UNIVERSITY OF BELGRADE

FACULTY OF CIVIL ENGINEERING

Aleksandar M. Sekulić

SPATIO-TEMPORAL INTERPOLATION OF CLIMATE ELEMENTS USING GEOSTATISTICS AND MACHINE LEARNING

Doctoral Dissertation

Belgrade, 2021

UNIVERZITET U BEOGRADU

GRAĐEVINSKI FAKULTET

Aleksandar M. Sekulić

PROSTORNO-VREMENSKA INTERPOLACIJA KLIMATSKIH ELEMENATA PRIMENOM GEOSTATISTIKE I MAŠINSKOG UČENJA

doktorska disertacija

Beograd, 2021

Mentor: V. prof. dr Milan Kilibarda, dipl. inž. geod.
Univerzitet u Beogradu, Građevinski fakultet, odsek za geodeziju i geoinformatiku

Članovi komisije: Prof. dr Branislav Bajat, dipl. inž. geod.
Univerzitet u Beogradu, Građevinski fakultet, odsek za geodeziju i geoinformatiku

V. prof. dr Jelena Luković, dipl. geogr.
Univerzitet u Beogradu, Geografski fakultet

Doc. dr Milutin Pejović, dipl. inž. geod.
Univerzitet u Beogradu, Građevinski fakultet, odsek za geodeziju i geoinformatiku

Doc. dr Mladen Nikolić, dipl. mat.
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane:

Dedicated to my family
and my girlfriend Miljana.

Acknowledgements

First of all, I want to express my utmost gratitude to my supervisor and a friend, Prof. Dr Milan Kilibarda for his mentorship and guidance from the very beginning of my PhD studies. He had a lot of patience and selflessly shared his great knowledge with me. He is a great support and to this day he still motivates me on a daily basis. I would also like to take this occasion to express my deep appreciation to my Prof. Dr Branislav Bajat, whose assistant I have been for six years, for giving me support and sharing his knowledge and experience with me. Without them I would never have got this dissertation done.

I am very thankful to Dr Gerard Heuvelink and Dr Mladen Nikolić for sharing their mathematical knowledge with me and helping in defining the ideas. Their observations, comments and suggestions were valuable for me and were of great influence on this work.

I would like to give thanks to Dr Jelena Luković, Dr Milutin Pejović, Dr Dragutin Protić, and Melita Perčec Tadić for their support, comments and suggestions that helped me in writing this dissertation.

I would like to acknowledge R, `R-sig-geo`, and `r-spatial` community for developing free and open tools for spatial modeling, then Edzer Pebesma, Roger Bivand, Gräler Benedikt, Tomislav Hengl, Leo Breiman, Marvin Wright, Jan Elseberg and all other researchers and developers of R packages that made this dissertation possible.

I would also extend my gratitude to all the colleagues and associates that made writing of the dissertation easier by taking over my duties when it was needed.

I would also like to thank my friend Maša Maksimović and a colleague Dr Nikola Tošić for linguistic corrections.

I also feel that I should thank my closest friends for their support.

Last, but not least, I am deeply thankful to my family, my father Miroljub, mother Violeta, and sisters Jelena and Milica, and my girlfriend Miljana for unconditional love and support over all these years. I could not do this without them. A special dedication goes to my nephew Uroš.

I would also like to dedicate this dissertation to my late grandpa Dragutin. I know this would mean a lot to him.

Aleksandar Sekulić

UNIVERSITY OF BELGRADE

Faculty of Civil Engineering

Department of Geodesy and Geoinformatics

Abstract

Spatio-temporal interpolation of climate elements using geostatistics and machine learning

High resolution daily maps for climate elements are a valuable source of information and serve as an input for climatology, meteorology, agriculture, hydrology, ecology, and many other research areas and disciplines. Spatio-temporal interpolation methods are often used for creation of daily maps for climate elements. In this research, already existing spatio-temporal geostatistical interpolation methods and newly developed spatio-temporal interpolation methods based on machine learning algorithms are applied to and evaluated on climate element case studies. A spatio-temporal regression kriging model for global land areas for mean daily temperature is simplified by using only a geometric temperature trend, digital elevation model, and topographic wetness index (without MODIS LST) as covariates and adapted for Croatian territories for the year 2008 in this dissertation. The leave-one-out and 5-fold cross-validation show that the accuracy of the model after adaptation is 97.8% in R^2 and 1.2 °C in RMSE, which is an improvement of 3.4% in R^2 and 0.7 °C in RMSE. The adapted daily mean temperature model also outperforms previously developed models for Croatia and shows similar or better accuracy in comparison with models for other local areas. The results show that the spatio-temporal regression kriging model for global land areas can be adapted to local areas using a national weather station network, thus providing more accurate daily mean temperature maps at a 1 km spatial resolution. The proposed adapted geostatistical model for Croatia still provides larger prediction errors in mountainous regions making it convenient for application in agricultural areas that are at lower altitudes.

A different approach to spatial or spatio-temporal interpolation of climate elements is to use machine learning algorithms together with spatial covariates. A novel Random Forest Spatial Interpolation (RFSI) methodology for spatial or spatio-temporal interpolation is proposed and evaluated in this dissertation. The RFSI methodology is based on the Random Forest algorithm that uses innovative spatial predictors: observations at n nearest locations and distances to them. The RFSI methodology is applied and evaluated in three case studies. In the first, a synthetic (simulated) case study, the accuracy of RFSI is compared with the accuracy of ordinary kriging, Random Forest for spatial prediction (RFsp), inverse distance weighting, nearest neighbour, and trend surface mapping interpolation methods. In this case study, RFSI outperforms nearest neighbour and trend surface mapping and has similar accuracy as RFsp and inverse distance weighting. RFSI is outperformed by ordinary kriging because this case study is created by geostatistical simulation and consequentially ordinary kriging is an optimal interpolation method in this case. In the following two real-world case studies, a daily precipitation for Catalonia for the 2016–2018 period and a daily mean temperature for Croatia for the year 2008, the accuracy of RFSI is compared with the accuracy of spatio-temporal regression kriging, inverse distance weighting, standard Random Forest and RFsp using a nested

k-fold leave-location-out cross-validation and RFSI outperformed all of them. RFSI is recommended for the interpolation of complex variables due to Random Forest's ability to model non-linear relations between covariates and target variables. RFSI can be used for spatial or spatio-temporal interpolation of any environmental variable.

Next, a MeteoSerbia1km dataset — a first gridded dataset for daily climate elements (maximum, minimum, and mean temperature, mean sea level pressure, and total precipitation) at a 1 km spatial resolution for Serbian territories for the 2000–2019 period — is created using RFSI methodology for spatio-temporal interpolation. Additionally, monthly and annual summaries and daily, monthly, and annual long term means maps of the climate elements are generated by aggregating the daily MeteoSerbia1km maps. The nested 5-fold leave-location-out cross-validation is used to access the accuracy of the MeteoSerbia1km daily dataset. The accuracy is high for daily temperature variables and sea level pressure and lower for daily precipitation which was expected due to its complexity. MeteoSerbia1km daily maps are further compared with the 10-km E-OBS daily maps and show high correlation with them except for daily precipitation.

The automation of the RFSI methodology is implemented within the R package `meteo`, in the form of four new R functions for creation, prediction, tuning, and cross-validation processes of RFSI model.

Key words: spatio-temporal interpolation, kriging, machine learning, random forest, RFSI, daily temperature, daily precipitation, MeteoSerbia1km, R, meteo

Scientific field: Geodesy

Scientific subfield: Modelling and management in geodesy

UNIVERZITET U BEOGRADU

Građevinski fakultet

Odsek za geodeziju i geoinformatiku

Sažetak

Prostorno-vremenska interpolacija klimatskih elemenata primenom geostatistike i mašinskog učenja

Gridovani podaci dnevnih klimatskih elemenata visoke rezolucije predstavljaju značajan izvor informacija koje se koriste kao ulazni podaci za analize u klimatologiji, meteorologiji, poljoprivredi, hidrologiji, ekologiji i ostalim istraživačkim oblastima i disciplinama. Prostorno-vremenske interpolacione metode često se koriste za kreiranje gridovanih dnevnih klimatskih elemenata. Globalni model prostorno-vremenskog regresionog kriginga za srednje dnevne temperature iznad površi Zemlje je pojednostavljen koristeći samo geometrijski temperaturni trend, digitalni model terena i topografski indeks vlažnosti (bez MODIS LST snimaka) kao prediktore i kalibrisan za područje Hrvatske koristeći podatke iz 2008 godine u ovoj disertaciji. Na osnovu prostorne kros-validacije, tačnost kalibrisanog modela iznosi R^2=97.8% i RMSE=1.2 °C, što je poboljšanje od 3.4% i 0.7 °C u odnosu na globalni model. Prilagođeni model srednjih dnevnih temperatura nadmašuje ostale već razvijene modele za područje Hrvatske u pogledu tačnosti i ima sličnu ili veću tačnost u odnosu na modele za druga lokalna područja ili države. Rezultati pokazuju da se globalni model prostorno-vremenskog regresionog kriginga može prilagoditi lokalnim područjima koristeći mrežu nacionalnih meteoroloških stanica i tako proizvesti gridovane podatke srednjih dnevnih temperatura veće tačnosti sa prostornom rezolucijom od 1 km. Kalibrisani model za područje Hrvatske još uvek ima manju tačnost u planinskim predelima, što ga čini pogodnim za primenu u poljoprivrednim područjima koja su na nižim nadmorskim visinama.

Algoritmi mašinskog učenja kombinovani sa inovativnim prostornim prediktorima predstavljaju novi oblik modela za prostornu ili prostorno-vremensku interpolaciju, koji mogu da se koriste i za interpolaciju klimatskih elemenata. U ovoj disertaciji je predstavljena i testirana inovativna *Random Forest Spatial Interpolation* (RFSI) metodologija za prostornu ili prostorno-vremensku interpolaciju. RFSI metodologija je bazirana na *Random Forest* algoritmu mašinskog učenja koji koristi inovativne prostorne prediktore: opažanja na n najbližih lokacija i rastojanja do njih. RFSI metodologija je primenjena i testirana na tri studije slučaja. U prvoj sintetičkoj studiji, koja predstavlja simulirani set podataka, tačnost RFSI metodologije je poređena sa tačnošću običnog kriging-a, *Random Forest for spatial prediction* (RFsp) metode, metode inverznih distanci (eng. *inverse distance weighting*), najbližeg suseda (eng. *nearest neighbour*) i mapiranja površi trenda (eng. *trend surface mapping*). U ovom slučaju, RFSI je pokazao veću tačnost u poređenju sa metodama najbližeg suseda i mapiranja površi trenda i sličnu tačnost kao RFsp i metoda inverznih distanci. Obični kriging je očekivano dao bolje rezultate od RFSI metodologije iz razloga što je simulirani set podataka kreiran geostatističkom simulacijom i samim tim obični kriging predstavlja optimalnu metodu interpolacije u ovom slučaju. U ostale dve studije slučaja, koje se odnose na dnevne količine padavina za područje Katalonije za 2016–2018 period i srednje dnevne temperature za područje Hrvatske za 2008 godinu, tačnost

RFSI metodologije je poređena sa tačnošću prostorno-vremenskog regresionog kriginga, metode inverznih distanci, standardnom *Random Forest* i RFsp metodom koristeći ugnježdenu prostornu kros-validaciju. RFSI metodologija je pokazala najbolje rezultate u ovim studijama. RFSI metodologija se preporučuje za interpolaciju složenih parametara zbog osobine *Random Forest* algoritma da može da modelira nelinearne veze između prediktora i modeliranog parametra. RFSI metodologija se takođe može koristiti za prostornu ili prostorno-vremensku interpolaciju bilo kog drugog parametra životne sredine.

Koristeći RFSI metodologiju za prostorno-vremensku interpolaciju, kreiran je MeteoSerbia1km set podataka koji predstavlja prvi set gridovanih dnevnih klimatskih elemenata (maksimalne, minimalne i srednje temperature, atmosferskog pritiska na nivou mora i količine padavina) sa prostornom rezolucijom od 1 km za područje Srbije za period 2000–2019. Agregacijom dnevnih gridovanih podataka dodatno su kreirani gridovani podaci mesečnih i godišnjih proseka (ukupne količine za padavine) i gridovani podaci dnevnih, mesečnih i godišnjih dugoročnih proseka klimatskih elemenata. Tačnost dnevnih MeteoSerbia1km gridovanih podaka je ocenjena pomoću ugnježdene prostorne kros-validacije. Tačnost dnevnih temperatura i atmosferskog pritiska na nivou mora je visoka, dok je tačnost dnevnih padavina očekivano nešto manja zbog složenosti samih padavina. Dnevni MeteoSerbia1km gridovani podaci su takođe poređeni sa E-OBS setom dnevnih gridovanih podataka sa prostornom rezolucijom od 10 km i pokazuju visok stepen korelacije, osim za padavine.

RFSI metodolgija je automatizovana i implementirana u okviru R paketa `meteo`, kroz četiri nove R funkcije za procese kreiranja, predikcije, kalibrisanja i kros-validacije RFSI modela.

Ključne reči: prostorno-vremenska interpolacija, mašinsko učenje, random forest, RFSI, dnevne temperature, dnevne padavine, MeteoSerbia1km, R, meteo

Naučna oblast: Geodezija

Uža naučna oblast: Modeliranje i menadžment u geodeziji

Abbreviations

AER	Adjusted Error Rate
AINA	Aggregation and Interpolation of NEX Archives
AMSV	Automated Meteorological Station network in Vojvodina region
ANN	Artificial Neural Networks
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
BLUP	Best Linear Unbiased Predictor
BPNN	Back Propagation Neural Network
C3S	Climate Change Service
CAMS	Copernicus Atmosphere Monitoring Service
CEMS	Copernicus Emergency Management Service
CarpatClim	Climate of the Carpathian region
CART	Classification And Regression Trees
CCC	Lin's Concordance Correlation Coefficient
CDATE	Cumulative Day from a Date
CHRS	Center for Hydrometeorology and Remote Sensing
CK	Co-Kriging
CMDT	Croatian Mean Daily Temperature dataset
CPC	Climate Prediction Center
CRAN	Comprehensive R Archive Network
CRS	Coordinate Reference System
CRU	Climatic Research Unit
CS	Cubic Splines
DEM	Digital Elevation Model
DOY	Day Of Year
ECA&D	European Climate Assessment & Dataset
E-OBS	Ensembles daily gridded OBServational dataset
ECMWF	European Centre for Medium-Range Weather Forecasts
ESA	European Space Agency
ESS	Error Sum of Squares
EUMETNET	network of European National Meteorological Services
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FCC	Federal Climate Complex
GAM	Generalized Additive Model
GBM	Gradient Boosting Machines
GCV	Generalized Cross-Validation
GDAL	Geospatial Data Abstraction Library
GERB	Geostationary Earth Radiation Budget
GHCN-daily	daily Global Historical Climatological Network

GLS	Generalized Least Square
GMT	Greenwich Mean Time
GPCC	Global Precipitation Climatology Centre
GPM	Global Precipitation Measurement
GRF	Geographical Random Forest
GSOD	Global Surface Summary of the Day
GSW	Generalized Split-Window
GTT	Geometric Temperature Trend
GWR	Geographically Weighted Regression
IDW	Inverse Distance Weighting
IJPS	Initial Joint Polar System
IK	Indicator Kriging
IMERG	Integrated Multi-satellitE Retrievals for GPM
INSOL	Total Potential Insolation
IQR	Interquartile Range
ISD	Integrated Surface Database
JMA	Japan Meteorological Agency
JRA-55	Japanese 55-year Reanalysis
KED	Kriging with External Drift
KMNI	Koninklijk Nederlands Meteorologisch Instituut
KNN	K Nearest Neighbour
LLOCV	Leave-Location Out Cross-Validation
LOO	Leave-One-Out
LTM	Long Term Means
LTOCV	Leave-Time-Out cross-validation
MAE	Mean Absolute Error
MARS	Meteorological Archival and Retrieval System
MASH	Multiple Analysis of Series for Homogenization
MERRA-2	Modern Era Reanalysis for Research and Applications Version-2
MeteoSerbia1km	daily gridded meteorological dataset at a 1-km spatial resolution across Serbia
Metop-SG	Metop-Second Generation
MFG	Meteosat First Generation
MISH	Meteorological Interpolation based on Surface Homogenized Data Basis
ML	Machine Learning
MLR	Multiple Linear Regression
MMC	Maximal Margin Classifier
MODIS LST	Moderate Resolution Imaging Spectroradiometer Land Surface Temperature
MSG	Meteosat Second Generation
MTG	Meteosat Third Generation
MVIRI	Meteosat Visible and Infrared Imager
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCDC	NOAA's National Climatic Data Center
NCEI	National Centers for Environmental Information
NCEP	National Centers for Environmental Prediction
NEX-GDM	NASA Earth Exchange Gridded Daily Meteorology
NN	Nearest Neighbour
NNI	Natural Neighbour Interpolation

NNRK	ANN Residual Kriging
NOAA	National Oceanic and Atmospheric Administration
NWM	Numerical Weather Model
OGC	Open Geospatial Consortium
OI	Optimal Interpolation
OK	Ordinary Kriging
OLS	Ordinary Least Squares
OOB	Out-Of-Bag
PCC	Pearson Correlation Coefficient
PDIR-Now	PERSIANN Dynamic Infrared Rain Rate near real-time
PERSIANN	Precipitation Estimation from Remotely Sensed Information using ANN
PERSIANN-CCS	PERSIANN-Cloud Classification System
PERSIANN-CDR	PERSIANN-Climate Data Record
PRCP	Precipitation
Prec-DWARF	Precipitation Downscaling With Adaptable Random Forests
PSL	Physical Sciences Laboratory
QA	Quality Assurance
QRF	Quantile Regression Forest
RAM	Random-Access Memory
RER	Reduction in the node's Error Rate
RF	Random Forest
RF-MEP	Random Forest based MErging Procedure
RFSI	Random Forest Spatial Interpolation
$RFSI_0$	RFSI without environmental covariates
RFsp	Random Forest for Spatial Predictions framework
RK	Regression Kriging
RKNNRK	Regression Kriging and ANN Residual Kriging
RMSE	Root Mean Square Error
RS	Remote Sensing
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SK	Simple Kriging
SLP	Sea Level Pressure
STRK	Spatio-Temporal Regression Kriging
STRK_Croatia	STRK model for Croatia
STRK_global	global STRK model
SV	Support Vectors
SVC	Support Vector Classifier
SVM	Support Vector Machines
SYNOP	Surface Synoptic Observations
TES	Temperature Emissivity separation
TIN	Triangulated Irregular Network
TMAX/Tmax	Maximum Temperature
TMEAN/Tmean	Mean Temperature
TMI	TRMM Microwave Imager
TMIN/Tmin	Minimum Temperature
TPS	Thin Plate Spline
TRMM	Tropical Rainfall Measuring Mission
TS	Trend Surface mapping

TS2	Trend Surface mapping of second-order
TSS	Total Sum of Squares
TWI	Topographic Wetness Index
UCI	University of California, Irvine
UK	Universal Kriging
UTC	Coordinated Universal Time
UTM	Universal Transverse Mercator
WCS	Web Coverage Service
WGS	World Geodetic System
WLS	Weighted Least Square
WMO	World Meteorlogical Organization
WMS	Web Map Service

Contents

Acknowledgements **v**

Abstract **vii**

Sažetak **ix**

Abbreviations **xi**

List of Figures **xix**

List of Tables **xxiii**

1 Introduction **1**
1.1 Motivation and problem statement . . . 1
1.2 Research objectives . . . 2
1.3 Research methodology . . . 3
1.4 Outline . . . 3

2 Spatial interpolation methods and their application to climate elements **5**
2.1 Introduction . . . 5
2.2 Station-based interpolation methods . . . 8
2.2.1 Deterministic Methods . . . 8
2.2.1.1 Nearest Neighbours . . . 8
2.2.1.2 Triangulated Irregular Network . . . 9
2.2.1.3 Natural Neighbour Interpolation . . . 9
2.2.1.4 Inverse Distance Weighting . . . 9
2.2.1.5 Trend Surface Mapping . . . 10
2.2.1.6 Splines and local trend surfaces . . . 10
2.2.1.7 Thin plate splines . . . 11
2.2.2 Geostatistical methods . . . 11
2.2.2.1 Ordinary kriging . . . 12
2.2.2.2 Simple kriging . . . 15
2.2.2.3 Indicator kriging . . . 15
2.2.2.4 Co-kriging . . . 15
2.2.2.5 Universal kriging . . . 16
2.2.2.6 Geostatistical simulations . . . 16
2.3 Covariate-based interpolation methods . . . 17
2.3.1 Linear regression methods . . . 17
2.3.1.1 Multiple linear regression . . . 17
2.3.1.2 Geographically weighted regression . . . 18
2.3.1.3 Generalized additive models . . . 18

2.3.2 Machine learning methods . . . 19
2.3.2.1 Random Forest, Gradient Boosting Machine and Cubist . . . 19
2.3.2.2 Artificial neural networks . . . 22
2.3.2.3 Support vector machines . . . 24
2.4 Combined methods . . . 25
2.4.1 Residual (regression) kriging . . . 25
2.4.2 Residual IDW . . . 27
2.4.3 Kriging with external drift . . . 27
2.4.4 Spatial machine learning methods . . . 27

3 Open daily climate datasets 29
3.1 Introduction . . . 29
3.2 Observational data . . . 30
3.2.1 OGIMET service . . . 31
3.2.2 GSOD . . . 31
3.2.3 GHCN-daily . . . 31
3.2.4 ECA&D . . . 32
3.3 Gridded data . . . 32
3.3.1 Station-based datasets . . . 33
3.3.1.1 E-OBS . . . 33
3.3.1.2 CPC . . . 33
3.3.1.3 CarpatClim . . . 33
3.3.2 Remote sensing products . . . 34
3.3.2.1 MODIS LST . . . 34
3.3.2.2 TRMM/IMERG . . . 35
3.3.2.3 PERSIANN . . . 36
3.3.2.4 EUMETSAT products . . . 36
3.3.3 Climate reanalysis datasets . . . 37
3.3.3.1 NOAA datasets . . . 38
3.3.3.2 ECMWF datasets . . . 39
3.3.4 Other environmental covariates . . . 39

4 Adaptation of global geostatistical mean daily temperature model to local areas 43
4.1 Introduction . . . 43
4.2 Study area and datasets . . . 45
4.2.1 Study area . . . 45
4.2.2 Datasets . . . 46
4.3 Methods . . . 47
4.3.1 Spatio-temporal regression-kriging . . . 47
4.3.2 Mean daily temperature model for global land areas . . . 48
4.3.3 Mean daily temperature model for Croatia . . . 48
4.3.4 Accuracy assessment . . . 49
4.4 Results . . . 49
4.4.1 Mean daily temperature model for global land areas and prediction . . . 49
4.4.2 Mean daily temperature model calibration for Croatia and prediction . . . 50
4.4.3 Accuracy assessment . . . 51
4.5 Discussion . . . 55
4.5.1 Global vs local model . . . 55
4.5.2 Mean daily temperature model for Croatia and comparison with other models 56
4.6 Conclusions . . . 59

5 Spatial and spatio-temporal interpolation using random forest 61
5.1 Introduction . . . 61
5.2 Materials and Methods . . . 63
5.2.1 Methodology . . . 63
5.2.1.1 Random Forest and RFsp . . . 64
5.2.1.2 Random Forest Spatial Interpolation . . . 64
5.2.2 Datasets and Covariates . . . 64
5.2.2.1 Synthetic Dataset . . . 64
5.2.2.2 Precipitation Dataset . . . 65
5.2.2.3 Temperature Dataset . . . 66
5.2.3 Accuracy Assessment . . . 66
5.2.3.1 Synthetic Case Study . . . 68
5.2.3.2 Real-World Case Studies . . . 69
5.3 Results . . . 69
5.3.1 Synthetic Case Study . . . 69
5.3.2 Precipitation Case Study . . . 73
5.3.2.1 Spatio-Temporal Regression Kriging (STRK) . . . 73
5.3.2.2 IDW and Random Forest Models . . . 74
5.3.2.3 Accuracy Assessment . . . 76
5.3.3 Temperature Case Study . . . 79
5.3.3.1 IDW and Random Forest Models . . . 79
5.3.3.2 Accuracy Assessment . . . 79
5.4 Discussion . . . 81
5.4.1 RFSI Performance . . . 81
5.4.2 Extensions and Improvements . . . 86
5.5 Conclusions . . . 86

6 Spatial and spatio-temporal interpolation of daily climate elements for Serbian territory at 1 km spatial resolution 89
6.1 Background & Summary . . . 89
6.2 Methods . . . 91
6.2.1 Study area . . . 91
6.2.2 Source data . . . 91
6.2.2.1 OGIMET . . . 91
6.2.2.2 DEM and TWI . . . 91
6.2.2.3 IMERG . . . 92
6.2.2.4 E-OBS . . . 92
6.2.2.5 Automated meteorological stations in Vojvodina region . . . 92
6.2.3 RFSI . . . 92
6.2.3.1 Model development and prediction . . . 93
6.2.3.2 Model tuning . . . 93
6.2.4 Modelling of daily meteorological variables . . . 93
6.2.4.1 Temperature . . . 93
6.2.4.2 Sea level pressure . . . 94
6.2.4.3 Precipitation . . . 94
6.3 Data Records . . . 95
6.4 Technical Validation . . . 96
6.4.1 Validation of daily datasets . . . 96
6.4.2 Comparison with E-OBS . . . 97
6.4.3 Test with stations in Vojvodina region . . . 98

6.5 Usage Notes . . . 99
6.6 Code availability . . . 100
6.7 Discussion and conclusions . . . 101

7 Automated spatio-temporal interpolation using R package meteo 105
7.1 Introduction . . . 105
7.2 R programming language . . . 106
7.3 R package meteo . . . 106
7.3.1 Related R packages . . . 107
7.3.1.1 sp . . . 107
7.3.1.2 spacetime . . . 107
7.3.1.3 gstat . . . 108
7.3.1.4 ranger . . . 108
7.3.1.5 nabor . . . 108
7.3.1.6 snowfall and doParallel . . . 109
7.3.1.7 rgdal and raster . . . 109
7.3.2 Spatio-temporal regression kriging . . . 109
7.3.2.1 Prediction . . . 110
7.3.2.2 Cross-validation . . . 112
7.3.3 Random Forest Spatial Interpolation . . . 114
7.3.3.1 Model development . . . 114
7.3.3.2 Prediction . . . 116
7.3.3.3 Model tuning . . . 118
7.3.3.4 Cross-validation . . . 120
7.4 Discussion and conclusions . . . 121

8 Discussion and conclusions 123

Bibliography 127

Biography 145

Prilozi 147

List of Figures

2.1 A map that shows the geostatistically simulated reality (SIM) and the locations of the 500 samples used to create prediction maps by all deterministic methods presented in this section (NN, TIN, NNI, IDW, TS2, CS, TPS). . . . 8
2.2 Sample and fitted semivariogram. . . . 13
2.3 The shape of linear, spherical, exponential, and Gaussian semivariogram models. . . . 14
2.4 RF algorithm scheme. . . . 20
2.5 ANN with two neuron layers in the hidden layer. . . . 23

3.1 DEMSRE3, TWISRE3, INMSRE3, and DICGSH1 provided by worldgrids.org. . . . 41

4.1 Spatial distribution of GSOD (blue circles) and CMDT (green squares) meteorological stations for mean daily temperature, CMDT stations which are included in GSOD dataset (orange diamond), and CMDT stations with missing DEM and TWI values (red triangles). . . . 46
4.2 DEM (left) and TWI (right) values for Croatia. . . . 47
4.3 The scatterplot of estimated mean daily temperature values from the trend for STRK_Croatia vs. observed values, (left). Histogram of the residuals from the trend for STRK_Croatia (right). It shows that residuals follow the normal distribution which justifies the use of the kriging. . . . 50
4.4 Sample semivariogram (left) and fitted sum-metric semivariogram (right) of the residuals from the trend model for STRK_Croatia. Semivariograms are presented in 3D. . . . 51
4.5 Annual average RMSE per station for testing of STRK_global predictions made by using GSOD stations (http://osgl.grf.bg.ac.rs/materials/tac_hr/). RMSE values are presented by the radius of the circles. . . . 52
4.6 Annual average RMSE per station. Results of LOO cross-validation, STRK_global on the left and STRK_Croatia on the right (http://osgl.grf.bg.ac.rs/materials/tac_hr/). RMSE values are presented by the radius of the circles. . . . 52
4.7 Scatter plot DEM vs annual average RMSE from LOO cross-validation, STRK_global on the left and STRK_Croatia on the right. Stations at altitudes above 1000 m (red) in the top right corner have the highest RMSEs. Notice the smaller scale on the y-axis for the STRK_Croatia model. . . . 53
4.8 Boxplot of the altitude per fold. . . . 53
4.9 Annual average RMSE per station for 5-fold cross-validation, STRK_global on the left and STRK_Croatia on the right (available at http://osgl.grf.bg.ac.rs/materials/tac_hr/). RMSE values are presented by the radius of the circles. . . . 54
4.10 Maps of predicted mean daily temperatures at 1 km spatial resolution using STRK_global on the left and STRK_Croatia on the right with CMDT stations for the first 4 days of January 2018 for Croatia. . . . 57

4.11 Time series of predictions from LOO cross-validation (red—STRK_global, blue—STRK_Croatia) and observations (green) for station Zavižan and station Zagreb-Maksimir. 58

5.1 Schematic representation of the RFSI algorithm. 65
5.2 GHCN-daily station locations on top of a digital elevation model of the study area (left) and histogram of daily precipitation for Catalonia (right). The histogram contains 92,404 GHCN-daily observations for the 2016—2018 period. 66
5.3 Maximum temperature (left), minimum temperature (middle) and IMERG precipitation estimates (right) for fourexample days, 1–4 January 2016. 67
5.4 Station locations in Croatia on top of a digital elevation model of the study area. . . 68
5.5 Comparison of average MAE estimated for each of the interpolation methods, for all nugget-to-sill ratios and ranges. Coloured bars are average MAE for test locations from 100 different simulations. Error bars are standard errors computed from 100 simulations. 70
5.6 Comparison of $R^2_{1:1}$ (top left), CCC (top right), MAE (bottom left) and RMSE (bottom right) estimated for each of the interpolation methods, for nugget-to-sill ratio 0.25 and range 200.Coloured bars are average accuracy metrics for test locations computed from 100 different simulations. Error bars are standard errors computed from 100 simulations. 71
5.7 Prediction maps made using 500 sample locations with nugget-to-sill ratio 0.25 and range 50, for one of the 100 simulated realities. The top left map (SIM) shows the simulated reality and the locations of the 500 samples. 72
5.8 Covariate importance plot for RFsp (left) and RFSI (right), for the case shown in Figure 5.7. The importance index is scaled to a maximum of 1, obs$i$ and dist$i$ represent observations and distances to the i-th nearest observation location, and layer.i represents buffer distances to the i-th observation location. 73
5.9 RFSI prediction maps made using 500 sample locations with nugget-to-sill ratio 0.25 and range 50, with different number of nearest locations (n).

. 74
5.10 RMSE vs number of nearest locations (n) used in RFSI for one simulation with nugget-to-sill ratio 0.25, ranges 50 and 200, using 100 (top), 500 (middle) and 2000 (bottom) sample locations. Larger discs represent the optimal number of nearest locations with minimum RMSE. 75
5.11 Histogram of STRK residuals. Residuals smallerthan −20 mm (0.2% of total residuals) and greater than +20 mm (1.2% of total residuals) are not shown. 75
5.12 STRK sample semivariogram and fitted sum-metric semivariogram. 76
5.13 Covariate importance plot for RF (left), RFsp (middle) and RFSI (right), for the precipitation case study. The importance index is scaled to a maximum of 1. The importance of covariates IMERG, TMAX, and TMIN is shown in red. 77
5.14 Scatter density plots of predictions vs. observations with 1:1 line for the precipitation case study. 78
5.15 Prediction maps of daily precipitation (mm) for the five models, for 1–4 January 2016. The bottom row shows the maximum observed precipitation for each day. 80
5.16 IQR of daily precipitation (mm) for the four models, for 1–4 January 2016. 81
5.17 Covariate importance plot for RF (left), RFsp (middle) and RFSI (right), for the temperature case study. The importance index is scaled to a maximum of 1. The importance of environmental covariates is shown in red. 82

5.18 Prediction maps of daily temperature (°C) for the four models, for 2 February 2008. The bottom row shows the maximum and minimum observed temperature. 83

6.1 SYNOP station locations used for making MeteoSerbia1km with DEM. 91
6.2 Prediction maps for all daily meteorological variables, for July 27, 2014. 95
6.3 Average RMSE per station for the 2000–2019 period, calculated from the nested 5-fold LLOCV. The units are °C for temperature, mbar for SLP and mm for PRCP. . . . 98
6.4 Predictions from the nested 5-fold LLOCV (red) and observations (black) for station Belgrade for year 2014. 99
6.5 Pearson correlation coefficient map between E-OBS and the daily MeteoSerbia1km datasets for Serbia. 100
6.6 Annual LTM maps for all daily meteorological variables. 102
6.7 Montlhy LTM maps for all daily meteorological variables, for January. 103
6.8 Montlhy LTM maps for all daily meteorological variables, for July. 104

7.1 An algorithm for STRK prediction using the `pred.strk` function. 111
7.2 An algorithm for k-fold LLOCV of the STRK model using the `cv.strk` function. . 113
7.3 An algorithm for RFSI model development using the `rfsi` function. 115
7.4 An algorithm for RFSI prediction using the `pred.rfsi` function. 117
7.5 An algorithm for tuning of the RFSI model using the `tune.rfsi` function. 119
7.6 An algorithm for nested n-fold LLOCV of the RFSI model using the `cv.rfsi` function. 121

List of Tables

4.1 Sum-metric semivariogram parameters for the STRK_global (Kilibarda et al. 2014). . 50
4.2 Sum-metric semivariogram parameters for the STRK_Croatia. 51
4.3 RMSE values [°C] for each month for STRK_global and STRK_Croatia obtained by LOO and 5-fold cross-validation and differences between them. 54

5.1 Distance calculation time and modelling time for RFSI and RFsp, and prediction time for RFSI, RFsp and OK. All results refer to the synthetic case study and represent the average computing time computed from 100 simulations. All calculations and time estimations were done on a personal computer with Intel® Core™ i7-7820X CPU @ 3.60GHz × 16 processor and 126 GB of RAM. 72
5.2 Sum-metric semivariogram parameters of the STRK model. 74
5.3 Optimized hyperparameters of IDW, RF, RFsp and RFSI for the precipitation case study. 76
5.4 Accuracy metrics of all six prediction methods as assessed using nested 5-fold LLOCV for the precipitation case study. 76
5.5 Performance of all models for precipitation below and above 1 mm. Values for the hits and misses are number of observations (obs.) for a given condition. Predictions (pred.) used in this table are from nested 5-fold LLOCV. Overall accuracy represents the percentage of correct classifications. Numbers in bold represent the best performance. 78
5.6 Optimized hyperparameters of IDW, RF, RFsp, and RFSI for the temperature case study. 79
5.7 Accuracy metrics of all six prediction methods as assessed using nested 10-fold LLOCV for the temperature case study. Note that accuracy metrics for STRK are taken from Hengl et al. (2012). 82

6.1 Summary statistics for the selected variables in OGIMET daily summaries for the 2000–2019 period. 92
6.2 Optimized hyperparameters for each of the daily meteorological variables. 94
6.3 MeteoSerbia1km dataset file naming convention. 96
6.4 Accuracy metrics for each meteorological variable for stations in Serbia, as assessed using the nested 5-fold LLOCV. 97
6.5 Confusion Matrix for PRCP RFSI classification model from the nested 5-fold LLOCV. Class 0 represents no precipitation, and class 1 represents precipitation occurrence. 97

Chapter 1

Introduction

1.1 Motivation and problem statement

"*Climate describes the average weather conditions for a particular location and over a long period of time.* " (World Meteorological Organization (WMO) 2018). According to WMO, the main climate elements are air temperature (including maximum and minimum temperatures), precipitation (rainfall, snowfall, etc.), humidity, wind (speed and direction), atmospheric pressure, evaporation, and solar radiation. Climate elements affect people in many ways. On the one hand, severe weather conditions, such as thunderstorms, snowstorms, tornadoes, hurricanes, and others, can cause harm to people. On the other hand, people benefit from good weather conditions. Accurate daily maps of climate elements (or daily gridded climate elements) are the basis for climate change analysis (Haylock et al. 2008; Lukovic et al. 2021) based on which near future weather can be predicted in order to help people in planning future activities. Daily maps of climate elements describe a complete spatial variability of the climate elements over an observed area and give the best estimates of climate elements at any spatial location away from weather stations (Haylock et al. 2008). Also, they mostly represent a regular time series covering larger time periods. As such, they allow climate change analysis over regions with a sparse distribution of weather stations and local areas completely without weather stations. Besides that, daily gridded climate elements are preferred over observations from weather stations in many research areas, such as agriculture, ecology, forestry, health and disease, meteorology, hydrology, transport, urban environments, and energy (Chapman and Thornes 2003).

Researchers use spatial or spatio-temporal interpolation methods (or just interpolators), to create daily gridded climatological datasets from observations at weather stations and other environmental covariates. The main advantage of spatial interpolators is that they take spatial dependency of the weather stations into account. Furthermore, spatio-temporal interpolators additionally consider temporal and spatio-temporal dependency.

Geostatistical interpolation methods (known also as kriging) are among the most popular in the last two decades. Spatio-temporal regression kriging (Heuvelink and Griffith 2010) is a kriging version for spatio-temporal variables, such as climate elements, that combines environmental covariates with spatio-temporal correlation between observations. Thanks to its implementation in the R (R Development Core Team 2012) package `gstat` (Pebesma 2004; Gräler et al. 2016), it was recently used for the interpolation of climate elements (Hengl et al. 2012; Kilibarda et al. 2014). Spatio-temporal regression kriging is a natural choice considering the fact that climate elements vary in space and time. Kilibarda et al. (2014) used spatio-temporal regression kriging for an interpolation of daily temperature (mean, maximum, and minimum) over global areas. The question is

if this global spatio-temporal geostatistical model for daily climate elements (temperature) can be adapted to local areas and a larger number of observations at weather stations and improve accuracy?

The problem with kriging is that it makes many assumptions and may not be applicable to the complex variables which do not satisfy a stationarity condition. Nowadays, machine learning (ML) algorithms have found application in all spheres of society and research areas (Samardžić-Petrović et al. 2016; Kovačević et al. 2020), and so in spatial interpolation. At the very beginning, ML algorithms were used to model complex and non-linear relations only between environmental covariates and a target variable. The problem is that ML algorithms are not explicitly spatial and the question is how to make them able to take spatial autocorrelation of the observations into account. There are several approaches to this problem in the literature. One approach is to apply geostatistics on residuals (Li et al. 2011; Appelhans et al. 2015; Seo et al. 2015; Xu et al. 2020), but the flaw is in that two independent models should be fitted (ML for trend and geostatistical for residuals) and predictions are also made in two steps, by summarizing predictions from ML and geostatistical models. Another approach is to add spatial coordinates, latitude and longitude (Li et al. 2011; He et al. 2016; Mohsenzadeh Karimi et al. 2018; Čeh et al. 2018; Georganos et al. 2019) or x- and y- coordinates in projection (Behrens et al. 2018), to the ML model as covariates. The problem with spatial coordinate covariates is that they might lead to artefacts in prediction maps (horizontal and vertical strips are visible). An ad-hoc solution to this is to use coordinates along several axes tilted at an oblique angle called "oblique geographic coordinates" (Møller et al. 2020), but to get rid of the artifact, a large number of these rotated coordinates have to be used. Spatial context can also be introduced through multi-model approach, where one ML model is fitted for each of the observation location and prediction is made using the ML model of the nearest observation (Georganos et al. 2019; Hashimoto et al. 2019), but as the number of observation locations increases, so does the number of ML models. One of the most promising approaches is to introduce innovative "spatial" covariates. So far these, so called spatial covariates have mostly been distance-based, such as distance-to-coast (Li et al. 2011), distance-to-closest dry grid cell (He et al. 2016), distances to the corners and center of a bounding box around the sampling locations (Behrens et al. 2018), "buffer distance maps" from observation points (Hengl et al. 2018), and others. The applicability of these approaches to the interpolation of climate elements has also been exploited and showed good results (Hengl et al. 2018; Zhu et al. 2019; Hashimoto et al. 2019). Predictions from geostatistical and most deterministic interpolation methods represent a (linear) combination of nearest observations. This is an important principle from perspective that nearer observations are more correlated than those that are further away. So far, this principle has not been included into ML algorithms for spatial interpolation, at least not in a direct way so that the ML predictions actually represent a combination of nearest observations. As this principle has already given the best results in geostatistical methods, a logical approach would be to evaluate it in a combination with ML algorithms.

Even though daily gridded climate datasets exist at a wide range of spatial and temporal resolutions (Sekulić et al. 2020b), there is a need for more accurate high-resolution localized data. Also, there is still no gridded daily climatological dataset at a 1 km spatial resolution for Serbian territories.

1.2 Research objectives

The main objective of this research is to improve existing spatial and spatio-temporal interpolation methods for climate elements, mainly by reaping the benefits of ML algorithms, more precisely of an Random Forest (RF) algorithm. The main contribution of this research is the development of a Random Forest Spatial Interpolation (RFSI) methodology which uses the RF algorithm together with observations at nearest stations and distances to them as spatial covariates. This methodology

is intended for spatial or spatio-temporal interpolation of not only climate elements, but all environmental variables. The RFSI model, unlike other deterministic and geostatistical models, benefits from the RF's ability to model complex and non-linear relations between all spatial and environmental covariates.

The research objectives are as follows:

1. To examine if global spatio-temporal geostatistical models for daily climate elements can be adapted to local areas and thus improve prediction accuracy. This objective represents a continuation of research conducted by Kilibarda (2013).

2. To develop and evaluate a spatial / spatio-temporal interpolation methodology based on the RF algorithm and observations at stations and distances to them, called RFSI, and compare its performance with commonly used spatial and spatio-temporal interpolators.

3. To analyse the influence of observations at stations and distances to them as spatial covariates on spatial or spatio-temporal interpolation of climate elements.

4. To create and evaluate the first gridded daily climatological dataset at a 1 km spatial resolution for Serbia based on the RFSI methodology.

5. To automate the process of spatial (spatio-temporal) interpolation with RFSI.

1.3 Research methodology

Given from the previous section, the dissertation relies on spatio-temporal regression kriging and ML algorithms, i.e. a newly developed RF-based RFSI methodology. Spatial interpolation methods, such as nearest neighbour, inverse distance weighting, trend surface mapping, ordinary kriging, standard RF, and RF for spatial prediction (RFsp) (Hengl et al. 2018) were used for performance comparison with RFSI. Climate element maps are visually analysed. Geostatistical and ML models for interpolation of daily climate elements are evaluated and compared by calculating accuracy metrics, such as the coefficient of determination (R^2), Lin's concordance correlation coefficient (CCC) (Lin 1989), mean absolute error (MAE), and root mean square error (RMSE), estimated from cross-validation procedure. The accuracy of the adapted spatio-temporal regression kriging model for the mean daily temperature for Croatia is assessed using a leave-location-out and a k-fold leave-location-out cross-validation, while a nested k-fold leave-location-out cross-validation (Meyer et al. 2018; Pejović et al. 2018) is used for the evaluation of the RFSI methodology.

1.4 Outline

The dissertation consists of eight chapters, including the Introduction and Discussion and conclusions chapters (Chapters 1 and 8).

The background information about the research topic, discussion about the problems that this dissertation deals with and outlines of the main research objectives have been described in this chapter (Chapter 1).

Chapter 2 gives a detailed overview of the commonly used methods for spatial interpolation of climate elements, emphasizing recently developed interpolation methods based on ML algorithms, so called "spatial machine learning methods", with their application to climate elements.

Chapter 3 presents the open daily observational and gridded climatological datasets together with environmental covariates that are used or discussed in this dissertation.

In Chapter 4, the global spatio-temporal regression kriging model for daily mean temperature at a 1 km spatial resolution (Kilibarda et al. 2014) is adapted to a case study for the territory of Croatia for the year 2008. The accuracy of the model is further accessed using leave-one-out and 5-fold cross-validations. This Chapter is based on an article Sekulić et al. (2020b).

Chapter 5 presents and analyses a new methodology, called RFSI, for spatial or spatio-temporal interpolation using the RF ML algorithm together with observations at the nearest locations and their distances from the prediction location as spatial covariates. The RFSI methodology is evaluated on three case studies: one synthetic (simulated) and two real-world climate element case studies — a daily precipitation for Catalonia, Spain, for the 2016–2018 period and a daily mean temperature for Croatia for the year 2008 (the same case study as in Chapter 4). In the synthetic case study, the accuracy of the RFSI is compared with ordinary kriging, RFsp (Hengl et al. 2018), and simple deterministic interpolation methods (inverse distance weighting, nearest neighbour, and trend surface mapping interpolation methods), while in the real-world case studies it is compared with spatio-temporal regression kriging, inverse distance weighting, standard RF, and RFsp interpolation methods using the nested k-fold cross-validation. This Chapter is based on an article Sekulić et al. (2020a).

Chapter 6 presents MeteoSerbia1km — a first gridded daily climatological dataset at a 1 km spatial resolution for Serbia, produced using the newly developed RFSI methodology. The dataset contains daily maps for five climate variables: maximum, minimum and mean temperature, mean sea level pressure, and total precipitation. The daily maps are further aggregated to produce monthly and annual summaries, daily, monthly, and annual long term means (LTM). The MeteoSerbia1km daily dataset is evaluated using the nested 5-fold leave-location-out cross-validation and compared with the Ensembles daily gridded observational dataset (E-OBS) dataset (Cornes et al. 2018).

Chapter 7 mainly describes the implementation of the RFSI methodology in the R package `meteo` (Kilibarda et al. 2014) in the form of four new functions for the automation of creation, prediction, tuning and cross-validation processes. It also presents an improvement of the existing function for prediction and newly developed function for cross-validation for the spatio-temporal regression kriging interpolation.

Chapter 8 gives a summary of the main conclusions from this dissertation and discusses future work.

Chapter 2

Spatial interpolation methods and their application to climate elements

This chapter gives a detailed overview of commonly used methods for spatial interpolation of climate elements and recently developed methods based on machine learning algorithms with their application to real-word case studies. Methods that interpolate based only on spatial dependencies between station observations are described first, followed by the methods based only on relations between covariates and the target variable. Lastly, an overview of methods that combine spatial correlation between station observations and covariates is given, with emphasis on literature review of state-of-the-art methods that use machine learning algorithms together with, so called, "spatial covariates" — spatial machine learning methods.

2.1 Introduction

Meteorological stations are a valuable source of information about climate elements. They collect observations (measurements or sample data) very frequently over time, from every 24 hours to 15 seconds. Even though the optimisation of stations sampling density and choosing optimal stations spatial locations can be done prior to development of the station network (Wadoux et al. 2020), it is not possible to cover the whole variability of the climate elements in the spatial domain due to practical reasons, especially in the case of a large area. No matter how many stations are in the field, there is always a lack of information about a climate element in-between the stations. In order to solve this problem, spatial interpolation is performed. Spatial interpolation is a process of prediction at unobserved spatial locations. For this purpose, a spatial interpolation model that takes advantage of spatial correlation between observations and/or relations between observations and environmental covariates, can be developed and evaluated. Spatial interpolation models can be extended to spatio-temporal interpolation models, which can further model temporal and spatio-temporal correlations between observations.

Using spatial or spatio-temporal interpolation (hereinafter referred to as interpolation) one can predict at single or multiple spatial or/and temporal locations. The main interpolation products are grids or maps, which represent georeferenced images with known pixel size (spatial resolution), where each pixel value represents a target variable value estimated using an interpolation model. The word "georeferenced" means that the spatial position of an image is known in a real world, i.e. each image pixel has its own geographical coordinates. Gridding is a synonym for interpolation and interpolation methods are often called interpolators.

In climatology and meteorology, interpolation methods are often used to produce gridded

datasets of climate elements. Beek (1991), Hartkamp et al. (1999), Tveito et al. (2006), Dobesch et al. (2007), and Sluiter (2009) gave an extensive overview of interpolation methods for climate elements. From these literature reviews it can be seen that various interpolation methods were used for the creation of climate elements gridded datasets. Also, the interpolation of climate elements has evolved over time and has become more complex, following the development and improvement of spatial interpolation methods in general. At the very beginning, in the early 90s, basic deterministic interpolators, such as Nearest Neighbour (NN), Trend Surface Mapping (TS), and moving averages, together with splines, were mostly used. Even though kriging was developed in the early 60s (Matheron 1963), as Beek (1991) observed: "*Little experience has been obtained with kriging in meteorological fields.*". In the review of Hartkamp et al. (1999), interpolators like Triangulated Irregular Network (TIN), Inverse Distance Weighted (IDW), Thin Plate Spline (TPS), and co-kriging (CK) were used the most. In the first decade of 21th century, kriging has become the most popular interpolator for climate elements, which is probably related to expansion of kriging implementation in various programming languages, such as R package `gstat` (Pebesma 2004; Gräler et al. 2016). Tveito et al. (2006), Dobesch et al. (2007), and Sluiter (2009) all concluded that various kriging versions, especially ones that introduce environmental covariates in the interpolation process, such as Universal Kriging (UK), Kriging with External Drift (KED), and Regression Kriging (RK), are often preferred solutions for the interpolation of climate elements. They also review physically based and methods specially developed for climate elements interpolation like PRISM (Daly et al. 1994), AURELHY (developed by Meteo France, Bénichou 1994), and Meteorological Interpolation based on Surface Homogenized Data Basis (MISH, developed by Hungarian Meteorological Service).

Back then, except for Artificial Neural Networks (ANN) (Tveito et al. 2006; Sluiter 2009), machine learning algorithms were rarely used for interpolation of climate elements. In the past years, ML algorithms have become increasingly popular and are often used in spatial interpolation (Li et al. 2011). They are also used in the interpolation of climate elements because they are capable of modeling complex and non-linear processes of climate elements. Many researchers compare ML methods with deterministic methods and kriging. For example, Appelhans et al. (2015) compared various ML, kriging methods, and their combinations in the form of residual kriging for spatial interpolation of monthly air temperature and concluded that ML methods, alone and combined with residual kriging, mostly outperform kriging. da Silva Júnior et al. (2019) compared IDW, Ordinary Kriging (OK), and two versions of RF ML algorithm for spatial interpolation of evapotranspiration in the northeast region of Brazil and showed that regular RF model outperformed both kriging and IDW.

In the past few years, a popular research topic in the field of spatial interpolation focuses mostly on how to include spatial context directly into ML models. By including it, more complex and non-linear spatial relations between observations along with environmental covariates can be modelled with one unique ML model. Quite a few newly developed ML methods for spatial interpolation were evaluated on climate elements case studies and show that these methods mostly give equal or better results than geostatistical interpolators (Hengl et al. 2018; Zhu et al. 2019; Hashimoto et al. 2019; Sekulić et al. 2020a). Nevertheless, geostatistical interpolators are still widely used because they have far more interpretative power than ML algorithms (Hengl et al. 2018). The ML methods for spatial interpolation are explained in Section 2.4.4.

Interpolation methods can be observed from various aspects and thus be divided into groups based on different criteria (Hartkamp et al. 1999; Tveito et al. 2006; Sluiter 2009; Li et al. 2011):

- *Regression vs Classification*: Regression interpolators predict numerical (continuous) values, while classification interpolators predict categorical (discrete) values (classes).
- *Global vs Local*: Global interpolators use all of the observations for the creation of a spatial interpolation model and prediction, while local interpolators use a limited number of neighbouring observations for the creation of one unique or multiple spatial interpolation models

and prediction. Global interpolators generally make a smoother interpolation surface and therefore are useful for investigation of long-range variations. In the cases where analysis of local anomalies is needed, local interpolators are used.

- *Exact vs Approximate*: Exact interpolators predict an identical value to an observation. Opposite from the exact interpolators are approximate interpolators, which assume uncertainty in predictions at observation locations.

- *Deterministic vs Stohastic*: Deterministic interpolators create an interpolation surface based on the geometric characteristics of observations, i.e. they use mathematical laws in order to determine weights for the creation of interpolation surface, while stochastic interpolators use statistics, i.e. probabilistic theory for the same.

- *Gradual vs Abrupt*: Gradual interpolators produce a smooth (gradual) surface, while abrupt interpolators produce a discrete (abrupt) surface.

- *Convex vs Non-convex*: On one hand, convex interpolators predict in the observation values domain, i.e. in-between minimum and maximum observations. On the other hand, non-convex interpolators can predict outside of the observation values domain.

- *Univariate vs Multivariate*: Univariate interpolators use observations of the target variable for prediction, while multivariate interpolators additionally use one or more auxiliary variables (co-variables) for the prediction of the target variable.

- *Linear vs Non-linear*: Linear interpolators make predictions based on a linear combination of covariates or observations, while non-linear are based on non-linear combinations.

Usually, interpolation methods are divided into groups of deterministic, stochastic (probabilistic), and combined (hybrid). Tveito et al. (2006) added three more groups: methods specially developed for meteorology and climatology, which are basically probabilistic methods (based on RK), then ANN (an ML method), and physical methods that are used for downscaling. Sluiter (2009) (based on Dobesch et al. 2007 and Tveito et al. 2006) omitted combined interpolation methods from classification and merged them with probabilistic ones, which is also a usual approach in literature. The approach used here is based on whether stations (observations) or covariates only, or their combinations are used for spatial interpolation:

- Station-based interpolation methods

- Covariate-based interpolation methods

- Combined interpolation methods

Commonly used interpolation methods for spatial interpolation of climate variables for each of these three groups are described in this chapter. Many interpolation methods described in this chapter have their own version applicable to spatio-temporal interpolation problems. Here, the focus is more on spatial interpolation methods. Some of kriging and ML spatio-temporal interpolation methods are presented in Chapters 4 and 5. Physical, downscaling, and special meteorological methods (Tveito et al. 2006) were out of the scope of this research because their application is quite specific.

2.2 Station-based interpolation methods

Station-based interpolation methods use only station observations in the spatial interpolation process. They can be divided into three groups:

- Deterministic methods
- Splines
- Geostatistical methods

2.2.1 Deterministic Methods

The idea behind deterministic interpolation methods is to create a continuous surface from observations using a mathematical function. A brief explanation of each method is given in the following Sections. Prediction maps of all deterministic methods are shown in Figure 2.1. Further details can be found in books by Burrough and McDonnell (1989) and Webster and Oliver (2007).

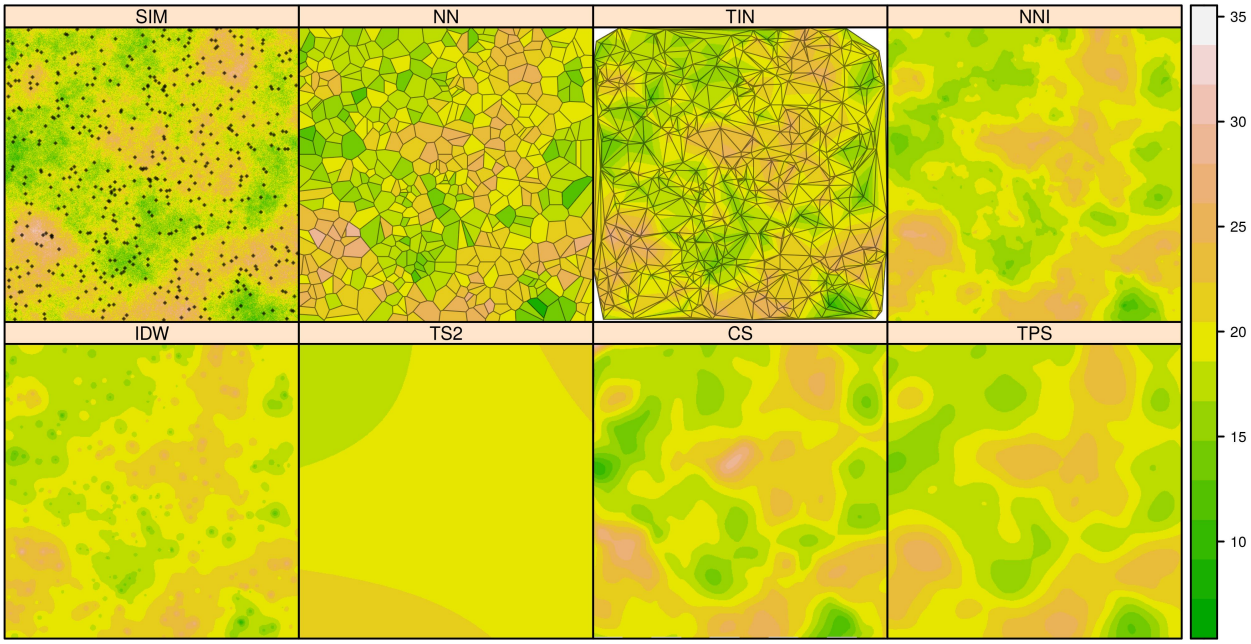


Figure 2.1: A map that shows the geostatistically simulated reality (SIM) and the locations of the 500 samples used to create prediction maps by all deterministic methods presented in this section (NN, TIN, NNI, IDW, TS2, CS, TPS).

2.2.1.1 Nearest Neighbours

Nearest neighbours interpolation simply assigns the value of the nearest measured point to a prediction location. The interpolated surface takes the form of Thiessen polygons (Thiessen 1911) or Voronoi diagrams (Figure 2.1, NN). NN is an exact and local interpolator. The disadvantage of NN is that it does not take the influence of the neighbouring observations into account.

NN is rarely used in the interpolation of climate elements. One can benefit from NN interpolation only in cases where there is a large number of observations. NN is still used in hydrology for

estimating areal precipitation (Tveito et al. 2006). Piper and Stewart (1996) created the first global daily temperature and precipitation dataset at 1 degree (∼100 km) spatial resolution for the year 1987, based on NN.

2.2.1.2 Triangulated Irregular Network

TIN interpolation creates a network of triangles so that vertices of triangles are observations (Figure 2.1, TIN). The triangles are created in a way that they are equilateral as possible and they do not contain any other observation. TIN uses the slope of the prediction location overlapping a triangle for prediction. There are different algorithms for TIN creation. Like NN, TIN is a simple, exact and local interpolator, but it uses more observations for interpolation, i.e. three observations, and creates a continuous interpolation surface.

As is the case for NN, it's application is limited, but still it can be used for the visual investigation of spatial patterns. IDW can be also used in cases where there is a large number of observations, i.e. if station density is high. Meteo Norway uses it for initial gridding of daily precipitation (Sluiter 2009).

2.2.1.3 Natural Neighbour Interpolation

Natural Neighbour Interpolation (NNI) (Sibson 1981) makes a prediction at a location as a weighted average of the nearest observation, so called "natural neighbours":

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^{n} w_i \cdot z(\mathbf{s}_i) \tag{2.1}$$

where $\hat{z}(\mathbf{s}_0)$ is the prediction at prediction location $\mathbf{s}_0$, w_i is a weight assigned to observation $z(\mathbf{s}_i)$ at location $\mathbf{s}_i$, and n is the number of the nearest observations, i.e. natural neighbours. NNI weights represents how much a Thiessen polygon at the prediction location $\mathbf{s}_0$ covers the surrounding Thiessen polygons at the observations locations $\mathbf{s}_i$:

$$w_i = \frac{P_i}{\sum_{i=1}^{n} P_i} \tag{2.2}$$

where P_i is the common area of a Thiessen polygon at the prediction location $\mathbf{s}_0$ and a Thiessen polygon at the observation location $\mathbf{s}_i$. By summarising all of the common areas, $\sum_{i=1}^{n} P_i$ actually represents the area of the Thiessen polygon at the prediction location $\mathbf{s}_0$. Also, the number of nearest observations, n, actually represents the number of Thiessen polygons around observations covered by a Thiessen polygon at the prediction location. NNI is also an exact and local interpolator, but different to NN and IDW, NNI is a smooth interpolator (Figure 2.1, NNI).

The National Center for Atmospheric Research (NCAR) uses NNI for the visualisation of interpolated climate elements (Hofstra et al. 2008).

2.2.1.4 Inverse Distance Weighting

IDW (Willmott et al. 1985), similar to NNI, makes a prediction at a location as a weighted average of the nearest observation (Eq. 2.1). IDW is named after the weights it uses, i.e., weights are inversely

related to distance:

$$w_i = \frac{d_i^{-p}}{\sum_{j=1}^{n} d_j^{-p}} \tag{2.3}$$

where d_i is the Euclidean distance between locations $\mathbf{s}_0$ and $\mathbf{s}_i$, and p is an exponent. The sum of the weights is equal to 1. These weights impose a greater influence to closer points relative to farther points; with a larger p exponent, the influence of nearer points becomes higher. IDW is an exact interpolator, but can be global or local. Here n represents the number of the nearest observations and, different from NNI, is set by the user. If n is set to be equal to the total number of observations, then we are talking about global IDW (Figure 2.1, IDW), and if n is less than the total number of observations, then we are talking about local IDW. NN is a special case of IDW in the limit when p approaches $+\infty$. The most commonly used IDW is the one where p equals 2. As for the NN, TIN, and NNI, there is no measure of error for IDW. Also, IDW weights do not take into account a configuration of the sampling, i.e. clustered observations have the same weights as an isolated observation.

IDW is widely used in the interpolation of climate elements, especially for precipitation (Dobesch et al. 2007). Jeong et al. (2020) used the average of IDW and PRISM predictions for daily precipitation estimation in order to solve the PRISM problem of precipitation overestimation. For example, Dodson and Marks (1997) used elevation in the form of hydrostatic and potential temperature equations in the IDW method to interpolate the minimum and maximum temperature at a 1 km resolution for the mountainous region of the US Pacific Northwest.

2.2.1.5 Trend Surface Mapping

Trend surfaces (Chorley and Haggett 1965) are linear regression models (see Section 2.3.1.1) in which geographic coordinates are used as covariates. For example, a second-order trend surface (TS2) uses a quadratic function of the x- and y-coordinates:

$$\hat{z}(s_0) = a \cdot s_{0,x}^2 + b \cdot s_{0,y}^2 + c \cdot s_{0,x} s_{0,y} + d \cdot s_{0,x} + e \cdot s_{0,y} + f \tag{2.4}$$

where a, b, c, d, e, and f are regression coefficients and $s_{0,x}$ and $s_{0,y}$ are the coordinates of the prediction location $\mathbf{s}_0$. The regression coefficients are usually estimated using ordinary least squares (see Section 2.3.1.1). TS are not popular because a higher-order trend surface is needed for complex variables (Figure 2.1, TS2). They are used mostly for discovering a long-range trend of the variable or for the interpolation of monthly and annual variables (Tveito et al. 2006).

2.2.1.6 Splines and local trend surfaces

Spline fits the surface through observations by series of m-order polynomial functions (called splines) and it is a global interpolation method (Mitas and Mitasova 1999). Based on m, splines can be:

- linear (m = 1)
- quadratic (m = 2)
- cubic (m = 3)

The polynomial function and m-1 derivatives are continuous at each observation for all splines. Therefore, for linear splines first derivative, for quadratic splines the second derivative, and for cubic splines (Figure 2.1, CS) the third derivative is continuous at each observation.

The local version of splines are local trend surfaces where the m-order spline is fitted for each observation location based on a limited number of neighbouring observations (Venables and Ripley 2002).

Splines are suitable for interpolation of monthly and annual climate elements, but are less suitable for high temporal resolution variables, such as daily and hourly climate elements (Sluiter 2009).

2.2.1.7 Thin plate splines

Thin plate spline (Wahba and Wendelberger 1980) is a spline-based interpolation method. Its name comes from bending a thin sheet of metal. As metal is rigid, the TPS resists bending by implying a roughness penalty that balances between surface smoothness and passing through observations. In order to control smoothness (and rigidity), TPS uses the λ parameter. Based on chosen λ parameter, TPS surface $f(\mathbf{s})$ is fitted so as to minimize the following energy function:

$$\min_f \sum_{i=1}^{n} (z(\mathbf{s}_i) - f(\mathbf{s}_i))^2 + \lambda \int f''(\mathbf{s})^2 d\mathbf{s} \tag{2.5}$$

where $z(\mathbf{s}_i) - f(\mathbf{s}_i)$ minimize the difference between observation and TPS function at location s_0 and $\int f''(\mathbf{s})^2 d\mathbf{s}$ is a roughness penalty function that penalizes the overall variability of TPS, i.e. maximizes the smoothness of TPS. Second derivative $f''(\mathbf{s})$ is used as a penalty because it is the measure of slope change, i.e. surface roughness. If λ equals 0, there is no roughness penalty and TPS surface will pass through observations (Figure 2.1, TPS). As λ parameter increases, smoothness of the TPS surface also increases. In practice, the optimal λ parameter can be estimated from the observation using a generalized cross-validation (GCV) (Wahba and Wendelberger 1980).

TPS was at first developed for spatial interpolation of climate elements (Wahba and Wendelberger 1980) and it is widely used for the interpolation of temperature (Jarvis and Stuart 2001; Stewart and Nitschke 2017) and precipitation (Hutchinson 1995; Tait et al. 2006; Hutchinson et al. 2009; Yuan et al. 2015). Haylock et al. (2008) used TPS for the interpolation of monthly temperature and precipitation, and then daily anomalies were interpolated by kriging.

2.2.2 Geostatistical methods

Earlier, the meaning of the word geostatistics was literally "statistics of the earth" or "statistics of the geo-sciences" (geology, geography, etc.). Nowadays, it has a different meaning — a statistics applied to spatial/spatio-temporal data. The most popular geostatistical interpolation method is kriging, and therefore kriging is often a synonym to geostatistics. Kriging was named after the south-African mining engineer Danie Krige who established the foundations of kriging (Krige 1951). Kriging is a stochastic interpolation method — it incorporates randomness in spatial/spatio-temporal interpolation. In the early '60s, Matheron (1963) introduced the mathematical basics of kriging with an application in geology. Kriging is based on an idea that closer observations are more similar and correlated than observations that are at a greater distance.

Unlike deterministic interpolation methods, kriging starts from the assertion that the observed reality is a realisation of a random field (Webster 2000). It uses the observations to estimate the parameters of this field, after which predictions are made. In other words, it assumes a geostatistical model and derives the optimal interpolation from it.

Many versions of station-based kriging exist (Li and Heap 2008; Webster and Oliver 2007). Some of them are:

- ordinary kriging — a basic form of kriging, for a variable with an unknown mean (Section 2.2.2.1)
- simple kriging — a kriging version for a variable with a known mean (Section 2.2.2.2)
- lognormal kriging — ordinary kriging of a lognormal transformation of the skewed and non-normal target variable (Webster and Oliver 2007)
- indicator kriging — a non-linear version of kriging used for binary variables (Section 2.2.2.3)
- disjunctive kriging — a non-linear version of kriging for finding a probability of a target variable exceeding a predefined threshold (Webster and Oliver 2007; Li and Heap 2008)
- probability kriging — combines co-kriging and indicator kriging, i.e. it is a co-kriging of target variable and its normalized rank order in form of a co-variable (Sullivan 1984)
- ordinary co-kriging — ordinary kriging for two or more spatially correlated variables (Section 2.2.2.4)
- Bayesian kriging — simple kriging that incorporates a prior knowledge about the trend (Omre 1987)
- block kriging — ordinary kriging used for prediction in a block

The most used kriging versions in the spatial interpolation of climate elements are ordinary, indicator, and co-kriging and they are explained in the following sections. Kriging can also be local or global, depending on whether a single global semivariogram or many local semivariograms for each of the observations is fitted (Hofstra et al. 2008). Also, extensions of kriging have been developed by introducing environmental covariates in the kriging process. Universal kriging, regression kriging, and kriging with external drift to name some of them. Universal kriging, as a station-based interpolation method (it uses geographical coordinates as covariates) is explained in Section 2.2.2.5, while regression kriging and kriging with external drift are explained in Section 2.4 where combined interpolation methods are presented.

Kriging and its versions are well described in many books, such as Isaaks and Srivastava (1989), Goovaerts (1997), Webster and Oliver (2007), and Chilès and Delfiner (2012).

2.2.2.1 Ordinary kriging

Ordinary kriging is the basic form of kriging, which, similarly to NNI and IDW, predicts $\hat{z}(s_0)$ as a linear combination of the observations where the sum of the weights is also equal to 1 (Eq. 2.1). However, unlike NNI and IDW, the weights (w_i) are derived from the degree of spatial correlation, as quantified by a semivariogram. The semivariogram is defined as:

$$\gamma_S(h) = \frac{1}{2} E(Z(\mathbf{s}) - Z(\mathbf{s}+h))^2 \tag{2.6}$$

where $\gamma_S(h)$ denotes the semivariance of observations at h units of a distance in space, E is a mathematical expectation, and $Z(\mathbf{s}) - Z(\mathbf{s}+h)$ is a difference between all observation pairs at spatial distance h. From the Eq. 2.6, the semivariogram actually describes a variance between pairs of observations at spatial distance h.

In general, a variable needs to meet two assumptions for OK: stationarity and isotropy. Webster and Oliver (2007) (in Section 4.3.1) said that: "*By stationarity we mean that the distribution of the*

random process has certain attributes that are the same everywhere". For OK, the stationarity assumption actually means that the mean and variance of the variable do not change over spatial and/or temporal domain, i.e. the same probability distribution function should be expected at every spatial and/or temporal location. Isotropy refers to a directionally independent variable (Sluiter 2009), i.e. that variable has uniform values in all directions. For OK, it means that semivariance ($\gamma_S(h)$) of the variable depends only on the magnitude of spatial distance between observations (h), and does not depend on the observation location ($\mathbf{s}$) and direction of h.

The sample (or experimental) semivariogram is estimated from the observations:

$$\hat{\gamma_S}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n} (z(\mathbf{s}_i) - z(\mathbf{s}_i - h))^2 \tag{2.7}$$

where $n(h)$ is the number of observation pairs at spatial distance h. A sample semivariogram is made for lag classes of h. In other words, semivariance $\gamma_S(h)$ is calculated for groups of observations at different spatial distance h ranges (e.g. for h ranges 0–10 km, 10–20 km, 20–30 km, etc.; blue points in Figure 2.2). From the sample semivariogram, various properties of the data, such as nugget, sill, and range, can be determined (Figure 2.2). The nugget represents semivariogram interception with $\gamma_S(h$ vertical axis. It shows semivariance for very short distances (i.e. distances that are smaller than the smallest distance between observations) or semivariance at an observation location caused by short-range spatial variability or a measurement error of the used meteorological instrument. The semivariogram range represents the distance h beyond which the observations become spatially independent, i.e. there is no spatial correlation between them. The semivariogram still represents the semivariogram value for the semivariogram range. For semivariogram modelling the partial sill, which represents the difference between sill and nugget, is used.

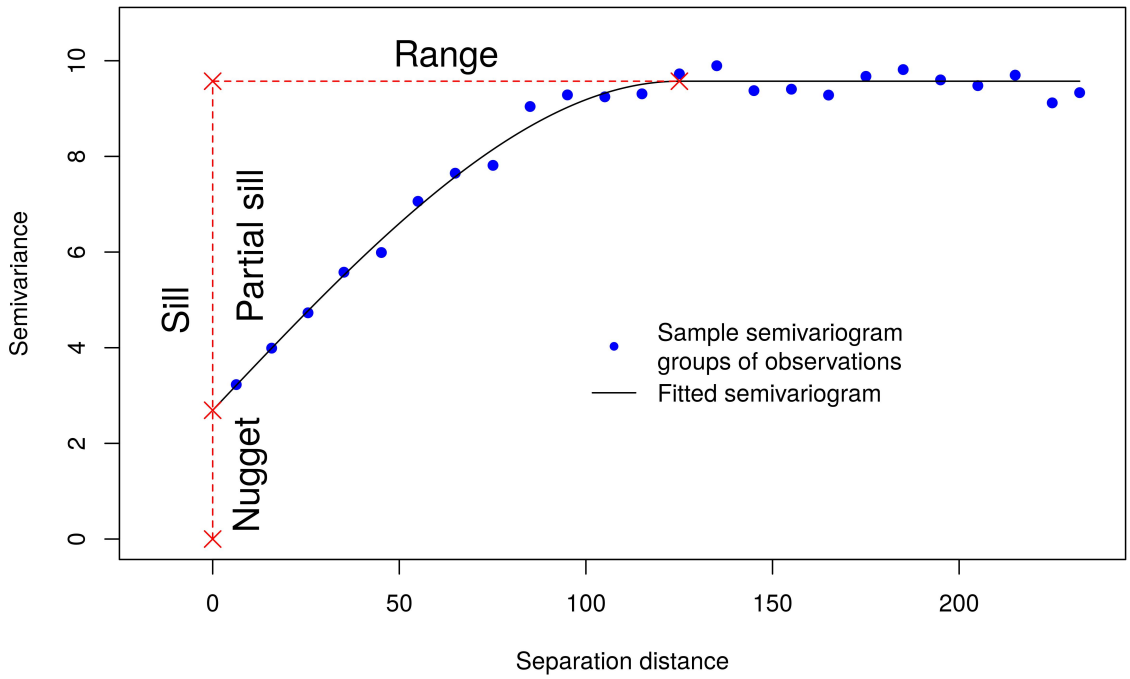


Figure 2.2: Sample and fitted semivariogram.

The final semivariogram is fitted through a sample semivariogram using a mathematical function, called semivariogram model (black line in Figure 2.2). The most common used mathematical function models are linear, spherical, exponential, and Gaussian (Figure 2.3). Besides these, other functions that can be used can be found in R package `gstat` (Pebesma 2004; Gräler et al. 2016), using `show.vgms` function. The semivariogram nugget, sill (partial sill) and range are used for

fitting the semivariogram and choosing the appropriate mathematical function.

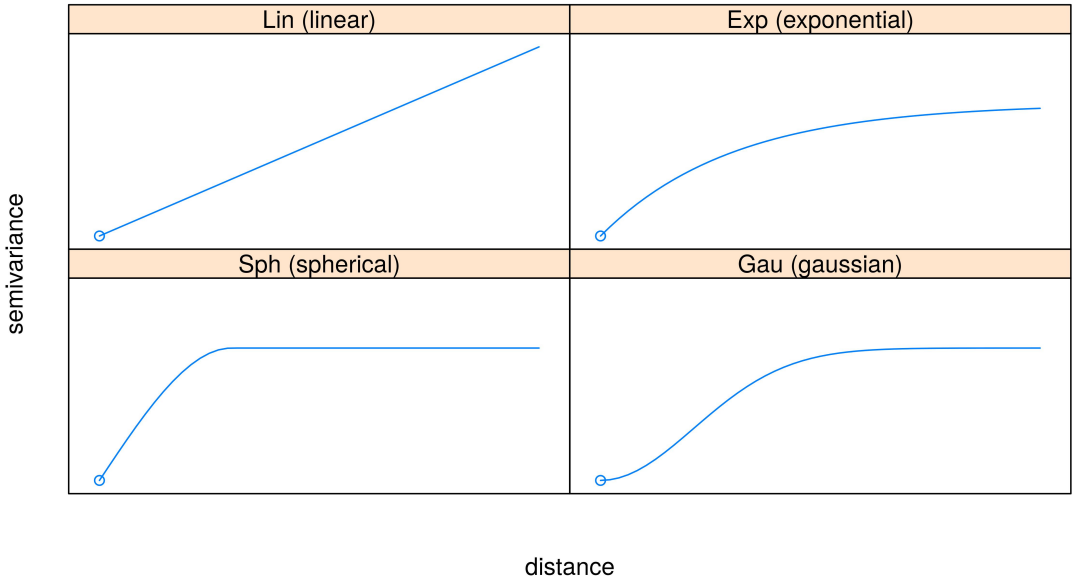


Figure 2.3: The shape of linear, spherical, exponential, and Gaussian semivariogram models.

The fitted semivariogram is then used for the calculation of kriging weights for each prediction location. The basic theory concerning the derivation of kriging weights is well explained in text books Goovaerts (1997), Webster and Oliver (2007), and Chilès and Delfiner (2012). Similar to IDW, OK weights tend to be relatively large when the separation distance between the observation and the prediction point is small, but they are also influenced by the spatial configuration of the observation points, and by the degree of short-distance spatial variation. To sum up, the OK weights are chosen such that the expected squared prediction error is minimized, under the condition of unbiasedness. The expected squared prediction error is known as the kriging variance and is also standardly computed in OK:

$$var[\hat{z}(s_0)] = 2\sum_{i=1}^{n} w_i \gamma_S(\mathbf{s}_i, \mathbf{s}_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma_S(\mathbf{s}_i, \mathbf{s}_j) \tag{2.8}$$

where $var[\hat{z}(s_0)]$ is the kriging variance at spatial location $\mathbf{s}_0$, $\gamma_S(\mathbf{s}_i, \mathbf{s}_0)$ is a semivariance between observations at spatial locations $\mathbf{s}_i$ and $\mathbf{s}_0$, and $\gamma_S(\mathbf{s}_i, \mathbf{s}_j)$ is the semivariance between observations at spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$. From the Eq. 2.8 it can be seen that kriging variance is independent of observations.

OK is widely used in climatology and meteorology, especially for the interpolation of daily climate elements. Courault and Monestiez (1999) used OK for the interpolation of daily maximum and minimum air temperatures in the southeast of France. Hunter and Meentemeyer (2005) used OK for the interpolation of daily precipitation and maximum and minimum temperatures for California, on grids with the spatial resolution of 2 km, for the 1980–2003 period. Stahl et al. (2006) compared 12 regression-based and weighted-based interpolation methods for the interpolation of daily maximum and minimum temperatures over British Columbia, Canada, and found that OK performs best. Hofstra et al. (2008) compared global and local kriging, two versions of angular distance weighting, natural neighbor interpolation, regression, 2D and 3D thin plate splines, and conditional interpolation in the case of daily precipitation, mean, minimum and maximum temperature, and sea

level pressure over Europe, for the 1961–1990 period. They concluded that, overall, global kriging performs the best.

The usage of OK is limited since the most of modelled variables are not stationary and isotropic in nature. In order to overcome this problem, other versions of kriging, described in the following sections, are developed.

2.2.2.2 Simple kriging

Simple kriging (SK) is a version of OK with a known trend (mean) of the target variable. Therefore, SK makes fewer assumptions than OK and thus improves kriging prediction performance. The SK prediction is still a linear combination of the observations (Eq. 2.1), but with an addition of the trend (μ):

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^{n} w_i \cdot z(\mathbf{s}_i) + (1 - \sum_{i=1}^{n} w_i) \cdot \mu \tag{2.9}$$

Another difference with OK is that the sum of the weights ($\sum_{i=1}^{n} w_i$) is not constrained to equals 1. Because of this, covariances are used for the calculation of weights instead of semivariances (Webster and Oliver 2007).

In reality, as for the climate elements, the trend of the target variable is often unknown, so OK is more used for spatial interpolation of climate elements. SK can be used in the form of residual kriging (Section 2.4.2), as Brinckmann et al. (2016) and Berezowski et al. (2016).

2.2.2.3 Indicator kriging

Indicator kriging (IK) is a version of kriging for modelling of binary variables (possible values are 0 and 1). A binary variable can be created from a continuous variable by applying some threshold. The values above the threshold get the value 1, and the values under the threshold get the value 0. The prediction can be made in the same way as OK (Section 2.2.2.1), or as SK (Eq. 2.9) if a sample mean value is taken for the trend (μ). The predictions are values between 0 and 1, actually showing the probability that the interpolated value is 1.

When transforming a continuous variable to a binary variable we might lose some information of the original data by choosing a different threshold and so splitting the data into two classes. In this case, disjunctive kriging is used because it provides a more sophisticated way of transforming continuous data to binary (Rivoirard 1994).

For climate elements, the most representative example is precipitation occurrence (Berezowski et al. 2016), where the threshold of 0 or 0.5 mm is applied to the precipitation amount (Hofstra et al. 2008). The results show a probability of precipitation occurrence.

2.2.2.4 Co-kriging

Co-kriging is a non-linear version of kriging that uses additional observational variables called co-variables. They are usually highly correlated with the target variable and with more samples in order to improve the prediction of the target variable. In order to consider a correlation between target and co-variables, a cross (or multivariate) semivariogram is introduced instead of a semivariogram. A cross semivariogram with one co-variable is defined as:

$$\gamma_{A,B}(h) = \frac{1}{2} E((Z_A(\mathbf{s}) - Z_A(\mathbf{s}+h)) \cdot (Z_B(\mathbf{s}) - Z_B(\mathbf{s}+h))) \tag{2.10}$$

where, analog to semivariogram (Eq. 2.6), $\gamma_{A,B}(h)$ denotes the cross-variogram value at h units of a distance in space, $Z_A(\mathbf{s}) - Z_A(\mathbf{s}+h)$ is the difference between all observation pairs at spatial distance h for the target variable, and $Z_B(\mathbf{s}) - Z_B(\mathbf{s}+h)$ is the same but for the co-variable. The CK prediction is calculated as a weighted sum of all variables, i.e. target variable and all co-variables:

$$\hat{z}_1(\mathbf{s}_0) = \sum_{l=1}^{v} \sum_{i=1}^{n_l} w_{li} \cdot z_l(\mathbf{s}_i) \tag{2.11}$$

where $\hat{z}_1(\mathbf{s}_0)$ is a prediction of the target variable at spatial location $\mathbf{s}_0$, v is the number of co-variables (including the target variable), n_l is the number of observations of the l-th co-variable, w_{li} is the weight for i-th observation of the l-th co-variable, and $z_l(\mathbf{s}_i)$ is the observation of the l-th co-variable at spatial location $\mathbf{s}_i$.

CK gives better results when there are more co-variables that are highly correlated with the target variable, but the CK model becomes complex and computationally time consuming. Schuurmans et al. (2007) used OK, CK, KED to interpolate daily precipitation, using precipitation observations as the main variable and precipitation radar data as the co-variable, over the Netherlands for the period March–October 2004. They concluded that CK and KED give more accurate predictions in the case of large extents.

2.2.2.5 Universal kriging

Universal kriging was first presented by Matheron (1963). UK is a similar method to regression kriging (explained in Section 2.4.1) and kriging with external drift (explained in Section 2.4.3), where the modelling of trend with linear regression is included in the kriging process. Computationally and in its original form, UK is a special case of kriging with external drift, where the trend is modeled as a function of coordinates (Hengl et al. 2012). Therefore, UK is purely a station-based interpolation method. Many authors use the term UK for kriging with external drift and regression kriging.

2.2.2.6 Geostatistical simulations

Even though kriging gives the best linear unbiased estimates of a specific variable, its predictions give smoothed surfaces (with the same values at observations locations) and a lot of spatial variation is lost, which is not what we would expect in reality. In other words, kriging variance does not represent the variance of the variable itself and, moreover, it is much smaller (Webster and Oliver 2007). To generate an interpolated surface that will have a spatial variation that we expect to happen in reality, i.e. to retain the variance of the variable (observations), geostatistical simulation methods are used. Webster and Oliver (2007) gave a definition of simulation: "*In geostatistics the term simulation is used to mean the creation of values of one or more variables that emulate the general characteristics of those we observe in the real world.*". Spatial variation of a specific variable can be characterized by its mean and semivariogram (or covariance) function. Geostatistical simulations use the mean and semivariogram (or covariance) function in order to create probable realizations of the specific variable with the same statistical characteristics. On one side, kriging gives the most accurate prediction (with minimum variance), and on the other side using simulation we keep the statistical characteristics and spatial variation of the variable.

Simulation can be unconditional and conditional. Unconditional simulation is the one where there are no other conditions other than the specified mean and semivariogram (or covariance) function. Conditional simulation adds a condition to keep the original values at the observation locations. There are several approaches for both, unconditional and conditional simulations. Some

of them are explained by Webster and Oliver (2007), Chilès and Delfiner (2012), and Bivand et al. (2013a). The simplest way to perform the unconditional simulation is by using the Monte Carlo simulation for the creation of random values over a grid, and then averaging the values inside a circle around each grid cell to create a spatially correlated surface (Webster and Oliver 2007). In Section 5.2.2.1, a sequential (unconditional) simulation algorithm is explained. Cornes et al. (2018) used the geostatistical simulation for spatial interpolation of climate elements, which is explained in Section 3.3.1.1.

2.3 Covariate-based interpolation methods

Covariate-based interpolation methods use only environmental covariates for spatial interpolation. Station observations are used here only for making a model, as a target variable, but not actually in spatial prediction. Two groups of covariate-based methods are methods based on linear regression and machine learning methods.

2.3.1 Linear regression methods

Linear regression methods model the target variable as a linear combination of one or more covariates. The three most used linear regression methods in the interpolation of climate elements are multiple linear regression, geographically weighted regression, and generalized additive models.

2.3.1.1 Multiple linear regression

Multiple linear regression (MLR) is a global interpolation method where the target variable is modelled as a weighted linear combination of covariates. MLR prediction at spatial location $\mathbf{s}_0$ is given by:

$$\hat{z}(\mathbf{s}_0) = \sum_{k=0}^{n_c} \beta_k x_k(\mathbf{s}_0) \tag{2.12}$$

where n_c is the number of covariates, β_k are regression coefficients estimated using ordinary least squares (OLS), β_0 is model intercept (by imposing f_0 is equal to 1), and $x_k(\mathbf{s}_0)$ are covariates values at spatial location $\mathbf{s}_0$. β_k the estimation of regression coefficients by OLS in matrix notation is represented as:

$$\hat{\beta}_{OLS} = (\mathbf{q}^T \cdot \mathbf{q})^{-1} \cdot \mathbf{q}^T \cdot \mathbf{z} \tag{2.13}$$

where $\hat{\beta}_{OLS}$ is the matrix of estimated β_k regression coefficients with dimensions $n \times 1$, $\mathbf{q}$ is the matrix of covariate values at observations locations with dimensions $n \times (n_c + 1)$, and $\mathbf{z}$ is the matrix of observations with dimensions $n \times 1$. MLR can be extended to the spatio-temporal domain by including spatio-temporal covariates, which is described in Section 4.3.1.

MLR is mostly used for temperature modelling or for aggregated climate elements. Kurtzman and Kadmon (1999) compared MLR models with splines and IDW for the interpolation of mean daily temperature values in Israel, and MLR outperformed both methods. Daly et al. (1994) used PRISM, a MLR-based model, to interpolate monthly and annual precipitation in the western US. PRISM was also applied by Schwarb et al. (2001) to create long-term precipitation grids for the European Alpine region. MLR is often used for trend modelling along with the modelling of residuals with a station-based interpolation method, such as kriging or IDW (see Sections 2.4.1, 2.4.1 and 2.4.3).

2.3.1.2 Geographically weighted regression

The main disadvantages of MLR for spatial interpolation are: (1) it is a global interpolator and (2) the spatial correlation between observations, i.e. the observations locations are not considered at all in the modelling and prediction process. Brunsdon et al. (1996) developed Geographically Weighted Regression (GWR) in order to overcome these problems.

To overcome the first MLR problem, GWR fits as many local MLR models as there are prediction locations. Starting from Eq. 2.12, GWR prediction at spatial location $\mathbf{s}_0$ is given by:

$$\hat{z}(\mathbf{s}_0) = \sum_{k=0}^{n_c} \beta_k(\mathbf{s}_0) x_k(\mathbf{s}_0) \tag{2.14}$$

where, different from Eq. 2.12, $\beta_k(\mathbf{s}_0)$ are local regression coefficients at spatial location $\mathbf{s}_0$. To overcome the second MLR problem, instead of OLS estimation, local regression coefficients $\beta_k(\mathbf{s}_0)$ are estimated by weighted least square (WLS):

$$\hat{\beta}_{WLS}(\mathbf{s}_0) = (\mathbf{q}^T \cdot \mathbf{W}(\mathbf{s}_0) \cdot \mathbf{q})^{-1} \cdot \mathbf{q}^T \cdot \mathbf{W}(\mathbf{s}_0) \cdot \mathbf{z} \tag{2.15}$$

where, different from Eq. 2.13, matrices $\mathbf{z}$ and $\mathbf{q}$ are created based on observations in a specified bandwidth from the spatial location $\mathbf{s}_0$ and corresponding covariate values, and $\mathbf{W}(\mathbf{s}_0)$ is the diagonal matrix of geographical weights between spatial location $\mathbf{s}_0$ and the observations in bandwidth. In practice, an optimal bandwidth is obtained by leave-one-out cross-validation. In its original form (Fotheringham et al. 1998), geographical weights in the $\mathbf{W}(\mathbf{s}_0)$ matrix are based on the Euclidean distance between spatial location $\mathbf{s}_0$ and the i-th observation in the bandwidth b:

$$w_i = exp\left(-\frac{d_i^2}{b^2}\right) \tag{2.16}$$

In order to make a prediction with GWR, firstly, local regression coefficients at prediction location are estimated using observations in the specified bandwidth. Then, the estimated local regression coefficients and covariates at a prediction location are used to make a prediction.

Wang et al. (2017) compared kriging, spline, IDW, MLR, and GWR for the interpolation of the monthly minimum, mean, and maximum temperature in China. Even though kriging performed better than GWR in warmer months, overall GWR outperformed all of the methods. Li et al. (2018b) created a gridded dataset of maximum and minimum daily temperature at a 1 km spatial resolution over the conterminous US for the 2003–2016 period using GWR.

2.3.1.3 Generalized additive models

Generalized additive model (GAM) (Hastie and Tibshirani 1986) introduces non-linearity in MLR by replacing covariates x_k and its corresponding regression coefficient β_k (Eq. 2.12) with a set of non-linear (smoothing) functions of one or more covariates. GAM can be seen as an MLR model (a linear regression model) of non-linear functions of covariates. This makes GAMs a more flexible model than MLR, with the ability to characterize non-linear relationships. GAM prediction is given by:

$$\hat{z}(\mathbf{s}_0) = \beta_0 + \sum_{k=0}^{n_c} f_k(\mathbf{s}_0) \tag{2.17}$$

where n_c is the number of non-linear functions f_k here (corresponds to the number of covariates in Eq. 2.12), $f_k(\mathbf{s}_0)$ are estimations of the non-linear function at spatial location $\mathbf{s}_0$. The word "additive"

comes from the contribution of each of the non-linear functions f_k to the final prediction. Smoothing splines are mostly used as non-linear functions. In order to fit the final model, GAM uses a backfitting algorithm where the fit of each non-linear function is repeatedly improved (James et al. 2013). After every iteration, the backfitting algorithm checks partial residuals r_i, one by one, and minimizes them:

$$r_i = z(\mathbf{s}_0) - \beta_0 + \sum_{k=0}^{n_c} f_k(\mathbf{s}_0) \tag{2.18}$$

GAM, where second-order polynomial functions and TPS were used as additive models, shown to be a good solution for modelling of monthly mean temperature for Taiwan's mountain regions in research by Guan et al. (2009). Aalto et al. (2013) came to the same conclusion for Finland, where GAM outperformed KED and GAM combined with residual kriging for monthly mean temperature. GAM models performed best in the case of Oregon, USA, where grids for maximum air temperature were created at a 1 km spatial resolution (Parmentier et al. 2014).

2.3.2 Machine learning methods

The application of ML in the spatial interpolation of climate elements is quite novel. Within the last decade or two, ML has become a popular tool for spatial interpolation of environmental variables, as well as in climatology and meteorology. Their advantage over previously explained methods is that they are mostly non-linear interpolators and can model complex relations between environmental covariates and a target variable.

The most popular ML algorithms used for spatial interpolation of climate elements are described in this section. Basic principles of each of the algorithms are given. More details on algorithm procedures can be found in Hastie et al. (2009), Kuhn and Johnson (2013), James et al. (2013), and Kanevski et al. (2009).

2.3.2.1 Random Forest, Gradient Boosting Machine and Cubist

All of the three methods are decision tree-based ML methods, but with differences in model fitting and prediction process.

RF is an ensemble ML algorithm based on decision trees and bagging (Breiman 1996, 2001). Decision Trees and Classification And Regression Trees (CART) (Breiman et al. 1984) are algorithms in which a prediction is made by a series of splitting rules. The splitting rules are represented by nodes, splitting rule decisions by branches, and final predictions by leaves. Building a CART is performed by splitting the data into two branches at each new node creation, until a stop criterion is satisfied. For each node, a feature (a synonym for covariate, but preferred nomenclature in ML) and a threshold for splitting are obtained by choosing these such that the variance of the data within the partitions obtained by the split is minimized. A prediction is made by moving through the nodes and branches and finally ending in one of the leaves. The benefits of CART compared with RF (explained below) are the low bias, simplicity, and ease of interpretation (James et al. 2013). However, they tend to overfit the training data and can be non-robust, which is manifested in a lower prediction accuracy.

In order to overcome the disadvantages of CART, bagging (bootstrap aggregation) was proposed by Breiman (1996). Bagging is an ensemble ML method that uses many weak learners, such as CART, and combines these into one stronger learner. Bootstrapping (sampling with replacement) is repeatedly used to sample the whole dataset and thus create a large number of weak learners. The prediction is represented by the average of the predictions from all weak learners. Thereby,

bagging reduces prediction error variance which makes the model more stable and more accurate.

RF (Breiman 2001) uses bagging and random feature selection in combination with CART as a weak learner. The problem with bagging is that bootstrapped samples may still be correlated if there are strong (dominant) features. This problem is mitigated by including random feature selection (Amit and Geman 1997) at each step during the creation of each CART. The number of features and the number of CARTs can be fine-tuned (the recommended number of features is $\sqrt{m}$ for classification and $\frac{m}{3}$ for regression, where m is the number of covariates). The overview of the RF algorithm is given in Figure 2.4.

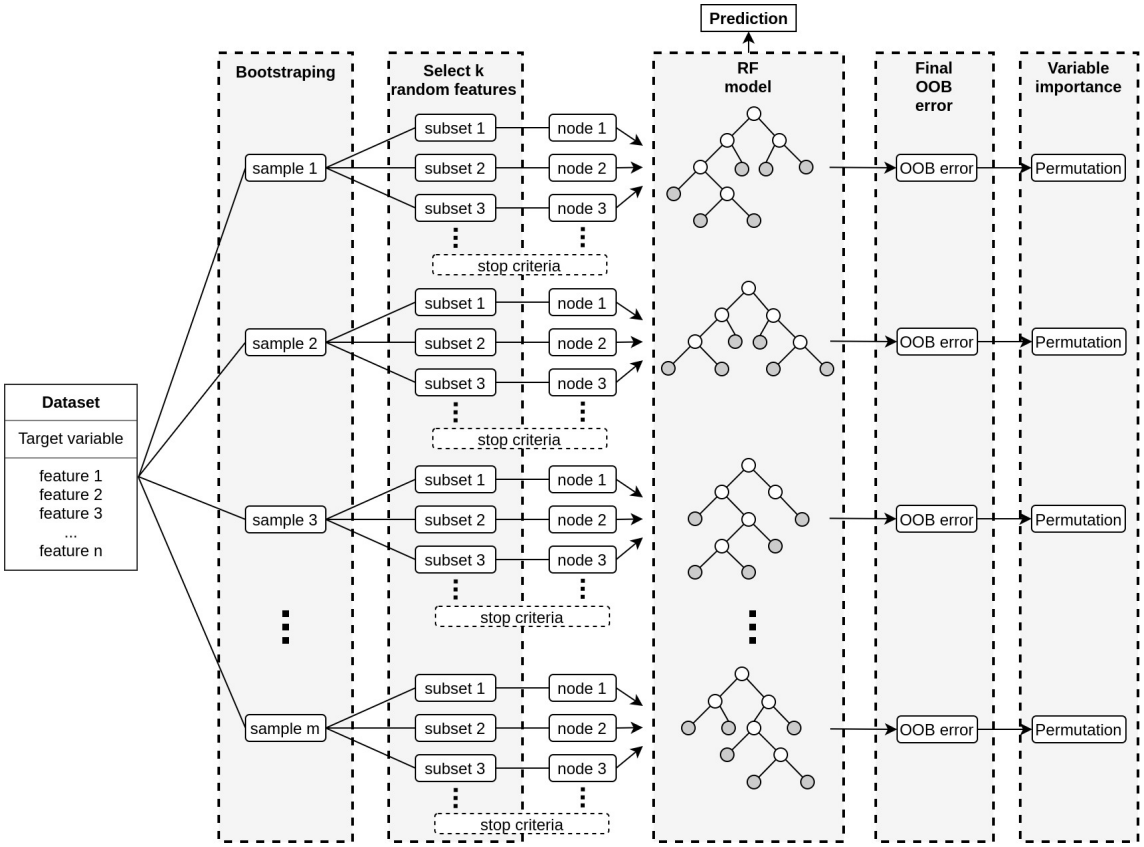


Figure 2.4: RF algorithm scheme.

In case of spatial interpolation, RF prediction at spatial location $\mathbf{s}_0$ represents an average of predictions from all decision trees:

$$\hat{z}(\mathbf{s}_0) = \frac{1}{B}\sum_{b=1}^{B}\hat{z}_b(\mathbf{s}_0) \tag{2.19}$$

where B is the number of decision trees and $\hat{z}_b(\mathbf{s}_0)$ is a prediction from b-th decision tree at spatial location $\mathbf{s}_0$. In case of classification, RF prediction is a mode from the prediction from all decision trees.

RF has an option for calculation of an out-of-bag (OOB) error (James et al. 2013). Observations that were not used for making a decision tree are called OOB observations. From the same decision tree, we can make predictions for the OOB observations. Next, these predictions, but from all decision trees, are averaged per OOB observation and are used together with corresponding OOB observations for the calculation of an OOB error. In case of spatial interpolation, OOB error does not show the spatial accuracy of the RF model, so spatial cross-validation is needed for spatial interpolation accuracy assessment. RF can measure variable importance by how much the total residual

sum of squares is decreased if a variable is chosen for a split in a decision tree, averaged over all of the decision trees (James et al. 2013).

As RF, Gradient boosting machines (GBM) (Friedman 2001) also use decision trees, but the decision trees are made in sequential order. Each new decision tree is made on the residuals from the previously fitted GBM model, i.e. from all previously fitted decision trees and, by doing so, slowly improves the accuracy of the GBM model. GBM prediction at spatial location $\mathbf{s}_0$ is given by:

$$\hat{z}(\mathbf{s}_0) = \sum_{b=1}^{B} \lambda \hat{z}_b(\mathbf{s}_0) \tag{2.20}$$

where B is the number of GBM decision trees, $\hat{z}_b(\mathbf{s}_0)$ is a prediction from b-th decision tree at spatial location $\mathbf{s}_0$, and λ is the shrinkage parameter which mostly ranges between 0.01 or 0.001 and sets the GBM learning speed. With large λ, the GBM model will learn faster and a small number of decision trees will be needed.

Cubist[1] is also a decision tree ensemble model, but has a different approach in comparison with RF and GBM in terms of decision tree splitting criterion and prediction. Different to RF and GBM, the cubist uses a reduction in the node's error rate criteria (RER) which represents the difference of standard deviation of the whole dataset before splitting (σ_z) and the weighted (by dataset size) average of standard deviations of the datasets after splitting, so called partitions (σ_{z_p}):

$$RER = \sigma_z - \sum_{p=1}^{P} \frac{n_p}{n} \cdot \sigma_{z_p} \tag{2.21}$$

where P is the number of partitions, n is the number of all samples, n_p is the number of samples of the p-th partition, and $\frac{n_p}{n}$ represent the weights.

The covariate with the largest reduction is chosen for splitting. Another novelty is that linear models, created based on split covariates from all parent nodes and a current node, are assigned to each node. The tree is growing until there is no reduction of error rate or not enough data. After the tree has grown, it is simplified by removing the nodes that are not decreasing an adjusted error rate (AER) previously computed for each node. AER is calculated based on the difference between observations and predictions in the node:

$$AER = \frac{n^* + n_c}{n^* - n_c} \sum_{i=1}^{n^*} |z_i - \hat{z}_i| \tag{2.22}$$

where n^* is the number of samples used for building the model, n_c is the number of covariates in the model. The $\frac{n^*+n_c}{n^*-n_c}$ term penalizes models with a large number of covariates.

Cubist also involves smoothing in the prediction process in order to avoid overfitting. A prediction from one decision tree is a weighted linear combination of predictions from all nodes linear models in the path from leaf to the initial node. Cubist, similar to GBM, uses a sequential series of decision trees, called committees, to make a prediction. The final prediction is made by averaging the predictions from committees (decision trees). Cubist can be represented by a rule-based model, a model consisting of many rules, and to each rule a multivariate linear model is assigned. The whole process of prediction is explained in detail by Kuhn and Johnson (2013).

From the presented tree-based methods, RF is the most used in climate elements interpolation. Just some of the most recent researches are mentioned here. Pang et al. (2017) used RF for down-

[1] https://www.rulequest.com/cubist-info.html

scaling of the daily mean temperature in the Pearl River basin in southern China. They show that, in that case, RF outperformed MLR, an artificial neural network, and a support vector machine. Mohsenzadeh Karimi et al. (2018) modelled long-term monthly air temperatures and they choose RF over a support vector machine and geostatistical methods. da Silva Júnior et al. (2019) showed that the RF model evapotranspiration in the northeast region of Brazil is better than IDW and OK. The main conclusion of Ruiz-Álvarez et al. (2019) study was that Random Forest produces the best results in comparison with Support Vector Machines, MLR and OK.

GBM performed the best in the interpolation of near-surface air temperature in Antarctica, with Moderate Resolution Imaging Spectroradiometer (MODIS) Land Surface Temperature (LST) as covariate, in comparison with RF and Cubist (Meyer et al. 2016). Fan et al. (2018) highly recommend GBM models for the interpolation of daily global solar radiation. dos Santos (2020) compared 54 regression models, mostly ML models, for the creation of a 1 km maximum daily temperature at a 1 km spatial resolution over London in summer for the 2006–2017 period and GBM outperformed all the methods.

Thanks to the R package `Cubist` (Kuhn and Quinlan 2020), Cubist recently became open-source. Since then, it was used in many studies, mostly for the interpolation of temperature, and gave good results. Emamifar et al. (2013) recommend Cubist for the interpolation of daily mean air temperature in the Khuzestan province (in the southwest of Iran), with MODIS LST as the covariate. Noi et al. (2017) came to the same conclusion for the case study in northwest Vietnam, especially for the mountainous areas. Méndez and Calvo-Valverde (2020) considered Cubist as the best approach for the creation of monthly air temperature grids over Costa Rica in comparison with RF, GAM, and geostatistical methods (OK and KED). This is because Cubist does not make assumptions on data normality and homoscedasticity. Appelhans et al. (2015) showed that regression trees (RF, GBM, and Cubist) perform better than any other ML or kriging interpolation method for monthly air temperature at Mt. Kilimanjaro, Tanzania. They propose Cubist with residual kriging as the best solution.

2.3.2.2 Artificial neural networks

The idea of an ANN is to simulate the information flow process in the brain. Artificial neurons are basic units of ANNs which imitates brain neurons. An artificial neuron can process incoming signals sent from other neurons and can send signals to the other neurons via connections (edges), which imitates brain synapses. In ANN, a sent signal from a specific neuron represents a value that is obtained by a non-linear function of incoming signals, i.e. incoming values. All connections have assigned weights, which imitate signals strength, and they are obtained in the ANN training process. Whether the signal will be sent from a neuron depends on whether the strength from all incoming signals cross some threshold value. All neurons are grouped into input, hidden, and output layers (Figure 2.5). Finally, a prediction is made by a let signal traveling from the input layer, through hidden layers, to the output layers.

In terms of spatial interpolation, covariates represent neurons in the input layer and observations represent neurons in the output layer. A linear combination of covariates at spatial location $\mathbf{s}_0$ transformed by some nonlinear function represents neurons in the hidden layer:

$$h_j(\mathbf{s}_0) = g(\sum_{k=0}^{n_c} \beta_{kj} x_k(\mathbf{s}_0)) \tag{2.23}$$

where n_c is the number of covariates, β_{kj} is the regression coefficient or weight which shows the influence of the k-th covariate on the j-th hidden neuron, $x_k(\mathbf{s}_0)$ is k-th covariate value at spatial

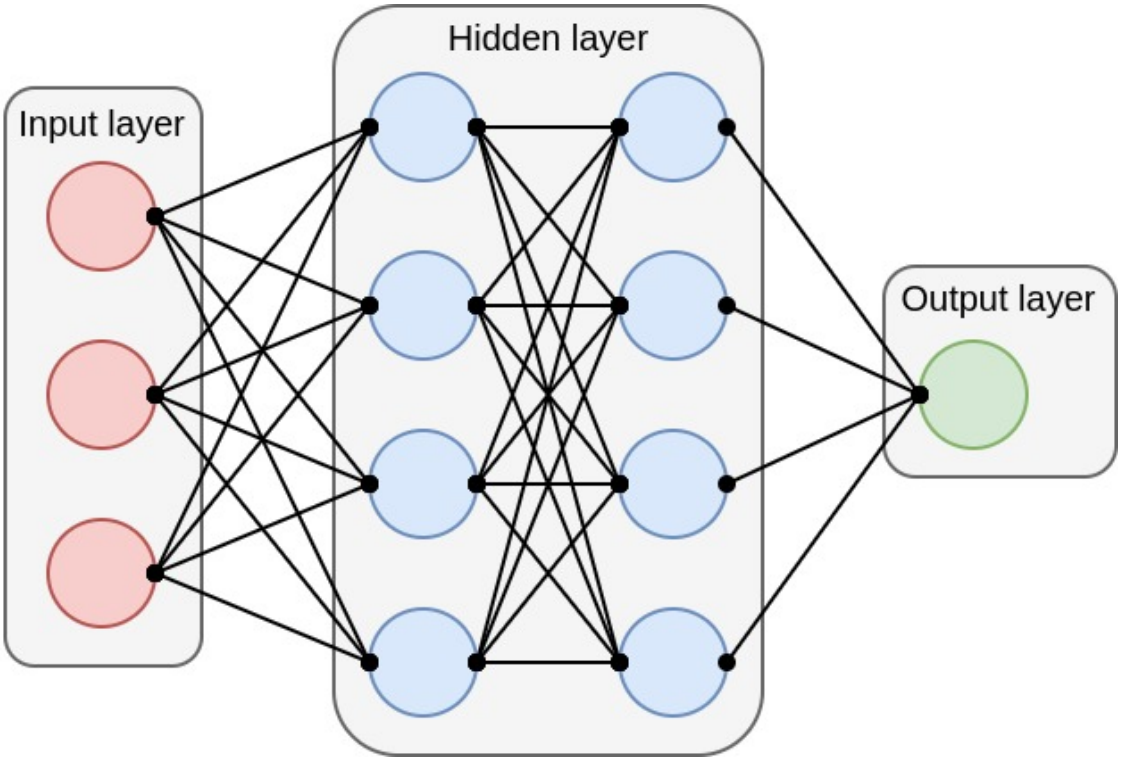


Figure 2.5: ANN with two neuron layers in the hidden layer.

location $\mathbf{s}_0$, and g is a non-linear function, called transfer or activation function, which usually gets for of sigmoidal function:

$$g_u = \frac{1}{1 + e^{-u}}. \tag{2.24}$$

If the g is chosen to be an identity function, ANN becomes an MLR problem.

Mathematically observed, the task of ANN is firstly to create new features (neurons) in the hidden layer from input covariates as non-linear functions of their linear combinations, and then to model a target variable from these new features in a similar manner. The spatial prediction of ANN at spatial location $\mathbf{s}_0$ is given as:

$$\hat{z}(\mathbf{s}_0) = \sum_{k=0}^{n_h} \alpha_k h_k(\mathbf{s}_0) \tag{2.25}$$

where n_h is the number of neurons in the hidden layer and α_k is the regression coefficient or weight which shows the influence of the k-th hidden neuron on the prediction. In the training phase, ANN is adjusting the difference between inputs and outputs (sum-of-squared errors) by a back-propagation algorithm (Rumelhart et al. 1986), which actually iteratively corrects the weights of the connections between neurons in input, hidden, and output layers, i.e. β and α weights. The back-propagation algorithm recalculates these weights by going backward through ANN, from the output layer to the input layer, and the iterative process stops when the sum-of-squared errors is not reduced any more. Often, ANN have more than one hidden layer. In that case, neurons in the next hidden layer represent a linear combination of neurons in the previous hidden layer. More details about the back-propagation algorithm and ANN can be found in Hastie et al. (2009), Kanevski et al. (2009).

ANN have been used for more than a decade in meteorology and climatology (Tveito et al. 2006). Tveito et al. (2006) also gave an extensive list of ANN applications in these areas. Rigol et al. (2001) used ANN for the first time in the spatial interpolation of the daily minimum air temperature and concluded that ANN has comparable accuracy with MLR or TPS with residual kriging.

2.3.2.3 Support vector machines

Support vector machine (SVM) algorithm (Vapnik 1995) represents a generalization of support vector classifier (SVC), which is a generalization of maximal margin classifier (MMC). These three algorithms are often mixed and called all together SVM. All three algorithms, in their original classification form, solve the problem of finding an optimal hyper-plane in the n-dimensional space (n is the number of covariates), which best separates binary observations into its two classes.

MMC is the most basic of all three and works in cases where the observation classes are separable by a linear hyper-plane. The minimal distance between observations and an observed separating hyper-plane is called the margin. The optimal separating hyper-plane, the so-called maximal margin hyper-plane, is the one with the largest margin, i.e. with the largest minimum distance to the observations. The observations that lie on the margin are called support vectors (SV), because they are actually vectors in an n-dimensional space and they "support" a separating hyper-plane in such a way that the hyper-plane position directly depends solely on them. If SV changes their position, the separating hyper-plane will also change position.

The problem with MMC is that it will not work in the case where observation classes are not clearly separable. For this case, the SVC algorithm, or soft margin classifier, that finds a linear separating hyper-plane that separates the classes with the smallest number of misclassified observations is used. Here, the optimal separating hyper-plane depends on the observations on the margin and the observations on the wrong side of their class margin. In this case, these observations are SV. The linear SVC model can be represented as a linear combination of the inner products of the observations:

$$f(x) = \sum_{i \in S} \alpha_i \sum_{j=1}^{n_c} x_j x_{ij} = \sum_{i \in S} \alpha_i \langle x, x_i \rangle \tag{2.26}$$

where S is a set of SV observations, n_c is the number of features (covariates), α_i is the coefficient for the i-th inner product $\langle x, x_i \rangle$ between a new point (at location x) and i-th observation (at location x_i), α_0 is the intercept (for $\langle x, x_0 \rangle = 1$), and x_j and x_{ij} are the j-th features of the new point and i-th observation. The α coefficients are estimated by using inner products between all observation pairs.

Linear separating SVC hyper-plane is often not a good solution. Therefore SVM extends SVC by finding a more flexible non-linear separating hyper-plane. The inner product of the SVC ($\langle x, x_i \rangle$, Eq. 2.26) is substituted with a non-linear kernel — "*a non-linear function that quantifies the similarity of two observations*" (James et al. 2013). The SVM non-linear model is given by:

$$f(x) = \sum_{i \in S} \alpha_i K(x, x_i) \tag{2.27}$$

where $K(x, x_i)$ is the kernel between the new point (at location x) and i-th observation (at location x_i). The kernel is mostly in the form of the polynomial kernel of degree d:

$$K(x, x_i) = (1 + \sum_{j=1}^{n_c} x_j x_{ij})^d \tag{2.28}$$

When the degree (d) of the kernel equals 1, SVM actually takes the form of SVC. If we use Eq. 2.26 to solve a spatial interpolation problem, then the spatial prediction from SVC at spatial location $\mathbf{s}_0$ is given by:

$$\hat{z}(\mathbf{s}_0) = \sum_{i \in S} \alpha_i K(x(\mathbf{s}_0), x(\mathbf{s}_i)) = \sum_{i \in S} \alpha_i (1 + \sum_{j=1}^{n_c} x_j(\mathbf{s}_0) x_j(\mathbf{s}_i))^d \tag{2.29}$$

where $K(x, x_i)$ is the kernel between the new point at spatial location $\mathbf{s}_0$) and i-th observation at spatial location $\mathbf{s}_i$), and $x_j(\mathbf{s}_0)$ and $x_j(\mathbf{s}_i)$ are the j-th covariate values at a new spatial location $\mathbf{s}_0$ and at i-th observation spatial location $\mathbf{s}_i$.

SVM was originally developed for classification and later adapted for regression problems. SVM is mostly used for interpolation, i.e. downscaling of precipitation (Tripathi et al. 2006; Anandhi et al. 2008), especially precipitation occurrence classification (Chen et al. 2010).

2.4 Combined methods

Combined interpolation methods use both environmental covariates and station observations for spatial interpolation. Most of the combined methods used in climatology and meteorology are two-step methods, where environmental covariates and covariate-based interpolators (Section 2.3) are used for modelling of the trend in the first step, and then in the second step any of station-based interpolators (Section 2.2), mostly kriging or IDW, are used for modelling of trend residuals. Besides the two-step approach, there are methods where interpolation is done in one step, i.e. with a unique spatial interpolation model. Kriging with external drift is the kriging version of one-step combined methods. Kriging with external drift is the kriging version for one step methods.

In the last few years, the increasingly popular topic in spatial interpolation is how to incorporate spatial context into ML methods. The literature review of newly developed spatial machine learning methods is given in Section 2.4.4.

2.4.1 Residual (regression) kriging

OK assumes second-order stationarity and, hence, that the mean of the underlying random function is constant. In order to include environmental covariates into the kriging modelling, residual kriging is used. Residual kriging or detrended kriging imposed itself as a two-step interpolation method that separates trend and residual modeling. The trend could be modeled using linear, machine learning, or any other regression technique.

The very first and most popular version of residual kriging is regression kriging (Hengl et al. 2007). Different from OK, the trend is assumed to be a linear combination of covariates, i.e. the trend is modelled using MLR (Ahmed and De Marsily 1987; Hengl et al. 2012; Kilibarda et al. 2014). These covariates must be known at all prediction locations and must be correlated with the dependent variable. Even though residual kriging and regression kriging are practically synonyms, residual kriging can be taken as a more general term than RK, where the trend can be modelled with any regression interpolation method. The word regression in RK refers to the MLR trend.

In a purely spatial variant of RK, the MLR and SK (or OK) are combined (Ahmed and De Marsily 1987; Odeh et al. 1995):

$$Z(\mathbf{s}) = m(\mathbf{s}) + V(\mathbf{s}) \tag{2.30}$$

where m is a deterministic component of the variable (trend) and is modeled using MLR (Eq. 2.12), and $V(\mathbf{s}, t)$ is a zero-mean spatial stochastic (regression) residual and is modeled using the SK (or OK), i.e. the spatial semivariogram (Eq. 2.6). Combining the Eqs. 2.12 and 2.1, RK prediction at a location s_0 is obtained as:

$$\hat{z}(\mathbf{s}_0) = \sum_{k=0}^{p} \beta_k x_k(\mathbf{s}_0) + \sum_{i=1}^{n} w_i \cdot z(\mathbf{s}_i) \tag{2.31}$$

where the β_k are estimated regression coefficients, the $x_k(\mathbf{s}_0)$ are covariates values at location s_0, p

is the number of covariates, w_i are kriging weights for the $z(\mathbf{s}_i)$ observations at n nearest locations.

MLR finds the relationships between covariates and the observed variable. Then, it uses the MLR model on the covariates to make a prediction at unknown locations. Even though multi-iteration generalized least squares (GLS) represent an optimal solution for MLR trend modeling, Kitanidis (1993) showed that OLS produces almost the same results as GLS in the case of kriging. Based on stationarity assumption, which means that the mean and variance of the residuals are constant throughout the spatial domain (the mean is zero), SK can be used for MLR residual modeling. The spatial correlation between the MLR residuals can be explained using a spatial semivariogram. Thus, RK provides better results than MLR and SK (or OK) used independently, except in special cases. For example, when MLR describes all of the variability of the observed variable and residuals have no spatial structure, then there is no need for SK (or OK). In the opposite case, when there are no relationships between the covariates and the observed variable, then only SK (or OK) could be performed (Zhu et al. 2013).

In recent years, RK has replaced OK as the main geostatistical interpolation technique (Hengl et al. 2012; Kilibarda et al. 2014). By modelling temporal and spatio-temporal correlation of the regression residuals, spatial RK can be extended to spatio-temporal RK. Spatio-temporal RK is explained in detail in Section 4.3.1.

RK is widely used in modelling of climate elements. Perčec Tadić (2010) used RK for mapping of twenty climatological variables at a spatial resolution of 1 km over Croatia, for the 1961–1990 period. Bajat et al. (2013) and Bajat et al. (2015) used RK for interpolation of annual LTM mean temperature and precipitation over Serbia, for the 1961–2010 and 1961–1990 periods, respectively. Wu and Li (2013) created a gridded temperature dataset over the US using regression kriging for interpolation of the average monthly temperature for January and July, 2010. Hengl et al. (2012) used spatio-temporal extension of regression-kriging in the interpolation of mean daily temperature over Croatia for the year 2008, while Kilibarda et al. (2014) did the same thing, but for maximum, minimum, and mean daily temperature at a spatial resolution of 1 km for the global land mass.

Except RK, many different approaches to residual kriging exist in literature. One approach is proposed by Haylock et al. (2008) for the interpolation of daily precipitation totals and monthly mean temperature. Monthly precipitation and temperature values were modelled with TPS. Then daily anomalies, in this case residuals, were modelled with IK and UK for precipitation and KED for temperature. Using the same methodology as Haylock et al. (2008), van den Besselaar et al. (2011) created a daily gridded data set for sea level pressure over Europe at 0.25 and 0.5° spatial resolution (the same as for E-OBS data, Section 3.3.1.1). Brinckmann et al. (2016) interpolated maximum, minimum, and mean daily air temperature and daily mean wind speed over Europe, for the 2001–2010 period at a spatial resolution of around 5km. They also used residual kriging, similar to Haylock et al. (2008) and van den Besselaar et al. (2011), with a difference of modelling of daily anomalies with SK. Krähenmann and Ahrens (2013) had a slightly different approach. They modelled gridding of daily maximum and minimum 2 m temperature monthly averages of maximum and minimum temperatures with RK, and then used SK for the interpolation of daily anomalies, for the Central European region and the Iberian Peninsula, for January and July of the 2009–2011 period.

Sun et al. (2015) used the so called geographically weighted regression kriging, where GWR was used for trend modelling and OK for residual modelling, for modelling of mean annual precipitation over China. This method gave the best prediction accuracy in comparison with MLR, GWR, and local RK.

Another approach is to interpolate the ML residuals using residual kriging (Li et al. 2011). Appelhans et al. (2015) combined Cubist and residual kriging approach to model monthly air temperature at Mt. Kilimanjaro, Tanzania. Xu et al. (2020) used a combination of RF and area-to-point kriging for residuals to downscale land surface temperature in Guangzhou, China. Seo et al. (2015) combined

ANN and residual kriging (NNRK) and also regression kriging and neural network residual kriging (RKNNRK) for the interpolation of precipitation. They showed that these two methods outperformed SK, OK and UK.

2.4.2 Residual IDW

As for the residual kriging, the trend and residuals are modelled in two steps. The only difference is that the residuals are modelled with IDW, i.e. w_i in Eq. 2.31 are IDW weights (Eq. 2.3).

Perry and Hollis (2005) created a monthly and annual gridded dataset for 36 climate variables at spatial resolution of 5 km over the UK, for the 1961–2000 period. The dataset is created using residual IDW, where the trend was modelled by MLR. Zhang et al. (2017) modelled long term mean annual precipitation over the Three Gorges Region basin, China and concluded that the hybrid SVM model, that uses SVM to model trend and IDW to model residuals, obtained superior results over IDW, OK, and RK.

2.4.3 Kriging with external drift

Kriging with external drift is a variant of kriging where linear regression is used for the modelling of the trend, but unlike RK, trend modelling is included in the kriging process and the computation is done in one step. Actually, KED prediction can also be presented with the Eq 2.1, but covariates are included in the weights (w_i) calculation process (the covariance matrix of residuals is extended with the covariates (Hengl et al. 2012)). Wackernagel (2003) started to use the term KED as an improved version of UK by introducing environmental covariates in the trend modeling instead of coordinates.

The terms RK, UK, and KED are often used interchangeably. Although these interpolation methods have differences in the means of computation, the predictions and accuracy of the predictions are the same (proof Hengl et al. 2007, Appendix).

KED is often used for the interpolation of temperature (Hudson and Wackernagel 1994; Roznik et al. 2019). Bostan et al. (2012) compared MLR, GWR, OK, RK, and KED for the interpolation of the average annual precipitation over Turkey. KED was the most accurate interpolator. Berezowski et al. (2016) used KED for the generation of grids for daily maximum and minimum air temperatures and precipitation totals at a spatial resolution of 5 km over the Vistula and Oder basins in Poland, for the 1951–2013 period.

2.4.4 Spatial machine learning methods

Until a few years ago, spatial interpolation with ML algorithms strictly relied on relations between environmental covariates and a target variable. More and more researchers are now trying to introduce spatial context in ML algorithms, mostly by inventing additional covariates, so called "spatial covariates", that are derived from spatial locations of observations. Unlike residual kriging and residual IDW, modelling and prediction processes of the spatial machine learning methods are done in one step, and can successfully model non-linear relations among all covariates together, spatial and non-spatial. Because of that, spatial machine learning methods can be useful for modelling of complex variables (e.g. precipitation).

Most of the newly developed frameworks for spatial interpolation with ML use the RF algorithm. The most simple approach is to use coordinates of observations, geographical (latitude and longitude) or in projection (x and y), as covariates (e.g. Li et al. 2011; Mohsenzadeh Karimi et al. 2018;

Behrens et al. 2018). da Silva Júnior et al. (2019) compared, so called, "Coordinate-based Random Forest" with IDW, OK, and RF for the interpolation of evapotranspiration, in the northeast region of Brazil, in January 2017. It turned out that Coordinate-based Random Forest did not perform any better than RF, and similar to IDW and OK. Furthermore, using coordinates as covariates can also cause orthogonal artifacts on a prediction map (Behrens et al. 2018; Hengl et al. 2018; Møller et al. 2020).

Some of the spatial covariates are evaluated on climate variables. He et al. (2016) introduced Prec-DWARF (Precipitation Downscaling With Adaptable Random Forests), where precipitation at adjacent grid cells are used as covariates for the downscaling of precipitation. Hengl et al. (2018) introduced "buffer distance maps" from observation points as spatial covariates in the RF model, and named this framework as Random Forest for spatial prediction — RFsp. RFsp was evaluated, among others, on precipitation case studies. Baez-Villanueva et al. (2020) created a Random Forest based MErging Procedure (RF-MEP) for the interpolation of daily precipitation over Chile for the 2000–2016 period. RF-MEP is actually a model based on RFsp (Hengl et al. 2018). Zhu et al. (2019) added weights based on altitude and distance differences between the target station and surrounding stations as covariates in SVM, ANN, and RF models. These models were named Geoi-SVM (Geo-Intelligent SVM), Geoi-BPNN (Geo-Intelligent Back Propagation Neural Network) and Geoi-RF (Geo-Intelligent RF), respectively, and were used for the interpolation of surface air temperature over China.

Some of the spatial covariates are evaluated on soil mapping case studies, but can easily be applied on climate elements and therefore are worth mentioning. Behrens et al. (2018) used coordinates in projection (x and z), distances to the corners and center of a bounding box around the sampling locations as covariates in the RF model for soil mapping. So far, the last published research by Møller et al. (2020) introduces coordinates along several axes tilted at an oblique angle, so called "oblique geographic coordinates", as covariates in RF, for digital soil mapping and additionally for precipitation. RF with oblique coordinates as covariates outperformed kriging and methods with distance-based covariates.

Another approach to include spatial context in ML algorithms is to fit multiple local models on different spatial locations, based on n nearest observations. The idea for this methodology comes from GWR. Georganos et al. (2019) proposed Geographical Random Forest (GRF), which works on this principle, for modelling population density in Dakar, Senegal. They used the spatially nearest ML model for prediction. A different approach is to use a weighted average of n spatially nearest ML models for prediction as Hashimoto et al. (2019) did with AINA methodology. They fit a multiple RF model for each grid cell, based on n nearest observations, and then a prediction is made with 16 surrounding RF models. This way they made gridded datasets at a 1 km spatial resolution for 30 daily climate variables over the conterminous United States, for the 1979–2017 period.

A literature review of the spatial machine learning methods is also given and further discussed in Chapter 5, where one of the main contributions of this dissertation, Random Forest Spatial Interpolation, is presented.

Chapter 3

Open daily climate datasets

Open daily climate datasets that are used or discussed in this dissertation are presented in this chapter. The most popular open repositories of global and regional observations at meteorological/climate stations (observational data) are described first. Then, some of the gridded daily climate datasets and their characteristics, such as creation methodology, versions, spatial resolutions, time periods they cover, etc., are presented. Gridded daily climate datasets are grouped into (1) station-based, (2) remote sensing-based, and (3) reanalysis datasets. Additionally, environmental covariates used in this dissertation are presented in the end.

3.1 Introduction

Open knowledge foundation[1] defined open data as: "Open data and content can be freely used, modified, and shared by anyone for any purpose". By Dietrich et al. (2015), open data have to follow next principles:

- *Availability and Accessibility*: the data must be available as a whole, preferably over the Internet.
- *Re-use and Redistribution*: the data must be in standardized form in order to be re-used and so that the results based on the data can be redistributed.
- *Universal Participation*: everyone must be able to re-use and redistribute the data.

The open data concept is important because of interoperability, i.e. to enable data exchange and collaboration between the data users. In other words, the main idea of open (science) data concept is to make observations and scientific results freely available to all kinds of users in general, so that they can be used and analysed further. The major user of open data is a scientific community which uses them to create new scientific results, i.e. new open data, that can be verified and reproduced.

Open data and open access data are practically the same terms, with the difference being that open access data has assigned copyright. This means that open access data, unlike open data, has redistribution constraints and has to be cited or acknowledged. Because of this insignificant difference, the term open data refers to both, open and open access data and can be seen as publicly available data.

Even though the idea of open data has been promoted for almost 70 years, the advent of the Internet gave this idea the support to be practically realized, because of the low costs and the Internet

[1] https://opendefinition.org/

availability. The establishment of the open data concept was actually connected to the climate data and formation of the World Data Centers, the World Data Centers for Meteorology and Geophysics among others, operated by the National Oceanic and Atmospheric Administration (NOAA), in the late 1950s (Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council 2008).

High-quality and high-resolution open climate data is widely used for research in various fields, such as meteorology, climatology, hydrology, ecology, agronomy, and others. Open climate data exists in two forms: (1) observations from weather stations (observational data) and (2) gridded climate data. Unlike observations from weather stations which represent time series of measurements collected at specific spatial locations (points), gridded climate data represents time series of grids (rasters) and so describes the variation of a climate variable in a whole spatial domain which is required in many applications (Abatzoglou 2013). Various methodologies are used to create gridded climate data. Three methodologies that are commonly used, (1) from weather stations, (2) from remote sensing (satellite) data, and (3) by climate reanalysis, are described in Section 3.3. Gridded climate data exists at different spatial (from 20 m to 500 km) and temporal resolution (from hourly to long term means products for a 30 year period) (Kilibarda et al. 2015; Sekulić et al. 2020b).

Mendelsohn et al. (2007) compared the performance in agriculture of observations at weather stations and satellite products for temperature and precipitation in Brazil, India, and the US. The satellite products gave better results in the case of temperature because they provide complete spatial coverage and observations at weather stations are sparsely spatially distributed, especially in rural areas. On the contrary, in the case of precipitation, satellite products could not measure precipitation accurately and so observations at weather stations are more preferred in this case. Kilibarda et al. (2015) did preliminary spatio-temporal analysis of global temperature stations and show that the spatial distribution of stations is mostly conditioned by environmental factors, such as population density and accessibility, which means that station density is lower in the areas at higher altitudes (mountains), polar areas, deserts, tropical forests, etc. Due to the fact that weather stations do not cover the spatio-temporal domain representative enough "from the point of view of spatio-temporal statistics" (Heuvelink et al. 2012), they concluded that spatio-temporal interpolation methods, such as RK, can create unbiased daily gridded temperature data.

Open climate data is available in different spatial support: global, regional, and local (national). The focus of this chapter is on global and European (regional) open daily climate data that was used or discussed in this dissertation.

3.2 Observational data

The most accurate and reliable climate data comes from observational data, i.e. observations at weather stations. The largest part of these observations exists on a national level, maintained by national (hydro) meteorological institutes. Often, these national repositories are not publicly available. Therefore, regional or global meteorological organizations, such as the Royal Netherlands Meteorological Institute (Dutch: Koninklijk Nederlands Meteorologisch Instituut - KMNI) and NOAA, have created regional or global repositories of daily observational data observations and made them publicly available, especially for research.

The focus of this section is on the open global and regional (European) daily observational data. Most of it is based on surface synoptic observations (SYNOP) from the WMO, including a portion of stations from national weather station networks.

3.2.1 OGIMET service

OGIMET[2] is a Weather Information Service which provides, among other data, historical daily summaries from the SYNOP reports for the period starting from the year 2000. SYNOP reports are meteorological alphanumeric messages for reporting observations, from more than 13,000 meteorological stations around the world. Reports are mostly available every 6 h (00, 06, 12 and 18 UTC), but for some stations every 3 or 1 h. The format of these reports is standardized and defined by the WMO. OGIMET collects SYNOP reports mainly from the NOAA FTP server.

Another similar service is Meteomanz[3].

3.2.2 GSOD

National Centers for Environmental Information (NCEI) — former NOAA's National Climatic Data Center (NCDC) — which is a part of US Federal Climate Complex (FCC), provides a Global Surface Summary of the Day (GSOD)[4] — a global dataset of the daily summaries (mean unless otherwise noted) of meteorological variables, namely: temperature (mean, maximum, and minimum), dew point, sea level pressure, pressure, visibility, wind speed (mean and maximum), maximum wind gust, precipitation amount, snow depth, indicator for occurrence of fog, rain or drizzle, snow or ice pellets, hail, thunder, tornado/funnel cloud. These variables are measured with high precision of a 0.1 variable unit, e.g. 0.1 °F (0.055 °C) for temperature, 0.1 inches (2.54 mm) for precipitation, 0.1 mbar for mean sea level pressure.

GSOD is made by aggregating global hourly SYNOP observations from more than 14.000 stations, stored in The Integrated Surface Database (ISD) maintained by US Air Force Combat Climatology Center, which is also part of FCC. The data is mostly available from over 9000 stations, covering a time period from 1929 to the present, with the most complete data starting from the year 1973. The daily summaries are available two days after an actual measurement was captured. Daily summaries are calculated only if there is a minimum of four observations at the station during the day, because of synoptic stations that measure four times a day. Since synoptic data follows Greenwich Mean Time (GMT), GSOD data is summarized each day at midnight by GMT. GSOD data undergoes extensive automated quality control of SYNOP reports and summaries.

Due to the fact that SYNOP data that is exchanged according to the WMO Resolution 40 (Cg-XII), WMO member countries can place restrictions on the use of GSOD data. But, in general, GSOD data is intended for "*free and unrestricted use in research, education, and other non-commercial activities*".

3.2.3 GHCN-daily

The Global Historical Climatology Network-Daily (GHCN-Daily) (Menne et al. 2012) is a dataset of daily climate summaries from more than 100,000 meteorological stations and more than 25 data sources in 180 countries and territories from all over the World. These sources mostly include GSOD stations, US stations, stations from an International collection outside of the US, as well as stations from National Meteorological and Hydrological Centers. The time period that they cover ranges from 1 year to 175 years, with the maximum station density starting from the 1960s. Same as GSOD, GHCN-Daily dataset is provided by NCEI. Daily climate summaries are available for 40 meteorological elements, where five of them are core variables: total daily precipitation, daily maximum

[2] https://www.ogimet.com/
[3] http://www.meteomanz.com/
[4] https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod

and minimum temperature, temperature at the time of observation, snowfall and snow depth. Most of the GHCN-daily stations, one half to two thirds approximately, measure total daily precipitation only.

Observations of GHCN-Daily dataset (started from the latest version 3) are updated every day from sources where this is possible. Then, the reconstruction of the dataset is usually obtained once a week in order to be synced with all of the data sources. Compared to GSOD, GHCN-daily undergoes a more detailed quality assurance (QA) check on a daily basis. QA checks start from simple checks like impossible time instances or values (exceeding variable limits), invalid characters, duplicates, etc. to more detailed data consistency checks. These QA checks are explained by Durre et al. (2008) and Durre et al. (2010). In the end, each of the observations will have been assigned a QA flag.

3.2.4 ECA&D

European Climate Assessment & Dataset (ECA&D)[5] is a project that has started in 2003 and was funded by the European Commission and EUMETNET (network of European National Meteorological Services), with the aim of collecting daily meteorological observations for temperature and precipitation across Europe and the Mediterranean in order to monitor and analyse climate changes and extremes (Klein Tank et al. 2002). From 2009 onwards, ECA&D has been completely funded by the KMNI, which was a project member from the beginning of the project. ECA&D now has the status of the regional climate center for Europe and the Middle East.

ECA&D collects daily observations from more than 20,000 stations across Europe, SYNOP stations and stations from the National Meteorological and Hydrological Services, observatories, and research centres, counting 79 participants in 65 countries. The data is collected for climate elements such as maximum, minimum, and mean temperature, sunshine, snow depth, precipitation, global radiation, sea level pressure, humidity, wind gust, speed, and direction, cloud cover. Around three fourths of the daily data are available for non-commercial research and education. This dataset covers the 1946–present period. The data is updated once a month by undergoing a two step quality control procedure, where the first step is to apply common homogeneity tests and the second step is to divide the data into three classes: OK, suspect, or missing. In the end, quality and homogeneity flags are attached to each observation. Two versions of ECA&D daily data are available, blended and non-blended. Blended data is a complete series of the observations, where the incomplete series are fulfilled with SYNOP data of the nearby stations, while non-blended data contains a series with missing values, i.e. the data series as provided by ECA&D participants.

ECA&D data are intended for non-commercial research and education use.

3.3 Gridded data

The main aim of the interpolation methods in Chapter 2 is to produce gridded datasets by assigning the interpolated variable values to pixels of the regular grid. Some of the station-based climate datasets are presented first. This is not the only approach to gridding the climate elements. The two other approaches, the one that applies algorithms to remote sensing data and the one that does the reanalysis of the various historical climate observations, are presented next. Some environmental covariates that are a useful source of information in the process of climate elements interpolation, are described also. Studies that address analysis of spatial and/or temporal variability and changes of the climate variables rely on gridded climate data.

[5] https://www.ecad.eu/

3.3.1 Station-based datasets

The first group of gridded datasets are station-based datasets. They use observations from weather stations and their spatial dependency to provide gridded climate data. They are created mostly using the station-based interpolators presented in Section 2.2. Time series of station observations are used to create a time series of gridded climate data, i.e. gridded climate datasets.

3.3.1.1 E-OBS

The blended ECA&D daily dataset is also used to provide E-OBS (Cornes et al. 2018) gridded land-only observational dataset over Europe for daily mean, minimum, and maximum temperature, precipitation amount, averaged sea level pressure, and solar radiation. E-OBS dataset is actually an ensemble dataset constructed through a conditional simulation procedure. For each of the 100 members of the ensemble, a spatially correlated random field is produced using a pre-calculated spatial correlation function. The mean across the members is calculated and is provided as the "best-guess" fields. E-OBS is a daily dataset, as ECA&D dataset, covers the whole of Europe (25°N-71.5°N; 25°W-45°E), with a spatial resolution of 0.1 degrees, which is approximately 10 km. It covers the 1950–present period and is updated twice a year.

E-OBS data is now available through Climate Data Store (CDS 2020) or through Copernicus Climate Change Service (C3S)[6]. As all ECA&D data, E-OBS data is intended for non-commercial research and educational use.

3.3.1.2 CPC

NOAA Climate Prediction Center (CPC) provides global gridded datasets for global maximum and minimum temperature (PSL 2020a) and precipitation (PSL 2020b). CPC datasets are at a spatial resolution of 0.5 degrees (∼50 km) and cover the time period from 1979 to the present.

Maximum and minimum temperature CPC datasets are created by the interpolation of anomalies at more than 6000 global stations, obtained from monthly values from the Climatic Research Unit (CRU), University of East Anglia, UK. Anomalies at stations are interpolated using the Shepard Algorithm, which is a distance-weight interpolator with directional correction (Cressman 1959; Shepard 1968). The precipitation CPC dataset is a product from the CPC Unified Precipitation Project that aims to create unified and improved quality precipitation datasets from all CPC sources. The precipitation CPC dataset is made by using the optimal interpolation (OI) objective analysis technique (Gandin 1965) on more than 16,000 stations.

These datasets are real-time updated. Because of their coarse resolution, they are used for climate monitoring and verification of forecast models.

3.3.1.3 CarpatClim

Climate of the Carpathian region (CarpatClim)[7] (Szalai et al. 2013) was a project that aims to improve climate data for the Carpathian Mountains and the Carpathian basin. The project participants were (hydro)meteorological services or institutes from Hungary, Croatia, Serbia, Romania, Ukraine, Slovakia, Poland, and Czech Republic. Firstly, daily observations at stations in the Carpathian region (from the participants) were collected. Next, these observations from different sources have to

[6] https://surfobs.climate.copernicus.eu/surfobs.php
[7] http://www.carpatclim-eu.org/pages/home/

be homogenized and then quality assessed, because different (hydro)meteorological services have different approaches in collecting the data and used meteorological instruments, quality control procedures, etc. For this purpose, Multiple Analysis of Series for Homogenization (MASH) procedures were used. Finally, MISH interpolation method (similar to RK, Section 2.4.1) was used to interpolate these homogenized observations into grids. This way, the main products of the CarpatClim project — daily gridded datasets for meteorological variables such as mean, maximum, and minimum temperature, total precipitation, 10 m wind direction and speed, sunshine duration, cloud cover, global radiation, relative humidity, surface air and vapour pressure, and snow depth — were created. The description of the MASH and MISH procedures can be found in the project deliverables.

The daily gridded datasets cover around 500 000 km2 in Europe (44°N-50°N; 17°E-27°E), at a spatial resolution of 0.1 degrees (∼10 km), and for the 1961–2010 period. These datasets are intended to be used for regional climate assessment and applied studies and for better understanding of the spatial and temporal climate processes of the Carpathian region. Based on the daily gridded datasets, the Climate Atlas of the Carpathian region was created.

3.3.2 Remote sensing products

Remote sensing data is retrieved from optical, radar or any other instruments (sensors), carried by satellites or placed on the ground. Most of the satellite sensors collect the data in a form of grid and so they already represent gridded data, but still not gridded climate data. In order to produce gridded climate datasets, different algorithms are applied over the satellite sensor bands or over a combination of satellite sensor bands and other ground sensors. MODIS LST is one of the most used remote sensing products in the modelling of daily temperature (see Chapter 4). Beside MODIS LST, other popular remote sensing products are presented in this section.

3.3.2.1 MODIS LST

MODIS is a sensor that operates on two satellites: Terra and Aqua, launched in December 1999 and May 2002, respectively, by the National Aeronautics and Space Administration (NASA). Terra and Aqua satellites are complementary in covering the whole Earth, where Terra is orbiting from north to south and passes the equator in the morning, and Aqua is orbiting from south to north and passes the equator in the afternoon.

MODIS LST is one of many MODIS products. MODIS LST are provided from both satellites, Terra (`MOD*`) and Aqua (`MYD*`). Currently, two MODIS LST products exist: `M*D11` (`MYD11/MYD11`) and newer `M*D21` (`MYD21/MOD21`). These two products have different methodologies for the creation of LST maps and exist on different production levels.

The `M*D11` products are created using the generalized split-window (GSW) (Wan and Dozier 1996) and day/night pair (Wan and Li 1997) algorithms. `M*D11` products are:

- Level 1B — a 5-minute swath (scene) of MODIS data, with the spatial resolution of 1 km at nadir and in a geographic projection (latitude, longitude)
- Level 2 — a 5-minute LST (geophysical) product made by using the GSW algorithm over Level 1B product, with the spatial resolution of 1 km and in a geographic projection (latitude, longitude)
- Level 3 — a LST product that has been temporally or spatially manipulated in a map projection. Level 3 products are:

- M*D11A1 — a daily LST product made by mapping the Level 2 product on a sinusoidal projection, with 1 km spatial resolution
- M*D11A2 — an 8-day LST product made by averaging two to eight days of the M*D11A1 products
- M*D11B1 — a daily LST product made by using the day/night algorithm, with 6 km spatial resolution and in the sinusoidal projection
- M*D11C1 — created from M*D11B1 product by resampling to the Climate Modeling Grid (CMG), with 0.05° spatial resolution and in the equal-angle geographic projection
- M*D11C2 — an 8-day composite of the M*D11C1 product
- M*D11C3 — a monthly composite of the M*D11C2 product

M*D21 products are created using the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) temperature emissivity separation (TES) algorithm (Islam et al. 2017) in order to overcome cold bias (of 3-5 K) in arid and semi-arid areas of the M*D11 products. The TES algorithm is a physics-based algorithm that simultaneously retrieves the LST and Emissivity from MODIS data. M*D21 Level 1B and Level 2 products are the same as for M*D11 products, except that for Level 2 products the TES algorithm is used instead of the GSW algorithm. M*D11 Level 3 products are:

- A1D/A1N — day (A1D) and night (A1N) daily LST products made by mapping the Level 2 product on a sinusoidal projection, with 1 km spatial resolution
- A2 — an 8-day LST product made by averaging two to eight days of the M*D21A1 products

MODIS LST products are widely used for the interpolation of temperature. MODIS LST application in the interpolation of daily temperature is given in the Introduction of Chapter 4.

3.3.2.2 TRMM/IMERG

Tropical Rainfall Measuring Mission (TRMM) (Huffman et al. 2007) was a NASA's satellite intended for the analysis of precipitation over the tropical and subtropical regions of Earth. The TRMM Microwave Imager (TMI) was used to measure microwave energy emitted by Earth and its atmosphere in order to estimate, among other parameters of the atmosphere, the rainfall intensity. It operated from 1997 to 2015.

Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (GPM), in short IMERG (Huffman et al. 2014), is an algorithm, made by NASA, that combines information from multiple sources, such as satellite microwave precipitation estimates, microwave-calibrated infrared satellite estimates, precipitation gauges, and other precipitation estimators to estimate precipitation over the majority of Earth's surface. IMERG provides gridded precipitation estimates at a spatial resolution of 0.1 degrees (∼10 km). Earlier versions of the IMERG dataset, based on GPM, were covering the period from 2014 to the present, but starting from version V06B, IMERG includes TRMM preprocessed data and now covering the period from June 2000 to the present.

IMERG data is available at three levels:

- Early run — available after 6 h. It only uses a forward propagation algorithm which does extrapolation in time.
- Late run — available after 18 h. Compared with Early run, it additionally has the data from 10 h after and then can use both forward and backward propagation algorithms that together do interpolation in time.

- Final run — available after 4 months. Precipitation estimates are additionally adjusted with ancillary data sets, such as monthly observations at the global Precipitation Climatology Centre (GPCC) stations, which are available months after. Sometimes, microwave overpasses can be late to be included in the Late run version, but can be included in the Final version.

Early and Late run products are available at 30-minute, 3-hour, daily, 3-day, 7-day, and 1-month temporal resolutions, while Late run products are available at 30-minute, daily, and monthly temporal resolutions.

3.3.2.3 PERSIANN

Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) (Nguyen et al. 2019) is an ANN classification and approximation system used to provide precipitation estimation based on long-wave infrared brightness temperature and daytime visible images provided by geostationary satellites. Furthermore, the PERSIANN system has adaptive training procedures that can update ANN with new incoming data. The PERSIANN system is developed by the Center for Hydrometeorology and Remote Sensing (CHRS), the University of California, Irvine (UCI).

Five PERSIANN precipitation estimation products are currently available:

- PERSIANN — a near real-time (2 days delay) basic product that covers the March 2000–present period.
- PERSIANN-Cloud Classification System (PERSIANN-CCS) — a real-time high-resolution product which additionally includes the cloud segmentation algorithm that classifies patches of clouds. It covers the January 2003–present period.
- PERSIANN-Climate Data Record (PERSIANN-CDR) — a product intended for long-term analysis. It uses the PERSIANN algorithm over GridSat-B1 infrared data and it is adjusted with GPCC monthly product (similar as IMERG final run). It is periodically updated and covers the January 1983–present period.
- PERSIANN Dynamic Infrared Rain Rate near real-time (PDIR-Now) — a real-time (15 to 60 minutes delay) high-resolution product based on real-time satellite precipitation monitoring system - iRain[8]. It covers the March 2000–present period.
- PERSIANN-CCS-CDR — a high spatial and temporal resolution product that combines the algorithms used for creation of CCS and CDR datasets. GridSat-B1 and NOAA CPC-4km dataset are used in CCS. It covers the January 2003–present period.

All of the products cover are 60°S to 60°N and have 1, 3, and 6-hourly, daily, monthly, and yearly products, except PERSIANN-CDR which does not have hourly products. PERSIANN and PERSIANN-CDR are at 0.25°, while other products are at a 0.04° (∼4 km) spatial resolution.

3.3.2.4 EUMETSAT products

The European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) is an intergovernmental organisation with 30 Member States, based in Darmstadt, Germany. EUMETSAT

[8]http://irain.eng.uci.edu

operates several satellite missions to monitor weather, climate, and environment. These satellite missions are Meteosat, Metop, Sentinel, and Jason.

Meteosat is a series of two-geostationary satellite systems that orbit over Europe, and Africa, and Indian Ocean at altitude 36,000 km. The Meteosat data is used for weather forecasting and climate monitoring. The Meteosat First Generation (MFG) satellites (Meteosat-1 to -7) were launched in 1977, but in March 2017 were all retired. MFG was equipped with a Meteosat Visible and Infrared Imager (MVIRI) sensor and provided images every half an hour. The current operating Meteosat Second Generation (MSG) satellites (Meteosat-8 to -11) were launched in 2004 in cooperation with the European Space Agency (ESA) and orbit over Europe and Africa (Meteosat-9 to -11), and over Indian Ocean (Meteosat-8), providing more frequent and improved images every 15 and 5 minutes (rapid service). All MSG satellites carry the main Spinning Enhanced Visible and Infrared Imager (SEVIRI) and the secondary Geostationary Earth Radiation Budget (GERB) instruments. Meteosat-11 will retire in the year 2033. Meteosat Third Generation (MTG) will be launched in early 2020 onward in cooperation with ESA, in order to continue MSG data collection until the 2040s. MTG will be equipped with an infrared sounder and the Copernicus Sentinel-4 Ultraviolet Visible and Near-infrared instrument.

Metop are polar- and low-orbiting satellites at the altitude of 817 km with the aim of collecting the data for the Pacific Ocean and continents of the southern hemisphere. As Meteosat, Metop data are intended for weather forecasting and climate monitoring. The current operating Metop satellites, Metop-A, -B and -C, were launched in 2006, 2012, and 2018, respectively, and carry eight different main instruments. Metop satellites are part of the Initial Joint Polar System (IJPS), a joint program with the NOAA. Metop-Second Generation (Metop-SG) satellites will be launched in mid 2020 onward, in order to continue Metop data collection until the 2040s. The Metop-SG A and Metop-SG B satellites will operate in three successive pairs and carry enhanced and new instruments, and the Copernicus Sentinel-5 instrument.

Sentinel is a marine and atmospheric satellite mission of Copernicus. Sentinel-3 and -6 are ocean monitoring satellites, while Sentinel-4 and -5 are instruments that will be carried by MTG and Metop-SG satellites and will monitor air quality, Sentinel-4 over Europe and Sentinel-5 in the atmosphere.

Jason is a series of low-orbit satellites, used for measuring mean sea level rise. The current Jason-3 satellite is orbiting at 1336 km altitude and carrying a radar altimeter that measures sea surface,wave height, and wind speed. Together with three previous US/European satellites (TOPEX-Poseidon, Jason-1, and -2), it creates a time series of global mean sea level measurements dating back to 1992.

EUMETSAT offers various products through Product Navigator[9]. 15-minute Meteosat data with a 4 km spatial resolution, such as the Land Surface Temperature - MSG (April 2009–present) and the Multi-Sensor Precipitation Estimate (April 2009–July 2019), can be aggregated to daily data. The EUMETSAT also offers the daily temporal resolution products, created by a specific algorithm. Some of them are Daily Land Surface Temperature - Metop (April 2017–present), Daily Shortwave Solar Irradiance - MSG (October 2011–December 2017), Daily Surface Solar Irradiance - MSG (October 2017–present), and Daily Evapotranspiration - MSG (December 2010–present).

3.3.3 Climate reanalysis datasets

The Numerical weather model (NWM) uses laws of physics to describe the dynamical behavior of the atmosphere. NWM is then used to predict the future states of the atmosphere based on the

[9]https://navigator.eumetsat.int/start

initial states (conditions) of the atmosphere. An initial state of the atmosphere is "measured" with various kinds of climate observations, such as observations at weather stations, satellite sensors, ground sensors, and others. These climate observations contain errors caused by the quality and accuracy of the instruments. In order to combine them so they can produce a stable initial state of the atmosphere, NWM uses a data assimilation process. The result of the data assimilation is the best fit of the NWM to the climate observations at a certain point in time.

Until now, many historical climate observations are collected, especially in the last decades. Climate observations for the present are more or less complete, but as we go back in time, there are fewer and fewer of them. Another problem is that these climate observations are not evenly distributed in the spatial and temporal domain. Climate reanalysis solves this problem. It is a method that combines all available historical climate observations with a single version of NWM. This way all of the historical climate observations are reanalysed with a consistent data assimilation procedure in order to provide consistent initial states for the next short-term forecasts, thus reconstructing spatial and temporal distribution of climate data on different pressure levels in the atmosphere. The final result is a consistent, spatially and temporally complete dataset of the past global weather. Climate reanalysis datasets topically cover several decades in time.

Climate reanalysis datasets are extensively used in climate change research and services, and also agriculture, water resources, and insurance.

Besides NOAA and the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis datasets presented here, there are many other sources of climate reanalysis. One such is the Japanese 55-year Reanalysis (JRA-55) provided by the Japan Meteorological Agency (JMA) and Modern Era Reanalysis for Research and Applications Version-2 (MERRA-2) provided by NASA.

3.3.3.1 NOAA datasets

The NOAA Physical Sciences Laboratory (PSL) provides gridded climate datasets with various methodologies and at various spatial and temporal resolution[10]. In section 3.3.1.2 the two NOAA CPC station-based daily gridded datasets, provided by PSL, are presented.

Among others, PSL provides global gridded reanalysis datasets, such as the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis 1 (Kalnay et al. 1996) and NOAA-CIRES 20th Century (20C) Reanalysis (Compo et al. 2011), which are available at daily temporal resolution. Both datasets are at a spatial resolution of 2.5 degrees ($\sim$210 km).

NCEP and NCAR were participated in the project called "Reanalysis" with the aim "*of producing a 40-year record of global analyses of atmospheric fields in support of the needs of the research and climate monitoring communities.*", for the 1957–1996 period. Nowadays, NCEP/NCAR reanalysis 1 covers the 1948–present period and assimilate and control the quality of land surface, ship, rawinsonde, pibal, aircraft, satellite, and other data. NCEP-DOE Reanalysis 2 (Kanamitsu et al. 2002) is an improved version of the NCEP/NCAR Reanalysis 1 that covers the 1979–present period. This new reanalysis updates the assimilation system, fixes errors, and updates parameterization of physical processes.

NOAA-CIRES 20C Reanalysis are global atmospheric circulation dataset that cover the period from the early 19th century to the 21st century (1850–2014). This dataset is intended for validation of daily climate model simulations of the 20th century. NOAA-CIRES 20C Reanalysis are created using the Ensemble Kalman Filter for data assimilation (Compo et al. 2011). Three versions of NOAA-CIRES 20C Reanalysis exist: V2, V2c, and V3, where each new version improves previous one with methodology improvement and new input datasets.

[10] https://www.psl.noaa.gov/data/gridded/index.html

3.3.3.2 ECMWF datasets

The European Centre for Medium-Range Weather Forecasts is a research institute that produces global numerical weather predictions and other various data for many users worldwide, based on its climate data archive which is the largest in the world. ECMWF also operates the Copernicus Atmosphere Monitoring Service (CAMS) and the C3S[11], and contributes to the Copernicus Emergency Management Service (CEMS). The ECMWF data can be retrieved through the Meteorological Archival and Retrieval System (MARS) via the MARS and Python client or web interface.

Besides various real-time and historical climate datasets, the ECMWF provides many reanalysis datasets. ERA-Interim (Dee et al. 2011) is an ECMWF global atmospheric reanalysis dataset at a spatial resolution of approximately 80 km, covering the period from January 1979 to August 2019. The ERA-Interim system uses a 4-dimensional variational analysis with a 12-hour analysis window to estimate a large number of atmospheric, land and oceanic climate variables.

ERA-Interim reanalysis are replaced with ERA5, an hourly reanalysis dataset (Muñoz Sabater 2019) for the 1979–present period with an improved data assimilation system and at a finer spatial resolution of 0.25° ($\sim$ 30 km). Besides an improvement in spatial and temporal resolution, ERA5 includes various newly reprocessed datasets and new instruments that were not available before, provides information about uncertainties for all variables at reduced spatial and temporal resolutions, and so provides better estimation for many climate parameters in comparison with ERA-Interim. While ERA-Interim, among others, has daily datasets, ERA5 has only hourly and monthly datasets. For daily climate analysis the ERA5 hourly dataset has to be aggregated to a daily temporal resolution. ERA5 is available through ECMWF C3S service.

3.3.4 Other environmental covariates

Environmental covariates are important gridded data for modelling the trend of climate variables. The most used are the digital elevation model (DEM)-derived environmental covariates presented below.

DEMSRE3 is a DEM at a spatial resolution of 1 km produced by combining NASA's Shuttle Radar Topography Mission (SRTM) 30+ (Rabus et al. 2003) and ETOPO DEM (Amante and Eakins 2009) provided by NCEI (NOAA). Another source of DEM at 1 km is WorldClim (Fick and Hijmans 2017).

TWISRE3 is a DEM product at a spatial resolution of 1 km derived from the SAGA GIS Topographic wetness index (TWI) (Beven and Kirkby 1979). TWI quantifies a topographic control on hydrological processes using a function of local upstream contributing area per unit (total catchment area divided by flow width) and local slope.

INMSRE3 is a mean potential incoming solar radiation product at a spatial resolution of 1 km derived in SAGA GIS (Böhner and Antonić 2009). Mean potential incoming solar radiation is a function of cloudiness (sky view factor), latitude, longitude, DEM, and other atmospheric inhomogeneities.

DICGSH1 is a product that represents distance to the nearest coast at a spatial resolution of 1 km. Distance-to-coast can be calculated using only the land boundaries, such as GADM or any other land boundaries. The process starts with making a union of land boundaries, then calculation a distance from each grid pixel to the nearest land boundary. Another source of distance-to-coast product is available by NOAA[12].

All the covariates (Figure 3.1) were downloaded from worldgrids.org (Reuter and Hengl 2012), a

[11] https://climate.copernicus.eu/

[12] https://catalog.data.gov/dataset/distance-to-nearest-coastline-0-01-degree-grid

repository of gridded global environmental covariates, mostly at a 1 km spatial resolution, intended for global soil mapping. This repository is not active any more, but worldgrids.org data archive is still maintained by Hengl (2018). An up-to-date version of most of the worldgrids.org data is now available at OpenLandMap[13] data portal.

[13]https://landgis.opengeohub.org

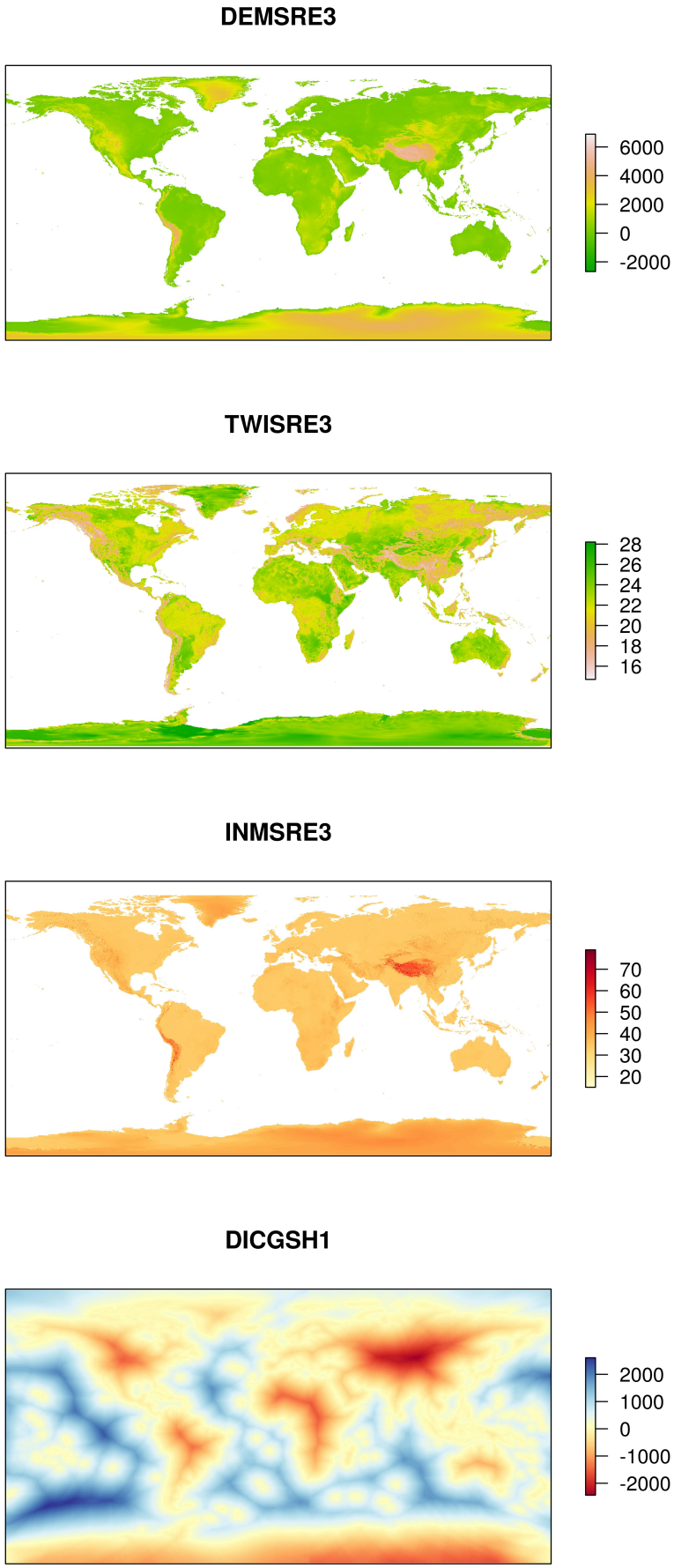


Figure 3.1: DEMSRE3, TWISRE3, INMSRE3, and DICGSH1 provided by worldgrids.org.

Chapter 4

Adaptation of global geostatistical mean daily temperature model to local areas[1]

High resolution gridded mean daily temperature datasets are valuable for research and applications in agronomy, meteorology, hydrology, ecology, and many other disciplines depending on weather or climate. The gridded datasets and the models used for their estimation are being constantly improved as there is always a need for more accurate datasets as well as for datasets with a higher spatial and temporal resolution. A spatio-temporal regression kriging model was developed for Croatia at 1 km spatial resolution by adapting the spatio-temporal regression kriging model developed for global land areas. A geometric temperature trend, digital elevation model, and topographic wetness index were used as covariates together with measurements from the Croatian national meteorological network for the year 2008. This model performed better than the global model and previously developed models for Croatia, based on MODIS land surface temperature images. The R 2 was 97.8 % and RMSE was 1.2 °C for leave-one-out and 5-fold cross-validation. The proposed national model still has a high level of uncertainty at higher altitudes leaving it suitable for agricultural areas that are dominant in lower and medium altitudes.

4.1 Introduction

High-resolution daily temperature gridded datasets are widely used for many purposes. They serve as input data for numerous models across various research fields, such as agronomy, meteorology, hydrology, ecology, and climatology. Researchers use spatial or spatio-temporal interpolation methods to create maps from point data and covariates. Nowadays, point data are available from weather stations on a global level (e.g., GHCN (Menne et al. 2012), GSOD[2]), regional level (ECA&D (Klein Tank et al. 2002)), and local (e.g., national hydrometeorological services) level. Furthermore, many of these point data sources have open data policy so they are easily accessible.

One needs to consider the extent, resolution, and support while performing an interpolation. In this case, the support is a time interval, an area, or a volume over which a measurement or prediction is made. A variety of gridded temperature datasets exists in various spatial and temporal resolutions and supports (an extensive list is available at https://psl.noaa.gov/data/gridded/). For example, researchers have investigated spatial ranges from 5° (Osborn and Jones 2014) to 250 m

[1]Based on article: Sekulić, A., Kilibarda, M., Protić, D., Tadić, M. P., & Bajat, B. (2020). Spatio-temporal regression kriging model of mean daily temperature for Croatia. Theoretical and Applied Climatology, 140(1–2), 101–114. https://doi.org/10.1007/s00704-019-03077-3 (Sekulić et al. 2020b)

[2]https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod

(Holden et al. 2016), temporal resolution ranges from 30 year period (PRISM Climate Group, Oregon State University[3]) to daily (Kilibarda et al. 2014) gridded datasets, spatial extent ranges from areas covering the whole world (Kilibarda et al. 2014) to relatively minute area extents (Rosenfeld et al. 2017), and finally temporal extent ranges from more than 50 years (Oyler et al. 2015) to a single year period (Parmentier et al. 2015). However, the global datasets are not optimal for most of the applications mentioned above due to their coarse spatial and temporal resolution and insufficient accuracy. Coarser spatial and temporal support leads to the averaging of spatial and temporal variability. This results in the omission of microclimatic areas and short time phenomena. Due to the shortcomings of global models, there is a need for the development of local models that can produce gridded datasets at a much finer spatial and temporal resolution with improved accuracy.

Longitude, latitude, and elevation are most commonly used covariates in temperature modeling—especially in linear models (Wu and Li 2013; Yuan et al. 2015). Generalized additive models based on longitude, latitude, and elevation gave the best results for generating a gridded daily dataset for maximum air temperature surfaces at 1 km spatial resolution for the state of Oregon, USA (Parmentier et al. 2014). The elevation is often an essential covariate due to the average temperature decreases with altitude. The elevation is used either directly in temperature models (e.g., Jarvis and Stuart 2001) or in the form of a topographic index or any other DEM derivatives. For example, Dodson and Marks (1997) used elevation in the form of hydrostatic and potential temperature equations in the inverse distance weighting method to interpolate the minimum and maximum temperature at 1 km resolution for the mountainous region in the US Pacific Northwest. However, many other covariates have been proven to be beneficial for temperature interpolation. Courault and Monestiez (1999) used general atmospheric circulation patterns along with elevation, Jarvis and Stuart (2001) introduced land cover as a covariate, specifically useful in modeling urban effects. In recent years, MODIS LST is widely used as one of the most important covariates for temperature interpolation. Zhu et al. (2013), Hengl et al. (2012), Kilibarda et al. (2014), Kilibarda et al. (2015), Williamson et al. (2014), Xu et al. (2014), Kloog et al. (2014), Parmentier et al. (2015), Huang et al. (2015), Stewart and Nitschke (2017), and Li et al. (2018a) used MODIS as the main covariate for their models. The MODIS LST is highly correlated with surface measured air temperature, where specifically daytime images are correlated well with maximum temperatures and nighttime images with minimum temperatures (Oyler et al. 2016). The problem with MODIS LST images is that they have spatial and/or temporal gaps that need to be filled. Filling the gaps using spatial or temporal interpolation together with the processing of images are computationally consuming processes. Proximity to the sea, land cover, vegetation indices, canopy height, cloud cover, etc. are also used as covariates in temperature modeling.

The most commonly used methods for interpolation of temperature involve distance criteria methods (Dodson and Marks 1997; Srivastava et al. 2009), splines—being mostly thin plate splines (Jarvis and Stuart 2001; Hutchinson et al. 2009; Yuan et al. 2015; Stewart and Nitschke 2017), regression and geostatistical methods (Courault and Monestiez 1999; Kurtzman and Kadmon 1999; Hunter and Meentemeyer 2005; Carrera-Hernández and Gaskin 2007; Haylock et al. 2008; Perčec Tadić 2010; Hengl et al. 2012; Wu and Li 2013; Krähenmann and Ahrens 2013; Kilibarda et al. 2014), and recent machine learning techniques (Xu et al. 2014; Gasch et al. 2015).

Kriging has become a very popular interpolation method for temperature and other meteorological variables due to its ability to take into account spatial correlation, to estimate target variables at unobserved locations, and to quantify the uncertainty associated with the estimator. Courault and Monestiez (1999) used OK to interpolate maximum and minimum temperatures at a 1 km spatial resolution for southeast France with the RMSE of 1–2 °C. Afterwards, RK was introduced, and it was proven that it gives better results than OK (Hunter and Meentemeyer 2005; Carrera-Hernández

[3]http://prism.oregonstate.edu/normals/

and Gaskin 2007) or any other interpolation methods like distance criteria, regressions, and splines (Hofstra et al. 2008). Haylock et al. (2008) interpolated, amongst other variables, the mean surface temperature for Europe at 25 km spatial resolution (E-OBS) by using kriging with an external drift on anomalies from monthly averages. Perčec Tadić (2010) made 20 climatological (climatological normals) maps for Croatia for the 1961–1990 period at a resolution of 1 km using regression kriging. Frick et al. (2014) provided gridded daily datasets of surface air temperatures for Germany at a 5 km spatial resolution using RK. Brinckmann et al. (2016) and Berezowski et al. (2016) interpolated the daily anomalies from the monthly averages using simple kriging for minimum and maximum temperatures for Europe. Spatio-temporal regression kriging (STRK) has recently become popular due to the development of `gstat` (Pebesma 2004; Graler et al. 2017) and `spacetime` (Pebesma 2012; Bivand et al. 2013a) packages in R. Graler et al. (2017) added extensions to the R package `gstat` for handling data formats from the R package `spacetime`, spatio-temporal semivariogram modeling, and spatiotemporal interpolation. Hengl et al. (2012) used STRK for interpolation of daily temperatures for Croatia and Kilibarda et al. (2014) for global land areas.

Nowadays, machine learning methods are becoming popular because they are easy to use and have decent accuracy performances. One of the reasons why ML methods were not used in this study is that they cannot be easily explained (black-box approach). Even though there are some initiatives to establish a framework for spatio-temporal interpolation using ML (Hengl et al. 2018), the accuracy is still lower in comparison with RK. As opposed to an ML approach, the use of STRK spatial and temporal correlations can be recognized and explained through semivariograms.

The first objective of this research is to examine the performance of the existing global STRK model (STRK_global, Kilibarda et al. 2014) over Croatia using an independent station dataset from dense Croatian national meteorological observing network. The second objective is to develop more accurate local model for mean daily temperature for Croatia based on smaller number of covariates (without MODIS LST) with respect to already existing model, i.e., Hengl et al. (2012). Finally, validation results of the developed local model will be compared and discussed in relation to the (1) existing STRK_global without high density station dataset from Croatia and (2) existing local model relying on MODIS data as covariate.

4.2 Study area and datasets

4.2.1 Study area

Although Croatia is a medium sized European country, the diverse topography, openness toward the Pannonian Plain and position on the eastern Adriatic coast characterizes the country with three main climatic regions: continental, mountainous, and maritime (Zaninović et al. 2008). This climate diversity has inspired the testing of different climatic (Antonić et al. 2001; Hengl et al. 2012) or physical models, where the research of the strong and gusty bora wind is amongst the most interesting examples (Bajić 1989; Belušić and Bencetić Klaić 2004; Horvath et al. 2009; Ivatek-Sahdan and Ivancan-Picek 2006). The diversity of climate conditions is explored and mapped in detail for the most recent standard climate normal 1961–1990 as reported in the climate atlas of Croatia (Zaninović et al. 2008), where the large range of values of different temperature parameters are presented, amongst which are the mean monthly and annual temperature, annual number of frosty, warm, and days with summer nights. In the recent decade, the observed climate change in the region is especially supported by pronounced warming and extended dry periods (Cindrić et al. 2010), which emphasize the need for spatio-temporal interpolation of temperature on fine spatial and daily temporal scale. The maps produced by these studies can serve as data sources for climate assessment and monitoring.

4.2.2 Datasets

Two different sets of the measurements from meteorological stations were used, namely GSOD (Section 3.2.2) and Croatian mean daily temperature dataset (CMDT). In addition, DEM and TWI were used as static covariates.

There are 48 GSOD stations in Croatia for the year 2008 (Figure 4.1, blue circles). GSOD was used in this study because it provides more measurements of the mean daily temperature than other open datasets (e.g., GHCN-daily, ECA&D), and it allows the prediction of mean air temperature with the STRK_global model to be independent with respect to CMDT. For the purpose of this study, the mean temperatures were converted from °F to °C.

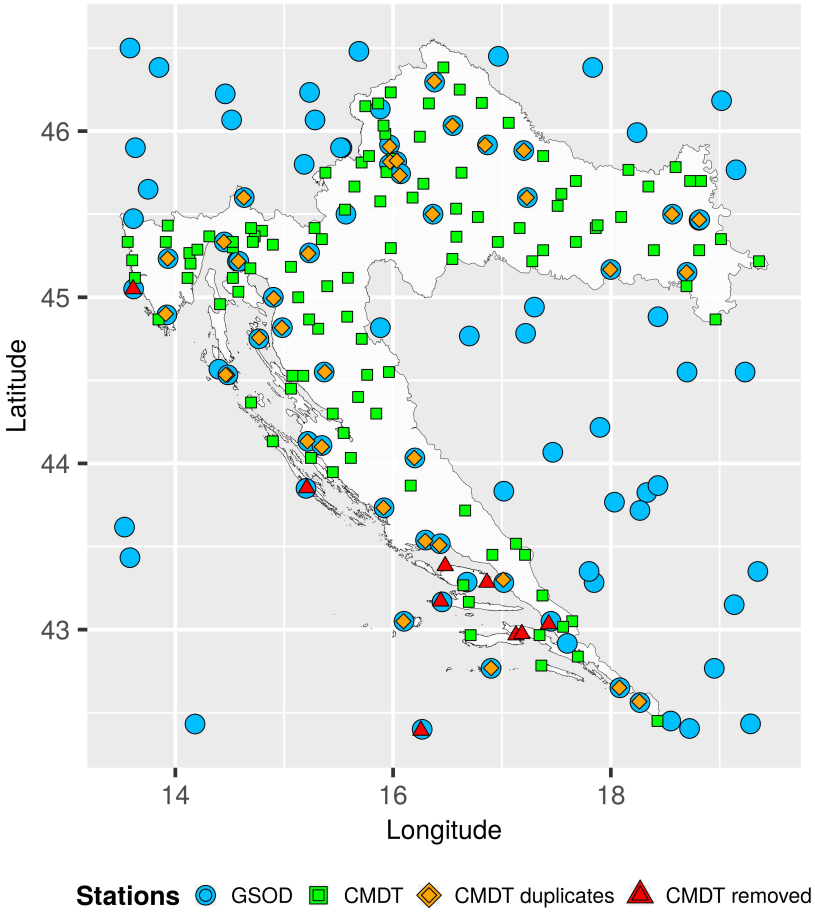


Figure 4.1: Spatial distribution of GSOD (blue circles) and CMDT (green squares) meteorological stations for mean daily temperature, CMDT stations which are included in GSOD dataset (orange diamond), and CMDT stations with missing DEM and TWI values (red triangles).

CMDT[4] provides data from 159 stations in Croatia (Figure 4.1, green squares). Furthermore, there are 57,282 measurements of the daily mean temperature available for the year 2008. A detailed description of this dataset is given by Hengl et al. (2012). The daily mean temperature is calculated as a weighted average of measurements taken at 07, 14, and 21 UTC. The precision of the measurements is 0.1 ∘C which is comparable with the GSOD dataset.

DEMSRE3 and TWISRE3 (DEM and TWI) at a 1 km spatial resolution, described in Section 3.3.4, were used as environmental covariates and are presented in Figure 4.2.

GSOD and CMDT datasets were stored in R STFDF objects (space-time full data frame) (Pebesma 2012), which are appropriate space-time objects, because the data exist for nearly all of the days at all of the stations' locations. For each CMDT station, DEM derivatives were extracted and added as an attribute to STFDF. Not all 157 CMDT stations were used for accuracy assessment. The

[4]http://spatial-analyst.net/book/HRclim200

9 coastal stations were not used for accuracy assessment (Figure 4.1, red triangles) because static covariates DEM and TWI were missing for these stations due to a poorly defined coastline on 1 km DEMSRE3. Furthermore, the stations located in the vicinity of 2 km from the GSOD stations, 37 of them that were considered as duplicates, were not used to test the STRK_global predictions made by GSOD (Figure 4.1, orange diamonds).

4.3 Methods

4.3.1 Spatio-temporal regression-kriging

STRK is an extension of RK interpolation method, described in Section 2.4.1. Besides the spatial component, it considers the influence of the time component and the space-time interaction on a prediction, i.e., it replaces spatial RK with spatio-temporal RK. STRK is a suitable candidate for the modeling of mean daily temperatures because of its ability to describe the spatio-temporal variability of a certain variable. Following the STRK interpolation method, the mean temperature variable $Z(\mathbf{s}, t)$ that varies over space ($\mathbf{s}$) and time (t) can be decomposed as (Heuvelink and Griffith 2010):

$$Z(\mathbf{s}, t) = m(\mathbf{s}, t) + V(\mathbf{s}, t) \tag{4.1}$$

In previous equation (Eq. 4.1), m is a deterministic component of the variable (trend) and is modeled using MLR (Section 2.3.1.1):

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

where the β_i are regression coefficients estimated using ordinary least squares and β_0 is model intercept (by imposing f_0 is equal to 1), the f_i are covariates that are known over the spatio-temporal domain, and p is the number of covariates (Eq. 4.2). $V(\mathbf{s}, t)$ is a zero-mean spatio-temporal stochastic residual and is modeled using a spatio-temporal sum-metric semivariogram (Graler et al. 2017):

V is a zero-mean spatio-temporal stochastic residual and is modeled using a spatio-temporal

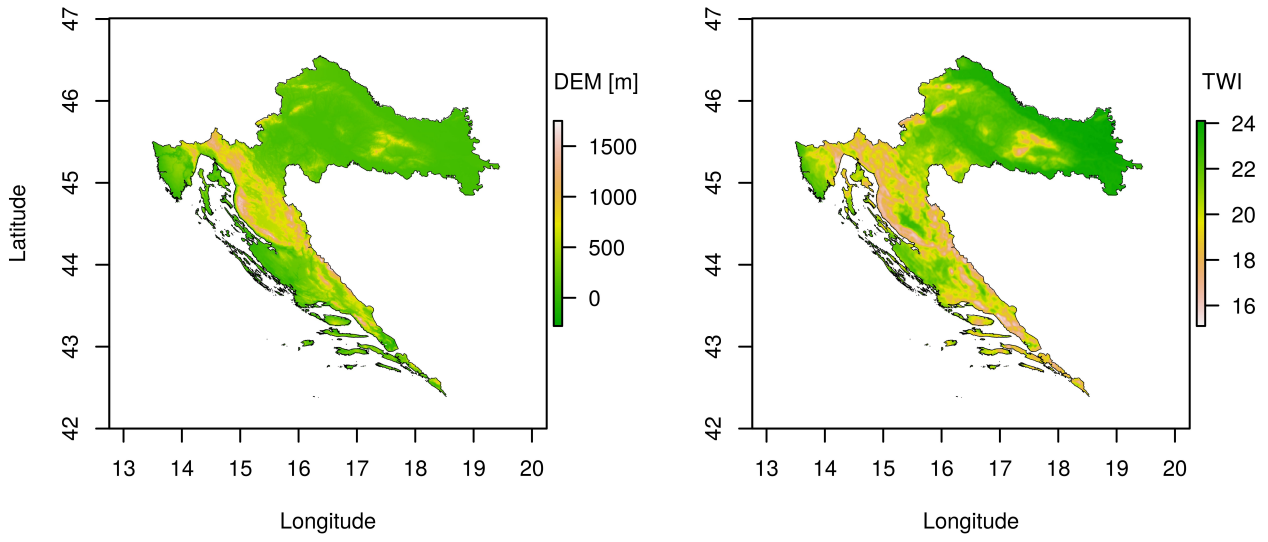


Figure 4.2: DEM (left) and TWI (right) values for Croatia.

sum-metric semivariogram (Graler et al. 2017):

$$\gamma(\mathbf{h}, u) = \gamma_S(\mathbf{h}) + \gamma_T(u) + \gamma_{ST}(\sqrt{\mathbf{h}^2 + (\alpha \cdot u)^2}) \tag{4.3}$$

where $\gamma(\mathbf{h}, u)$ denotes the semivariance of residuals at $\mathbf{h}$ units of a distance in space and u units of a distance in time, γ_S, γ_T are purely spatial and temporal components, γ_{ST} is the space-time interaction component and α is a spatio-temporal anisotropy ratio which converts units of temporal separation (u) into spatial distances ($\mathbf{h}$) (Eq.4.3).

4.3.2 Mean daily temperature model for global land areas

The mean daily temperature gridded dataset for Croatia at a 1 km spatial resolution was produced using the STRK_global implemented in the R package `meteo` (Kilibarda et al. 2014) and GSOD stations. The `tmeanGSODECAD_noMODIS` model from the `tregcoef` data of the R package `meteo` were used for trend estimation. Then the `tmeanGSODECAD_noMODIS` fitted semivariogram from the `tvgms` data of the R package `meteo` was used for residual prediction. An up-to-date version of R package `meteo` is available for download at https://r-forge.r-project.org/projects/meteo/.

Geometric temperature trend (GTT), DEM, and TWI are covariate layers used for MLR. The only dynamic covariate layer is GTT proposed by Kilibarda et al. (2014). GTT is a function of latitude (ϕ) and the day of the year (day). GTT is defined with the following function (Eqs. 4.4 and 4.5) for the mean daily temperature:

$$GTT = 30.4 \cos\phi - 15.5(1 - \cos\theta) \sin|\phi| \tag{4.4}$$

where θ is:

$$\theta = (day - 18)\frac{2\pi}{365} + 2^{1-sgn(\phi)}\pi \tag{4.5}$$

The original model for global land areas uses MODIS LST as a covariate (Kilibarda et al. 2014). However, the MODIS LST daily images have spatial gaps while MODIS LST 8-day images have temporal gaps due to cloud contamination, so those gaps need to be filled. Since the idea of our research was to develop a simple, accurate, and fast model for mean daily air temperature estimation, MODIS LST data are omitted. The STRK_global model is explained in detail by Kilibarda et al. (2014).

4.3.3 Mean daily temperature model for Croatia

In order to make a better estimation of the mean daily temperature for Croatia at a 1 km resolution, an adaptation of the presented STRK_global for mean daily temperature was made. The STRK_Croatia was developed using the data from CMDT. This dataset contains observations from more than 150 stations, which is about three times the amount compared with 48 GSOD stations used for the making of the STRK_global (Kilibarda et al. 2014). A trend model was made using the same covariates applied in the STRK_global: GTT, DEM, and TWI. Consequently, a spatio-temporal sum-metric semivariogram was made for the residuals calculated at the stations locations.

The trend modeling, estimation of the sample semivariogram, and fitting of the spatio-temporal semivariogram are performed in the R software (R Development Core Team 2012) using the `lm` base function and the `vgmST` and `fit.StVariogram` functions from the R package `gstat`. The code is available at http://osgl.grf.bg.ac.rs/materials/tac_hr/. The STRK_Croatia is now available in R package `meteo`, i.e., trend in `tregcoef` data and spatio-temporal semivariogram in `tvgms` data named `hr`. It was used to produce a local mean daily temperature gridded dataset for Croatia

for year 2008.

4.3.4 Accuracy assessment

The accuracy of the STRK_global and STRK_Croatia was assessed by leave-one-out (LOO) and stratified 5-fold cross-validation. Before that STRK_global predictions made using GSOD stations were compared with the CMDT data. Five stratified folds were created using modified `stratfold3d` function of the R package `sparsereg3D`[5] (Pejović et al. 2018). This function creates stratified folds in three steps:

1. Stations are clustered using k-means clustering according to spatial location,
2. Each cluster is split to folds, stratified according to the altitude of the station,
3. Each final fold is obtained by merging one fold from each cluster.

Each of the folds was used once for cross-validation based on the data from the other four folds. This method was chosen because temperature observations in CMDT, as well as in GSOD, were not represented well enough in the areas at higher altitudes. Rather than LOO cross-validation, stratified cross-validation was used for two reasons. (1) It is better in terms of bias and variance (Kohavi 1995), and (2) it has an ability to separate data in such a way that each fold of the data is a representative sample of the whole dataset with regard to altitude and spatial distribution of the stations.

The coefficient of determination (R^2) and root mean squared error were calculated as performance measures for both (STRK_global and STRK_Croatia) examined models. Also, the annual average RMSE per test or cross-validated station was calculated in order to find a cause of the worst results which occur at some stations. All of the figures used to present annual average RMSE per station are available as interactive maps at http://osgl.grf.bg.ac.rs/materials/tac_hr/. These maps were produced by R package `plotGoogleMaps` (Kilibarda and Bajat 2012).

4.4 Results

4.4.1 Mean daily temperature model for global land areas and prediction

The STRK_global was already implemented in the R package `meteo` (`tmeanGSODECAD_noMODIS` model, Kilibarda et al. 2014), so predictions were made for the limited area of Croatia. The spatio-temporal trend model for STRK_global is given by Eq. 4.6:

$$trend = -2.44 + 1.02 \cdot GTT + 0.0004 \cdot DEM - 0.025 \cdot TWI \tag{4.6}$$

The parameters of the fitted sum-metric semivariogram are shown in the Table 4.1.

Kilibarda et al. (2015) found that GTT by itself explains 75% of mean daily temperature variations with a standard error of $\pm$ 5.7 °C which makes GTT the most important covariate of the model. Furthermore, they concluded that there is no pure temporal correlation between the residuals and also that the temporal correlation is caught by the spatio-temporal component.

Mean daily temperatures at a 1 km spatial resolution for Croatia for the year 2008 were estimated using above described STRK_global model and both GSOD and CMDT dataset. These datasets are

[5] https://github.com/pejovic/sparsereg3D

Table 4.1: Sum-metric semivariogram parameters for the STRK_global (Kilibarda et al. 2014).

Component	Nugget [°C^2]	Sill [°C^2]	Range	Function	Anisotropy ratio
Spatial	2.24	30.55	5130 km	Spherical	n/a
Temporal	0.00	0.00	0.1 days	Spherical	n/a
Spatio-temporal	0.59	9.74	2242 km	Spherical	501 km/day

available at http://osgl.grf.bg.ac.rs/materials/tac_hr/ in GeoTIFF format. Predictions for each pixel are made using the 30 nearest GSOD or CMDT stations, and observations at these stations are not only for a specific day, but also for the day before.

4.4.2 Mean daily temperature model calibration for Croatia and prediction

The estimated spatio-temporal trend for the STRK_Croatia is defined as:

$$trend = 18.73 + 0.86 \cdot GTT + 0.0092 \cdot DEM - 0.606 \cdot TWI \tag{4.7}$$

This trend explains about 80% of the variation of the mean daily temperature with RMSE = 3.5 °C, and GTT by itself explains 74% of the mean daily temperature variation with a standard error of ± 4 °C.

In Figure 4.3 the scatterplot of observations and predictions is presented (left). Residuals from the trend are normally distributed allowing for the kriging interpolation (Figure 4.3, right). In Figure 4.4 sample semivariogram and fitted sum-metric semivariogram are presented. The sample semivariogram shows that there is obviously a spatio-temporal correlation between the residuals and on account of this spatio-temporal kriging that is applicable. The parameters of the fitted sum-metric semivariogram are shown in Table 4.2.

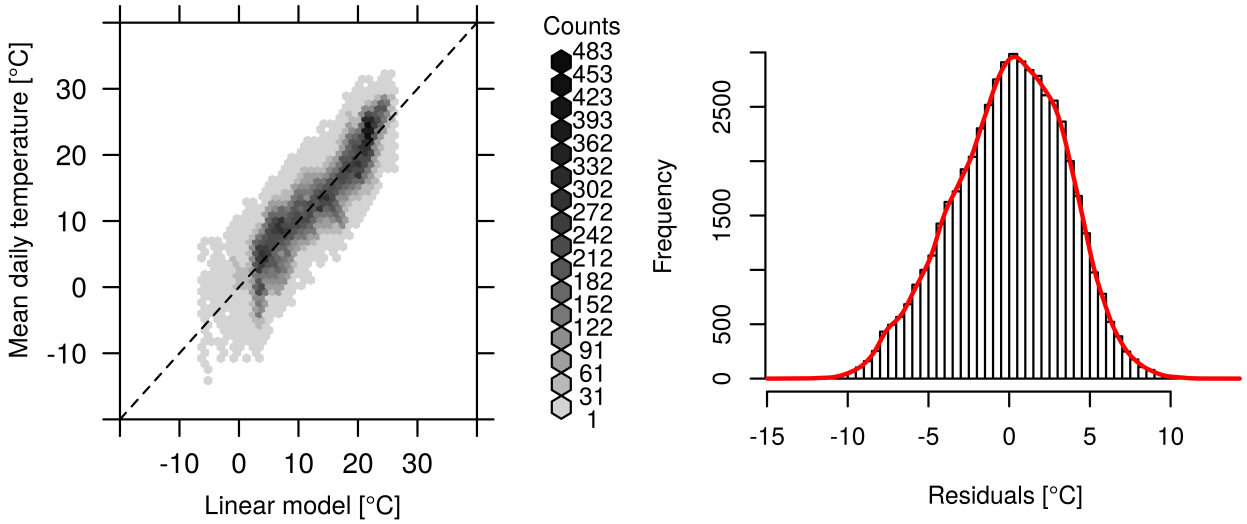


Figure 4.3: The scatterplot of estimated mean daily temperature values from the trend for STRK_Croatia vs. observed values, (left). Histogram of the residuals from the trend for STRK_Croatia (right). It shows that residuals follow the normal distribution which justifies the use of the kriging.

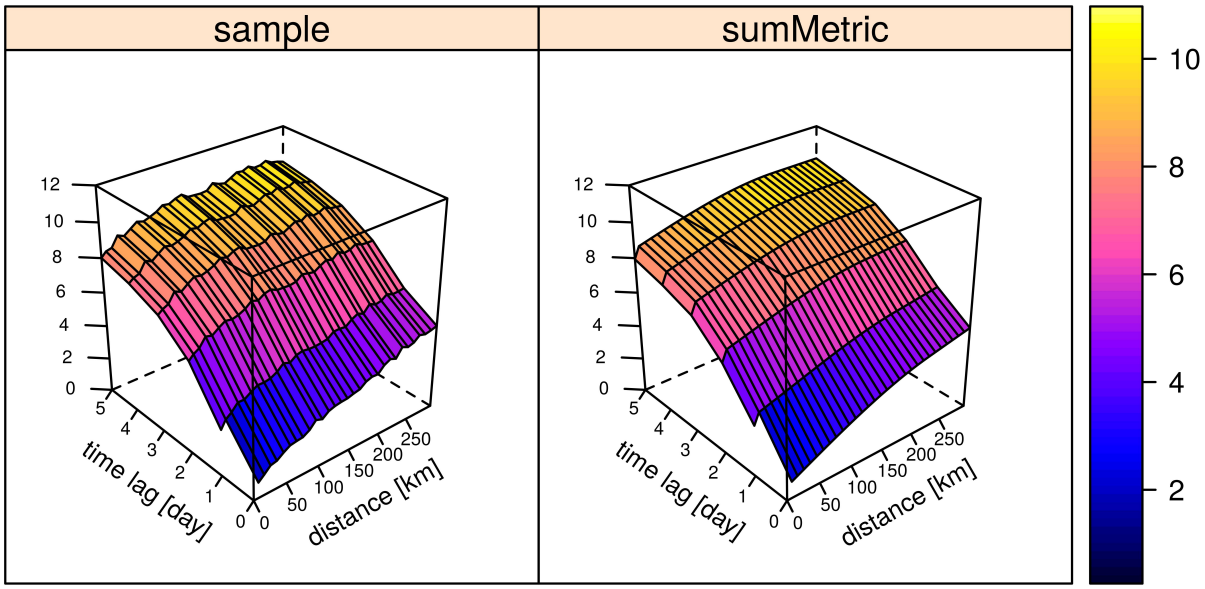


Figure 4.4: Sample semivariogram (left) and fitted sum-metric semivariogram (right) of the residuals from the trend model for STRK_Croatia. Semivariograms are presented in 3D.

Table 4.2: Sum-metric semivariogram parameters for the STRK_Croatia.

Component	Nugget [°C²]	Sill [°C²]	Range	Function	Anisotropy ratio
Spatial	0.56	1.61	221 km	Spherical	n/a
Temporal	0.00	3.60	7.4 days	Spherical	n/a
Spatio-temporal	0.27	4.58	830 km	Spherical	248 km/day

Mean daily temperatures at a 1 km spatial resolution for Croatia for the year 2008 were estimated using above described STRK_Croatia and CMDT dataset. They are also available at http://osgl.grf.bg.ac.rs/materials/tac_hr/ in GeoTIFF format. Predictions for each pixel were made using the 30 nearest CMDT stations and observations from them for a specific day and previous 6 days as it could be inferred from the range in the temporal component of the sum-metric semivariogram (Table 4.2).

The STRK_Croatia is also added to the R package `meteo`.

4.4.3 Accuracy assessment

STRK_global predictions based on GSOD stations was tested with CMDT. It is important to emphasize that these 111 stations from the CMDT were not used in the making of the STRK_global. The R^2 of the test is 92.9% and RMSE is 2.1 °C. The annual average RMSEs per station are presented in the Figure 4.5. The test shows that R^2 is about 4% lower than for cross-validation (96.6%, Kilibarda et al. 2014), and RMSE is in a range of the result for the cross-validation for the whole world (2.4 °C, Kilibarda et al. 2014) and averaged for Croatia (2 °C, http://dailymeteo.org/node/3 — not active anymore). These results are explainable by larger number of stations used by Kilibarda et al. (2014) since they merged ECA&D dataset with GSOD.

The LOO cross-validation was performed for STRK_global and STRK_Croatia models with 148

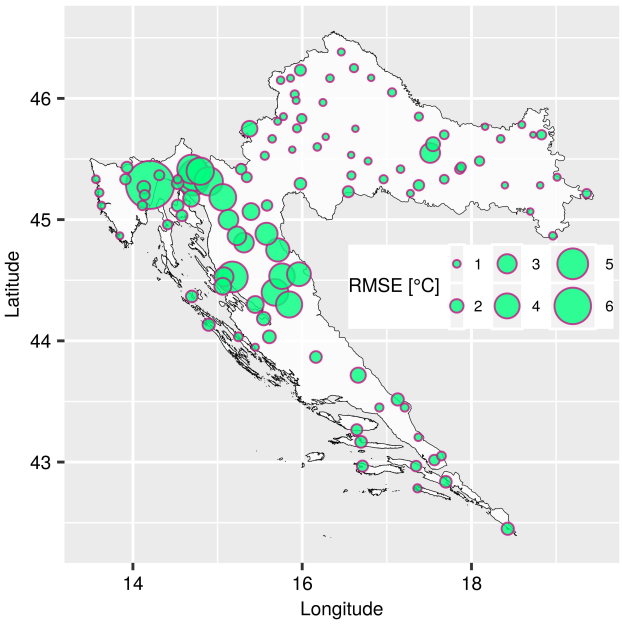


Figure 4.5: Annual average RMSE per station for testing of STRK_global predictions made by using GSOD stations (http://osgl.grf.bg.ac.rs/materials/tac_hr/). RMSE values are presented by the radius of the circles.

CMDT stations (nine stations without DEM and TWI were excluded). The R^2 of STRK_global and STRK_Croatia equals 94.4% and 97.8%, respectively, while the RMSE equals 1.9 °C and 1.2 °C, respectively. The annual average RMSE per station is presented in the Figure 4.6. For the STRK_Croatia, three stations at altitudes higher than 1000 m got the highest RMSEs and they are around 3 °C (Figure 4.7). All the other stations at altitudes lower than 1000 m got an RMSE less than 2.5 °C.

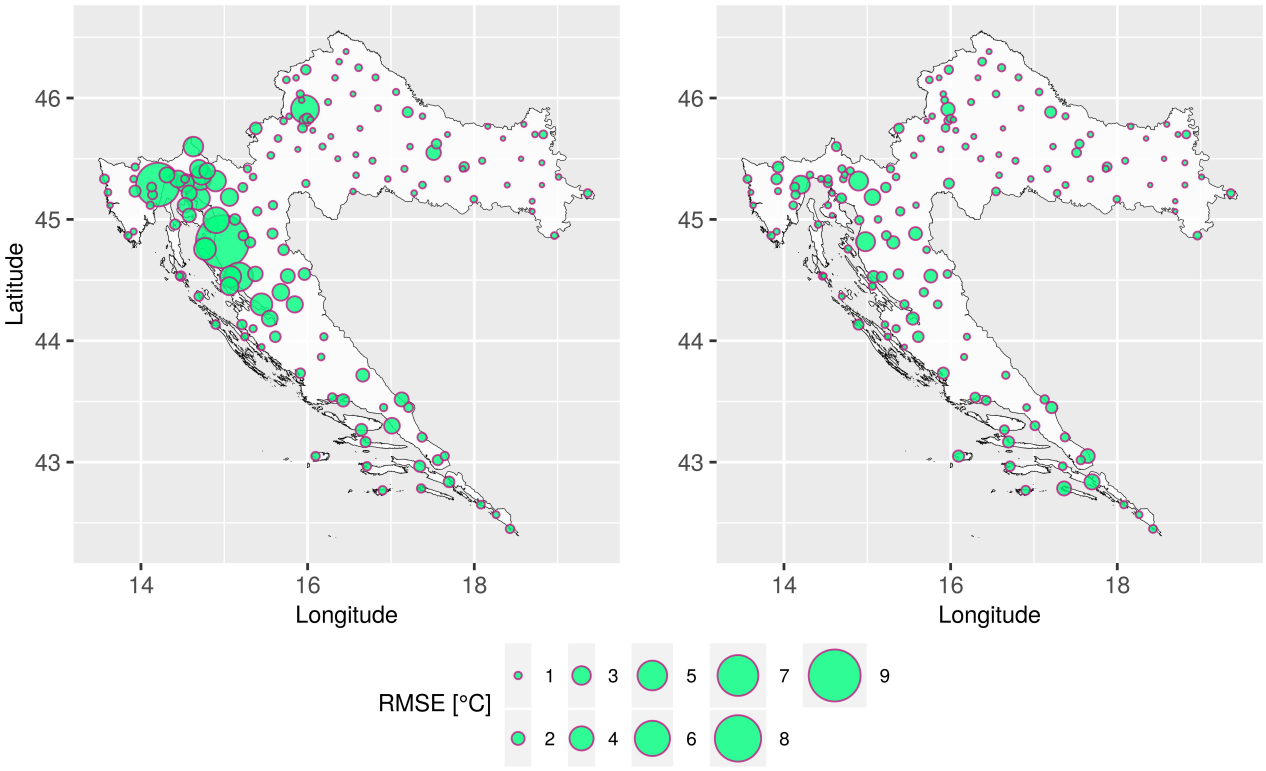


Figure 4.6: Annual average RMSE per station. Results of LOO cross-validation, STRK_global on the left and STRK_Croatia on the right (http://osgl.grf.bg.ac.rs/materials/tac_hr/). RMSE values are presented by the radius of the circles.

The 5-fold stratification folds from 148 CMDT stations are shown in Figure 4.8. Each of the folds

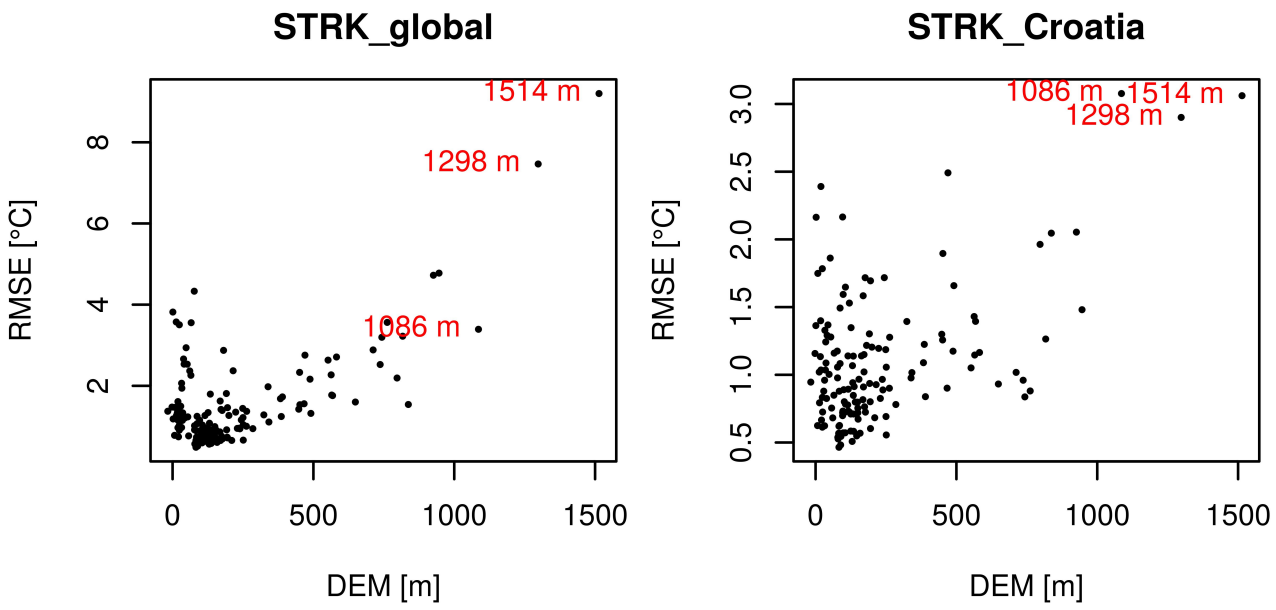
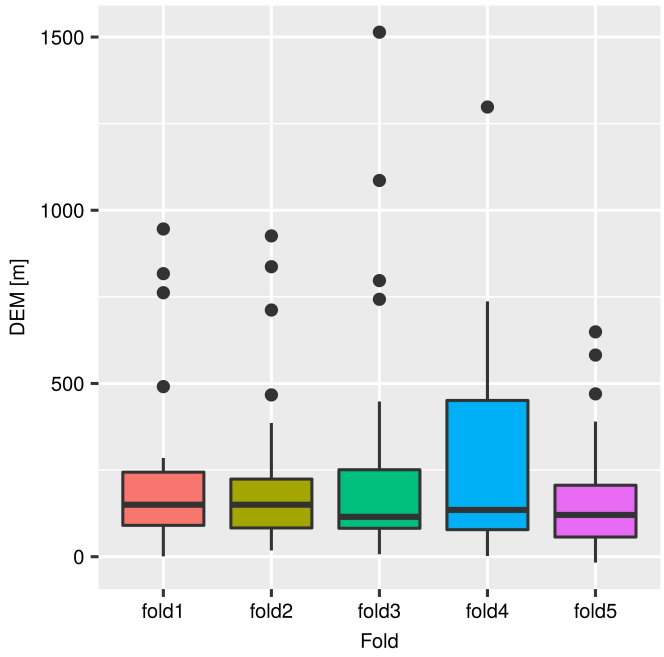


Figure 4.7: Scatter plot DEM vs annual average RMSE from LOO cross-validation, STRK_global on the left and STRK_Croatia on the right. Stations at altitudes above 1000 m (red) in the top right corner have the highest RMSEs. Notice the smaller scale on the y-axis for the STRK_Croatia model.

is a representative sample of the entire dataset considering the elevations and spatial distribution of the stations and considering that the median and mean of the folds do not differ more than 10 to 20 m from the median and mean of the whole dataset. The spatial distribution of the stations per fold is presented in the Figure 4.9.

Figure 4.8: Boxplot of the altitude per fold.



The results from the stratified 5-fold cross-validation show that the STRK_global explains about 95.3% of the variation with 1.7 °C RMSE, while proposed STRK_Croatia explains about 98.2% of the variation with 1.1 °C RMSE. These results are in agreement with the LOO cross-validation. The RMSE per station are presented in the Figure 4.9 with different color coding for each fold.

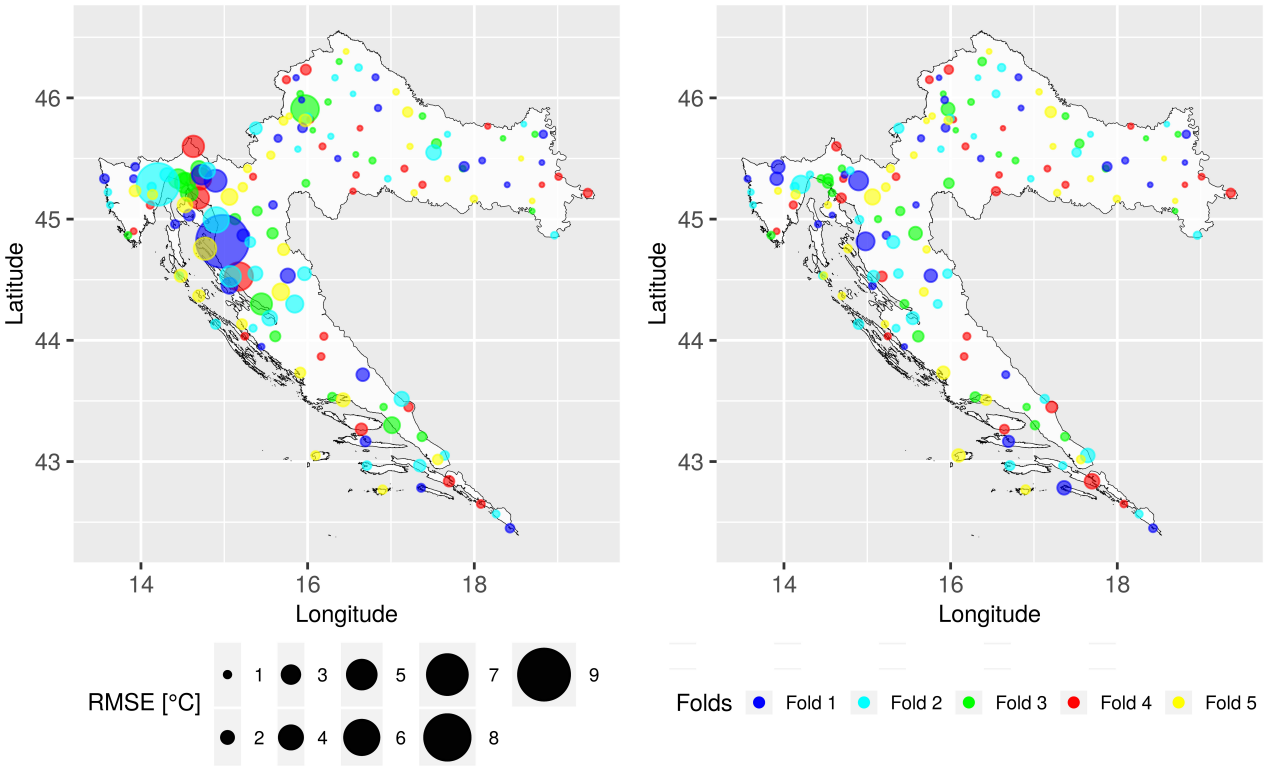


Figure 4.9: Annual average RMSE per station for 5-fold cross-validation, STRK_global on the left and STRK_Croatia on the right (available at http://osgl.grf.bg.ac.rs/materials/tac_hr/). RMSE values are presented by the radius of the circles.

The accuracy per month for both models was also assessed (Table 4.3). The STRK_global does not show noticeable seasonal differences in average monthly RMSEs. On the other hand, STRK_Croatia shows noticeably larger RMSEs in cold season with the largest RMSE in January and smaller RMSEs in warm season with the smallest RMSE in April. The improvements in changing from a global to local model are also larger in warm season.

Table 4.3: RMSE values [°C] for each month for STRK_global and STRK_Croatia obtained by LOO and 5-fold cross-validation and differences between them.

Model	**STRK_global**		**STRK_Croatia**		**Difference**	
Cross-validation	LOO	5-fold	LOO	5-fold	LOO	5-fold
January	1.87	1.91	**1.51**	**1.52**	**0.36**	**0.40**
February	1.84	1.86	1.38	1.38	0.46	0.48
March	1.89	1.91	1.03	1.04	0.86	0.87
April	1.84	1.85	**0.93**	**1.95**	**0.91**	**0.90**
May	1.80	1.81	1.02	1.03	0.78	0.78
June	1.78	1.80	1.00	1.00	0.78	0.80
July	1.88	1.88	1.04	1.04	0.84	0.84
August	1.87	1.89	1.14	1.16	0.73	0.73
September	**1.92**	**1.93**	1.08	1.09	0.84	0.84
October	**1.77**	**1.79**	1.21	1.22	0.56	0.57
November	1.87	1.90	1.23	1.26	0.64	0.64
December	1.91	1.94	1.13	1.16	0.78	0.78

4.5 Discussion

4.5.1 Global vs local model

Besides GTT, DEM and TWI proved to be significant covariates in the STRK_Croatia trend model (Eq. 4.7). They have a larger influence in the prediction of the STRK_Croatia compared with the one for STRK_global as is supported by the t test for the significance of the regression coefficients. The significant difference between the trend of the STRK_Croatia and the STRK_global is in the intercept value. The aforementioned value is 18.73 °C for Croatia, which is significantly higher compared with −2.43 °C for the STRK_global. This can be explained by the mean annual temperature that is higher for Croatia (around 13 °C) than for the entire world (around 1 °C).

A striking difference between the STRK_Croatia and STRK_global fitted semivariograms is that a temporal semivariogram component appears in the STRK_Croatia. This means that there is a pure temporal correlation between the data in the range of 7 days. Nugget effects from the spatial and spatio-temporal components indicate that the short-range variability is 0.3 °C. This, in turn, shows that there is room for model improvement because the precision of the measurements being 0.1 °C is declared, which suggests that the stations with lower precision are presented in the data or are themselves potential outliers in data. As expected, the ranges and the anisotropy ratio (excluding the temporal component) for the STRK_Croatia are lower than for the STRK_global due to the higher density of stations and smaller spatial extent.

Accuracy assessment shows that the STRK_Croatia, which is an adaptation of the STRK_global, significantly improves interpolation accuracy by 3.3% in R^2 and 0.6 °C in RMSE. When comparing accuracies per month (Table 4.3), the STRK_Croatia performs better than STRK_global in each month and the improvements are larger in warm parts of the year. The largest improvement is for April, from 1.84 °C to 0.93 °C and the smallest for January, from 1.87 °C to 1.51 °C RMSE. The average monthly RMSEs of the STRK_Croatia, for those that are larger in the cold season compared with the warm one, indicate that there are still some influences that modifies winter temperatures (like e.g., cold air pool and temperature inversions) that cannot be explained by the model. Similar conclusions were obtained in Hiebl et al. (2009) and in Perčec Tadić (2010) when comparing monthly normals and in Hiebl and Frei (2016) when comparing daily minimum and daily maximum temperatures. In these papers, the cold months/season had prediction errors that were larger than in warm months/season. The adjustment of the global model and a benefit of the larger observations density become obvious if we take a look at predictions in Figure 4.10. The STRK_Croatia model shows a more pronounced spatial variability, especially in the mountainous regions. On the other hand, the STRK_global smooths the prediction because it was trained on the sparser station network for the whole world (Kilibarda et al. 2014). Further on, the spatial range of 221 km for the STRK_Croatia semivariogram is much shorter than 5130 km for STRK_global. Also, the spatial nugget of 2.24 °C for the STRK_Croatia semivariogram is much larger than 0.56 °C for STRK_global. This results in a loss of local variability and accuracy in the STRK_global. For the STRK_global, the highest errors occur in the western part of Croatia and near the coastline (Figures 4.7 and 4.10) because it represents a mountainous region (Figure 4.2). The STRK_Croatia managed to reduce errors not only in that region but for the whole area of Croatia. However, the error in the mountainous region is still higher compared with the other parts of Croatia. Figure 4.11 shows time series of predictions from LOO cross-validation and observations for Zavižan (1514 m) and Zagreb-Maksimir (121 m) stations. It can be noticed that both STRK_global and STRK_Croatia predict mean daily temperature with high accuracy at lower altitude (Figure 4.11, Zagreb-Maksimir), which confirms claim by Kilibarda et al. (2014) that STRK_global performs better for areas at lower altitude. STRK_Croatia predictions are much closer to observations with slight underestimation while STRK_global mostly overestimates

mean daily temperature for stations at higher altitude (Figure 4.11, Zavižan). This improvement at higher altitude was expected because STRK_global models variability of mean daily temperature for the whole world, while STRK_Croatia tends to explain variability just for Croatia. It looks like that STRK_Croatia predictions are STRK_global predictions shifted to observations values (Figure 4.11, Zavižan) with small adjustments. This shift is a consequence of the shift in trend models, i.e., trend for STRK_Croatia performs better than trend for STRK_global. Another reason is that residuals for STRK_global and STRK_Croatia follow the same spatio-temporal patterns, even though residuals from STRK_global are larger than STRK_Croatia (Figure 4.10).

4.5.2 Mean daily temperature model for Croatia and comparison with other models

The mean temperature for Croatia has already been modeled by regression-kriging and the same dataset of CMDT stations in a previous study (Hengl et al. 2012). Latitude, longitude, DEM, topographically weighted distance from the coastline, and TWI were used as static, and DEM-derived total potential insolation (INSOL) and MODIS LST images as dynamic covariates in that study. Hengl et al. (2012) explained 86% of variation with 3.4 °C RMSE by MLR using these covariates, which is slightly better compared with 80% of the variation with 3.5 °C RMSE for the STRK_Croatia. This result may be explained by the larger number of covariates used and the effect of dynamic predictors in a model. Consequently, the fitted semivariograms are also different. The STRK_Croatia fitted semivariogram has lower nuggets, sills, and ranges, and the spatio-temporal component is also more significant. However, the overall accuracy is improved by 1.2 °C in RMSE and 7% in R^2 (RMSE = 2.4 °C and R^2 = 91% in Hengl et al. 2012), even though the MODIS LST images were omitted. The trend model proposed by Hengl et al. (2012), which includes MODIS, already explained a lot of spatial patterns and there was not much spatio-temporal relation left for SK to model. On the other hand, the simple STRK_Croatia trend model performed slightly worse but the fitted semivariogram explained more spatio-temporal variation. GTT explains a lot of temperature variation, which is comparable with MODIS LST. However, GTT obviously leaves a stronger spatio-temporal relation between residuals that can be explained by kriging.

When comparing the accuracy of other local (country) models at 1 km spatial resolution, STRK_Croatia performs similar or even better than some of them. Frei (2014) interpolated daily temperature at 1 km spatial resolution for Switzerland (European Alps) using nonlinear profiles and non-Euclidean distances and Rosenfeld et al. (2017) applied linear mixed effect models (3-step model with MODIS LST) for Israel. They both achieved RMSE of around 1 °C and R^2 of around 97%. One must keep in mind that both Switzerland and Israel cover a smaller area (around 41,000 and 21,000 km^2 , respectively) than Croatia (57,000 km^2), while the number of stations used for model development were comparable or larger (100 and 239, respectively). Nonetheless, the results of the STRK_Croatia are in the range of this accuracy. Huang et al. (2015) used linear regression models with MODIS LST as a covariate for central China's Shaanxi Province, and Janatian et al. (2017) used a similar method with 11 more covariates in the eastern region of Iran. The accuracies of these two models (RMSE ranged from 2.5–3.5 °C and R^2 was around 90%) are lower compared with STRK_Croatia because of a much larger area of these two countries and due to the fact that only around 20 stations were used for model development. Extensive research is available that interpolates daily minimum and maximum temperatures at 1 km spatial resolution for different areas. For example, Jarvis and Stuart (2001) performed the analysis for England and Wales, Zhu et al. (2013) for Xiangride River basin in the north Tibetan Plateau, Parmentier et al. (2014, 2015) for the state of OR, USA, Oyler et al. (2015) and Li et al. (2018a) for the conterminous USA, Hiebl and Frei (2016) for Austria. RMSE values were around 1–3 °C and R^2 did not exceed 97%. Some of the models performed even better (RMSE below 1 °C) but the reason for this was due to a larger number of stations

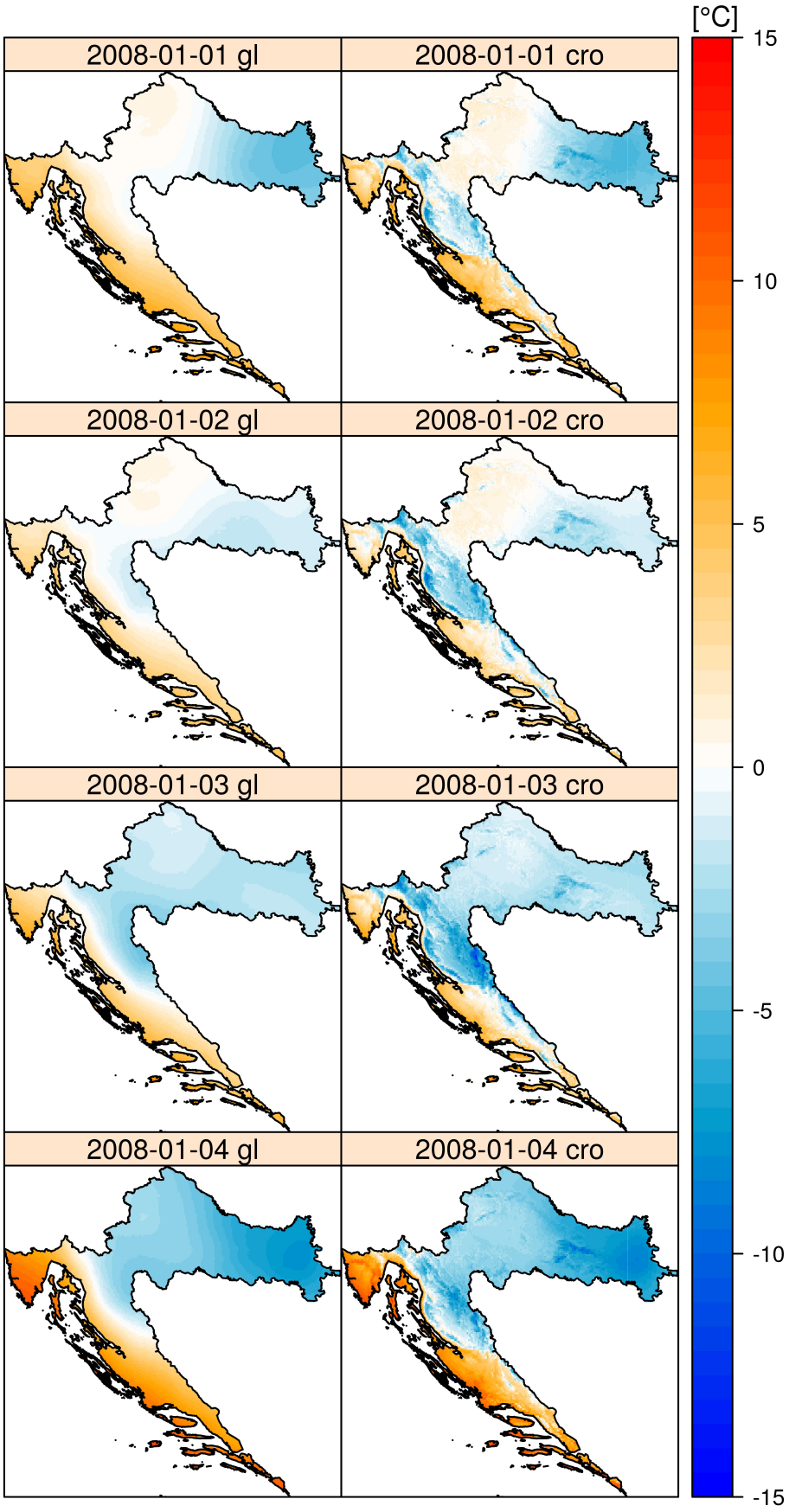


Figure 4.10: Maps of predicted mean daily temperatures at 1 km spatial resolution using STRK_global on the left and STRK_Croatia on the right with CMDT stations for the first 4 days of January 2018 for Croatia.

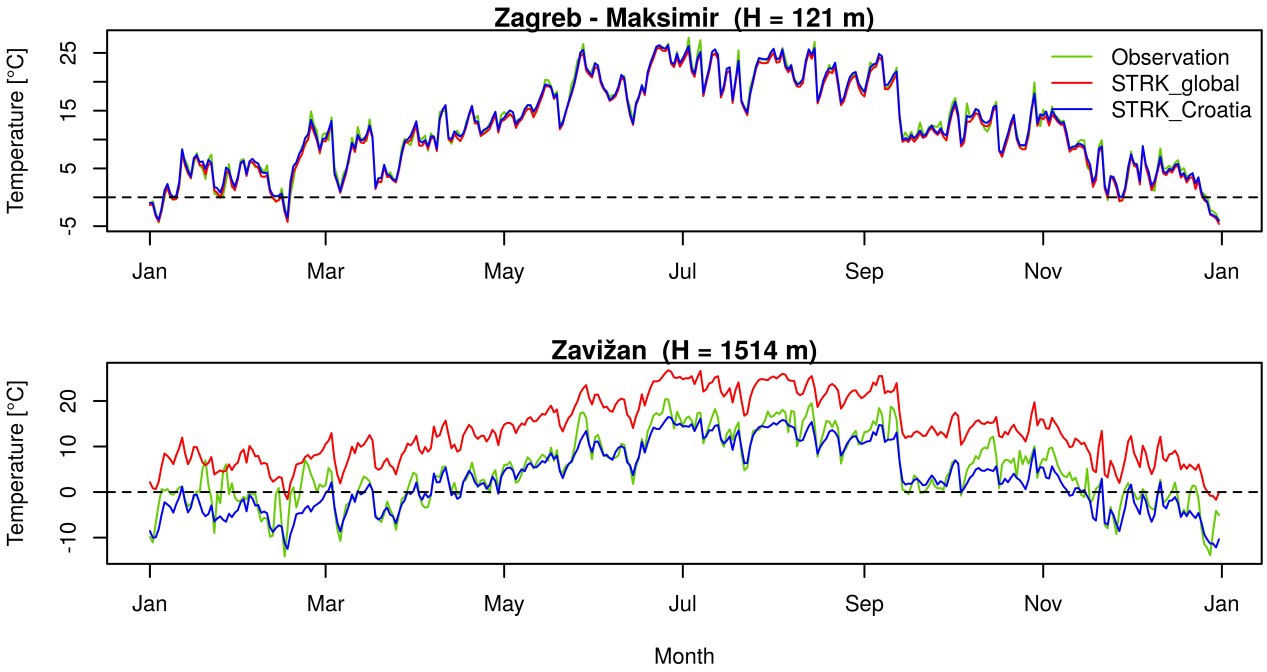


Figure 4.11: Time series of predictions from LOO cross-validation (red—STRK_global, blue—STRK_Croatia) and observations (green) for station Zavižan and station Zagreb-Maksimir.

available. Others modeled daily temperature at coarser spatial or temporal resolution. Benali et al. (2012) provided weekly 1 km mean temperature estimations for Portugal (RMSE was 1.33 °C and R^2 was 94.1%). Frick et al. (2014) provided 5 × 5 km gridded daily datasets of surface air temperature for Germany (RMSE was 1.39 °C and R^2 was 98.3%). Brinckmann et al. (2016) provided daily mean temperature dataset for Europe at 5 km spatial resolution (RMSE was 1–2 K and R^2 was 90%). Most of the above mentioned models are generally more complex or they use a large number of covariates including MODIS LST, which has a well-known problem with missing values and cloudiness. However, their accuracy is not better in comparison with STRK_Croatia. As a result, STRK_Croatia is recommended as a simple framework not only for mean but also for maximum and minimum temperature interpolation that can be applied to other countries or local areas.

There is still some room for model improvement in terms of mean daily temperature prediction at higher altitudes (specifically over 1000 m altitudes). Microclimate at higher altitudes is more complex. Also, insufficient number of stations and their distribution at higher altitudes do not cover temperature variability that could be explained by STRK (Kilibarda et al. 2015). Model underperformance and station deficiency problem at higher altitudes are also confirmed by Hengl et al. (2012). Many other publications point to the same problem. Perčec Tadić (2010) mapped monthly means of 20 climatological parameters, including the mean temperature, for the 1961–1990 period for Croatia with a resolution of 1 km. She proved that mapping accuracy is lower at higher altitudes due to the station deficiency problem. Dodson and Marks (1997) also confirmed that interpolation of the temperatures on higher altitudes will be biased toward temperature at lower elevations. Stahl et al. (2006) compared 12 interpolation methods for interpolating daily maximum and minimum temperatures over British Columbia , Canada, and the main conclusion was that the prediction was better with a denser distribution of stations on higher altitudes. Many others, like Krähenmann and Ahrens (2013) and Frei (2014) also have drawn the same conclusion. Benali et al. (2012) suggest that MODIS LST can improve accuracy in areas with low station density. MODIS LST could explain microclimatic conditions but the model will become more complex. The stations of neighboring countries are even more likely to improve the STRK_Croatia due to the fact that they have a significant impact

on the prediction for the areas near the Croatian border.

4.6 Conclusions

Considering both accuracy and model simplicity, the STRK_Croatia has proved to be a good solution for production of high resolution mean daily temperature grids for local areas. Compared with STRK_global (Kilibarda et al. 2014), the improvement was made in 3.4% in R^2 and 0.7 °C in RMSE. Suggested methodology uses only tree covariates, DEM, TWI, and GTT, and improves overall accuracy by 7% in R^2 and 1.2 °C in RMSE in comparison with the study (Hengl et al. 2012), where seven covariates, including MODIS LST images, were used. Having in mind that all covariates (DEM, TWI, and GTT) used in our study are available in real-time, the proposed STRK_Croatia can be used for obtaining real-time temperature grids, which is not the case with models based on MODIS LST images. Most of existing temperature models are generally more complex or they use large number of covariates that also include MODIS LST. However, in most cases, their accuracy is lower in comparison with STRK_Croatia. Nonetheless, accuracy assessment shows that the STRK_Croatia model still does not perform well enough for the prediction of mean daily temperatures at higher altitudes ($>$ 1000 m) by reporting similar errors as before with spatial or spatio-temporal interpolation methods for this area. Additional stations and measurements at higher altitudes and stations from countries around Croatia and MODIS LST could improve prediction accuracy at higher altitudes. This limitation makes the model most suitable for application on lower elevations such as in agriculture, health care, spatial planning, tourism, etc. Future research should focus on the enhancement of model prediction accuracy at higher altitudes. The proposed framework for the development of the STRK model could be applicable to any local area not only for mean but also for daily maximum and minimum temperature.

Chapter 5

Spatial and spatio-temporal interpolation using random forest[1]

For many decades, kriging and deterministic interpolation techniques, such as inverse distance weighting and nearest neighbour interpolation, have been the most popular spatial interpolation techniques. Kriging with external drift and regression kriging have become basic techniques that benefit both from spatial autocorrelation and covariate information. More recently, machine learning techniques, such as random forest and gradient boosting, have become increasingly popular and are now often used for spatial interpolation. Some attempts have been made to explicitly take the spatial component into account in machine learning, but so far, none of these approaches have taken the natural route of incorporating the nearest observations and their distances to the prediction location as covariates. The value of including observations at the nearest locations and their distances from the prediction location by introducing Random Forest Spatial Interpolation (RFSI) was explored in this research. RFSI was compared with deterministic interpolation methods, ordinary kriging, regression kriging, Random Forest and RFsp in three case studies. The first case study made use of synthetic data, i.e., simulations from normally distributed stationary random fields with a known semivariogram, for which ordinary kriging is known to be optimal. The second and third case studies evaluated the performance of the various interpolation methods using daily precipitation data for the 2016–2018 period in Catalonia, Spain, and mean daily temperature for the year 2008 in Croatia. Results of the synthetic case study showed that RFSI outperformed most simple deterministic interpolation techniques and had similar performance as inverse distance weighting and RFsp. As expected, kriging was the most accurate technique in the synthetic case study. In the precipitation and temperature case studies, RFSI mostly outperformed regression kriging, inverse distance weighting, random forest, and RFsp. Moreover, RFSI was substantially faster than RFsp, particularly when the training dataset was large and high-resolution prediction maps were made.

5.1 Introduction

Spatial and spatio-temporal interpolation of natural and socio-economic variables are important in many scientific fields. Some basic interpolation techniques are nearest neighbour (Thiessen 1911), inverse distance weighting (Willmott et al. 1985), and trend surface mapping (Chorley and Haggett 1965). In the 1980s, geostatistical interpolation (kriging) (Matheron 1963) was introduced. This turned out to be a major improvement because kriging takes into account spatial correlation and

[1]Based on article: Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random Forest Spatial Interpolation. Remote Sensing, 12(10), 1687. https://doi.org/10.3390/rs12101687 (Sekulić et al. 2020a)

quantifies the interpolation error through the kriging standard deviation. Kriging is the Best Linear Unbiased Predictor (BLUP) for spatial data under certain stationarity assumptions (Goovaerts 1997). It is also very flexible because there are many variants that can deal with specific cases, such as anisotropy, non-normality, and information contained in covariates (Diggle and Ribeiro 2007; Webster and Oliver 2007).

However, kriging also has disadvantages. It can be computationally demanding, makes many assumptions, and it may not be easy to come up with a sound geostatistical model that fits all types of data well (Hengl et al. 2018). It is also not well suited for incorporating the abundance of covariate information that is available nowadays. An important issue is that it is difficult to define a geostatistical model for data that cannot easily be transformed to normality. To solve this challenge, indicator kriging was developed (Journel 1983); however, it is cumbersome and not model-based (i.e., it does not use formal statistical methods derived for an explicit and complete statistical model, see Diggle and Ribeiro (2007). The Generalized Linear Geostatistical Model (Diggle and Ribeiro 2007) is statistically sound but still limited in the type of distributions it can handle, and in addition it is technically very complex. For example, it is far from obvious how variables with many zeroes and extreme values, such as in the case of precipitation, can be modelled geostatistically. Even though annual and monthly precipitation can still have zero values in arid regions and exhibit strong positive skewness, spatial interpolation using kriging is less problematic in these cases than for daily or hourly precipitation, because temporally aggregated precipitation tends more to the normal distribution. However, when mapping hourly or daily precipitation, spatial variability is higher, the stationarity assumption becomes questionable, and the distribution of precipitation becomes skewed, and has a lot of zeroes (Carrera-Hernández and Gaskin 2007; Castro et al. 2014). Similar problems may occur with kriging air quality indices or concentrations of pollutants in ground- and surface water (Gräler et al. 2013). In these situations, kriging may not be a good choice.

In recent years, more and more use is being made of machine learning techniques for spatial interpolation (Li and Heap 2014). ML heavily relies on the strength of the relation between the dependent variable and covariates and can produce remarkably accurate results if this correlation is strong. Nowadays remote sensing (RS) based covariates are abundant and this has given a boost to ML for spatial and spatio-temporal mapping. One of the strengths of ML is that it is very flexible and not restricted to linear relations, as in linear regression, regression kriging, and kriging with external drift (Li et al. 2011; Appelhans et al. 2015; Hengl et al. 2015; Kirkwood et al. 2016; Hashimoto et al. 2019). ML for spatial interpolation is used in many fields, including soil science, climatology, geology, econometrics, spatial planning, and land use mapping. For example, Kirkwood et al. (2016) used quantile regression forests to map soil geochemical variables in southwest England and obtained more accurate results compared with ordinary kriging. The authors concluded that eventually the spatial autocorrelation of the target variable was entirely captured by the auxiliary variables. Kirkwood et al. (2016) and Veronesi and Schillaci (2019) gave an extensive overview of the application of ML in soil mapping. Mohsenzadeh Karimi et al. (2018) compared ML methods and reported that random forest was superior to support vector machines (SVM) and artificial neural networks in estimating long-term monthly air temperature. Hashimoto et al. (2019) proposed a NASA Earth Exchange Gridded Daily Meteorology (NEX-GDM) RF model for mapping daily precipitation (among other meteorological variables) at 1 km spatial resolution using satellite, re-analysis, radar, and topography data for the conterminous United States, from 1979 to 2017.

Despite the increased use and mapping successes, most of the RF frameworks for spatial interpolation do not take into account that the observations are geo-referenced and may be spatially correlated. In other words, they do not fully exploit the available spatial information. Some approaches to include a geographic context into ML were to introduce longitude and latitude as covariates (Li et al. 2011; He et al. 2016; Mohsenzadeh Karimi et al. 2018; Čeh et al. 2018; Georganos et al. 2019), as well as to use distance-to-coast (Li et al. 2011) and distance-to-closest dry grid cell as

covariates (He et al. 2016). He et al. (2016) also used precipitation at adjacent grid cells as covariates for downscaling precipitation using random forest. Behrens et al. (2018) used x- and y-coordinates and distances to the corners and center of a bounding box around the sampling locations as covariates. Hengl et al. (2018) introduced RFsp, which uses buffer distance maps from observation points as covariates. The authors showed that adding these covariates improved prediction and produced results that mimic kriging. Zhu et al. (2019) proposed an ML model which considers autocorrelation to reconstruct surface air temperature data at high spatial resolution across China. They added weights based on altitude and distance differences between the target station and surrounding stations as covariates. Georganos et al. (2019) proposed Geographical Random Forest as a function of RS covariates for modelling population density in Dakar, Senegal. This methodology imitates geographically weighted regression by fitting local RF models for each observation location using the covariates from n nearest observations as training data, while for prediction the closest RF model is used. Hashimoto et al. (2019) proposed the AINA methodology, which is similar to the method of Georganos et al. (2019), with the difference being that Hashimoto et al. (2019) fitted models to grid cells and made predictions by weighing 16 surrounding RF models.

However, to the best of our knowledge, none of the current approaches that aim to include geographical context in ML explicitly included the actual observations at the nearest locations of the prediction location as covariates. This is quite surprising because it seems to be a natural choice to include them; it is the very basis of kriging and most deterministic interpolation methods.

With this in mind, the objectives of this paper were: (1) to introduce Random Forest Spatial Interpolation (RFSI), i.e., RF which includes the neighbouring observations and their distances to the prediction location as covariates, and (2) to evaluate the performance of RFSI against simple deterministic interpolation techniques (NN, TS, and IDW), kriging, standard RF, and RFsp. For this purpose, we first define the RFSI approach and give a brief overview of existing, alternative interpolation methods. Next, we analyse its performance using a synthetic case study where realities were simulated from normally distributed stationary random fields, with a known semivariogram. In such a case it is known that kriging is optimal. The performance of RFSI in this case was evaluated and compared with the performance of OK, RFsp, IDW, NN, and TS. Finally, RFSI was applied to two real-world case studies, a daily precipitation dataset for Catalonia for the years 2016–2018 and a mean daily temperature dataset for Croatia for the year 2008 (i.e., the same dataset as used in Hengl et al. (2012) and compared its performance to STRK, IDW, standard RF and RFsp by using nested k-fold cross-validation.

A complete script in R (R Development Core Team 2012) and datasets for prediction and benchmarking of the prediction efficiency are available and can be obtained via the GitHub repository at https://github.com/AleksandarSekulic/RFSI.

5.2 Materials and Methods

5.2.1 Methodology

Interpolation methods that were used in the following experiments: simple deterministic interpolation methods, such as NN, IDW, and TS of the second order (TS2), and OK are explained in detail in Sections 2.2.1 and 2.2.2.1. In the following section, a short description of RF which was also already explained in Section 2.3.2.1 is given, and then Section 5.2.1.2 describes the main contribution of this research - Random Forest Spatial Interpolation.

5.2.1.1 Random Forest and RFsp

Random Forest (Breiman 2001) is an ensemble ML algorithm that uses a large number of decision trees (Breiman et al. 1984) made on the subsets of observations and covariates. RF actually evolved from bagging (Breiman 1996, 2001) and they actually work the same, except that RF introduced the random feature (covariate) selection in the process of making the subsets based on which decision trees are made and. By doing that, RF decision trees become uncorrelated. RF is already explained in detail in Section 2.3.2.1. Even more detailed explanation of CART, bagging and RF can be found in James et al. (2013).

The overall RF model predictions can be written as

$$\hat{z}(s_0) = f(x_1(s_0), x_2(s_0), \ldots, x_m(s_0)) \tag{5.1}$$

where the $x_i(s_0)(i = 1, \ldots, m)$ are covariates at location s_0. RF has an option for measuring variable importance, which quantifies how much each feature influences the RF model accuracy. RF can also be used to assess accuracy based on out-of-bag error statistics (James et al. 2013).

RFsp is a straightforward extension of RF, which includes buffer distance maps to all observation locations as covariates (Hengl et al. 2018). Each buffer distance map is obtained by calculating Euclidean distances from the centers of all prediction pixels to the center of the pixel in which an observation location falls. Thus, in RFsp there are as many buffer distance maps as there are observations.

5.2.1.2 Random Forest Spatial Interpolation

Spatial autocorrelation between observations is not included in standard RF, other than indirectly through spatial correlation in covariates. Considering that nearby observations carry information about the value at a prediction location, additional covariates were incorporated in the RF model. The added covariates are defined as the observations at the n nearest locations and the distances from these locations to the prediction location. Hence, the RFSI model is as follows:

$$\hat{z}(s_0) = f(x_1(s_0), \ldots, x_m(s_0), z(s_1), d_1, z(s_2), d_2, z(s_3), d_3, \ldots, z(s_n), d_n) \tag{5.2}$$

where $s_i(i = 1, \ldots, n)$ is the i-th nearest observation location from s_0 and $d_i = |s_i - s_0|$.

The workflow of the RFSI algorithm is presented in Figure 5.1. For each training location, the n nearest locations are derived and their observations and distances to the training location are included as covariates, along with other environmental covariates. Prediction is made in the same way: for each prediction location, the observations of and distances to the n nearest locations are used.

5.2.2 Datasets and Covariates

5.2.2.1 Synthetic Dataset

The sequential simulation algorithm of the R package `gstat` (Pebesma 2004) was used to generate realisations of a stationary random field. This algorithm randomly visits each simulation location (i.e., grid node) in the study area and simulates a value based on the conditional Gaussian distribution, conditioned on already simulated values and the known semivariogram and mean. For more details see Bivand et al. (2013a).

All simulations were performed over a 500 $\times$ 500 regular grid (250,000 pixels). We imposed a

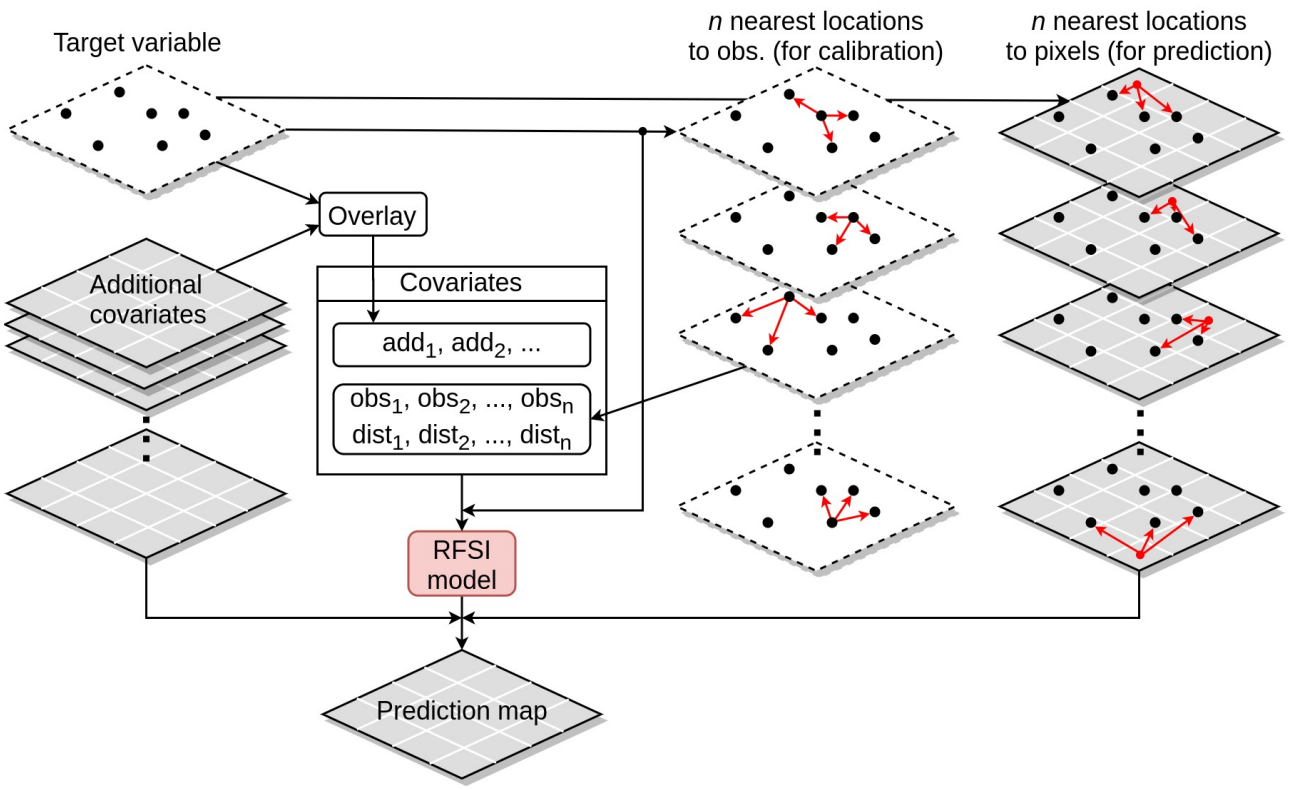


Figure 5.1: Schematic representation of the RFSI algorithm.

mean of 20 and used spherical semivariograms with a sill of 10 units, semivariogram ranges of 50 and 200 units, and nugget-to-sill ratios of 0.00, 0.25 and 0.50. To speed up simulation, the maximum number of conditioning data was set to 50 (i.e., the nearest 50 points). For each of the six semivariogram combinations, 100 different simulations were performed. As explained later, this was done to eliminate unwanted effects of incidental characteristics of single realisations on the results.

5.2.2.2 Precipitation Dataset

Catalonia is an autonomous region in the north-east of Spain that covers 32,108 km^2 (Figure 5.2). Catalonia was chosen as a study area because it has a well-established network of meteorological stations and observations are freely available through the GHCN-daily (Menne et al. 2012, Section 3.2.3). The Catalonia station dataset that was used to model daily precipitation with the tested methodologies consists of observations from 87 GHCN-daily stations for a three-year period, from 2016 to 2018. All observations which failed any of the GHCN-daily quality assurance checks (2948, 3.1% of the total) were removed from the dataset. Coordinates were reprojected from WGS 84 global reference system to UTM zone 31N projection (which is appropriate for Catalonia) before computing Euclidean distances to nearest stations, as required in RFSI. The station locations and a histogram of the observations are shown in Figure 5.2. About 69% (63,880 of a total of 92,404 observations) of the GHCN-daily precipitation data are zero. The maximum observed daily precipitation amount is 220.9 mm.

Three environmental covariates were included in the kriging and RF models in the precipitation case study.

The IMERG (Huffman et al. 2014, Section 3.3.2.2) late run version V06A precipitation estimates were used in this case study. We did not use the final run because this incorporates the GHCN-daily station precipitation, which is the dependent variable we aim to predict. IMERG estimates are a

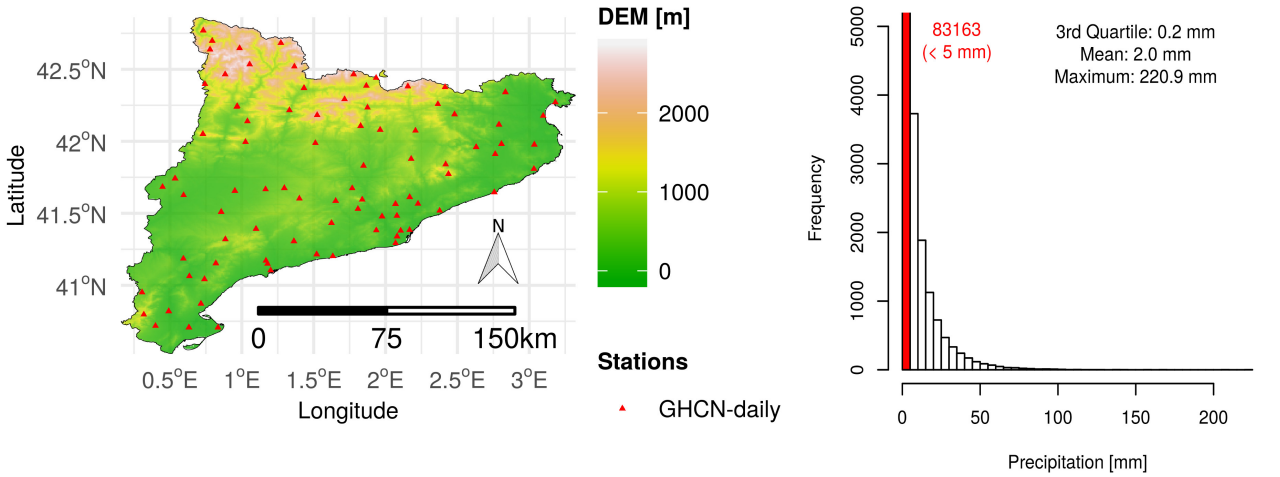


Figure 5.2: GHCN-daily station locations on top of a digital elevation model of the study area (left) and histogram of daily precipitation for Catalonia (right). The histogram contains 92,404 GHCN-daily observations for the 2016—2018 period.

space-time covariate with a spatial resolution of 10 km and temporal resolution of one day (Figure 5.3).

Space-time daily covariates, maximum (TMAX) and minimum temperature (TMIN) (Figure 5.3) estimated with models proposed by Kilibarda et al. (2014) were also used. Including DEM as a covariate was also tested, but this did not improve model accuracy, presumably because the effect of elevation was already accounted for by the other three covariates.

5.2.2.3 Temperature Dataset

The Croatian temperature dataset consists of 57,282 observations from 159 stations for the year 2008, provided by the Croatian National Meteorological Service. The station locations are shown in Figure 5.4. The minimum and maximum observed daily temperature values are −14.1 °C and 32.6 °C, respectively. Station coordinates are in UTM zone 33N projection. Covariates used to model mean daily temperature were latitude, longitude, distance-to-coastline, elevation, seasonal fluctuation, insolation (total incoming solar radiation), and MODIS LST images (insolation and MODIS LST images are space-time covariates). A detailed description of this dataset and covariates is given in Hengl et al. (2012).

5.2.3 Accuracy Assessment

The following accuracy metrics were used for all three case studies: coefficient of determination ($R^2_{1:1}$), Lin's concordance correlation coefficient (Lin 1989), mean absolute error, and root mean square error. Because the coefficient of determination used here should not be confused with the square of the Pearson correlation between observed and predicted values, we denote it as $R^2_{1:1}$ and define it as:

$$R^2_{1:1} = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(z(s_i) - \hat{z}(s_i))^2}{\sum_{i=1}^{n}(z(s_i) - \overline{z}(s_i))^2} \tag{5.3}$$

where ESS is the Error Sum of Squares, TSS the Total Sum of Squares, and $\overline{z}(s_i)$ the mean of the observations. In the synthetic case study, the accuracy metrics were calculated for all prediction locations, since the "true" value is known for all pixels. In the real-world case studies, a cross-

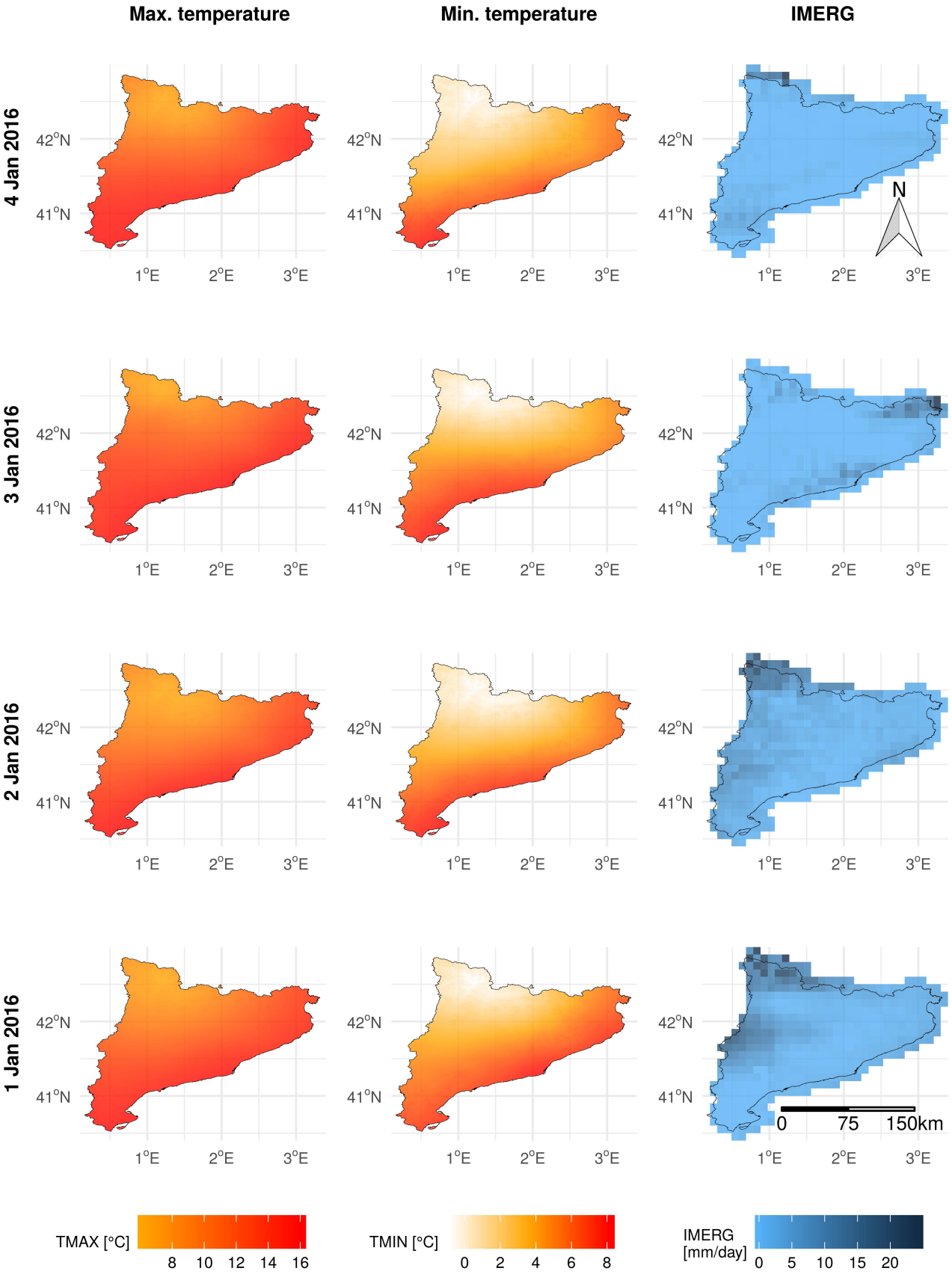


Figure 5.3: Maximum temperature (left), minimum temperature (middle) and IMERG precipitation estimates (right) for fourexample days, 1–4 January 2016.

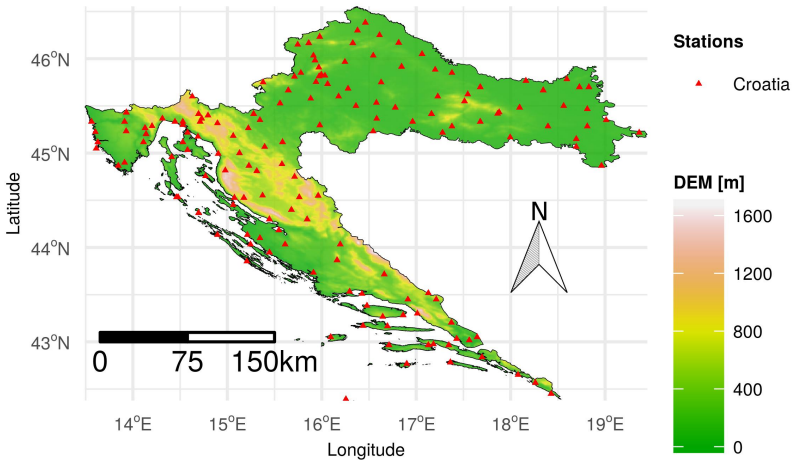


Figure 5.4: Station locations in Croatia on top of a digital elevation model of the study area.

validation approach was used as explained in Section 5.2.3.2 below.

5.2.3.1 Synthetic Case Study

Each of the 600 simulated datasets (100 different simulations for each of six semivariogram combinations) was randomly split in two: a sample dataset and a test dataset. An advantage of the synthetic case is that the reality for the entire study area is known. This means that accuracy metrics can be computed by comparing predictions with observations on a test dataset that comprises the entire area (except the relatively small training dataset), instead of using cross-validation. In this way, the accuracy metrics are no longer estimates, but true metrics, calculated without error.

For kriging and deterministic interpolation methods, the sample dataset was used to generate predictions. To eliminate the effect of semivariogram estimation errors, the model parameters that were used to generate the simulations were used for kriging. For each semivariogram case, the spatial interpolations were done for all 100 realisations, accuracy metrics computed over the test dataset, and averaged over all 100 cases. This was done to avoid accuracy metrics being influenced by incidental characteristics of a single realisation.

For both RF models (RFsp and RFSI), the sample dataset was used as training data for model calibration. The sample dataset (and/or their locations) was also used to define the additional covariates specific to these methods. Splitting was done six times with different sizes of the sample dataset: 100, 200, 500, 1000, 2000, and 5000 locations (0.04%, 0.08%, 0.20%, 0.40%, 0.80%, 2.00% of the total, respectively). In this way we could also analyse the sensitivity of the accuracy metrics of all interpolation methods to the number of sample locations. RFsp and RFSI were trained by the R package `ranger` (Wright and Ziegler 2017). Spatial covariates, i.e., observations and (Euclidean) distances to the nearest locations were calculated with the `knn` function of the R package `nabor` (Elseberg et al. 2012) and R package `doParallel` (Microsoft Corporation and Steve Weston 2019).

None of the RF hyperparameters were tuned, because this would be too computationally demanding, given that 600 simulations were done. Also, the results with tuned hyperparameters were checked for some simulations and were found not to be significantly different from those obtained with default hyperparameter values. A total of 250 trees (`ntree` parameter in R) were used for modelling RFsp and RFSI. Random feature selection (`mtry` parameter in R) for RFSI modelling was done with one third of the covariates (the default value). For RFsp, `mtry` was set to two-thirds of the number of covariates, as recommended by Hengl et al. (2018). The additional covariates used in RFSI were derived from the 25 nearest locations. IDW predictions were made by the `idw` function from R package `gstat`, using the 25 nearest observations and setting the exponent parameter p to

2. NN predictions were also made using the `idw` function from R package `gstat`, by setting the number of nearest observations to 1. TS predictions were made using the R `lm` function. Kriging was done using the `krigeST` function from R package `gstat`.

5.2.3.2 Real-World Case Studies

In the precipitation case study, the accuracy was assessed using a "target-oriented" cross-validation strategy (Meyer et al. 2018), i.e., by a nested 5-fold leave-location out cross-validation (LLOCV). For the temperature case study, a nested 10-fold LLOCV was used, as done in Hengl et al. (2012), enabling a comparison of results. Leave-location-out means that entire stations (with all their observations) were assembled in the same fold. Thus, the data were first split into K (five or ten) main folds, where K $-$ 1 folds comprised a calibration dataset and the remaining fold a test dataset. Next, the calibration dataset was split into K nested folds to estimate the hyperparameters using a standard LLOCV and fit the model. The test dataset was then used to assess the performance of the model. The advantage of nested LLOCV over standard LLOCV is that the data of the test fold are not used to tune the RF hyperparameters (Pejović et al. 2018). The hyperparameters for the final RF models were then calculated based on standard LLOCV, i.e., without nested folds (their role is just to approximate the accuracy of the final model). The same approach was used for STRK, where each calibration dataset was used to fit a linear regression trend and the residual semivariogram. Final accuracy metrics were calculated based on the predictions from all test datasets (i.e., K main folds).

RF hyperparameters, number of variables to possibly split at each node (`mtry`), minimal node size (`min.node.size`) and ratio of observations-to-sample in each decision tree (`sample.fraction`) were tuned for RF, RFsp, and RFSI models. Additionally, the number of nearest stations to be included (n) was tuned for RFSI. The number of trees (`num.trees`) hyperparameter was set to 250. The number of nearest stations n and p exponent were also tuned for IDW. The `stratfold3d` function of the R package `sparsereg3D`[2] was used to create K main folds for nested LLOCV with equally spatially distributed locations (by longitude and latitude).

In the case of OK and RK, the kriging prediction error was also characterized by the kriging standard deviation (Goovaerts 1997). In case of RF, prediction uncertainties were quantified using Quantile Regression Forest (QRF) (Meinshausen 2006). The interquartile range (IQR) was calculated as:

$$IQR = \hat{z}_{q=0.75} - \hat{z}_{q=0.25} \tag{5.4}$$

where $\hat{z}_{q=0.75}$ and $\hat{z}_{q=0.25}$ are QRF predictions of the 0.75 and 0.25 quantiles, respectively (i.e., upper and lower quartiles). Assuming that the kriging prediction errors are normally distributed, the kriging IQR can be calculated as $1.35 \cdot sd$, where sd is the kriging standard deviation.

5.3 Results

5.3.1 Synthetic Case Study

Average MAE values over 100 simulations per interpolation method for each of the six semivariogram combinations are presented in Figure 5.5. Plots with $R^2_{1:1}$, CCC, and RMSE are only presented for one of the six cases (Figure 5.6), because these have similar patterns as MAE. The results are presented in the form of bar charts, grouped by the size of the sample dataset. Each individual plot represents one of six semivariogram combinations.

[2]https://github.com/pejovic/sparsereg3D

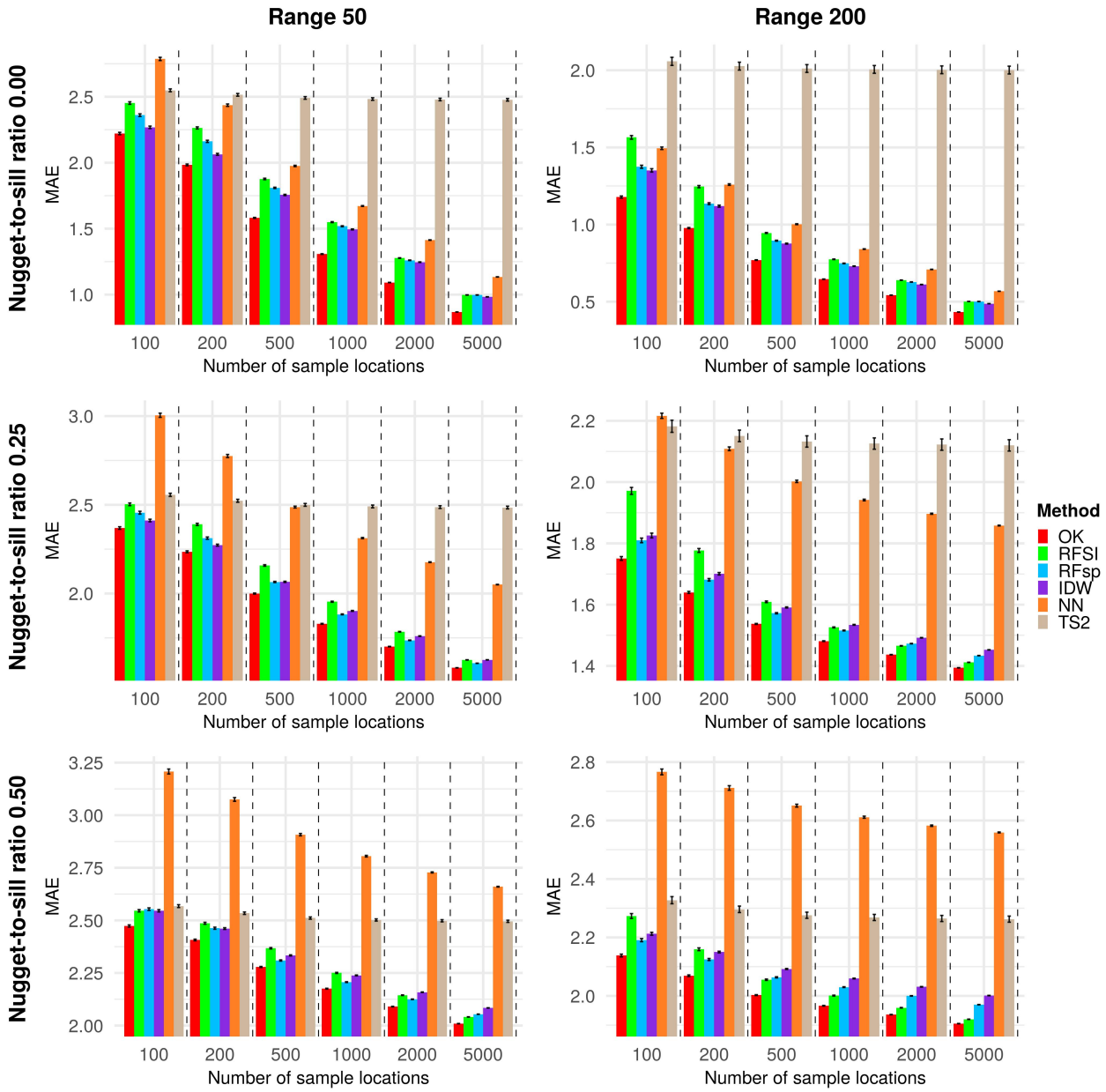


Figure 5.5: Comparison of average MAE estimated for each of the interpolation methods, for all nugget-to-sill ratios and ranges. Coloured bars are average MAE for test locations from 100 different simulations. Error bars are standard errors computed from 100 simulations.

Figure 5.5 shows, as expected, that OK was the best predictor in all cases. IDW, RFsp, and RFSI had similar performance and were the most accurate after OK. IDW was the best (after OK) in case of a low nugget, whereas RFsp and RFSI were better for higher nugget-to-sill ratios, especially if the range was large. In case of a low nugget, when there is a lack of noise, spatial variation was smooth and well captured by IDW. The difference between RFsp and RFSI was small in most cases. When the number of sample locations increases, the difference between the RF models (RFsp and RFSI) and OK decreases, faster for the 0.25 nugget-to-sill ratio case than for the 0.00 nugget-sill ratio case. NN and TS overall had poor performance. The reason for this is that NN uses only the nearest observation, which is a poor strategy, particularly in the case of a large nugget. The disadvantage of TS is that it has only a few global parameters. For this reason it cannot benefit from large sample datasets.

Table 5.1 shows the average distance calculation time, modelling time, and prediction time for

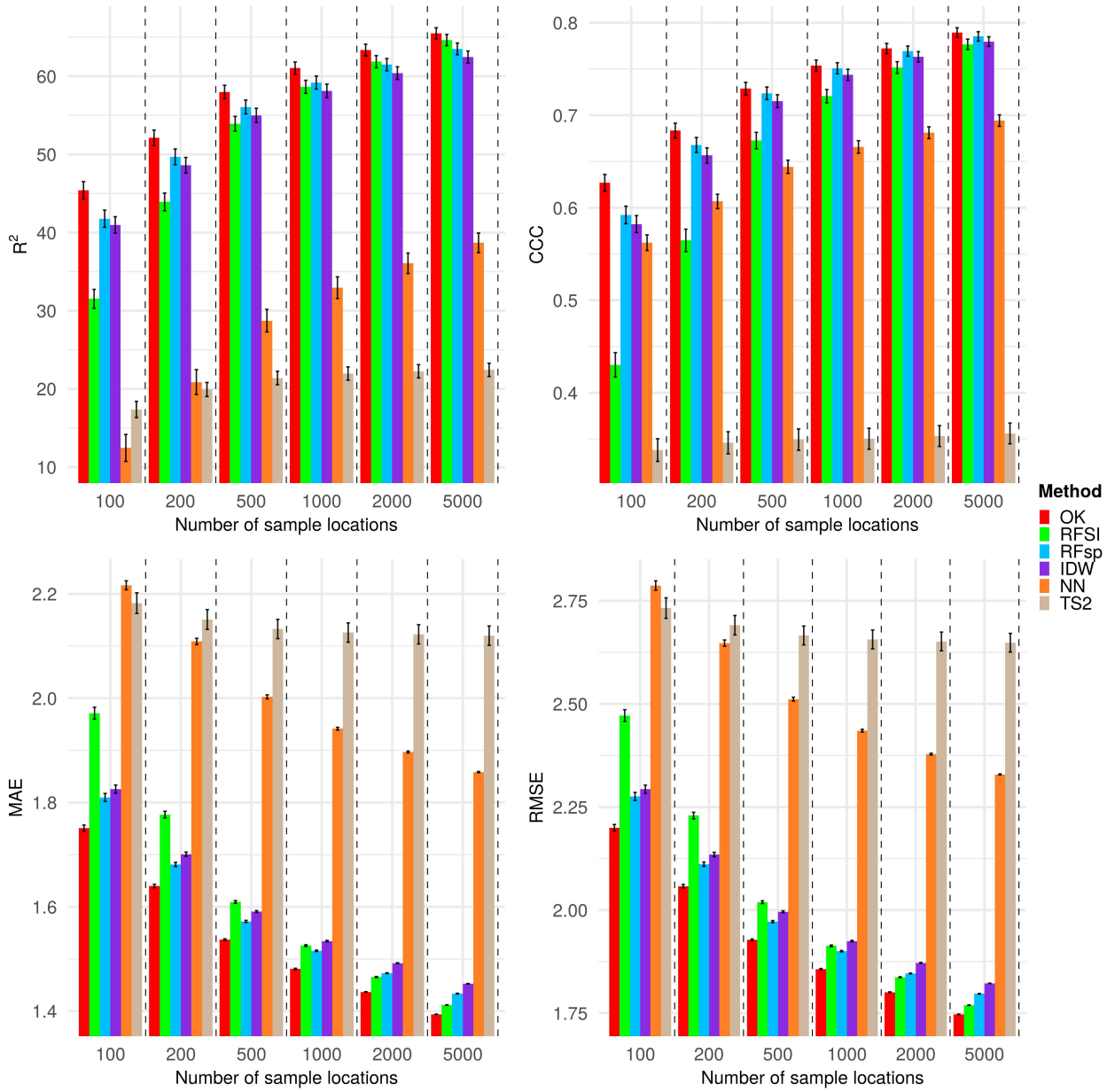


Figure 5.6: Comparison of $R^2_{1:1}$ (top left), CCC (top right), MAE (bottom left) and RMSE (bottom right) estimated for each of the interpolation methods, for nugget-to-sill ratio 0.25 and range 200.Coloured bars are average accuracy metrics for test locations computed from 100 different simulations. Error bars are standard errors computed from 100 simulations.

RFsp and RFSI, for all semivariogram cases. RFSI was much faster than RFsp in all cases, especially for large sample datasets. RFSI calculates distances to the n nearest locations, whereas RFsp creates a covariate raster with distances for each sample location. This also means that RFsp is a memory consuming process. If there is a large number of locations (more than 1000), sometimes the entire RAM memory was used and the calculation process slowed down significantly. The prediction computing time of RFSI was similar or even smaller compared with that of local OK.

Prediction maps of one randomly selected simulation for the 0.25 nugget-to-sill ratio, 50 range, and 500 sample locations case are presented in Figure 5.7. As expected, TS produces a very smooth surface. Also, typical Thiessen polygons are visible in the NN prediction maps. IDW, OK, RFsp and RFSI prediction maps have similar patterns, although they vary in degree of noisiness.

Table 5.1: Distance calculation time and modelling time for RFSI and RFsp, and prediction time for RFSI, RFsp and OK. All results refer to the synthetic case study and represent the average computing time computed from 100 simulations. All calculations and time estimations were done on a personal computer with Intel® Core™ i7-7820X CPU @ 3.60GHz × 16 processor and 126 GB of RAM.

Criteria	**Method**	**Number of Points**					
		100	**200**	**500**	**1000**	**2000**	**5000**
Distance calculation time [s]	RFsp	24.98	47.75	114.42	263.08	477.37	3832.88
	RFSI	1.40	1.48	1.62	1.65	1.69	1.75
Modelling time [s]	RFsp	0.06	0.27	2.35	13.50	71.73	498.21
	RFSI	0.02	0.04	0.09	0.20	0.42	1.18
Prediction time [s]	OK	5.25	5.72	6.38	6.81	7.11	8.03
	RFsp	5.47	9.57	22.30	46.32	70.58	312.12
	RFSI	2.93	3.37	4.05	4.74	5.60	6.83

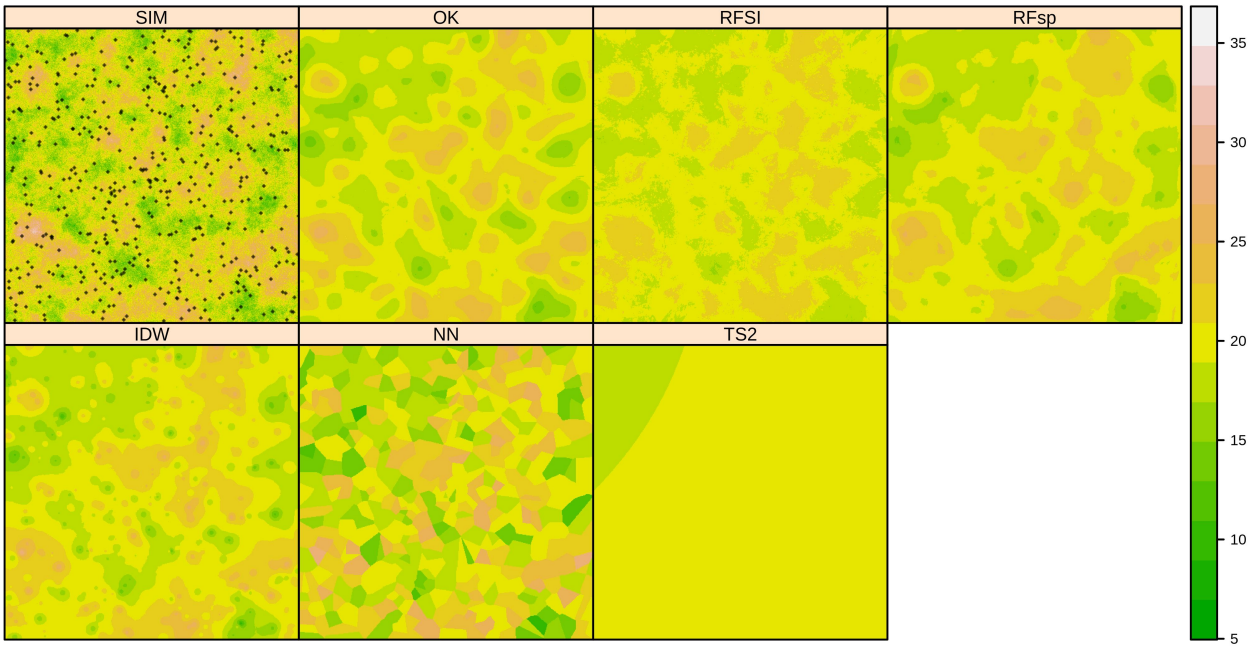


Figure 5.7: Prediction maps made using 500 sample locations with nugget-to-sill ratio 0.25 and range 50, for one of the 100 simulated realities. The top left map (SIM) shows the simulated reality and the locations of the 500 samples.

The top ten most important covariates for RFSI are all nearest observations, with the highest importance for the very nearest observations (Figure 5.8). This clearly shows that in RFSI distances are less important than observations. Figure 5.8 was created based on the realisation shown in Figure 5.7. Other realisations and semivariogram cases were also checked and had similar results for RFSI. The type of feature importance used was `impurity`, which means that the importance of the feature was represented by how much the overall variance decreased by using that feature when partitioning the instances (Wright and Ziegler 2017). Furthermore, the feature importance index was scaled to a maximum of 1.

To evaluate the sensitivity of RFSI to the choice of the number of nearest locations (n), RFSI prediction maps obtained with different numbers of nearest locations were compared. Figure 5.9 shows that by increasing the number of nearest locations, prediction maps become smoother. This figure

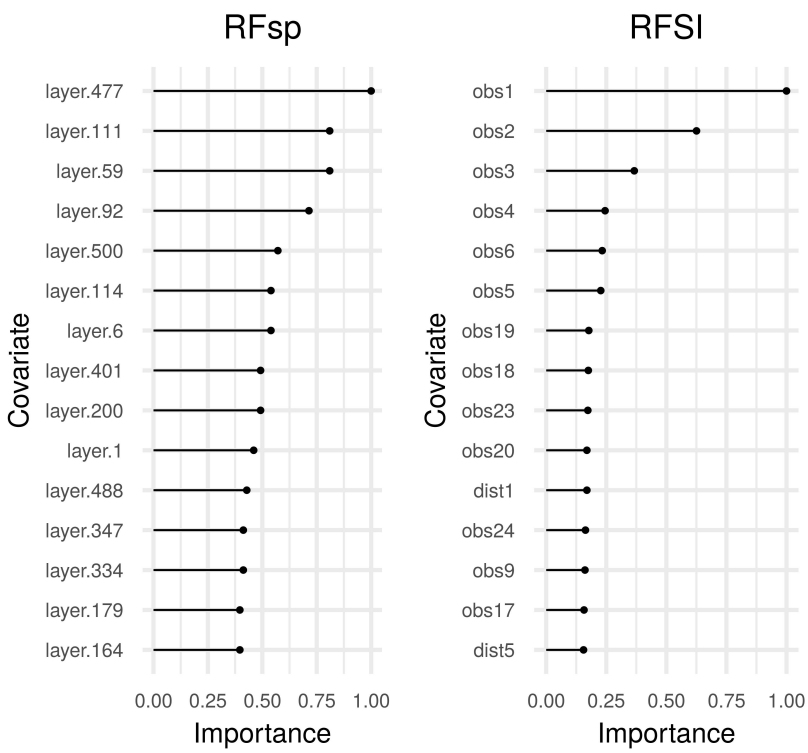


Figure 5.8: Covariate importance plot for RFsp (left) and RFSI (right), for the case shown in Figure 5.7. The importance index is scaled to a maximum of 1, obs$i$ and dist$i$ represent observations and distances to the i-th nearest observation location, and layer.i represents buffer distances to the i-th observation location.

refers to the case shown in Figure 5.7. Furthermore, by increasing the spatial range and sample size, the optimal value of n increases (Figure 5.10). After reaching the optimal value, the accuracy mostly stayed constant. The exception was a case with range 50 and 100 sample locations, because the sample size was small and there were insufficient data for modelling a variable with a small spatial correlation length.

5.3.2 Precipitation Case Study

Since precipitation varies both in space and time, the precipitation case study is referred to as spatio-temporal interpolation. The performance of RFSI was compared with STRK, IDW, standard RF and RFsp. Other deterministic interpolation methods (NN, TS) were not taken into consideration because these were already outperformed in the synthetic case and cannot easily take environmental covariates into account.

5.3.2.1 Spatio-Temporal Regression Kriging (STRK)

STRK was done in a similar way as in Hengl et al. (2012) and Kilibarda et al. (2014). First, a multiple linear regression model was used to fit a trend function, and then, the regression residuals were interpolated using spatio-temporal ordinary kriging. Using the R `lm` function the RK trend was given by:

$$trend_{RK}(s,t) = 6.466 + 0.055 \cdot IMERG(s,t) - 0.499 \cdot TMAX(s,t) + 0.478 \cdot TMIN(s,t) \quad (5.5)$$

The trend model explained 40.9% of the variation of the daily precipitation. The residual standard deviation was 5.3 mm. Residuals of −124.6 mm and 202.2 mm occurred and were the consequence of precipitation extremes. More than 98% of the residuals were between −20 mm and +20 mm. Log-transformation of the precipitation data prior to modelling was tried, but this did not improve results.

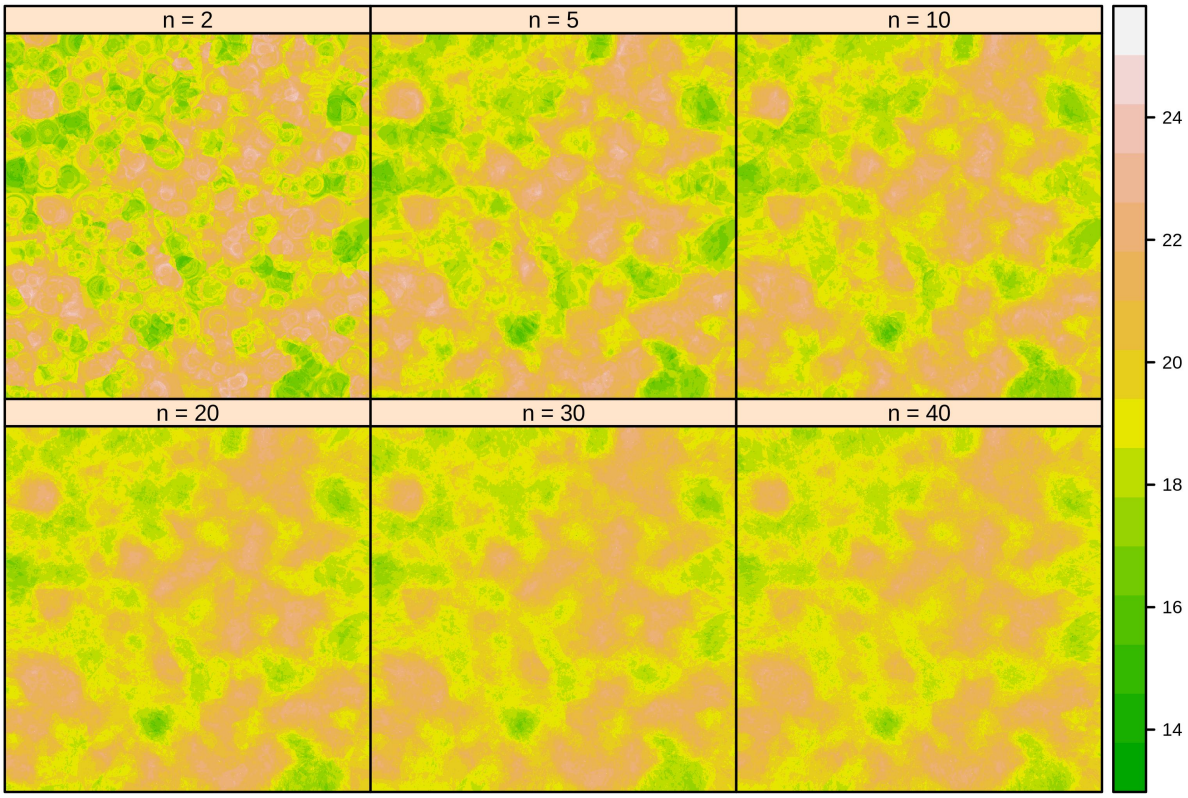


Figure 5.9: RFSI prediction maps made using 500 sample locations with nugget-to-sill ratio 0.25 and range 50, with different number of nearest locations (n).

The extremes can be a problem for RK, but it should be noted that daily precipitation was purposely chosen as a real-world case study because it is difficult to model geostatistically. A histogram of the residuals is presented in Figure 5.11. The residual sample and fitted sum-metric semivariogram are given in Figure 5.12. The sum-metric semivariogram (Heuvelink et al. 2017), which is the sum of three semivariograms that model spatial, temporal and spatio-temporal correlation, was fitted using the R package `gstat`. Table 5.2 shows the parameters of the fitted sum-metric semivariogram. Note that residual temporal correlation was negligible and limited to only a few days, whereas residual spatial correlation was considerable and reached the sill at about 100 km. STRK predicted negative precipitation values in some instances. In those cases the prediction was set to zero.

Table 5.2: Sum-metric semivariogram parameters of the STRK model.

Component	Nugget [mm²]	Sill [mm²]	Range	Function	Anisotropy Ratio
Spatial	0.00	0.89	218.8 km	Spherical	n/a
Temporal	1.63	4.15	2.6 days	Spherical	n/a
Spatio-temporal	9.51	11.30	91.7 km	Spherical	120 km/day

5.3.2.2 IDW and Random Forest Models

The optimized hyperparameters for IDW and final RF models are presented in Table 5.3.

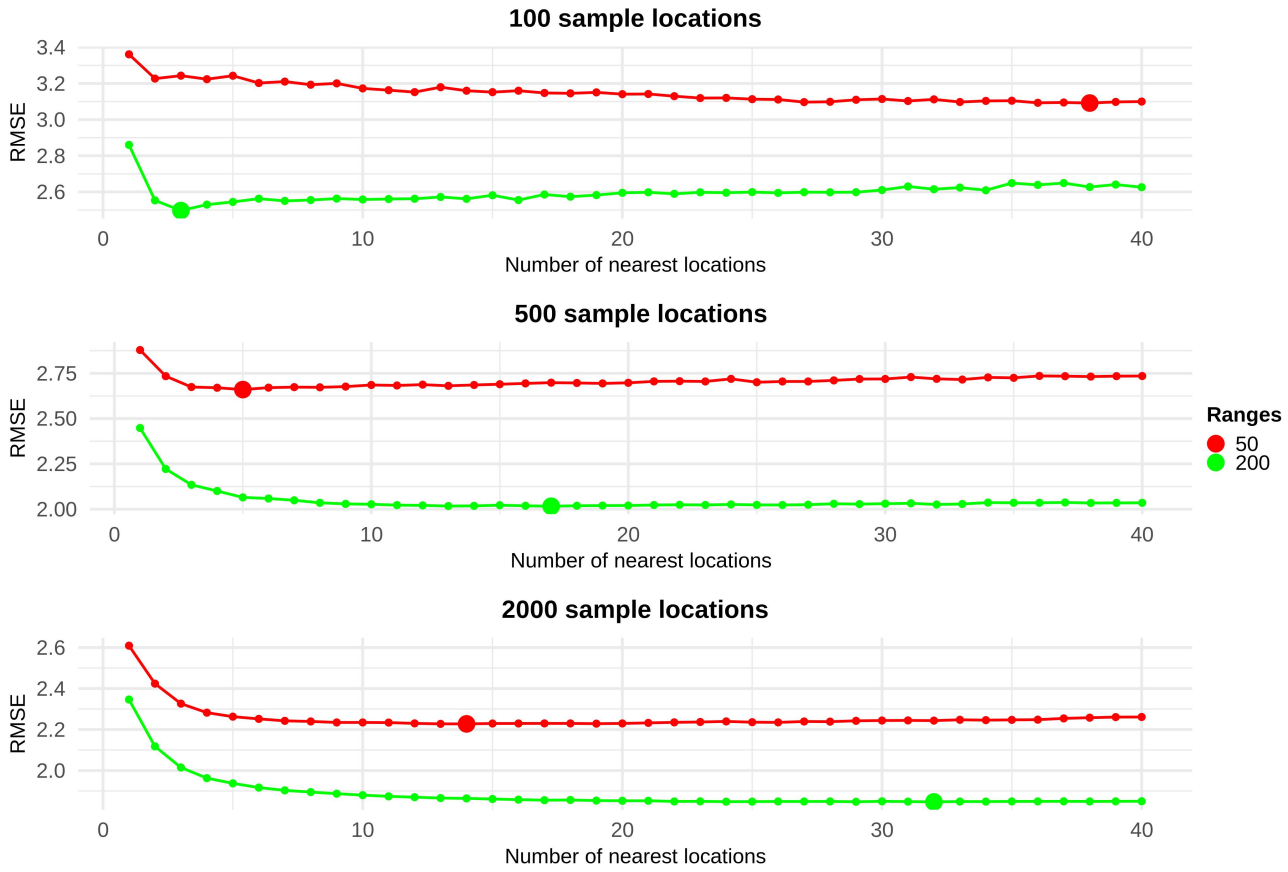
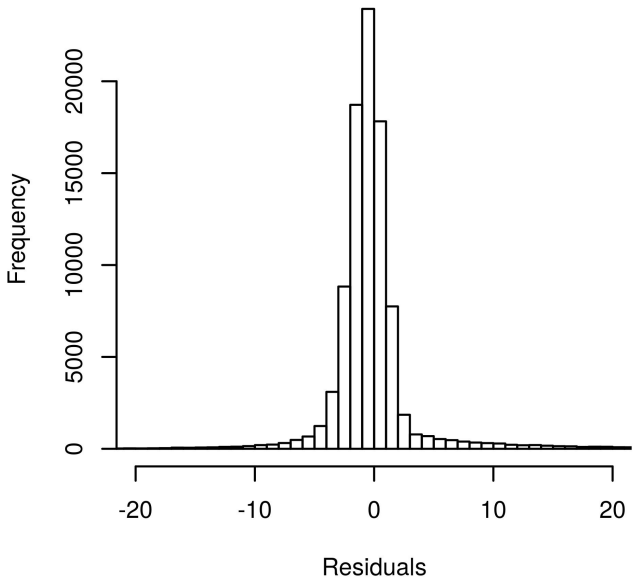


Figure 5.10: RMSE vs number of nearest locations (n) used in RFSI for one simulation with nugget-to-sill ratio 0.25, ranges 50 and 200, using 100 (top), 500 (middle) and 2000 (bottom) sample locations. Larger discs represent the optimal number of nearest locations with minimum RMSE.

Figure 5.11: Histogram of STRK residuals. Residuals smallerthan −20 mm (0.2% of total residuals) and greater than +20 mm (1.2% of total residuals) are not shown.



As in the synthetic case (Figure 5.8), the first few nearest observations, sorted by order, are the most important covariates of the RFSI model (Figure 5.13). IMERG is the most important covariate for RF and RFsp, followed by TMAX and TMIN. The spatial covariates (i.e., distance from stations) have negligible importance in RFsp. IMERG, TMAX, and TMIN are more important than distance covariates for RFSI but substantially less important than the nearest observations.

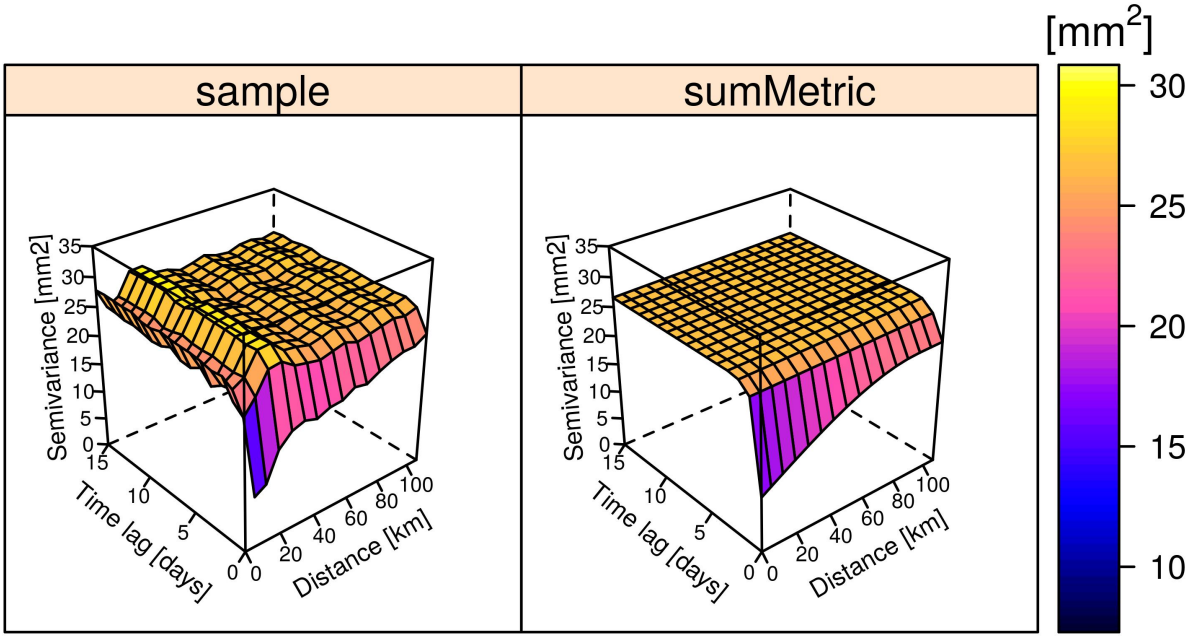


Figure 5.12: STRK sample semivariogram and fitted sum-metric semivariogram.

Table 5.3: Optimized hyperparameters of IDW, RF, RFsp and RFSI for the precipitation case study.

Model	mtry	min.node.size	sample.fraction	n	p
IDW	n/a	n/a	n/a	13	2.2
RF	2	20	0.65	n/a	n/a
RFsp	58	4	0.29	n/a	n/a
RFSI	4	6	0.95	7	n/a

5.3.2.3 Accuracy Assessment

Table 5.4 shows the accuracy metrics for all five models. In addition, RFSI without environmental covariates ($RFSI_0$) was also evaluated. RF exhibited the worst performance as it used fewer covariates than RFsp and RFSI and cannot benefit from residual spatial autocorrelation. RFsp had higher accuracy than RF because it includes buffer distances, but was much less accurate than STRK, IDW, RFSI, and $RFSI_0$. Apparently, STRK, IDW, RFSI, and $RFSI_0$ were more able to capture residual spatial autocorrelation than RFsp. RFSI also outperformed STRK, which may be due to the fact that RFSI is much more flexible in modelling the relation between the environmental covariates and daily precipitation. Interestingly, IDW and $RFSI_0$ performed quite well.

Table 5.4: Accuracy metrics of all six prediction methods as assessed using nested 5-fold LLOCV for the precipitation case study.

Method	$R^2_{1:1}$ [%]	CCC	MAE [mm]	RMSE [mm]
STRK	67.5	0.815	1.2	3.9
IDW	69.6	0.820	1.1	3.8
RF	49.4	0.674	1.7	4.9
RFsp	53.3	0.690	1.6	4.7
RFSI	69.5	0.820	1.1	3.8
$RFSI_0$	68.6	0.814	1.2	3.9

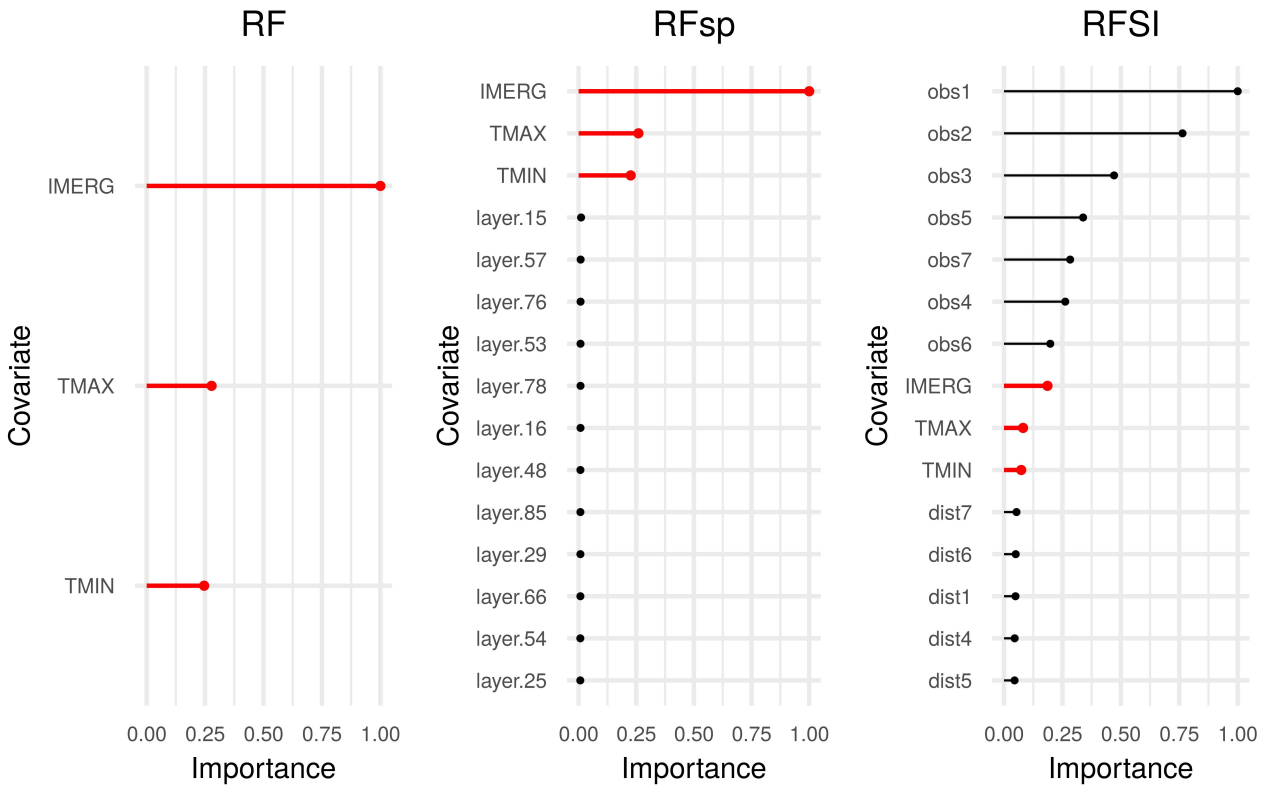


Figure 5.13: Covariate importance plot for RF (left), RFsp (middle) and RFSI (right), for the precipitation case study. The importance index is scaled to a maximum of 1. The importance of covariates IMERG, TMAX, and TMIN is shown in red.

Given that environmental covariates had low importance in the precipitation case study (Figure 5.13), one may ask whether these covariates were informative at all. The results of the standard RF model show that they do have value, because the $R^2_{1:1}$ of RF was 49.4% (Table 5.4). However, the same table shows that the $R^2_{1:1}$ of RFSI with and without environmental covariates only had a small difference of 1%. This indicates that in the precipitation case study, spatial autocorrelation is dominant over environmental covariates, so that using neighbouring observations and their distances alone explains a large part of the variation, after which adding environmental covariates has little added value. This was also confirmed by the relatively high accuracy of IDW interpolation. Note, however, that these results depend on sampling density and may turn out differently in other cases.

Scatter density plots of predictions against observations from nested LLOCV are presented in Figure 5.14. Point clouds for RF and RFsp are more dispersed, which agrees with the higher MAE and RMSE, and lower $R^2_{1:1}$ and CCC, in comparison with STRK and RFSI. Another reason why IDW performed as well as STRK and RFSI might be that IDW managed to model zeros well. Table 5.5 shows the number of hits and misses for predicting zero and non-zero precipitation of all models. Note that 1 mm is taken as a threshold for zero precipitation, because a "dry day" is defined as a day with less precipitation than 1 mm (Tank et al. 2009). IDW and RFSI had the best overall accuracy. STRK, IDW, and RFSI modelled zeros best, while RFSI and $RFSI_0$ were better in modelling precipitation above 1 mm.

Predictions made at 1 km spatial resolution for four example days (Figure 5.15) show that RF and RFsp over-predicted precipitation extremes (34.2 mm and 30.0 mm on 4 January). RFSI predicted a maximum precipitation of 13.8 mm, STRK 13.6 mm and IDW 14.2 mm on 4 January. An advantage of RFSI and other RF models in comparison with RK and STRK is that these do not extrapolate

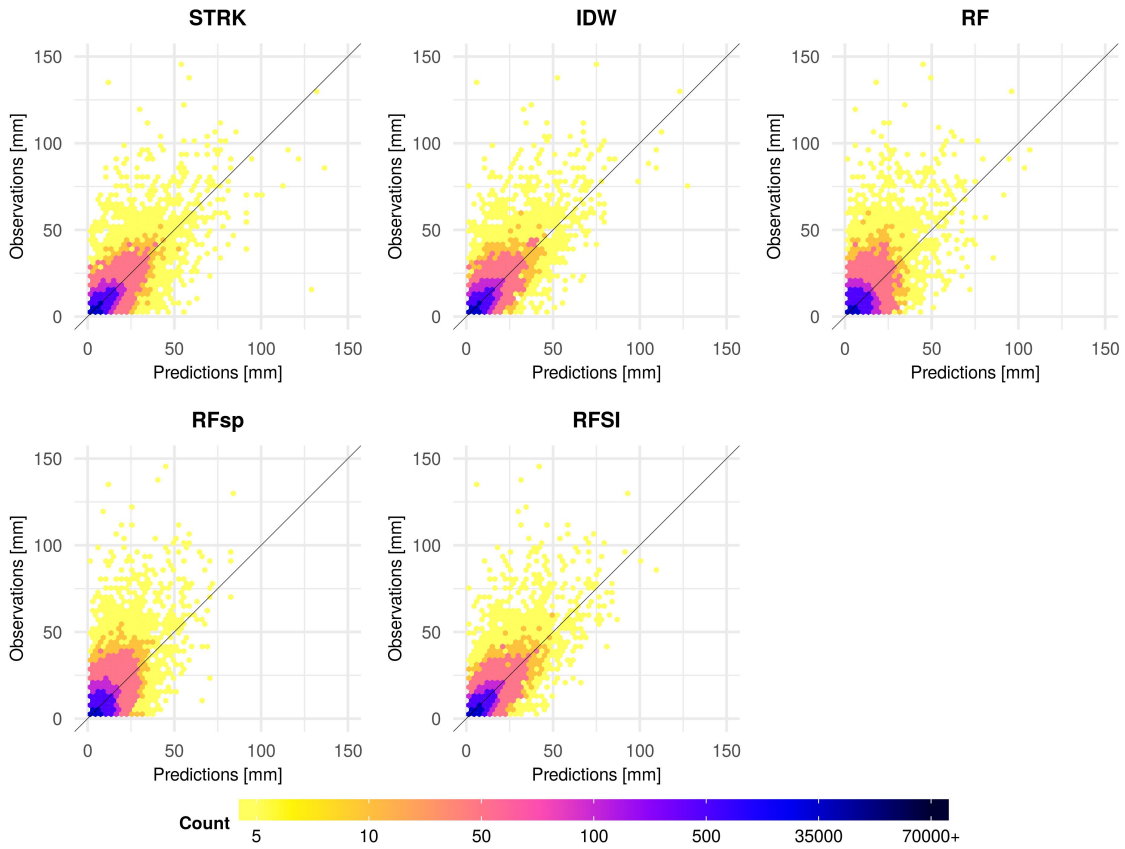


Figure 5.14: Scatter density plots of predictions vs. observations with 1:1 line for the precipitation case study.

Table 5.5: Performance of all models for precipitation below and above 1 mm. Values for the hits and misses are number of observations (obs.) for a given condition. Predictions (pred.) used in this table are from nested 5-fold LLOCV. Overall accuracy represents the percentage of correct classifications. Numbers in bold represent the best performance.

		Hits		Misses		RMSE [mm]		Overall
Method	**obs.**	< 1 **mm**	≥ 1 **mm**	< 1 **mm**	≥ 1 **mm**	< 1 **mm**	≥ 1 **mm**	**Accuracy**
	pred.	< 1 **mm**	≥ 1 **mm**	≥ 1 **mm**	< 1 **mm**			**[%]**
STRK		68,272	15,894	6,307	1,847	1.2	8.6	91.2
IDW		**68,398**	16,164	**6,181**	1,577	**1.0**	**8.4**	**91.6**
RF		63,382	14,914	11,197	2,827	1.7	10.6	84.8
RFsp		64,235	15,273	10,344	2,468	1.6	10.2	86.1
RFSI		68,031	16,524	6,548	1,217	**1.0**	**8.4**	**91.6**
$RFSI_0$		67,917	**16,535**	6,662	**1,206**	**1.0**	8.5	91.5

and do not give negative precipitation predictions. STRK predicted negative precipitation in 35.8% of all cases, with a minimum of −29.9 mm. As mentioned in Section 5.3.2.1, all negative predictions were replaced with zeros. IMERG has a low spatial resolution (Figure 5.3), which leads to a blocky structure in all prediction maps in Figure 5.15, except for RFSI and IDW. IMERG patterns are most noticeable in RF and RFsp predictions, especially on 4 January, because IMERG is their most important feature (Figure 5.13).

The location-specific prediction uncertainty of RF, RFsp, and RFSI was quantified using QRF and displayed together with the STRK IQR in Figure 5.16. The large nugget of the residual semivariogram means that the STRK IQR is substantial everywhere, since it cannot be smaller than the square root of

the nugget variance, multiplied by 1.35. The STRK IQR is fairly constant over space, with somewhat lower values near station locations and somewhat larger values in areas that have a low station density. The IQRs of the RF models have much larger spatial variation: these models do not assume stationarity of the model residual and as a result the IQR is small for zero and low precipitation amounts, whereas it is large on days and in areas with a high precipitation amount, as observed in Hengl et al. (2018). The IQRs of RF and RFsp are much larger than those of RFSI for days with large precipitation amounts.

5.3.3 Temperature Case Study

For this case study, an STRK model was not fitted because it was previously done in Hengl et al. (2012). For the other interpolation methods, the same modelling approach was used as in the precipitation case study.

5.3.3.1 IDW and Random Forest Models

The optimized hyperparameters for the final RF models and IDW are presented in Table 5.6.

Table 5.6: Optimized hyperparameters of IDW, RF, RFsp, and RFSI for the temperature case study.

Model	mtry	min.node.size	sample.fraction	n	p
IDW	n/a	n/a	n/a	11	1.8
RF	6	3	0.85	n/a	n/a
RFsp	154	2	0.77	n/a	n/a
RFSI	5	15	0.90	10	n/a

Seasonal fluctuation, MODIS LST images, insolation and distance-to-coastline were the most important covariates for RF and RFsp (Figure 5.17). Similarly to what was observed for the synthetic and precipitation case studies, the first few nearest observations were the most important covariates for RFSI in this case, followed by MODIS LST images, seasonal fluctuation, DEM and insolation. Distance from stations for RFsp and RFSI were less important than the nearest observations and environmental covariates.

5.3.3.2 Accuracy Assessment

The accuracy metrics for all six models are presented in Table 5.7. As for the precipitation case study, $RFSI_0$ was also evaluated. STRK had the worst performance, possibly because the separable STRK model used in Hengl et al. (2012) is quite restrictive and may not provide a realistic approximation of the true, underlying spatio-temporal structure. IDW and $RFSI_0$ had lower accuracy compared with all other RF models, because they could not benefit from covariates. RF benefited from covariates more than in the precipitation case study. Buffer distance covariates did not give an added value and thus RFsp performed worse than RF. At the same time, RFSI benefited from nearest observation covariates more and therefore outperformed all other methods.

Predictions made at 1 km spatial resolution for February 2, 2008 are shown in Figure 5.18. IDW predictions are the smoothest. All RF methods (RF, RFsp and RFSI) show similar patterns of influence of the most important covariates, especially seasonal fluctuation, MODIS LST images and insolation.

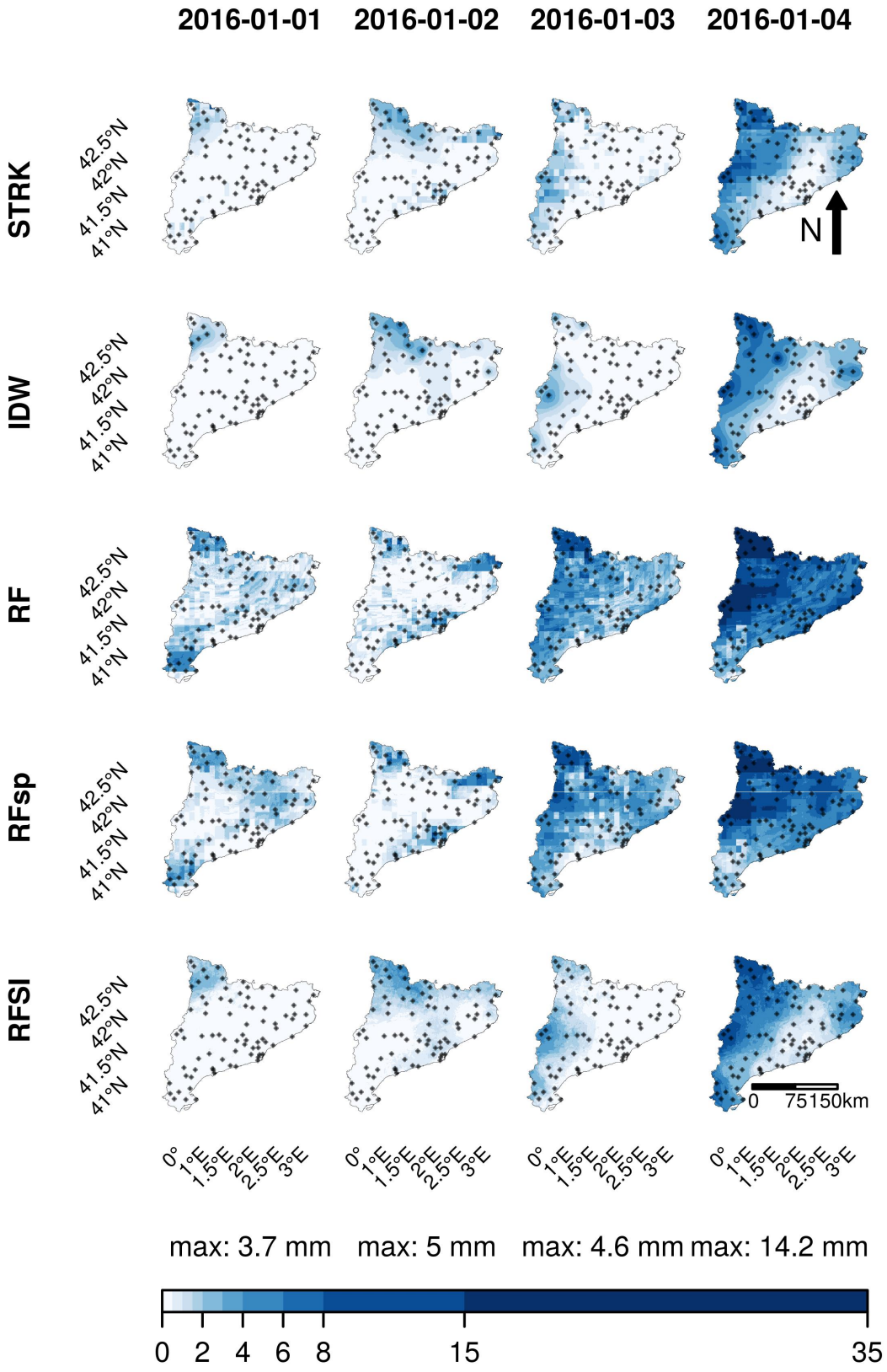


Figure 5.15: Prediction maps of daily precipitation (mm) for the five models, for 1–4 January 2016. The bottom row shows the maximum observed precipitation for each day.

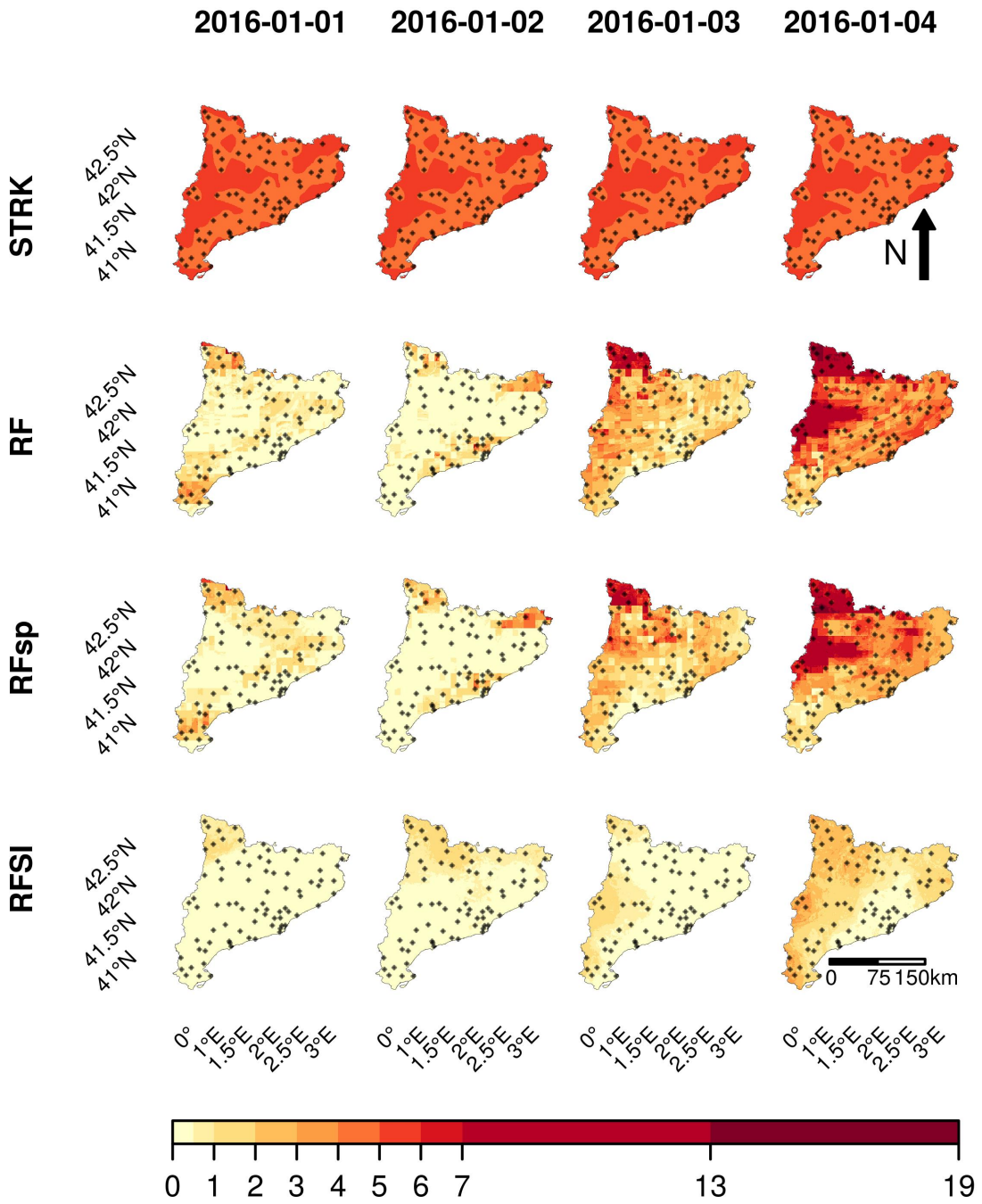


Figure 5.16: IQR of daily precipitation (mm) for the four models, for 1–4 January 2016.

5.4 Discussion

5.4.1 RFSI Performance

In the synthetic case study, OK performed the best because the realities were created using OK simulation. Note that the accuracy of OK was probably overestimated because we ignored semivariogram estimation error. The effect of that error may be substantial in case of small sample sizes (Webster and Oliver 2007, Chapter 6). IDW, RFsp and RFSI had similar performance and were slightly worse

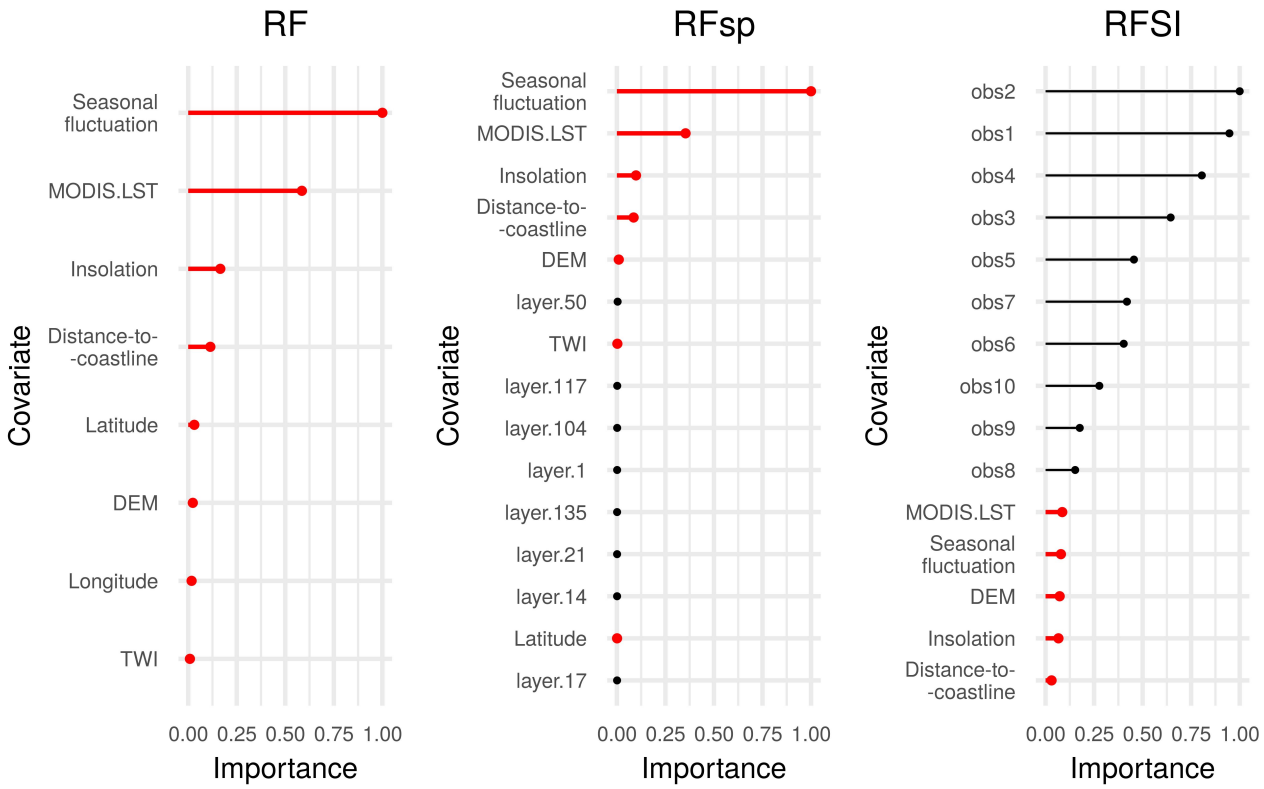


Figure 5.17: Covariate importance plot for RF (left), RFsp (middle) and RFSI (right), for the temperature case study. The importance index is scaled to a maximum of 1. The importance of environmental covariates is shown in red.

Table 5.7: Accuracy metrics of all six prediction methods as assessed using nested 10-fold LLOCV for the temperature case study. Note that accuracy metrics for STRK are taken from Hengl et al. (2012).

Method	$R^2_{1:1}$ [%]	CCC	MAE [mm]	RMSE [mm]
STRK	91.0	n/a	n/a	2.4
IDW	95.0	0.974	1.2	1.8
RF	95.7	0.978	1.1	1.6
RFsp	95.5	0.976	1.1	1.6
RFSI	96.6	0.983	1.0	1.4
$RFSI_0$	94.9	0.974	1.2	1.8

than OK. Worse performance of IDW compared to OK in synthetic case studies was also found in Zimmerman et al. (1999), MacCormack et al. (2013), and Nevtipilova et al. (2014).

IDW, RFsp, and RFSI performed differently for different semivariogram nugget-to-sill ratios, ranges, and sample sizes (Figure 5.5). IDW outperformed RFsp and RFSI in the case of low nugget. This might be because in the synthetic case, where realities are simulations from normally distributed stationary random fields, the best interpolator (i.e., OK) is linear. This indicates that non-linear interpolators, such as RFsp and RFSI, have no clear advantage over linear interpolators, such as IDW.

IDW weights are large for near observations, which is the best strategy in case of strong spatial autocorrelation. This explains why IDW performs well in case of a zero nugget. IDW performance deteriorates if the nugget-to-sill ratio is large, because in such case IDW assigns too much weight to

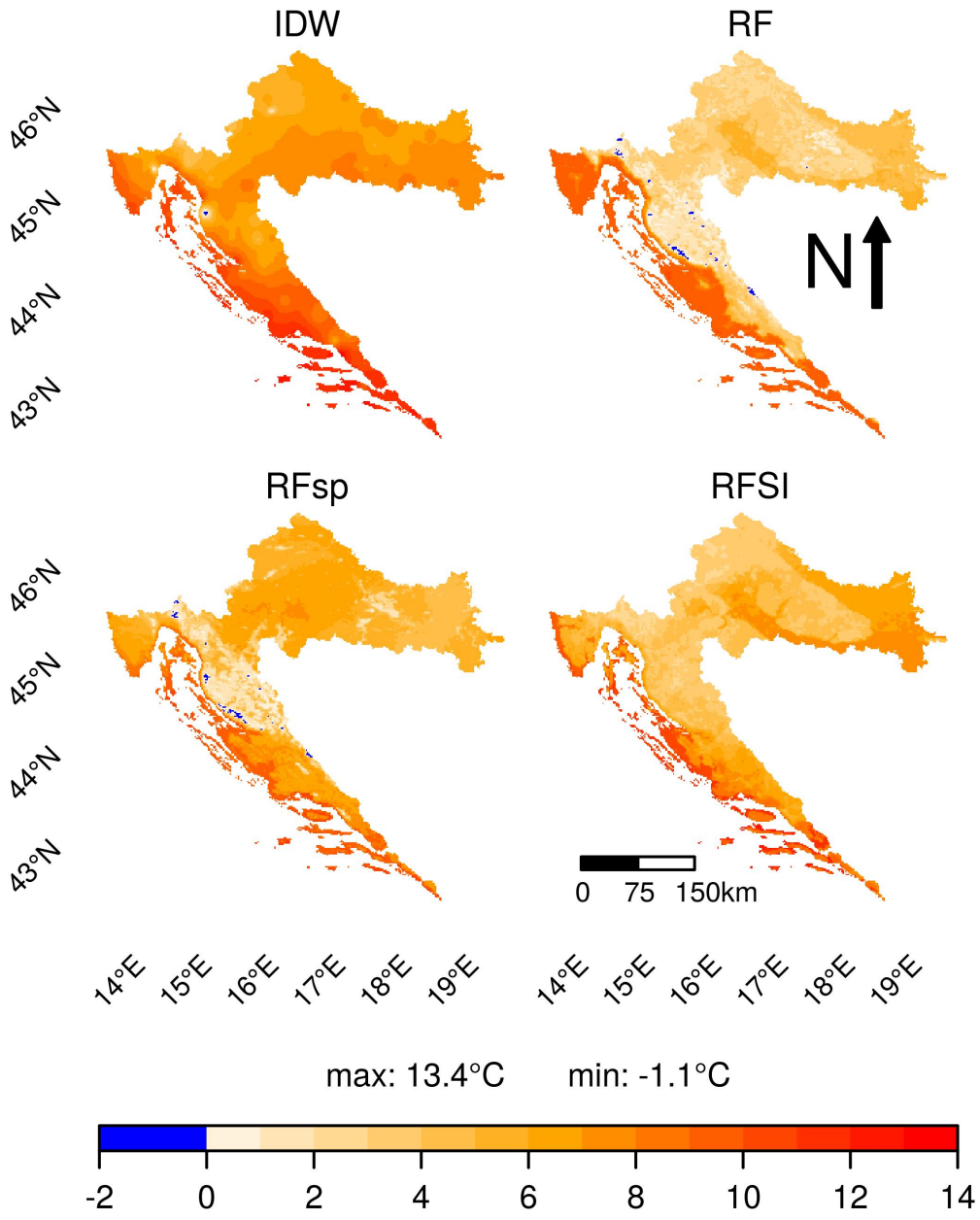


Figure 5.18: Prediction maps of daily temperature (°C) for the four models, for 2 February 2008. The bottom row shows the maximum and minimum observed temperature.

near observations. This effect is strongest in case of a large semivariogram range, because in such case distant observations carry more information than when the semivariogram range is small. RFsp and RFSI are better able to incorporate the effect of a large nugget-to-sill ratio, but only when the sample size is sufficiently large, so that there are enough calibration data to train the model.

Figure 5.5 also shows that NN has the worst performance if the nugget-to-sill ratio and semivariogram range are large, because in this case the nearest observation captures only a small part of the available information. It was already noted in Section 5.3.1 that TS has poor performance compared to other methods in case of large sample sizes, because it has only a few global parameters and hardly benefits from the extra information in large sample datasets. However, in case of a large nugget-to-sill ratio, it still outperforms NN because in such case short-distance spatial variation (i.e.,

"noise") affects NN much more than TS.

When comparing IDW, RFsp and RFSI, the advantage of RFSI over RFsp is its computational speed, especially in case of large datasets (Table 5.1), while its advantage over IDW is that environmental covariates can be added to the model.

The influence of the n parameter, that is the number of nearest locations, was also evaluated on the performance of RFSI in the synthetic case study. The optimal value of n depended on the degree of spatial correlation and the sample size. Optimal values of n were large when the sample size and semivariogram range were large (Figure 5.10). Figure 5.10 also shows that the effect of n was not that large, provided it is not too small, because after an initial decrease the RMSE was fairly constant. We therefore recommend that initial values for tuning the n parameter are between 5 and 35. If n is not tuned, a value of 25 seems sufficient. Clearly, this specific needs to be further researched, but our results suggest that extending the number of nearest locations to more than 25 will not improve the results significantly, because the added value of extra neighbours becomes smaller and smaller as new neighbours are added. Similar results are found in kriging, where limiting the local search neighbourhood to the nearest 25 or 50 observations is often done to save computing time. This hardly deteriorates the kriging prediction accuracy because kriging weights quickly converge to zero when there are many other observations closer to the prediction location (Webster and Oliver 2007, Chapter 8). In fact, we observed a similar effect in the RFSI importance plots (Figures 5.8, 5.13, and 5.17).

In the real-world case studies, the reality was not simulated from a geostatistical model as in the synthetic case, which means that STRK does not have to be the best interpolation method (Makridakis et al. 2018). Observations and distances to the nearest locations showed to be valuable spatial covariates for RFSI. RFSI combined with other environmental covariates (e.g., IMERG, MODIS LST) significantly improved prediction performance, mainly because standard RF did not capture all spatial and spatio-temporal correlation. Furthermore, RFSI outperformed STRK and RFsp.

In the precipitation case study, IDW had similar performance as RFSI, and outperformed all other methods, including STRK and RFsp. Malamos and Koutsoyiannis (2016), Liao et al. (2018), Qiao et al. (2019), and Long et al. (2020) compared OK and IDW (among other methods) and also reported that IDW had similar performance and sometimes outperformed kriging in real-world case studies. Note also that the number of environmental covariates was fairly small in this study and did not add much information in cases in which neighbourhood observations were available. Thus, interpolation methods that make use of environmental covariates did not benefit much in this case study.

In the temperature case study, RFSI outperformed IDW because in this case the environmental covariates were more important than in the precipitation case study. But IDW was better than STRK, possibly because STRK was limited to a separable covariance model. RFSI was also better than RFsp, which confirmed that there are cases where using the nearest observations as covariates in RF has truly added value.

Comparison of RFSI and RFSI_0 showed that adding environmental covariates did not increase performance much in the precipitation case study, while it did improve prediction accuracy considerably in the temperature case study. The difference lies in whether the environmental covariates have added value to the information already provided by the nearest observations. In the precipitation case, neighbouring observations and their distances alone already explained a large part of the variation, after which adding environmental covariates had little added value. Note, however, that this does not mean that the environmental covariates carry no information about precipitation. The results of the standard RF model show that they do have value, because the $R^2_{1:1}$ of RF was 49.4% (Table 5.4). But the small difference of less than 1% in the $R^2_{1:1}$ of RFSI with and without environmental covariates shows that environmental covariates were no longer important once neighbouring observations were available, as confirmed by the covariate importance plot (Figure

5.13). For the temperature case study, neighbouring observations were also more important than environmental covariates, but less so than in the precipitation case study. Including environmental covariates could still improve performance considerably (Table 5.7). This shows that it is useful to include environmental covariates as well as nearest observations and their distances in RFSI. Depending on the case, RFSI will determine from the training data which of the two information sources is most important and make predictions based on that.

Spatial interpolation methods tend to smooth the reality because both linear and non-linear averaging of observations produce predictions that on average are closer to the mean of the observations and miss the extremes. A typical example of this is OK, which produces smooth maps, particularly in a case where the nugget-to-sill ratio is high, while the reality is quite noisy in that case. Predictions of RF models also have smaller variance than the observations, as confirmed by the scatter density plots shown in Figure 5.14. The more accurate the spatial interpolation method, the closer the predictions are to the observations and the less smoothing will occur. Thus, in the precipitation case study IDW and RFSI had the lowest smoothing effect, and in the temperature case study RFSI had less smoothing than the other interpolation methods. While there are ways to decrease the degree of smoothing by combining interpolation and stochastic simulation (e.g. Goovaerts 2000), this comes at the expense of an increased MAE.

In summary, RFSI has a number of important advantages over STRK and RFsp:

1. RFSI is much closer to the philosophy of spatial interpolation than standard RF and RFsp. RFSI uses observations nearby in a direct way to predict at a location. RFsp uses a much more indirect way to include the spatial context in RF prediction. In fact, RFSI mimics kriging much more than RFsp, with the additional advantage that it is not restricted to a weighted linear combination of neighbouring observations.

2. Compared to kriging, RFSI is easier to fit, because there is no need for semivariogram modelling and stringent stationarity assumptions.

3. RFSI provides a model with more interpretative power than RFsp, i.e., the importance of the first, second, third, etc., nearest observations can be assessed and compared with each other (Figure 5.8) and with the importance of environmental covariates (Figures 5.13 and 5.17). RFsp variable importance shows how important buffer distances from observation points are, but this is difficult to interpret, because it is unclear why certain buffer distance layers have high importance and others do not. However, it should be noted that feature importance is difficult to measure objectively in cases where covariates are cross-correlated and their influence may be masked by other covariates.

4. RFSI has several orders of magnitude better scaling properties than RFsp. In RFsp the number of spatial covariates equals the number of observations, whereas in RFSI it is optimized and fairly independent of the number of observations.

5. Hengl et al. (2018) recommended using RFsp for fewer than 1000 locations. For more than 1000 locations RFsp becomes slow because buffer distances cannot be computed quickly (Table 5.1). The calculation of spatial covariates needed to apply RFSI, (Euclidean) distances and observations to the nearest locations, is not computationally extensive.

6. RFsp cannot be spatially cross-validated properly, i.e., with nested LLOCV. Considering that in nested LLOCV entire stations are held out, the buffer distance covariates in the test dataset (consisting of one main fold) and nested folds of the calibration dataset (consisting of the other folds) are not the same. Therefore, RFsp hyperparameters tuned on the nested folds with one set of buffer distance covariates can be a poor choice to make predictions on the test dataset.

5.4.2 Extensions and Improvements

RFSI predicts in the sample dataset value domain. This can be a disadvantage in the case of new observations that are out of the sample dataset value domain. This is a well-known extrapolation problem of RF (Hengl et al. 2018, Figure 14; Behrens et al. 2018; Hashimoto et al. 2019). Another similar potential problem relates to distances. When predicting at a location where distances to the nearest observations are smaller or larger than the distances used to develop the RFSI model, the prediction will be made in the same way as for the lowest or largest distance to the nearest observation. A solution for these problems would be to fit the RFSI model again or to fit extra trees to the RFSI model with the new observations and distances. Furthermore, the spatial sampling design may be optimised for RFSI (Wadoux et al. 2019).

The distances to the nearest observations had low importance in RFSI. It seems that distances to the nearest locations are still not used optimally in RFSI. They were always significantly less important than observations at nearest locations, possibly because distance information is indirectly incorporated in the order of the observations. Possible improvements could be to not only consider Euclidean distance, but also take direction into account (anisotropy) and local observation density.

Currently, RFSI is a methodology for spatial interpolation, even though it can be applied to spatio-temporal data, as was done in the real-world case studies. Future work may be oriented to the extension of RFSI to the space-time domain by including the nearest temporal observations and temporal distances as covariates. Some temporal covariates, such as day of year — DOY (He et al. 2016), cumulative day from a date — CDATE (Hengl et al. 2018), and month of the year (Mohsenzadeh Karimi et al. 2018) were already used in RF models and gave good results. Another possible improvement of RFSI could be the use of ensemble ML techniques, e.g., SVM and RF could be combined for classification and regression problems. Ensemble ML tends to perform at least as well as the best ML algorithm in the ensemble (Davies and van der Laan 2016).

Finally, the main goal of this research was not to mimic kriging, but to develop a different method that might outperform kriging in cases where the kriging assumptions are violated. More case studies are needed to evaluate the general performance of RFSI, however the three case studies in this research provide sufficient evidence that RFSI has merit. No one-size-fits-all algorithm exists. The choice of the optimal method for spatial interpolation depends on the case study, spatial structure, and the behaviour of the data and covariates. Thus, there is much to say for having a large variety of interpolation methods to choose from, and we have confidence that RFSI is a valuable extension of the spatial interpolation toolbox.

5.5 Conclusions

In this study, a novel spatial interpolation method, RFSI, was introduced. It was shown that it can produce accurate spatial interpolation results. RFSI prediction maps had higher accuracy than simple deterministic interpolation methods such as nearest neighbour and trend surfaces interpolation, and were generally comparable to or performed better than kriging, IDW, RF, and RFsp. Nearest observations and distances to nearest observations are of great value for RFSI. An initial hypothesis of this research, that RFSI can identify an optimal combination of nearest observations for prediction at unknown locations, was shown to be correct. Unlike kriging, RFSI is not limited to using only linear combinations of observations. RFSI has no stringent stationarity assumptions and can model non-linearity between covariates and the target variable. This makes it suitable for modelling complex variables with zero-inflated and skewed distributions and in cases where the stationarity condition is not satisfied. Furthermore, RFSI can be used to investigate the importance of nearest

observations by specifying their variable importance, which is difficult with existing RF methods for spatial interpolation. There is still room for improvement, especially in including distances in a more direct way and incorporating a temporal component into the RFSI.

Chapter 6

Spatial and spatio-temporal interpolation of daily climate elements for Serbian territory at 1 km spatial resolution[1]

In this study, the first daily gridded meteorological dataset at a 1-km spatial resolution across Serbia for the 2000–2019 period, named MeteoSerbia1km, was produced. The dataset consists of five daily variables: maximum, minimum and mean temperature, mean sea level pressure, and total precipitation. In addition to daily summaries, monthly and annual summaries, daily, monthly, and annual long term means were produced. Daily gridded data were interpolated using the Random Forest Spatial Interpolation methodology based on using nearest observations and distances to them as spatial covariates, together with environmental covariates to make a random forest model. The accuracy of the MeteoSerbia1km daily dataset is assessed using nested 5-fold leave-location-out cross-validation. All temperature variables and sea level pressure showed high accuracy, whereas the accuracy of total precipitation was lower, due to its nature. MeteoSerbia1km was also compared with the E-OBS dataset with a coarser resolution: both datasets showed similar coarse-scale patterns for all daily meteorological variables, except for total precipitation. As a result of its high resolution, MeteoSerbia1km is suitable for exhaustive environmental analyses.

6.1 Background & Summary

Daily meteorological observations are available from various sources, such as GHCN-daily (Menne et al. 2012), GSOD[2], ECA&D (Klein Tank et al. 2002), and OGIMET[3]. However, there is no information from these sources on daily meteorological variable values at unobserved locations, and so gridded meteorological datasets are made. Daily gridded meteorological datasets are essential input for numerous models and analyses across various research fields. For example, daily meteorological gridded dataset are used in agriculture for yield estimation (Marshall et al. 2018; Lin et al. 2020), occurrence of insect pests and disease (Juran et al. 2020), and crop growth (de Wit and van Diepen 2008), in meteorology (Haslinger et al. 2014), hydrology (Lee et al. 2019), ecology (Abatzoglou 2013), climate and climate change (Sippel et al. 2020), risk assessment (Petritsch and Hasenauer 2014), and forestry (McAlpine et al. 2018).

[1]Based on article: Sekulić, A., Kilibarda, M., Protić, D., & Bajat, B. (2021?) A high-resolution daily gridded meteorological dataset for Serbia made by Random Forest Spatial Interpolation. *Under review. Submitted to Scientific Data.*

[2]https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod

[3]https://www.ogimet.com/

Various sources of daily gridded meteorological datasets exist on global and regional levels which cover the territory of Serbia. MODIS LST (Wan 2006), TRMM / IMERG (Huffman et al. 2014), and PERSIANN (Nguyen et al. 2019), at spatial resolutions of 1 km, 0.1 degrees (∼10 km), and 0.04 degrees (∼4 km), respectively, are datasets made by algorithms based on remote sensing products. Climate Prediction Center global temperature (PSL 2020a) and precipitation (PSL 2020b), E-OBS (Cornes et al. 2018), and CarpatClim (Szalai et al. 2013) are station-based datasets, where CPC datasets are global at a spatial resolution of 0.5 degrees (∼50 km), whereas E-OBS covers the whole of Europe and CarpatClim covers 500 000 km2 in Europe, both at a spatial resolution of 0.1 degrees (∼10 km). However, with respect to Serbia, CarpatClim covers only its northern part. The third group of daily gridded meteorological datasets are reanalysis products. Some of the global NOAA products are NCEP/NCAR (Kalnay et al. 1996) and NOAA-CIRES 20th Century Reanalysis (Compo et al. 2011) at a spatial resolution of 2.5 degrees (250 km). ERA-Interim (Dee et al. 2011) is an ECMWF reanalysis dataset at a spatial resolution of 80 km, covering the period from 1979. ECMWF also provides an ERA5 hourly reanalysis dataset (Muñoz Sabater 2019) for the same time period as ERA-Interim, but at a finer spatial resolution (0.25 degrees), which can be aggregated to a daily dataset. The most of daily gridded datasets on global and regional levels produced at coarser spatial resolution can hardly represent localized meteorological patterns, which is their main limitation. MODIS LST is at finer spatial resolution (1 km), but daily products do not cover the entire spatial domain. Therefore, there is a need for localised meteorological gridded datasets at finer spatial resolutions. High-resolution daily gridded meteorological datasets are available for other regions (Hutchinson et al. 2009; Herrera et al. 2012; Xavier et al. 2016; Yanto et al. 2017; Nashwan et al. 2019; Werner et al. 2019; Razafimaharo et al. 2020), but this is the first one that refers to Serbia.

With this in mind, we developed the MeteoSerbia1km dataset, the first daily gridded meteorological dataset at a 1-km spatial resolution across Serbia, for the 2000–2019 period. The MeteoSerbia1km dataset consists of daily maximum, minimum and mean temperature (Tmax, Tmin, Tmean), mean sea level pressure (SLP), and total precipitation (PRCP). For this purpose Random Forest Spatial Interpolation methodology — RFSI (Sekulić et al. 2020a) was used. RFSI was selected as it combines environmental covariates and observations at nearest stations to predict values at unobserved locations. Additionally, monthly and annual averages and daily, monthly, and annual long term means were made by averaging (or summing for PRCP) MeteoSerbia1km dataset. The accuracy of the MeteoSerbia1km daily grids were assessed by nested k-fold cross-validation. Because daily meteorological datasets for Serbia do not exist and there is no reference point, MeteoSerbia1km was compared with the E-OBS daily dataset at a spatial resolution of 10 km. MeteoSerbia1km was also tested with independent station observations.

As daily gridded meteorological datasets mostly cover a longer period of time, they can help in understanding the behaviour of meteorological variables in both spatial and temporal domains. The newly developed MeteoSerbia1km dataset is suitable for localized environmental and microclimate analyses, precision agriculture, forestry, regional and urban planning, hydrological analysis, risk management in Serbia. MeteoSerbia1km dataset is freely available in the GeoTIFF format. Daily products will be frequently updated. The dataset will also be promoted in the future by improving the RFSI methodology, adding additional environmental covariates and including national meteorological observations.

6.2 Methods

6.2.1 Study area

Serbia is a medium sized Southeastern European country that covers an area of 88,361 km^2, i.e., around 18% of the Balkan Peninsula (18.8 °–23.0 °E longitude, 41.8 °–46.2 °N latitude). It is characterized by a complex topography (Figure 6.1), where northern parts are within the Pannonian Plain, and southern parts are crossed with several mountain systems. The mean altitude of Serbia is 473 m, ranging from 29 m in the northeast to 2,656 m on Prokletije Mountain in the southwest (Bajat et al. 2013). There are three main types of climate in Serbia, from north to south: continental, moderate continental, and modified Mediterranean climate. Precipitation is unevenly distributed with an average amount of 739 mm, whereas the average temperature for the 1961–2010 period was 10.4 °C (Bajat et al. 2015).

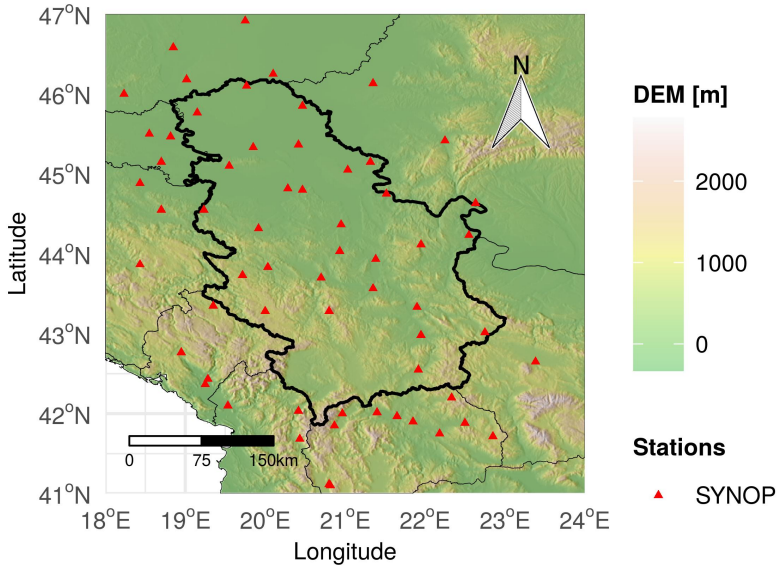


Figure 6.1: SYNOP station locations used for making MeteoSerbia1km with DEM.

6.2.2 Source data

6.2.2.1 OGIMET

OGIMET (Section 3.2.1) daily summaries from 61 SYNOP stations, wherefrom 28 are in Serbia, were used for spatial interpolation of meteorological variables (Figure 6.1). The remaining 33 stations in a 100-km buffer around the Serbian border were used for a more accurate spatial interpolation, especially in the areas near the Serbian border.

The outliers for OGIMET precipitation daily summaries were detected and removed as the observations which were four times larger than (a) the maximum of the surrounding observations, i.e. observations in a radius of 100 km and (b) the corresponding E-OBS value (see section 6.2.2.4).

Summary statistics for each of the meteorological parameters is given in Table 6.1.

6.2.2.2 DEM and TWI

DEMSRE3 and TWISRE3 (DEM and TWI) at a 1 km spatial resolution, described in Section 3.3.4, were used as environmental covariates.

Table 6.1: Summary statistics for the selected variables in OGIMET daily summaries for the 2000–2019 period.

Parameter	Tmax [°C]	Tmin [°C]	Tmean [°C]	SLP [mbar]	PRCP [mm]
Minimum	−22.2	−34.8	−24.8	967.4	0.0
1st quartile	9.7	0.5	5.0	1012.5	0.0
Median	18.3	6.9	12.3	1016.5	0.0
Mean	17.6	6.4	11.8	1017.1	2.0
3rd quartile	25.8	12.7	18.9	1021.4	1.0
Maximum	45.9	30.8	35.4	1077.8	198.0

6.2.2.3 IMERG

The IMERG (Huffman et al. 2014, Section 3.3.2.2) final run version V06B precipitation estimates were used for PRCP model development. IMERG estimates are a space-time covariate with a spatial resolution of 10 km and temporal resolution of one day. Earlier versions of the IMERG dataset, based on GPM, were covering the 2014–present period, but starting from version V06B, IMERG includes TRMM preprocessed data going back to June 2000. Herein, the IMERG dataset was used as a coarser scale covariate for precipitation. Therefore, the IMERG dataset was resampled to a 1-km spatial resolution using bilinear interpolation and DEM as a base layer.

6.2.2.4 E-OBS

E-OBS (Cornes et al. 2018, Section 3.3.1.1) is an ensemble dataset constructed through a conditional simulation procedure. Because E-OBS is based on observations from ECA&D and SYNOP meteorological stations, it was used for comparison with the daily MeteoSerbia1km dataset and detection of precipitation outliers.

6.2.2.5 Automated meteorological stations in Vojvodina region

Automated meteorological station network in Vojvodina region (AMSV)[4] collects hourly data for temperature (Tmax, Tmin, Tmean), dew point, PRCP, relative humidity, etc., mostly for the period starting from March 2005. AMSV daily summaries from 55 stations were used for independent test of MeteoSerbia1km in Vojvodina region, specifically Tmax, Tmin, Tmean, and PRCP.

6.2.3 RFSI

RFSI (Sekulić et al. 2020a, Section 5.2.1.2) is a novel methodology for spatial interpolation based on the random forest machine learning algorithm (Breiman 2001). In comparison with other random forest models for spatial interpolation, RFSI uses additional spatial covariates: (1) observations at n nearest locations and (2) distances to them, in order to include spatial context into the random forest. RFSI model predictions can be written as:

$$\hat{z}(s_0) = f(x_1(s_0), \ldots, x_m(s_0), z(s_1), d_1, z(s_2), d_2, z(s_3), d_3, \ldots, z(s_n), d_n) \tag{6.1}$$

where $\hat{z}(s_0)$ is the prediction at prediction location s_0, $x_i(s_0)(i = 1, \ldots, m)$ are environmental covariates at location s_0, $z(s_i)$ and d_i are spatial covariates $(i = 1, \ldots, n)$, where $z(s_i)$ is the i-th

[4] http://www.pisvojvodina.com/Shared_Documents/AMS_pristup.aspx

nearest observation from s_0 at location s_i and $d_i = |s_i - s_0|$. These spatial covariates proved to be valuable extensions for the random forest algorithm in improving spatial accuracy. A detailed description of RFSI, performance and implementation procedure is provided by Sekulić et al. (2020a).

6.2.3.1 Model development and prediction

In order to prepare the data for RFSI modelling, all of the environmental covariates were overlaid with training observation locations, for each day. Then, RFSI spatial covariates were created in the following way: for each day and for each training observation location, n nearest training observation locations were found and n pairs of covariates—observations at n nearest locations and distances to them—were calculated. Extracted overlaid values and n pairs of spatial covariates were assigned to the corresponding observations making a dataset which was then used to fit an RFSI model.

Predictions were made in a similar manner as the development of the RFSI model. For each of the desired prediction days and locations (in this case pixels of the target grid), environmental covariates were extracted and observations at n nearest training locations and distances to them were calculated. Then, predictions were made using extracted values and n pairs of spatial covariates and an already fitted RFSI model. The entire process of making an RFSI model and prediction is already presented in Section 5.2.1.2 (Figure 5.1). It should be noted that the RFSI model can handle both regression and classification tasks.

6.2.3.2 Model tuning

In order to achieve the best possible prediction accuracy, hyperparameters for the RFSI models were tuned. The tuned hyperparameters were the number of variables to possibly split at each node (`mtry`), minimal node size (`min.node.size`) and ratio of observations-to-sample in each decision tree (`sample.fraction`), and the number of nearest observations (`n.obs`). The number of trees (`ntree`) hyperparameter was fixed and set to 250, according to Sekulić et al. (2020a), as a larger value of `ntree` would not improve the RFSI model accuracy. The splitrule (`splitrule`) hyperparameter was also fixed and set to be `variance` for regression tasks, and `gini` index for a classification task.

The hyperparameters were tuned using 5-fold leave-location-out cross-validation. Here "leave-location-out" means that observations from one station (location) were in the same fold. By doing so, the targeted spatial prediction accuracy was assessed (Meyer et al. 2018). Many different combinations of hyperparameters were tested and for each combination, 5-fold LLOCV was performed. In other words, for each of the hyperparameter combinations, the entire dataset was split into 5 folds. Each of the folds once represented a test fold, while the four remaining folds were used to fit the RFSI model with a hyperparameter combination. Finally, RMSE was adopted as a criterion for the selection of optimal hyperparameters.

6.2.4 Modelling of daily meteorological variables

6.2.4.1 Temperature

Modelling of daily temperature variables, Tmax, Tmin, and Tmean, is a pure regression task. All daily temperature RFSI models are as follows:

$$T_{max,min,mean}(s_0) = f_R(DEM, TWI, GTT, DOY, IDW, z(s_1), d_1, \ldots, z(s_{n.obs}), d_{n.obs}) \quad (6.2)$$

where $T_{max,min,mean}(s_0)$ is the daily temperature (Tmax, Tmin, and Tmean) prediction at prediction location s_0, f_R denotes an RFSI regression model, GTT is the geometric temperature trend, a function of latitude and day of the year (which was shown to be a valuable covariate for Tmax, Tmin and Tmean) (Kilibarda et al. 2014), DOY is a temporal covariate, i.e., the day of the year, IDW is a local inverse distance weighting prediction based on $n.obs$ number of nearest observations (excluding observed location).

The tuned hyperparameters for each of the daily temperature models are given in Table 6.2. IDW exponent (p) was also tuned. The $n.obs$ hyperparameter was 10 for Tmax and 9 for Tmin and Tmean models.

Table 6.2: Optimized hyperparameters for each of the daily meteorological variables.

Variable	mtry	min.node.size	sample.fraction	n.obs	p
Tmax	7	15	0.98	10	2.9
Tmin	4	11	0.93	9	2.2
Tmean	7	14	1.00	9	3.0
SLP	6	11	0.91	9	3.5
PRCP classification	3	2	0.70	9	n/a
PRCP regression	7	11	0.93	6	3.3

6.2.4.2 Sea level pressure

Modelling of daily SLP was also a pure regression task. The SLP RFSI model has fewer covariates than corresponding temperature models:

$$SLP(s_0) = f_R(DEM, DOY, IDW, z(s_1), d_1, \ldots, z(s_9), d_9) \tag{6.3}$$

where $SLP(s_0)$ is the daily SLP prediction at prediction location s_0.

The tuned hyperparameters for daily SLP model are given in Table 6.2. The $n.obs$ hyperparameter was 9.

6.2.4.3 Precipitation

PRCP was modelled in two steps, i.e. with two models: (1) classification model for daily precipitation occurrence and (2) regression model for daily precipitation amount, denoted as:

$$\begin{aligned} PRCP(s_0) = f_C(DEM, T_{max}, T_{min}, SLP, IMERG, DOY, z(s_1), d_1, \ldots, z(s_9), d_9) \cdot \\ f_R(DEM, T_{max}, T_{min}, SLP, IMERG, DOY, IDW, z(s_1), d_1, \ldots, z(s_6), d_6) \end{aligned} \tag{6.4}$$

where $PRCP(s_0)$ is the daily PRCP prediction at prediction location s_0, f_C denotes the PRCP RFSI classification model with 0 and 1 as possible classes, $Tmax$, $Tmin$, and SLP are corresponding daily predictions from the MeteoSerbia1km dataset at location s_0, and $IMERG$ is the corresponding overlayed value from the IMERG dataset at location s_0. Both precipitation models were fitted on the entire dataset with the same covariates. This means that zero precipitation observations were included in the regression model fitting. A reason for this was to include zero precipitation proximity into the regression model. As seen from Eq. 6.4, in PRCP prediction, the regression model was applied only in the locations where the classification model predicted the precipitation occurrence (class 1).

The tuned hyperparameters for both daily PRCP classification and regression models are given in Table 6.2. The $n.obs$ hyperparameter for the classification model was 9, and for the regression model was 6.

6.3 Data Records

MeteoSerbia1km is a high-resolution daily meteorological gridded dataset for Serbia, consisting of Tmean, Tmax, Tmin, SLP and PRCP variables, for the 2000–2019 period. As an example, prediction maps for July 27, 2014 are presented in Figure 6.2. In addition, monthly and annual averages (totals for PRCP) were generated by aggregating daily datasets. Then, daily, monthly, and annual LTM are generated by averaging daily, monthly and annual datasets. Since the first five months of the year 2000 were missing from the IMERG dataset, the daily and monthly PRCP averages start from June, 2000. Furthermore, the daily and monthly PRCP LTMs were calculated without the first five months of the year 2000, and PRCP annual averages and LTM were calculated without the year 2000. Additionally, only the data for leap years were available for generation of daily LTM for February 29.

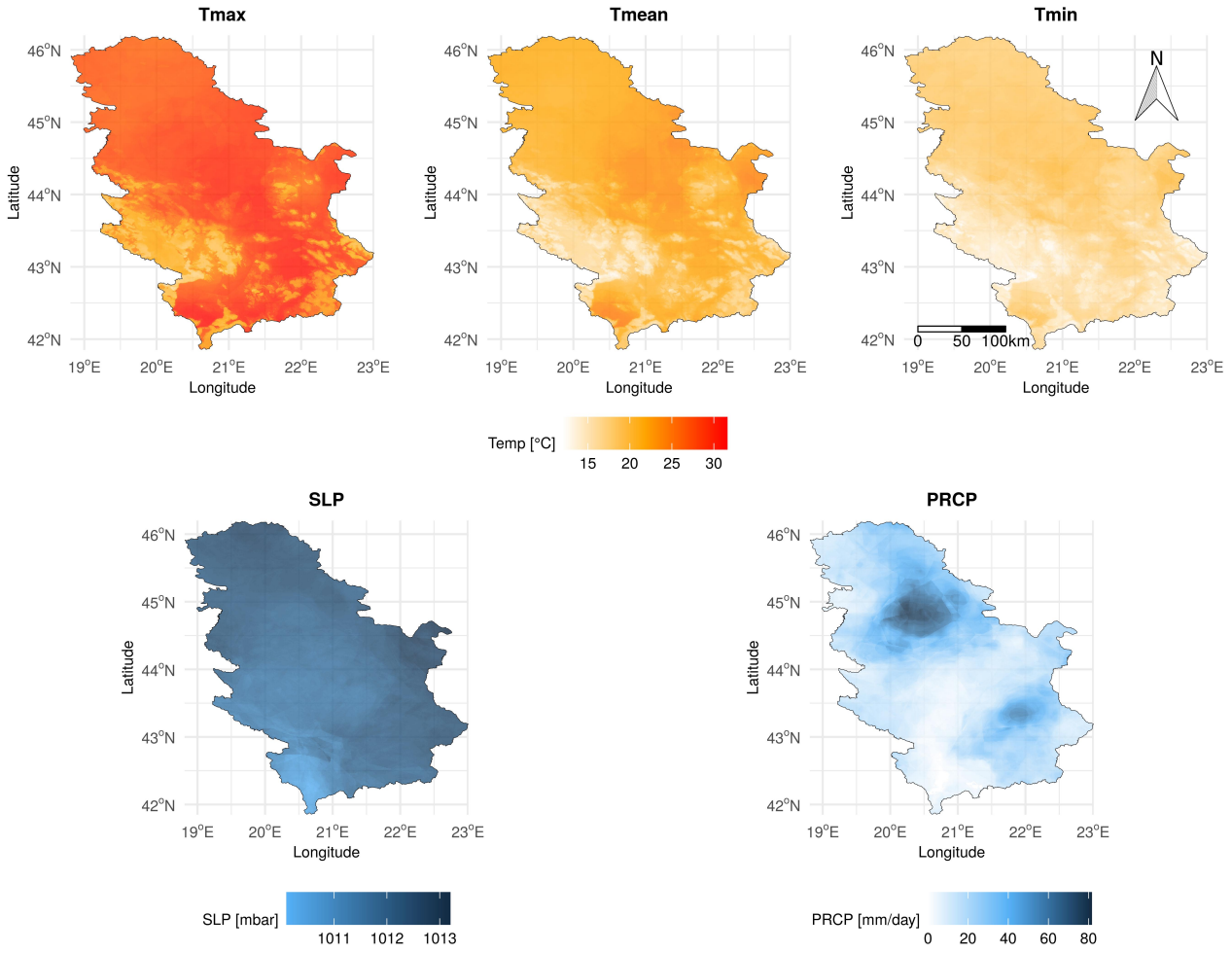


Figure 6.2: Prediction maps for all daily meteorological variables, for July 27, 2014.

The OpenStreetMaps country border[5] of Serbia was used to ensure that the MeteoSerbia1km dataset covers the territory of Serbia. The entire dataset is at a 1-km spatial resolution, and is avail-

[5]https://osm-boundaries.com/

able in both, WGS84 and UTM34N projections. The dataset is stored in the GeoTIFF (`.tif`) format. Units of the dataset values are

- temperature (Tmean, Tmax, and Tmin) - tenths of a degree in the Celsius scale (°C)
- SLP - tenths of a mbar
- PRCP - tenths of a mm

Furthermore, all dataset values are stored as integers (`INT32` data type) in order to reduce the size of the GeoTIFF files, i.e., temperature values should be divided by 10 to obtain degrees Celsius, and the same for SLP and PRCP to obtain millibars and millimeters.

The adopted file naming convention is provided in Table 6.3. It should be noted that the naming convention is different for different products with different temporal resolutions.

Table 6.3: MeteoSerbia1km dataset file naming convention.

Product	File nomenclature	Example
Daily averages	var_{time period}_{yyyymmdd}_{proj}.tif	tmax_day_20000101_wgs84.tif
Monthly averages	var_{time period}_{yyyymm}_{proj}.tif	tmax_mon_200001_wgs84.tif
Annual averages	var_{time period}_{yyyy}_{proj}.tif	tmax_ann_2000_wgs84.tif
Daily LTM	var_ltm_{time period}_{mmdd}_{proj}.tif	tmax_ltm_day_0101_wgs84.tif
Monthly LTM	var_ltm_{time period}_{mm}_{proj}.tif	tmax_ltm_mon_01_wgs84.tif
Annual LTM	var_ltm_{time period}_{proj}.tif	tmax_ltm_ann_wgs84.tif

The dataset can be downloaded from ZENODO Sekulić et al. (2020)[6], year by year.

6.4 Technical Validation

6.4.1 Validation of daily datasets

The daily MeteoSerbia1km dataset was validated using nested 5-fold LLOCV, which combines nested k-fold (Pejović et al. 2018) and leave-location-out cross-validation. For nested 5-fold LLOCV, similarly as for the regular 5-fold LLOCV, the entire dataset was split into five folds. Each of the folds was once used for testing, while the four remaining folds were used for hyperparameter tuning with regular 5-fold LLOCV (see the Model tuning section). Four accuracy metrics, namely, coefficient of determination (R^2), Lin's concordance correlation coefficient (Lin 1989), mean absolute error and root mean square error were calculated for all daily meteorological variables, for stations in Serbia (Table 6.4). The SLP model had the highest accuracy, especially for stations in Serbia, followed by Tmax and Tmean. This is due to the fact that SLP and temperature are continuous variables and have strong spatial autocorrelation. Tmin showed slightly lower accuracy than Tmax and Tmean, and PRCP showed the lowest accuracy, which was also reported in similar studies (Cornes et al. 2018; Dhakal et al. 2020). Furthermore, LLOCV accuracy is lower for stations outside of Serbia because of the well-known edge effect interpolation problem. Therefore, including the stations outside of Serbia into LLOCV would not give an objective accuracy assessment of the MeteoSerbia1km dataset and would even deteriorate accuracy.

Accuracy of the two-step PRCP model with a classification model and a unique PRCP regression model was the same. The advantage of the PRCP two-step-step model with classification is that zero

[6]http://doi.org/10.5281/zenodo.4058167

Table 6.4: Accuracy metrics for each meteorological variable for stations in Serbia, as assessed using the nested 5-fold LLOCV.

Variable	R^2 [%]	CCC	MAE	RMSE
Tmax	97.4	0.987	1.1 °C	1.7 °C
Tmin	93.7	0.968	1.4 °C	2.0 °C
Tmean	97.4	0.987	1.0 °C	1.4 °C
SLP	99.1	0.996	0.5 mbar	0.7 mbar
PRCP	63.8	0.784	1.1 mm	3.1 mm

PRCP values were predicted as exact zeros. Cohen's kappa coefficient (Cohen 1960) for the PRCP RFSI classification in Serbia was 0.779. The confusion matrix is shown in Table 6.5. For the case where the observed values were zero (class 0), only 4.21% of the final predicted values were larger than 1 mm and 0.44% of them were larger than 5 mm. For the opposite case where the predicted values were zero (class 0), only 3.94% of the observed values were larger than 1 mm and 0.86% of them were larger than 5 mm.

Table 6.5: Confusion Matrix for PRCP RFSI classification model from the nested 5-fold LLOCV. Class 0 represents no precipitation, and class 1 represents precipitation occurrence.

		Observation	
		0	1
Prediction	0	108248 (93.40%)	11591 (16.35%)
	1	7651 (6.60%)	59298 (83.65%)

The average RMSE per station for the entire time period is presented in Figure 6.3. Stations at the highest altitudes, Kopaonik (1711 m) and Crni Vrh (1037 m), had the largest average RMSE for all temperature variables. Additionally, Sjenica (1038 m) and Zlatibor (1029 m) had large average RMSE for Tmin which is the reason for the lower accuracy in comparison with Tmax and Tmean. On the one hand, microclimatic conditions at higher altitudes affect temperature behaviour so that overall spatial autocorrelation, and therefore the accuracy, is lower. On the other hand, the accuracy is higher at lower altitudes, especially in Vojvodina region, the northern part of Serbia. This makes temperature datasets particularly suitable for agriculture. Average RMSE for SLP is low and equally distributed on the territory of Serbia, which is confirmed by overall high accuracy (Table 6.4). Average RMSE for PRCP is also equally distributed over the territory of Serbia. Time series of predictions from nested 5-fold LLOCV and observations for the Belgrade station, for year 2014, are presented in Figure 6.4. The figure shows that differences between observations and predictions for Tmax, Tmean, and SLP are minor, whereas those for Tmin are somewhat larger, mostly because Tmin is slightly underestimated, as reflected in the lower accuracy in comparison with Tmax and Tmean (Table 6.4). For PRCP, the days without precipitation are predicted well, whereas the days with precipitation are slightly underestimated.

6.4.2 Comparison with E-OBS

The E-OBS dataset was taken as a benchmark dataset because it was made by geostatistical simulation, i.e., spatial interpolation from ECA&D stations, which also includes SYNOP stations. The daily MeteoSerbia1km dataset was aggregated to a 10-km spatial resolution in order to be comparable with the E-OBS dataset. Pearson correlation coefficients (PCC) between E-OBS and the daily MeteoSerbia1km dataset aggregated to a 10-km spatial resolution were calculated. PCCs calculated for

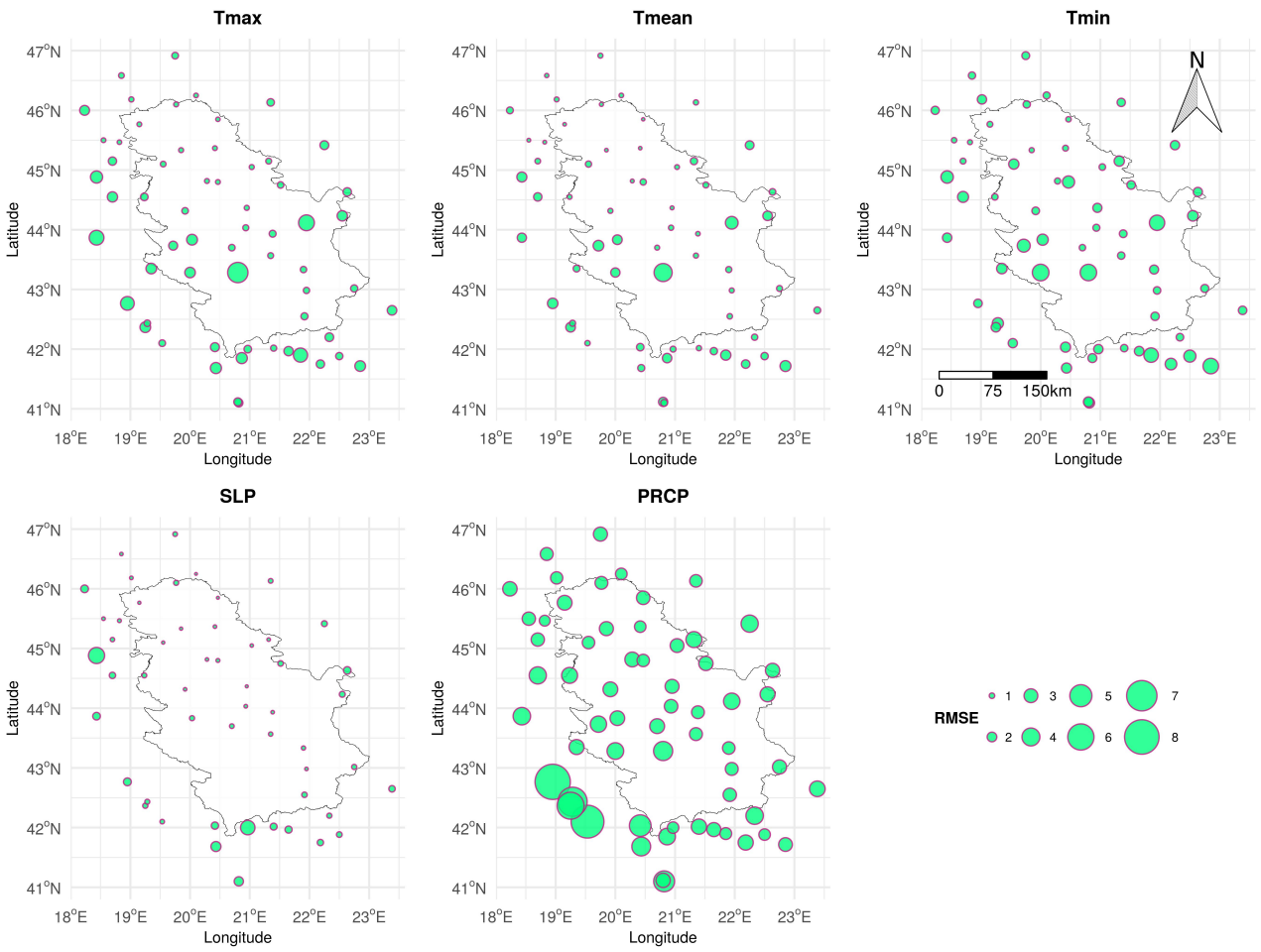


Figure 6.3: Average RMSE per station for the 2000–2019 period, calculated from the nested 5-fold LLOCV. The units are °C for temperature, mbar for SLP and mm for PRCP.

each E-OBS raster pixel for each meteorological variable are presented in Figure 6.5. The MeteoSerbia1km dataset shows an overall high correlation with the E-OBS dataset for Tmax, Tmin, Tmean, and SLP (0.992, 0.989, 0.993, and 0.922 respectively) and similar coarse-scale spatial patterns, with slightly lower correlation around stations Kopaonik and Crni Vrh (Figure 6.5) where the LLOCV accuracy was the lowest (Figure 6.3). Correlation for SLP was lower in the southwestern part of Serbia, probably because of the lack of SYNOP SLP stations in that area (Figure 6.3). The MeteoSerbia1km dataset showed the lowest correlation with the E-OBS dataset for PRCP (0.551). The main reason for this is that precipitation is a complex variable and different models can produce significantly different results. Another reason is that E-OBS methodology does not include IMERG which is an important predictor for the PRCP model and, consequently, predictions follow IMERG patterns. Bearing in mind that accuracy of MeteoSerbia1km and E-OBS PRCP models does not differ much in RMSE and MAE, RFSI PRCP can be valuable for the areas where E-OBS cannot contribute or where a finer spatial resolution of 1 km is needed. Hence, MeteoSerbia1km dataset describes local variation of daily PRCP in Serbia better than E-OBS.

6.4.3 Test with stations in Vojvodina region

MeteoSerbia1km was also tested with independent AMSV stations that were not used for making RFSI models. RMSE between AMSV stations and the corresponding MeteoSerbia1km values over Vojvodina region for the 2005–present period for Tmax, Tmin, Tmean, and PRCP was 1.6 °C, 1.8 °C

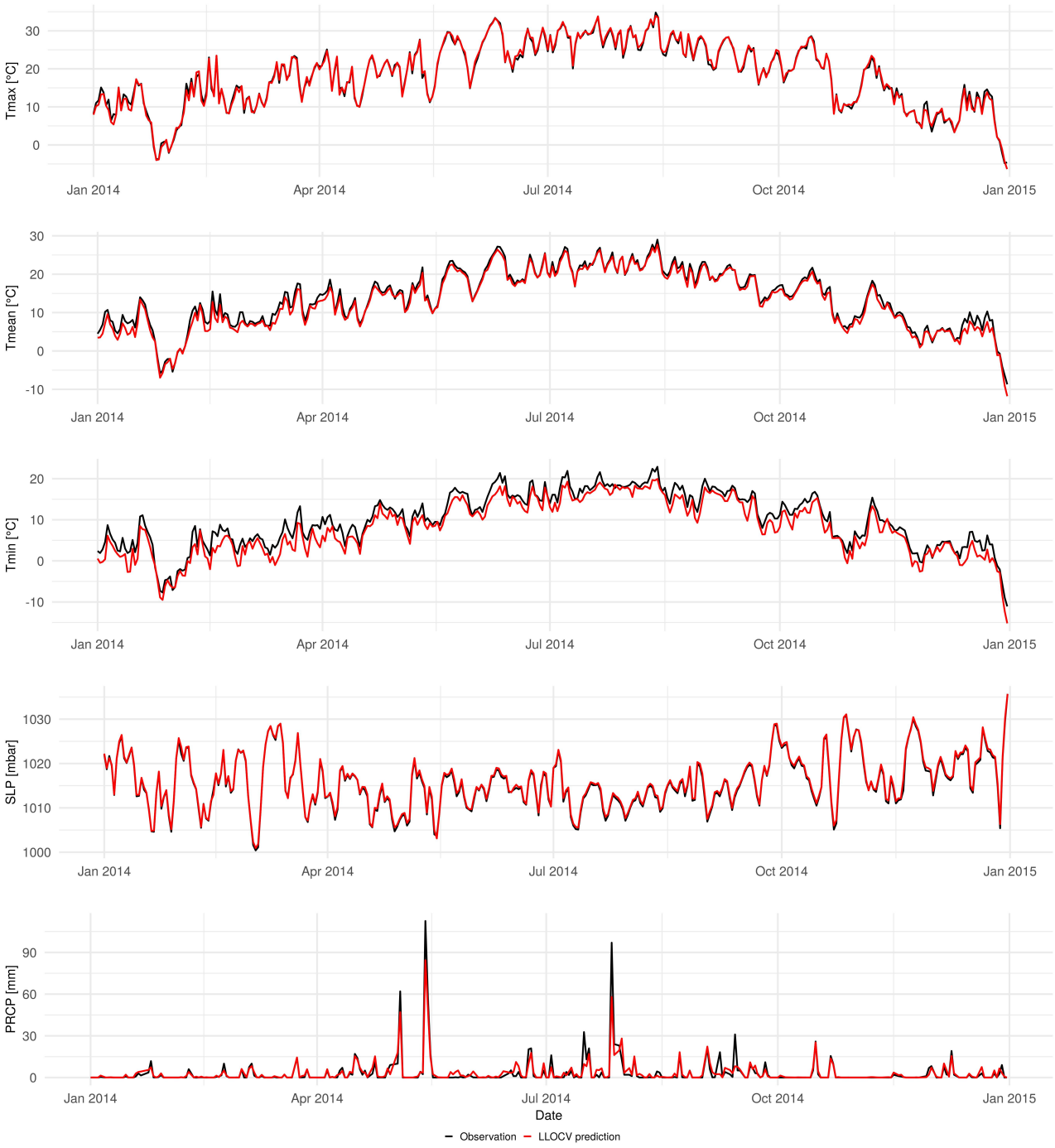


Figure 6.4: Predictions from the nested 5-fold LLOCV (red) and observations (black) for station Belgrade for year 2014.

1.2 °C and 3.7 mm, respectively. In comparison with the results from LLOCV for the entire Serbia (Table 6.4), accuracy of MeteoSerbia1km temperature variables is slightly better, while accuracy of MeteoSerbia1km PRCP is slightly worse. Lower RMSE for PRCP can be taken as a consequence of more dense network of AMSV stations than OGIMET stations and large spatial variability of PRCP.

6.5 Usage Notes

MeteoSerbia1km is the first high-resolution daily gridded meteorological dataset for Serbia at a 1-km spatial resolution. The dataset can be used in a wide range of topics such as agriculture, insurance,

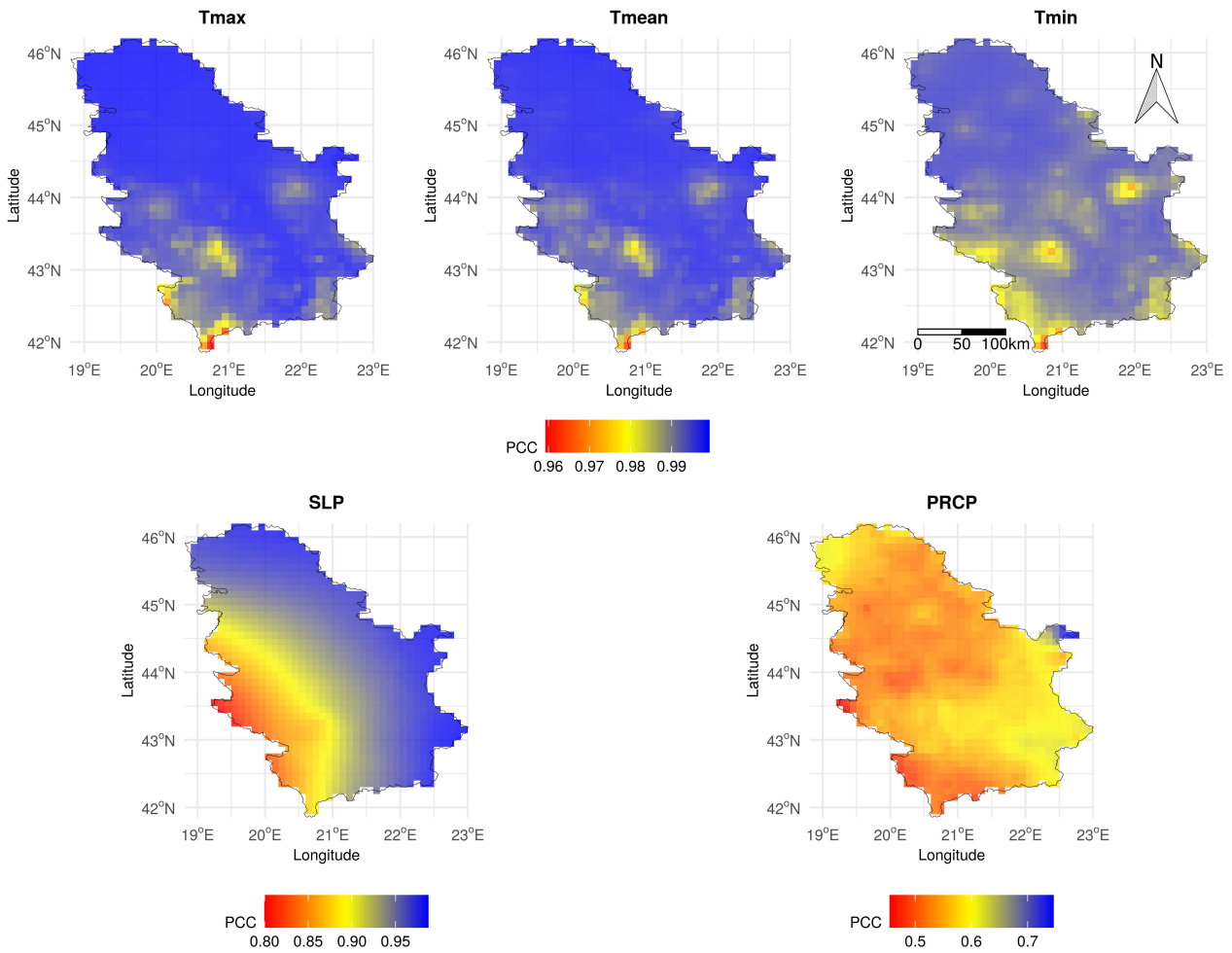


Figure 6.5: Pearson correlation coefficient map between E-OBS and the daily MeteoSerbia1km datasets for Serbia.

forestry, climatology, meteorology, hydrology, ecology, soil mapping, urban planning, or any other research field that needs gridded data with a high spatial resolution.

MeteoSerbia1km is in the GeoTIFF format which makes it interoperabile with any GIS software, such as SAGA GIS[7], QGIS[8], ArcGIS[9], etc. It should be noted that MeteoSerbia1km values are multiplied by 10, so they should be divided by 10 to obtain values in basic units (°C, mbar and mm). Finally, predictions of some days may show artifacts due to misrepresentation of meteorological stations.

The data are freely available under the Creative Commons Licence: CC BY 4.0.

6.6 Code availability

The R programming language (R Development Core Team 2012), version 3.6.1, was used for the automation of the entire process for making the MeteoSerbia1km dataset, using the following packages: `climate` (Czernecki et al. 2020), `meteo` (Kilibarda et al. 2014), `nabor` (Elseberg et al. 2012), `CAST` (Meyer 2018), `caret` (Kuhn 2019), `sp` (Pebesma and Bivand 2005; Bivand et al.

[7]http://www.saga-gis.org/
[8]http://www.qgis.org
[9]https://www.arcgis.com/

2013b), `spacetime` (Pebesma 2012; Bivand et al. 2013b), `gstat` (Pebesma 2004; Gräler et al. 2016), `raster` (Hijmans 2019), `rgdal` (Bivand et al. 2019), `doParallel` (Microsoft Corporation and Steve Weston 2019), `ranger` (Wright and Ziegler 2017), `plyr` (Wickham 2011), `ggplot2` (Wickham 2016).

To automate the development, tuning, cross-validation and prediction processes for the RFSI method, five additional R functions were created and added to the `R meteo` package (Kilibarda et al. 2014)[10][11]:

- `near.obs` - for finding n nearest observations and distances to them from desired locations,
- `rfsi` - for RFSI model fitting,
- `tune.rfsi` - for RFSI model tuning,
- `cv.rfsi` - for RFSI model cross-validation,
- `pred.rfsi` - for RFSI model prediction.

In order to make this work reproducible, a complete script in R and datasets used for modelling, tuning, validation, and prediction of daily meteorological variables are available via the GitHub repository at https://github.com/AleksandarSekulic/MeteoSerbia1km.

6.7 Discussion and conclusions

Annual LTM maps are presented in Figure 6.6. It can be seen that Serbia has high temperature variability, which Bajat et al. (2015) also concluded. Influence of regional topography, i.e. DEM, is clearly present for all of the climate elements, especially for temperature and PRCP. Temperature decreases, while SLP and PRCP increases with increase of altitude. Similar patterns are visible in comparison of MeteoSerbia1km annual LTM Tmean with one created by Bajat et al. (2015) (Figure 2). Bajat et al. (2015) created an annual LTM map for Tmean in Serbia, for the 1961–2010 period, using RK. The Tmean values (Figure 6.6) are the highest in the lowlands of Vojvodina region where Pannonian Plain dominates the relief, Velika Morava River Valley, and Kosovo Region. In the mountainous areas of the southwestern and southeastern parts of the country occupied with several mountain systems (the Carpathian, Balkan, and Rhodope Mountains), Tmean values are the lowest with weakly clustered spatial patterns. Montlhy LTM for January and July are presented in Figures 6.7 and 6.8. The annual, January (Figure 6.7) and July (Figure 6.8) Tmean, Tmax, and Tmin are showing a similar pattern of change with the highest values in the lowlands and the lowest in the mountains.

Regarding the annual LTM PRCP map, MeteoSerbia1km is comparable with the map for the 1961–1990 period, provided by Bajat et al. (2013) (Figure 9) using RK. Geographic distribution of precipitation also shows variation over the country (Figure 6.6). Contrary to the Tmean map, higher PRCP amount values are in the southern/south-western mountainous part of Serbia, while lower PRCP amount values are in the northern flat part. The wettest area is in the mountains southwest (Prokletije Mountain) and in southern parts (Šara Mountain) of the country with annual average amount of precipitation over 1,200 mm. The driest part is the northern part of Serbia which extends into the Vojvodina region with amounts of less than 600 mm a year. Western parts of the country are wetter than the eastern parts (Bajat et al. 2013). The LTM map provided by Bajat et al. (2013) is more detailed because they had more than 1000 observations. The average precipitation in January

[10] https://github.com/AleksandarSekulic/Rmeteo
[11] http://r-forge.r-project.org/projects/meteo/

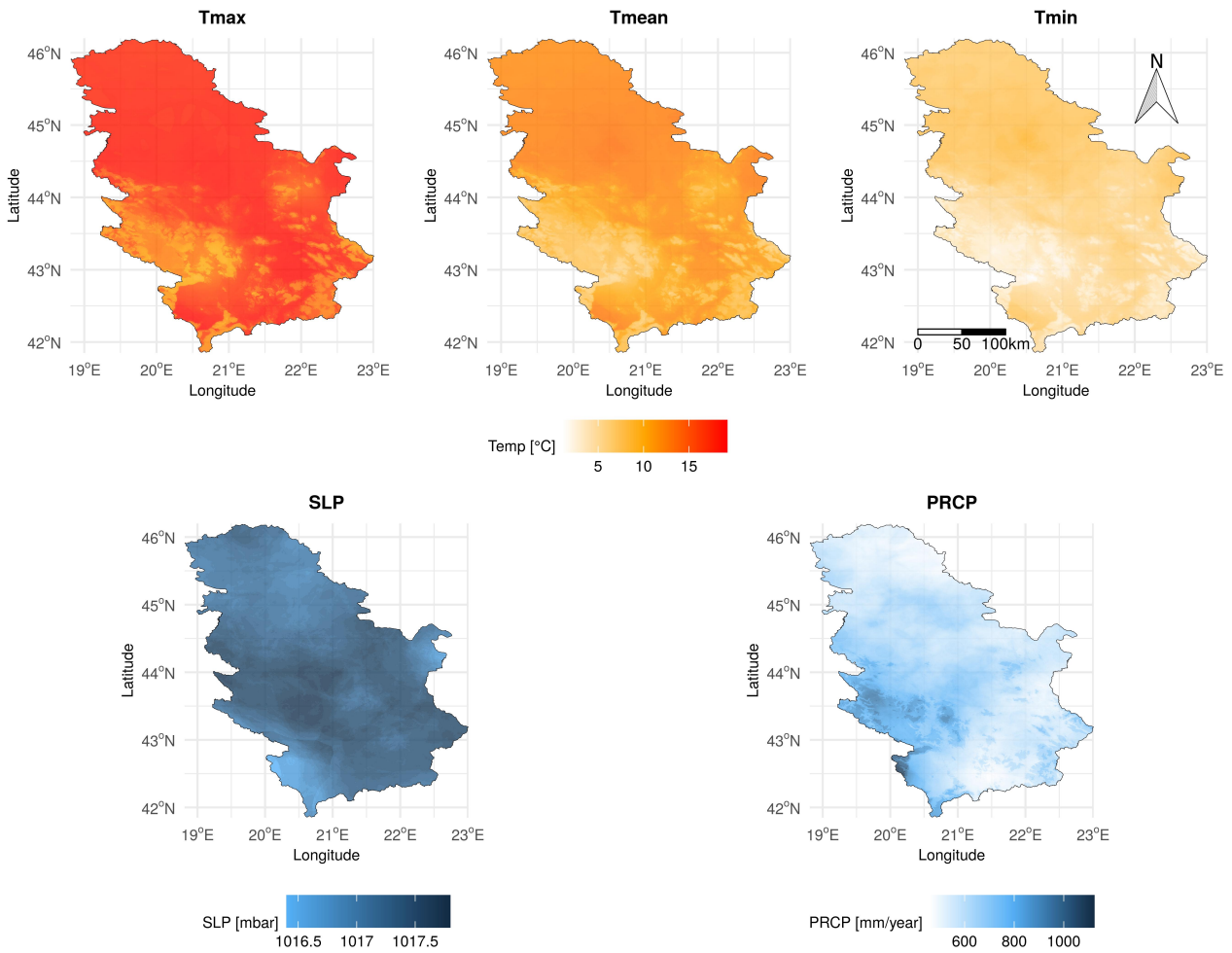


Figure 6.6: Annual LTM maps for all daily meteorological variables.

follows the pattern of annual mean precipitation (Figure 6.7) while average precipitation in July (Figure 6.8) shows opposite spatial pattern with relatively wet rainfall in the northern and central parts, and dry Kosovo region.

To conclude, MeteoSerbia1km is the first gridded dataset for daily climate elements at a 1 km spatial resolution for Serbian territory, for the twenty year period (2000–2019). The accuracy of the MeteoSerbia1km daily dataset is comparable with the regional daily gridded datasets E-OBS, at coarser 10 km spatial resolution. The accuracy of the nested 5-fold LLOCV is confirmed with the independent test with AMSV stations in Vojvodina region. MeteoSerbia1km annual LTMs shows similar spatial structure as in the previous PRCP and Tmean studies for Serbia (Bajat et al. 2013, 2015). MeteoSerbia1km will be a useful source of information for agronomists, climatologists, hydrologists, insurance companies, and others.

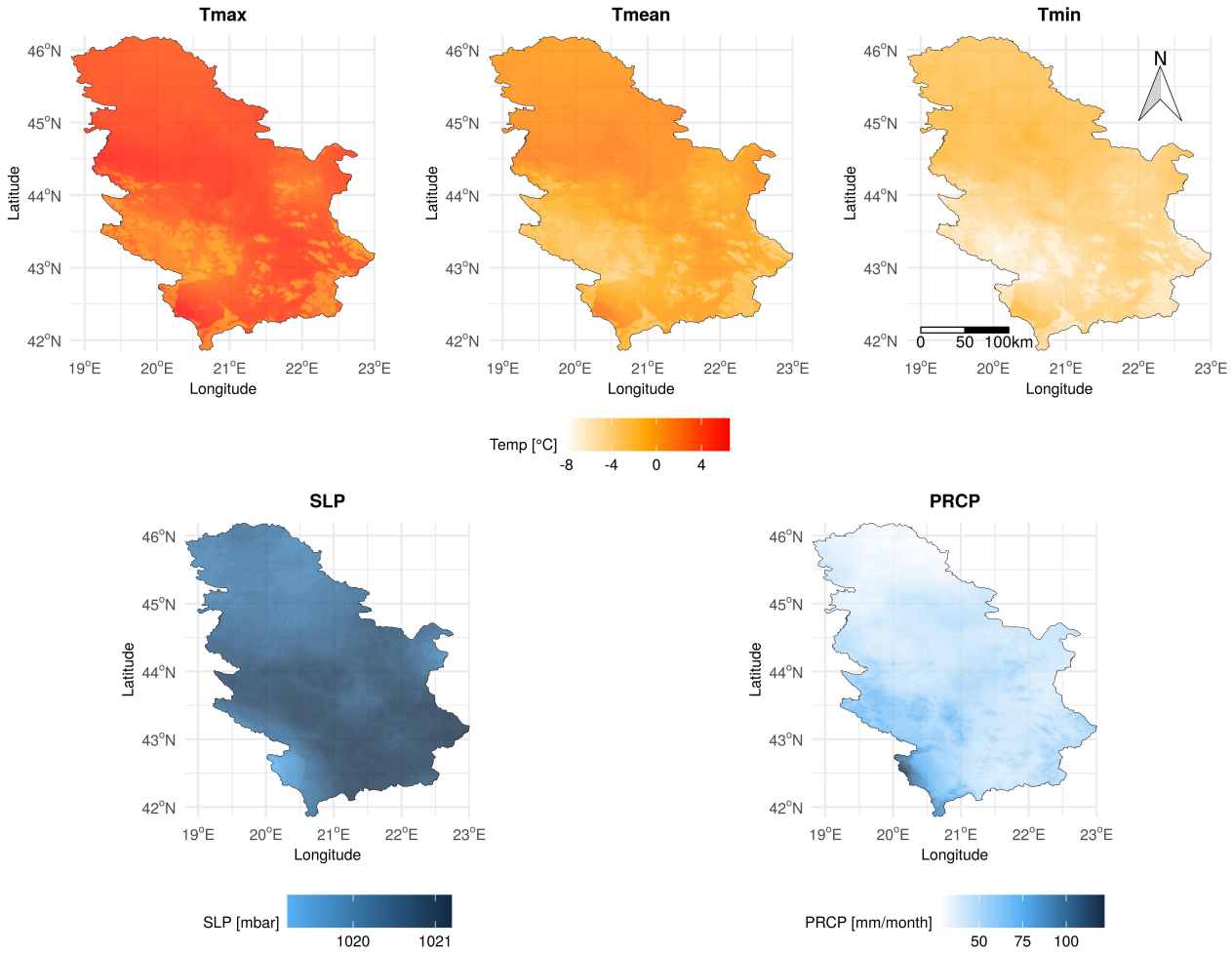


Figure 6.7: Montlhy LTM maps for all daily meteorological variables, for January.

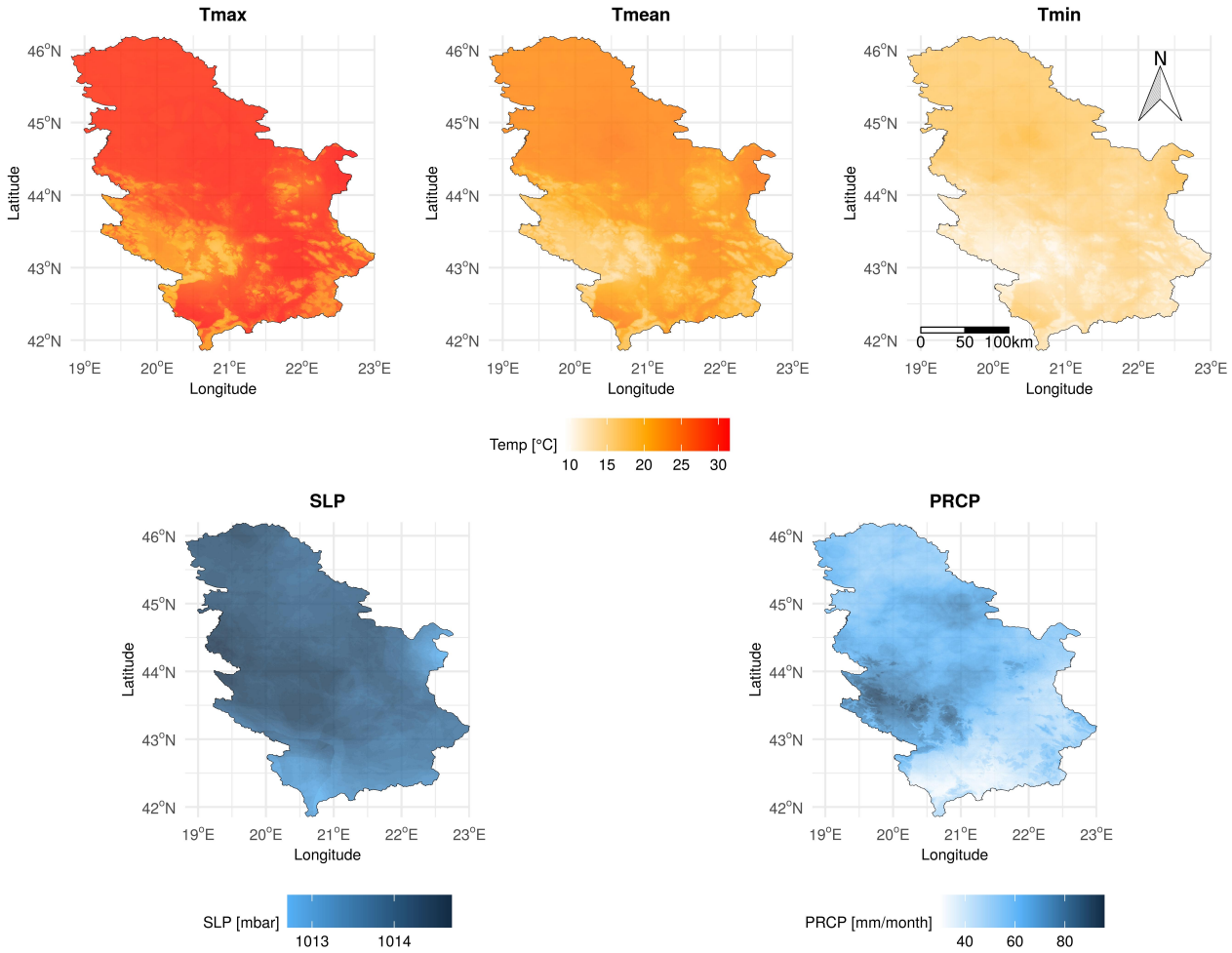


Figure 6.8: Montlhy LTM maps for all daily meteorological variables, for July.

Chapter 7

Automated spatio-temporal interpolation using R package meteo

This chapter, firstly, presents the R package `meteo` (Kilibarda et al. 2014) for the RFSI and spatio-temporal geostatistical interpolations and the R packages that it relies on. Secondly, it presents functions from the `meteo` package: (1) for the automation of prediction and cross-validation processes for the spatio-temporal regression kriging and (2) for the automation of creation, prediction, tuning and cross-validation processes of the newly developed RFSI methodology are presented.

7.1 Introduction

Twenty years ago, Bivand and Gebhardt (2000) examined a transition from the S and commercial `S-Plus` (Becker and Chambers 1984) to the open-source R environment (R Development Core Team 2012) in terms of handling and analysing spatial data. Many packages for spatial statistics and geostatistics were already ported to R and so became freely available to the wider range of potential users. One of the first packages of that kind was a `spatial` package (Venables and Ripley 2002), developed in 1997. Furthermore, the `spacetime` package (Pebesma 2012) extended R applicability from spatial to spatio-temporal statistics.

Now, twenty years later, the community for spatial statistics is still growing, mostly owing to the R packages such as `sp` (Pebesma and Bivand 2005; Bivand et al. 2013b), `rgdal` (Bivand et al. 2019), `raster` (Hijmans 2019), and `gstat` (Pebesma 2004) and the R's open-source concept which give the environment for providing reproducible researches (Bivand 2020). Almost 900 R packages use spatial classes from the `sp` package (Bivand 2020) and the `meteo` package is one such package. Bivand (2020) also emphasises the importance of the transition from the `sp`, `rgdal`, and `raster` to the newly developed `sf` (Pebesma 2018a), `stars` (Pebesma 2020), and `terra` (Hijmans 2020) packages in the future. The main reason for this transition is that the new packages are easier to maintain than the old one and the transition process is not problematic. Many new packages for the visualization of spatial data, such as tmap (Tennekes 2018, 2020) and mapview (Appelhans et al. 2020), relies on these new spatial packages.

Many ML algorithms are also implemented in R[1] and are used for spatial interpolation (see Sections 2.3.2 and 2.4.4). For example, the most popular implementations of the RF algorithm are `randomForest` (Liaw and Wiener 2002) and `ranger` (Wright and Ziegler 2017). The `caret` package (Kuhn 2019) provides a set of functions that makes modelling with various ML algorithms

[1] https://cran.r-project.org/web/views/MachineLearning.html

easier. Researchers often develop packages for new spatial interpolation methodologies (Hengl et al. 2018; Georganos et al. 2019; Baez-Villanueva et al. 2020; Møller et al. 2020) that heavily rely on the R ML packages.

The R package `meteo` implements both spatio-temporal geostatistics and RF-based spatial interpolation (RFSI) for climate and other environmental variables. It is mainly based on `sp`, `spacetime`, `gstat`, and `ranger` packages.

7.2 R programming language

As defined by the R Development Core Team (2012) and on the official R project website[2], "*R is a language and environment for statistical computing and graphics*". R originated from S language and environment developed mainly by John Chambers together with Rick Becker and Allan Wilks, employees of the Bell Laboratories company (Becker and Chambers 1984; Becker et al. 1988; Chambers and Hastie 1992; Chambers 1998). R and S are practically the same programming language with a difference in implementation. Most of S code can be run in R without any change. R is a GNU[3] project which means that R is an open source software under the General Public License. In other words, R is an open source version of S. R has a possibility for calling functions written in C, C++, and `Fortran` code which can speed-up computation in many cases.

The main reason why R succeeded as a project is its functional extensibility through a package oriented system. An R package is a set of functions, data, and help that are related to a specific task. R environment comes with several base packages that cover basic statistical and linear algebra computations, creation of graphics, and other similar functionalities. Besides base R packages, there is a long list of user-created R packages that are intended for a specific use. These user-created R packages are available from The Comprehensive R Archive Network (CRAN)[4] main repository and several other repositories, such as `r-forge`[5], `GitHub`[6], etc., and can be installed if needed.

7.3 R package meteo

Initially, the R package `meteo` was developed by Kilibarda et al. (2014) in order to provide functionalities for automated spatio-temporal interpolation of climate variables. However, it can also be used for the spatio-temporal interpolation of any other environmental variable. It contains a set of functions for automated interpolation with STRK and for preparing the space-time data and covariates, together with global STRK (regression and semivariogram) models for maximum, minimum, and mean daily temperatures (Kilibarda et al. 2014) created based on publicly available climate data, such as GSOD, ECA&D, GHCN-daily (Section 3.2), and environmental covariates such as MODIS LST (Section 3.3.2.1), DEM and TWI (Section 3.3.4).

In this dissertation, R package `meteo` is updated with:

- an improvement of STRK prediction and cross-validation functions,
- a STRK model for mean daily temperature for Croatia (Section 4.4.2),

[2]https://www.r-project.org/
[3]https://www.gnu.org/
[4]https://cran.r-project.org/
[5]https://r-forge.r-project.org/
[6]https://github.com/

- functionalities for spatial interpolation with RFSI methodology (Section 5.2.1.2).

The latest version of the R package `meteo` can be downloaded from `r-forge`[7] and `GitHub`[8].

7.3.1 Related R packages

7.3.1.1 sp

The R package `sp` (Pebesma and Bivand 2005; Bivand et al. 2013b) provides classes, methods, and functions for handling spatial data. The spatial data classes are implemented for the following spatial objects:

- points: `SpatialPoints` and `SpatialMultiPoints` classes
- lines: `SpatialLines` class
- polygons: `SpatialPolygons` class
- grids: `SpatialPixels` and `SpatialGrid` classes

All of these classes have corresponding `*DataFrame` classes (e.g. `SpatialPointsDataFrame`), that, besides spatial location information, contain additional attributes for spatial objects in R's `data.frame` format — a table where columns represent attributes and rows represent object instances. All of the `sp` classes are based on a `Spatial` class that has common spatial methods. This package has methods and functions for plotting the spatial data, spatial selection, taking a subset of data, retrieving spatial information (e.g. coordinates, projection, etc.), and others.

7.3.1.2 spacetime

The R package `spacetime` (Pebesma 2012) provides classes, methods, and functions for handling spatio-temporal data. It actually extends the `sp` package by adding a temporal dimension to the spatial data using the `xts` class of the R package `xts` (Ryan and Ulrich 2020). The space-time classes are:

- spatio-temporal full grids: STF
- spatio-temporal sparse grids: STS
- spatio-temporal irregular grids: STI
- spatio-temporal trajectories: STT

The same as for `sp` package, these classes have their corresponding *DF classes (e.g. STFDF) that also store attributes in `data.frame` format and has a ST class that has common spatio-temporal methods for all `spacetime` classes. This package has methods and functions for plotting the spatio-temporal data as map sequences, spatio-temporal selection and subsetting, retrieving spatial information, and others.

[7] http://r-forge.r-project.org/projects/meteo/
[8] https://github.com/AleksandarSekulic/Rmeteo

7.3.1.3 gstat

R package `gstat` (Pebesma 2004) is the most popular package for spatial and spatio-temporal geostatistical interpolation (modelling, prediction, and simulation). It contains functionalities for performing various univariable and multivariable (co-) kriging versions:

- (residual/cross) semivariogram modelling using sample semivariograms and fitting of parametric models
- applying geometric anisotropy, i.e. directional semivariograms
- restricted maximum likelihood fitting of partial sills
- plotting of sample (cross) semivariograms and fitted semivariograms
- SK, OK, UK, KED, and their co-kriging versions
- (sequential) Gaussian (co)simulation
- indicator (co)kriging and sequential indicator (co)simulation
- local and global kriging
- block (co)kriging or simulation for rectangular or irregular blocks.

An input for most of these functions are objects of `sp` classes.

Since the year 2010, when the development of `spacetime` package started, `gstat` package has also acquired functionalities for spatio-temporal kriging (Gräler et al. 2016). This means that STRK can also be performed, combining the `lm` function (`stats` package) with spatio-temporal kriging for residuals. Recently, `gstat` package can handle novel spatio-temporal classes from R packages `sf` (Pebesma 2018b) and `stars` (Pebesma 2020).

7.3.1.4 ranger

`ranger` (Wright and Ziegler 2017), short for "RANdom forest GEneRator", is a fast implementation of random forests (Breiman 2001, Section 2.3.2.1) for high dimensional data written in C++ and also available as a package in R. It deals with classification and regression in RF and also has functionalities for Random Survival Forests (Ishwaran et al. 2008), extremely randomized trees (Geurts et al. 2006), and quantile regression forests (Meinshausen 2006).

Wright and Ziegler (2017) compared `ranger` with various RF implementations, including the widest used `randomForest` package (Liaw and Wiener 2002), and showed that `ranger` is so far "*the fastest and most memory efficient implementation of RF to analyze data*". They also recommend an R version of `ranger` because it is easy to use and is as fast as a C++ implementation.

7.3.1.5 nabor

R package `nabor` (Elseberg et al. 2012) is actually an R wrapper for `libnabo` — a fast K Nearest Neighbour (KNN) library for low dimensions (2D and 3D). Its `knn` function is so far and to our knowledge, the fastest implementation of the KNN algorithm and is, as such, used in the calculation of Euclidean distances to n nearest stations for the RFSI interpolation methodology (Section 5.2.1.2).

7.3.1.6 snowfall and doParallel

`snowfall` (Knaus 2015) is a R package for parallel computing. It wraps over a more complex `snow` (Tierney et al. 2018) package for parallel computing and so makes parallel computing using clusters easier.

The `doParallel` package (Microsoft Corporation and Steve Weston 2019) is even faster and easier to use than the `snowfall` package. It combines a `parallel` package (R Core Team 2020), which is also based on the `snow` package, to provide a parallel computing backend for a `foreach` (Microsoft and Weston 2017) package, which creates loops, using its `%dopar%` function.

These packages are intended for inexperienced users in the area of parallel computing in R.

7.3.1.7 rgdal and raster

The `rgdal` package (Bivand et al. 2019) is a wrapper over two libraries: (1) a Geospatial Data Abstraction Library (GDAL) (GDAL/OGR contributors 2020) which is a translator library for raster and vector geospatial data formats, and (2) a PROJ library (PROJ contributors 2020) which has functionalities to transform geospatial coordinates between various coordinate reference systems (CRS), including cartographic projections and geodetic transformations. Various formats of raster (gridded data) and vector data can be imported into an R environment as objects of `sp` classes and, in the opposite direction, can be exported from objects of `sp` classes.

The `raster` package (Hijmans 2019) is used for "*Reading, writing, manipulating, analyzing and modeling of spatial data.*", but mostly for raster data. It has GIS-alike functions for the manipulation of raster data, such as raster algebra, raster modifications, operations with vector data, and others.

`raster` package provides three main R raster classes:

- `RasterLayer` - a single-layer raster data,
- `RasterBrick` - a multi-layer raster data from a single file,
- `RasterStack` - a multi-layer raster data from many files,

and it uses the `rgdal` package to read and write raster data.

These two packages are actually not included in the R package `meteo`, but are relevant for producing maps and importing the data into an R environment.

7.3.2 Spatio-temporal regression kriging

R package `gstat` provides functions for spatio-temporal kriging, UK, and KED, but there is no support for STRK (Section 4.3.1). STRK can be performed in two steps, fitting the MLR trend model first with the `lm` function (`stats` package) and then fitting a spatio-temporal semivariogram on residuals. Prediction is also made in two steps, summarising a prediction from the MLR trend model and a residual prediction from spatio-temporal OK model. R package `meteo` provides an automated function for precisely these shortcomings of the `gstat` package. The `pred.strk` function of R package `meteo` does a STRK prediction in one step with the previously fitted MLR trend model and spatio-temporal semivariogram on residuals. Unlike the`gstat` package, `pred.strk` function (`meteo`) gives a possibility for accuracy assessment of the STRK model using leave-one-station-out cross-validation.

The description of the new `pred.strk` function of the R package `meteo` is given in this Section. The new `pred.strk` function works only for STRK prediction as the code is more optimised and it now supports more different input and output formats (classes) as well as parallel processing with `doParallel` package. Accuracy assessment of the STRK model is improved and changed to k-fold leave-location-out cross-validation and is implemented in a new `cv.strk` function. The old `pred.strk` function is still available in R package `meteo` and it was renamed to `pred.strk.old`.

The `pred.strk` and `cv.strk` functions were used for the Croatian mean daily temperature case study (Section 4.4.2).

7.3.2.1 Prediction

An MLR trend model and a spatio-temporal semivariogram over the MLR trend residuals have to be fitted before the STRK prediction obtained through the `pred.strk` function. The MLR trend model can be fitted using the `lm` function (`stats` package). The spatio-temporal sample semivariogram can be obtained using the `variogramST` function (`gstat` package) and can be fitted using the `fit.StVariogram` function (`gstat` package). The `pred.strk` function has the following arguments:

```
pred.strk(data,
          zcol=1,
          data.staid.x.y.time = c(1,2,3,4),
          obs,
          obs.staid.time = c(1,2),
          stations,
          stations.staid.x.y = c(1,2,3),
          newdata,
          newdata.staid.x.y.time = c(1,2,3),
          zero.tol=0,
          reg.coef,
          vgm.model,
          sp.nmax=20,
          time.nmax=2,
          by="time",
          tiling= FALSE,
          ntiles=64,
          output.format = "STFDF",
          parallel.processing = FALSE,
          pp.type = "snowfall",
          cpus=detectCores()-1,
          computeVar=FALSE,
          progress=TRUE,
          ...)
```

The algorithm of the `pred.strk` function is given in the Figure 7.1.

In the beginning, an input spatio-temporal data is prepared. Observations at stations can be given in four different ways: as a `data` argument of (1) `STFDF` or (2) `STSDF` class (`spacetime` package), (3) `data.frame` class, or (4) through `obs` and `stations` arguments of `data.frame` class. `zcol` argument shows the position of a target variable in `data` object

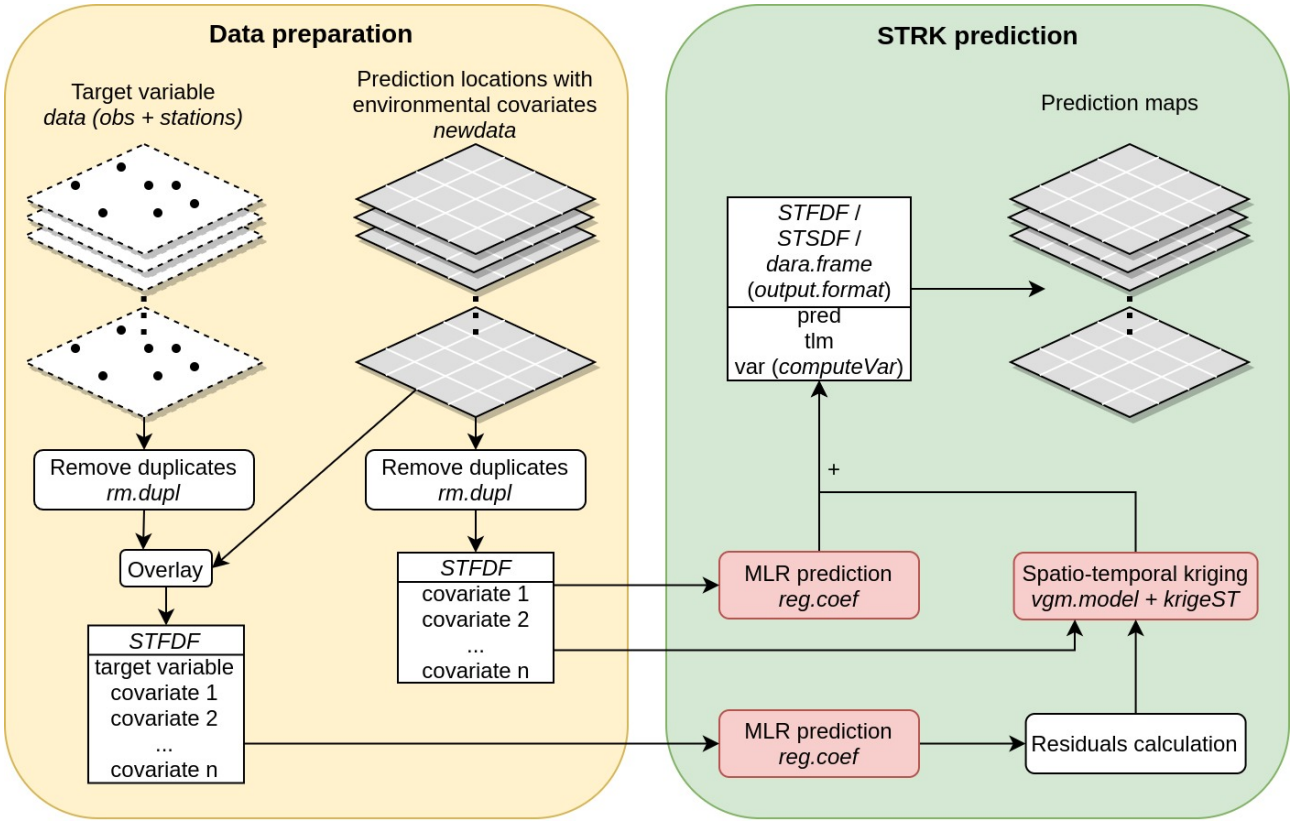


Figure 7.1: An algorithm for STRK prediction using the `pred.strk` function.

and `data.staid.x.y.time` argument shows the position of the station ID (staid), longitude (x), latitude (y), and time columns in `data` object only if it is of the `data.frame` class. Optionally, the `data` object has columns with covariate values, named regression coefficients in the `reg.coef` object. The argument `obs.staid.time` shows the position of the station ID (staid) and time columns in the `obs` object, and the argument `stations.staid.x.y` shows the position of the station ID (staid), longitude (x), and latitude (y) columns in the `stations` object. Prediction locations can be given in three different ways just like (with) the first three ways for the observations at the stations. These are a `newdata` argument of (1) `STFDF`, (2) `STSDF`, or (3) `data.frame` class where `newdata.staid.x.y.time` argument does the same thing as the `data.staid.x.y.time` argument does for `data` object. In addition, the `newdata` object has to have columns with covariate values, named regression coefficients in the `reg.coef` object. All of the spatial duplicates, i.e. the point pairs with equal spatial coordinates, are removed with a `rm.dupl` function from the `meteo` package. A `zero.tol` argument sets distance value below (or equal to) which spatial locations are considered duplicates. If it is set to zero, there will be no duplicates. Both `data` and `newdata` objects are converted to the `STSDF` class using the `meteo2STFDF` function (`meteo` package) at the end of this stage.

Next, the MLR trend prediction is performed. The MLR trend coefficients, that can be obtained from an `lm` object with the `coefficients` method, are given in the `reg.coef` argument. If any of the covariates in the `reg.coef` are missing from the `newdata` object, an error is raised. If there are no covariates in the `data` object, the function firstly checks if the covariates can be obtained from the `newdata` object by doing an spatio-temporal overlay. If the covariates are still missing from the `data` object, a spatio-temporal OK is performed. The spatial locations or time instances without covariate values are further removed from the `data` and `newdata` objects.

The residuals at the stations are estimated by subtracting observations with MLR trend prediction. Spatio-temporal kriging is performed using the `krigeST` function from the R package

gstat and an already fitted spatio-temporal semivariogram in the StVariogramModel class (gstat package), given through the vgm.model argument, over the station residuals. Additionally, the number of nearest spatial and temporal locations used for prediction can be set with arguments sp.nmax and time.nmax. There is also an option to perform kriging in the ntiles number of tiles of newdata object (tiling argument) which can speed up the whole process significantly. Kriging variance can also be calculated by setting the computeVar argument to be TRUE. The kriging process can be performed sequentially or parallelly, setting the parallel.processing argument to FALSE or TRUE respectively, looping through time instances or through spatial locations of the newdata object by setting the by argument to time or station respectively. If the parallel processing option is chosen, additional arguments, such as wheather snowfall or doParallel package will be used (pp.type argument) and number of used CPUs (cpus argument), can be set. Additional arguments (...) of the krigeST function can also be set.

Finally, the residuals estimated by spatio-temporal kriging are added to the MLR trend prediction for the newdata object. The output of the pred.strk function is or an object of STFDF, STSDF or data.frame class (depends on output.format argument), with the following columns:

- pred – STRK predictions,
- tlm – MLR trend predictions,
- var – kriging variance (if computeVar=TRUE).

Whether the progress of the STRK prediction process will be shown or not, is set with the argument progress.

7.3.2.2 Cross-validation

The k-fold LLOCV is used for the accuracy assessment of the spatial models, where all of the observations from one station are in the same fold. As for the pred.strk function, a MLR trend model and a spatio-temporal semivariogram over the MLR trend residuals have to be fitted before using the cv.strk function. The cv.strk function has the following arguments:

```
cv.strk(data,
        zcol=1,
        data.staid.x.y.time = c(1,2,3,4),
        obs,
        obs.staid.time = c(1,2),
        stations,
        stations.staid.x.y = c(1,2,3),
        zero.tol=0,
        reg.coef,
        vgm.model,
        sp.nmax=20,
        time.nmax=2,
        type = "LLO",
        k = 5,
        seed = 42,
        folds,
```

```
fold.column,
refit = TRUE,
output.format = "STFDF",
parallel.processing = FALSE,
pp.type = "snowfall",
cpus=detectCores()-1,
progress=TRUE,
...)
```

The algorithm of the `cv.strk` function is given in Figure 7.2.

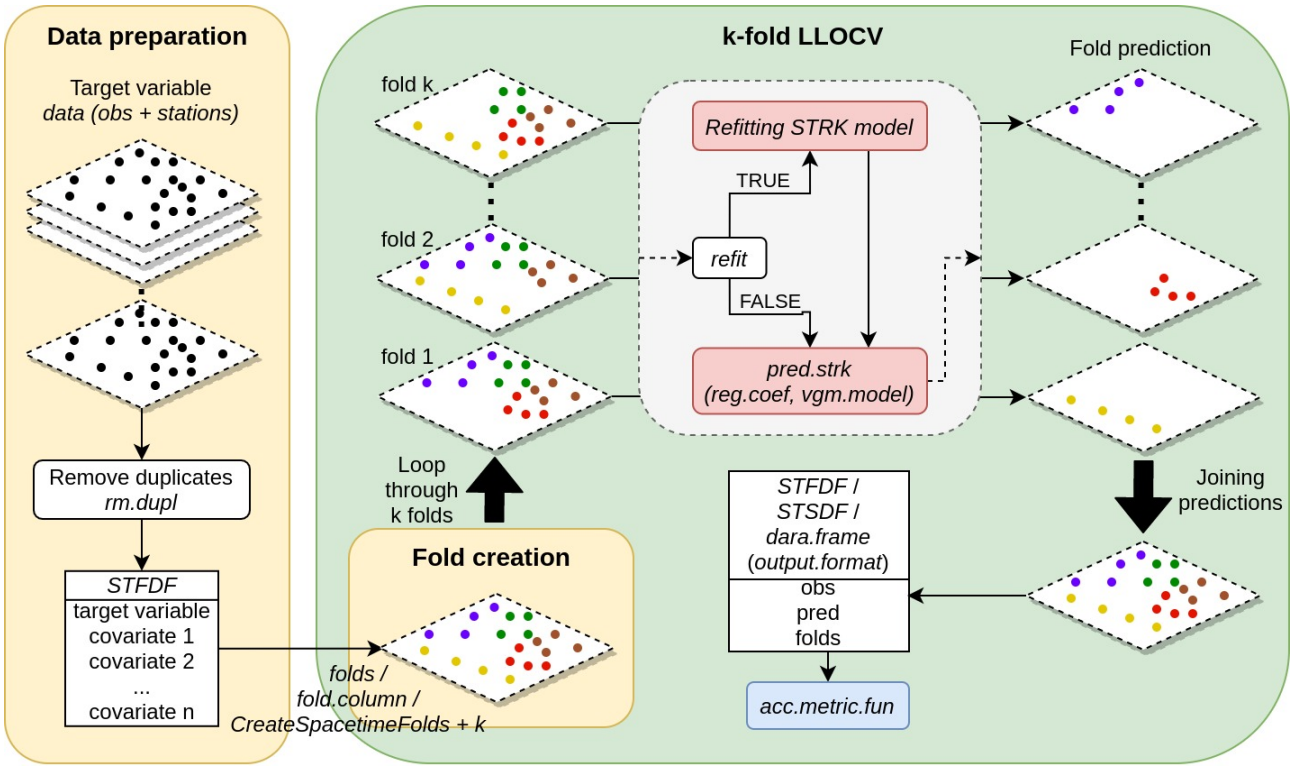


Figure 7.2: An algorithm for k-fold LLOCV of the STRK model using the `cv.strk` function.

Input object classes and a data preparation process of the observations at stations is the same as for the`pred.strk` function. The only difference being that covariates, named regression coefficients in the `reg.coef` object, have to be present in the `data` object (or in the `obs` object) which means that a spatial overlay with covariates has to be done first. If any of the covariates are missing, a spatio-temporal OK is performed.

The creation of the folds can be done in three ways. The first way is for the user to set the fold vector using the argument `folds` (the length has to be the same as the length of the observations). The second way is to use a user-specified column from the `data` objects (or `obs` object), set by the `fold.column` argument. The third way is to set the number of folds (`k`) and `seed`, and then the `CreateSpacetimeFolds` of the R package CAST (Meyer 2018) is used for the random creation of the `k` folds. Currently, only LLOCV is implemented in the `cv.strk` function. In the future, leave-time-out and leave-location-time-out cross-validation will be implemented and will be set through the `type` argument.

Once the folds are created, a LLOCV process is performed. The `cv.strk` function loops through each of the folds where the observations from the current fold are used for validation and the

observations from the remaining folds (k−1 folds) are used in the STRK prediction process of the observations from the current fold. If the `refit` argument is set to TRUE, for the each fold, the STRK model (MLR trend model and spatio-temporal semivariogram) is always refitted with the data from the remaining folds. For this purpose, the same covariates as in the `reg.coef` argument are used for fitting the MLR trend model and the `vgm.model` is used as the initial semivariogram for fitting. The `pred.strk` function is used for fold STRK prediction. Looping through the folds is done sequentially and the parallel processing arguments: `parallel.processing`, `pp.type`, and `cpus`, refer to the `pred.strk` function. The additional arguments (`...`) of the `pred.strk` (`krigeST`) function can also be set.

In the end, all of the observations and corresponding LOOCV predictions are put in the one unique object of the `STFDF`, `STSDF` or `data.frame` class, setting the `output.format` argument, with the following columns:

- `obs` — observations,
- `pred` — predictions from the k-fold LLOCV,
- `folds` — folds used for the k-fold LLOCV.

The `progress` argument can be used to set the showing of the progress of the LLOCV process. The `meteo` package provides an `acc.metric.fun` function that can be used for the calculation of the standard classification and regression accuracy metrics, where inputs are observations and predictions from LLOCV.

7.3.3 Random Forest Spatial Interpolation

RFSI (Sekulić et al. 2020a; Section 5.2.1.2) is a novel methodology for spatial interpolation that uses observations at the nearest stations and the distances to them as covariates in the RF model. In order to create, validate, and make a prediction from an RFSI model, four new functions are added to R package `meteo`:

- `rfsi` — RFSI model creation,
- `pred.rfsi` — RFSI prediction,
- `tune.rfsi` — tuning of RFSI model,
- `cv.rfsi` — nested k-fold LLOCV of RFSI model.

The RFSI functions of the `meteo` package relies extensively on the RF algorithm of the `ranger` package (Section 7.3.1.4).

The description of these four functions is given in this Section. These functions were used for the modelling of daily precipitation for Catalonia (Section 5.3.2), daily mean temperature for Croatia (Section 5.3.3, and climate elements for Serbia (Chapter 6).

7.3.3.1 Model development

The `rfsi` function is used for creation of the RFSI model and has the following arguments:

```
rfsi(formula,
     data,
     data.staid.x.y.time = c(1,2,3,4),
     obs,
     obs.staid.time = c(1,2),
     stations,
     stations.staid.x.y = c(1,2,3),
     zero.tol = 0,
     n.obs = 10,
     # time.nmax,
     avg = FALSE,
     increment = 10000,
     range = 50000
     direct = FALSE,
     use.idw = FALSE,
     idw.p = 2,
     s.crs = NA,
     t.crs = NA,
     cpus = detectCores()-1,
     progress = TRUE,
     ...)
```

The algorithm of the `rfsi` function is given in Figure 7.3.

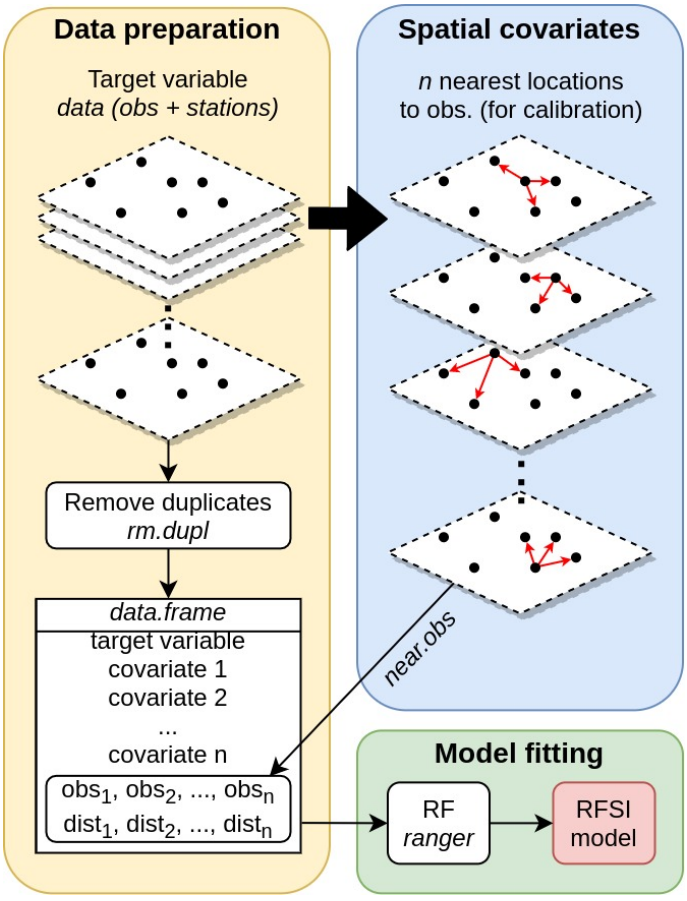


Figure 7.3: An algorithm for RFSI model development using the `rfsi` function.

The preparation of the observations at the stations is done in the same way as for the *cv.strk* function. The observations at stations together with covariates can be of STFDF, STSDF,

or `data.frame` and imported to the function through the `data` argument (or `obs` and `stations` arguments). In addition, the `data` object can be of `SpatialPointsDataFrame` or `SpatialPixelsDataFrame` (`sp package classes`) when pure spatial interpolation is performed, for only one time instance. RFSI currently does only a spatial interpolation, but can be applied to spatio-temporal variables. In that case, RFSI assumes that the spatial process for such variables is the same (does not change) over time.

The `data` object coordinates are reprojected from the source CRS (`s.crs`) to the target CRS in the projection (`t.crs`) using the `spTransform` function (`sp` package), unless they are already in the projection. This is necessary for the calculation of the Euclidean distances between observations. The source CRS is taken from the `data` object if it exists, otherwise it is taken from the `s.crs` argument. If one of the `s.crs` and `t.crs` arguments is empty (NA), the coordinates of the `data` object are taken as they are already in the projection and used for the Euclidean distances calculation as they are.

Next, the `n.obs` nearest observations and distances to them are calculated for each observation (and each time instance) using the `near.obs` function from R package `meteo`. The `near.obs` function is based on the `knn` function of the `nabor` package. The output of the `near.obs` function is a `data.frame` where the first `n.obs` columns are the Euclidean distances to the `n.obs` nearest stations and next `n.obs` columns are observations at the `n.obs` nearest stations, and the rows are given observations. The `near.obs` function works parallelly always with a set `cpus` number of cores, because without parallel processing and in the case of large time series of observations, it would be a time consuming process. Additional spatial covariates can also be calculated with the `near.obs` function, such as averages in circles with different radius around observations the (if `avg` argument is TRUE, based on the radius `increment` and maximum `range`), the nearest observations in four mathematical quadrants of observations the (if `direct` argument is TRUE), the IDW predictions at observation locations the (if `use.idw` argument is TRUE, with the IDW weight power of `idw.p`).

After the calculation of spatial covariates, the RFSI model, in essence an RF model, is fitted using the `ranger` function from the same name package and the `formula` argument (of `formula` class). The `formula` contains only environmental covariates, without spatial covariates which are implied by setting the `n.obs`, `avg`, etc. arguments. If the `formula` argument is `z 1`, the RFSI model is fitted using only the spatial covariates. The `ranger` function already works parallelly with the `cpus` number of cores. Additional arguments (`...`) of the `ranger` (`ranger` package) function can also be set. As for the previously described functions, if the `progress` is set to TRUE, the progress of the whole process will be printed. The `rfsi` function returns an RFSI model of the `ranger` class.

7.3.3.2 Prediction

Before the RFSI prediction using the `pred.rfsi` function, an RFSI model, based on which the prediction will be made, has to be fitted with the `rfsi` function. The `pred.rfsi` function is used for prediction from the RFSI model and has the following arguments:

```
pred.rfsi(model,
          data,
          zcol=1,
          data.staid.x.y.time = c(1,2,3,4),
          obs,
          obs.staid.time = c(1,2),
          stations,
```

```
stations.staid.x.y = c(1,2,3),
newdata,
newdata.staid.x.y.time = c(1,2,3),
zero.tol=0,
s.crs=NA,
newdata.s.crs=NA,
t.crs=NA,
output.format = "data.frame",
cpus=detectCores()-1,
progress=TRUE,
...)
```

The algorithm of the `pred.rfsi` function is given in Figure 7.4.

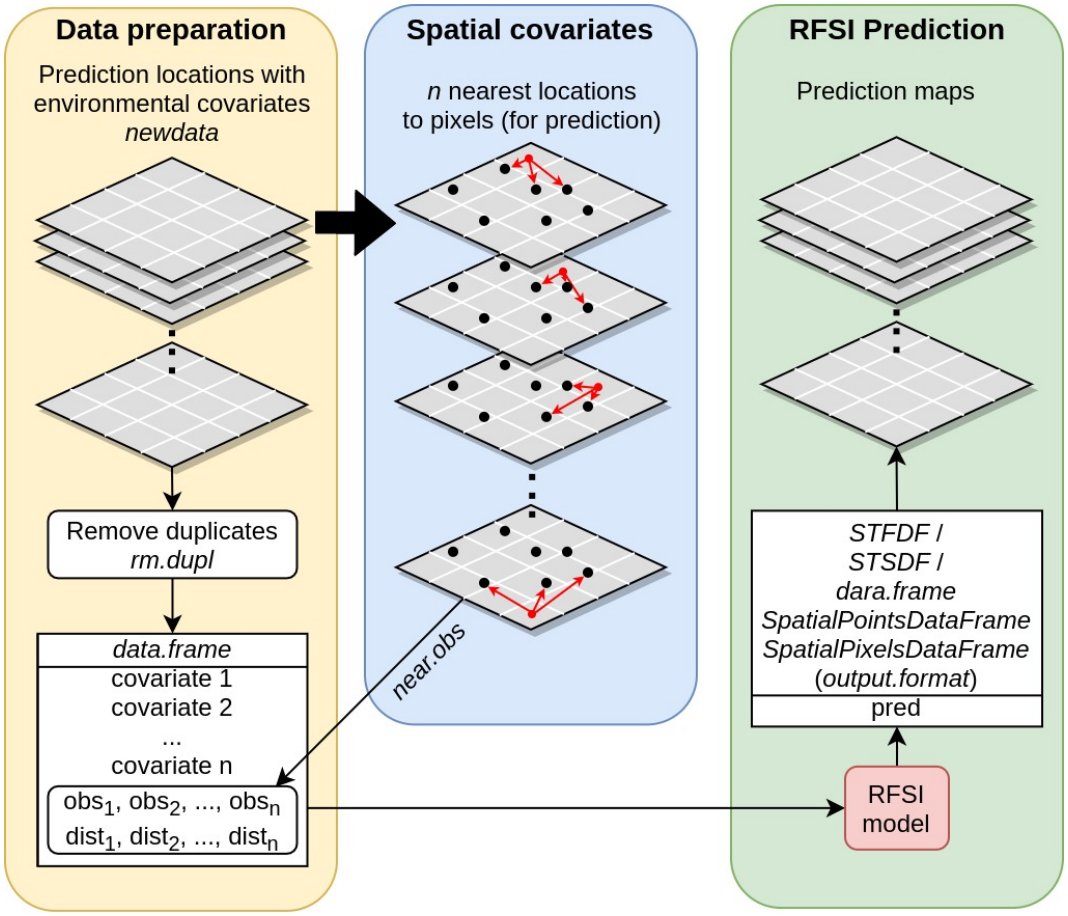


Figure 7.4: An algorithm for RFSI prediction using the `pred.rfsi` function.

Input object classes and preparation of the observations at stations (`data` or `obs` and `stations`) data is done in the same way as the preparation of the observations at stations for the `rfsi` function. Input object classes and preparation of the prediction locations (`newdata`) is done in a similar way as for the observations at stations, including the reprojection of the coordinates. The difference is that the prediction locations can have a different source CRS (`newdata.s.crs`), but both, observations and prediction locations, have one unique target CRS (`t.crs`).

The full list of covariates, both spatial and non-spatial, is extracted from the RFSI `model`, so there is no need for the `formula` argument here. The spatial predictors (the nearest observations and distances to them, averages in circles with different radiuses, etc.) are now calculated for prediction locations (`newdata`), in a way described for the `rfsi` function. After this, prediction is

performed based on the RFSI model in the `model` argument and `predict()` function of the `ranger` package. Additional arguments `...` can be passed to the `predict()` function. The progress of the prediction process can be followed if the `progress` argument is set to TRUE. The `pred.rfsi` function returns and object of `data.frame`, `SpatialPointsDataFrame`, `SpatialPixelsDataFrame`, `STFDF`, or `STSDF` class, depending on set `output.format` argument, with predictions (`pred` column).

7.3.3.3 Model tuning

In order to optimally fit a RFSI model, various RF and other hyperparameters can be tuned. The hyperparameters that can be tuned are four RF hyperparameters: number of trees (`num.trees`, number of variables to possibly split at each node (`mtry`, minimal node size (`min.node.size`, ratio of observations-to-sample in each decision tree (`sample.fraction`, and splirule (`splitrule`); and two RFSI parameters: the number of the nearest stations (`n.obs`) and the power of the IDW weights (`idw.p`). The hyperparameters are tuned using the standard k-fold LLOCV. The `tune.rfsi` function is used for the tuning of the RFSI model and has the following arguments:

```
tune.rfsi(formula,
          data,
          data.staid.x.y.time = c(1,2,3,4),
          obs,
          obs.staid.time = c(1,2),
          stations,
          stations.staid.x.y = c(1,2,3),
          zero.tol=0,
          use.idw = FALSE,
          s.crs=NA,
          t.crs=NA,
          tgrid,
          tgrid.n=10,
          tune.type = "LLO",
          k = 5,
          seed=42,
          folds,
          fold.column,
          acc.metric,
          fit.final.model=TRUE,
          cpus=detectCores()-1,
          progress=TRUE,
          ...)
```

The algorithm of the `tune.rfsi` function is given in Figure 7.5.

Input objects classes (`STFDF`, `STSDF`, `data.frame`, `SpatialPointsDataFrame`, or `SpatialPixelsDataFrame`) and preparation of the observations at stations with environmental covariates together with coordinate reprojections is the same as for the `rfsi` function. The covariates of the tuned RFSI, alongside spatial covariates, are given within the `formula` argument.

The creation of the folds process is presented in the `cv.strk` function description. The folds can be created by the user, using the `folds` or `folds.column` argument, or randomly, using

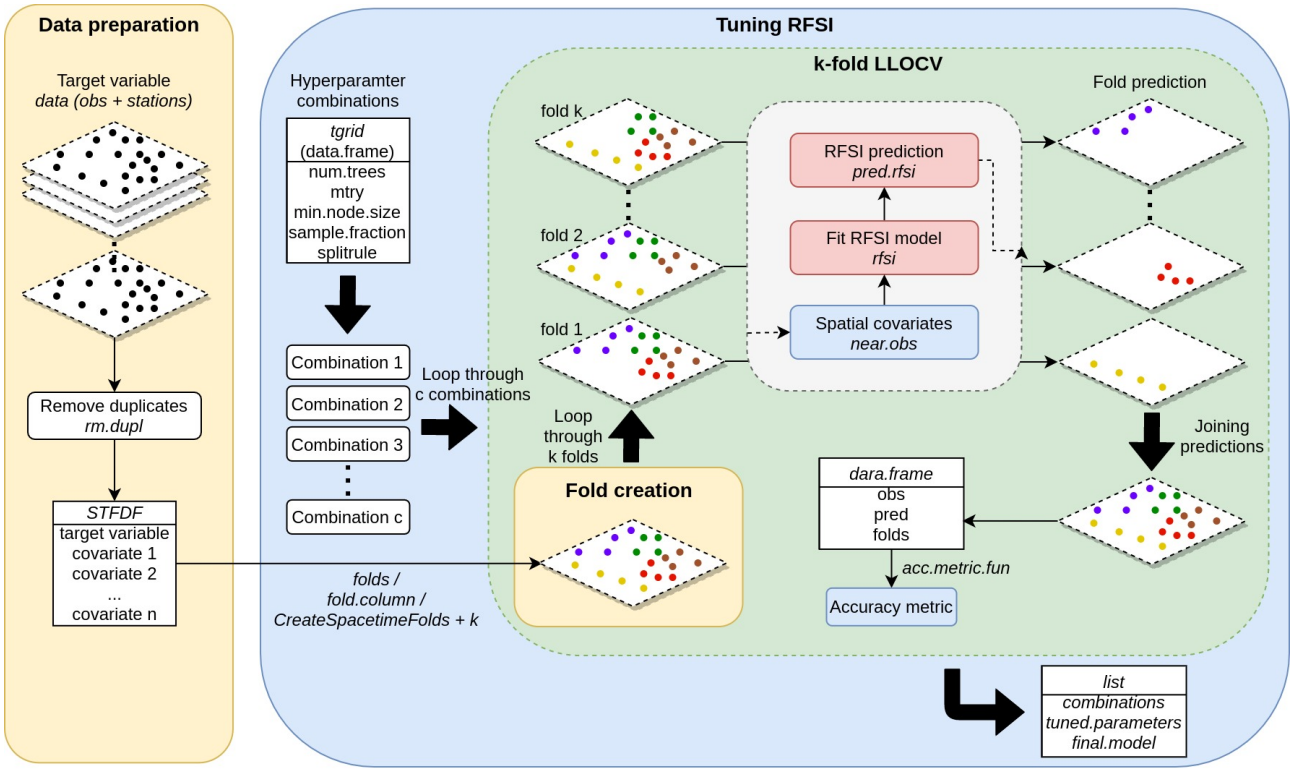


Figure 7.5: An algorithm for tuning of the RFSI model using the `tune.rfsi` function.

the `k` and `seed` arguments. Currently, only LLOCV is implemented.

The tuning process is done for the hyperparameter combinations in the `tgrid` argument of the `data.frame` class. Only hyperparameters that are present in the `tgrid` are tuned, while others are set to default values. The `idw.p` parameter is tuned only if the `use.idw` argument is set to TRUE. A number of random hyperparameter combinations in the `tgrid` can be taken using the `tgrid.n` argument. The `tune.rfsi` function loops through the hyperparameter combinations. For each of the combinations the k-fold LLOCV (set by `tune.type` argument) is performed, i.e. each fold is once used for validation, where the prediction is done using the `pred.rfsi` function, and the remaining folds are used to fit RFSI model using the `rfsi` function. The `cpus` argument set the number of cores used for parallel processing in the `rfsi` and `pred.rfsi` functions. After k-fold LLOCV for one hyperparameter combination, the specified accuracy metric (`acc.metric`) is calculated using the `acc.metric.fun` function and assigned to the hyperparameter combination. Additional arguments (`...`) can be passed to the `ranger` function. Finally, the hyperparameter combination with the best accuracy metric is taken as optimal and a final RFSI model is fitted based on it, if the `fit.final.model` argument is set to TRUE. The progress of the tuning process also can be followed (`progress`).

The `tune.rfsi` function returns a `list` with the following elements:

- `combinations` – `data.frame` of all hyperparameter combinations with chosen accuracy metric,
- `tuned.parameters` – the optimal hyperparameter combination,
- `final.model` – final RFSI model, if `fit.final.model`=TRUE.

7.3.3.4 Cross-validation

The `cv.rfsi` function is used for the nested k-fold LLOCV of the RFSI model. The difference between the standard and nested k-fold cross-validation is that, in the case of the nested k-fold cross-validation, the observed fold is validated on the model that is tuned on the remaining $k-1$ folds using the standard k-fold cross-validation (Section 5.2.3.2). The `cv.rfsi` function has the following arguments:

```
cv.rfsi(formula,
        data,
        data.staid.x.y.time = c(1,2,3,4),
        obs,
        obs.staid.time = c(1,2),
        stations, # data.frame(id,x,y)
        stations.staid.x.y = c(1,2,3),
        zero.tol=0,
        use.idw=FALSE,
        s.crs=NA,
        t.crs=NA,
        tgrid,
        tgrid.n=10,
        tune.type = "LLO",
        k = 5,
        seed=42,
        folds,
        fold.column,
        acc.metric,
        output.format = "data.frame",
        cpus=detectCores()-1,
        progress=TRUE,
        ...)
```

The algorithm of the `cv.rfsi` function is given in Figure 7.6.

Classes of the input objects and the data preparation process with coordinate reprojections (from `s.crs` to `t.src`) is the same as for the `tune.rfsi` and `rfsi` functions. The covariates of the RFSI model being cross-validated, alongside spatial covariates, are given within the `formula` argument. The number of the nearest observations (`n.obs`) and the power of the IDW weights are given in the `tgrid` argument.

The creation of the main folds for the k-fold nested LLOCV can be done in three ways, the same as for the `tune.rfsi` function: (1) using the `folds` argument, (2) `folds.column` argument, or (3) randomly, using the `k` and `seed` arguments.

The `cv.rfsi` function loops through k main folds and each fold is used once for prediction (i.e. validation) from the RFSI model tuned on the data from the remaining $k-1$ folds. Tuning of these "nested" RFSI models is done with standard the k-fold LLOCV using the `tune.rfsi` function for the `tgrid.n` number of hyperparameter combinations defined in the `tgrid` argument and also by checking the specified accuracy metric in the `acc.metric` argument. The nested folds are created randomly with the `CreateSpacetimeFolds` of the R package CAST (Meyer 2018) and `k` and `seed` arguments. Currently, only the k-fold LLOCV (`tune.type` argument) is implemented in the `tune.rfsi` function. The prediction for each main fold is done using the

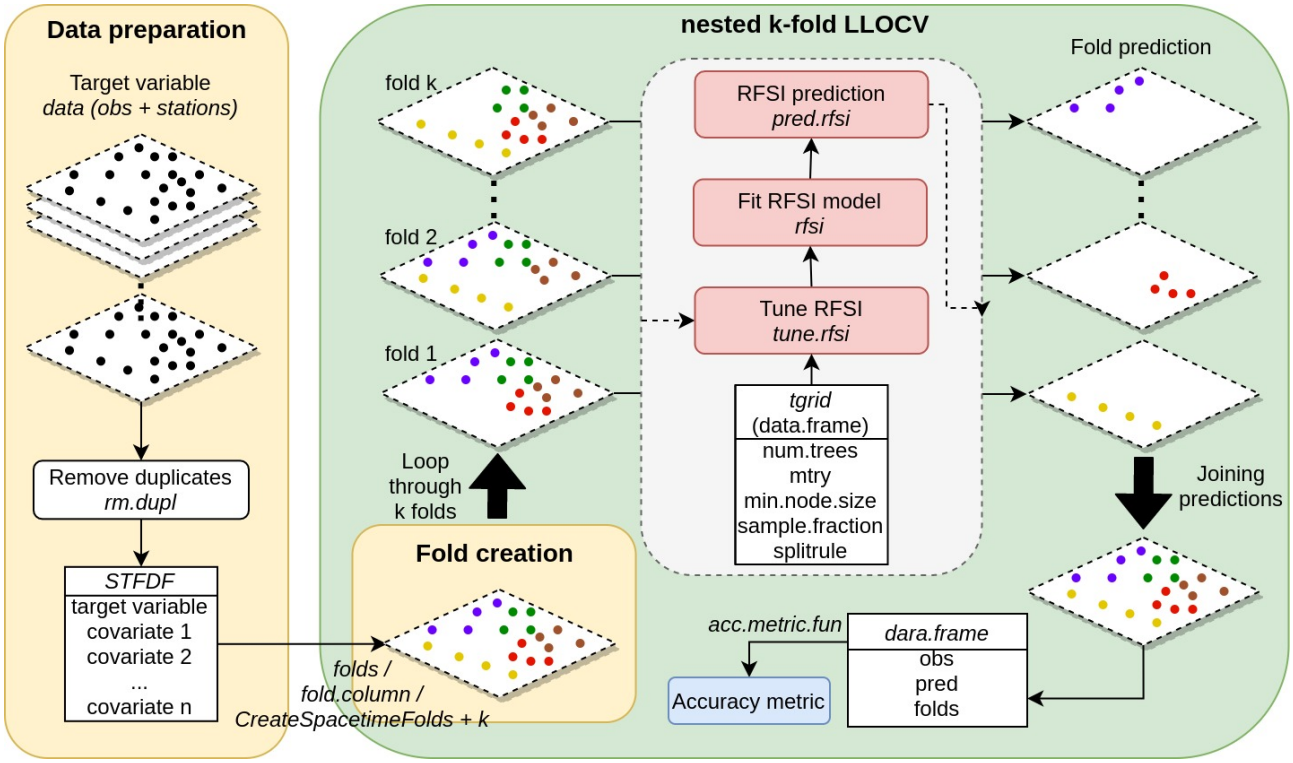


Figure 7.6: An algorithm for nested n-fold LLOCV of the RFSI model using the `cv.rfsi` function.

`pred.rfsi` function. The `cpus` argument set the number of cores used for parallel processing in the `tune.rfsi` and `pred.rfsi` functions. Additional arguments (`...`) can be passed to the `ranger` function.

As for the `cv.strk` function, all of the observations and corresponding predictions from the nested k-fold LOOCV are put in the one unique object of STFDF, STSDF or `data.frame` class, depending on the `output.format` argument, with the following columns:

- `obs` – observations,
- `pred` – predictions from the k-fold nested LLOCV,
- `folds` – folds used for the nested k-fold LLOCV.

Also, the `acc.metric.fun` function can be used for the estimation of the standard accuracy metrics.

7.4 Discussion and conclusions

The R package `meteo` implements functions for the automation of STRK and RFSI interpolations. The `pred.strk` function for STRK prediction is improved and the `cv.strk` function is created for k-fold LLOCV. Also, four new functions for creation (`rfsi`), prediction (`pred.rfsi`), tuning with k-fold LLOCV (`tune.rfsi`), and validation with the nested k-fold LLOCV (`cv.rfsi`) of the RFSI model are implemented. These `meteo` functions are easy to use and significantly reduce the amount of code and time that would be spent on implementation of these two interpolation methods. They also have support for standard spatial and spatio-temporal classes from the `sp` and

`spacetime` classes. Besides climate variables, the STRK or RFSI interpolations of the `meteo` package can be applied to any other environmental covariates.

Additional functionalities can be added to the `meteo` package and there is still room for improvements. Firstly, the transition to the `sf` and `stars` classes has to be done, but with maintaining support for the `sp` and `spacetime` classes. Additional spatial covariates can be added to the RFSI interpolation and some existing spatial covariates, such as an average in circles with different radiuses, have to be tuned and included in the cross-validation process. The RFSI can be extended from spatial to spatio-temporal interpolation by introducing observations at nearest stations and distances to them from previous time instances (e.g. days). Besides the nested k-fold LLOCV, two other "target-oriented" cross-validation approaches (Meyer et al. 2018), leave-time-out (LTOCV) and leave-location-and-time-out cross-validation (LLTOCV), have to be implemented. There is also room to speed up the tuning process by optimizing the way that the `tune.rfsi` function chooses the potential hyperparameter combination. Therefore less hyperparameters combinations will be looped. Eventually, RFSI methodology could possibly be integrated with the `ranger` function.

Chapter 8

Discussion and conclusions

Overall, the contribution of this dissertation is reflected in the improvement of spatio-temporal interpolation of climate elements using geostatistical and machine learning models. This has been accomplished through adopting global geostatistical models to local areas and developing an innovative spatio-temporal interpolation method based on the Random Forest machine learning algorithm, named Random Forest Spatial Interpolation (RFSI).

The spatio-temporal regression kriging model for global land areas is refitted on a local Croatian weather station network and as a result improves accuracy of daily mean temperature maps at a 1 km spatial resolution. The accuracy of the adapted geostatistical model for local areas, assessed using the leave-one-out cross-validation, was 97.8% in R^2 and 1.2 °C in RMSE, which is an improvement of 3.4% in R^2 and 0.7 °C in RMSE in comparison with the original global geostatistical model. This showed that global daily geostatistical models can be applied to local areas with denser weather station networks and produce more accurate daily maps of climate elements, at least for daily temperature. The spatio-temporal regression kriging model for the mean daily temperature is rather simple because it includes only three covariates. From them two are static and they are DEM derivatives (DEM and TWI) and one is dynamical (GTT) and can easily be calculated. Therefore, this model can be used for obtaining near real-time daily temperature maps with high accuracy. The simplified daily temperature model for Croatia mostly outperforms existing models for Croatia and other local areas. Still, the adapted model does not solve the problem of lower accuracy in the mountainous region caused by the lack of weather stations. Additional observations at higher altitudes could be used for the calibration of the geostatistical model in order to improve accuracy at higher altitudes. Another solution would be to include MODIS LST into the model, but this will increase the complexity of the temperature model and cause a delay in daily temperature map creation. The problem with the accuracy at higher altitudes is a topic for a future study. The approach of adopting global spatio-temporal geostatistical models for daily climate elements for the creation of more accurate localized maps can be applied to any climate element other than temperature.

The RFSI methodology for spatial or spatio-temporal interpolation, which is based on the Random Forest algorithm and observations at nearest stations and distances to them in the form of spatial covariates, was developed and described in Chapter 5. In the synthetic case study, RFSI outperforms simple deterministic interpolation methods. It was also shown that RFSI can be used to produce accurate maps of daily climate elements in the daily precipitation amount and daily mean temperature case studies. In these case studies the RFSI methodology outperforms spatio-temporal regression kriging, inverse distance weighting, standard random forest, and RFsp interpolation methods. New spatial covariates, observations at nearest stations and distances to them, are most credited for this because, unlike existing machine learning methods for spatial interpolation, they introduce spatial context in machine learning (Random Forest) model in a similar manner

as kriging. Unlike kriging, RFSI prediction can be seen as a non-linear combination of spatial and other environmental covariates thanks to the Random Forest algorithm. According to that, it is recommended to use RFSI for spatio-temporal interpolation of complex and non-stationary climate elements, such as precipitation. The introduction of the new spatial covariates further allows the Random Forest algorithm to decide whether spatial correlation or correlation with environmental covariates has more influence on prediction. Besides interpolation of climate elements, RFSI can be applied on interpolation of soil, pollutants, population density, or any other environmental parameter.

Certainly, the RFSI methodology is expected to be improved in the future and applied to various environmental case studies. The RFSI does not fully exploit distances to the stations (spatial covariates) and there is room for a methodological improvement. In essence, the RFSI is a spatial interpolation methodology and temporal component (e.g. spatial covariates from several days before) is planned to be incorporated in the future. Also, there is a possibility to develop a multivariate RFSI version to support modelling of relations between co-variables. Spatial covariates are calculated before creation of the RFSI model and this can potentially cause the circular reasoning problem. The solution to this problem would be to incorporate spatial covariates calculation in the bagging process of the Random Forest algorithm. Finally, spatial covariates can be used in different machine learning algorithms than Random Forest and thus providing innovative frameworks for spatio-temporal interpolation.

Machine learning algorithms, especially ensemble machine learning algorithms, are breaking boundaries in the area of spatial and spatio-temporal interpolation and could potentially replace the traditional interpolation methods in the future due their robustness and ability to assimilate a large number of spatial and environmental covariates. Regardless of their high accuracy, machine learning algorithms are often a ”black-box”, so the simple deterministic and interpretative geostatistical methods will be in use for a long time.

The RFSI methodology was applied to spatio-temporal interpolation of daily climate elements, namely maximum, minimum and mean temperature, mean sea level pressure, and total precipitation, for the Serbian territory, for the 2000–2019 period. The resulting MeteoSerbia1km dataset is the first gridded daily climatological dataset at a 1 km spatial resolution for Serbia. The results show that RFSI can provide high-level climatological maps with accuracy comparable to the 10-km daily E-OBS dataset. Daily maps and aggregated climatological products (monthly, annual summaries and daily, monthly, and annual long term means) of MeteoSerbia1km can be applied to climate change, agricultural and many other research areas. MeteoSerbia1km dataset can be improved using a larger number of local weather stations in Serbia and following the improvement of RFSI methodology in the future.

The implemented functions for creation, prediction, tuning, and cross-validation of the RFSI model in the R package `meteo` facilitate and automate the use of the RFSI methodology and provide support for standard R spatial and spatio-temporal classes. This will increase the availability of RFSI and the number of its applications. These RFSI functions are still under development and will be updated in the future with newer R spatial classes and improved workflows in order to speed them up, especially tuning and cross-validation processes.

As it was already mentioned, future work will be oriented to improving the accuracy of the spatio-temporal regression kriging model for daily temperatures in the areas at higher altitudes and improving the RFSI methodology, mainly in terms of addressing its shortcomings and extension to spatio-temporal and multivariate interpolation. RFSI will further be evaluated on various environmental case studies. Along with RFSI improvements, functions for the RFSI methodology in the R package `meteo` and MeteoSerbia1km dataset will also be updated. In the future, the RFSI is planned to be used for spatio-temporal interpolation of daily climate elements at a global scale with a 1 km

spatial resolution. At first, daily maps of climate elements will be provided for the 2000–present period and then will go further into the past, as long as a sufficient number of weather stations is available. Due to lower density of the weather station network in the past, two or more models based on different weather station sources and environmental covariates will be developed for each daily climate element if needed. Daily maps will be created on a daily basis with the least possible latency and will also be aggregated to monthly and annual summaries and long term means products. The final product arising from this dissertation will be a WEB GIS portal, named WorldDailyMeteo, that will serve these high-resolution global daily climatological maps according to the Open Geospatial Consortium (OGC) standards such as Web Map Service (WMS) and Web Coverage Service (WCS). Besides visualisation of the climatological maps, WorldDailyMeteo will provide other functionalities to look up to e.g. OpenLandMap service[1], such as standard GIS functionalities, point queries on the daily climatological maps time series with chart representation, downloading entire or subset of climatological maps, and other functionalities suitable for climate analysis. The biggest challenge will be to find the fastest way for daily maps creation and to optimally design the WorldDailyMeteo portal to serve a large number of potential users.

[1] https://openlandmap.org/

Bibliography

Aalto, J., Pirinen, P., Heikkinen, J., and Venäläinen, A. (2013). Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models. *Theor. Appl. Climatol.*, 112(1-2):99–111.

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.*, 33(1):121–131.

Ahmed, S. and De Marsily, G. (1987). Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resour. Res.*

Amante, C. and Eakins, B. (2009). Etopo1 1 arc-minute global relief model: Procedures, data sources and analysis. *NCEI* `https://www.ngdc.noaa.gov/mgg/global/`.

Amit, Y. and Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Comput.*

Anandhi, A., Srinivas, V. V., Nanjundiah, R. S., and Nagesh Kumar, D. (2008). Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *Int. J. Climatol.*, 28(3):401–420.

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. *Ecol. Modell.*, 138(1-3):255–263.

Appelhans, T., Detsch, F., Reudenbach, C., and Woellauer, S. (2020). mapview: interactive viewing of spatial data in r. https://CRAN.R-project.org/package=mapview. R package version 2.9.0.

Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., and Nauss, T. (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spat. Stat.*, 14:91–113.

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Beck, H. E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K., Giraldo-Osorio, J. D., and Xuan Thinh, N. (2020). RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements. *Remote Sens. Environ.*, 239:111606.

Bajat, B., Blagojević, D., Kilibarda, M., Luković, J., and Tošić, I. (2015). Spatial analysis of the temperature trends in Serbia during the period 1961–2010. *Theor. Appl. Climatol.*, 121(1-2):289–301.

Bajat, B., Pejović, M., Luković, J., Manojlović, P., Ducić, V., and Mustafić, S. (2013). Mapping average annual precipitation in Serbia (1961–1990) by using regression kriging. *Theor. Appl. Climatol.*, 112(1-2):1–13.

Bajić, A. (1989). Severe bora on the northern adriatic. part i: Statistical analysis. *Hrvatski meteorološki časopis*, 24:1–9.

Becker, R. and Chambers, J. (1984). *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, CA, USA.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language*. Chapman & Hall, London.

Beek, E. G. (1991). Spatial interpolation of daily meteorological data: theoretical evaluation of available techniques. Report 53.1, DLO The Winand Staring Centre., Wageningen, The Nederlands.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., and MacMillan, R. A. (2018). Spatial modelling with Euclidean distance fields and machine learning. *Eur. J. Soil Sci.*, 69(5):757–770.

Belušić, D. and Bencetić Klaić, Z. (2004). Estimation of bora wind gusts using a limited area model. *Tellus A Dyn. Meteorol. Oceanogr.*, 56(4):296–307.

Benali, A., Carvalho, A., Nunes, J., Carvalhais, N., and Santos, A. (2012). Estimating air surface temperature in Portugal using MODIS LST data. *Remote Sens. Environ.*, 124:108–121.

Bénichou, P. (1994). Cartography of Statistical Pluviometric Fields with an Automatic Allowance for Regional Topography. In *Glob. Precipitations Clim. Chang.*, pages 187–199. Springer Berlin Heidelberg, Berlin, Heidelberg.

Berezowski, T., Szcześniak, M., Kardel, I., Michałowski, R., Okruszko, T., Mezghani, A., and Piniewski, M. (2016). CPLFD-GDPT5: High-resolution gridded daily precipitation and temperature data set for two largest Polish river basins. *Earth Syst. Sci. Data*, 8(1):127–139.

Beven, K. J. and Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.*, 24(1):43–69.

Bivand, R. and Gebhardt, A. (2000). Implementing functions for spatial statistical analysis using the R language. *J. Geogr. Syst.*, 2(3):307–317.

Bivand, R., Keitt, T., and Rowlingson, B. (2019). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. `https://CRAN.R-project.org/package=rgdal`. R package version 1.4-4.

Bivand, R. S. (2020). Progress in the R ecosystem for representing and handling spatial data. *J. Geogr. Syst.*

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013a). *Applied Spatial Data Analysis with R*. Springer New York, New York, NY.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013b). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Böhner, J. and Antonić, O. (2009). Land-Surface Parameters Specific to Topo-Climatology. In Hengl, T. and Reuter, H. I., editors, *Dev. Soil Sci. Geomorphometry Concepts, Software, Appl.*, chapter 8, pages 195–226. Elsevier Ltd.

Bostan, P., Heuvelink, G., and Akyurek, S. (2012). Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *Int. J. Appl. Earth Obs. Geoinf.*, 19:115–126.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24(2):123–140.

Breiman, L. (2001). Random Forests. *Mach. Learn.*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Routledge.

Brinckmann, S., Krähenmann, S., and Bissolli, P. (2016). High-resolution daily gridded data sets of air temperature and wind speed for Europe. *Earth Syst. Sci. Data*, 8(2):491–516.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geogr. Anal.*, 28(4):281–298.

Burrough, P. A. and McDonnell, R. A. (1989). *Principles of geographical information systems*. Oxford University press, Oxford.

Carrera-Hernández, J. and Gaskin, S. (2007). Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. *J. Hydrol.*, 336(3-4):231–249.

Castro, L. M., Gironás, J., and Fernández, B. (2014). Spatial estimation of daily precipitation in regions with complex relief and scarce data using terrain orientation. *J. Hydrol.*, 517:481–492.

CDS (2020). E-obs. *Climate Data Store (CDS)* `https://doi.org/10.24381/cds.151d3ec6`.

Čeh, M., Kilibarda, M., Lisec, A., and Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS Int. J. Geo-Information*, 7(5):168.

Chambers, J. M. (1998). *Programming with Data: A Guide to the S Language*. Springer-Verlag, New York.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*. Chapman & Hall, London.

Chapman, L. and Thornes, J. E. (2003). The use of geographical information systems in climatology and meteorology. *Prog. Phys. Geogr.*, 27(3):313–330.

Chen, S.-T., Yu, P.-S., and Tang, Y.-H. (2010). Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *J. Hydrol.*, 385(1-4):13–22.

Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty: Second Edition*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Chorley, R. J. and Haggett, P. (1965). Trend-Surface Mapping in Geographical Research. *Trans. Inst. Br. Geogr.*, 37:47–67.

Cindrić, K., Pasarić, Z., and Gajić-Čapka, M. (2010). Spatial and temporal analysis of dry spells in Croatia. *Theor. Appl. Climatol.*, 102(1-2):171–184.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1):37–46.

Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council (2008). *Earth Observations from Space*. National Academies Press, Washington, D.C.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, C. f. O. . A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J. (2011). The Twentieth Century Reanalysis Project. *Q. J. R. Meteorol. Soc.*, 137(654):1–28.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M., and Jones, P. D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *J. Geophys. Res. Atmos.*, 123(17):9391–9409.

Courault, D. and Monestiez, P. (1999). Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast France. *Int. J. Climatol.*, 19(4):365–378.

Cressman, G. P. (1959). An operational objective analysis system. *Mon. Weather Rev.*, 87(10):367–374.

Czernecki, B., Głogowski, A., and Nowosad, J. (2020). Climate: An R Package to Access Free In-Situ Meteorological and Hydrological Datasets For Environmental Assessment. *Sustainability*, 12(1):394.

da Silva Júnior, J. C., Medeiros, V., Garrozi, C., Montenegro, A., and Gonçalves, G. E. (2019). Random forest techniques for spatial interpolation of evapotranspiration data from Brazilian's Northeast. *Comput. Electron. Agric.*, 166:105017.

Daly, C., Neilson, R. P., and Phillips, D. L. (1994). A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. *J. Appl. Meteorol.*, 33(2):140–158.

Davies, M. M. and van der Laan, M. J. (2016). Optimal Spatial Prediction Using Ensemble Machine Learning. *Int. J. Biostat.*, 12(1):179–201.

de Wit, A. and van Diepen, C. (2008). Crop growth modelling and crop yield forecasting using satellite-derived meteorological inputs. *Int. J. Appl. Earth Obs. Geoinf.*, 10(4):414–425.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, 137(656):553–597.

Dhakal, K., Kakani, V. G., Ochsner, T. E., and Sharma, S. (2020). Constructing retrospective gridded daily weather data for agro-hydrological applications in Oklahoma. *Agrosystems, Geosci. Environ.*, 3(1).

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., and Zijlstra, T. (2015). *Open Data Handbook Documentation.* Open Knowledge.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics.* Springer Series in Statistics. Springer New York, New York, NY.

Dobesch, H., Dumolard, P., and Dyras, I. (2007). *Spatial Interpolation for Climate Data.* ISTE, London, UK.

Dodson, R. and Marks, D. (1997). Daily air temperature interpolated at high spatial resolution over a large mountainous region. *Clim. Res.*, 8(1):1–20.

dos Santos, R. S. (2020). Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. *Int. J. Appl. Earth Obs. Geoinf.*

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S. (2010). Comprehensive Automated Quality Assurance of Daily Surface Observations. *J. Appl. Meteorol. Climatol.*, 49(8):1615–1633.

Durre, I., Menne, M. J., and Vose, R. S. (2008). Strategies for Evaluating Quality Assurance Procedures. *J. Appl. Meteorol. Climatol.*, 47(6):1785–1791.

Elseberg, J., Magnenat, S., Siegwart, R., and Andreas, N. (2012). Comparison of nearest-neighbor-search strategies and implementations for efficient shape registration. *J. Softw. Eng. Robot.*, 3(1):2–12.

Emamifar, S., Rahimikhoob, A., and Noroozi, A. A. (2013). Daily mean air temperature estimation from MODIS land surface temperature products based on M5 model tree. *Int. J. Climatol.*, 33(15):3174–3181.

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., and Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Convers. Manag.*, 164:102–111.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.*, 37(12):4302–4315.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. E. (1998). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester.

Frei, C. (2014). Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol.*, 34(5):1585–1605.

Frick, C., Steiner, H., Mazurkiewicz, A., Riediger, U., Rauthe, M., Reich, T., and Gratzki, A. (2014). Central European high-resolution gridded daily data sets (HYRAS): Mean temperature and relative humidity. *Meteorol. Zeitschrift*, 23(1):15–32.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 29(5):1189–1232.

Gandin, L. S. (1965). Objective Analysis of meteorological fields. *Isr. Progr. Sci. Transl.*, page 242.

Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., and Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The Cook Agronomy Farm data set. *Spat. Stat.*, 14:70–90.

GDAL/OGR contributors (2020). GDAL/OGR geospatial data abstraction software library.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.*, 0(0):1–16.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.*, 63(1):3–42.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press, Oxford.

Goovaerts, P. (2000). Estimation or simulation of soil properties? An optimization problem with conflicting criteria. *Geoderma*, 97(3-4):165–186.

Graler, B., Pebesma, E., and Heuvelink, G. (2017). *Encyclopedia of GIS*. Springer International Publishing, Cham.

Gräler, B., Rehr, M., Gerharz, L., and Pebesma, E. (2013). Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009. *ETC/ACM Tech. Pap. 2012/8*.

Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218.

Guan, B., Hsu, H., Wey, T., and Tsao, L. (2009). Modeling monthly mean temperatures for the mountain regions of Taiwan by generalized additive models. *Agric. For. Meteorol.*, 149(2):281–290.

Hartkamp, a. D., De Beurs, K., Stein, A., and White, J. W. (1999). *Interpolation Techniques for Climate Variables Interpolation*. NRG-GIS Series 99-01. D.F.: CIMMYT, Mexico.

Hashimoto, H., Wang, W., Melton, F. S., Moreno, A. L., Ganguly, S., Michaelis, A. R., and Nemani, R. R. (2019). High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States. *Int. J. Climatol.*, 39(6):2964–2983.

Haslinger, K., Koffler, D., Schöner, W., and Laaha, G. (2014). Exploring the link between meteorological drought and streamflow: Effects of climate-catchment interaction. *Water Resour. Res.*, 50(3):2468–2487.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Stat. Sci.*, 1(3):297–310.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, New York, NY.

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 1950–2006. *J. Geophys. Res.*, 113(D20):D20119.

He, X., Chaney, N. W., Schleiss, M., and Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resour. Res.*, 52(10):8217–8237.

Hengl, T. (2018). Worldgrids archived layers at 1 km to 20 km spatial resolution (version v0.2). *ZENODO* `http://doi.org/10.5281/zenodo.1637816`.

Hengl, T., Heuvelink, G. B., and Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301–1315.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes de Jesus, J., Tamene, L., and Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PLoS One*, 10(6):e0125814.

Hengl, T., Heuvelink, G. B. M., Perčec Tadić, M., and Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theor. Appl. Climatol.*, 107(1-2):265–277.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., and Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518.

Herrera, S., Gutiérrez, J. M., Ancell, R., Pons, M. R., Frías, M. D., and Fernández, J. (2012). Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02). *Int. J. Climatol.*, 32(1):74–85.

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. In *Spat. Des.*, pages 207–230. John Wiley & Sons, Ltd, Chichester, UK.

Heuvelink, G. B. M., Pebesma, E., and Gräler, B. (2017). Space-Time Geostatistics. In Shekhar, S., Xiong, H., and Zhou, X., editors, *Encycl. GIS*, pages 1919–1926. Springer International Publishing, Cham.

Hiebl, J., Auer, I., Böhm, R., Schöner, W., Maugeri, M., Lentini, G., Spinoni, J., Brunetti, M., Nanni, T., Tadić, M., Perčec Bihari, Z., Dolinar, M., and Müller-Westermeier, G. (2009). A high-resolution 19611990 monthly temperature climatology for the greater Alpine region. *Meteorol. Zeitschrift*, 18(5):507–530.

Hiebl, J. and Frei, C. (2016). Daily temperature grids for Austria since 1961—concept, creation and applicability. *Theor. Appl. Climatol.*, 124(1-2):161–178.

Hijmans, R. J. (2019). raster: Geographic Data Analysis and Modeling. `https://CRAN.R-project.org/package=raster`. R package version 2.9-23.

Hijmans, R. J. (2020). terra: spatial data analysis. https://CRAN.R-project.org/package=terra. R package version 0.8-6.

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.*, 113(D21):D21110.

Holden, Z. A., Swanson, A., Klene, A. E., Abatzoglou, J. T., Dobrowski, S. Z., Cushman, S. A., Squires, J., Moisen, G. G., and Oyler, J. W. (2016). Development of high-resolution (250 m) historical daily gridded air temperature data using reanalysis and distributed sensor networks for the US Northern Rocky Mountains. *Int. J. Climatol.*, 36(10):3620–3632.

Horvath, K., Ivatek-Šahdan, S., Ivančan-Picek, B., and Grubišić, V. (2009). Evolution and Structure of Two Severe Cyclonic Bora Events: Contrast between the Northern and Southern Adriatic. *Weather Forecast.*, 24(4):946–964.

Huang, R., Zhang, C., Huang, J., Zhu, D., Wang, L., and Liu, J. (2015). Mapping of Daily Mean Air Temperature in Agricultural Regions Using Daytime and Nighttime Land Surface Temperatures Derived from TERRA and AQUA MODIS Data. *Remote Sens.*, 7(7):8728–8756.

Hudson, G. and Wackernagel, H. (1994). Mapping temperature using kriging with external drift: Theory and an example from scotland. *Int. J. Climatol.*, 14(1):77–91.

Huffman, G. J., Bolvin, D. T., and Nelkin, E. J. (2014). Integrated Multi-satellitE Retrievals for GPM (IMERG), Final Run, version V06B. `ftp://arthurhou.pps.eosdis.nasa.gov/gpmdata/`. Accessed: 31 July, 2019.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J. Hydrometeorol.*, 8(1):38–55.

Hunter, R. D. and Meentemeyer, R. K. (2005). Climatologically Aided Mapping of Daily Precipitation and Temperature. *J. Appl. Meteorol.*, 44(10):1501–1510.

Hutchinson, M. F. (1995). Interpolating mean rainfall using thin plate smoothing splines. *Int. J. Geogr. Inf. Syst.*, 9(4):385–403.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., and Papadopol, P. (2009). Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum–Maximum Temperature and Precipitation for 1961–2003. *J. Appl. Meteorol. Climatol.*, 48(4):725–741.

Isaaks, E. and Srivastava, R. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860.

Islam, T., Hulley, G. C., Malakar, N. K., Radocinski, R. G., Guillevic, P. C., and Hook, S. J. (2017). A Physics-Based Algorithm for the Simultaneous Retrieval of Land Surface Temperature and Emissivity From VIIRS Thermal Infrared Data. *IEEE Trans. Geosci. Remote Sens.*, 55(1):563–576.

Ivatek-Sahdan, S. and Ivancan-Picek, B. (2006). Effects of different initial and boundary conditions in ALADIN/HR simulations during MAP IOPs. *Meteorol. Zeitschrift*, 15(2):187–197.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Janatian, N., Sadeghi, M., Sanaeinejad, S. H., Bakhshian, E., Farid, A., Hasheminia, S. M., and Ghazanfari, S. (2017). A statistical framework for estimating air temperature using MODIS land surface temperature data. *Int. J. Climatol.*, 37(3):1181–1194.

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. *J. Appl. Meteorol.*, 40(6):1075–1084.

Jeong, H.-G., Ahn, J.-B., Lee, J., Shim, K.-M., and Jung, M.-P. (2020). Improvement of daily precipitation estimations using PRISM with inverse-distance weighting. *Theor. Appl. Climatol.*, 139(3-4):923–934.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *J. Int. Assoc. Math. Geol.*, 15(3):445–468.

Juran, I., Grubišić, D., Štivičić, A., and Čuljak, T. G. (2020). Which factors predict stem weevils appearance in rapeseed crops? *J. Entomol. Res. Soc.*, 22(2):203–210.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.*, 77(3):437–471.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., and Potter, G. L. (2002). NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Am. Meteorol. Soc.*, 83(11):1631–1644.

Kanevski, M., Pozdnoukhov, A., and Timonin, V. (2009). *Machine learning for spatial environmental data: Theory, applications and software*. EPFL Press.

Kilibarda, M. (2013). *AUTOMATED MAPPING OF CLIMATIC VARIABLES USING SPATIO-TEMPORAL GEOSTATISTICAL METHODS*. Phd thesis, Faculty of Civil Engineering, University of Belgrade.

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. B. M., Gräler, B., Pebesma, E., Perčec Tadić, M., and Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *J. Geophys. Res. Atmos.*, 119(5):2294–2313.

Kilibarda, M., Tadić, M. P., Hengl, T., Luković, J., and Bajat, B. (2015). Global geographic and feature space coverage of temperature data in the context of spatio-temporal interpolation. *Spat. Stat.*, 14:22–38.

Kirkwood, C., Cave, M., Beamish, D., Grebby, S., and Ferreira, A. (2016). A machine learning approach to geochemical mapping. *J. Geochemical Explor.*, 167:49–61.

Kitanidis, P. K. (1993). Generalized covariance functions in estimation. *Mathematical Geology*, 25(5):525–540.

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.

Kloog, I., Nordio, F., Coull, B. A., and Schwartz, J. (2014). Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. *Remote Sens. Environ.*, 150:132–139.

Knaus, J. (2015). snowfall: Easier cluster computing (based on snow). R package version 1.84-6.1.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proc. 14th Int. Jt. Conf. Artif. Intell. - Vol. 2*, pages 1137– 1143, San Francisco. Morgan Kaufmann Publishers Inc.

Kovačević, J., Cvijetinović, Ž., Lakušić, D., Kuzmanović, N., Šinžar-Sekulić, J., Mitrović, M., Stančić, N., Brodić, N., and Mihajlović, D. (2020). Spatio-temporal classification framework for mapping woody vegetation from multi-temporal sentinel-2 imagery. *Remote Sens.*, 12(17):1–23.

Krähenmann, S. and Ahrens, B. (2013). Spatial gridding of daily maximum and minimum 2 m temperatures supported by satellite observations. *Meteorol. Atmos. Phys.*, 120(1-2):87–105.

Krige, D. G. (1951). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *J. South. African Inst. Min. Metall.*, 52(6):119–139.

Kuhn, M. (2019). caret: Classification and Regression Training. `https://CRAN.R-project.org/package=caret`. R package version 6.0-84.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York, New York, NY.

Kuhn, M. and Quinlan, R. (2020). *Cubist: Rule- And Instance-Based Regression Modeling*. R package version 0.2.3.

Kurtzman, D. and Kadmon, R. (1999). Mapping of temperature variables in Israel: a comparison of different interpolation methods. *Clim. Res.*, 13(1):33–43.

Lee, M., Im, E., and Bae, D. (2019). Impact of the spatial variability of daily precipitation on hydrological projections: A comparison of GCM- and RCM-driven cases in the Han River basin, Korea. *Hydrol. Process.*, 33(16):2240–2257.

Li, J. and Heap, A. D. (2008). *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia, Canberra, Australia, record 200 edition.

Li, J. and Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environ. Model. Softw.*, 53:173–189.

Li, J., Heap, A. D., Potter, A., and Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environ. Model. Softw.*, 26(12):1647–1659.

Li, X., Zhou, Y., Asrar, G. R., and Zhu, Z. (2018a). Creating a seamless 1 km resolution daily land surface temperature dataset for urban and surrounding areas in the conterminous United States. *Remote Sens. Environ.*, 206(January):84–97.

Li, X., Zhou, Y., Asrar, G. R., and Zhu, Z. (2018b). Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. *Remote Sens. Environ.*, 215:74–84.

Liao, Y., Li, D., and Zhang, N. (2018). Comparison of interpolation models for estimating heavy metals in soils under various spatial characteristics and sampling methods. *Trans. GIS*, 22(2):409–434.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255.

Lin, T., Zhong, R., Wang, Y., Xu, J., Jiang, H., Xu, J., Ying, Y., Rodriguez, L., Ting, K. C., and Li, H. (2020). DeepCropNet: a deep spatial-temporal learning framework for county-level corn yield estimation. *Environ. Res. Lett.*, 15(3):034016.

Long, J., Liu, Y., Xing, S., Zhang, L., Qu, M., Qiu, L., Huang, Q., Zhou, B., and Shen, J. (2020). Optimal interpolation methods for farmland soil organic matter in various landforms of a complex topography. *Ecol. Indic.*, 110(November 2019):105926.

Lukovic, J., Chiang, J., Blagojevic, D., and Sekulić, A. (2021). A later onset of the rainy season in California. *Geophys. Res. Lett.*

MacCormack, K. E., Brodeur, J. J., and Eyles, C. H. (2013). Evaluating the impact of data quantity, distribution and algorithm selection on the accuracy of 3D subsurface models using synthetic grid models of varying complexity. *J. Geogr. Syst.*, 15(1):71–88.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS One*, 13(3):e0194889.

Malamos, N. and Koutsoyiannis, D. (2016). Bilinear surface smoothing for spatial interpolation with optional incorporation of an explanatory variable. Part 2: Application to synthesized and rainfall data. *Hydrol. Sci. J.*, 61(3):527–540.

Marshall, M., Tu, K., and Brown, J. (2018). Optimizing a remote sensing production efficiency model for macro-scale GPP and yield estimation in agroecosystems. *Remote Sens. Environ.*, 217:258–271.

Matheron, G. (1963). Principles of geostatistics. *Econ. Geol.*, 58(8):1246–1266.

McAlpine, C. A., Johnson, A., Salazar, A., Syktus, J., Wilson, K., Meijaard, E., Seabrook, L., Dargusch, P., Nordin, H., and Sheil, D. (2018). Forest loss and Borneo's climate. *Environ. Res. Lett.*

Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999.

Mendelsohn, R., Kurukulasuriya, P., Basist, A., Kogan, F., and Williams, C. (2007). Climate analysis with satellite versus weather station data. *Clim. Change*, 81(1):71–83.

Méndez, M. and Calvo-Valverde, L. (2020). Comparison performance of machine learning and geo-statistical methods for the interpolation of monthly air temperature over Costa Rica. *IOP Conf. Ser. Earth Environ. Sci.*, 432:012011.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An Overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Ocean. Technol.*, 29(7):897–910.

Meyer, H. (2018). CAST: 'caret' Applications for Spatial-Temporal Models. `https://cran.r-project.org/package=CAST`. R package version 0.3.1.

Meyer, H., Katurji, M., Appelhans, T., Müller, M., Nauss, T., Roudier, P., and Zawar-Reza, P. (2016). Mapping Daily Air Temperature for Antarctica Based on MODIS LST. *Remote Sens.*, 8(9):732.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ. Model. Softw.*, 101(November):1–9.

Microsoft and Weston, S. (2017). foreach: Provides foreach looping construct for r. R package version 1.4.4.

Microsoft Corporation and Steve Weston (2019). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. `https://CRAN.R-project.org/package=doParallel`. R package version 1.0.15.

Mitas, L. and Mitasova, H. (1999). Spatial Interpolation. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geogr. Inf. Syst. Princ. Tech. Manag. Appl.*, chapter 34, pages 481–492. Wiley.

Mohsenzadeh Karimi, S., Kisi, O., Porrajabali, M., Rouhani-Nia, F., and Shiri, J. (2018). Evaluation of the support vector machine, random forest and geo-statistical methodologies for predicting long-term air temperature. *ISH J. Hydraul. Eng.*, 00(00):1–11.

Møller, A. B., Beucher, A. M., Pouladi, N., and Greve, M. H. (2020). Oblique geographic coordinates as covariates for digital soil mapping. *SOIL*, 6(2):269–289.

Muñoz Sabater, J. (2019). ERA5-Land hourly data from 1981 to present.

Nashwan, M. S., Shahid, S., and Chung, E.-S. (2019). Development of high-resolution daily gridded temperature datasets for the central north region of Egypt. *Sci. Data*, 6(1):138.

Nevtipilova, V., Pastwa, J., Boori, M. S., and Vozenilek, V. (2014). Testing Artificial Neural Network (ANN) for Spatial Interpolation. *J. Geol. Geosci.*, 03(02):1–9.

Nguyen, P., Shearer, E. J., Tran, H., Ombadi, M., Hayatbini, N., Palacios, T., Huynh, P., Braithwaite, D., Updegraff, G., Hsu, K., Kuligowski, B., Logan, W. S., and Sorooshian, S. (2019). The CHRS Data Portal, an easily accessible public repository for PERSIANN global satellite precipitation data. *Sci. Data*, 6(1):180296.

Noi, P., Degener, J., and Kappas, M. (2017). Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data. *Remote Sens.*, 9(5):398.

Odeh, I., McBratney, A., and Chittleborough, D. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67(3-4):215–226.

Omre, H. (1987). Bayesian kriging-Merging observations and qualified guesses in kriging. *Math. Geol.*, 19(1):25–39.

Osborn, T. J. and Jones, P. D. (2014). The CRUTEM4 land-surface air temperature data set: construction, previous versions and dissemination via Google Earth. *Earth Syst. Sci. Data*, 6(1):61–68.

Oyler, J. W., Ballantyne, A., Jencso, K., Sweet, M., and Running, S. W. (2015). Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *Int. J. Climatol.*, 35(9):2258–2279.

Oyler, J. W., Dobrowski, S. Z., Holden, Z. A., and Running, S. W. (2016). Remotely Sensed Land Skin Temperature as a Spatial Predictor of Air Temperature across the Conterminous United States. *J. Appl. Meteorol. Climatol.*, 55(7):1441–1457.

Pang, B., Yue, J., Zhao, G., and Xu, Z. (2017). Statistical Downscaling of Temperature with the Random Forest Model. *Adv. Meteorol.*, 2017:1–11.

Parmentier, B., McGill, B., Wilson, A., Regetz, J., Jetz, W., Guralnick, R., Tuanmu, M.-N., Robinson, N., and Schildhauer, M. (2014). An Assessment of Methods and Remote-Sensing Derived Covariates for Regional Predictions of 1 km Daily Maximum Air Temperature. *Remote Sens.*, 6(9):8639–8670.

Parmentier, B., McGill, B. J., Wilson, A. M., Regetz, J., Jetz, W., Guralnick, R., Tuanmu, M.-N., and Schildhauer, M. (2015). Using multi-timescale methods and satellite-derived land surface temperature for the interpolation of daily maximum air temperature in Oregon. *Int. J. Climatol.*, 35(13):3862–3878.

Pebesma, E. (2012). spacetime: Spatio-temporal data in R. *Journal of Statistical Software*, 51(7):1–30.

Pebesma, E. (2018a). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

Pebesma, E. (2018b). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

Pebesma, E. (2020). stars: spatiotemporal arrays, raster and vector data cubes. https://CRAN.R-project.org/package=stars. R package version 0.4-3.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Comput. Geosci.*, 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.

Pejović, M., Nikolić, M., Heuvelink, G. B., Hengl, T., Kilibarda, M., and Bajat, B. (2018). Sparse regression interaction models for spatial prediction of soil properties in 3D. *Comput. Geosci.*, 118(March):1–13.

Perry, M. and Hollis, D. (2005). The generation of monthly gridded datasets for a range of climatic variables over the UK. *Int. J. Climatol.*, 25(8):1041–1054.

Perčec Tadić, M. (2010). Gridded Croatian climatology for 1961–1990. *Theor. Appl. Climatol.*, 102(1-2):87–103.

Petritsch, R. and Hasenauer, H. (2014). Climate input parameters for real-time online risk assessment. *Nat. Hazards*, 70(3):1749–1762.

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

PROJ contributors (2020). PROJ coordinate transformation software library.

PSL (2020a). CPC Global Daily Temperature. *Physical Sciences Laboratory (PSL), NOAA* `https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html`.

PSL (2020b). CPC Global Unified Gauge-Based Analysis of Daily Precipitation. *Physical Sciences Laboratory (PSL), NOAA* `https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html`.

Qiao, P., Li, P., Cheng, Y., Wei, W., Yang, S., Lei, M., and Chen, T. (2019). Comparison of common spatial interpolation methods for analyzing pollutant spatial distributions at contaminated sites. *Environ. Geochem. Health*, 41(6):2709–2730.

R Core Team (2020). R: A language and environment for statistical computing.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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 J. Photogramm. Remote Sens.*, 57(4):241–262.

Razafimaharo, C., Krähenmann, S., Höpp, S., Rauthe, M., and Deutschländer, T. (2020). New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS). *Theor. Appl. Climatol.*

Reuter, H. I. and Hengl, T. (2012). Worldgrids — a public repository of global soil covariates. In Minasny, B., Malone, B. P., and McBratney, A. B., editors, *Digit. soil assessments beyond Proc. fifth Glob. Work. Digit. Soil Mapping, Sydney, Aust. 10-13 April 2012*. CRC Press.

Rigol, J. P., Jarvis, C. H., and Stuart, N. (2001). Artificial neural networks as a tool for spatial interpolation. *Int. J. Geogr. Inf. Sci.*

Rivoirard, J. (1994). *Introduction to disjunctive kriging and non-linear geostatistics*. Oxford University Press, Oxford.

Rosenfeld, A., Dorman, M., Schwartz, J., Novack, V., Just, A. C., and Kloog, I. (2017). Estimating daily minimum, maximum, and mean near surface air temperature using hybrid satellite models across Israel. *Environ. Res.*, 159(March):297–312.

Roznik, M., Brock Porth, C., Porth, L., Boyd, M., and Roznik, K. (2019). Improving agricultural microinsurance by applying universal kriging and generalised additive models for interpolation of mean daily temperature. *Geneva Pap. Risk Insur. - Issues Pract.*, 44(3):446–480.

Ruiz-Álvarez, M., Alonso-Sarria, F., and Gomariz-Castillo, F. (2019). Interpolation of instantaneous air temperature using geographical and MODIS derived variables with machine learning techniques. *ISPRS Int. J. Geo-Information*.

Rumelhart, D., Hinton, G., and Williams, R. (1986). *Learning Internal Representations By Error Propagation*. The MIT Press.

Ryan, J. A. and Ulrich, J. M. (2020). xts: extensible time series. https://CRAN.R-project.org/package=xts. R package version 0.12-0.

Samardžić-Petrović, M., Dragićević, S., Kovačević, M., and Bajat, B. (2016). Modeling Urban Land Use Changes Using Support Vector Machines. *Trans. GIS*, 20(5):718–734.

Schuurmans, J. M., Bierkens, M. F. P., Pebesma, E. J., and Uijlenhoet, R. (2007). Automatic Prediction of High-Resolution Daily Rainfall Fields for Multiple Extents: The Potential of Operational Radar. *J. Hydrometeorol.*, 8(6):1204–1224.

Schwarb, M., Daly, C., Frei, C., and Schär, C. (2001). Mean annual and seasonal precipitation throughout the European Alps 1971-1990. In Sperafico, R., Weingartner, R., and Leibundgut, C., editors, *Hydrol. Atlas Switzerland. Landeshydrologie und Geol.*, chapter 2.6 and 2. Institute of Geography of University Bern.

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., and Bajat, B. (2020a). Random Forest Spatial Interpolation. *Remote Sens.*, 12(10):1687.

Sekulić, A., Kilibarda, M., Protić, D., Tadić, M. P., and Bajat, B. (2020b). Spatio-temporal regression kriging model of mean daily temperature for Croatia. *Theor. Appl. Climatol.*, 140(1-2):101–114.

Sekulić, A., Kilibarda, M., Protić, D., and Bajat, B. (2020). Meteoserbia1km: the first daily gridded meteorological dataset at a 1-km spatial resolution across serbia for the 2000–2019 period. *ZENODO* `http://doi.org/10.5281/zenodo.4058167`.

Seo, Y., Kim, S., and Singh, V. P. (2015). Estimating Spatial Precipitation Using Regression Kriging and Artificial Neural Network Residual Kriging (RKNNRK) Hybrid Approach. *Water Resour. Manag.*, 29(7):2189–2204.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proc. 23rd ACM Natl. Conf. Assoc. Comput. Mach.*, pages 517–524, Princeton, New Jersey, USA. ACM Press.

Sibson, R. (1981). A Brief Description of Natural Neighbour Interpolation. In Barnett, V., editor, *Interpret. Multivar. data*, pages 21–36. John Wiley & Sons, Ltd, Chichester, UK.

Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E., and Knutti, R. (2020). Climate change now detectable from any single day of weather at global scale. *Nat. Clim. Chang.*, 10(1):35–41.

Sluiter, R. (2009). Interpolation methods for climate data — literature review. Intern rapport, KNMI Royal Netherlands Meteorological Institute, De Bilt, The Netherlands.

Srivastava, A. K., Rajeevan, M., and Kshirsagar, S. R. (2009). Development of a high resolution daily gridded temperature data set (1969-2005) for the Indian region. *Atmos. Sci. Lett.*, 10(4):249–254.

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. *Agric. For. Meteorol.*, 139(3-4):224–236.

Stewart, S. B. and Nitschke, C. R. (2017). Improving temperature interpolation using MODIS LST and local topography: a comparison of methods in south east Australia. *Int. J. Climatol.*, 37(7):3098–3110.

Sullivan, J. (1984). Conditional Recovery Estimation Through Probability Kriging — Theory and Practice. In Verly, G., David, M., Journel, A. G., and Marechal, A., editors, *Geostatistics Nat. Resour. Charact.*, pages 365–384. Springer Netherlands, Dordrecht.

Sun, W., Zhu, Y., Huang, S., and Guo, C. (2015). Mapping the mean annual precipitation of China using local interpolation techniques. *Theor. Appl. Climatol.*, 119(1-2):171–180.

Szalai, S., Auer, I., Hiebl, J., Milkovich, J., Radim, T., Stepanek, P., Zahradnicek, P., Bihari, Z., Lakatos, M., Szentimrey, T., Limanowka, D., Kilar, P., Cheval, S., Deak, G., Mihic, D., Antolovic, I., Mihajlovic, V., Nejedlik, P., Stastny, P., and Mikulov, J. (2013). Climate of the Greater Carpathian Region. Final technical report, European Commission, Joint Research Centre (JRC).

Tait, A., Henderson, R., Turner, R., and Zheng, X. (2006). Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *Int. J. Climatol.*, 26(14):2097–2115.

Tank, A. K., Zwiers, F. W., and Zhang, X. (2009). Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation. Technical Report WCDMP-No. 72, WMO-TD No. 1500, World Meteorological Organization.

Tennekes, M. (2018). tmap : Thematic Maps in R. *J. Stat. Softw.*, 84(6).

Tennekes, M. (2020). tmap: thematic maps. https://CRAN.R-project.org/package=tmap. R package version 3.1.

Thiessen, A. H. (1911). Precipitation averages for large areas. *Mon. Weather Rev.*, 39(7):1082–1089.

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2018). snow: Simple network of workstations. R package version 0.4-3.

Tripathi, S., Srinivas, V., and Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: A support vector machine approach. *J. Hydrol.*, 330(3-4):621–640.

Tveito, O. E., Wegehenkel, M., van der Wel, F., and Dobesch, H. (2006). The use of geographical information systems in climatology and meteorology. Final report, COST Action 719.

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.*, 116(D11):D11110.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer New York, New York, NY.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York.

Veronesi, F. and Schillaci, C. (2019). Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. *Ecol. Indic.*, 101(December 2018):1032–1044.

Wackernagel, H. (2003). *Multivariate Geostatistics*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wadoux, A. M., Brus, D. J., and Heuvelink, G. B. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*, 355:113913.

Wadoux, A. M., Heuvelink, G. B., Uijlenhoet, R., and de Bruin, S. (2020). Optimization of rain gauge sampling density for river discharge prediction using Bayesian calibration. *PeerJ*, 8:e9558.

Wahba, G. and Wendelberger, J. (1980). Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation. *Mon. Weather Rev.*, 108(8):1122–1143.

Wan, Z. (2006). MODIS land surface temperature products users' guide. *ICESS, Univ. Calif.*

Wan, Z. and Dozier, J. (1996). A generalized split-window algorithm for retrieving land-surface temperature from space. *IEEE Trans. Geosci. Remote Sens.*, 34(4):892–905.

Wan, Z. and Li, Z.-L. (1997). A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. *IEEE Trans. Geosci. Remote Sens.*, 35(4):980–996.

Wang, M., He, G., Zhang, Z., Wang, G., Zhang, Z., Cao, X., Wu, Z., and Liu, X. (2017). Comparison of Spatial Interpolation and Regression Analysis Models for an Estimation of Monthly Near Surface Air Temperature in China. *Remote Sens.*, 9(12):1278.

Webster, R. (2000). Is soil variation random? *Geoderma*, 97(3-4):149–163.

Webster, R. and Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. Statistics in Practice. John Wiley & Sons, Ltd, Chichester, UK.

Werner, A. T., Schnorbus, M. A., Shrestha, R. R., Cannon, A. J., Zwiers, F. W., Dayon, G., and Anslow, F. (2019). A long-term, temporally consistent, gridded daily meteorological dataset for northwestern North America. *Sci. Data*, 6(1):180299.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *J. Stat. Softw.*, 40(1).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Williamson, S., Hik, D., Gamon, J., Kavanaugh, J., and Flowers, G. (2014). Estimating Temperature Fields from MODIS Land Surface Temperature and Air Temperature Observations in a Sub-Arctic Alpine Environment. *Remote Sens.*, 6(2):946–963.

Willmott, C. J., Rowe, C. M., and Philpot, W. D. (1985). Small-Scale Climate Maps: A Sensitivity Analysis of Some Common Assumptions Associated with Grid-Point Interpolation and Contouring. *Am. Cartogr.*, 12(1):5–16.

World Meteorological Organization (WMO) (2018). *Guide to Climatological Practices WMO-No. 100*. World Meteorological Organization (WMO), Geneva, Switzerland, 2018 edition.

Wright, M. N. and Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.*, 77(1).

Wu, T. and Li, Y. (2013). Spatial interpolation of temperature in the United States using residual kriging. *Appl. Geogr.*, 44:112–120.

Xavier, A. C., King, C. W., and Scanlon, B. R. (2016). Daily gridded meteorological variables in Brazil (1980-2013). *Int. J. Climatol.*, 36(6):2644–2659.

Xu, J., Zhang, F., Jiang, H., Hu, H., Zhong, K., Jing, W., Yang, J., and Jia, B. (2020). Downscaling Aster Land Surface Temperature over Urban Areas with Machine Learning-Based Area-To-Point Regression Kriging. *Remote Sens.*, 12(7):1082.

Xu, Y., Knudby, A., and Ho, H. C. (2014). Estimating daily maximum air temperature from MODIS in British Columbia, Canada. *Int. J. Remote Sens.*, 35(24):8108–8121.

Yanto, Livneh, B., and Rajagopalan, B. (2017). Development of a gridded meteorological dataset over Java island, Indonesia 1985–2014. *Sci. Data*, 4(1):170072.

Yuan, W., Xu, B., Chen, Z., Xia, J., Xu, W., Chen, Y., Wu, X., and Fu, Y. (2015). Validation of China-wide interpolated daily climate variables from 1960 to 2011. *Theor. Appl. Climatol.*, 119(3-4):689–700.

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

Zhang, X., Liu, G., Wang, H., and Li, X. (2017). Application of a Hybrid Interpolation Method Based on Support Vector Machine in the Precipitation Spatial Interpolation of Basins. *Water*, 9(10):760.

Zhu, W., Lű, A., and Jia, S. (2013). Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products. *Remote Sens. Environ.*, 130:62–73.

Zhu, X., Zhang, Q., Xu, C.-Y., Sun, P., and Hu, P. (2019). Reconstruction of high spatial resolution surface air temperature data across China: A new geo-intelligent multisource data-based machine learning technique. *Sci. Total Environ.*, 665:300–313.

Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M. P. (1999). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Math. Geol.*, 31(4):375–390.

Biography

Aleksandar Sekulić was born on January 30th, 1991 in Pančevo, Serbia. He finished the "Stevica Jovanović" elementary school in 2005 in Pančevo and "Uroš Predić" Grammar school (natural sciences and mathematics) in 2009 also in Pančevo. He took part in the state competition in mathematics twice and won one 2nd, and one 3rd place in regional competitions. In 2009, he started study of Geodesy at University of Belgrade, Faculty of Civil Engineering, Department of Geodesy and Geoinformatics. In 2012 he finished BSc studies with an average grade of 9,47 (max 10.00) and thesis titled "*3D Modeling in Software Environment Cadcorp 7.1*". In 2014 he finished MSc studies with an average grade of 9,72 (max 10.00) and thesis titled "Application of Transport Network and pgRouting Module for the Purposes of Vehicle Routing". During his studies he received three awards for excellent results during the study from Faculty, and in 2014, he received the award of the Institute of Geodesy and Geoinformatics for the best master thesis at the Department of Geodesy and Geoinformatics of the 2013/2014 school year.

In 2014, he started his PhD studies, within the same University. During the studies he passed exams with the average grade of 9.88 (max 10.00) and started with research related to spatio-temporal interpolation of daily climate elements. In January 2021, he submitted PhD dissertation entitled "*Spatio-temporal interpolation of climate elements using geostatistics and machine learning*". During his PhD study Aleksandar Sekulić published as author or co-author papers related to spatio-temporal interpolation of climate elements: 4 journal papers (from SCI list), 1 in Serbian journals, 7 international conference papers, 1 Serbian conference paper, and 2 technical solutions.

Since November 2014, he has been a teaching assistant in the field of Engineering Geodesy, GIS and Geostatistics at University of Belgrade, Faculty of Civil Engineering, Department of Geodesy and Geoinformatics. He participated as a researcher in 2 Horizon 2020 projects funded by the European Commission (APOLLO and BEACON), in 2 research projects funded by the Serbian Ministry of Science (III47014 and CERES), and one ERASMUS+ project (GEOWEB).

Prilozi

Prilog 1: Izjava o autorstvu

Prilog 2: Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Prilog 3: Izjava o korišćenju

Изјава о ауторству

Име и презиме аутора Александар Секулић

Број индекса 908/14

Изјављујем

да је докторска дисертација под насловом

SPATIO-TEMPORAL INTERPOLATION OF CLIMATE ELEMENTS USING GEOSTATISTICS AND MACHINE LEARNING (PROSTORNO-VREMENSKA INTERPOLACIJA KLIMATSKIH ELEMENATA PRIMENOM GEOSTATISTIKE I MAŠINSKOG UČENJA)

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

Потпис аутора

У Београду, _____ _____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Александар Секулић

Број индекса 908/14

Студијски програм Геодезија и геоинформатика

Наслов рада SPATIO-TEMPORAL INTERPOLATION OF CLIMATE ELEMENTS USING GEOSTATISTICS AND MACHINE LEARNING (PROSTORNO-VREMENSKA INTERPOLACIJA KLIMATSKIH ELEMENATA PRIMENOM GEOSTATISTIKE I MAŠINSKOG UČENJA)

Ментор В. проф. др Милан Килибарда, дипл. инж. геод.

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива 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 аутора

У Београду, _____ _____

Изјава о коришћењу

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

SPATIO-TEMPORAL INTERPOLATION OF CLIMATE ELEMENTS USING GEOSTATISTICS AND MACHINE LEARNING (PROSTORNO-VREMENSKA INTERPOLACIJA KLIMATSKIH ELEMENATA PRIMENOM GEOSTATISTIKE I MAŠINSKOG UČENJA)

која је моје ауторско дело.

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

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, _____ _____

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 прерада**. 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 прерада**. 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.