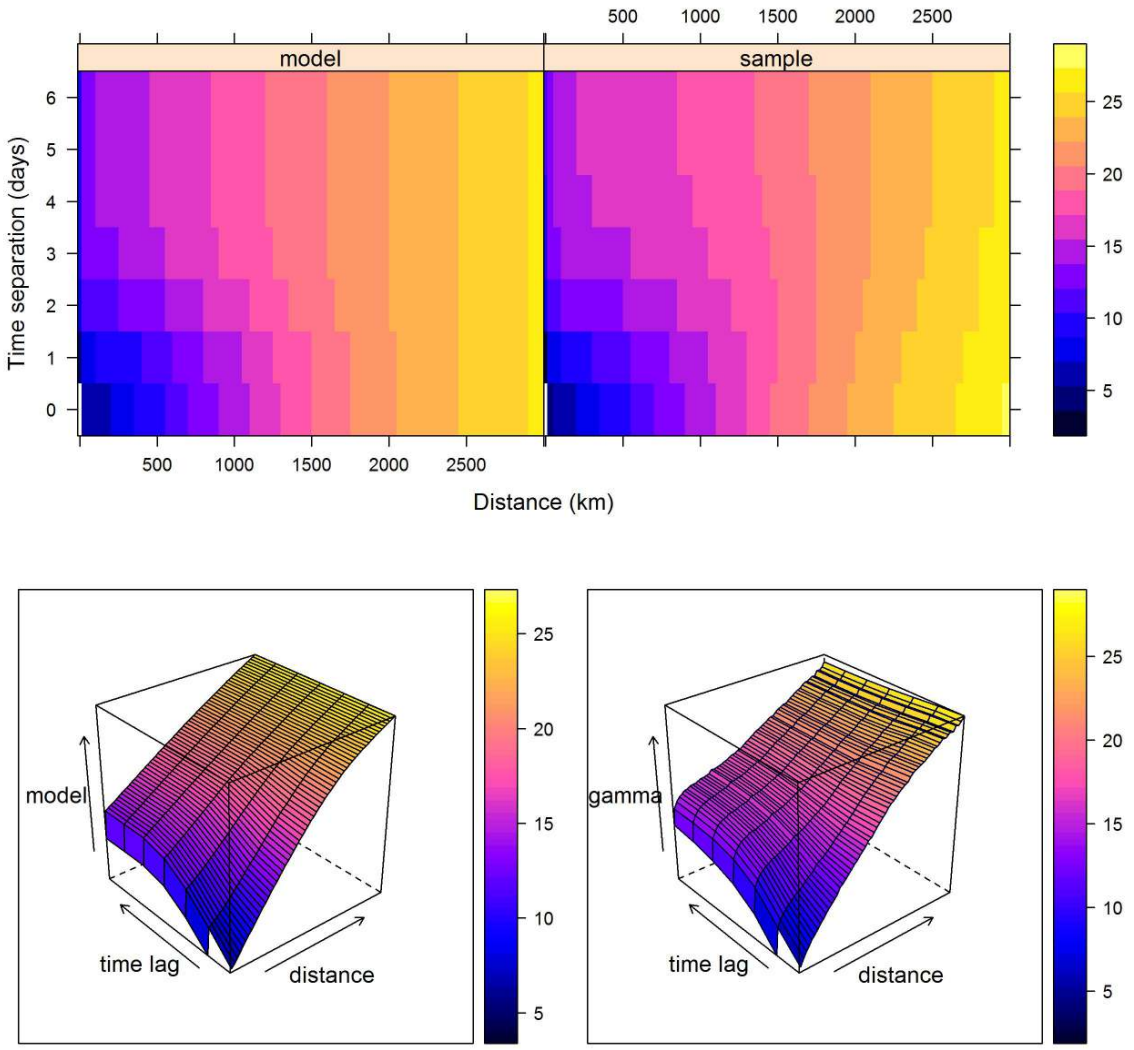

FIGURE 3.16: Fitted sum metric model (left) and sample variogriom (right) of linear regression residuals of mean daily temperature observation on 8-day MODIS spline images. The variogram surface is presented in 2D (above) and 3D manner (below).

is 583 km/day, this allowing to scale time separation in spatial distance for purpose of calculating space-time joint components of covariance. Therefore, space-time correlation of mean daily temperature residuals on MODIS LST spline images modeled with sum-metric covariance structure contains 7 parameters describing correlation at any space-time distances.

The results of cross-validation confirm that the spatio-temporal prediction model of mean

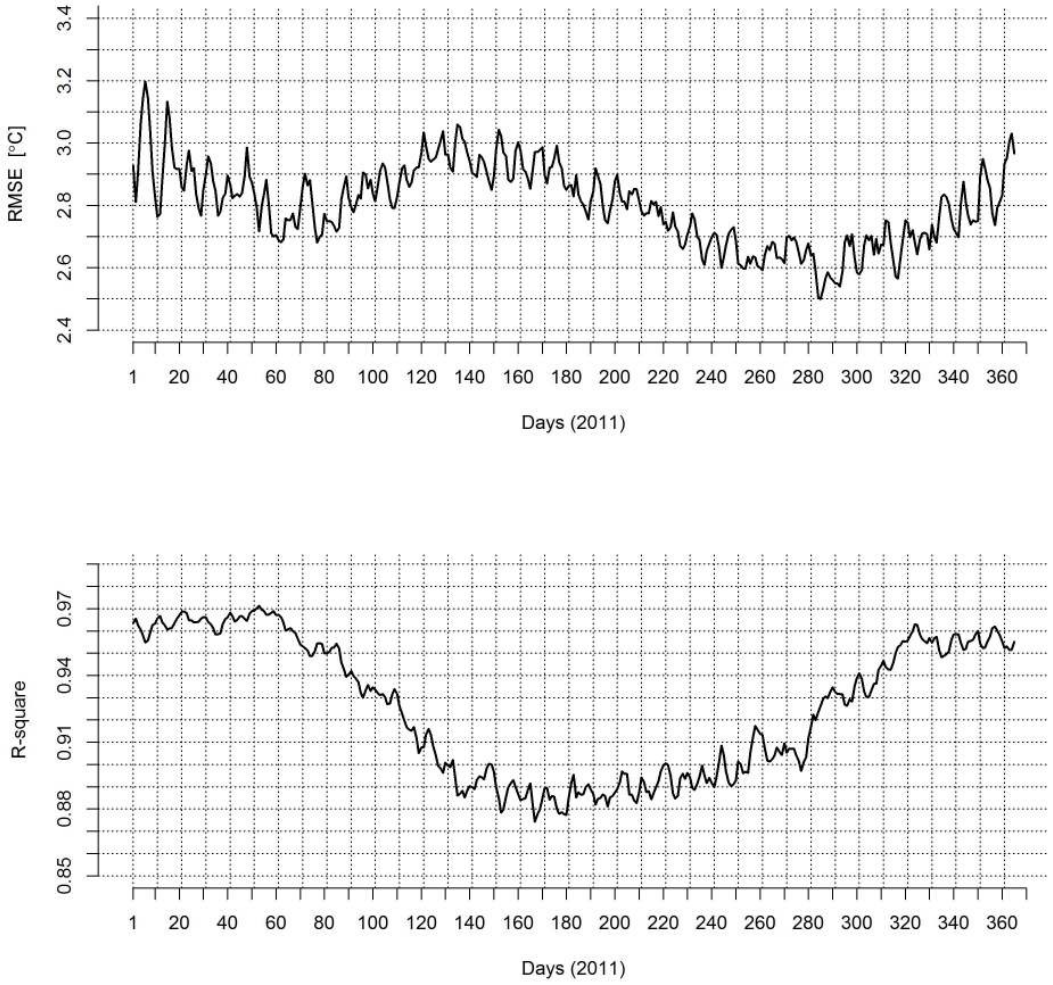

FIGURE 3.17: Daily temporal variation for RMSE and $R-square$ for year 2011.

daily air temperature can explain between 87.4–97.1% of variability in the observed values, with an average of 92.5% (see $R-square$ in Figure 3.17). Figure 3.17 further shows that average root mean square error per day (RMSE) varies from 2.5°C to 3.2°C with overall average 2.8.

Annual changes of RMSE and $R-square$ with the largest errors during winter on southern hemisphere when the global range of minimum and maximum mean daily temperature on the global land mask is the highest. RMSE calculated for latitude bands shown in

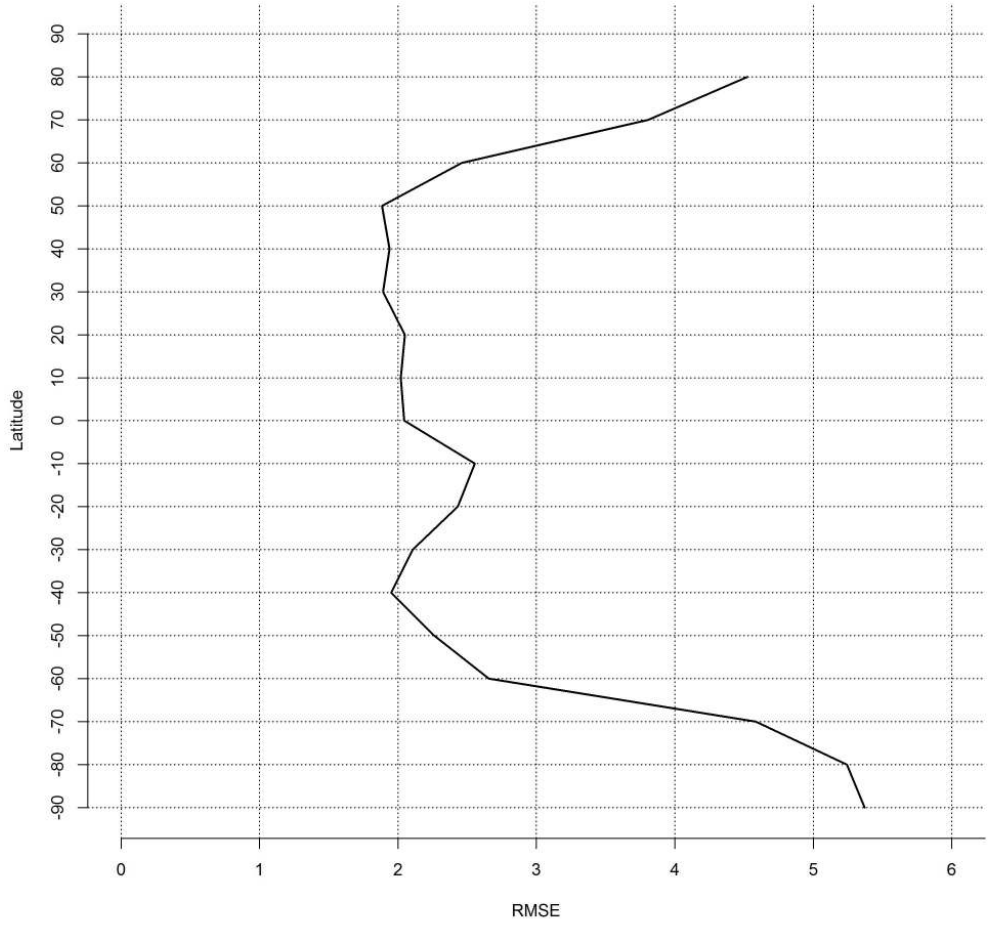

FIGURE 3.18: Spatial variation of (RMSE) for different latitudes (aggregated per 1 degree).

Figure 3.18 illustrates that $[-70 - -90]°$ latitudes have the highest RMSE, the same plot shows that for the latitude range $[0, 50]°$ global spatio-temporal model has the highest precision, in average less than 2°C, what is expected because the number of stations in this range is the highest (Figure 3.12).

Spatial distribution of RMSE calculated per year for each station (Figure 3.19) indicates that there are several regions where model is critically inaccurate. In the map shown in Figure 3.19, 47 stations, 0.5% of total points involved in cross-validation have an RMSE above 6°C (with an average of 6.8°C), 10% of total points have an RMSE in range from 3

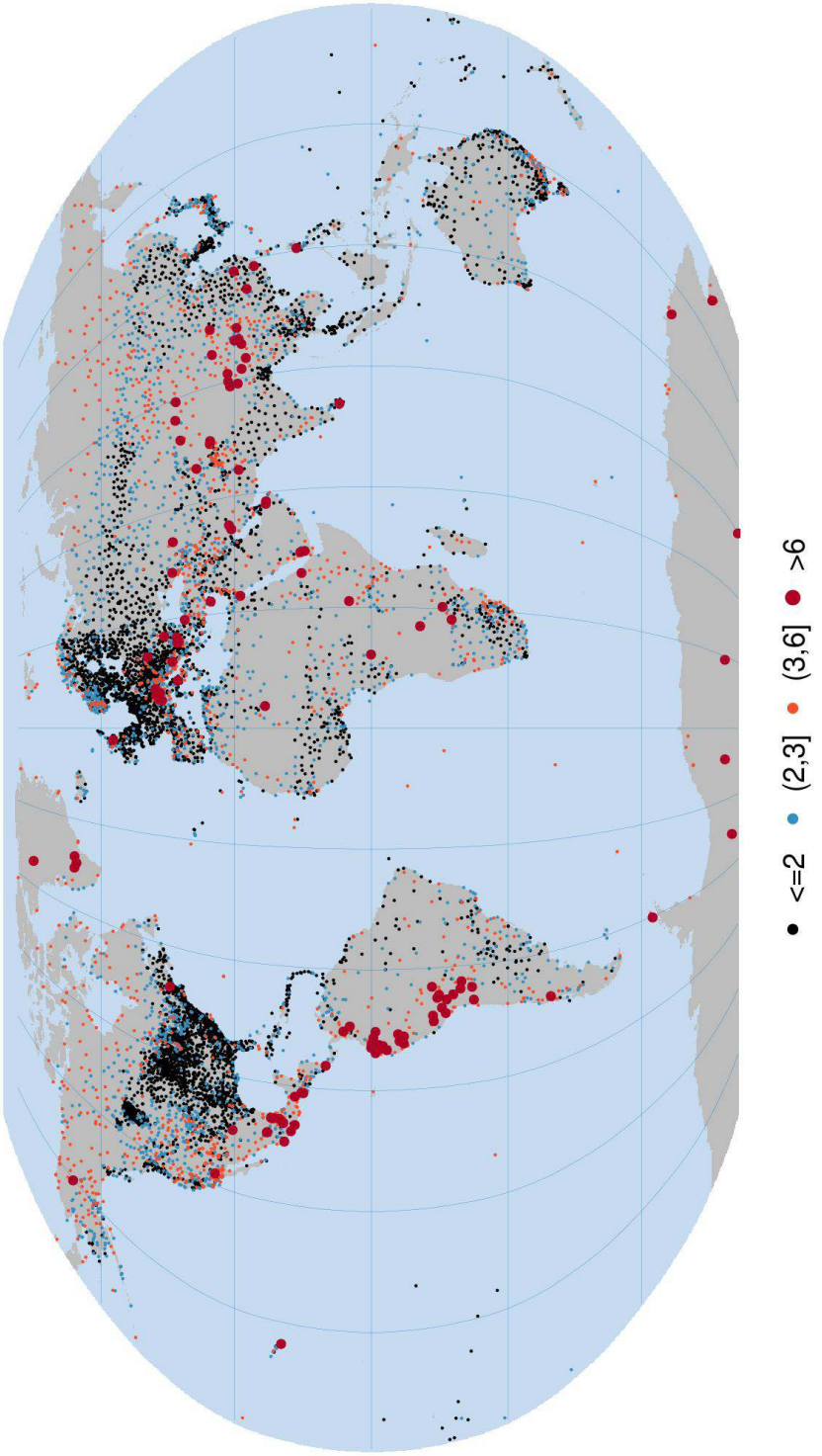

FIGURE 3.19: Map of the cross-validation errors (RMSE) averaged per year for each station. Red circles indicate cross-validation outliers estimated using a spatio-temporal regression-kriging model. Clusters of red circles indicate problematic areas.

to 6°C (with an average of 3.8°C). The most of the stations with RMSE higher then 3°C are located in Antarctica, mountains regions and in the sparsely populated areas with low station density coverage. Black circles in Figure 3.19 are points with high accuracy i.e. RMSE <2 (63% of total points) and with an average of RMSE $=1.5$°C. Finally, average RMSE for the global land mass of 2.8°C, but for USA and Europe the accuracy of interpolation by using this data set and spatio-temporal regression kriging model is less than 2°C in average. The interactive web map of the stations with RMSE produced with R package plotGoogleMaps (Kilibarda and Bajat, 2012), with cross-validating values against observation is provide on the `http://dailymeteo.org/WorldMap-RMSE-STRK-MeanT-on-MODIS.htm`.

3.7 Discussion and conclusions

All the data required to produce WorldDailyMeteo images are basically ready to be used. The remaining issues are: how to fit global spatio-temporal models and run automated geostatistics, and how accurate can we expect to get? The preliminary results of this research indicate that, before the production of daily estimates of meteorological variables at 1 km resolution can become operational, a care needs to be taken to account for spatial, temporal and feature space clustering of the meteorological networks. The expected accuracy probable can be better if more static and dynamic covariate layers are included, but sparsely covered areas in geographical and feature space still would be areas with double lower accuracy.

Temporal coverage analysis of the publicly available meteorological station data sets indicates that the GSOD and ECA&D data sets together represent respectable source of meteorological data for spatial-temporal analysis at daily resolution (especially from year 1960), however the distribution through time and space is unequal. What worries especially is that there is a drop in the observation density in the last decade. Even more distinct clustering can be observed in geographical and feature spaces. Analysis of distribution of stations locations in feature space shows that the spatial distribution of the meteorological network is distinctly controlled by environmental factors with an AUC=0.84 (estimated using cross-validation), which means that most of the distribution of the stations can be explained by population density and accessibility. Figure 3.14 also clearly

shows that the stations frequency is the highest near coast lines and at the lower elevations. Thus, unequally station distribution in feature space coverage, proved by MaxEnt results, can lead to inaccurate prediction in higher elevation and sparse populated areas.

For mean daily temperature measurements at stations we have further fitted a spatio-temporal model using the 8-day MODIS LST time series of images and got an average accuracy of about 2–3°C when assessed using cross-validation (which confirms the results of some local studies by Hengl et al. (2012), Heuvelink et al. (2012) and Neteler (2010)). This is promising as it indicates that indeed highly accurate maps of daily temperatures could be produced at high spatial resolution using the global spatio-temporal models. Figure 3.19 also shows that the outliers are distinctly grouped in areas poorly covered with meteorological stations. The second group of outliers we observed were in the mountain regions i.e. areas frequently covered with snow. This corresponds to the work of Neteler (2010) who experienced similar difficulties of working with dynamic snow cover on mountain tops. All this indicates that the produced spatio-temporal models will have serious problems for all areas that have been under-sampled in geographical or feature space.

Figure 3.15 also indicates that temperatures from $[-10, 50]$°C range are much better correlated with the MODIS LST images and hence easier to map than temperatures within this range. Similar results were reported, for example, by Wan et al. (2004); Yoo et al. (2011). Lack of stations in polar areas and in large mountain chains remains probably the most serious problem for global meteorological mapping (as illustrated in Figures 3.19, 3.18). The current meteorological station networks (Figure 3.1, 3.3, 3.2) probably do not represent the true complexity of climate at higher elevations, deserts, tropical forests, that is in sparsely populated areas (25–50% of the land surface). Topography, arrangement and orientation of mountain ranges are the key climatic factors in land areas (Hartmann, 1994; Beniston, 2006), hence station networks should follow this complexity in order to allow for a truly global assessment. It is not realistic to expect that the density of the meteorological stations will change in the years to come, and that these geographical coverage gaps will be solved. What is more interesting to the research community is probably how to overcome these problems with using appropriate statistical methods and the higher precision remote sensing technologies.

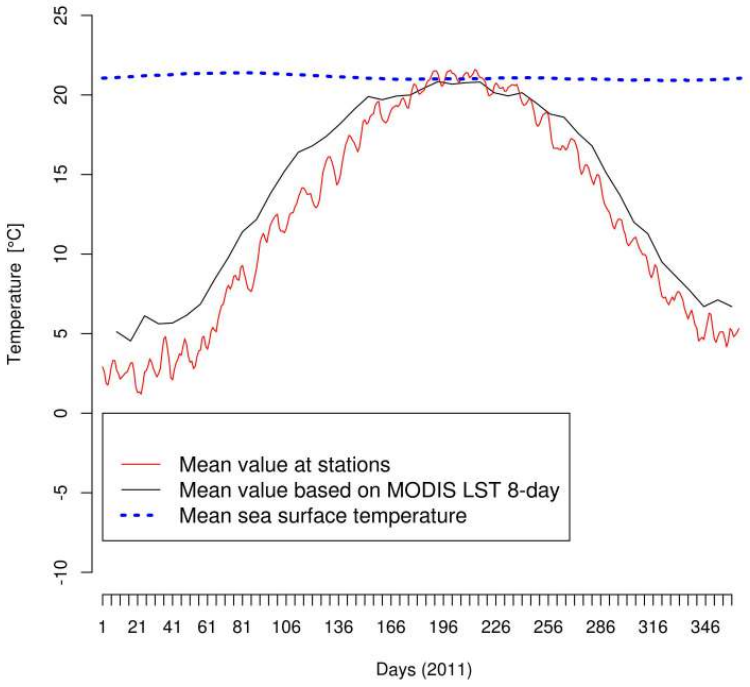

FIGURE 3.20: Mean daily temperatures for all stations in year 2011 (red), as compared to the mean 8-day temperature estimated based on the MODIS LST product (black), and the long-term sea surface daily temperatures obtained from `http://discover.itsc.uah.edu/amsutemps/` (blue).

Remote sensing is the future of global meteorology, even if the original RS imagery is noisy and produces biased estimates. For example, surface temperatures estimated by MODIS LST product can contain significant noise and artefacts. On the other hand, they can be used to get a more representative estimate of the global temperatures (*'the complete picture'*), which is impossible to achieve by using ground data only. Figure 3.20, showing differences between the global mean daily temperature from three sources, illustrates what we mean by bias. There are two logical explanations of the differences in the three curves in Figure 3.20. First, the mean sea surface temperature is relatively constant with slightly higher values in March (2011); both MODIS LST mean temperature and the mean temperature at the stations follow the same pattern that reflects the seasonality in the northern hemisphere as most of the land mass falls in the northern hemisphere. Second, meteorological stations do not cover land mass uniformly so that the mean can

not represent global mean temperature and therefore the values are in average somewhat higher than the MODIS LST temperatures. In general, accuracy of station's observations is higher than of the MODIS LST estimated values, but the station's observations provide a biased estimate of the global mean daily temperature because of clustering of points. If one would try to estimate global land temperature only based on the values on meteorological stations that he/she would probably make a systematic error (under-estimation) of 2-3°C. On the other hand, MODIS 8-day satellite images with 95% of coverage of land mask lack the fine precision of the daily measurements, providing land surface temperature. All this indicates that there is indeed a need for spatio-temporal regression-kriging methods that can produce calibrated images of daily air temperatures, so that also the global daily average can be estimated in an unbiased manner.

To conclude: the observed high temporal, spatial and feature space clustering of meteorological stations potentially represent a limitation of these data sets that could complicate fitting of accurate global spatio-temporal models. This does not imply that no reliable models can be fitted using these data sets, but that sophisticated spatio-temporal techniques need to be used that can account for the data clustering, spatially if remote sensing and/or monthly images are not used as predictor. Spatio-temporal regression kriging model can provide the most realistic estimate of the uncertainty, so that also an unbiased estimates of the global and local land air temperature and other meteorological variables can be produced. The presented model can be used for calibration of 8-day MODIS LST images by using station observation resulting with daily global images of mean daily air temperature at 1 km scale. This would be the first global daily images at very high spatial and temporal resolution (1 km spatial and 1 day temporal resolution).

Chapter 4

Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution[1]

Around 9000 stations from merged GSOD and ECA&D daily meteorological data sets were used to build spatio-temporal geostatistical models and predict daily air temperature at ground resolution of 1 km for the global land mass. Predictions were made for the mean, maximum and minimum temperature using spatio-temporal regression-kriging with a time series of MODIS 8 day images, topographic layers (DEM and TWI) and a geometrical temperature trend as covariates. The model and predictions were built for the year 2011 only, but the same methodology can be extended for the whole range of the MODIS LST images (2001–today). The accuracy of predicting daily temperatures has been assessed using leave-one-out cross-validation; the standard approach is extended with block approach. The values were aggregated for blocks of land of size 500×500 km to account for geographical point clustering of station data. All computations were implemented in the R environment for statistical computing by combining functionality of the gstat package (geostatistical modelling), rgdal and raster packages (raster data loading and analysis), and snowfall package (cluster computing). The results show that the average accuracy for

[1] Based on article: Kilibarda M, Hengl T, Gerard B.M. H, Benedikt G, Edzer P, Perčec Tadić M, Branislav B (2013?) Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *in review; Journal of Geophysical Research, D: Atmospheres*

predicting mean, maximum and minimum daily temperatures is RMSE = ±2°C for areas densely covered with stations, and between ±2°C and ±4°C for areas with lower station density. The lowest prediction accuracy was observed in highlands (> 1000 m) and in Antarctica with a RMSE around 6°C. This automated geostatistical framework could be used to produce global archives of daily temperatures (new generation WorldClim repository) and to feed various global environmental models.

4.1 Introduction

Records from near-surface weather stations are the foundation of climate research (Peterson and Vose, 1997). These stations still provide the most accurate and most reliable measurements of weather. Ground station measurements are not only important because of their high accuracy, they are also the only available records of spatial and temporal variation of climatic variables before the first satellite based observations came available in the 1960s.

Station observations are the main source of input of spatial interpolation that predict climatic variables on raster grids, while some interpolation methods also make use of additional, auxiliary information such as remotely sensed images and/or topographic layers. In-depth reviews of interpolation methods commonly used in meteorology and climatology are given by Price et al. (2000), Jarvis and Stuart (2001), Tveito et al. (2006) and Stahl et al. (2006), just to mention the most recent publications in the field. The literature, in general, shows that interpolation techniques used in meteorology and climatology are quite diverse: they range from nearest neighbour methods, splines, regression and kriging, to neural networks and machine learning techniques.

Hofstra et al. (2008) made a comparison of six interpolation methods for prediction of daily precipitation, mean, minimum and maximum temperature, and sea level pressure from station data in Europe, and compared anomalies interpolation relative to the long-term monthly mean (1960-1990). The result showed that global kriging on anomalies was the best overall performing method. Besides using ordinary kriging on anomalies for predicting daily values at regional scales, a method known as *"regression-kriging"* (RK) (also known as *"Kriging with External Drift"* and/or *"Universal kriging"*) has been

widely recognized as a flexible and a well-performing technique for unbiased spatial prediction of meteorological and environmental variables (Carrera-Hernández and Gaskin, 2007; Haylock et al., 2008; Hengl et al., 2012).

To our knowledge, no group has previously attempted to interpolate daily values of meteorological variables using spatio-temporal regression-kriging with time-series of remote sensing based covariates, at fine resolution of 1 km. The challenges to achieve this are significant, both from a methodological and technological perspective. The global land mask contains about 149 million pixels at 1 km resolution, which means that predicting daily values for 10 years would result in about 4 TB of data. Fitting of geostatistical models with millions of point pairs, and predicting at such large grid stacks can only be achieved by intelligent programming that avoids memory limit problems and computations that take weeks or longer.

In this study, we present an automated mapping framework for producing predictions of daily mean, minimum and maximum air temperatures using spatio-temporal regression-kriging implemented in the R environment for statistical computing via the `gstat` (Pebesma, 2004) and `stats` (R Development Core Team, 2012) packages. As inputs we use a collection of publicly available daily records from NCDC's Global Surface Summary of Day and European Climate Assessment & Dataset, and a time series of MODIS LST 8 day images and topographic layers as covariates. The research reported herein focuses on the year 2011 for practical purposes and assumes that the same methodology can be extended for the whole range of MODIS LST images (2001–today). The input data sets and methods are described in Section 4.2. The results of model fitting, cross-validation and validation are presented in Section 4.3, whereas summary results are given in the final 4.4 "Discussions and conclusions".

4.2 Data and Methodology

4.2.1 Merged global station data set

GSOD and ECA&D daily meteorological data sets have been merged to produce a consistent global station data set. A large portion of the station data had missing values, but

all stations were used in the interpolation procedure. Even though the meteorological services responsible for collecting the data also perform at least basic quality control, a small portion of the stations from the merged data sets contained clear gross errors and needed to be cleaned. The gross errors were detected using the following procedure: an initial spatio-temporal model was first fitted using all data followed by analysis of cross-validation predictions at the station location. The high cross-validation root mean square errors (derived from yearly residuals at certain station) for a number of stations suggest that there could be some gross errors in the data. This was usually confirmed with abrupt jumps in observation values in time series plots; thereby showing observations against cross-validation prediction. We decided to remove all stations that had cross-validation residual higher than $\pm$ 15°C, as these clearly contain errors in the data set. After all data filtering, the final set contained about 9000 stations from merged GSOD and ECA&D data sets.

4.2.2 Covariates: remote sensing images and DEM-derivatives

4.2.2.1 National Aeronautics and Space Administration (NASA)

The Moderate Resolution Imaging Spectroradiometer (MODIS) images of the Terra and Aqua Earth Observing System (EOS) platforms provide the possibility for retrieving atmospheric and oceanographic variables using different techniques. There are many data products from MODIS observations describing land (temperature, land cover), oceans (sea surface temperature, optical thickness) and atmosphere (water vapor, cloud product, atmospheric profiles) that can be used for studies at different scales ranging from local to global. Land Surface Temperature (LST) data and images obtained from MODIS thermal bands that are distributed by the Land Processes Distributed Active Archive Center (LP DAAC FTP) of the US Geological Survey are very often used in meteorological studies (Coll et al., 2009).

In addition to the MODIS Level 2 or higher level data, there are also MODIS Level 1 data that are distributed trough the LAADS portal hosted at the Goddard Space Flight Center. The MODIS LST data are available on a daily basis and have spatial resolutions of 1×1 km (Coll et al., 2009). The nominal accuracy of the MODIS LST product is $\pm$1°K.

However, some validations reported accuracy statistics smaller better than 1°K in clear sky conditions within the temperature range of -10°C to 50°C (Yoo et al., 2011). MODIS LST data and/or images are one of the most often used and best documented publicly available remote sensing products in the world.

In this work, we only use Level 3 MODIS LST 8 day composite images to improve spatial predictions of mean, minimum and maximum daily temperature despite the fact that daily day-time and night-time MODIS LST images are also available. Notably, the correlation with ground data would probably be more significant if we would use day-time MODIS LST images for maximum temperature prediction and night-time images for minimum temperature prediction. However, day/night images contain many missing pixels that ultimately limit their usability for global mapping.

The original 8 day MODIS LST images were converted from degrees Kelvin to Celsius. The original images still contained 0–15% of missing pixels due to clouds or other reasons which were replaced using the SAGA GIS function *"Close Gaps"*. This function uses splines as a robust method for filling gaps in areas with sparse or irregularly spaced data points (Neteler, 2010). Furthermore, the 8 day images were disaggregated in the time dimension through the use of splines for each pixel. As a result, the daily coverage was obtained.

4.2.2.2 DEM derivatives

The elevation data used in this study were obtained from the WorldGrids.org portal. The data set is derived as a combination of the publicly available SRTM 30+ and ETOPO data sets, and is commonly referred to as DEMSRE.

The *'SAGA Wetness Index'* is based on a modified catchment area calculation implemented in SAGA GIS (Böhner et al., 2008). The Global SAGA Wetness Index used in this paper was produced by Milan Kilibarda and Tomislav Hengl and the processing script is available via the WorldGrids.org data portal.

4.2.3 Spatio-temporal regression kriging

Consider the problem of describing the spatio-temporal process of a continuous variable Z. Z varying over space and time, e.g. temperature varies in space from one location to another and in time from one point in time to another. The statistical model of such a process is typically composed of the sum of a trend and a stochastic residual (Burrough, 1998; Heuvelink and Griffith, 2010; Hengl et al., 2012):

$$Z(\mathbf{s},t) = m(\mathbf{s},t) + \varepsilon'(\mathbf{s},t) + \varepsilon''(\mathbf{s},t) \tag{4.1}$$

where $\mathbf{s},t$ is the space-time continuum, m is the trend component, $\varepsilon'(\mathbf{s},t)$ is the spatio-temporally correlated stochastic component and $\varepsilon''(\mathbf{s},t)$ is the uncorrelated noise. The phenomenon Z is observed at a finite set of points in space and time. An interpolation technique is required in order to predict Z at an unobserved location or time. Geostatistical interpolation techniques start by defining a model that describes the degree of variation of the variable of interest in space and time, then followed by characterizing its relationship with explanatory variables that are denoted as *'covariates'*.

The global trend of Z can often be explained using covariates known over the spatio-temporal domain, e.g. part of the variation of temperature can be explained using climatic factors (static) such as latitude and elevation, TWI and time dependent predictors like day of year and space and time dependent MODIS LST. It is convenient to represent the relationship between the dependent variable and the covariates using a linear model. The linear trend model is given by:

$$m(\mathbf{s},t) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{s},t), \tag{4.2}$$

where the β_i are unknown regression coefficients, the f_i covariates that must be exhaustively known over the spatio-temporal domain, and p is the number of covariates. Covariate f_0 is taken as unity, resulting in β_0 representing the intercept.

The linear model for the mean daily temperature is given as a multiple linear regression on the covariate layers (described in section 4.2.2). In addition to these covariate layers,

we assume that the global daily temperature is a function of geometrical position of a particular location on Earth and day of the year. We call this a *'geometrical temperature trend'*. The geometrical temperature trend for the mean temperature was modelled as a function of the day of the year and latitude (ϕ):

$$t_{geom} = 30.4\cos\phi - 15.5(1-\cos\theta)\sin|\phi|, \tag{4.3}$$

where θ is derived as:

$$\theta = (day-18)\frac{2\pi}{365} + 2^{1-\mathrm{sgn}(\phi)}\pi. \tag{4.4}$$

The number 18 represents the coldest day in the northern and warmest day in the southern hemisphere and was derived empirically by graphical inspection of mean daily temperature plots from stations in the northern and southern hemisphere. The sgn denotes the signum function that extracts the sign of a real number. Parameters 30.4°C and 15.5°C of the geometric temperature trend were calculated by least squares fitting on circa 44 million daily temperature observations from 2000 to 2011. These two numbers are, in fact, similar to the mean yearly temperature on the Equator and the mean global Earth temperature.

The linear model for the minimum daily temperature uses the same covariates as the linear multivariate model for mean daily temperature. The geometrical temperature trend for minimum daily temperature was:

$$t_{geom} = 24.2\cos\phi - 15.7(1-\cos\theta)\sin|\phi|, \tag{4.5}$$

The geometrical temperature trend for maximum daily temperature was:

$$t_{geom} = 37\cos\phi - 15.4(1-\cos\theta)\sin|\phi|, \tag{4.6}$$

The parameters 24.2°C and 15.7°C in Eq.(4.5), and 37°C and 15.4°C in Eq.(4.6) were also derived by the method of least square estimation based on the 11 years of observation from 2000 to 2011.

In recent years, many covariate layers with fine resolution have become available and regression models can often explain a significant part of the observed variation. However, in practice, the trend cannot explain all variation even though predictors are spatially, temporally and spatio-temporally varying. The residuals of the regression model might show spatio-temporal dependencies, which suggests that a spatio-temporal variogram may be estimated from the residuals at observation locations and used to krige the residuals. To make this explicit, we write the variable of interest as a sum of the trend and space-time residual:

$$Z(\mathbf{s},t) = m(\mathbf{s},t) + V(\mathbf{s},t), \tag{4.7}$$

where V is a zero-mean stochastic residual.

To proceed with the estimation of the spatio-temporal covariance structure of V, we assume it to be stationary and spatially isotropic. In other words, we assume that the variance of V is constant and that the covariance of V at points $(\mathbf{s},t)$ and $(\mathbf{s}+\mathbf{h},t+u)$ only depend on their separation distance (h,u), where h is the Euclidean distance $|\mathbf{h}|$. These assumptions might be hard to fulfil for the random field Z but are more likely to be realistic for the residuals. The spatio-temporal covariances are usually described using a spatio-temporal variogram, which measures the average dissimilarity between data separated in the spatio-temporal domain using the distance vector (h,u) defined as:

$$\gamma(h,u) = \frac{1}{2}\mathbb{E}(V(\mathbf{s},t) - V(\mathbf{s}+\mathbf{h},t+u))^2 \tag{4.8}$$

where $\mathbb{E}$ denotes mathematical expectation.

The residual V may be thought of as comprising three components: spatial, temporal, and spatio-temporal interaction (Heuvelink et al., 2012). The sum-metric variogram structure, that considers these three components as mutually independent, is defined as (Heuvelink et al., 2012):

$$\gamma(h,u) = \gamma_S(h) + \gamma_T(u) + \gamma_{ST}(\sqrt{h^2 + (\alpha \cdot u)^2}), \tag{4.9}$$

where $\gamma(h,u)$ denotes the semi-variance of V with h units of distance in space and u units of distance in time, γ_S, γ_T are purely spatial and temporal components, and γ_{ST} is the space-time interaction component. The spatio-temporal anisotropy ratio α converts units of temporal separation (u) into spatial distances (h). The spatio-temporal sum-metric variogram model can be seen as a surface with ten parameters; three parameters for each variogram component (sill, nugget, range) and the spatio-temporal anisotropy parameter α. Semivariances (and covariances) can be estimated for any spatio-temporal separation distance (h,u) once these parameters are estimated from the observed residuals. In turn, these can be used in spatio-temporal kriging to compute the best linear unbiased predictor (i.e., with minimum expected mean squared error) for any space-time point where V (and Z) was not observed. The formulas of kriging in the spatio-temporal domain do not differ fundamentally in a mathematical or statistical sense from those of spatial kriging (Heuvelink et al., 2012):

$$\hat{V}(\mathbf{s_0}, \mathbf{t_0}) = \mathbf{c_0}^{\mathbf{T}} \mathbf{c}^{-\mathbf{1}} \overline{\mathbf{V}}, \tag{4.10}$$

where $\mathbf{c}$ is the $n \times n$ variance-covariance matrix of the residuals at the n observation space-time points, as derived from the spatio-temporal variogram, $\mathbf{c_0}$ is a vector of covariances between the residuals at the observation and prediction points, $\mathbf{T}$ denotes matrix transpose, and $\overline{\mathbf{V}}$ is a vector of residuals (see Eq. 4.7) at the n observation points.

The final prediction of variable Z at location $(\mathbf{s}_0, t_0)$ is defined as:

$$\hat{z}(\mathbf{s}_0, t_0) = \hat{m}(\mathbf{s}_0, t_0) + \hat{V}(\mathbf{s}_0, t_0). \tag{4.11}$$

where $\hat{m}(\mathbf{s}_0, t_0)$ is the estimated multiple linear regression trend. The regression coefficients are estimated in the usual way, using where possible Generalized Least Squares or Ordinary Least Squares (Pinheiro and Bates, 2009). Note that 'regression-kriging' specifically implies that the regression modelling and residual kriging parts are addressed separately: we first produced predictions for the regression part (see Eq. 4.2), followed by

extracting residuals for all observations and finally fitting a global sum-metric variogram model. The residuals were then interpolated and added to the predicted trend.

Spatio-temporal regression-kriging has made a breakthrough in the past decade with theoretical concepts and providing various examples of applications (Gething et al., 2007; Heuvelink and Griffith, 2010; Heuvelink et al., 2012; Gräler et al., 2011; Hengl et al., 2012). Here, we extend the spatio-temporal regression-kriging framework that combines ground observation together with MODIS 8 day images, as was presented in a Croatian case study (Hengl et al., 2012), to a global data set and hyper-resolution data. We implemented all the computing in the R environment for statistical computing (R Development Core Team, 2012) by combining functionality of gstat package (geostatistical modelling), rgdal and raster packages (raster data loading and analysis), and snowfall package (cluster computing). We used the gstat package (Pebesma, 2004) that is also capable of working with spatio-temporal data sets defined in spacetime package (Pebesma, 2012) for variogram model fitting. The sample variograms were estimated with spatial lags of 50 km and time lags of 1 day. Because this is a global point data set, all distances were calculated as great circle distances in the WGS84 coordinate reference system.

4.2.4 Accuracy assessment

Two approaches were applied for assessing the accuracy of the predictions made for the daily temperature of the global land surface as obtained with spatio-temporal regression-kriging. These were: (1) cross-validation (CV), and (2) comparison with GHCN-M monthly data. For validation using GHCN-M data, we predicted values at daily resolution and then aggregated these predictions to monthly averages. Stations from the GHCN-M dataset that were closer than 50 km to any station used in this study were excluded in order to avoid station overlap and to increase the independence of the validation data to obtain more objective results.

For cross-validation, we use the leave-one-out cross-validation method. This works as follows: the method predicts a complete annual time series of daily temperature at all stations using only observations from neighbouring stations (i.e. the 35 nearest observation were sampled for current date and the day before and after the current one, resulting in the collection of around 105 observations that exclude data from the target station itself).

The predicted values were compared with the actual observations of the target stations to derive cross-validation statistics. The final accuracy of the model is assessed using the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{m} \cdot \sum_{j=1}^{m} \left[\hat{T}(\mathbf{s}_j, t_j) - T(\mathbf{s}_j, t_j)\right]^2} \tag{4.12}$$

where $\hat{T}(\mathbf{s}_j, t_j) - T(\mathbf{s}_j, t_j)$ is the difference between the cross-validation prediction and the observed temperature at spatio-temporal location $(\mathbf{s}_j, t_j)$, and m is the number of observations for the station. The derived RMSE per station were then exported to KML and HTML formats to allow for visual exploration of errors in space and time domains. These visualizations can be accessed via the `http://dailymeteo.org` website.

Because stations are heavily clustered (Kilibarda et al., 2013c), the global RMSE mostly depends on the accuracy in areas with a high station density. In order to obtain a more objective measure of accuracy that accounts for this point clustering the block aggregated RMSE for 500 by 500 km blocks of land prepared in Sinusoidal equal area projections is analysed. The regression-kriging cross-validation statistics were first calculated in the WGS84 coordinate reference system using geodetic line distances, and then were re-projected to Sinusoidal projection for the block aggregation.

4.3 Results

4.3.1 Mean daily temperature interpolation

4.3.1.1 Linear regression for mean daily temperature

Figure 4.1 shows the geometrical trend values against observed temperatures for two sample stations. Surprisingly, the geometrical temperature trend already explains 75% of daily temperature variation with a standard error of ± 5.7°C.

Figure 4.2 shows the dissagregated MODIS LST 8 day layer (MODIS spline) against observed temperature for two stations. The linear regression model using only MODIS

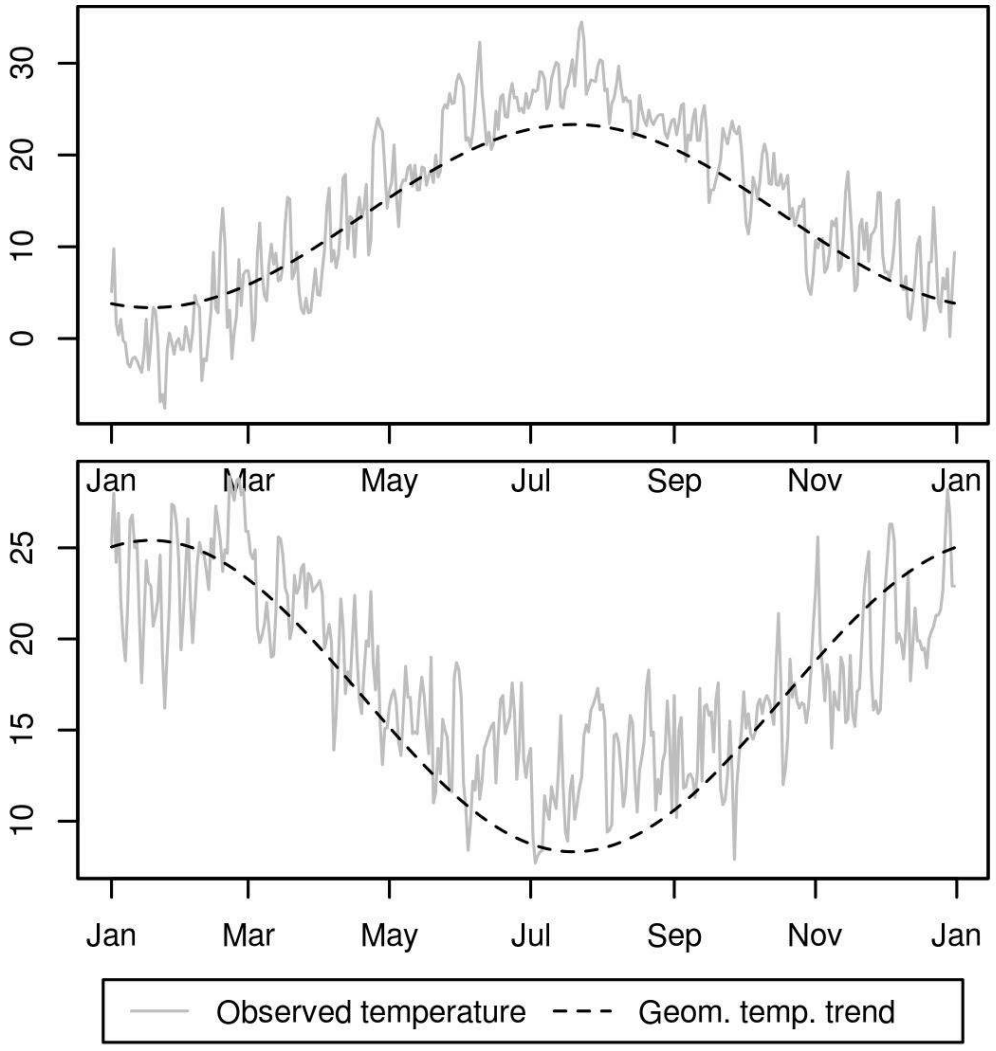

FIGURE 4.1: Mean daily temperature observation in 2011 (gray solid line) and geometrical temperature trend (black dashed line). PHILADEPLHIA, USA ($\lambda = -75, \phi = 39.993$) (top). BUNBURY, Western Australia ($\lambda = 115.65, \phi = -33.35$) (bottom).

LST spline images explains 80% of the variability in mean daily temperature values for the year 2011. Thus, MODIS LST spline images are significant estimators of the daily temperature with an average precision of ±5.2°C. Again, this precision is lower than the one reported by Wan et al. (2004) because we use 8 day composites and not daily MODIS LST images in order to reach full land coverage.

The DEM and TWI layers also appeared to be significant covariate layers even though we expected that MODIS LST will account for the variation of temperature with elevation. We suspect the main reason for some elevation dependency was left unexplained in MODIS LST is related with the fact that this is a cloud free product so it is likely to underestimate winter temperatures (due to strong radiative cooling in cloud free situations) and overestimate summer temperatures (Van De Kerchove et al., 2013). As a consequence, during winter days/nights surface observed temperatures would be higher under the clouds due to suppressed radiative cooling and our MODIS LST gap-filling procedure would probably underestimate temperature in these areas. During summer, under the clouds in the mountains observed surface temperature would be lower, while over gap-filling procedure would result in higher temperatures. Since this two processes are mainly elevation dependent this could be accounted for with DEM and TWI covariate layers.

The final multiple linear model with four covariates explains 84.2% of the variation and associated standard deviation of ±4.6°C. Figure 4.2 shows plots of modelled against observed temperature for the same stations as used in previews figures.

Figure 4.3 presents the general relationship between the observed temperature and linear model on the full data set used for spatio-temporal modelling. Note that the residuals are in general normally distributed around the regression line and no heteroscedasticity can be observed.

4.3.1.2 Spatio-temporal variogram model for mean daily temperature

ÂňThe right-hand side of Figure 4.4 shows the 2D and 3D sample space-time variogram. The fitted model (ten variogram parameters described in Section 4.2.3 with the *fit.StVariogram* function in `gstat`) is shown in the left-hand side of Figure 4.4. Table 4.1 summarizes the

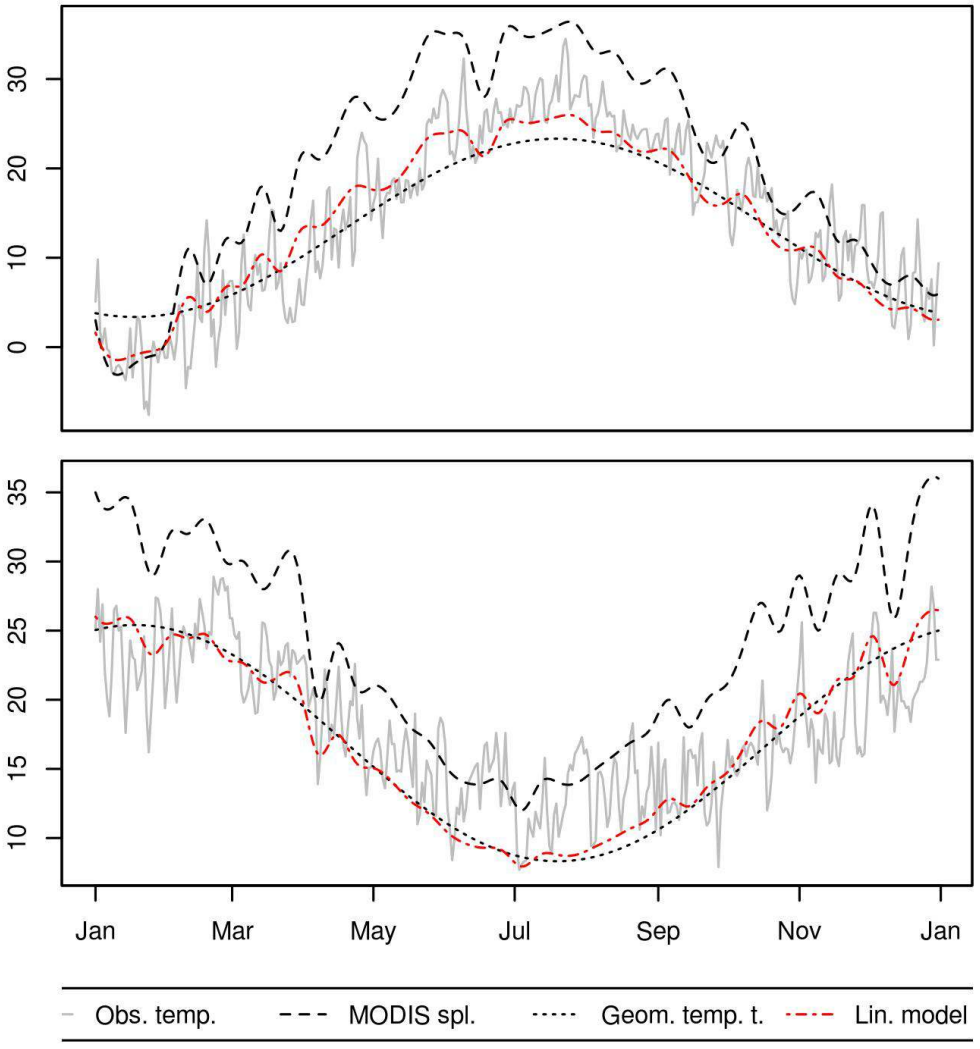

FIGURE 4.2: Mean daily temperature observation in 2011 (gray solid line) and multivariate linear model of mean daily temperature (red dashed line) on MODIS LST spline (black dashed line), geometrical temperature trend (black dotted line), elevation and topographic wetness index. PHILADEPLHIA, USA ($\lambda = -75, \phi = 39.993$) (top). BUNBURY, Western Australia ($\lambda = 115.65, \phi = -33.35$) (bottom).

parameter estimates of the sum-metric variogram model. Note that all variogram components were modelled as spherical functions.

The Figure 4.4 indicates that regression residuals have clear correlations both in space and time and therefore spatio-temporal kriging of residuals is certainly applicable. The fitted

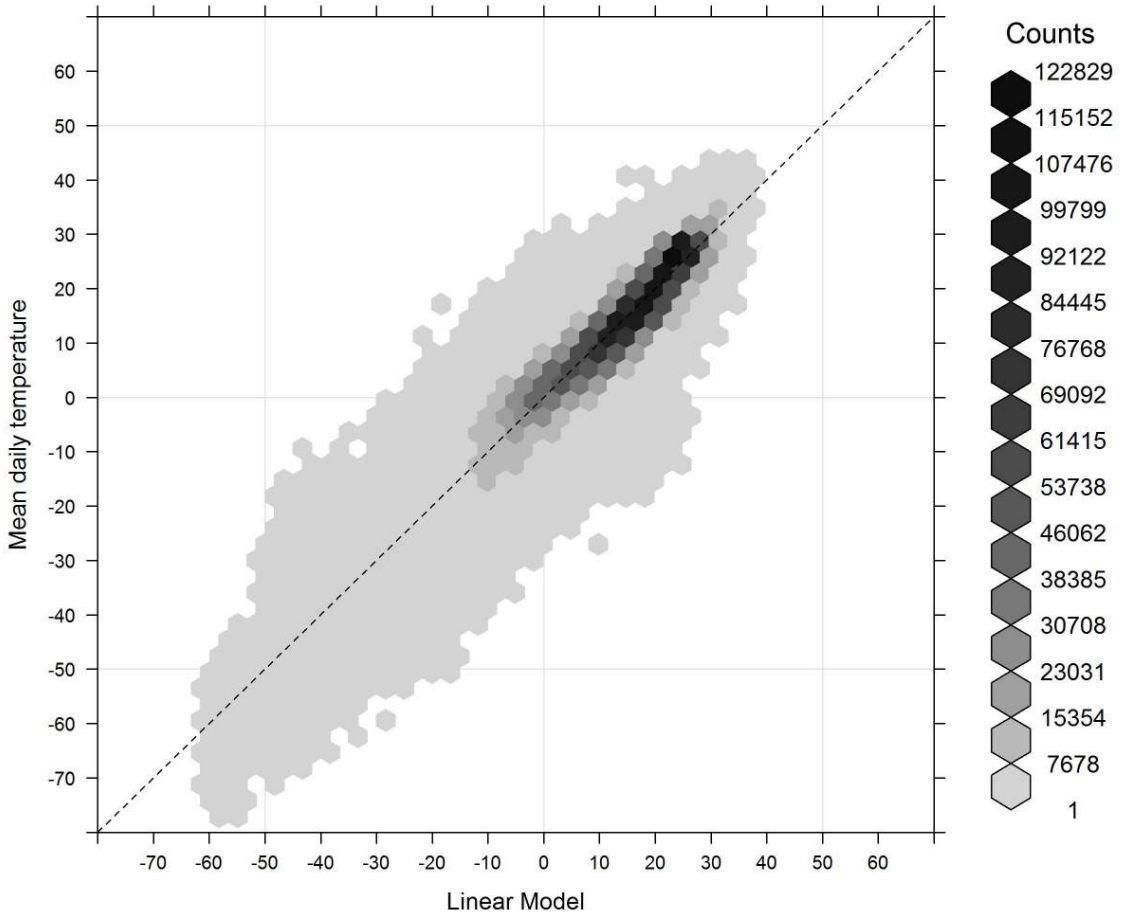

FIGURE 4.3: Scatter plot showing the general relationship between mean daily temperature and multivariate linear model prediction of mean daily temperature. The dashed line is the 1:1 relationship.

spatio-temporal variogram parameters of the mean daily temperature residuals show a significant purely spatial variogram component, while the purely temporal component is zero and temporal variability is only contained in the space-time interaction component. This suggests that the temporal pattern in mean temperature is probably already sufficiently captured by the regression model. Current residuals are correlated with residuals from day after (or before) and correlation depends on space-time distance. But only temporal separation between any two stations (without knowing spatial distance) doesn't explain

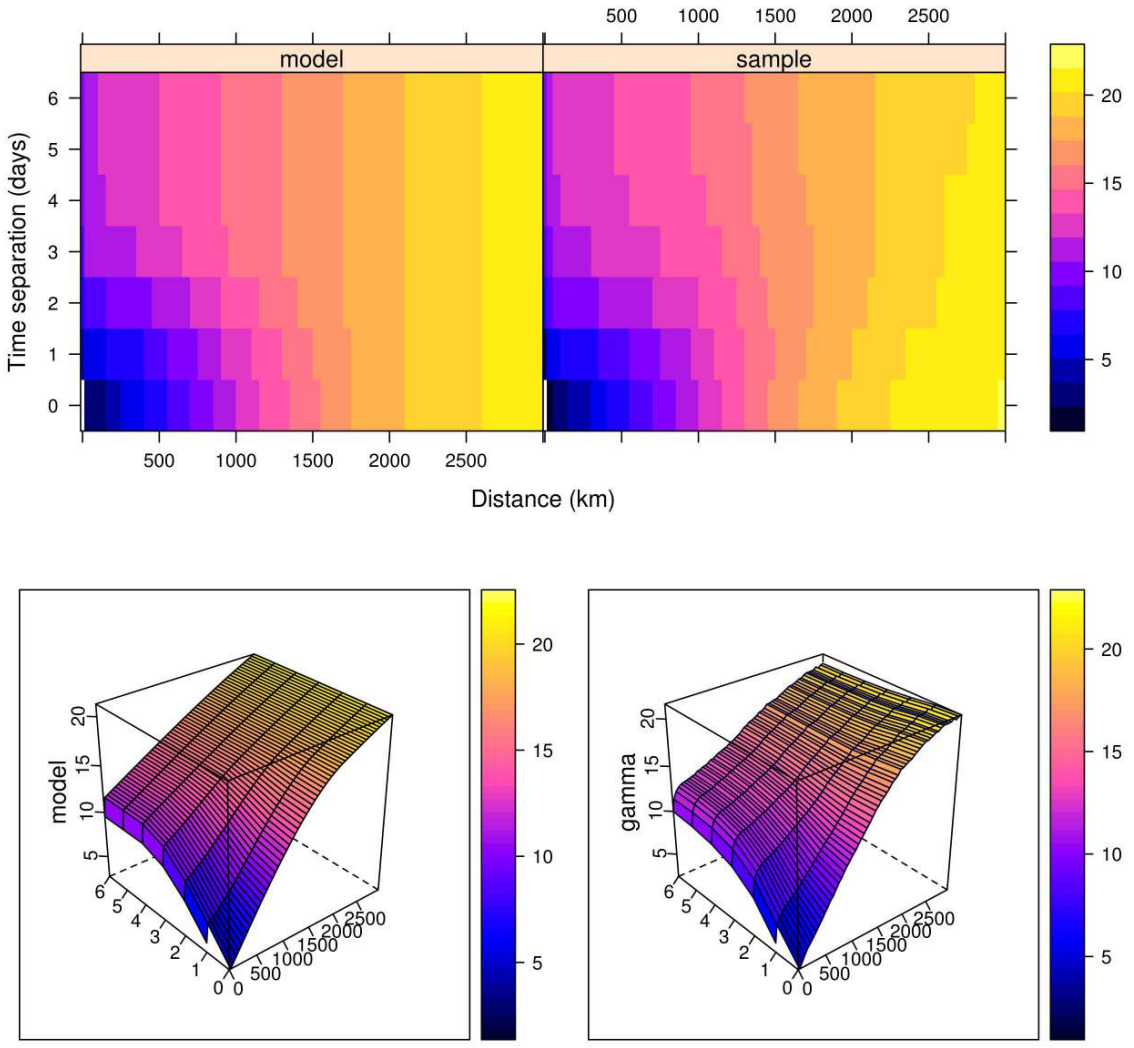

FIGURE 4.4: Fitted sum-metric model (left) and sample variogriom (right) of residuals from multiple linear regression of mean daily temperature on MODIS, geometrical temperature trend, elevation and topographic wetness index. The variogram surface is presented in 2D (top) and 3D (bottom).

TABLE 4.1: Parameters of the fitted sum-metric variogram model for mean daily temperature regression residuals, each component (see Eq. 4.9) is modelled using a spherical function.

	Nugget	Sill	Range parameter	Anisotropy ratio
spatial	1.934	14.13	5903 km	
temporal	0	0	0 days	
space-time	0.474	9.065	2054 km	497 km/day

even part of spatio-temporal correlation. Contrary, only spatial distance between two stations (without knowing temporal separation) explain part of spatio-temporal correlation. The short distance variation (nugget effect) in both the purely spatial and spatio-temporal components indicates that the model can't give better precision than $\pm$ 1.5°C globally (for interpolation at daily resolution). The range parameters are very large (especially pure spatial range) showing that the residuals are correlated within wide zones reaching sill after 6000 km. Thus, the local neighbourhoods need to be selected in a way that reflects the spatial and temporal ranges. Only few temporal instances will be selected while the spatial selection spans several hundred kilometres. This is captured by the spatio-temporal anisotropy as well suggesting that a station with a temporal lag of one day exhibits a similar correlation as a station about 500 km apart.

4.3.1.3 Accuracy assessment: mean daily temperature

An interpolated map of daily mean temperature for the first and second January of 2011 is shown in Figure 4.5, daily maps for the year 2011 at 1 km spatial resolution in GeoTiff format are available for download via `http://www.dailymeteo.org`. The mean daily temperature map of coterminous USA for the first 4 days in January 2011 is presented in Figure 4.6.

The cross-validation results on the complete data set showed a RMSE=2.4°C for global land areas including Antarctica, with R-square of 96.6%. The block aggregated RMSE results shows that the average accuracy is a bit worse (RMSE = 2.8 , see also map Figure 4.7). As mentioned previously, the global block aggregated RMSE gives a more

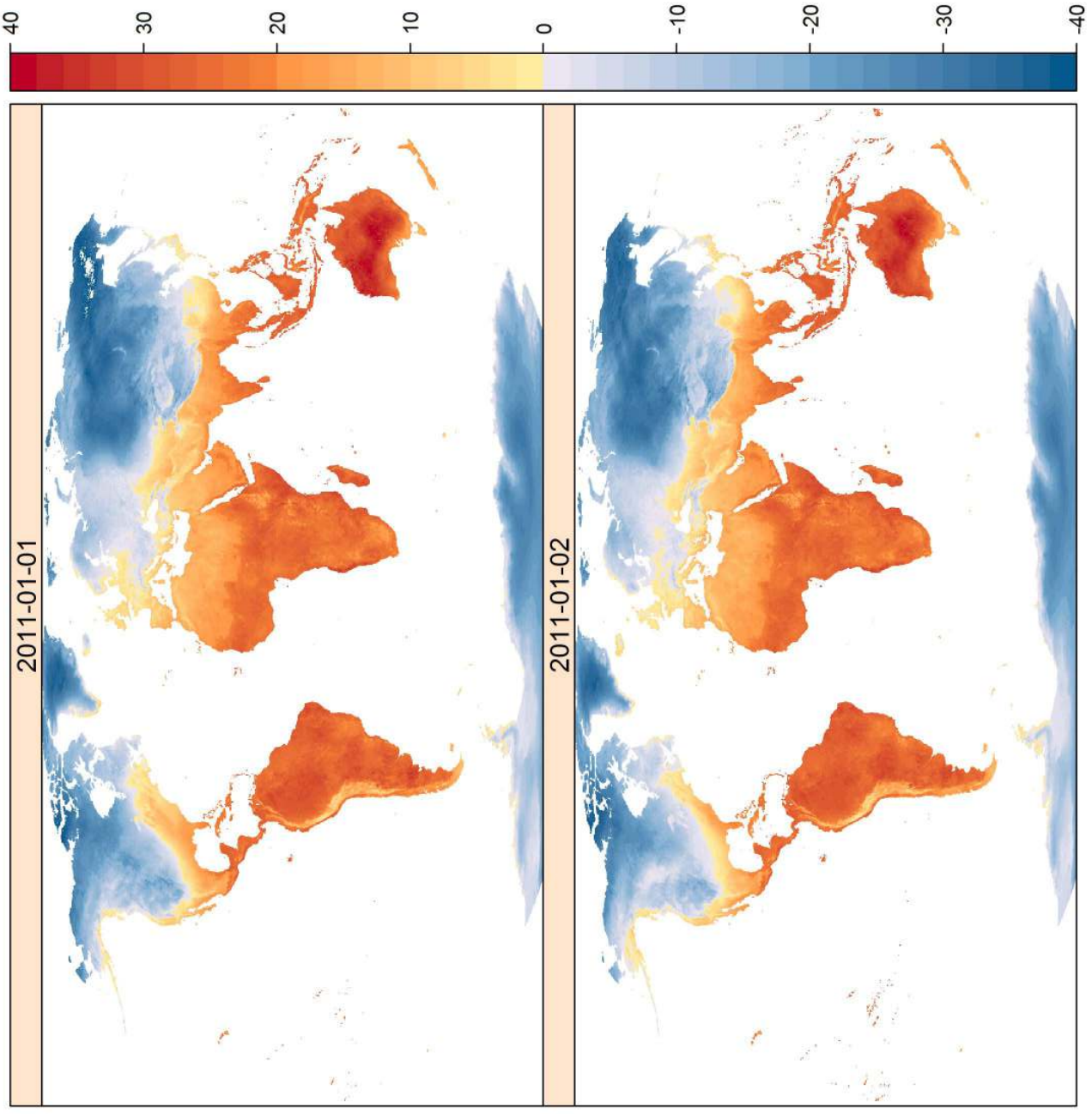

FIGURE 4.5: Map of mean daily temperature, interpolated by using spatio-temporal regression kriging on the GSOD and ECA&D stations observation. The map is Robinson projection.

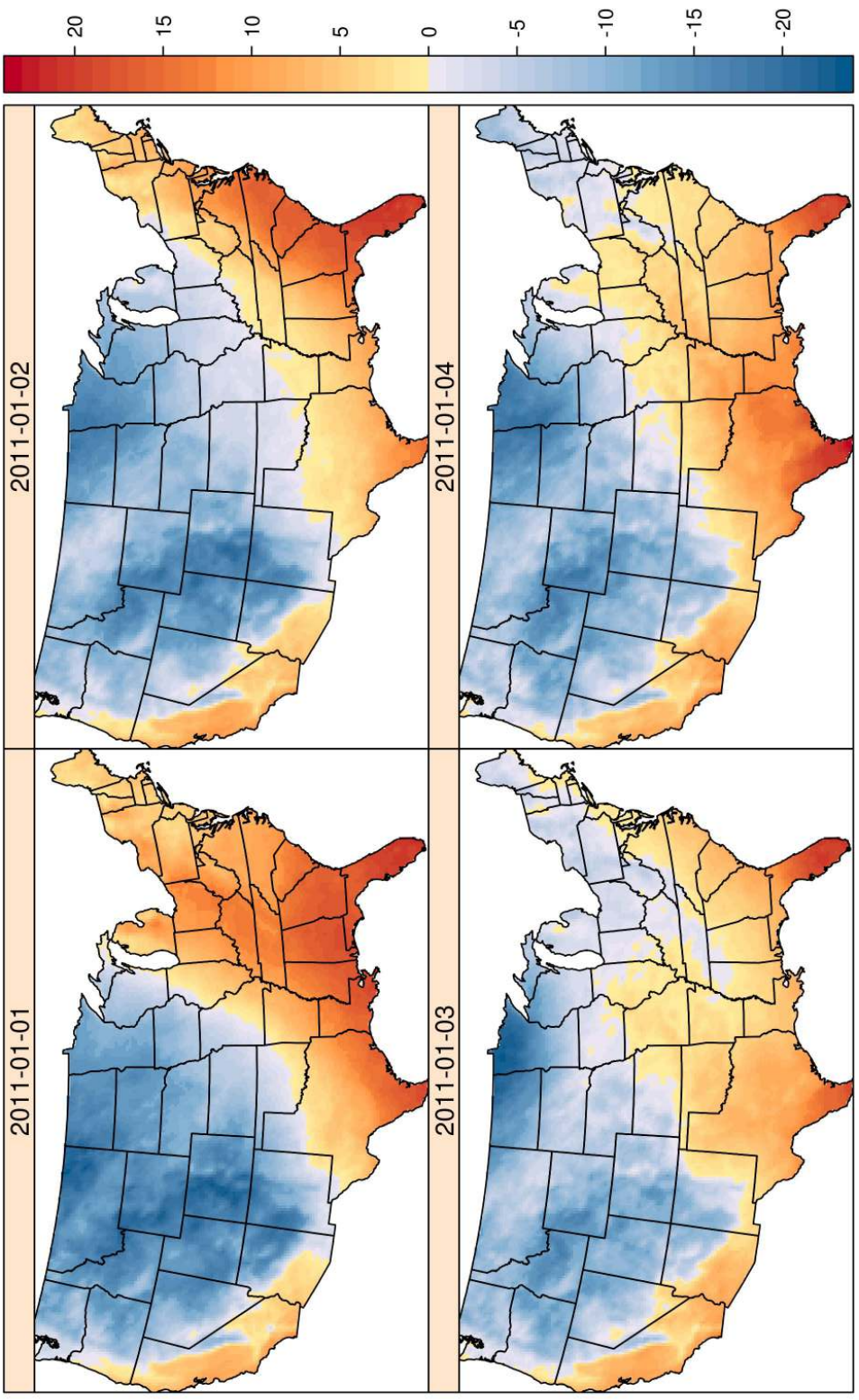

FIGURE 4.6: Map of mean daily temperature for the first 4 days in January 2011, covering coterminous USA in Albers equal-area conic projection.

objective global measure of accuracy. Thus, the actual RMSE is half a degree larger than RMSE calculated as a simple mean from all stations.

The monthlyRMSE obtained from cross-validation of monthly aggregated observations with cross-validation prediction is 1.7°C. This is an important result because it indicates that the model can be used for monthly image production (aggregation of daily gridded data). The yearly RMSE is 1.4°C. The spatial distribution of RMSE calculated per station (yearly average of squared daily cross-validation residuals, which is a daily quality measure) is shown in Figure 4.8. In this figure, the stations with RMSE <2°C represent 59% of the total number of stations (black dots), and 26% of stations have an RMSE between 2°C and 3°C. Figure 4.8 is also provided as an interactive map produced with the R package plotGoogleMaps (Kilibarda and Bajat, 2012), and is available via `http://www.dailymeteo.org`.

Observed and cross-validated values for two stations are shown in Figure 4.9. Considering the fact that cross-validation predictions are made using only 35 neighbouring stations (in spatial and 3 days in temporal domain) without any observation from the validation station, the spatio-temporal regression-kriging model is an accurate tool for filtering missing values in time series of mean daily temperatures.

The spatial distribution of RMSE can also be aggregated in the spatial domain by region or country. The aggregated results show that the smallest RMSE=1°C is achieved in the Netherlands, whereas Europe on average performs with an RMSE=1.6°C. Other results for large countries and regions are Russia (c.a. 2.2°C), USA 1.8°C, South America 3.1°C, while Antarctica has the highest RMSE with 5.9°C. An interactive map of spatially aggregated RMSE at the country level is also available via `http://www.dailymeteo.org`.

The RMSE on the GHCN-M data is 1.5°C and spatial distribution of RMSE calculated per year (yearly average of squared monthly validation residuals) for each station is shown on Figure 4.10 (an interactive map is available at `http://www.dailymeteo.org`). This map shows that 48% of predicted points have prediction accuracy smaller than 1°C. GHCN-M stations at a monthly resolution have an accuracy between 1 and 2°C for 40% of the points.

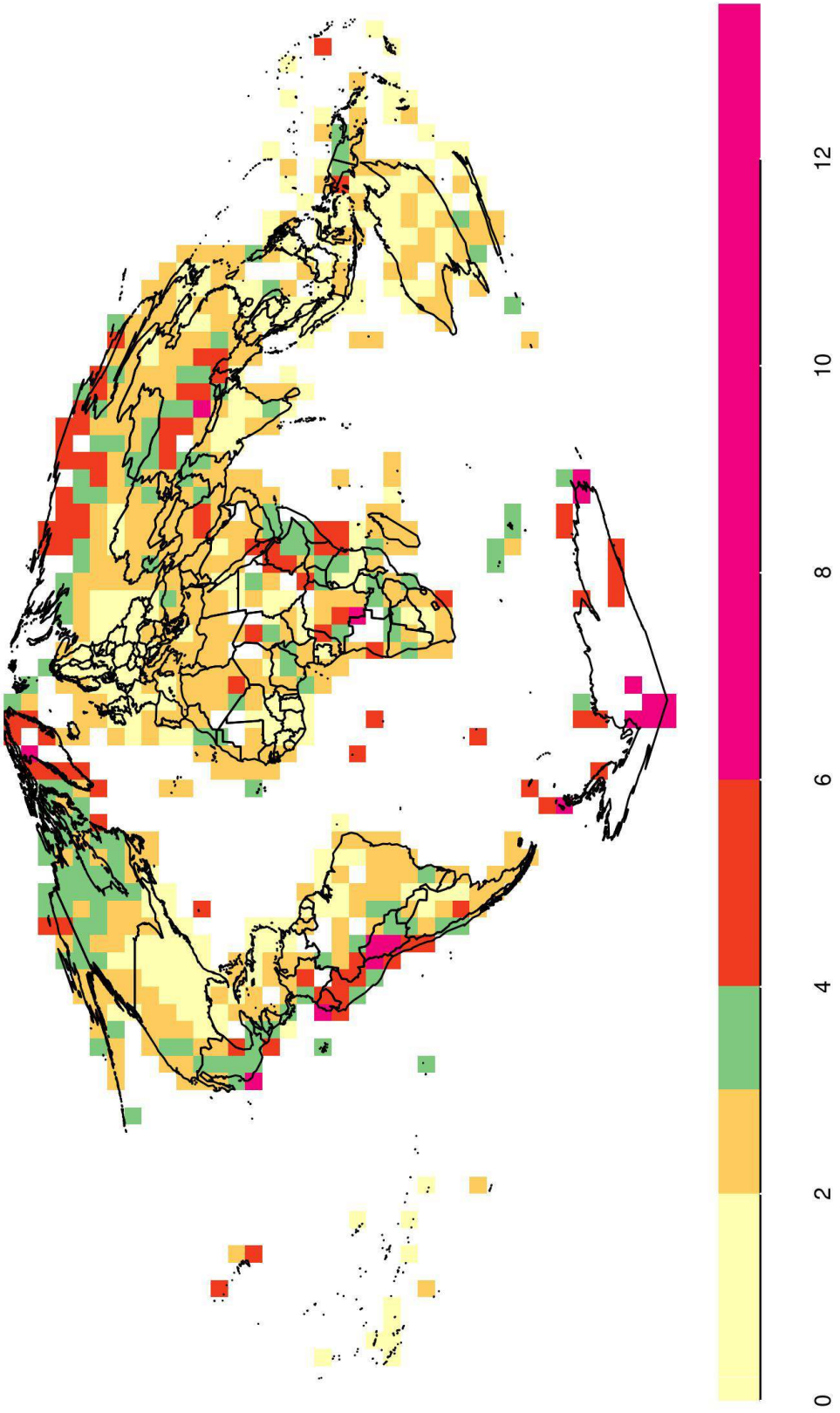

FIGURE 4.7: Map of mean daily temperature cross-validation errors (RMSE) aggregated to 500 by 500 km blocks. Block aggregation is made on equal area Sinusoidal projection.

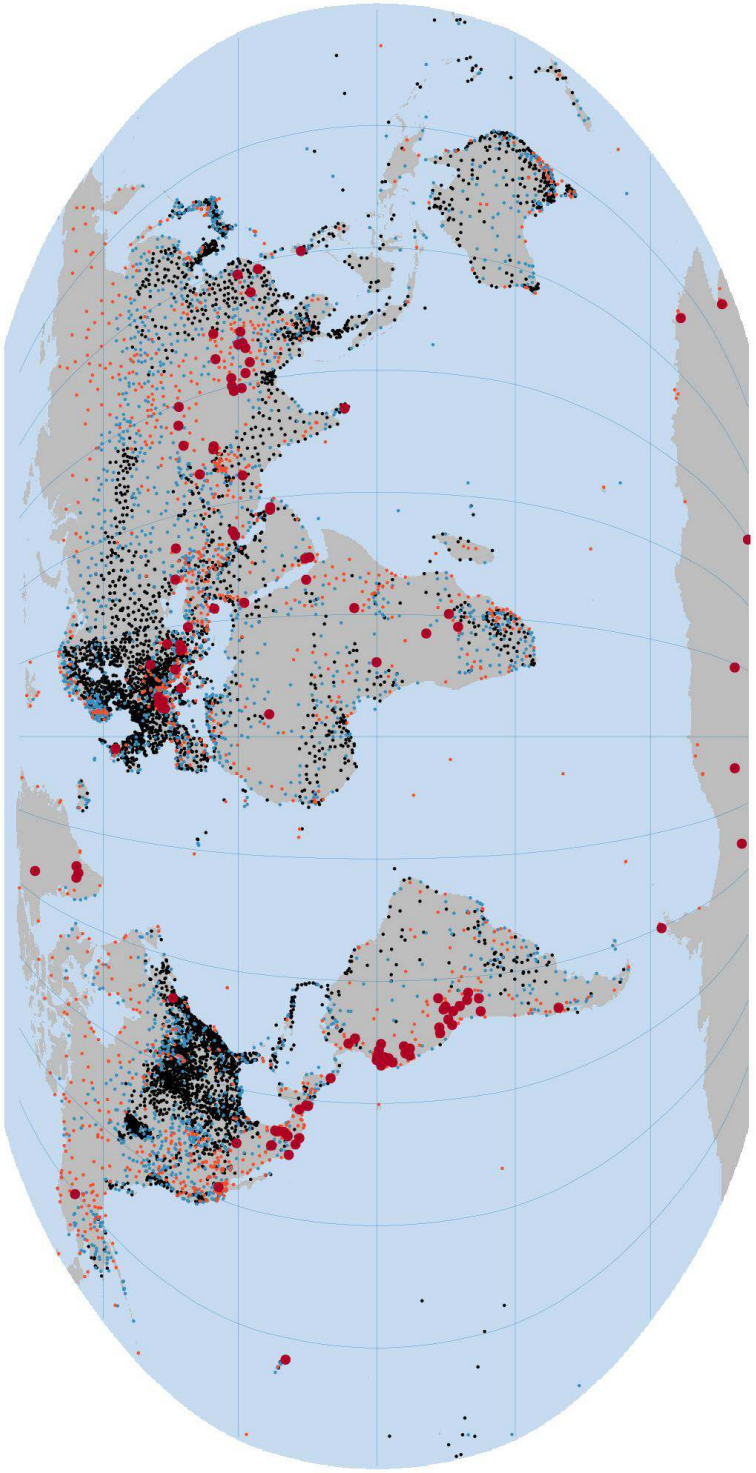

FIGURE 4.8: Map of mean daily temperature cross-validation errors (RMSE) averaged per year for each station. Red circles indicate cross-validation outliers with RMSE > 6. Clusters of red circles indicate problematic areas, partly presence of gross error in observation time series.

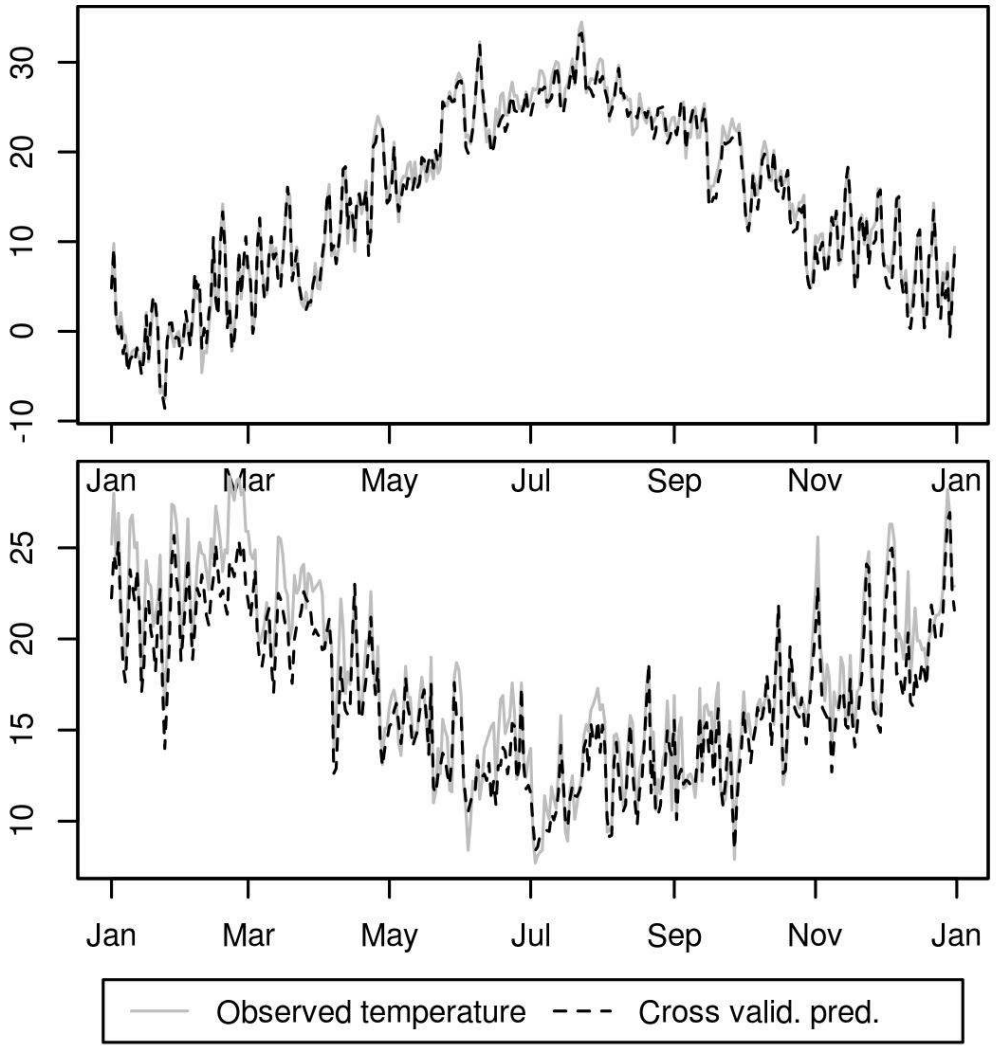

FIGURE 4.9: Comparison of mean daily temperature observations in 2011 (gray solid line) and space-time regression kriging cross-validation prediction of mean daily temperature (black dashed line). PHILADEPLHIA, USA ($\lambda = -75, \phi = 39.993$) (top). BUNBURY, Western Australia ($\lambda = 115.65, \phi = -33.35$) (bottom).

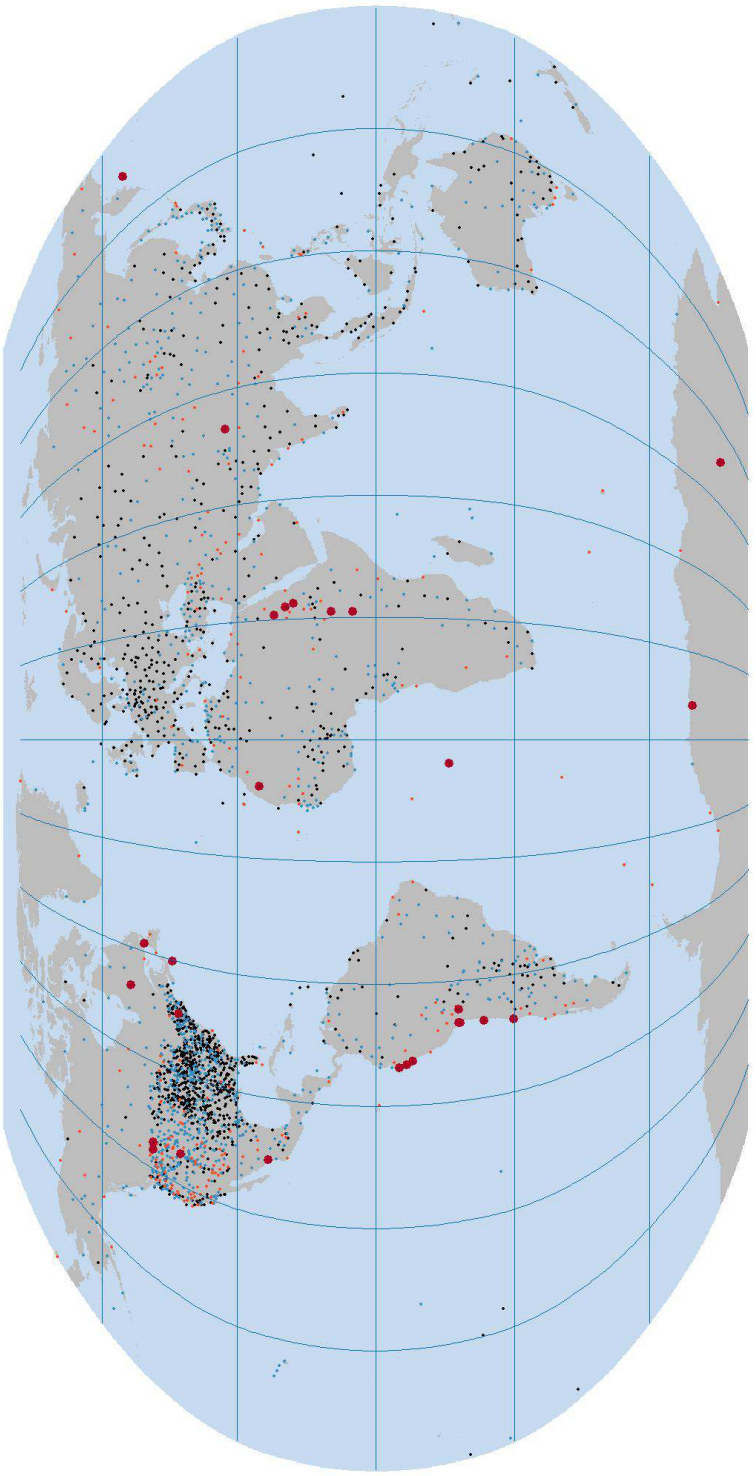

FIGURE 4.10: Map of the validation errors (RMSE) averaged per year for each station, calculated by using GHCN-M stations which were not used for model and prediction.

4.3.2 Minimum daily temperature interpolation

4.3.2.1 Linear regression model for minimum daily temperature

The geometrical temperature trend explains about 72% of the minimum daily temperature variations for 2011 with a standard error of ±6°C, Figure 4.11 shows geometrical trend against observation. The results of regression modelling based on MODIS LST spline images explains 70% of the variability in minimum daily temperature values for the year 2011 with an average precision of ±6.3°C; thus performing somewhat worse than for mean temperature.

DEM and TWI layers also showed to be highly significant covariates for minimum daily temperature. The final linear model with four covariates explains 77% of the variation with a standard error of ±5.5°C. Figure 4.11 shows a plot of the modelled geometrical trend for minimum daily temperature and MODIS LST spline values against observed temperature on the same stations.

4.3.2.2 Spatio-temporal variogram model for minimum daily temperature

The spatio-temporal variogram is modeled in the same way as was described for mean daily temperature. The variogram for minimum daily temperature has similar parameters as the mean daily temperature (see Table 4.2 and Figure 4.12). Again, the pure spatial component exists and pure temporal one doesn't. The pure spatial component shows spatial dependence of regression residuals across the Globe (pure spatial range is 5725 km) at any time separation, whereas complete temporal variability of residuals is contained in the spatio-temporal interaction part of the variogram structure. The nugget parts of these components are around 3.5°C^2, which is higher than in the mean temperature case and suggests that short range variability in space and time of minimum temperature regression residuals is significantly higher than for the mean temperature so extreme temperatures being harder to predict.

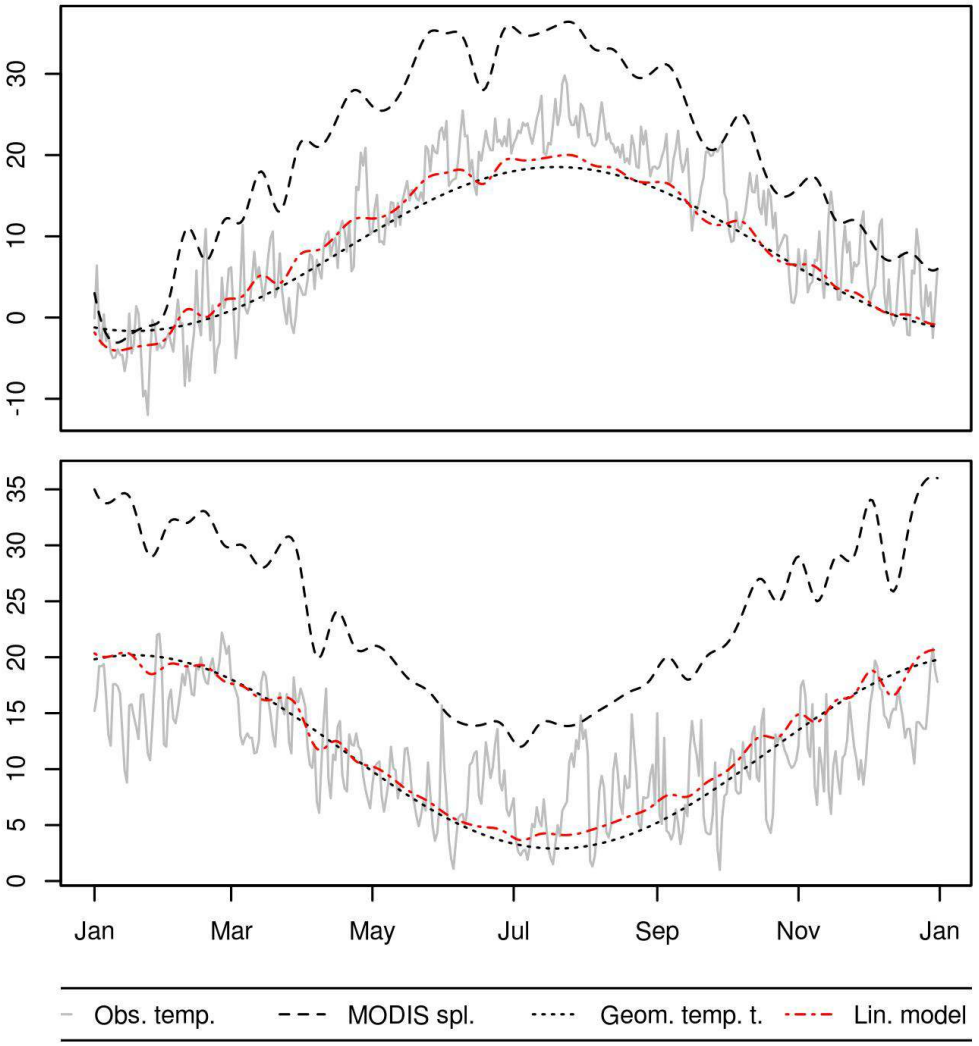

FIGURE 4.11: Minimum daily temperature observation in 2011 (gray solid line) and multivariate linear model of minimum daily temperature (red dashed line) on MODIS LST spline (black dashed line), geometrical temperature trend (black dot line), elevation and topographic wetness index. PHILADEPLHIA, USA ($\lambda = -75, \phi = 39.993$) (top). BUNBURY, Western Australia ($\lambda = 115.65, \phi = -33.35$) (bottom).

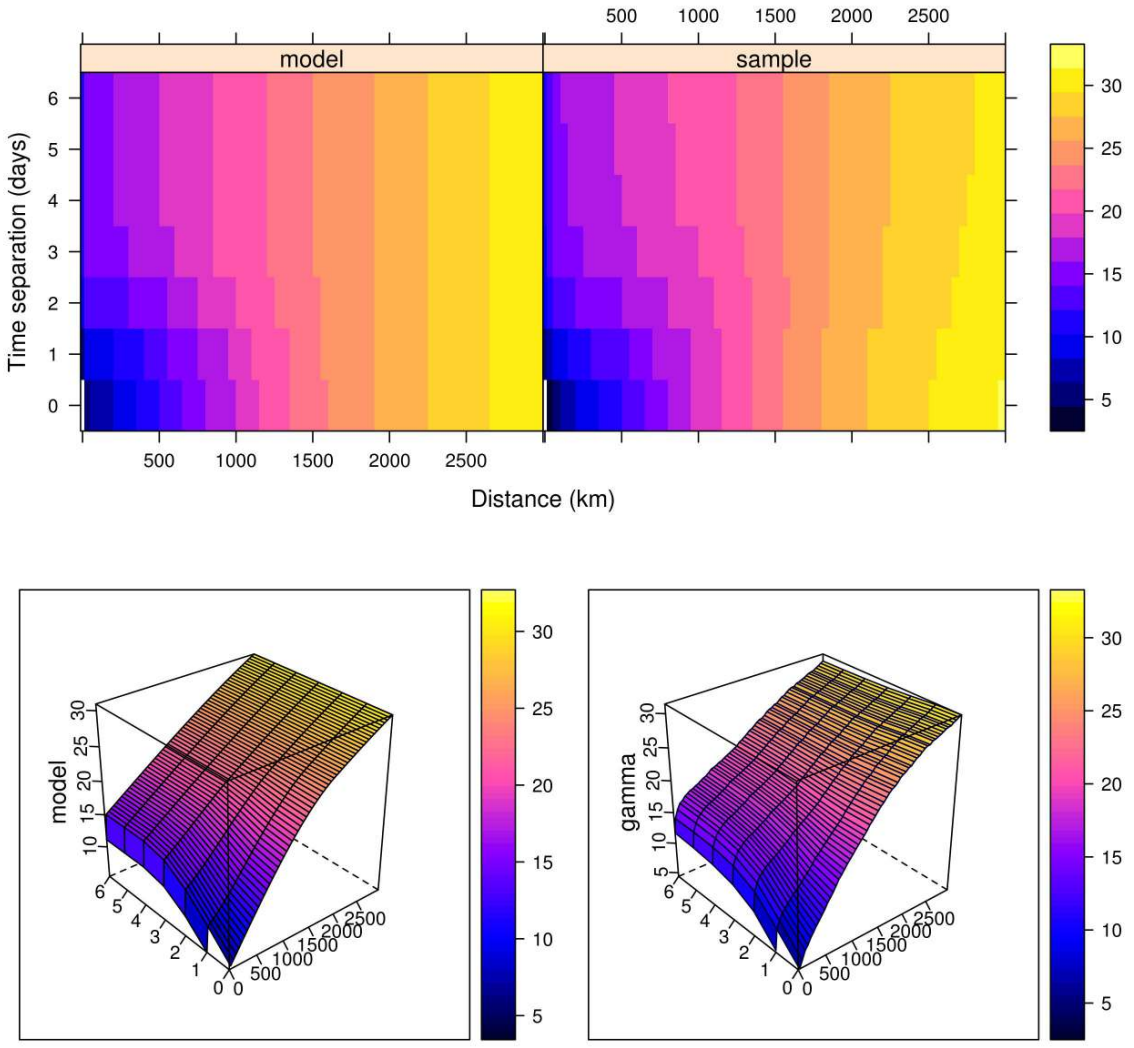

FIGURE 4.12: Fitted sum metric model (left) and sample variogriom (right) of residuals from multiple regression of minimum daily temperature observation on MODIS, geometrical temperature trend, elevation and topographic wetness index. The variogram surface is presented in 2D (top) and 3D manner (bottom).

TABLE 4.2: Parameters of the fitted sum-metric variogram model for minimum daily temperature regression residuals, each component (see Eq. 4.9) is modeled using a spherical variogram function.

	Nugget	Sill	Range parameter	Anisotropy ratio
spatial	3.695	22.682	5725 km	
temporal	0	0	0 days	
space-time	1.67	9.457	1888 km	485 km/day

4.3.2.3 Accuracy assessment: minimum daily temperature

The results of cross-validation for minimum temperature produced a RMSE=2.7°C for global land areas including Antarctica, with R-square of 94.2%. Monthly RMSE obtained from the cross-validation of monthly aggregated observation and cross-validation prediction is 2°C, yearly RMSE is 1.7°C. The spatial distribution of RMSE calculated per station (yearly average of squared daily cross-validation residuals, daily quality measure) for each station is shown in Figure 4.13, where the stations with RMSE $<$2°C are represented with 40% of the total number of stations (black dots), and 2°C$<$ RMSE $<$3°C with 35% (blue dots), 23% of points are with 3°C$<$ RMSE $<$6°C, and 200 stations are with RMSE $>$6°C.

The spatial distribution of RMSE also shows lower accuracy than predictions of the mean temperature in general. The aggregated results show that the lowest RMSE=1.4°C is achieved in the Netherlands, Europe without Russia (c.a. 2.7°C) an RMSE of around 2.3°C, USA 2.3°C, South America 3.1°C, Antarctica has again the highest RMSE=4.7°C.

4.3.3 Maximum daily temperature interpolation

4.3.3.1 Linear regression model for maximum daily temperature

The geometrical temperature trend in the linear model explains 75% of maximum daily temperature variation for 2011 with a standard error of $\pm$6.6°C. Figure 4.14 shows the

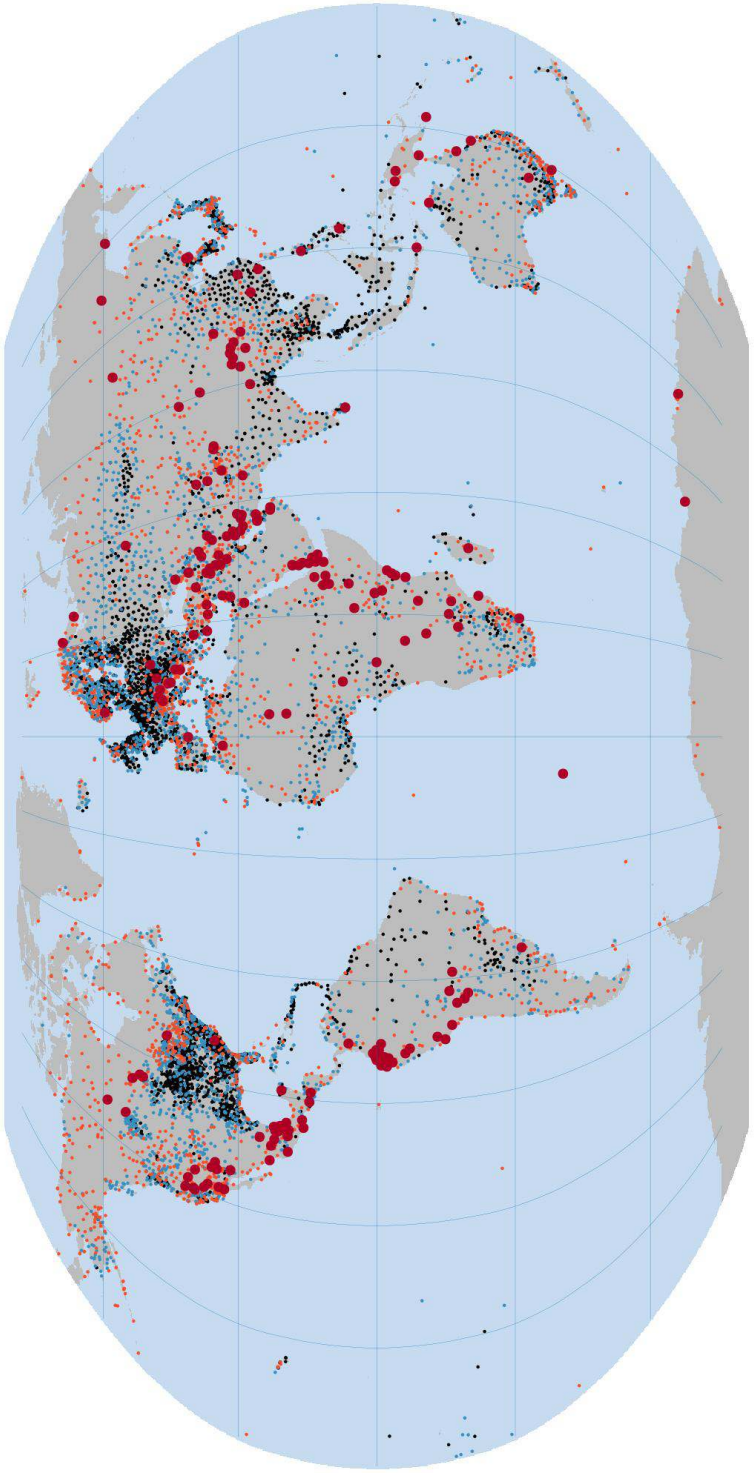

FIGURE 4.13: Map of minimum daily temperature cross-validation errors (RMSE) averaged per year for each station. Red circles indicate cross-validation outliers with RMSE > 6. Clusters of red circles indicate problematic areas, partly presence of gross error in observation time series.

geometrical trend compared against observation. The geometrical trend results are comparable to the results of modelling the mean temperature and are hence better than the results for modelling the minimum daily temperature case.

The regression modelling only with MODIS LST spline images already explains 84.5% of the variability in maximum daily temperature values for the year 2011 with an average precision of ±5.2°C. MODIS LST 8 day images are the best predictor for the maximum daily temperature when compared to actual mean and minimum daily temperatures. DEM and TWI layers also were significant covariate layers for the maximum daily temperature. The final linear model with four covariates explains 86.7% of variation with standard deviation of ±4.8°C, Figure 4.14 shows the modelled linear regression line, the geometrical trend for the maximum daily temperature and the MODIS LST spline values against observed temperature for the same stations.

4.3.3.2 Spatio-temporal variogram model for maximum daily temperature

Table 4.3 summarizes the parameters of the spatio-temporal variogram model, as in the previous variograms the components of variogram are spherical functions. Similar as for minimum daily temperature the nugget effect of pure spatial component is higher showing that this model can not achieve a better accuracy than the model for mean daily temperature.

TABLE 4.3: Parameters of the fitted sum-metric variogram model for minimum daily temperature regression residuals, each component (see Eq. 4.9) is modeled using a spherical variogram function.

	Nugget	Sill	Range parameter	Anisotropy ratio
spatial	2.8722	8.314	4930 km	
temporal	0	0	0 days	
space-time	1.750	11.175	2117 km	527 km/day

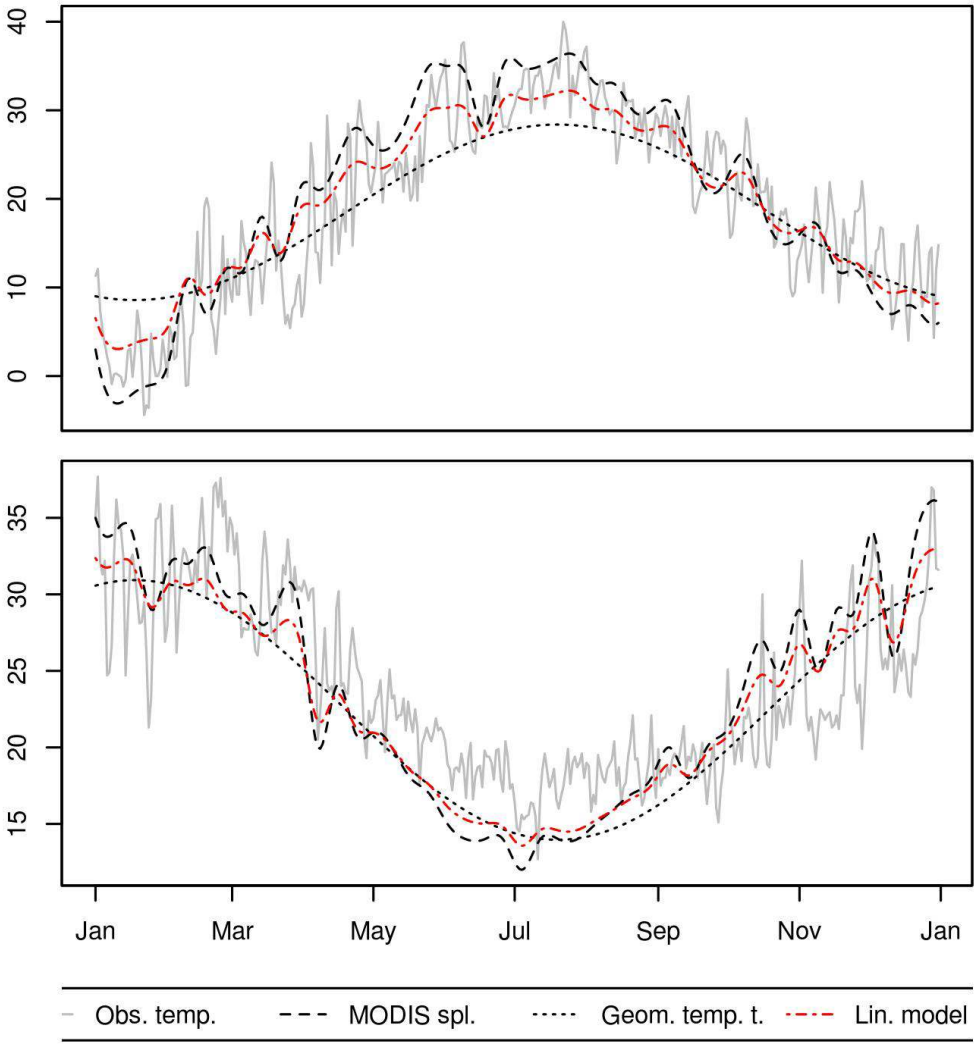

FIGURE 4.14: Maximum daily temperature observation in 2011 (gray solid line) and multivariate linear model of maximum daily temperature (red dashed line) on MODIS LST spline (black dashed line), geometrical temperature trend (black dotted line), elevation and topographic wetness index. PHILADEPLHIA, USA ($\lambda = -75, \phi = 39.993$) (top). BUNBURY, Western Australia ($\lambda = 115.65, \phi = -33.35$) (bottom).

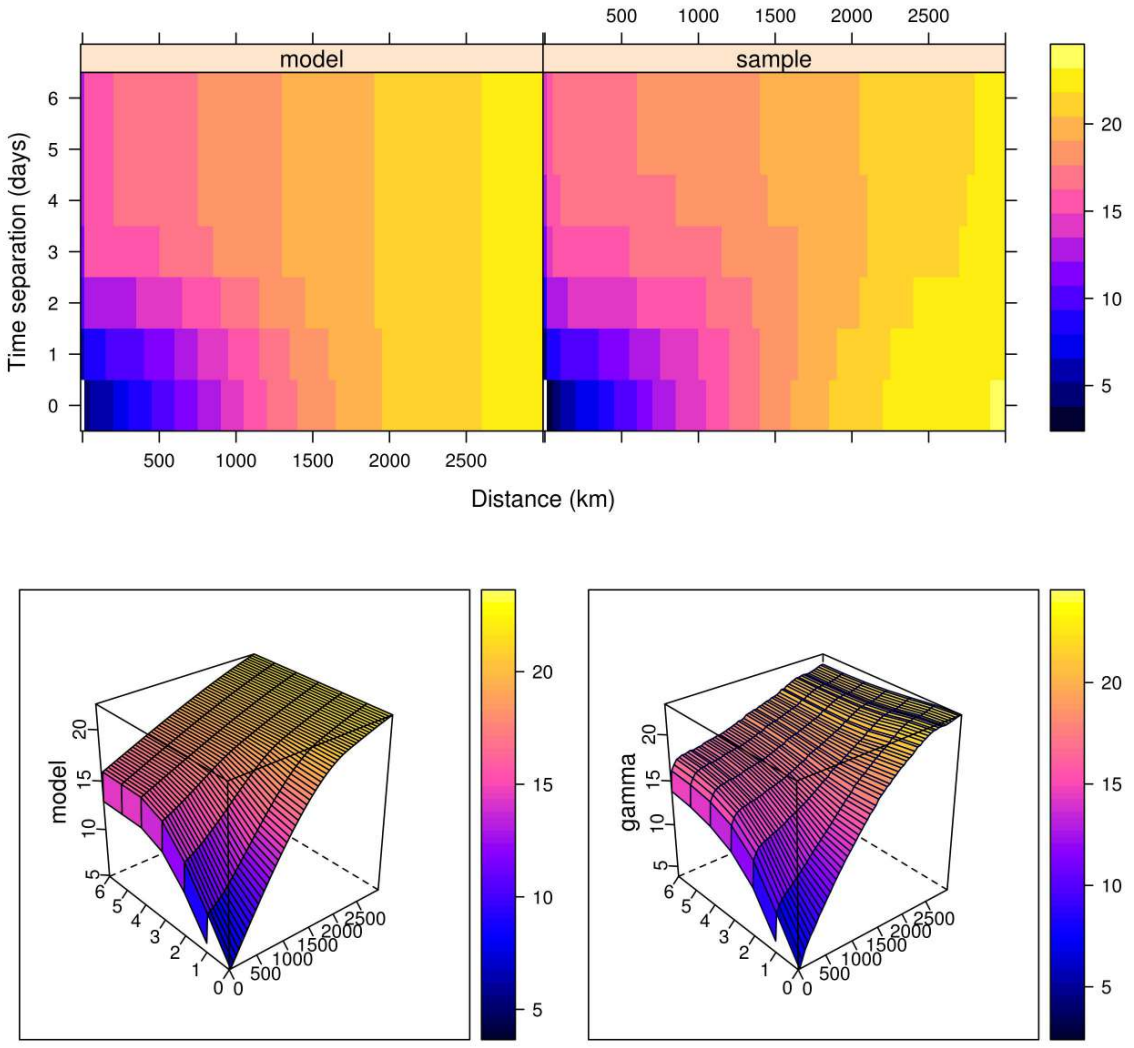

FIGURE 4.15: Fitted sum-metric model (left) and sample variogriom (right) of residuals from multiple linear regression of maximum daily temperature on MODIS, geometrical temperature trend, elevation and topographic wetness index. The variogram surface is presented in 2D (above) and 3D (below).

4.3.3.3 Accuracy assessment: maximum daily temperature

Results of cross-validation for maximum daily temperature on the complete data set gave a RMSE = 2.6C for global land areas including Antarctica, with R-square 95.9%. Monthly RMSE obtained from cross-validation of monthly aggregation of observation and cross-validation prediction is 1.9°C, and yearly RMSE is 1.6°C. Spatial distribution of RMSE calculated per station (yearly average of squared daily cross-validation residuals, daily quality measure) for each station is shown in Figure 4.16, where the stations with RMSE <2°C are represented with 41% of total number of stations (black dots), and 2°C< RMSE <3°C with 41% (blue dots), 16.6% of points are with 3°C< RMSE <6°C, and 106 stations are with RMSE >6°C.

The spatial distribution of RMSE also shows lower accuracy than for mean daily temperature. The aggregated results show that the best RMSE=1.3°C is achieved in the Netherlands, Europe without Russia (ca. 2.7°C) achieves around 2.1°C, USA 2.1°C, South America 3.2°C, Antarctica has the highest RMSE=5°C.

4.4 Discussion and Conclusions

In this paper we have demonstrated how dense publicly available ground station data together with a time series of remote sensing images and covariates at 1 km resolution can be used to predict mean, minimum and maximum daily temperature for the global land mass in space and time. The obtained global models for mean, minimum and maximum temperature were further used to produce gridded images of daily temperatures at very high spatial and temporal resolution. We achieved an average prediction accuracy of about 2–3°C for daily temperature prediction when assessed using cross-validation (which confirms the results of some local studies by Hengl et al. (2012), Heuvelink et al. (2012) and Neteler (2010)). This is promising as it indicates that highly accurate maps of daily temperatures can be produced at high spatial resolution using global spatio-temporal models. Figures 4.8, 4.13, 4.16 also show that the outliers are distinctly grouped in areas that are poorly covered with meteorological stations and in mountain regions, i.e. areas frequently covered with clouds or snow. This agrees with findings of Neteler (2010), who experienced similar difficulties in working with dynamic snow cover on mountain tops.

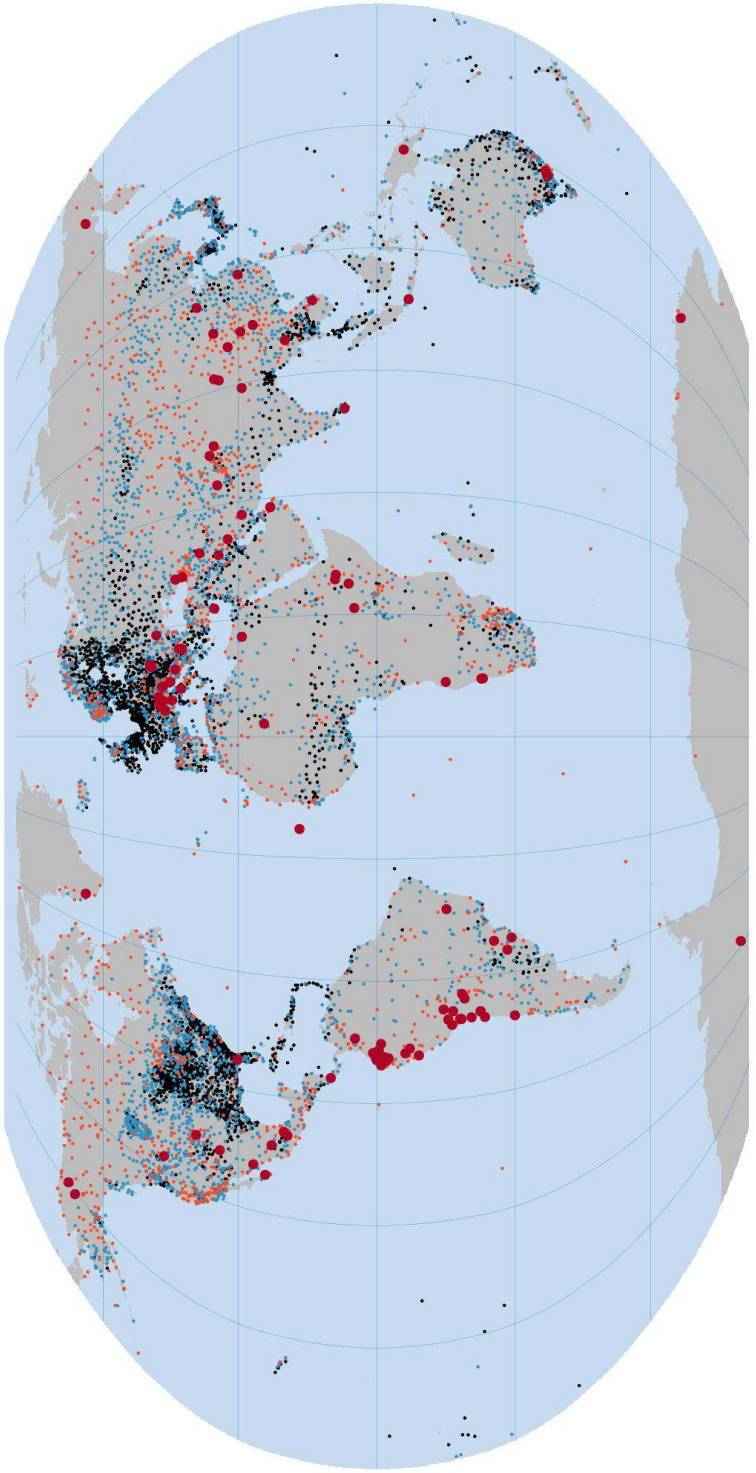

FIGURE 4.16: Map of maximum daily temperature cross-validation errors (RMSE) averaged per year for each station. Red circles indicate cross-validation outliers with RMSE > 6. Clusters of red circles indicate problematic areas, partly presence of gross error in observation time series.

During the model fitting, we discovered that the GSOD point data sets still contain many artifacts and possible gross errors. We removed a small portion of obvious errors, but surely there is even more noise in this data set. It was beyond the scope of this study to identify and remove all errors. Station data filtering should probably be performed by the organizations that collected the data because they have expert knowledge on the measurements and stations. In that context, our proposed methodology for cross-validation provides a tool to detect stations with potential errors in time series, and we recommend that responsible organizations use it to detect errors and clean up their data sets. Furthermore, by overlaying the point data and WorldGrids.org covariates, we were able to detect stations with inaccurate locations. This is especially important for stations in mountainous regions, which proved to be very important for model building as the error of predicting temperature increases with elevation.

It is worth noting that the presented global regression-kriging models can also be used to produce maps of associated uncertainty at very high spatio-temporal resolution in addition to providing estimates of the values of target variables. Basically, by using the global models presented in this paper, one can get an unbiased prediction of daily air temperatures for any place on the global land mask (at support size of 1 km) and for any day of the year for the period from the beginning of the MODIS mission until today.

The geometrical temperature trend (Eq.4.3) presented in this paper turned out to be a crucial covariate. Alone, it can explain more than 70% of the temperature variation. This indicates that a similar model without remote sensing images can be made for the daily temperature interpolation for the period when MODIS images were not available and would not perform much worse than a model that includes MODIS data as a covariate. The fitted spatio-temporal global models for mean, minimum and maximum daily temperature could be used as a tool for disaggregation of MODIS 8 day images to daily images and for the calibration of land surface temperatures (conversion to air temperature).

Chapter 5

Meteo package for automated spatio-temporal mapping

This Chapter describes the R package meteo that is under development. The package purpose is to provide functionalities for the automated mapping of meteorological observations using spatio-temporal regression kriging. The package contains regression and variogram models that were presented and described in Chapter 4. The models were fitted using publicly available data sets (see Chapter 3). Spatio-temporal regression kriging is implemented in a way that can be used for large amounts of data. Detection of outliers, which are based on iterative cross-validations, is also implemented in the package. In addition to the implemented methods, the package performance is presented through the case study of mapping mean daily temperatures in Serbia.

5.1 Introduction

The most powerful R package available for geostatistical analysis is gstat, which was developed for applied geostatistics Pebesma (2004). Many spatial geostatistics techniques (including ordinary, universal kriging, block kriging, kriging in a local neighbourhood, variogram cloud diagnostics, variogram modelling, multivariable variogram modelling,

cokriging and simulation)are available to the broad community of geoscientists. The techniques of spatio-temporal variogram fitting and implementation of global spatio-temporal ordinary kriging has recently been developed. Spatio-temporal regression kriging prediction and cross validation have not been implemented in gstat, yet.

The package meteo Kilibarda et al. (2013b)[1] has been implemented in the R environment for statistical computing (R Development Core Team, 2012). It combines functionalities of the gstat, rgdal (Bivand et al., 2013) and raster (Hijmans and van Etten, 2013) packages (raster data loading and analysis) and snowfall (Knaus, 2013) package (cluster computing). This package provides an automated framework for various tasks including: spatio-temporal regression kriging interpolation of ground based observations interpolation and de-trended using covariates, e.g. satellites products and DEM derivatives. Global temperature models are stored in the package (see Chapter 4 for models details) such that the prediction (interpolation) can be performed without fitting spatio-temporal regression and variogram models.

The automated spatio-temporal kriging interpolation procedure is a data driven approach designed for mapping with little or no human interaction. Hengl (2009) describes automated mapping, as an evolving technique that encompasses the future of mapping frameworks:

"We can conclude that an unavoidable trend in the evolution of spatial prediction models will be a development and use of fully-automated, robust, intelligent mapping systems. Systems that will be able to detect possible problems in the data, iteratively estimate the most reasonable model parameters, employ all possible explanatory and empirical data, and assist the user in generating the survey reports."

The meteo package endeavours in this direction and includes the additional paradigm of using a global model as the target meteorological/climatic variable. Currently, automated mapping with the meteo package can be decomposed in chunks:

1. defining input observations and covariates;
2. use of pre-calculated global models;

[1] https://r-forge.r-project.org/projects/meteo/

3. detecting and/or removing outliers;
4. creation of final prediction (and its export to GIS formats);
5. cartographic visualisation of results and/or creation of web maps (e.g. by using R package plotGoogleMaps (Kilibarda and Bajat, 2012) for automatic creation of interactive web maps).

In addition, meteo offers the possibility of using user defined covariates, regressions and variograms; thereby giving more flexibility of using the package in a semi-automated approach.

5.2 R enviroment and related packages

5.2.1 R enviroment

As stated in the Introductory section of the R Language Definition on-line manual (R Development Core Team, 2012), R is a system for statistical computation and graphics, which provides, among other things, programming facilities, high-level graphics, interfaces to other languages, and debugging facilities. R implements a language similar to the S language that was originally developed by John Chambers (Becker and Chambers, 1984). The main difference is in the license statement because, R is a free and open source software under the terms of the GNU General Public License in contrast to the S language. The syntax of the R language is analogous to the C programming language. However, a fully functional interpreter permits the creation of functions and calculations within an environment defined by a command line window or a graphical user interface (Grunsky, 2002). R is organized as a collection of packages designated for specific tasks.

The R package system has been one of the key factors in the overall success of the R project (R Development Core Team, 2012). The R contains the base system which enables statistical computation, linear algebra computation, graphics creation, and other similar features. A package is a related set of functions, help, and data files that have been bundled up together. Packages in R are similar to modules in Perl, libraries in C/C++, and classes

in Java. It is not necessary to install the specific packages if they do not part into the user's computing and analysing interests.

The set of developed packages are especially interesting for the meteo package. The R developers have written the R package sp to extend R with classes and methods for spatial data (Pebesma and Bivand, 2005). Classes specify a structure and define how spatial data are organised and stored. Methods are instances of functions specialised for a particular data class (Bivand et al., 2008). Another important package used in this study is the rgdal package. This package uses functions of the Geospatial Data Abstraction Library to read and write GIS data with options of handling a coordinate referent system (CRS). This package allows the user to define CRS for spatial object. CRS might be obtained directly from the data if data are imported from the GIS file by the rgdal package. Performing transformations among different CRSs is available using the PROJ4 library[2], which is implemented in the rgdal package. A very efficient tool for raster manipulation is the raster package (Hijmans and van Etten, 2013), which provides functionalities for reading, writing, manipulating, analysing and modelling of a gridded spatial data. Packages spacetime and gstat will be briefly described in further text.

5.2.2 Package spacetime

Spatio-temporal data have been used in meteorological/climatic mapping for a long time. However, they have not been defined as an object with structured spatial and temporal elements and bound together as a spatio-temporal data model. The examples of those data are: time series of weather measurements from ground stations in regions of interest, satellite images of weather, etc. Spatio-temporal data have mainly been used and analysed separately, whereby the spatial aspect is analysed first and the temporal aspect afterwards or reversed. Such data has not been included in an integral modelling approach (Bivand et al., 2008). The lack of GIS data models and software for storing, handling and analysing spatio-temporal data were the main reason for the described data processing approach outlined in this Chapter.

[2] http://trac.osgeo.org/proj/

Package spacetime provides classes and methods for different types of spatio-temporal data that are implemented in R. Spatio-temporal data types implemented in spacetime, include: space-time regular lattices, sparse lattices, irregular data, and simple trajectories (Pebesma, 2012). In addition, the utility functions for plotting data as map sequences (lattice or animation) or multiple time series and methods for spatial and temporal selection and subsetting and spatial and temporal overlay are provided in this package.

A *STFDF-class* is used in the meteo package for storing, overlaying and manipulation of spatio-temporal data. *STFDF-class* is a data model referred to as a full space-time grid. It contains observation data (stored as *data.frame* objects that are presented in analogue form as a spreadsheet in Excel), spatial features (points, lines, polygons, grid cells) as *sp* objects and time information data and time as a vector. This data model (class) implies that each spatial location contains observations for each time instance. Therefore, the number of observations is a product of the number of locations and number of time instances. Unobserved space-time locations (e.g. missing observation on certain day at meteorological station) are stored as a missing value *NA* in an observation table. This class is suitable for storing meteorological/climatic data from both ground stations and remote sensing data.

5.2.3 Package gstat

The gstat package provides a wide range of univariable and multivariable geostatistical functions for modelling, predicting and simulation, whereby the package sp provides general purpose classes and methods for defining, importing/exporting and visualizing spatial data (Pebesma, 2004). The package allows one to calculate sample variograms, fit valid models, show variograms, calculate (pseudo) cross-variograms, fit valid linear models, and calculate/fit directional variograms and variogram models (anisotropy coefficients are not fitted automatically).

The development of the spacetime package has already started in 2010 and the gstat function has been adapted for spatio-temporal mapping (Pebesma, 2013). This package can be used for spatio-temporal geostatistics, estimated sample spatio-temporal variograms, spatio-temporal variogram fitting and ordinary global spatio-temporal kriging.

The *variogramST* function calculates empirical spatio-temporal variograms using an object of *STFDF-class* as input data. The resulting variogram can be visualised as a surface (see Figure 5.1) or an image. Spatio-temporal model fitting is provided by the *fit.StVariogram* function. It fits a spatio-temporal variogram model of a given type from a spatio-temporal sample variogram. Different variogram models can be defined using the *vgmST* function. A variogram model (separable, product-sum, sum-metric, ect.) determines a structure of the space-time covariance model, e.g. the sum-metric structure is defined using the Equation 4.9 that contains pure spatial, pure temporal and spatio-temporal components.

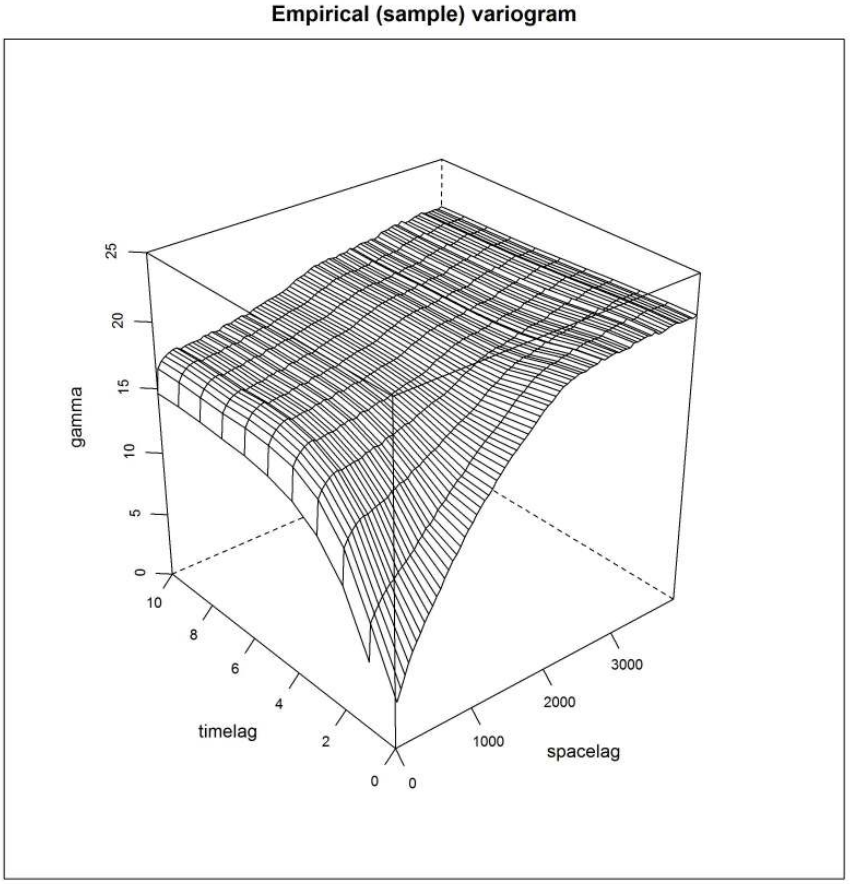

FIGURE 5.1: Spatio-temporal sample (experimental) variogram surface.

A function *krigeST* provides the particular implementation of global spatio-temporal ordinary kriging. At the moment, *krigeST* does not support block kriging or kriging in a local neighbourhood and it does not provide simulation.

The prediction over a large area with lots of spatio-temporal observations cannot be performed using the generic *krigeST* function because of the intensity of computation efforts. The alternative solution could be:

- dividing area of interest in smaller parts (tiles), as well as observations points;
- use *krigeST* for prediction and ;
- mosaic back tiles to area of interpolation.

Spatio-temporal regression kriging using `gstat` also must be performed separately: the first step isto produce predictions for the regression part using the regression function, the second part involves extracting residuals for all observations and finally fitting a global sum-metric variogram model. The residuals then should be interpolated and added to the predicted trend. For large areas and lots of observations, the prediction of residuals using ordinary kriging needs to be completed tile by tile. The `meteo` package automates the procedure for regression kriging prediction over large areas.

5.3 Development of meteo package

The `meteo` package contains several functions that aim to fit table data to spatio-temporal objects. Such fitting is necessary for acquiring a spatio-temporal kriging prediction. The spatio-temporal regression kriging function (*pred.strk*) is the most important part of the package. This function can perform predictions fully based on the *krigeST* function without any simplification of the kriging procedure. The use of this function for "large data" sets (even few hundreds of observations) need to be performed with tilling and simplified local spatio-temporal kriging with a fast neighbouring searching algorithm implemented in `meteo`.

5.3.1 Simplified searching algorithm for spatio-temporal kriging

Fast spatio-temporal regression kriging implemented in `meteo` applies a tiling procedure for prediction. The area is divided into tiles (smaller parts, see Figure 5.2)by the tiling

function, which is implemented in the meteo package. For each tile, the nearest spatio-temporal observations are selected according to distance from tile's centroids. Subsequently, spatio-temporal regression kriging estimates values within each tile on the base of nearest selected observations. Thus, within each tile, all estimates are calculated by using global kriging from previously selected observations. The procedure differs from traditional kriging in a local neighbourhood approach (which uses the neighbours observations while searching for an algorithm for each location) in that the number of spatial searches for nearest observations is reduced.

Figure 5.3 shows a tiling and searching algorithm in a graphical manner. For this explanatory example, ten nearest spatial locations are selected for the each tile. *Tiles* i, j are coloured in black and green and contain around 4,000 unmeasured locations points (regular grid at 1 km). The predictor function at any location from *tile i* uses 10 nearest observations in the space domain, and only two of these ten observations are not used in the*tile* j. In other words, *tile* j, uses 8 common observations for both tiles.

After tiling and selection of the nearest spatial locations, a prediction is performed for all target temporal instances. Accordingly, the procedure spares time spent in spatial searching because the prediction is performed for each of the time instances in a row and is based on the initial neighbourhood selections. Therefore, the full advantage of this approach is evident when the prediction is performed for longer periods of time (e.g. for month or year period). For example, the reduced number of local searches (just for one tile containing n_{points}) for a spatio-temporal prediction that is completed for a one year period, is defined as:

$$n_{red} = n_{points} \cdot n_{times} \tag{5.1}$$

where n_{red} is a number reduced searches, n_{times} is a number of time instances of the target spatio-temporal prediction and n_{points} is a number of spatial locations in one tile. The number of tiles should be defined and depend on multiple criteria, e.g. observation density over an area of interpolation, the number of points for prediction, the number of nearest observations that should be used for kriging, and target savings in computation etc.

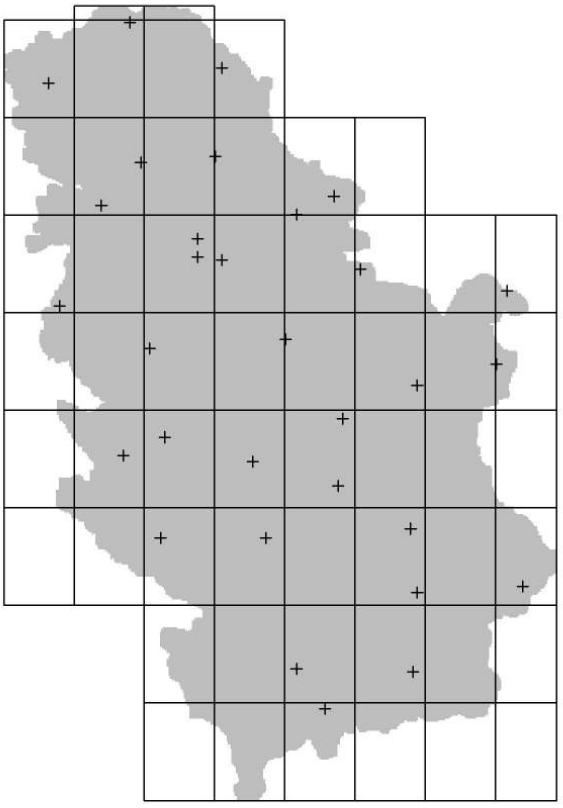

FIGURE 5.2: Plot of tiles over domain of interpolation over with observations.

The automated selection of an optimal number of tiles within the domain of interpolation is still open for question. For example, the Figure 5.3 depicts an area that is divided into 56 tiles for the territory of Serbia. Figure 5.2 shows all tiles and observations over tiles. It is clear that many tiles 'share' nearest observations and therefore potential artefacts in the edge line appear mostly due to the presence of outliers in observations or when observations are heavily clustered. The `meteo` package offers option for double tiling (two different networks of tiles) followed by averaging the results of predictions derived

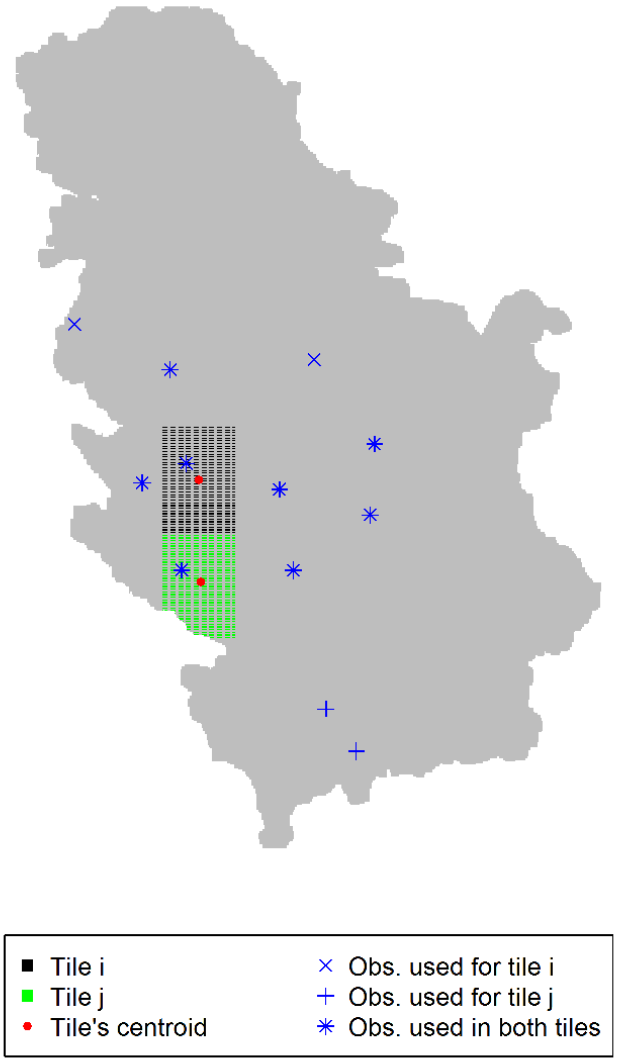

FIGURE 5.3: Plot of tiles together with accompaning observations used for spatio-temporal regression kriging.

from different tiling systems to avoid artefacts.

5.3.2 Outliers detection based on cross-validation

During the model fitting (in Chapter 3 and Chapter 5), it was discovered that the GSOD point data set still contain many artefacts and possible gross errors. The small portion

of obvious errors was removed based on the comparison of a cross-validation prediction with the observations. The example of outliers detected by cross-validation is given in Figure 5.4. Stations with RMSE averaged per year from daily residuals higher than 15°C are selected as potential outliers. Visual inspection shows that most of selected stations are obvious outliers, but the fact is that they also increase hardly RMSE of neighbours.

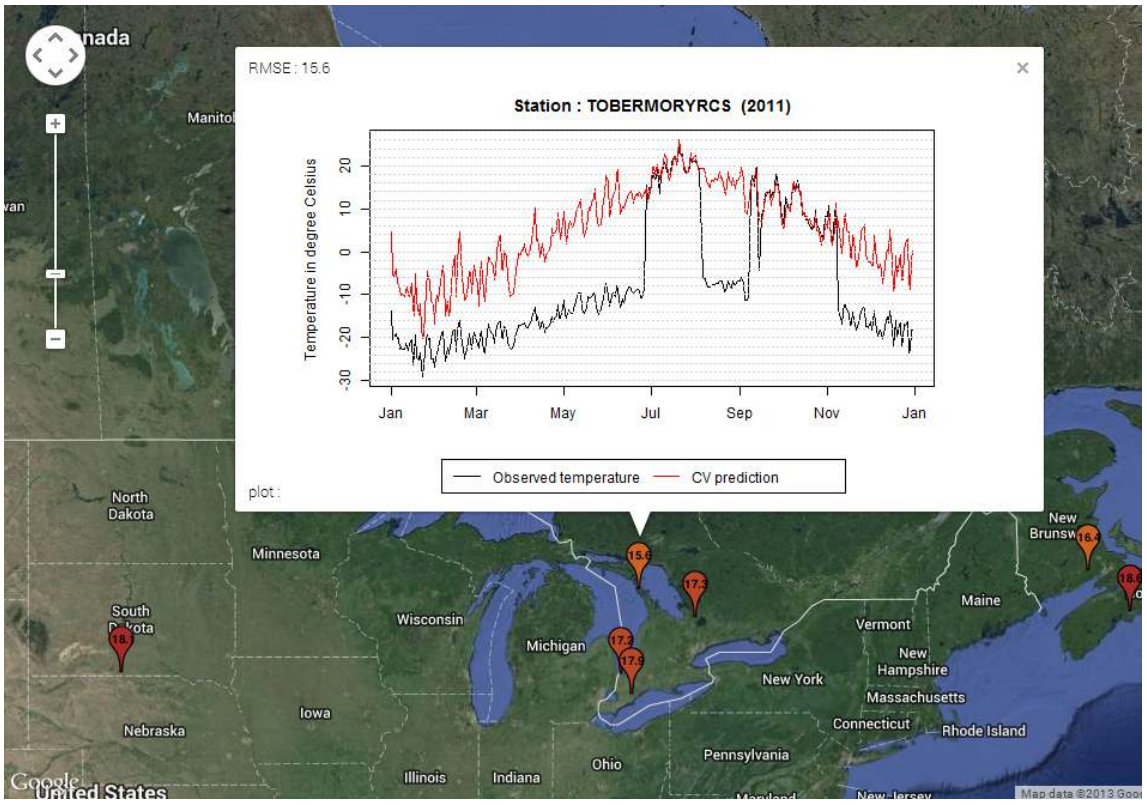

FIGURE 5.4: Exapmle of outliers detected based on cross-validation. Point labels show RMSE averaged per year from daily residuals. Map presents only sample of potential outliers from GSOD data set in 2011.

The outliers' detection algorithm firstly performs cross-validation and detects the station with the highest residual. If the residual is higher than the *a priori* defined threshold value, the station is removed from the dataset and new cross-validation is performed. Again, the station with the highest residual is compared against the threshold value and the iterative process is repeated while no more residuals exceed the defined threshold.

The described method is implemented in the `meteo` package (as part of the *pred.strk* function) and the package can perform detection and removal of outliers based on the defined

threshold. This method should be described in detail and should be tested with simulated and real data in further work.

5.4 Case study: Automated mapping mean daily temperature in Serbia

A collection of stations from GSOD and ECA&D data sets descried in Section 3.2 were used for mapping the mean daily temperatures in Serbia from 2011-07-05 to 2011-07-08. Observation data (for July 2011) are stored in the `meteo` package as table data (*data.frame*) for the purpose of demo examples. Corresponding spatial information are stored in the package as the same class. A function *meteo2STFDF* creates spatio-temporal objects from two data tables. The first table must contain at least three columns (attributes): time, station id and observation. The second table with station information must at least contain: station id and coordinates.

```
data(dtempc)
data(stations)
temp<- meteo2STFDF(dtempc,stations)
```

Covariates also need to be transformed into spatio-temporal *STFDF-class* objects. Covariates for Serbia (2011-07-05 to 2011-07-08) are stored in the package and contain two dynamic covariates (geometrical temperature trend, splined MODIS LST, see Chapter 4) and two static covariates DEM and TWI. Figure 5.5 shows a spatio-temporal plot of the splined MODIS LST over the domain of interpolation.

The geometrical-temperature trend is shown on Figure 5.6. Static covariates are shown in Figure 5.7.

Static and dynamic covariates used in this example were stored as one object of the*STFDF-class* and was named *regdata*. The following command produces spatio-temporal regression kriging prediction for the period between 2011-07-05 and 2011-07-08.

```
res= pred.strk(temp, newdata= regdata[,1:4], threshold.res=10 )
```

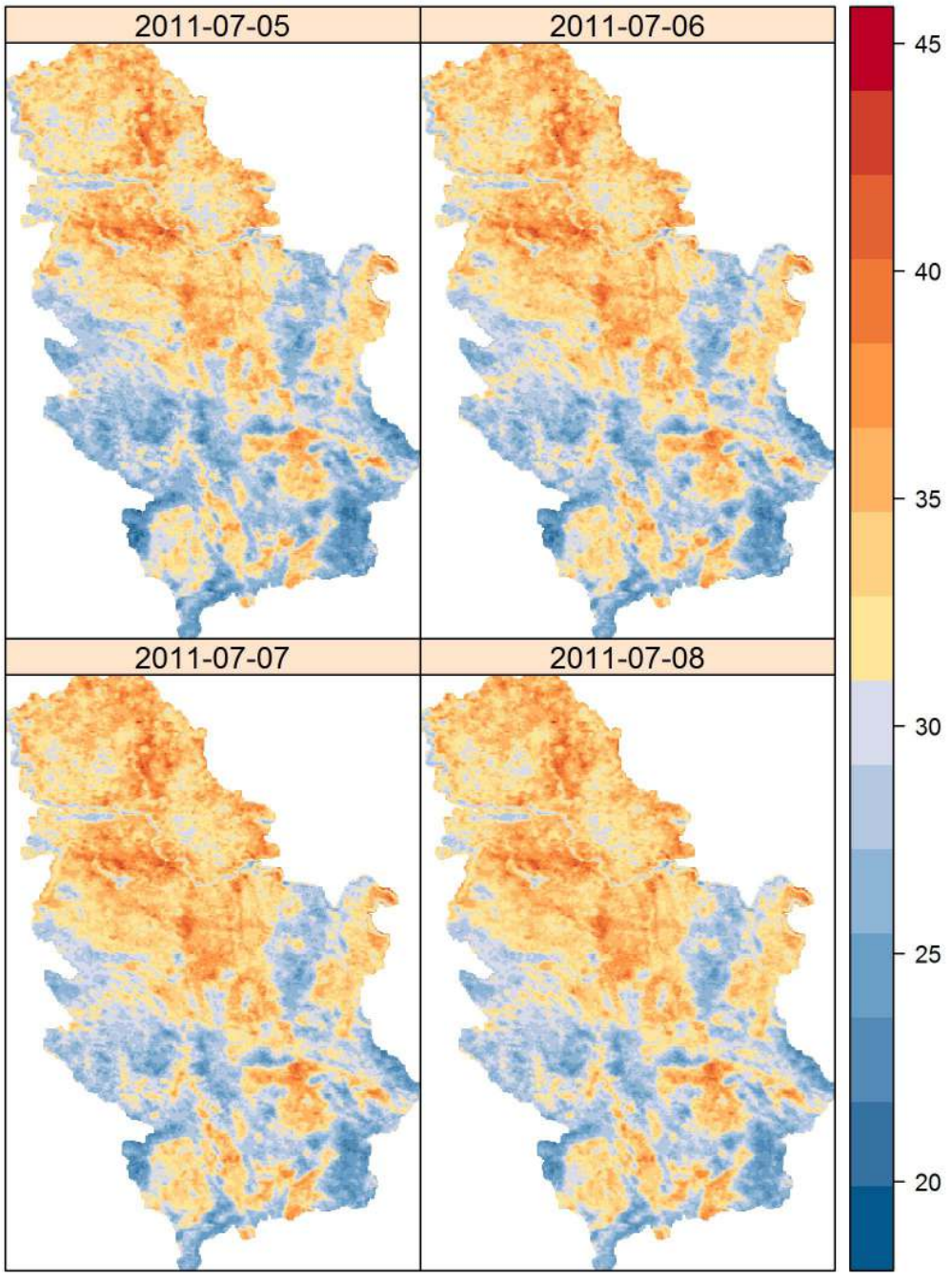

FIGURE 5.5: Splined MODIS LST 8-day images in Serbia (2011-07-05 to 2011-07-08).

The *temp* object contains spatio-temporal observations and *regdata* defines covariates and the frame for prediction. The prediction has been estimated for each spatio-temporal

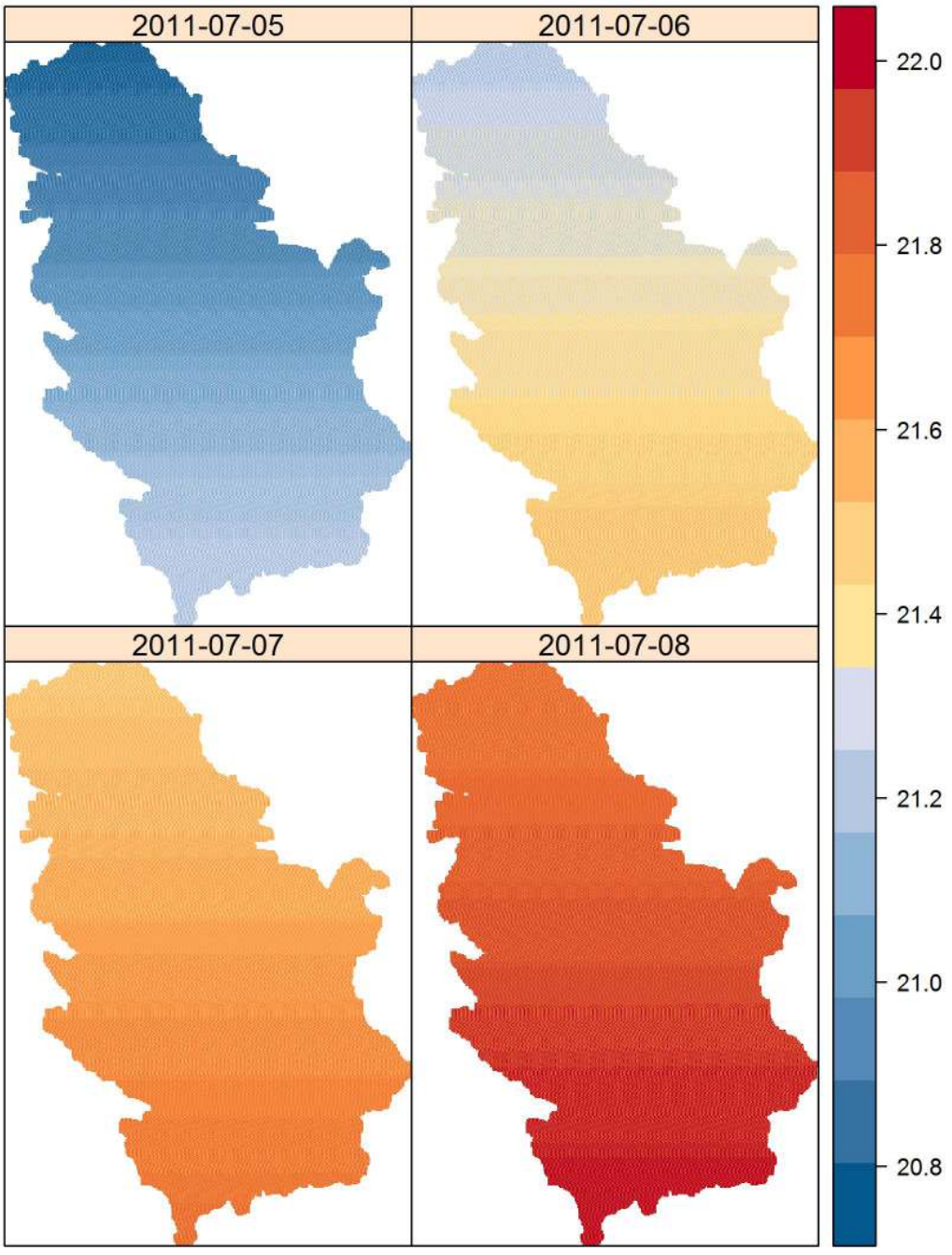

FIGURE 5.6: Geometrical-temperature trend in Serbia (2011-07-05 to 2011-07-08).

points defined in the*regdata* space-time grid. The global regression model fitted and described in Chapter 4 is presented as a variogram model for mean temperature and was

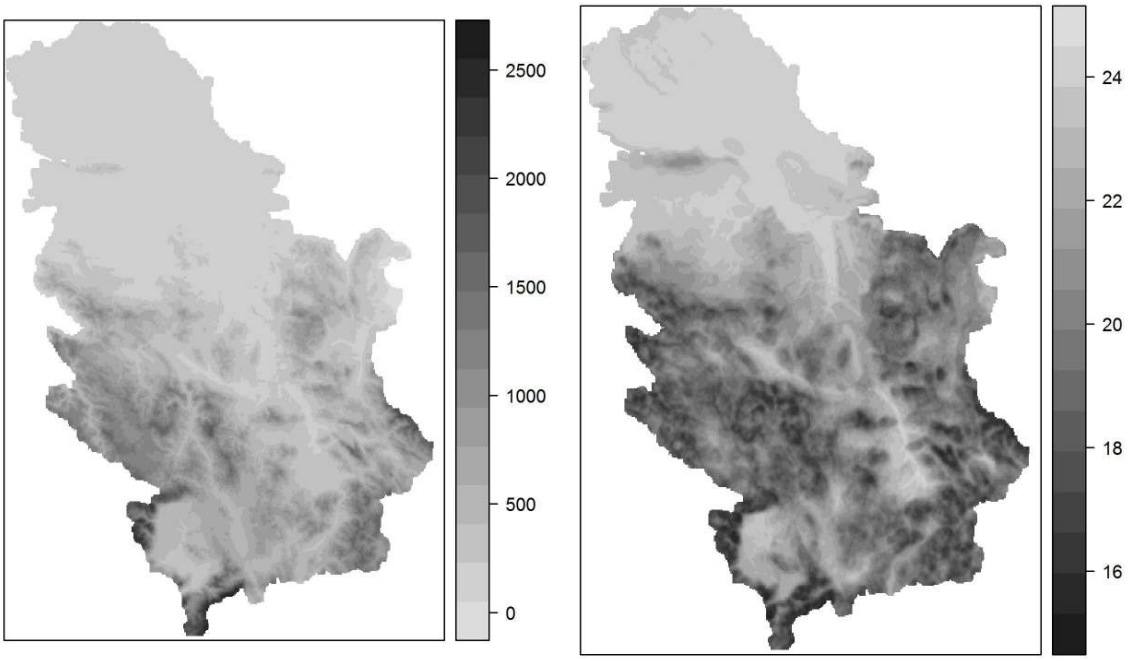

FIGURE 5.7: (left) Digital elevation model. (right) SAGA topographic wetness index

specified in the function as a default setting. The threshold value was specified with an argument *threshold.res*.

The resulting object *res* is a list of particular results:

- an object of *STFDF-class* with column contains prediction of mean daily teperature;
- cross validation information for points used in prediction;
- removed locations as spatial object, showing spatial locations of removed stations;
- removed locations with observations as an object of *STFDF-class*.

The prediction of mean daily temperature (Figure 5.8) was produced based only on the observations of 27 stations and the trend part was computed using covariates previously described that uses a regression model within the function. The total number of 27 stations was selected from the 35 stations that were available in the pool. A total of 8 stations were removed by function as outliers that were determined based on the defined threshold and iterative cross-validation process described in Section 5.3.2.

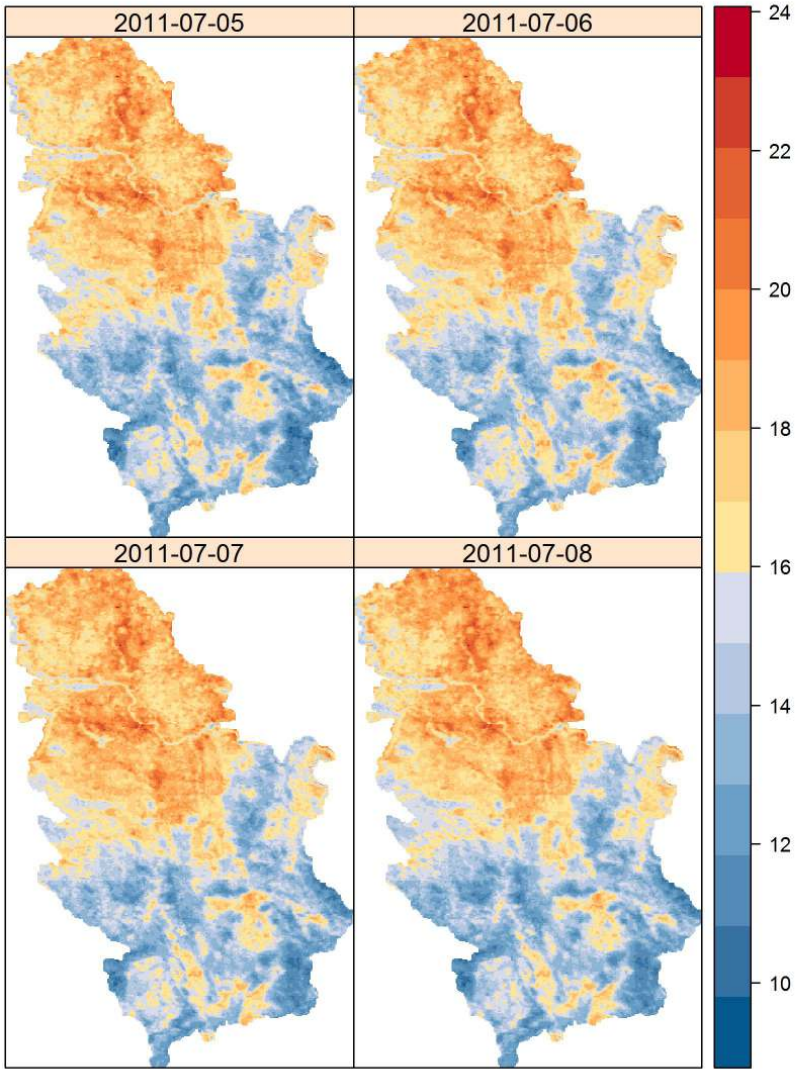

FIGURE 5.8: Prediction of mean daily temperature for Serbia (from 2011-07-05 to 2011-07-08) produced by automated mapping.

Detected outliers are showed in the spatio-temporal plot shown in Figure 5.9 and the multi-panel time series plot. The plot shows 7 out of 8 outliers because the station name "BELGRADE(OBSERVATORY)" contains measurement showing -47°C temperature that is obviously an error and would cause the *values* axis to be incorrectly scaled on the plot.

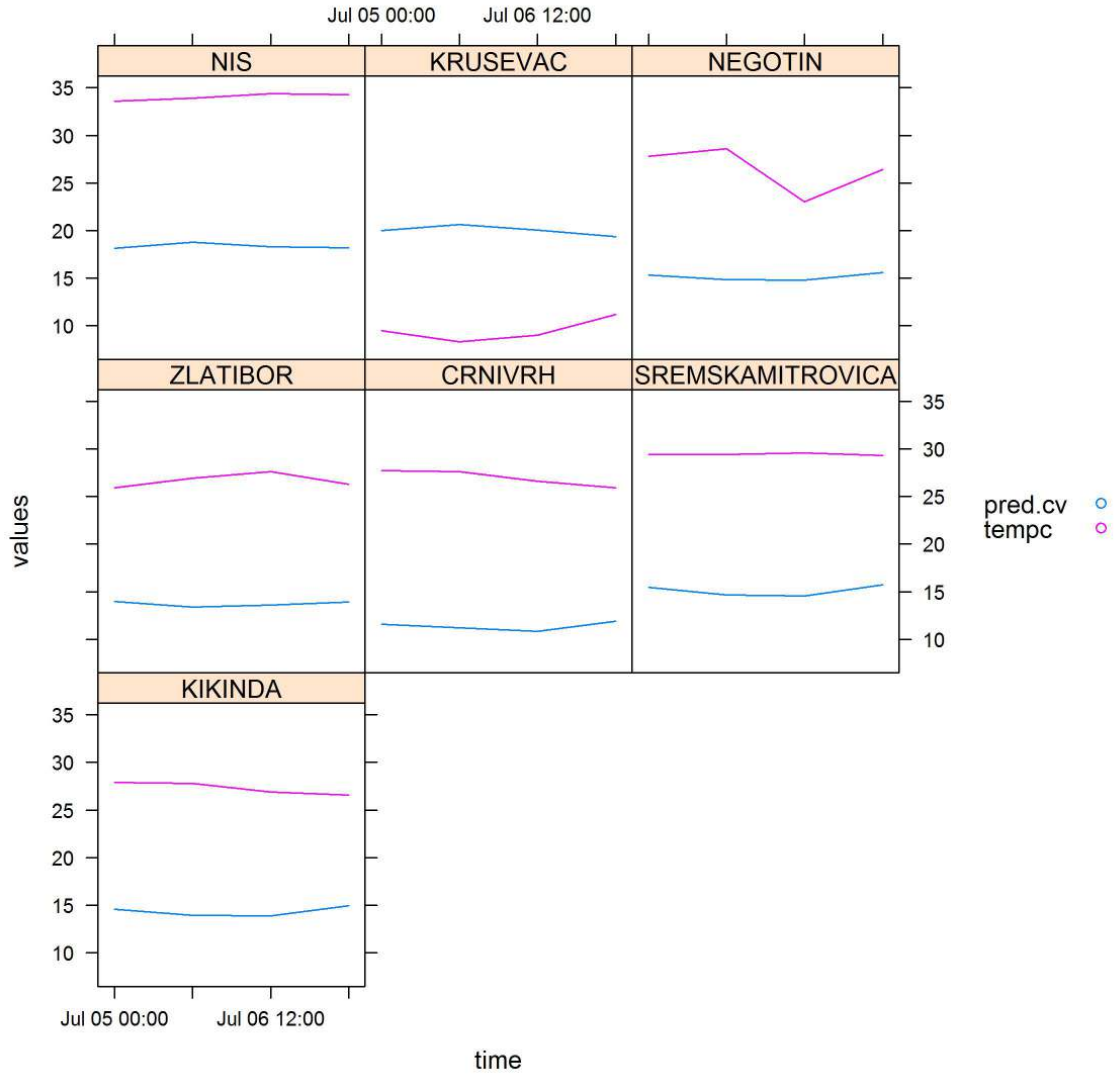

FIGURE 5.9: Outliers detected using detection based on cross-validation.

The detected outliers were removed but could be analysed individually. For example, the first station from the Figure 5.9 entitled "NIS" is obtained from the GSOD data set and the same station obtained from the ECA&D data set has around 15°C lower mean daily temperatures(see Figure 5.10). The lower temperatures from ECA&D are more correlated with covariates than the rest of the used observations. Therefore, the station from GSOD that was from the same location was detected as an outlier and presented residuals around 15°C, see Figure 5.9.

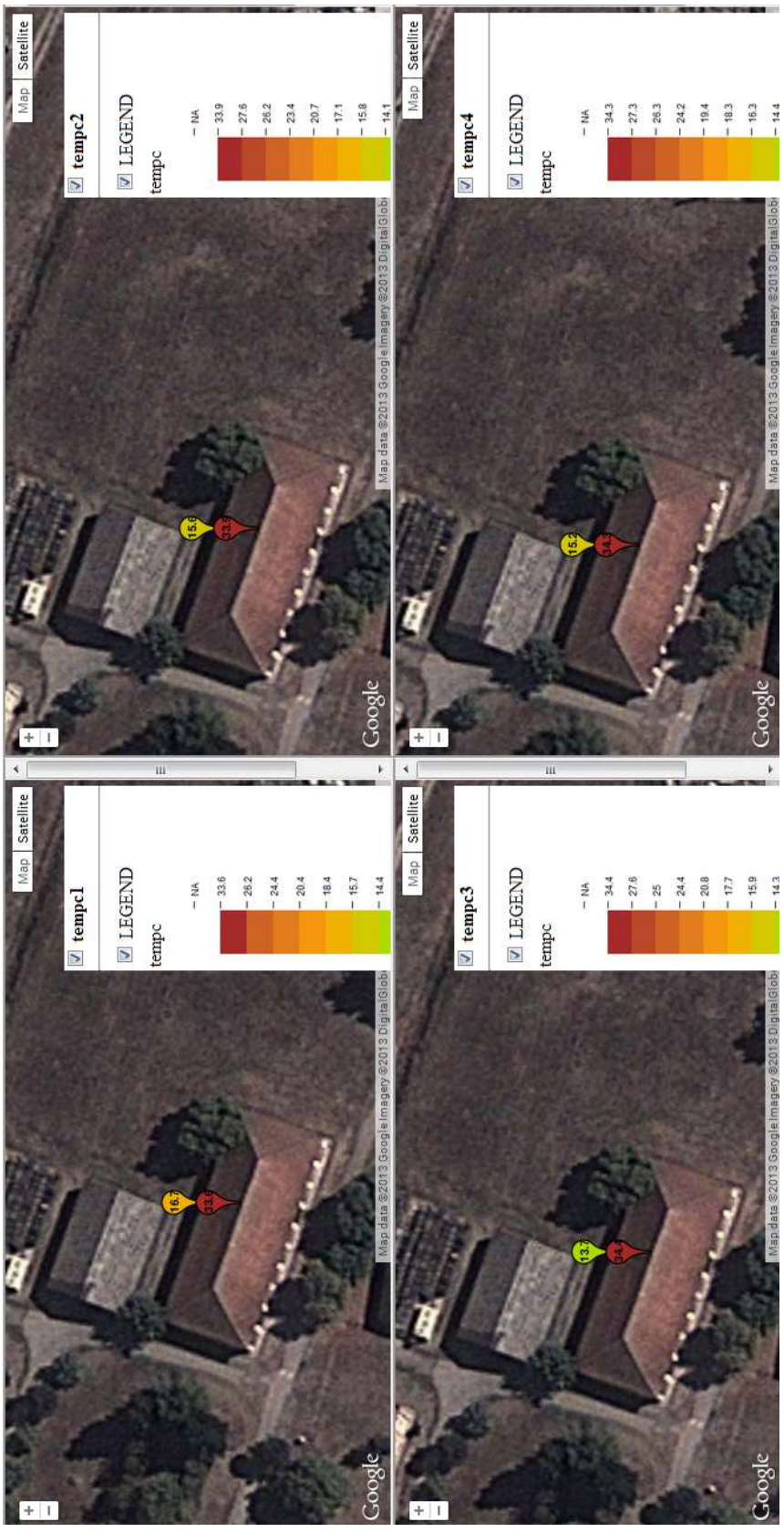

FIGURE 5.10: Removed station "NIS" from GSOD data set (redish) compared with same station from ECA&D data, for period from 2011-07-05 to 2011-07-08.

5.5 Discussion and conclusion

The mapping framework described in this Chapter enables the use of spatio-temporal regression kriging for meteorological mapping. The Implementation of a fast searching algorithm provides an advantage in computing when completing interpolations over a large spatio-temporal grid. The advantage is especially noticeable if the grid of points contains larger time series (e.g. predictions made for the area of interpolation over a year period where each location contains around 365 observations).

The automated mapping framework presented herein is still under development and a lot of functionalities need to be implemented in the future. There are still many open questions related to a) an optimal number of tiles for the domain of interpolation, b) the choice of an optimal threshold for the detection of outliers, and c) incorporating a function for downloading ground station observations from data providers. Likewise, the development of procedures for downloading and mosaicking remote sensing imagery and their organisation in an appropriate space-time object would be useful for many meteo/climatic applications.

Filtering missing pixels in MODIS LST 8-day images through the use of spatial splines also needs to be implemented in the package. Similarly, temporal disaggregation from 8-day images to daily images using splines (in the temporal domain) might be offered as an automated procedure.

Automated mapping using a global model incorporated in the mapping framework is a new approach in the automated mapping field. The global model should be iteratively improved with increasing availability (and/or quality) of observations both from ground stations and/or from remote sensing data. Therefore, global modelling of processes (that can be modelled with spatio-temporal kriging) could be performed similarly by storing the global model within automated mapping frameworks.

Chapter 6

Spatio-temporal visualisation of meteorological data using plotGoogleMaps [1]

Google Maps are increasingly used for communication throughout many map-based services and are often embedded on third-party websites via the Google Maps API. The main objective of this study is to develop a solution for the easy creation of an interactive web map, with a base map supplied by Google, where all map elements and additional functionalities are handled by just one line of code. The present solution for the automatic creation of a complete web map is the R package that is based on the Google Maps API, plotGoogleMaps. This tool provides a new interactive plot device for handling the geographic data for web browsers. It also offers a complete map in the HTML format, which has become a regular medium for cartographic communication. HTML as a multimedia medium gives new possibilities in the visualisation of spatial and spatio-temporal data. The tool plotGoogleMaps is developed in the R software language and is designed for the automatic creation of web maps. This chapter discusses plotGoogleMaps applications in meteorological spatial and spatio-temporal mapping.

[1] Mostly based on article: Kilibarda M, Branislav B (2012) plotGoogleMaps: The R-based web-mapping tool for thematic spatial data, *Geomatica* , 2012, 66, 37-49

6.1 Introduction

Although the Internet has been in existencesince the late 60's, the widespread use of the World Wide Web (Web)was estabished in the mid-90âĂŹs and shortly after has become a foremost medium for cartographers (Peterson, 2007). The real tipping point in the usage of geographic information on the Web was the year 2005. In June 2005, Google released the Google Maps Application Programming Interface (API), which allows a combination of geographic information from a variety of sources and formats. One of the most important capabilities of the API is the generation of mashup maps, which is the product of the combination of geographic data from one source with a map from another source (Miller, 2006; Haklay et al., 2008; Gartner, 2009). The mashup maps are easy to create and can be implemented in any web page for free and without any technical specification and requirements whatsoever, thereby resulting in an increased web mapping popularity. This progress, together with the popularity of web mapping and its application, is clarified by Haklay (Haklay et al., 2008):

"These rapid developments in web mapping and geographic information use are enabled and facilitated by global trends in the way that individuals and communities use the Internet and new technologies to create, develop, share and use information (including geographic information), through innovative, often collaborative, applications."

This change in direction of Web philosophy from communication media to contribution media is named Web 2.0. The Web 2.0 term and concept was first coined and described by Tim O'Reilly in 2005 [2] and later in publication (O'Reilly, 2007).

The collaborative nature of the Web 2.0 environment allows data production to be shared among many individuals (Feick and Deparday, 2010). Goodchild (2007) described the term 'Web Mapping 2.0" as an important part of the Web 2.0 concept. Integration and visualization of different geographic information on base maps (such as Google Maps/Earth, Virtual Earth, or Yahoo Map), is the core of 'Web Mapping 2.0'. The most significant part of Web Mapping 2.0 corresponds to Google Maps/Earth services. Google Earth/Maps is

[2] `http://oreilly.com/web2/archive/what-is-web-20.html`

ground-breaking software that excelled in at least five categories: availability of application, high quality background maps, single coordinate system, web-based data sharing, popular interface and availability of API services (Hengl, 2009).

Google Maps API has encouraged a considerable number of users, with intermediate and advanced programming knowledge, to build their own applications using Google Maps data as visualization interfaces (Gibin et al., 2008). According to BuiltWith Trends statistics [3] the number of Google Maps websites usage was over 800,000 and, comprised mainly of thematic cartography.

A thematic map displays the spatial pattern of social or physical phenomena such as population density, life expectancy or climate change. Thematic mapping has a long history in geography and one important part of presenting thematic data is the provision of high quality base maps that allow integration into the Google Map interface through the Google Maps API. The London Profiler (Gibin et al., 2008) presents geographic information as a series of choropleth maps on top of Google Maps. This example is an exception from most mashups because they mostly display spatial point data (push pins).

The existing solution of using a Google Maps image as a background for plotting spatial data is the general concept of the RgoogleMaps R package (Loecher, 2013) that is based on Google Static Maps API. The Google Static Maps API [4] allows the embedding of a Google Maps image in the user's webpage without requiring JavaScript or any dynamic page loading. This package provides static maps without interactive tools such as data pan or zoom control and with a constrained quality of the Google background map. The maximum zoom level, provided by Google Static Maps API, concurs with the maximum size limitation of 640 x 640 pixels.

The other package with similar functionalities, which provides an interface between the R and the Google Visualisation API, is called the googleVis (Gesmann and de Castillo, 2011). The Google Visualisation API offers interactive charts that can be embedded into web pages. The googleVis package contains options to produce map mashups based on Google Maps API. The input data for the package is the data frame with marked columns

[3] `http://trends.builtwith.com/websitelist/Google-Maps`
[4] `https://developers.google.com/maps/documentation/staticmaps/`

that are related to location information. This is a typical package used for handling spatial data and their visualization.

The objective of plotGoogleMaps (Kilibarda, 2013) is to provide a solution for the easy creation of an interactive web map, with a base map supplied by Google, where all map elements and additional functionalities are handled by just one line of code. The obtained result is an interactive map rendered in a web browser. The automatic creation of a complete web map, which is based on the Google Maps API (the HTML file with CSS styling and Java Script functionality) , is provided by the R package plotGoogleMaps . The package provides a solution to create and visualize vector and raster data, to map features, plot choroplet maps, and include proportional symbols. The version of package 2.0 is extended to accommodate the visualisation of spatio-temporal classes in the form of afull space-time grid (*STFDF*) and the visualization of unstructured spatio-temporal data (*STIDF*), for more datails see Pebesma (2012). The growing popularity of the R language was the driving force behind the development of the plotGoogleMaps tool, particularly among academic and expert communities and especially in the field of spatial data analyses. The goal of the presented work is to adopt and apply map design principles in order to create mashups and to focus on the minimization of coding and scripting. This would enable the creation of mashups based on Google Maps without any Internet programming knowledge. Thus, the creation of web maps becomes a plot facility for R users. The web maps created by plotGoogleMaps package could be used as a temporary result for spatial visualization (spatio-temporal) generated on local machines or published on any web page. The next section contains a brief description of theoretical and technical background for the development of the plotGoogleMaps package, Web 2.0 and AJAX, Google Maps API. In addition to this section, there is a description of package functionalities and an explanation of how these functions were programmed. Details concerning the source code and instructions on how to get this package are also included. The paper continues with examples of practical package applications in meteorological/climatic mapping.

6.2 Package plotGoogleMaps and underling web technologies

6.2.1 Web 2.0 and AJAX

Best (2006) defined the main characteristics of Web 2.0, and in a few words are listed as: a rich user experience, user participation, dynamic content, metadata (tagging, as semantic enrichment), web standards (e.g. W3C[5] recommendations) and scalability. Examples of Web 2.0 applications embrace social networking sites, video sharing, wikis, blogs, etc. The Web 2.0 concept is based on the AJAX technology.

Asynchronous JavaScript and XML, or shorter, AJAX (Schutta and Asleson, 2005) is the name given to a set of modern web application development technologies that were previously known as the dynamic HTML (DHTML) and remote scripting. The fact that this web application has a similar speed to standard desktop applications makes up the core benefit of this technology. Traditional web pages, created in HTML, were static, non flexible, and hardly adopted for any dynamic content. AJAX, on the other hand, as the base of Web 2.0 concept, was built for use in dynamic, interactive, and efficient web pages with high performance. Web pages without AJAX were slow; user interaction with the website required significant web server-side resources such that the server needed to send a complete web page if just one part of the web page was changed and sent back to the user. If the user drew a point on the web map, the server would send back the whole redrawn map with the new point icluded. AJAX brings a new concept such that the user interaction is left to a user's computer, which works only with a changeable part of the web page. Therefore, even if the user operates with one point only, it would not be necessary to redraw the whole map again. AJAX-based geographical applications significantly improve the usability of web mapping (Skarlatidou and Haklay, 2006; Haklay and Zafiri, 2007; Haklay et al., 2008; Kilibarda et al., 2010). Apart from AJAX, APIs have also influenced web mapping strongly. API is a set of routines, protocols, and tools for building software applications. The most popular web mapping APIs are: Google Maps API, Yahoo! Maps API, Microsoft Virtual Earth API, AOL MapQuest API (Gartner, 2009).

[5] urlhttp://www.w3.org/

6.2.2 Google Maps API

Google Maps API is a set of predefined JavaScript classes that are designed for embedding the Google Maps site into an external website. The resut of this process is such that additional geographical data could be overlaid over a basic Google Map. These results are possible to realize even if the creator is not an expert in web programming, although basic knowledge in JavaScript programming language, XML, Ajax and XHTML is required. Google Maps API, compared to the relative complexity of Open Geospatial Consortium (OGC) standards, is much easier for implementation. Google Maps API provides mapping functionality and high-resolution background data, but map mashups implementation still requires some web programming knowledge. It enables a combination of geographic information from a variety of sources and formats. GIS data objects, such as vectors, points, polylines, polygons or raster are represented in the mashups as Google Maps API Java Script objects. For that reason, it is necessary to transform the GIS data into Java Script objects that are appropriate to Google Maps API. These objects could be cartographically represented with point symbols, (although with some limitation) . Polylines and polygonâĂŹs line representations could be defined with outline width, color, transparency and fill color for the polygon area. Transforming the GIS data into Google Maps API objects and defining a cartographic representation for every single object is time consuming and rather difficult, especially for someone who has no experience in web programming. The developed R package, plotGoogleMaps, offers an easy creation of mashups as local files or files ready to be published on the Web.

6.2.3 Development of plotGoogleMaps

The newly developed R package, plotGoogleMaps, based on AJAX and Google Maps API service, produces HTML file map mashups (web maps) with Google Map high-resolution background data and additional data layers. The first version of the package was developed in 2010 and the current 2.0 version was published in 2013 (Kilibarda, 2013).

The package depends on two packages for spatial and spatio-temporal data handling, sp (Pebesma and Bivand, 2005; Bivand et al., 2008) and spacetime (Pebesma, 2012).

Another package significant for the automatic cartographic re-projection within plotGoogleMaps is rgdal R package (Bivand et al., 2013). These sets of developed packages are especially interesting for geoscientists. The R developers have written the R package sp to extend R with classes and methods for spatial data (Pebesma and Bivand, 2005). Classes specify a structure and deïňĄne how spatial data are organized and stored. Methods are instances of functions, specialized for a particular data class (Bivand et al., 2008). The package rgdal provides functionalities for importing and exporting the most popular formats of GIS data in R. This package uses functions of the Geospatial Data Abstraction Library [6] to read and write the GIS data and includes options for handling a coordinate referent system (CRS). With the rgdal package, users can optionally define a CRS or inherit it from the input data as well as perform data transformations between different CRSs using the PROJ4 library [7], implemented as part of rgdal. The imported GIS data are converted into sp objects and are used for handling the vector and raster (grid) data in R. These functionalities enable the use of a very large amount of the GIS data format as an input for the plotGoogleMaps, previously read via rgdal. Input data, sp or spacetime object with defined CRS, is the only mandatory argument in the plotGoogleMaps functions. Hence, different GIS formats of input data are read in R and are afterwards based on a predefined visualization method; those data are mapped as web map (Figure 6.1).

The plotGoogleMaps contains functionalities from PROJ4 library, which performs coordinate transformations from source CRS to WGS84 CRS that is used for spatial data handling by Google Maps. Google Maps API allows for additional spatial data handling, in the form of XML, KML, and GeoRSS, but some visualization functionalities, as well as interaction with attribute data in the form of Google Maps API *InfoWindow* object, and similar, are difficult to be controlled. Another solution is to use the data in the form of predefined JavaScript classes of vector data primitives; point, line and polygon data and raster overlay. This approach is also implemented in plotGoogleMaps. It means that every single primitive is separated from the spatial object and its geometry is translated into a JavaScript object. Attribute data for every single feature is converted into a JavaScript *InfoWindow* object; its activation is available by clicking on the related feature on the produced web map. Additional visualization options supported by Google Maps API objects such as outline width, color, and transparency can be specified in plotGoogleMaps

[6] `http://www.gdal.org/`
[7] `http://trac.osgeo.org/proj/`

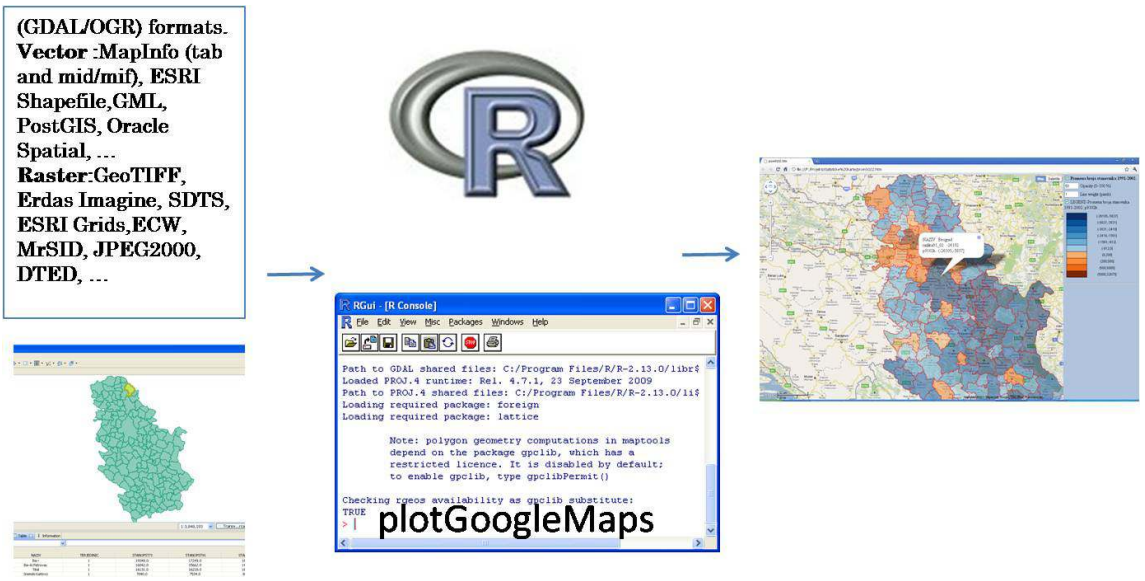

FIGURE 6.1: Simplified workflow for web map production by using plotGoogleMaps.

functions. The visualization of mandatory parameters is easy to set and is achieved by using optional arguments in plotGoogleMaps functions. Therefore, plotGoogleMaps writes object arguments in every JavaScript object and in the final HTML file. Google Maps API provides the majority of Google Maps utilities including pan, zoom, background layer control, and scale bar. Map utilities are controlled by optional arguments in plotGoogleMaps functions. Similarly, map width, background color, layer name, legend name, default background map, etc., can be set by using optional arguments in the plotGoogleMaps function. Some advanced utilities for interactive controls of additional layers are provided by defaults in plotGoogleMapsincluding: layer appearance, line width, transparency, and legend colors display control. Spatial data, with visualization parameters and utilities, are written in the HTML file with JavaScript and CSS elements. Thus, in the Rlanguage, plotGoogleMaps with only one function, and with few arguments, may produce many lines of codes in few languages for Web programming (Figure 6.2).

The map mushup (6.2- right bottom window) is produced by the R command:

```
> plotGoogleMaps(meuse, filename='myMap.htm')
```

The command contains two arguments: (1) the first is the set of spatial data, named Meuse, that contains 155 points with 12 soil properties in the form of attributes,(2) the second is

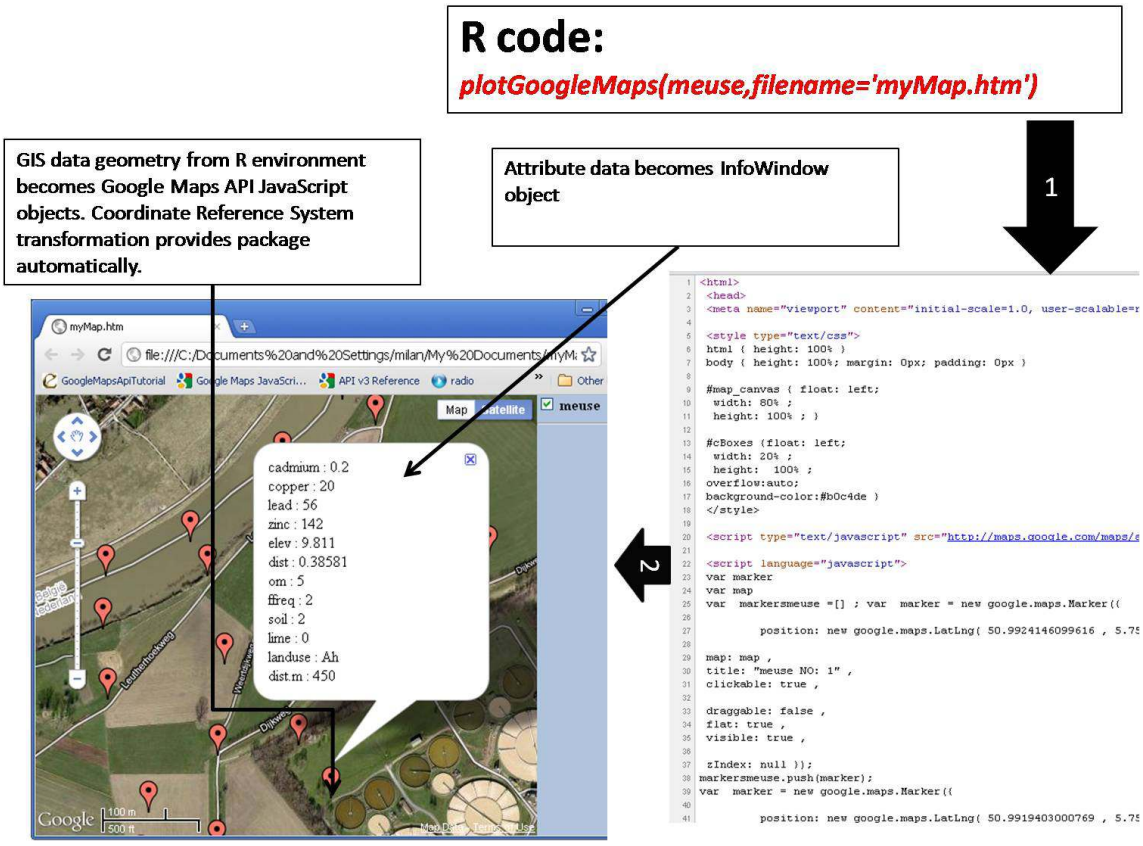

FIGURE 6.2: Plotting vector point data. Just one line of R code substitutes many lines of HTML with JavaScript and CSS code.

an optional argument with a file name of the output map mashup. This function contains many optional arguments. For example, the description of the function arguments may be provided by using the "help" function in R or can be found on the package web page[8] or in the reference manual. The same web page contains the package source; the source code is open and available under the GPL license. Technical details about the used solutions in the package could be obtained from the source code.

Generally, the idea of package implementation is denoted as the automatic production of HTML map mashups using spatial or spatio-temporal data objects. The R command used in the previous example produces an HTML file from the Meuse data. The output map also contains CSS elements and JavaScript scripts. The control of CSS elements is available through the use of optional function arguments in order to set the dimensions of

[8] `http://cran.r-project.org/web/packages/plotGoogleMaps/index.html`

a map area. Since the optional arguments are not set in this example, the resulting HTML file contains default CSS styling settings. The rest of the produced HTML is JavaScript that contains Google Maps API and a set of JavaScript functionalities related to layer control options. The Meuse data was transformed from native CRS to WGS 84 and every single point was translated to a Google Maps Marker object, i.e. JavaScript object, used in Google Maps API. The base map is set, by default, to be a Google hybrid map where the initial zoom and central points of the base map depend on the Meuse bounding box. The attribute data, for every single point separately, are converted to *google.maps.InfoWindow* objects, with an associated *listener* function included to handle the click event on the marker.

The next section contains examples focusing on the package application in meteorological/climatic visualisation.

6.3 Implementation and applications

The functionalities of the package that are used for the production of web maps of climatic spatial and spatio-temporal data are presented through the following case studies - examples. The applications illustrated in this paper pertain to different studies concerning spatial and spatio-temporal data analysis.

6.3.1 Spatio-temporal visualisation of climatic variables

The package plotGoogleMaps plots spatial objects of *sp* and *spacetime* classes over Google Maps using very simple syntax and only one or a few lines of R code. Figure 4.8 shows results of an accuracy assessment presented in Chapter 4. The map of mean daily temperature cross-validation errors (RMSE) is averaged per year for each station. Daily residuals should be very interesting, but the traditional way to represent daily residuals associated with a multi-panel plot (365 plots for year 2011) and space-time cross section plot (e.g., space on the x-axis and time on the y-axis) or animated plot. The traditional approach for daily residual visualisation is cumbersome to be both informative and intuitive at the same time. Thanks to the multimedia nature of the HTML, the package plotGoogleMaps

uses traditional spatial plots together with a time series multivariate plot implemented in zoo (Zeileis and Grothendieck, 2005) package, see Figure 6.3.

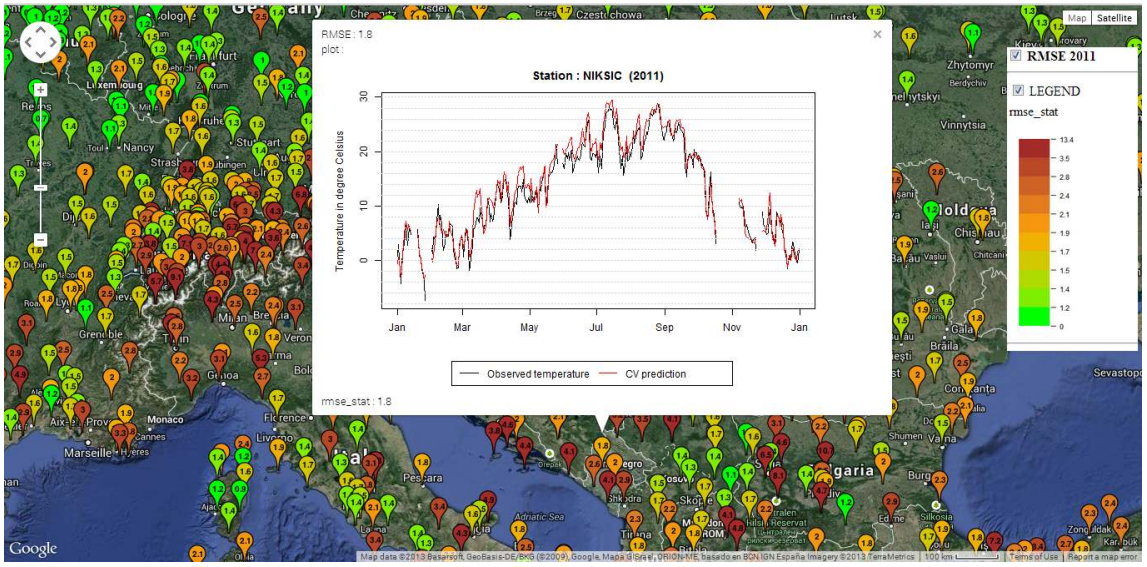

FIGURE 6.3: RMSE map. Space-time regression kriging of mean daily temperature observations on 8-day on MODIS 8 day images, topographic layers (DEM and TWI) and a geometrical temperature trend; `http://dailymeteo.org/`

In multi-panel (lattice) plots, panels share x- and y-axis, a common legend, and the strip above the panel indicates what the panel is about (Bivand et al., 2008). Similar, plot is implemented in plotGoogleMaps, as well, giving the possibility to visually compare data from few time instances. The interactive nature of the produced map provides an opportunity to inspect additional attributes by opening more than one *google.maps.InfoWindow* simultaneously. A Multi-panel plot of temperature data in Serbia for 2011-07-05 and 2011-07-06 is provided as a web map. The map showed in Figure 6.4 is produced simply with the following code:

```
> stplotGoogleMaps(MeanTemp,zcol='tempc', mapTypeId='ROADMAP',w='49%',h='100%')
```

A similar line of code produces the spatio-temporal visualisation of predicted mean temperatures in Serbia for 2011-07-05 and 2011-07-06. The additional interactive control utilities: layer appearance, transparency line width, and legend colors display can be set by adding additional arguments to plotGoogleMaps plotting functions. The color coding system for map design in R is supported by RColorBrewer package (Brewer et al.,

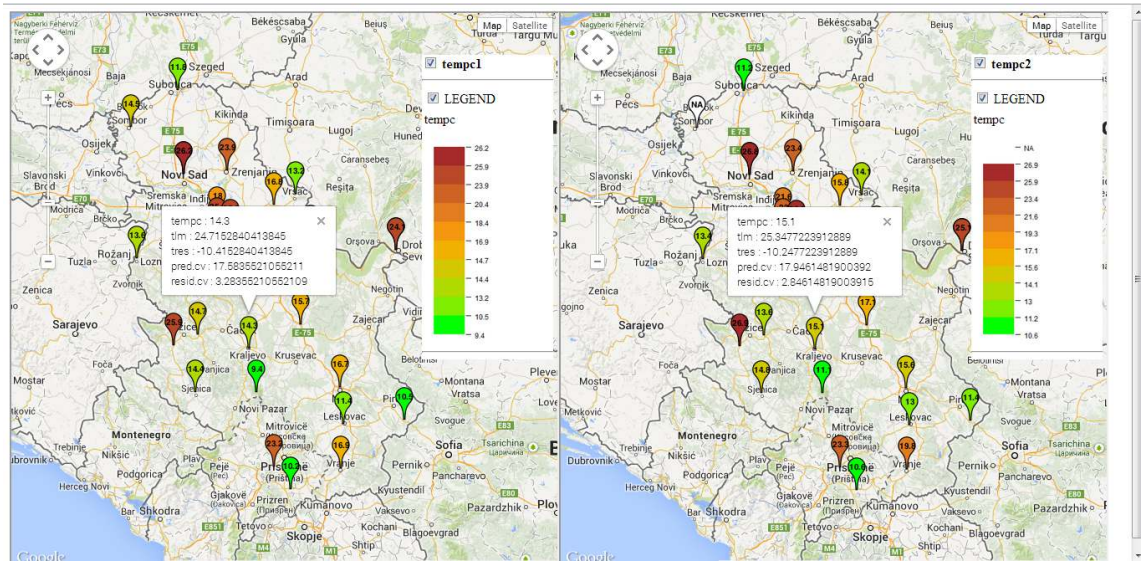

FIGURE 6.4: Mean daily temperature observations from GSOD and ECA&D data sets for Serbia for 2011-07-05 and 2011-07-06.

2003). It provides palettes for drawing nice maps. This package was derived from the ColorBrewer website `http://colorbrewer2.org/`. ColorBrewer is an online tool that helps chose a colour palette according to the number of data classes and the nature of data (matched with sequential, diverging and qualitative schemes). The colors obtained from RColorBrewer are used for the color scheme in Figure 6.5.

```
blues=colorRampPalette(brewer.pal(9, "PuBu")[c(8,5,2)] , space = "Lab")
reds=colorRampPalette(brewer.pal(9, "YlOrRd")[c(2,5,8)] , space = "Lab")
stplotGoogleMaps(Prediction, w='49%', h='100%', colPalette= c(blues(7),reds(7)))
```

Quantitative point symbols, such as proportional symbols of varying sizes that are used to symbolize totals at a point, are available in the plotGoogleMaps. The most frequently used shapes are circles, however squares and triangles are also possible solutions offered by the plotGoogleMaps function. A spatial plot with proportional symbols is presented in Figure 6.6.

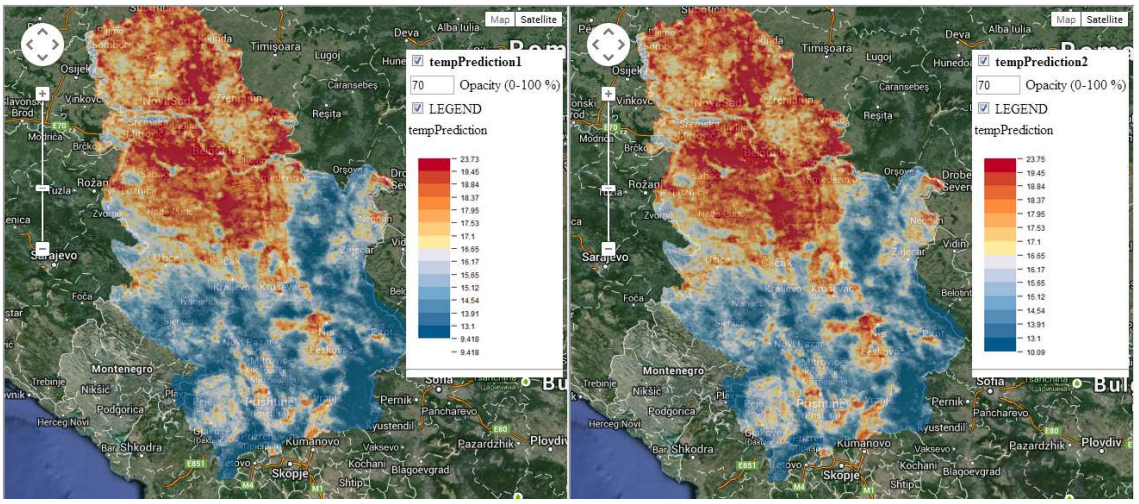

FIGURE 6.5: Mean daily temperature images for Serbia for 2011-07-05 and 2011-07-06.

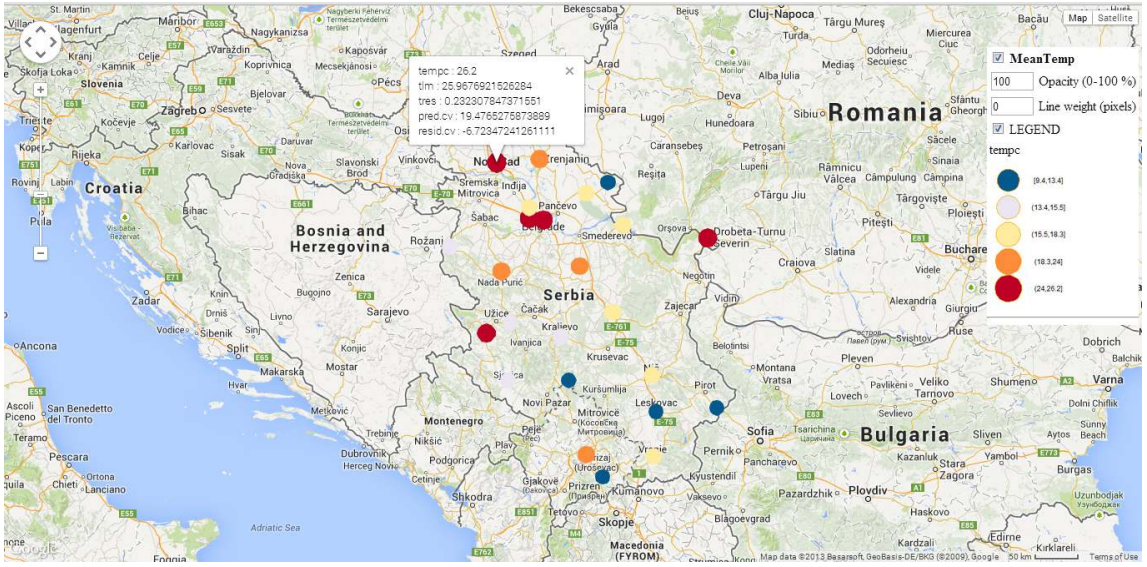

FIGURE 6.6: Mean daily temperature observations from GSOD and ECA&D data sets for Serbia for 2011-07-05. Proportional symbols.

6.3.2 Real-time visualisation of meteorological observations

The package can be implemented for the visualisation of meteorological observations in real-time. The example of using the package for this purpose is illustrated in the presentation of meteorological data in Catalonia (Spain), see website `http://meteo4u.com/`.

Figure 6.7 shows a temperature map where Google Maps markers are represented as numbers that indicate actual temperatures in near real-time, which were obtained from a few different sources. The map mushup (Figure 6.7) shows detailed station information in the form of a *'tooltip'* window. Detailed information are also available as Google Maps API JavaScript object, *google.maps.InfoWindow*, appearing after the user clicks on a specific station point. As a result, the obtained map mushup is interactive, intuitive and functional thanks to underlying Google Maps API and predefined JavaScript functions created by plotGoogleMaps. The map creation could be automated using the R server and maps are changed as data are updated.

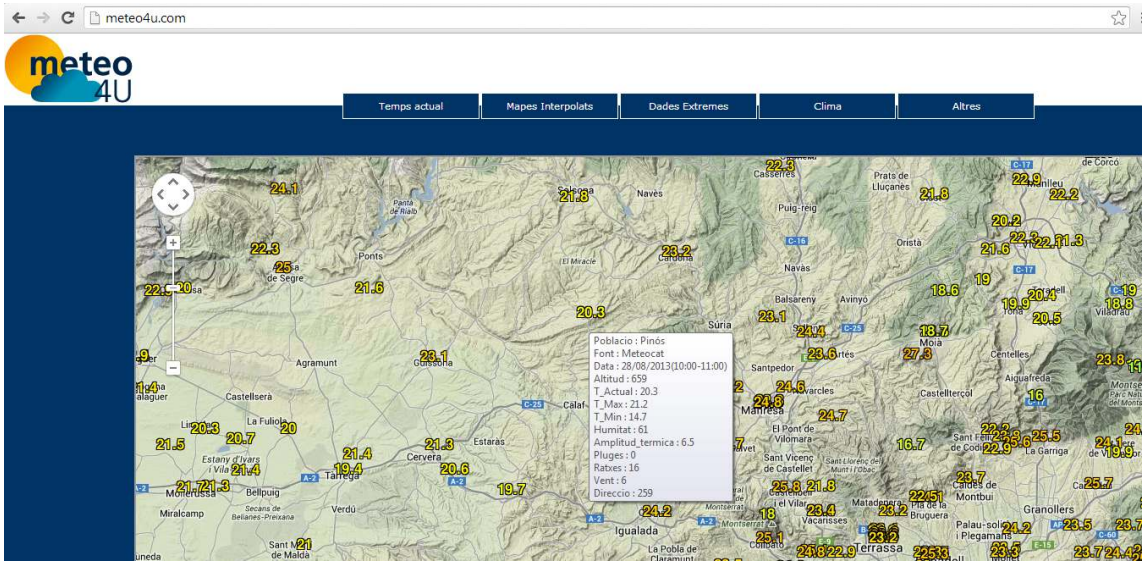

FIGURE 6.7: Plotting temperature data using plotGoogleMaps. Map mushup is availible on `http://meteo4u.com/`.

The main function in the package *vectorsSP* creates a radius vector from point data in form of *SpatialLinesDataFrame* class depending on radius and azimuth. This function is very appropriate for mapping wind speed and wind direction observation. Figure 6.8 shows a real life application of the automated visualisation of wind observations.

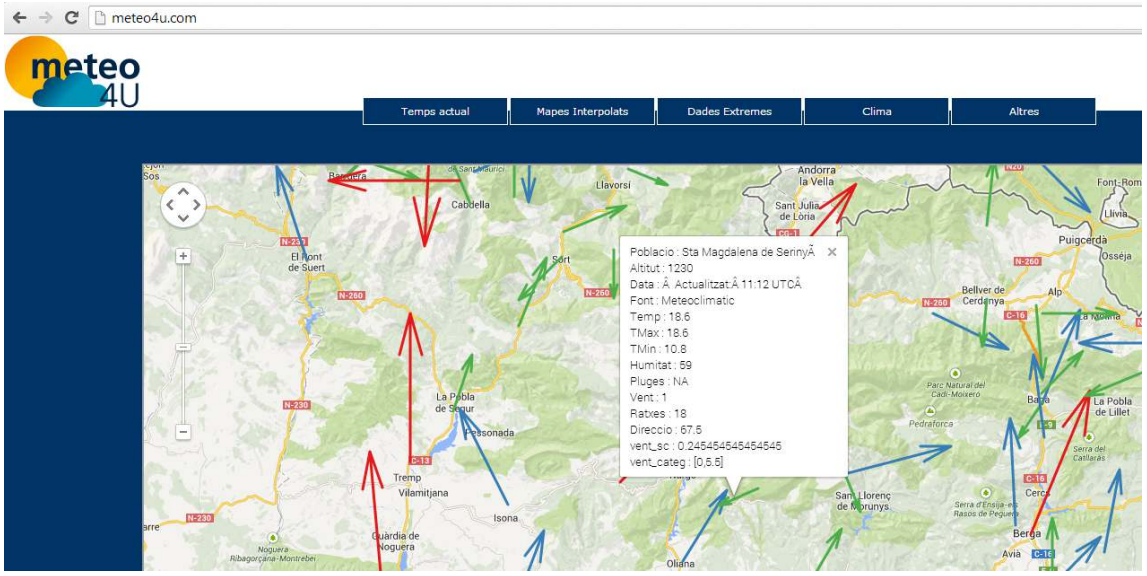

FIGURE 6.8: Plotting wind observation by using proprortional symbols depending on wind speed. The orientation of the radius vectors depends on wind direction. Map mushup is availible on `http://meteo4u.com/`.

6.3.3 Spatial visualisation of rainfall trends in Serbia

The spatial pattern of annual, seasonal and monthly rainfall trends in Serbia are examined by Luković et al. (2013). The study used data from 63 weather stations between the period of 1961–2009. Precipitation trends are depicted with different colors in Figure 6.9 (positive trends are blue and negativetrends are red) and are overlaid with proportional bubble symbols presenting the trend values of each considered station. The bubbles outlined with black circles in the maps represent stations with statistically significant trends at the confidence level of 97.5 %. The web maps are available on `http://www.grf.bg.ac.rs/~bajat/Trends.htm`.

Kilibarda et al. (2013c,a) use plotGoogleMaps for scientific communication and visualisation in meteo/climatic mapping, see `http://dailymeteo.org/` .

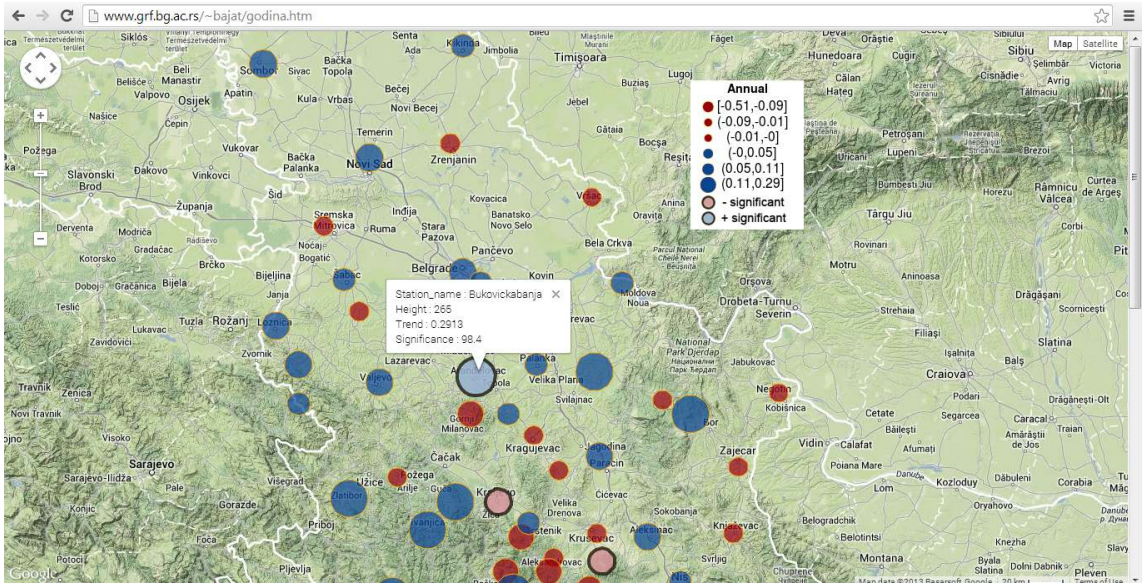

FIGURE 6.9: Spatial distribution of rainfall trends in Serbia from 1961 to 2009, annual map. The bubbles with blackoutlined circles represent stations with significant positive and negative trends at the confidence level of 97.5 %. The web map is availible on `http://www.grf.bg.ac.rs/~bajat/Trends.htm`.

6.4 Discussion and Conclusions

The plotGoogleMaps is a free and open source software solution for the simple creation of rich interactive web maps. In this case, plotGoogleMaps uses the web browser as a plotting device instead of the default R graphic device. Therefore, it offers more advantages when compared to the classical R plotting device environment. For an example, significant advantages include high quality of background Google layers for better abstractions of geographical reality, spatial data exploration functionality and map interactivity (navigation control, pan, zoom, attribute info windows, etc). The package plotGoogleMaps is a tool that can be used for plotting meteorological/climatic spatial and spatio-temporal data for internal use or for website.

This package promotes the creation of interactive maps in user friendly environments where a map is stored in the HTML format. Map sharing with non GIS map users is easier and it is not necessary to use the GIS software. Sharing is simply achieved using aweb browser and the map remains interactive and simpler than in professional software. Also, R users can use this package instead of standard plot functions because it provides

a faster preview of the mapped data in relation to geographic reality provided by Google Maps.

Google Maps API is not suitable for handling large amounts of data and consequently plotGoogleMaps has the same constraint. One of the possible alternative *open-source* solutions might be the combination of a server side software (e.g. Geoserver) and a client side software (e.g. Openlayers), according to OGC standards[9]. This solution requires more specific GIS knowledge and a greater understanding of software and standards for establishing a mapping framework. The implementation of automatic web map creation would have more requirements and depend more on server software components.

[9] `http://www.opengeospatial.org/standards`

Chapter 7

Discussion and conclusion

This work was conducted in part thanks to organizations such as the national Meteorological Services and WMO, National Aeronautics and Space Administration (NASA), National Climatic Data Center (NCDC), European Climate Assessment and Dataset Project (ECA&D), Global Precipitation Climatology Centre (GPCC), European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), and the United States National Oceanic and Atmospheric Administration (NOAA). The meteorological data provided by these organizations is available to the public and to the research community. The spatio-temporal models were based on publicly available ground observations together with publicly available time series of atmospheric and surface reflectance images (MODIS, Meteosat, GOES, GMS). Such data can be used to produce a new generation of detailed daily global maps of meteorological variables. Even though the positions of meteorological stations are representative to describe the weather and climate in some neighbourhood, the global geographical or feature space coverage is not representative from the point of view of spatio-temporal statistics requirements or sampling strategies (Heuvelink et al., 2012).

Global spatio-temporal analysis of publicly available data sets (a collection of GSOD and ECA&D) shows that the observed high temporal, spatial and feature space clustering of meteorological stations potentially represent a limitation of these data sets and could further complicate the fitting of accurate global spatio-temporal models. This implies that sophisticated spatio-temporal techniques need to be used that can account for the

data clustering. The use of remote sensing and/or monthly images as covariates is one solution to overcome clustering issues. The spatio-temporal regression kriging model uses covariates for de-trending and is followed by the interpolation of regression residuals . This interpolation step uses the covariance model integrally in space and time through the incorporation of a spatio-temporal variogram. The applied covariance model in this thesis takes into account pure spatial, temporal and spatio-temporal components of variability. Such treatment can provide the most realistic estimate of the uncertainty so that unbiased estimates of the global and local land air temperature and other meteorological variables can be also be produced.

The presented model can be used for calibration of 8-day MODIS LST images by integrating station observations together with geometrical temperature trends, elevations and the topographic wetness index. The result of this treatment would afford the first global daily air temperature images at very high spatial and temporal resolution (1 km spatial and 1 day temporal resolution). The geometrical temperature trend (Eq. 4.3) presented in this thesis could be a crucial covariate for real-time mapping or for temperature interpolation for the dates before MODIS LST images has been launched.

Furthermore, globally fitted models of daily temperatures could be used for regional or local studies, e.g. where a limited number of ground observations are available so that some referent global model of spatio-temporal variability is required. The models described in this thesis can be obtained by installing the `meteo` package that has been mostly created and maintained by the author.

Clearly, the presented computational framework could be used to produce a global archive of the mean, minimum and maximum temperature images. The daily maps of temperature could also serve as raster files in a similar fashion as climate layers from the WorldClim project (Hijmans et al., 2005). This would require HUGE data storage and serving capacities considering the amount of output pixels (10 years by 365 days by 3 meteorological variables plus uncertainty maps). The service should also offer Web GIS functionalities implemented through OGC standards.

This study discovers that the GSOD point data sets still contain many artefacts and possible gross errors. Therefore, the mapping accuracy can be improved by filtering station

observations. Concerning the number of stations, this procedure should be done by using some automated method. One of the main objectives for the further development of the meteo package will be incorporation of an automated and tested algorithm for the detection of outliers.

A future plan is to use the publicly available data sets shown in this thesis to model and interpolate daily meteorological variables such as: precipitation, wind speed, snow depth, meteorological indicators etc. at the spatial resolution of 1 km and temporal support of 1 day. WordDailyMeteo could be extended to offer all meteorological variables contained in publicly available data sets.

Presented approach in this thesis could be also used for climatic mapping. Figure 7.1 shows general climatic-mapping scheme based on spatio-temporal daily mapping in contrast to classical climatic mapping approach. Classical approach assumes aggregation of meteorological measurements to climatic variable and than spatial modelling and prediction. WorldDailyMeteo approach, is based on the aggregation of daily maps, archiving and offering daily weather images but also offering monthly, yearly or other climatic global maps at very high spatial resolution. The climatic maps would be based on daily spatial estimates, therefore, all daily measurements even from stations with time series covering short period are incorporated into final climatic estimate.

What would WorldDailyMeteo offer that other services do not provide:

1. WorldDailyMeteo maps would be of high spatial detail (1 km), high temporal resolution (1 day; 10 years of maps) and with a global coverage;
2. WorldDailyMeteo predictions would be based on using state of the art geostatistical methods (linear or GLM-based spatio-temporal regression-kriging with time series of predictors - MODIS and similar RS images);
3. All target meteorological variables would be mapped using automated mapping frameworks with a single global model for each target meteorological variable;
4. All target meteorological maps could be aggregated to climatic maps, at very various temporal support (monthly, yearly, etc);

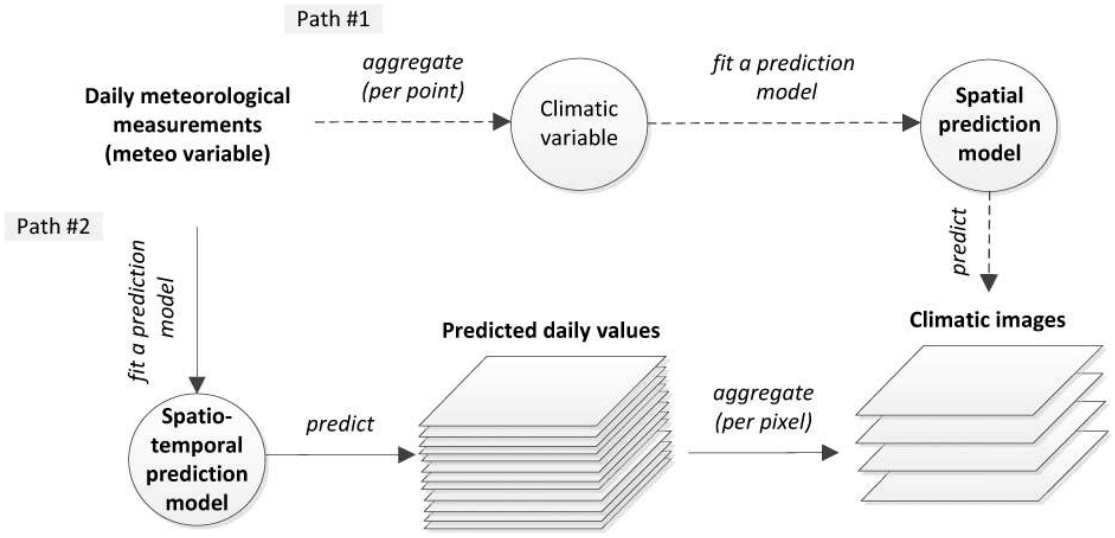

FIGURE 7.1: A general spatio-temporal prediction framework. (Path#1) Classical climatic mapping approch. (Path#2) Daily mapping and climatic aggregation, WorldDailyMeteo approch.

5. The produced time series (10 years of daily images) of the target meteorological parameters could be analysed using time-series / Fourier analysis algorithms. We could extract global, regional and local components of dynamics of meteorological variables (per pixel);

Very high spatio-temporal resolution dataset offered from WordDailyMeteo, based on cleaned publicly available data, could be used for analysing extremes of climate, in environmental modelling, precise agriculture, hydrological modelling, terrestrial biospheric modelling, ect.

Bibliography

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A. M. G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Rupa Kumar, K., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D. B., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M., and Vazquez-Aguirre, J. L. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research, D: Atmospheres*, 111(D5):n/a–n/a.

Amante, C. and Eakins, B. W. (2009). ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. *NOAA Technical Memorandum NESDIS NGDC*, 24.

Antonić, O., Križan, J., Marki, A., and Bukovec, D. (2001). Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks. *Ecological Modelling*, 138(1):255–263.

Baddeley, A. and Turner, R. (2006). Modelling spatial point patterns in R. In Baddeley, A., Gregori, P., Mateu, J., Stoica, R., and Stoyan, D., editors, *Case Studies in Spatial Point Process Modeling*, volume 185 of *Lecture Notes in Statistics*, pages 23–74. Springer New York.

Bader, M. Y. and Ruijten, J. J. (2008). A topography-based model of forest cover at the alpine tree line in the tropical Andes. *Journal of Biogeography*, 35(4):711–723.

Bajat, B., Hengl, T., Kilibarda, M., and Krunić, N. (2011). Mapping population change index in southern serbia (1961–2027) as a function of environmental factors. *Comput. Environ. Urban.*, 35(1):35–44.

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., and Ziese, M. (2012). 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 Syst. Sci. Data Discuss.*, 5(2):921–998.

Becker, R. A. and Chambers, J. M. (1984). *S: an interactive environment for data analysis and graphics*. CRC Press.

Benichou, P. and Le Breton, O. (1987). Prix Norbert Gerbier 1986: Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques (Incorporating topography in statistical mapping of precipitation fields, in English). *Météorologie*, (19).

Beniston, M. (2006). Mountain weather and climate: A general overview and a focus on climatic change in the Alps. *Hydrobiologia*, 562:3–16.

Best, D. (2006). Web 2.0 Next big thing or next big Internet bubble? In *Lecture Web Information Systems*. Eindhoven: Technische Universiteit Eindhoven.

Bivand, R., Keitt, T., and Rowlingson, B. (2013). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 0.8-6.

Bivand, R., Pebesma, E., and Rubio, V. (2008). *Applied spatial data analysis with R*. Use R Series. Springer, Heidelberg.

Boehner, J. and Antonic, O. (2009). Chapter 8 land-surface parameters specific to topoclimatology. In Hengl, T. and Reuter, H. I., editors, *Geomorphometry Concepts, Software, Applications*, volume 33 of *Developments in Soil Science*, pages 195 – 226. Elsevier.

Böhner, J., Blaschke, T., and Montanarella, L., editors (2008). *SAGA — Seconds Out*, volume 19. Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie, Hamburg.

Brewer, C. A., Hatchard, G. W., and Harrower, M. A. (2003). Colorbrewer in print: a catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32.

Brohan, P., Kennedy, J., Harris, I., Tett, S., and Jones, P. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res. D: Atmos.*, 111(D12):D12106.

Burrough, P., editor (1998). *Principles of Geographical Information Systems*. Oxford University Press, Oxford, 2nd edition.

Burrough, P. A. and Mcdonnell, R. A. (1998). *Principles of Geographical Information Systems*. Oxford University Press, USA, 2 edition.

Butler, D. (2006). Virtual globes: The web-wide world. *Nature*, 439:776–778.

Caesar, J., Alexander, L., and Vose, R. (2006). Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *J. Geophys. Res. D: Atmos.*, 111(D5):D05101.

Carrera-Hernández, J. and Gaskin, S. (2007). Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. *Journal of Hydrology*, 336(3):231–249.

Chapman, A. D. (2005). *Principles of data quality, version 1.0.* Report for the Global Biodiversity Information Facility. Australian Biodiversity Information Services, Toowoomba South.

Christensen, R. (2001). *Advanced linear modeling: multivariate, time series, and spatial data; nonparametric regression and response surface maximization*. Springer.

Christy, J. R., Spencer, R. W., and Braswell, W. D. (2000). MSU Tropospheric Temperatures: Dataset Construction and Radiosonde Comparisons. *Journal of Atmospheric and Oceanic Technology*, 17(9):1153–1170.

Coll, C., Wan, Z., and Galve, J. M. (2009). Temperature-based and radiance-based validations of the V5 MODIS land surface temperature product. *Journal of Geophysical Research*, 114(D20):D20102.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.

Daly, C., Gibson, W. P., Taylor, G. H., Johnson, G. L., and Pasteris, P. (2002). A knowledge-based approach to the statistical mapping of climate. *Climate research*, 22(2):99–113.

Daly, C., Taylor, G., and Gibson, W. (1997). The prism approach to mapping precipitation and temperature. In *Proceedings, 10th Conference on Applied Climatology, American Meteorology Society*, pages 10–12.

Davis, G. (2007). History of the noaa satellite program. *J. Appl. Remote Sens.*, 1(1):012504–012504–18.

Demyanov, V., Kanevski, M., Chernov, S., Savelieva, E., and Timonin, V. (1998). Neural network residual kriging application for climatic data. *Journal of Geographic Information and Decision Analysis*, 2(2):215–232.

Di Luzio, M., Johnson, G. L., Daly, C., Eischeid, J. K., and Arnold, J. G. (2008). Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous united states. *J. Appl. Meteor. Climatol.*, 47(2):475 – 497.

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns*. Mathematics in biology. Edward Arnold, New York. 2nd edition.

Dubayah, R. and Rich, P. M. (1995). Topographic solar radiation models for gis. *International Journal of Geographical Information Systems*, 9(4):405–419.

Elith, J., H. Graham, C., P. Anderson, R., Dudik, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., SoberÃşn, J., Williams, S., S. Wisz, M., and E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. (2011). A statistical explanation of maxent for ecologists. *Divers. Distrib.*, 17(1):43–57.

Environment and Natural Resources (2001). FAOCLIM 2: world-wide agroclimatic data. Working paper no. 5 (cd-rom), FAO, Rome.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45(2):n/a–n/a.

Feick, R. and Deparday, V. (2010). Evaluating selected visualisation methods for exploring vgi. *Geomatica*, 64(4):427–437.

Feidas, H., Makrogiannis, T., and Bora-Senta, E. (2004). Trend analysis of air temperature time series in greece and their relationship with circulation using surface and satellite data: 1955–2001. *Theoretical and Applied Climatology*, 79:185–208.

Frei, C. and Schaer, C. (1998). A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology*, 18(8):873–900.

Gartner, G. (2009). Applying web mapping 2.0 to cartographic heritage. *e-Perimetron*, 4(4):234–239.

Gesmann, M. and de Castillo, D. (2011). Using the google visualisation api with r. *The R Journal*, 3(2):40–44.

Gething, P., Atkinson, P., Noor, A., Gikandi, P., Hay, S., and Nixon, M. (2007). A local space–time kriging approach applied to a national outpatient malaria data set. *Computers & geosciences*, 33(10):1337–1350.

Gibin, M., Singleton, A., Milton, R., Mateos, P., and Longley, P. (2008). An exploratory cartographic visualisation of London through the Google Maps API. *Applied Spatial Analysis and Policy*, 1(2):85–97.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Gräler, B., Gerharz, L., and Pebesma, E. (2011). Spatio-temporal analysis and interpolation of PM10 measurements in Europe. *ETC/ACM Technical Paper*, 10.

Grunsky, E. (2002). R: a data analysis and statistical programming environment–an emerging tool for the geosciences. *Computers & Geosciences*, 28(10):1219–1222.

Haklay, M., Singleton, A., and Parker, C. (2008). Web mapping 2.0: The neogeography of the geoweb. *Geography Compass*, 2(6):2011–2039.

Haklay, M. and Zafiri, A., editors (2007). *Usability engineering for GIS Ű Learning from a snapshot*. Presented at the 23rd International Cartographic Congress, Moscow, Russia, 4Ű10 August 2007.

Hartkamp, A. D., De Beurs, K., Stein, A., and White, J. W. (1999). *Interpolation techniques for climate variables*. CIMMYT Mexico, DF.

Hartmann, D. (1994). *Global physical climatology*. Academic Press, San Diego.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M. (2008). A european daily high-resolution gridded data set of surface temperature and precipitation for 1950âĂŞ2006. *J. Geophys. Res. D: Atmos.*, 113(D20):D20119.

Hengl, T. (2007). *A practical guide to geostatistical mapping of environmental variables*, volume 140.

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. University of Amsterdam, Amsterdam, 2nd edition.

Hengl, T., Heuvelink, G. B., Perčec Tadić, M., and Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theoretical and Applied Climatology*, 107:265–277.

Heuvelink, G. B. M. and Griffith, D. A. (2010). Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. *Geographical Analysis*, 42(2):161–179.

Heuvelink, G. B. M., Griffith, D. A., Hengl, T., and Melles, S. J. (2012). *Sampling Design Optimization for Space-Time Kriging*, pages 207–230. John Wiley & Sons, Ltd.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.*, 25:1965–1978.

Hijmans, R. J. and van Etten, J. (2013). *raster: Geographic data analysis and modeling*. R package version 2.1-25.

Hofierka, J. and Suri, M. (2002). The solar radiation model for open source gis: implementation and applications. In *Proceedings of the Open source GIS - GRASS users conference*, pages 1–19.

Hofstra, N., Haylock, M., New, M., Jones, P., and Frei, C. (2008). Comparison of six methods for the interpolation of daily, european climate data. *J. Geophys. Res. D: Atmos.*, 113(21).

Hsu, K.-l., Gao, X., Sorooshian, S., and Gupta, H. V. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, 36(9):1176–1190.

Isaaks, E. and Srivastava, R. (1988). Spatial continuity measures for probabilistic and deterministic geostatistics. *Mathematical Geology*, 20(4):313–341.

Isaaks, E. H. and Srivastava, R. M. (1989). *Applied geostatistics*. Oxford University Press.

Jarvis, C. H. and Stuart, N. (2001). A comparison among strategies for interpolating maximum and minimum daily air temperatures. Part II: The interaction between number of guiding variables and the type of interpolation method. *Journal of Applied Meteorology*, 40(6):1075–1084.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P. (2012). Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research*, 117(D5):D05127.

Kanevski, M., Pozdnoukhov, A., and Timonin, V. (2009). *Machine Learning for Spatial Environmental Data: theory, applications and software*. Epfl Press.

Kidd, C. (2001). Satellite rainfall climatology: a review. *Int. J. Climatol.*, 21(9):1041–1066.

Kiktev, D., Sexton, D. M., Alexander, L., and Folland, C. K. (2003). Comparison of modeled and observed trends in indices of daily climate extremes. *Journal of Climate*, 16(22):3560–3571.

Kilibarda, M. (2013). *plotGoogleMaps: Plot SP or ST(STIDF,STFDF) data as HTML map mashup over Google Maps*. R package version 2.0.

Kilibarda, M. and Bajat, B. (2012). Plotgooglemaps: The R-based web-mapping tool for thematic spatial data. *Geomatica*, 66(1):37–49.

Kilibarda, M., Hengl, T., Heuvelink, G., Graeler, B., Pebesma, E., Perčec Tadić, M., and Bajat, B. (2013?a). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres*, in review.

Kilibarda, M., Hengl, T., Pebesma, E., and Graeler, B. (2013b). *meteo: Spatio-temporal analysys and mapping of metorological observations*. R package version 0.1-0.

Kilibarda, M., Perčec Tadić, M., Hengl, T., Luković, J., and Bajat, B. (2013?c). Publicly available global meteorological data sets: sources, representation, and usability for spatio-temporal analysis. *International Journal of Climatology*, in review.

Kilibarda, M., Protić, D., and Nestorov, I. (2010). Application of Google Maps API service for creating web map of information retrieved from CORINE land cover databases. *Glasnik Srpskog geografskog drustva*, 90(4):103–114.

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and Petrovic, P. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol.*, 22(12):1441–1453.

Kleiner, K. (2011). Data on demand. *Nature Clim. Change*, 1:10–12.

Knaus, J. (2013). *snowfall: Easier cluster computing (based on snow).* R package version 1.84-4.

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., and Rennie, J. (2011). An overview of the global historical climatology network monthly mean temperature data set, version 3. *Journal of Geophysical Research*, 116(D19):D19121.

Leemans, R. and Cramer, W. P. (1991). The IIASA database for mean monthly values of temperature, precipitation, and cloudiness on a global terrestrial grid. *INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS, Laxenburg, Austria.*

Legates, D. and Willmott, C. (1990). Mean seasonal and spatial variability in global surface air temperature. *Theoretical and Applied Climatology*, 41(1-2):11–21.

Li, J. and Heap, A. D. (2008). *A review of spatial interpolation methods for environmental scientists*. Geoscience Australia, Canberra.

Loecher, M. (2013). *RgoogleMaps: Overlays on Google map tiles in R*. R package version 1.2.0.3.

Luković, J., Bajat, B., Blagojević, D., and Kilibarda, M. (2013). Spatial pattern of recent rainfall trends in serbia (1961–2009). *Regional Environmental Change*, pages 1–11.

Mendelsohn, R., Kurukulasuriya, P., Basist, A., Kogan, F., and Williams, C. (2007). Climate analysis with satellite versus weather station data. *Clim. Change*, 81:71–83.

Menne, M., Imke, D., Vose, R., Gleason, B., and Houston, T. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910.

Miller, C. C. (2006). A beast in the field: The google maps mashup as gis/2. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 41(3):187–199.

Mitas, L. and Mitasova, H. (1999). Spatial interpolation. *Geographical Information Systems: Principles, Techniques, Management and Applications, Wiley*, 481.

Mitchell, T. D. and Jones, P. D. (2005). An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, 25(6):693–712.

Neteler, M. (2010). Estimating daily land surface temperatures in mountainous environments by reconstructed modis lst data. *Remote Sensing*, 2(1):333–351.

New, M., Hulme, M., and Jones, P. (1999). Representing twentieth-century space-time climate variability. part i: Development of a 1961-90 mean monthly terrestrial climatology. *J. Clim.*, 12:829 – 856.

New, M., Hulme, M., and Jones, P. (2000). Representing twentieth-century space-time climate variability. part ii: development of 1901-96 monthly grids of terrestrial surface climate. *J. Clim.*, 13:2217–2238.

Ohring, G. and Booth, A. (1995). The noaa pathfinder program. *Adv. Space Res.*, 16(10):15 – 20.

Ohring, G., Gallo, K., Gruber, A., Planet, W., Stowe, L., and Tarpley, J. D. (1989). Climate and global change: Characteristics of noaa satellite data. *Eos Trans. AGU*, 70(41):895–895.

O'Reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17.

Pebesma, E. (2012). spacetime: spatio-temporal data in R. *J. Stat. Softw.*, 51:1–30.

Pebesma, E. (2013). Spatio-temporal geostatistics using gstat. *R viggnete for gstat package*.

Pebesma, E., Cornford, D., Dubois, G., Heuvelink, G., Hristopoulos, D., Pilz, J., Stoehlker, U., Morin, G., and Skoien, J. (2010). Intamap: the design and implementation of an interoperable automated interpolation web service. *Computers & Geosciences*, ?:?

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30(7):683–691.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in r. *R News*, 5(2):9–13.

Peterson, M. (2007). The internet and multimedia cartography. In Cartwright, W., Peterson, M., and Gartner, G., editors, *Multimedia Cartography*, pages 35–50. Springer Berlin Heidelberg.

Peterson, T. C. and Vose, R. S. (1997). An overview of the global historical climatology network temperature database. *Bull. Am. Meteorol. Soc.*, 78(12):2837–2849.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecol. Modell.*, 190(3–4):231–259.

Pinheiro, J. and Bates, D. (2009). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer.

Piper, S. C. and Stewart, E. F. (1996). A gridded global data set of daily temperature and precipitation for terrestrial biospheric modeling. *Global Biogeochemical Cycles*, 10(4):757–782.

Price, D. T., McKenney, D. W., Nalder, I. A., Hutchinson, M. F., and Kesteven, J. L. (2000). A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agricultural and Forest meteorology*, 101(2):81–94.

Prigent, C. (2010). Precipitation retrieval from space: An overview. *C. R. Geoscience*, 342(3-4):380–389.

Proedrou, M., Theoharatos, G., and Cartalis, C. (1997). Variations and trends in annual and seasonal air temperatures in greece determined from ground and satellite measurements. *Theoretical and Applied Climatology*, 57:65–78.

R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabus, B., Eineder, M., Roth, A., and Bamler, R. (2003). The shuttle radar topography mission - a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57(4):241 – 262.

Ripley, B. D. (1981). *Spatial Statistics*. Wiley, New York.

Rohde, R., Muller, R., Jacobsen, R., Muller, E., and Perlmutter, S. e. a. (2012). A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinfor Geostat: An Overview*, 1.

Santer, B. D., Wigley, T. M. L., Gaffen, D. J., Bengtsson, L., Doutriaux, C., Boyle, J. S., Esch, M., Hnilo, J. J., Jones, P. D., Meehl, G. A., Roeckner, E., Taylor, K. E., and Wehner, M. F. (2000). Interpreting differential temperature trends at the surface and in the lower troposphere. *Science*, 287(5456):1227–1232.

Schutta, N. T. and Asleson, R. (2005). *Foundations of AJAX*. Apress.

Seong, J. C., Mulcahy, K. A., and Usery, E. L. (2002). The sinusoidal projection: A new importance in relation to global image data. *The Professional Geographer*, 54(2):218–225.

Skarlatidou, A. and Haklay, M. (2006). Public web mapping: preliminary usability evaluation. *Presented at GIS Research UK 2005, Nottingham, UK*.

Smith, T. M., Arkin, P. A., Bates, J. J., and Huffman, G. J. (2006). Estimating bias of satellite-based precipitation estimates. *J. Hydrometeorol.*, 7(5):841–856.

Sorensen, R., Zinko, U., and Seibert, J. (2006). On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrology and Earth System Sciences Discussions*, 10:101–112.

Stahl, K., Moore, R., Floyer, J., Asplin, M., and McKendry, I. (2006). Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agricultural and Forest Meteorology*, 139(3):224–236.

Struzik, P., Stancalie, G., Danson, F., Toulios, L., Dunkel, Z., and E., T. (2011). Study of satellite data availability and their resolution in time and space for the assessment of climate change and variability impacts on agriculture. ch. in satellites data availability, methods and challenges for the assessement of climate change and and variability impacts on agriculture. In Toulios, L. and Stancalie, G., editors, *Satellite Data Availability, Methods and Challenges for the Assessment of Climate Change and Variability Impacts on Agriculture*, pages 3–26. COST Secretariat / The Council of the European Union.

Thorne, P. W., Parker, D. E., Christy, J. R., and Mears, C. A. (2005). Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Am. Meteorol. Soc.*, 86(10):1437–1442.

Toulios, L. and Stancalie, G., editors (2011). *Satellite data availability, methods and challenges for the assessment of climate change and variability impacts on agriculture*. COST Secretariat / The Council of the European Union.

Tsai, V. J. D. (1993). Delaunay triangulations in tin creation: an overview and a linear-time algorithm. *International journal of geographical information systems*, 7(6):501–524.

Tveito, O., Wegehenkel, M., van der Wel, F., and Dobesch, H. (2006). Spatialisation of climatological and meteorological information with the support of GIS (Working Group 2). *The Use of Geographic Information Systems in Climatology and Meteorology, Final Report*, pages 37–172.

Van De Kerchove, R., Lhermitte, S., Veraverbeke, S., and Goossens, R. (2013). Spatio-temporal variability in remotely sensed land surface temperature, and its relationship with physiographic variables in the Russian Altay Mountains. *International Journal of Applied Earth Observation and Geoinformation*, 20:4–19.

Van den Besselaar, E. J. M., Haylock, M. R., van der Schrier, G., and Klein Tank, A. M. G. (2011). A european daily high-resolution observational gridded data set of sea level pressure. *J. Geophys. Res. D: Atmos.*, 116(D11):n/a–n/a.

Venables, W. N., Ripley, B. D., and Venables, W. (1994). *Modern applied statistics with S-PLUS*, volume 250. Springer-verlag New York.

Vose, R. S., Division., O. R. N. L. E. S., (U.S.), G. C. R. P., of Energy. Office of Health, U. S. D., Research., E., (U.S.), C. D. I. A. C., and Martin Marietta Energy Systems, I. (1992). *The Global historical climatology network : long-term monthly temperature, precipitation, sea level pressure, and station pressure data*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory; Available to the public from N.T.I.S., Oak Ridge, Tenn.; Springfield, VA.

Wan, Z., Zhang, Y., Zhang, Q., and Li, Z.-L. (2004). Quality assessment and validation of the modis global land surface temperature. *International Journal of Remote Sensing*, 25(1):261–274.

Webster, R. and Oliver, M. A. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.

Wessel, P. and Smith, W. H. (1996). A global, self-consistent, hierarchical, high-resolution shoreline database. *J. Geophys. Res. B: Solid Earth*, 101(B4):8741–8743.

WMO-No. 837 (1996). Exchanging meteorological data: guidelines on relationships in commercial meteorological activities - wmo policy and practice. Report, World Meteorological Organisation, Geneva.

Yang, Y., Wilson, L., and Wang, J. (2010). Development of an automated climatic data scraping, filtering and display system. *Computers and Electronics in Agriculture*, 71(1):77 – 87.

Yoo, J.-M., Won, Y.-I., Cho, Y.-J., Jeong, M.-J., Shin, D.-B., Lee, S.-J., Lee, Y.-R., Oh, S.-M., and Ban, S.-J. (2011). Temperature trends in the skin/surface, mid-troposphere and low stratosphere near korea from satellite and ground measurements. *Asia-Pacific Journal of Atmospheric Sciences*, 47:439–455.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, 14(6):1–27.

Biography

Milan Kilibarda was born in Nikšić, Montenegro, on August 15th, 1983. He completed elementary school (in 1998) and gymnasium for Natural sciences and mathematics in 2002 in Nikšić. He was awarded with degree "Luča", for all excellent marks in gymnasium. In 2002, he started study of Geodesy at University of Belgrade, Faculty of Civil Engineering, Department of Geodesy and Geoinformatics. In 2007, he finished study with average mark 9.41 (max 10.00). He defended diploma exam entitled *"Possibility of implementation LPIS and CwRS systems for control of subventions in agriculture in Republic of Serbia"*. During the study he received two awards for excellent results during the study from Faculty, and in 2007, he received award from University as the best student from Faculty of Civil Engineering in year 2007.

In 2007, he enrolled in PhD study, within the same University. During the study he completed exams and started with research related to space-time interpolation of daily meteorological variables for global land areas. In September 2013, he submitted PhD thesis entitled *"Automated mapping of climatic variables using spatio-temporal geostatistical methods"*. During his PhD study Milan Kilibarda published as author or co-author papers related to spatial analyses and visualisation: 4 journal papers (from SCI list), 3 papers in international peer-reviewed journals, 2 in Serbian journals, 2 chapters in monography, 12 international conference papers.

Since January 2008, he has been a teaching assistant in the field of Cartography at Faculty of Civil Engineering, Department of geodesy and geoinformatics, University of Belgrade. He was involved as researcher in 3 research projects founded by Serbian Ministry of Science, and he gave consulting in 2 innovative pilot projects in Serbia (iSCOPE and eEnviPer) co-founded by European Commission (CIP-ICT-PSP program).

Прилог 1.

Изјава о ауторству

Потписани Милан Килибарда

број индекса 33/07

Изјављујем

да је докторска дисертација под насловом

„АУТОМАТСКО КАРТИРАЊЕ КЛИМАТСКИХ ВАРИЈАБЛИ ПРИМЕНОМ ПРОСТОРНО-ВРЕМЕНСКИХ ГЕОСТАТИСТИЧКИХ МЕТОДА“

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 04.09.2013. године

Милан Килибарда

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Милан Килибарда

Број индекса 33/07

Студијски програм Геодезија и геоинформатика

Наслов рада „АУТОМАТСКО КАРТИРАЊЕ КЛИМАТСКИХ ВАРИЈАБЛИ ПРИМЕНОМ ПРОСТОРНО-ВРЕМЕНСКИХ ГЕОСТАТИСТИЧКИХ МЕТОДА“

Ментор В.проф. др Бранислав Бајат, дипл. геод. инж

Потписани

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 04.09.2013. године

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

„АУТОМАТСКО КАРТИРАЊЕ КЛИМАТСКИХ ВАРИЈАБЛИ ПРИМЕНОМ ПРОСТОРНО-ВРЕМЕНСКИХ ГЕОСТАТИСТИЧКИХ МЕТОДА“

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 04.09.2013. године

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.