Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Capturing local-scale temperature patterns through machine-learning-based downscaling and bias-correction of global climate data

Joel-Philip Jäschke

Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Capturing local-scale temperature patterns through machine-learning-based downscaling and bias-correction of global climate data

Author:	Joel-Philip Jäschke
Examiner:	Prof. Dr. Felix Dietrich
Advisor:	Dr. Sebastian Rupprecht (Munich Re)
Submission Date:	May 28th, 2024

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

May 28th, 2024 Joel-Philip Jäschke

Acknowledgments

I would like to express my deepest appreciation to my supervisor Dr. Sebastian Rupprecht for his continuous guidance, all the weekly meetings and discussions we had—which I will really miss—and for giving me the chance to work on such an interesting topic and providing all the necessary resources. I had a great time at Munich Re and learned about so many new topics!

I also wish to thank Prof. Dr. Felix Dietrich for taking interest and acting as an examiner for this thesis, and for all the feedback he provided as well as the new topics he introduced me to.

Lastly, I would like to thank my girlfriend and my parents for their support and for always brightening my mood, even when something did not work out as planned. Thank you!

Abstract

To effectively manage impacts of rising global temperatures, high-resolution climate data is required. Reanalysis, which combines numerical weather predictions with observations, provides the most precise data on the global state of the atmosphere, the land surface and the oceans, but still exhibits biases and lacks fine-scale detail due to its coarse horizontal resolution. Current downscaling and bias correction methods mostly use regression models and interpolated observations to improve the reanalysis, but often fail to adequately represent urban temperature profiles.

This thesis addresses these limitations by testing the ability of various machine learning models to downscale ERA5 reanalysis daily maximum temperature fields from their native $25\,\mathrm{km}$ to an improved $5\,\mathrm{km}$ horizontal resolution, combined with a bias-correction against surface temperature observations, achieved through fusing reanalysis and remote sensing indicators. To reduce artificial bias during training, a lapse rate correction between observed and modelled quantities is employed. In addition, Urban Heat Island anomalies are derived for selected cities and the model's ability to reconstruct fine-scale urban temperature profiles is assessed against high-resolution numerical simulations.

Results of model application indicate significant improvements globally, with a complete removal of bias and a 10–30% reduction of mean absolute error, depending on the region, when assessed against an independent set of observations. Comparing results with high-resolution numerical simulations in urban areas, the model exhibits root mean square errors between 1.0–$1.5\,^{\circ}\mathrm{C}$, albeit with notable spatial bias.

In conclusion, the results demonstrate the effectiveness of the developed model in downscaling and bias-correcting reanalysis temperature, but show the need for further refinement at city level to address the spatial bias in urban areas. The results represent an advancement over existing reanalysis data, particularly in capturing sub-gridscale detail in urban areas where explicit city parametrizations are currently lacking or non-existent in reanalysis products.

Contents

Acknowledgements vii

Abstract ix

1 Introduction 1

2 Theoretical Background and Related Work 3
2.1 Numerical Weather Prediction and Observation . . . 3
2.1.1 Numerical Weather Models . . . 3
2.1.2 In-Situ Observations . . . 5
2.1.3 Remote-Sensing Observations . . . 6
2.1.4 Reanalysis Models . . . 7
2.2 Selected Machine Learning Models . . . 8
2.2.1 Decision Trees . . . 9
2.2.2 Random Forests . . . 10
2.2.3 Gradient Boosting . . . 11
2.2.4 Evaluation of Machine Learning Methods . . . 13
2.3 Urban Heat Islands . . . 14
2.4 Related Work . . . 15

3 Machine Learning for Temperature Downscaling 19
3.1 Methodology . . . 19
3.1.1 Preparation of Training Data . . . 19
3.1.2 Feature Selection . . . 23
3.1.3 Model Selection . . . 24
3.1.4 Performance Evaluation of Gridded Data . . . 26
3.1.5 Determination of Urban Heat Island Effect . . . 27
3.2 Results . . . 28
3.2.1 Results of Model Selection . . . 28
3.2.2 Influence of In-Situ Observation Density . . . 32
3.2.3 Performance Comparison of Raw and Improved Data . . . 34
3.2.4 Ability of Model to add Sub-gridscale Variability . . . 42
3.2.5 Assessment of Model Physicality . . . 47
3.2.6 Derivation of Canopy Urban Heat Island Anomaly . . . 48
3.2.7 Model Performance without Remote Sensing Observations . . . 49
3.3 Discussion . . . 50

4 Conclusion **57**

Bibliography **59**

Appendix **71**
A Environmental Lapse Rate Factors 71
B Hyperparameter Optimization 72
C Additional Results from Model Selection 74
D Additional Results from Temperature Field Evaluation 75
E Reanalysis-only Model 76

1 Introduction

Modern numerical weather prediction (NWP) systems are capable of simulating hundreds of variables related to the atmosphere, the land surface and the oceans on a global scale with a high degree of accuracy. However, their spatial resolution is often limited by computational resources, restricting even the most advanced operational global systems today to a $25\,\mathrm{km}$ horizontal grid spacing at best. But cities, where more than 4.4 billion people live today, require finer resolutions because many processes that affect the urban climate occur at scales well below $25\,\mathrm{km}$. In addition to the lack of resolution, NWP systems exhibit inherent systematic errors due to imperfect approximations of physical processes, which may cause their results to be biased and can skew analysis using these fields as ground-truth.

To overcome such issues, a post-processing step is often applied that interpolates NWP fields of a variable to a higher spatial resolution and uses auxiliary fields, interpolated from surface observations of the same quantity, to match the statistical distribution of the NWP-produced variable to that of the interpolated observations. Problematically, the interpolation is strongly contingent on the density and distribution of ground observations, and areas with sparse observation networks may have to rely on a single measurement of a physical quantity to represent areas of hundreds or thousands of square kilometers, which is often unrealistic. This leads to issues not only in regions with sparse observation networks, but also in areas where the variable of interest is subject to rapid change. An example of this is urban temperature, which is affected by anthropogenic heat emissions and complex aerodynamic flows due to the urban topography.

Therefore, this thesis aims to investigate whether machine learning models can be used to substitute this post-processing step by combining NWP data with satellite observations, geographical data and ground observations to bias-correct and downscale NWP inputs. The objectives of this thesis are to gather quantitative evidence about the ability of various machine learning models to remove the bias and improve the resolution of daily maximum temperature fields produced by NWP systems and to assess errors both at global and local scales. Furthermore, this thesis will answer the question of whether such models can add information to the coarse-scale inputs in such a way that they faithfully reproduce physics at smaller scales. To this end, outputs will be compared with high-resolution city-scale numerical simulations in order to quantify potential errors and biases in the model results. Based on these findings, a case study on the determination of Urban Heat Islands from gridded data will be conducted, which is one of the use cases that benefits from the higher resolution and better representation of urban temperature profiles in cities.

In Chapter 2, an overview of the data sources that are available for the fusion approach will be given and various machine learning methods that will be evaluated to achieve the goals mentioned above will be introduced. A brief introduction of the Urban Heat Island, its origins and implications, will be given, which will be followed by a short summary of the current state of the art. In Chapter 3, the implementation will be discussed, covering the determination of training features, preparation and pre-processing of the data and methods to determine an appropriate model for the problem, followed by the introduction of an improved method for the derivation of Urban Heat Islands from gridded data and ways to assess errors of geospatial data. Following that, the accuracy of the previously-produced daily maximum temperature maps will be assessed, both at a global scale with respect to reanalysis data and in-situ observations, but also at a local scale against high-resolution city-scale weather simulations. Finally, the findings will be summarized and a short outlook on future research topics that can be derived from the results of this work will be given in Chapter 4.

2 Theoretical Background and Related Work

This chapter aims to provide a short overview of current numerical methods for simulating weather, along with methods and observations used to approximate the true state of the atmosphere, consisting for example of surface-level weather stations and satellite observations. The reader will also be introduced to reanalysis, which is the process of combining numerical weather models and observations in one continuous process to represent the best-known state of the atmosphere. Furthermore, tree-based machine learning methods will be presented, which will serve as the basis for downscaling and bias correction in this thesis. Finally, the origin and implications of the Urban Heat Island effect will be detailed and overview of the current state of the art is given.

2.1 Numerical Weather Prediction and Observation

Knowing the current state of the atmosphere is crucial, with short-term applications including the tracking of extreme events, such as heat waves or tropical cyclones, while over longer time frames, the effect of a changing climate on the human population can be analyzed. To make good predictions, NWP systems rely on complex numerical models and an accurate initial value of the atmosphere. This task is an initial-value problem [1, 2], as at some point the current state has to be known in order to initialize the model. However, as every initial state will contain errors due to measurement uncertainty and the atmosphere is an inherently chaotic system, an NWP model left free-running would accumulate errors and eventually produce unusable results [3].

Furthermore, the imperfect nature of observations is amplified by issues such as spatial representativity of point observations [4, 5] or limited availability in the case of satellite data. Additionally, satellite data is often limited to a single overpass of a location per day and the interference of clouds in remotely-sensed quantities causes large amounts of observations to be discarded [6].

Therefore, it is easy to see why good models are insufficient on their own and why observations of the true state are vital to both continuously nudge simulations towards the physical truth [3] and to reduce initial uncertainties [7].

2.1.1 Numerical Weather Models

Such as most physical processes, Bjerknes [1] postulated in 1904 that weather can also be modeled by a set of physical laws, mathematically described through various equations.

These equations consist of the Navier-Stokes and mass continuity equations, the ideal gas law and the first fundamental law of thermodynamics. Since six out of seven equations in the resulting system are partial differential equations without a known analytic solution, a numerical solution procedure is necessary. This requires both a spatial and temporal discretization of equations at discrete points, turning the solution into a area-wise average [4], as models have to abide physical conservation laws for modeled quantities, requiring surface-integrated fluxes into and out of the discretization cells. While this makes computing a solution tractable, it also requires a trade-off where processes happening at scales below the resolution of the discretization require some form of closure term in order to be resolved [7]. Such unresolved processes, originating from friction, turbulence, condensation and radiative heating as well as cooling, are accounted for by source terms for mass, momentum and heat through so-called *parametrizations*, relating how unresolved scales interact with resolved scales. To make predictions about the atmosphere, the discretized equations can be combined with an initial value, representing the initial state of the atmosphere, and evolved forward in time to yield a likely future state of the atmosphere. A few notable mentions of such NWP models include the Weather Research and Forecasting (WRF) model [8], representing an open-source model that is commonly used for research purposes and mostly applied at regional scales, and the Europen Center for Medium-Range Weather Forecast's (ECMWF) Integrated Forecasting System (IFS), representing a model used operationally to make predictions for the entire globe.

Generally, there is a difference between global- and regional-scale NWP models. The latter refer to models with domains covering only a fraction of the globe, such as individual cities, provinces or countries. Compared to global models, regional models additionally require boundary conditions, as they need to know the evolution of the atmosphere outside of their domain. By only covering part of the globe, the smaller spatial domain allows regional models to run at higher spatial resolutions, given the same amount of compute as a global model. Ultimately though, both regional and global models are affected by a lack of computational resources, which limits the horizontal resolution they can achieve. More importantly, there is also a difference with respect to the required parametrizations of a model. While some processes, such as cloud micro-physics, are calculated similarly in both global and regional models, i.e. they are *grid-scale invariant*, different processes like deep convection, which often only occur in fractions of a grid cell at resolutions used for global models, require different parametrizations in global NWPs compared to regional models [7].

This effect can for example be observed in Wedi *et al.* [9] who ran the IFS at a horizontal grid spacing of 9 and 1.4 kilometers, respectively, and compared differences between the coarse-resolution model with the parametrization for deep convection enabled and the high-resolution model with the same parametrization disabled. They found great agreement between the two, even though one of the most important parametrizations was disabled in the high-resolution model, emphasizing both the importance of accurate subgrid parametrizations in a coarse-scale model while also showing the merit of running at higher resolutions.

2.1.2 In-Situ Observations

While NWP systems make it possible to simulate the evolution of weather, surface-level observations represent a point source of information about the instantaneous state of the atmosphere. Measurable quantities typically include temperature, liquid and solid precipitation, wind speed and wind direction as well as direct and indirect solar radiation, recorded through manual or automatic weather stations. These quantities are measured through different instruments, such as thermometers for temperature, anemometers for wind speed and pyranometers for incident shortwave radiation.

Most national meteorological services (NMS) operate their own networks of in-situ observations. They are typically located in networks of strategically placed stations in order to represent underlying phenomena fully while requiring the least amount of stations [10, 11, 12]. Some examples include MeteoFrance's RADOME network and MeteoSwiss's SwissMetNet, providing some or all of the above-mentioned physical quantities at varying time intervals. Compared to national networks, there also exist global data sets that aggregate network data from various NMS and alleviate some of the issues of working with national-scale data, such as searchability due to language barriers, varying data formats and access restrictions. Notable examples include the *Global Historical Climate Network daily* (GHCNd) [13], the *Global Hourly - Integrated Surface Database* (ISD) [14] and the *Global Surface Summary of Day* (GSOD) [15].

When working with station observations, it is important to consider inhomogeneity and representativity of contained time series. *Inhomogeneity* refers to non-climatic changes in station records that may stem from relocation or a change of instrument, causing jumps or artificial trends in the recorded data [16, 17, 18]. This is especially critical when examining long timeseries at single locations, as systematic changes may erroneously be considered as a climatic signal, whereas in reality, it may just represent expanding urbanization causing rising temperatures near a station location for example. While in theory, the use of a fully homogenized timeseries is the only true solution, this is hard to achieve in practice. Wijngaard *et al.* [19] assessed the homogeneity of a data set containing historical weather station records for Europe and found that when applying a set of four homogeneity tests for the period between 1901–1999, 94 % of temperature series were flagged as `doubtful` or `suspect`, indicating a high likelihood of some form of inhomogeneity being present in the data. One should note that using inhomogeneous series for interpolation does not necessarily lead to useless resulting fields, but rather that care needs to be taken when using such data for trend analysis [20, 21]. *Representativity* refers to the issue that when extrapolating measurements from the position of the observation to nearby locations, careful handling is required as quantities can vary greatly over a short range if environmental conditions are complex [22, 4]. Daly *et al.* [22] found that at small distances between 1 and 10 km from a station, elevation and terrain-induced effects can greatly influence temperature patterns and cause sharp gradients due to effects such as cold-pool formation in valleys.

Another important aspect is the amount of quality control (QC) applied to measurements. Weather stations can record wrong values for all sorts of reasons, such as electronic failures, accidental measurement interference or transmission issues. In order to catch such erroneous values, QC methods are mandatory. A large suite of such tests is for example documented in Durre *et al.* [23], and range from simple tests, such as checking that minimum temperatures are lower than maximum temperatures on a day at a station, to complex spatial consistency tests that compare surrounding values to find spatial outliers. However, even strict QC is not always a guarantee for correct data. While GHCNd applies very strict QC measures, as documented in Durre *et al.* [23], most of its temperature measurements in the continental United States come from the COOP network [24], which is run by volunteers. Davey *et al.* [25] found that many stations in this network do not meet quality standards required for correctly recording various quantities such as temperature. This means that while measurements may be available at a station and may pass all QC procedures, their amplitude could still not be representative of the underlying physical state. Furthermore, COOP does not enforce standardized reporting times, leading to a potential mismatch of values from adjacent days being reported as the value for the current day [26, 27]. After all, correcting for such effects is highly dependent on the available measurement metadata and hard to do consistently at a global scale.

2.1.3 Remote-Sensing Observations

While weather stations provide point information at the surface, Remote Sensing (RS) describes the process of measuring physical quantities from airborne and spaceborne platforms, such as planes and satellites. They produce spatial representations of some measured quantity at various spatial resolutions and various degrees of spatial coverage, reaching from very local representations recorded by planes to global coverage provided by satellites. Typically, satellites provide daily overpasses, but this can vary a lot with some satellites providing new data every few minutes for small regions, but some only passing over a location every week. In the realm of weather prediction, the Television Infrared Observation Satellite 1 (TIROS-1), launched in 1960, was the first weather satellite, equipped with optical cameras for taking pictures of cloud cover across the earth. Since then, a multitude of new systems have been launched, providing invaluable data for analysis of the earth's surface and assimilation into operational NWP systems alike [28]. For the most part, they are equipped with spectral imagers, although some systems, such as the Advanced Scatterometer (ASCAT) onboard the MetOp satellites, also provide microwave measurements of temperature profiles and wind vectors in the atmosphere [29] or the Soil Moisture Active Passive (SMAP) satellite [30], which provides soil moisture measurements.

An issue that affects many measurements taken by spaceborne instruments in particular is the effect of cloud cover. Satellite instruments that operate in the visible and infrared wavelength range cannot see through clouds, often rendering large parts of their data unusable [6]. While microwave-based sensor are not directly affected, as clouds are

mostly transparent for microwave radiation, they still require special consideration for cloud-affected pixels due to higher-order interactions between clouds and the measured quantity [31, 32].

One way of dealing with cloud cover in observations is gap-filling. Gap-filling describes the process of reconstructing cloud-covered pixels by using surrounding values as well as temporal behavior to derive the value at the surface [33, 34, 35]. This yields a spatially and temporally consistent field with the effect of clouds removed, but comes at the cost of increased levels of uncertainty.

In the case of land surface temperature, special care needs to be taken to avoid clear-sky bias in the resulting data [36, 37]. Clear-sky bias refers to the phenomenon where reconstructed fields may be biased high due to the fact that only values unaffected by clouds can be used for reconstruction, which tend to be slightly different than those covered by clouds. Although there exist ways around this issue, such as the approach by Jia *et al.* [38], who utilized a full surface energy balance to properly consider the effect of clouds, they require considerably more processing and auxiliary data.

2.1.4 Reanalysis Models

NWP models require an accurate guess for the initial state of the atmosphere in order to be able to make reasonable predictions about the future. One challenge that comes with this is the continuous integration of new measurements into the model. Trying to create a new guess of the current state of the atmosphere every couple of timesteps purely from observations is an inefficient use of information and often infeasible due to a lack of available data.

This is where the so-called process of reanalysis comes in. In reanalysis, a NWP system is initialized with an initial state and then continuously fed with new observations that are merged with the model state to nudge it towards the observed state. This is done through data assimilation, which was historically based on algorithms utilizing the principle of Optimal Interpolation [39, 40], but has since then evolved to use 3D- and 4D-variational methods, such as defined by Le Dimet and Talagrand [41]. Data assimilation methods work by considering both measurements from a large variety of sources along with predicted model state and combine the information from both sources, all while considering uncertainties and spatial representativity from observations and model fields alike [40]. Modern methods are very flexible and allow for a wide range of different measurement types to be assimilated, such as weather station measurements, radiosondes and airborne and spaceborne observations, either directly or indirectly.

The most notable reanalysis nowadays is ERA5 [42], which is widely regarded as the best available global representation of the historical state of the atmosphere, the land surface and the ocean. ERA5 is based on ECMWF's IFS and utilizes 4D-variational ensemble data assimilation for the atmosphere with a separate data assimilation system for land-based variables, relying on 2D Optimal Interpolation. It is run operationally at the ECMWF and produces data continuously, with a latency of a few weeks for current data to be available.

Especially the assimilation of satellite data in NWP systems has led to great improvements in forecast quality, making errors on the southern hemisphere comparable to those in the northern hemisphere [43]. The former was historically much worse, simply due to a smaller number of in-situ observations, making estimates of the true state of the atmosphere error-prone. While modern reanalysis systems, such as the one used to produce ERA5, are highly flexible, they still only assimilate a small fraction of data available [7]. As covered previously, this is due to effect of clouds in satellite observations and how they affect cloudy-sky pixels. In order to overcome these issues, there are ongoing efforts to also assimilate cloud-affected observations in an all-sky fashion [31]. The first operational assimilation of cloudy-sky remote sensing measurements happened in 2009 inside ECMWF's operational forecasting system where Bauer *et al.* [44] introduced a unified algorithm for the assimilation of clear-sky, cloudy-sky and precipitation-affected observations, but has now evolved to be one of the most important improvements to forecast skill [31].

Even though modern reanalysis systems have improved tremendously over the past decades, they are still not without their issues. For many applications, their coarse resolution of approx. $25\,\mathrm{km}$ in the case of ERA5 is simply insufficient, both due to errors introduced through imperfect parametrizations as well as the raw resolution just being insufficient to discern smaller-scale spatial features. Additionally, even ERA5 still contains some systematic errors, such as a residual bias in its temperature fields [45, 46, 47] that needs to be dealt with. The next section will introduce machine learning methods that are then going to be used in the development of a system to increase the resolution of ERA5 fields as well as help deal with the systematic errors.

2.2 Selected Machine Learning Models

To solve the issue of bias-correcting and downscaling variables from weather models, machine learning (ML) models are used in this thesis. ML models can learn hidden patterns from a set of training data, thereby serving as a black-box approximation of processes with potentially unknown structure. This has made them a valuable tool for researchers from all areas of science in the past decades.

An important concept in machine learning is the Bias-Variance tradeoff. This tradeoff refers to a balance between two sources of error in a model, the bias and variance. Bias stems from a model that is too simple to reproduce the full hidden complexity in the training set and gives too little attention to the training samples, often leading to a model that underfits the data and produces consistently high errors on both seen and unseen samples [48]. Variance refers to a very flexible model that can accurately fit the training set, but is very sensitive to small changes in the data. This may originate from a model that learns noise patterns in the data, allowing it to achieve very low errors on the training set, but leads to a poor ability to generalize to unseen data. The latter issue is commonly known as overfitting in applied machine learning [48]. An ideal model is exactly complex enough to capture the full hidden complexity in the data, therefor minimizing its bias, while also

maximizing its ability to generalize to unseen data, minimizing its variance. Getting to such a model is a process of trial and error where tools, such as Cross-Validation, can be used during parameter tuning to guide practitioners to a close-to-optimal model.

Achieving the goals of bias-correction and downscaling in this thesis involve the solution of a regression problem. Therefore, for the next sections, it is assumed that the mapping

$$f : X \to Y \tag{2.1}$$

can be used to model the problem, where $X = \{\boldsymbol{x}_i\}_{i=1}^N$ represents the set of N feature vectors in the training set, with each $\boldsymbol{x}_i \in X$ being a vector of p scalar features $\boldsymbol{x}_i = \{x_1^i, x_2^i, .., x_p^i\}$, while $Y = \{y_i\}_{i=1}^N$ represents the set of scalar target values. These are combined into a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. The goal is to find a function f that optimally fits the training set.

2.2.1 Decision Trees

Decision trees, as introduced by Breiman *et al.* [49], describe a supervised machine learning model that can be used for classification and regression. According to the problem statement introduced above, the focus will lie on regression trees, but most concepts transfer to classification problems. The subsequent algorithmic formulation is inspired by Hastie *et al.* [50]. The goal is to now find a tree $T(x; \Theta_T)$ that optimally solves $f_T : X \to Y$ which, once trained, can be used to predict y^* for an unknown input $\boldsymbol{x}^*$ and is defined by Θ_T, which refers to the structure of the tree, i.e. on which features and at which values to split the training data. In order to find an optimal tree, the minimization problem is posed as

$$\Theta_T = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, T(\boldsymbol{x}_i; \Theta)).$$

Commonly, the squared error loss $L(y, x) = (y - x)^2$ is used. Ideally, one would minimize the expression

$$\underset{R_j}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{y_i \in R_j} (y_i - \widehat{y}_{R_j})^2 \quad \text{with} \quad \dot{\bigcup}_{j \in I} R_j = \mathcal{D},$$

where R_j represent p-dimensional boxes that partition the training set $\mathcal{D}$,

$$\widehat{y}_{R_j} = \frac{1}{|R_j|} \sum_{y_i \in R_j} y_i$$

is the average prediction for all training samples falling inside that box and $|R_j|$ being the number of points falling into that box. Finding an optimal partition yields the lowest global squared error.

However, this computation is impractical, as the boxes R_j can have arbitrary shapes and testing all combinations is computationally intractable. To overcome this issue, regression

trees use a binary recursive partitioning scheme to split the input data, producing a decision node with two children, each again being a decision node up to the last layer which consists of leaf nodes, containing regression predictions. For their construction, a greedy approach is chosen, which produces locally-optimal maximum likelihood estimations. At each node, the training set is split into the following two subsets

$$R_l = \{\mathcal{D}|X_j \leq s\} \quad \text{and} \quad R_r = \{\mathcal{D}|X_j > s\}$$

for each feature j. Then, the following expression is minimized at each node:

$$\underset{j,s}{\operatorname{argmin}} \left(\sum_{y_i \in R_l} (y_i - \widehat{y}_{R_l})^2 + \sum_{y_i \in R_r} (y_i - \widehat{y}_{R_r})^2 \right).$$

Once the ideal feature and split have been found, the set is split into two disjoint subsets and the process is repeated for each subset. This is done until some stopping criterion is met, such as reaching a maximum predefined depth, a minimum number of samples per node or the gained cost reduction through another split being below some threshold.

Advantages of Decision Trees are their easy interpretability, fast training, no assumptions about the distribution of inputs and the ability to quantify feature importance. Disadvantages include high variance, where a small change in the input data can result in an entirely different tree, a lack of smoothness of the output and issues capturing additive structures in the input.

2.2.2 Random Forests

Random Forests, also known as Decision Forests, were first introduced by Breiman [51] and improve upon some of the drawbacks experienced with Decision Trees by combining bagging and a random tree creation method. Bagging, short for bootstrap aggregating, refers to the creation of B new, equally-sized, training sets $\{\mathcal{D}_b\}_{b=1}^{B}$ by sampling the original set $\mathcal{D}$ with replacement. Using these bootstrap samples, B estimators are fit on the separate bootstrap sets $\mathcal{D}_b$, producing a set of predictors $\{f_b(x)\}_{b=1}^{B}$. To make predictions with this model, one can simply average predictions from all estimators in this set f_B to get an ensemble prediction

$$f_B(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x).$$

A big advantage of bagging is the reduction of variance, which especially benefits estimators, such as Decision Trees, that naturally have a high variance.

The second idea is to only use a random subset of $m \leq p$ predictors at each node for deciding on a split point during tree creation. Common values for m are $m = \sqrt{p}$ or $m = \log_2 p + 1$, but can go as low as $m = 1$. Therefore, all trees $\{T(x; \Theta_b)\}_{b=1}^{B}$ in a Random

Forest are built on a training subset $\mathcal{D}_b$ and each node of a tree is split on a random subset of features. This yields a predictor of the form

$$f_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b),$$

similar to how predictions were made using bagging.

One of the major features of Random Forests is their ability to measure the out-of-bag (OOB) error. The OOB error refers to the idea of letting a Random Forest make predictions using its training set, but to only consider predictions from each tree $T(\boldsymbol{x};\Theta_b)$, for which $(\boldsymbol{x}, y) \notin \mathcal{D}_b$, to measure the error committed by trees on left-out training samples. This allows the user to diagnose how well the ensemble of individual trees performs on unseen data without having to create a separate testing set.

Random Forests keep many of the advantages of Decision Trees, such as fast training and a built-in measure of feature importance while also being hard to overfit. Additionally, they give the user a tool to assess their error without the need for cross validation through the OOB error. However, disadvantages include their memory usage, which can get quite large when using large training sets and large numbers of trees, and harder interpretability compared to plain Decision Trees.

2.2.3 Gradient Boosting

Boosting methods approach the solution of a problem of the form of Equation (2.1) in a different manner, by combining a set of weak learners to create a single strong learner. Gradient Boosting methods, as first introduced by Friedman [52], are part of the family of Boosting methods. The approach taken when fitting a Gradient Boosting model, again inspired by Hastie *et al.* [50], is that base models are successively fitted in a stage-wise process. While the focus in this thesis lies on using Decision Trees $T(x;\Theta)$ as a base regression model, the general approach is easily transferable to other base learners. The problem to be solved is the one defined in Equation (2.1) and the following algorithm is based on the work of Friedman [52]. To measure the goodness of fit of the approximation, a loss function is required, yielding the problem

$$\underset{f}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i)) \tag{2.2}$$

to be solved. For the sake of simplicity, the loss function is assumed to be the squared error loss

$$L(y, x) = \frac{1}{2}(y - x)^2.$$

Using the fact that each prediction $f(\boldsymbol{x}_i)$ is just a scalar, Equation (2.2) can be treated like a numerical minimization problem with an initial guess

$$f_0 = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

and the update rule given by

$$f_m(x) = f_{m-1}(x) + h_m(x),\ m \geq 1.$$

Choosing h_m comes down to the minimization method used, but in the case of Gradient Descent, as commonly used in many methods, h_m is given by

$$h_m(x) = -\nabla_m \cdot \gamma_m = -\left.\frac{\partial L(y, f(x))}{\partial f(x)}\right|_{f=f_{m-1}} \cdot \gamma_m = (y - f_{m-1}(x))\gamma_m.$$

As can be seen, when using the squared error loss, the update rule is based on the approximation residual of the product of the previous step and some step length. Under the assumption of a discrete set of inputs, one would get

$$\boldsymbol{h_m} = \{\gamma_m(y_i - f_{m-1}(\boldsymbol{x}_i))\}_{i=1}^{N}$$

as a set of local directions of steepest descent of the training data. Using this formulation, however, the model would not be able to generalize to unseen data, as its gradients depend on training targets y. To remedy this issue, the gradient term $-\nabla_m$ is approximated through a base regressor that is trained to predict approximation residuals y given inputs x, decoupling it from the training target values. The trained regressor $\widehat{t}_m = T(\boldsymbol{x}_i; \Theta_m)$ with

$$\Theta_m = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N} (-\nabla_{m,i} - T(\boldsymbol{x}_i; \Theta))^2$$

consequently learns to approximate the local direction of steepest descent from the data. Once fitted, the step length γ_m can be determined as

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \widehat{f}_{m-1}(\boldsymbol{x}_i) + \gamma \widehat{t}_m(\boldsymbol{x}_i))$$

through simple line search. Therefore, the final update rule for discrete inputs becomes

$$\widehat{f}_m(x) = \widehat{f}_{m-1}(x) + \gamma_m \widehat{t}_m(x),$$

which is repeated in a stage-wise process for $m = 1, ..., M$. In general, the more learners are fitted, the better the prediction becomes.

2.2.4 Evaluation of Machine Learning Methods

For the introduced models, it is important to find a set of parameters such that each model performs close to the minimum of the Bias-Variance curve. In order to find such a model, one needs tools to assess both how well it generalizes, but also how accurate it can make predictions. A useful tool for the first issue is Cross Validation. Cross Validation, also referred to as n-fold Cross Validation [53], with $2 \leq n \leq N$, refers to the process of partitioning the data set $\mathcal{D}$ into n subsets and then successively training a model on $n-1$ subsets, while validating its performance on the remaining subset. By comparing errors on the training and validation subsets, it is easy to see whether errors increase noticeably between training and validation sets, indicating overfitting of the model.

As this thesis is concerned with geospatial problems, improvements upon just splitting the training set $\mathcal{D}$ randomly can be made. If there exists structure in the training data, such as data points originating from a small number of spatially distributed sources, such as weather stations, it may be beneficial to use the source as an indicator for partitioning the training and validation sets. By having partitions that contain mutually exclusive weather stations, the model's ability to generalize to unseen locations not contained in the training data can easily be tested.

While Cross Validation provides insights into the model's ability to generalize from the training set, tools to assess the accuracy are still necessary. In this thesis, root mean squared error (RMSE), mean absolute error (MAE), mean error (ME) and the coefficient of determination R^2 are used to quantify the mismatch between predicted and real temperatures at weather stations.

The MAE is defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(\boldsymbol{x}_i)|$$

and measures the absolute deviation of predictions from the true value.

The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2}$$

and is more sensitive to outliers compared to the MAE due to the square. Both the MAE and RMSE do not make any differentiation between over- and under-prediction of quantities.

The ME refers to a measure of bias of model predictions with respect to target values. It is defined as

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N} y_i - f(\boldsymbol{x}_i)$$

and does not calculate the absolute difference but instead calculates signed differences compared to the MAE. Values close to zero mean that the model makes unbiased predic-

tions, while large positive or negative values indicate a model that commits a systematic under- or overestimation, respectively. In light of applications discussed in later chapters, ME is especially important as one of the goals is to apply a bias correction to the model data, i.e. to remove bias present in the inputs from the final results. Ideally, the model is also able to reduce some systematic errors from the measures, therefor reducing MAE and RMSE metrics on top of removing the bias.

R^2 refers to a measure of the amount of variance in the target values that can be explained through the predictive features X used during training. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

where

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

An R^2 value of zero indicates a model that always predicts $\bar{y}$ while a value of one indicates a perfect model. Values can be smaller than zero, indicating a model that makes predictions worse than the least-squares fit of the data.

2.3 Urban Heat Islands

An application that greatly benefits from an improved horizontal resolution of gridded temperature fields is the determination of the Urban Heat Island (UHI) effect in cities. The UHI refers to the effect of cities warming up more than their rural surroundings [54]. Generally, one has to differentiate between surface UHI (SUHI) and canopy UHI (CUHI), where the former refers to a land surface temperature gradient between city and rural surroundings while the latter refers to a surface air temperature gradient between city and rural surroundings. This distinction is important, as SUHI and CUHI are driven by markedly different processes and exhibit different behavior with respect to their temporal occurrence and magnitude [55].

The process of UHI formation is quite complex and influenced by many factors, such as land cover, vegetation density and aerodynamic surface roughness due to buildings [54]. Excess heating of surface air is mainly connected to two processes, absorption of shortwave radiation by the urban surface and direct emission of longwave radiation through anthropogenic processes. Shortwave radiation emitted by the sun is absorbed by urban structures, amplified by the fact that many materials present in cities have a low albedo, which refers to the amount of incoming radiation reflected by a material, causing a large amount of incident shortwave radiation to be absorbed. Energy accumulated by the environment is then re-emitted into the atmosphere in the form of long-wave radiation, causing heating of the surface air layer. The second contributor is longwave radiation directly emitted into the environment by anthropogenic processes, such as heating and cooling,

transportation, machinery and power generation. These emissions stem from inefficiencies and conversion losses in thermodynamic processes, i.e. no electrical system being 100 % efficient, therefore emitting longwave radiation generated through resistive heating of electrical circuits for example.

The Urban Heat Island effect has both economical and health-related implications. As found in Heaviside *et al.* [56], the UHI effect amplifies summertime temperature extremes further in cities, which globally represent the locations with the highest population density, causing even higher excess mortality increases due to heat-related stress. Similar findings were also made by Goggins *et al.* [57] and Smargiassi *et al.* [58], who found a statistically significant link between increased inner-city surface temperatures and heat-related mortality. Furthermore, an increasing UHI effect causes higher utilization of air conditioning, which at the same time amplifies UHI intensity even more due to electrical losses and increased energy demands. Due to continued urbanisation, a population that is becoming older and a changing climate, future health-related risks are projected to become even worse [59].

To lower the impact of UHI on the population, various mitigation measures exist. These can be broadly divided into reducing the amount of shortwave radiation absorbed by cities and increasing the evaporative cooling potential through increased vegetation cover and water availability. The former method works by using highly-reflective building materials, such as white roofs with an albedo of greater than 70 %, as found in Li *et al.* [60], causing a noticeable SUHI and CUHI decrease. The latter works by increasing the amount of vegetation present in cities, commonly done through using so-called green roofs [60, 61]. Green roofs refer to rooftops covered with vegetation, providing both increased thermal resistance and latent heat capacity [62] and serve the same purpose of lowering both SUHI and CUHI in cities. As a nice side effect, they also increase biodiversity [61].

2.4 Related Work

A short overview of the existing literature is given now which introduces works with a similar goal of improving the resolution and bias-correcting, among other things, daily maximum temperature fields produced by ERA5. Downscaling and bias-correction of reanalysis data has been and is continuing to be done quite regularly, with some notable data sets being *CHELSA-W5E5* [63], *MSWX* [64] and *EM-Earth* [65]. While all these produce a downscaled and bias-corrected version of the ERA5 reanalysis, in the case of air temperature, they do so largely through regressing against elevation and using high-resolution climatology to correct statistical distributions. The idea of using the relationship of temperature and elevation to produce downscaled temperature maps dates back to 1997 where it was introduced by Dodson and Marks [66]. Both CHELSA-W5E5 and MSWX rely on environmental lapse rates (ELR) and high-resolution elevation to downscale their temperature, where ELRs describe a trend of decreasing temperature with increasing elevation [67]. Once temperature is brought to a higher resolution using coarse and fine eleva-

tion data, they then use additional data sets, such as CRU TS 4.03 climatology [68] for CHELSA-W5E5 and CHELSA climatology [69] for MSWX, to bias-correct daily temperatures by matching their monthly distribution to the ones provided by climatology data. While this approach serves the goal of achieving a high-resolution, bias-corrected data set, it falls short with respect to representing urban temperature fields.

Gridded data sets, such as CRU TS and CHELSA, are based on an interpolation of weather station observations. One issue with this approach is that if no observations are located inside cities, the climatology may underestimate inner-city temperature due to just interpolating rural values across cities extents. Combined with the fact that ERA5 as of now does not have an explicit city representation in its land surface scheme [27], effects of UHI may go unnoticed, even after applying a bias correction and downscaling due to systematic under-representation of UHI effects in the first place.

There also exist data sets that do not make use of reanalysis data directly, such as the work by Kilibarda *et al.* [70], who produced a global, $1\,\mathrm{km}$ resolution map of daily minimum and maximum air temperature. They use Kriging for interpolation with 8-day composite MODIS land surface temperature images, various environmental covariates, such as elevation, and a geometric temperature trend for the interpolation. Zhang *et al.* [71] produced a global map of maximum and minimum temperatures using a model named Spatially Varying Coefficient Model with Sign Preservation [72], which uses elevation and a gap-filled land surface temperature as covariates. While both methods improve upon previous data by utilizing satellite data as a source of sub-gridscale variability, the method by Kilibarda *et al.* only uses 8-day composite land surface temperature, which they had to spatially interpolate to fill gaps due to cloud cover. Additionally, these 8-day composites may not provide the same value for interpolation as a gap-filled daily LST product would do, simply due to not providing daily-varying surface temperature. The work by Zhang *et al.* [71] used such a gap-filled product, namely data based on the works by Zhang *et al.* [33], but overall only considers two indicators, which is likely to be insufficient in many areas of the world, as the link between land surface temperature and air temperature has been shown to be highly non-linear, depending on atmospheric conditions, surface configuration, terrain complexity and many more [73, 74].

A recent publication by Zheng *et al.* [75] introduces yet another methodology, applying a method that is quite similar to what is outlined in this thesis. They test various machine learning methods to derive high-resolution temperature maps from a combination of reanalysis and satellite data, such as reanalysis surface air temperature and gap-filled satellite land surface temperature. However, they only produce data for Eurasia, which does not meet the criterion of a global data set, as targeted by this thesis. This work will not only differ from Zheng *et al.* by producing a global data set, but also use a different set of input features, not rely on spatial clustering of the domain with different models per region and finally a different pre-processing procedure is applied to remove artificial bias.

Similarly, Bouallègue *et al.* [76] tested the ability of Linear Regression, Random Forests and Neural Networks to correct forecast errors in the operational IFS model. While this is different from the hindcast correction outlined in this work, the approach is similar in that

it uses fields produced by the NWP system to make predictions about the error.

Note that the examples above by no means represent an exhaustive survey of available data. Generally though, many data sets with the goal of providing high-resolution temperature maps can be divided into those directly based on reanalysis data that are bias-corrected using long-term climatology and those generated by using a regression model and different auxiliary variables, like elevation, land surface temperature and land use.

Finally, with respect to representing UHI, there exist global data sets that quantify the SUHI effect, derived from MODIS land surface temperature and various algorithms, such as NASA's Socioeconomic Data and Applications Center (SEDAC) Global Urban Heat Island map [77] and the Yale Center for Earth Observation (YCEO) Surface Urban Heat Island data [78, 79]. However, at this point in time, no global (gridded) data set providing CUHI estimates exists and most analyses appear to rely on local pairs of stations with one being located inside of a city and the other one located in its rural surroundings.

3 Machine Learning for Temperature Downscaling

With the necessary theory covered, this chapter now details the approach used for producing downscaled and bias-corrected daily maximum air temperature maps. The next sections cover model and feature selection and detail how to correctly assess the error of spatial data, followed by the introduction of a methodology for deriving Urban Heat Islands from gridded temperature fields. Afterwards, results of the study are presented and further details on the model performance for bias correction and downscaling of daily maximum temperature are provided. Finally, results are discussed and limitations of the model are detailed.

3.1 Methodology

Before applying any data-driven modelling approach, it is necessary to prepare and preprocess the data that is used later. This involves quality control, feature selection and proper consideration of physical and spatial relations between variables. In a subsequent step, ways to assess which model performs best are needed, given the large variety of algorithms available today. The nature of geospatial data simultaneously requires more elaborate ways of error assessment, both during model selection and later evaluation of the results, entailing tasks like the correct handling of data with varying spatial density. Such special treatment is required along all stages of the modelling process and also finds its use during the derivation of Urban Heat Islands. A new algorithm is detailed, based on previous work by Valmassoi and Keller [80], that can be used derive UHI intensity from coarse gridded maps while also accounting for various terrain effects.

3.1.1 Preparation of Training Data

In order to train a machine learning model, it is necessary to generate a set of training data. To achieve this, a number of input data sets are pre-processed to account for various deficiencies and data format differences. One of the most important stages in this process consists of the preparation of the training target variable, which acts as a source of truth for the quantity in question during model training. As described in Section 2.1.2, weather stations are a perfect fit for this application, as they measure various atmospheric variables with very high accuracy. To get such observations at a global scale, daily maximum air temperature data from both GHCNd and GSOD are downloaded from their respective

sources. Problematically, GSOD contains a higher number of outliers and erroneous values when compared to GHCNd, likely stemming from the less restrictive quality control process. As the model's predictive quality greatly depends on the quality of its input data, a subset of quality control algorithms from [23] is therefore implemented, which are also used to quality-control GHCNd data. The subset of implemented algorithms from [23] is listed in Table 3.1, where a focus is put on those that flagged the highest number of measurements in the publication, namely the internal and temporal consistency checks. For a more thorough explanation, the reader is deferred to [23].

Table 3.1: Subset of quality control checks implemented for GSOD, as documented in [23].

Group	**Name of check**
Basic integrity checks	Naught check
	Repeated month in year
	Repeated month across years
	min = max in month
	World record exceedance
	Identical-value streak check
Internal & temporal consistency	Iterative consistency check
	Spike-Dip check
	Lagged temperature range check

After application of these algorithms to the GSOD data, the improved set of observations better matches the error rate of GHCNd data.

As both GHCNd and GSOD are merged into one large training set, it is further necessary to filter out stations contained in both, as there is no guarantee that stations are mutually exclusive between the two. Due to its stricter quality standards, GHCNd is deemed to be the preferred source of information, even after correction of the GSOD stations. A spatial buffer zone of $5000\,\mathrm{m}$ is created around each station location in GHCNd and all GSOD stations located inside of the buffer are discarded. The value of $5000\,\mathrm{m}$ is chosen by visual inspection of duplicate station locations, but different values for this threshold may cause less false-positive duplicate stations to be dropped, thus artificially reducing the data.

After dropping all duplicate stations, a final check from [23] is implemented. The spatial corroboration test checks whether temperature anomalies with respect to their long-term station mean are similar between neighboring stations in a certain search radius and flags those anomalies that deviate from their spatial neighbors. The motivation to repeat this test is that both GHCNd and GSOD may contain stations where a station from one data set individually may not have any neighbors inside the above-mentioned search radius. By merging both GSOD and GHCNd, such isolated stations may now have neighbors inside the radius that originate from the other data set. Subsequently, the corroboration check may find new outlier values because it now has access to more spatial neighbors, consequently improving the classification of errors and spatial consistency of the merged data.

With the target variable out of the way, the next step is to prepare the spatial and temporal predictor fields used during model training. This process mainly involves downloading source data, changing the data format, such that all values are available as netCDF files, and aggregating all fields to a common spatial and temporal resolution. In the case of reanalysis data from ERA5, which is only available as hourly data, this requires taking the daily maximum of hourly temperatures, calculating the average instantaneous power for fluxes and deriving the daily means for all remaining variables. Wind speed w is then derived from the zonal wind component u and the meridional wind component v as

$$w = \sqrt{u^2 + v^2}.$$

All satellite-derived values, such as land surface temperature from [33] or the MODIS leaf area index are also transformed to a common latitude-longitude grid, as some of them are only available in the sinusoidal equal-area projection. For variables only available every couple of days, their values are interpolated bilinearly between available time steps. The gap-filled land surface temperature from [33] further requires insertion of leap days, as the data set only provides 365 steps per year, irrespective of whether the year is a leap year. This is achieved by linearly interpolating values between the 365^{th} day of the leap year and the first day of the following year. A full list of all processed variables can be found in Table 3.2.

Table 3.2: A list of all input features considered for model training with spatial and temporal resolution ($\boldsymbol{\Delta x}$ and $\boldsymbol{\Delta t}$, respectively). Auxiliary features may not have a spatial or temporal resolution, or they may be constant in time, in which case no resolution is specified. A check mark in the **Used** field indicates that the feature is kept after feature selection.

Category	**Variable**	**Description**	$\boldsymbol{\Delta x}$	$\boldsymbol{\Delta t}$	**Used**
Reanalysis	tasmax	Maximum air temperature	0.25°	1 day	✓
	ws	Mean wind speed	0.25°	1 day	✓
	sshf	Surface sensible heat flux	0.25°	1 day	✓
	oro	Model orography	0.25°	-	✓
Satellite	lst_day	Gap-filled MODIS daily max LST	1 km	1 day	✓
	lst_night	Gap-filled MODIS daily min LST	1 km	1 day	✓
	cfc	CLARA cloud fraction	0.25°	1 day	✓
	lai	MODIS Leaf Area Index	1 km	8 day	
Geographical	lat	Latitude	-	-	✓
	lon	Longitude	-	-	✓
Temporal	cos_doy	Cosine of day of year	-	1 day	✓
	yr	Year	-	1 year	
Auxilliary	popd	Population density	1 km	-	
	urbf	Urban Fraction	1 km	-	
	lulc	Land Use Land Cover	1 km	-	

Once all the preparatory work is finished, a set of all potential training features can be created by sampling their values across the entire time span at observation locations. At this point, a decision on the target resolution has to be made in order to sample values in a physically-consistent way. For this thesis, a target spatial resolution of $5\,\mathrm{km}$ and temporal resolution of 1 day is chosen, however, other values are possible. This yields a good compromise between the required storage of the resulting maps and the ability to make out fine-grained spatial features in the maps. Variables with a finer resolution than the target are averaged to the target resolution while those with a coarser resolution are not treated further. After matching the resolution, all spatial input features are then bilinearly interpolated to the location of the weather stations in the merged and filtered set of GHCNd and GSOD stations with the exception of daily reanalysis maximum temperature and orography.

As mentioned in Section 2.1.2, weather stations provide a very local measurement of a variable while NWP fields provide grid-cell averages. Therefore, the assumption that a temperature measurement made by a weather station located somewhere inside a grid cell of the reanalysis is comparable to the reanalysis prediction may introduce an artificial bias [4] that makes it hard for the model to learn the true systematic error structure, but instead forces it to focus on the error introduced by failure to consider how representative the quantity measured by a weather station is for the entire reanalysis grid cell it is contained in. It was previously established in Section 2.4 that temperature has a strong relationship with elevation. This relation is used to overcome the issue of station representativity by applying a simple elevation matching strategy. A lapse rate adjustment, as introduced in [66], is applied to temperature observations in order to simulate their measurements being taken at the same height as the grid-cell average height of the reanalysis instead of their native elevation. When applying this adjustment, one gets a corrected temperature $\mathrm{T}_{\mathrm{cor}}$ defined as

$$\mathrm{T}_{\mathrm{cor}} = \mathrm{T}_{\mathrm{obs}} - \Gamma \Delta h, \tag{3.1}$$

where Γ refers to the lapse rate and $\Delta h = h_{\mathrm{station}} - h_{\mathrm{reanalysis}}$ to the elevation difference between the station and the reanalysis elevation. In general, a default lapse rate Γ of $-0.0065\,\mathrm{K\,m^{-1}}$ is often assumed in literature [66, 81], however, different values may offer improved performance in certain regions [67, 82]. Although Dutra *et al.* [82] recommend to use dynamically-derived lapse rates from ERA5 temperature fields at various altitudes for the sake of consistency, they found no benefit in performance using them compared to the static $-0.0065\,\mathrm{K\,m^{-1}}$ rate, which is also found in independent tests repeated in this thesis for which the results can be found in Table A.1. For all subsequent steps, $\Gamma = -0.0065\,\mathrm{K\,m^{-1}}$ is therefore assumed. A similar approach of matching station and background elevations is often utilized in literature to both directly produce improved data sets as well as in operational data assimilation systems for the production of reanalysis data. Lussana *et al.* [83] used it to fuse observations with NWP background fields, using 2D Optimal Interpolation in Norway to directly produce improved fields for temperature, while the ECMWF's Land-DA [84] operational data assimilation system, based

on the work of Brasnett [85], uses it to integrate in-situ observations in their operational models to improve their forecasts.

Finally, using the lapse rate adjustment methodology, reanalysis maximum temperature and orography, a high-resolution elevation and the station maximum temperature and elevation, the set of training features can be assembled. This was done by adding one feature corresponding to a corrected reanalysis temperature through using Equation (3.1) and considering the reanalysis orography and the high resolution elevation, as well as a corrected in-situ temperature measurement of the weather station by again applying Equation (3.1), but this time using the station elevation and high-resolution elevation, respectively. The latter feature will then be the training target for the final model. While transforming both temperatures using an intermediate elevation may seem redundant for training, it bears advantages during inference. Here, it can be seen as a physically-informed interpolation of temperature, as instead of just bilinearly interpolating ERA5 temperatures to station locations, the elevation difference between the ERA5 orography and high-resolution elevation slightly alter the reanalysis temperatures, thus adding sub-gridscale features from a digital elevation model to raw reanalysis fields. A similar step is also used in [63] to downscale temperatures. Without this step, the only other source of such variability would be the MODIS land surface temperature, but especially in mountainous and topographically complex terrain, this may suffer from miss-representation due to other aspects affecting the land surface temperature, such as vegetation, land use and snow cover. To conclude the pre-processing and generation of training data, stations are randomly assigned a label, splitting them into training, validation and test stations. From all available stations, $70\,\%$ are training stations, $10\,\%$ are validation stations and the remaining $20\,\%$ are test stations. This is done in order to have independent stations available for hyper-parameter optimization and validation that the model never saw during training.

3.1.2 Feature Selection

With all input data prepared, it is necessary to come up with objective measures to derive a subset of training features that are truly relevant to the model performance. Finding a suitable subset of candidate training features is done by a combination of screening publications with similar goals of downscaling daily maximum temperature as well as physical process knowledge. Equations used in NWP models to derive air temperature can be used to give a first indication for which variables affect the underlying process, which can then be used to guide the selection of training features for the model. Here, the formulation used by WRF to derive surface air temperature T_{2m}, taken from Grossman *et al.* [86], is used

$$T_{2m} = T_s - \frac{H}{\rho c_p C_H},$$

where T_s refers to the land surface temperature, H to the sensible heat flux, ρ being the density and c_p the heat capacity of air, $\kappa = 0.4$ von Kármán's constant and

$$C_H = \frac{\kappa \times u_{*2m}}{ln\left(\frac{2m}{z_{0,T}}\right) - \Psi_H},$$

with u_{*2m} being the friction velocity, $z_{0,T}$ the roughness length for heat and Ψ_H the integrated universal function for heat at the surface. While it is outside the scope of this thesis to go into detail on the formulation, it is a good indicator for understanding the processes affecting air temperature near the surface.

All features in Table 3.2 are determined to be correlated with the analytical formulation from the WRF model. Terms, such as the land surface temperature, wind speed and the sensible heat flux, are best directly acquired from reanalysis, whereas u_{*2m}, $z_{0,T}$ and Ψ_h indicate parametric formulations that relate wind, or more generally advection of air parcels, and heat fluxes with terrain and landuse. Features like the MODIS leaf area index, land use and the auxiliary parameters, such as population density and urban fraction, are supposed to represent sub-gridscale variability of the terrain. When combined with wind speed, they are expected to be proxies for parameters such as u_{*2m}, $z_{0,T}$ and Ψ_h. Finally, cloud coverage affects the net radiation budget at the earth's surface and is therefore incorporated into the training.

It is important to note that more features do not guarantee better performance of the model and can even deteriorate its performance when redundant features are present that just add noise and may cause the model to overfit [87]. Consequently, it is necessary to filter out features that truly help a model to make good predictions. This task can be achieved in many ways, such as using Recursive Feature Elimination or Sequential Feature Selection (SFS). In this thesis, a bottom-up approach using Sequential Feature Selection is used. During SFS, the algorithm first trains the model on a single feature, trying all available features sequentially, and selects the one that produces the model with the lowest errors as measured by some metric, such as the RMSE, on the validation set. Once the best individual feature is found, this step is then repeated with the remaining features, but now training the model on the combination of previously selected and one previously unselected feature. This is repeated until some stopping condition is reached, such as a maximum number of features or new features not improving the validation score by some threshold. In this thesis, a threshold-based stopping criterion is chosen, using the RMSE and a minimum improvement of 0.01.

3.1.3 Model Selection

Using the set of features determined in the previous step, an objective selection procedure for determining an optimal model is now necessary. All further tasks are done using the `Python` programming language. For this thesis, four tree-based methods, namely Random Forests (RF), Extra Trees (ET), Gradient Boosting (GB) and Histogram Gradient Boosting (HGB), are tested, along with a Multilayer Perceptron (MLP) as a simple

neural network. The tree-based models of the `scikit-learn` package [88] are used, whereas the MLP is based on a custom implementation using the `swimnetworks` package [89]. In order to evaluate the skill as well as the practicality of each model, hyper-parameter optimization is conducted for all of them. This is achieved through applying the `GridSearchCV` algorithm from `scikit-learn`, a grid search procedure with cross validation, using stations located inside of a test region. The domain comprises Germany and parts of neighboring countries and is temporally subset to contain data from the years 2010–2014, although any other region and time frame could be used as well. As grid search tests all combinations of parameters, consequently incurring a lot of training cycles for each model, and considering the large amount of training data from all global stations across a time span of 18 years, it is necessary to limit the region and time frame to keep the total runtime to a reasonable level. The subsetted data is then split based on the label assigned to each station during preparation. All stations in the validation set are used to assess model skill during cross validation, whereas those in the training set are used for model training. As mentioned previously, using data from stations that models have never seen before during training to assess their accuracy is a good measure of true predictive skill, as it shows how well they are able to extrapolate spatially. During grid search, each model is scored based on its achieved RMSE, MAE and R^2 values on the validation set.

For tree-based models, tunable parameters affecting the model performance include the maximum depth of trees, minimum samples per split or the number of features to use for splitting at each decision tree node. Simple dense MLPs offer less hyper-parameters, with the most important being the number of layers, the width of each layer and the activation function. The full list of tested parameters for every model can be found in Appendix B. Before the models can be trained on the data, it is necessary to consider the numerical properties of each feature. While variables from Table 3.2 are predominantly numerical, land cover represents categorical classes and is therefore a categorical feature. There exist various ways to work with categorical values, such as One-Hot encoding, where each category is assigned its own feature, with a one indicating that a sample is of that class while a zero indicates the opposite. Alternatively, some models allow direct specification of categorical columns during training, such as the `HistGradientBoostingRegressor` model from `scikit-learn`, and handle them internally. One-Hot encoding was applied for all models without native support for categorical features. Furthermore, while this is irrelevant for tree-based methods, normalizing features for MLPs beforehand can greatly improve the training procedure. Especially when training features vary drastically in amplitude, normalization helps during gradient descent by prohibiting features with large amplitude to overpower those with small amplitude, but of potentially higher importance [50]. An interesting fact about the `swimnetworks` package is that it removes the need for data normalization when using MLPs. For a thorough explanation, the reader is deferred to [89]. From a usability perspective, this allows the user to train the model as he or she would any other tree-based model, which are known to be insensitive to scaling of the input features.

Besides performance on the problem, it is also important to quantify practicality of the

resulting machine learning models. Practicality in this sense refers to aspects such as time to train, size of the trained model stored on disk or memory required to load the model from disk. Sometimes, a user may wish to train the model and only later run inference with it or keep the model around for regulatory reasons. When the numerically-optimal model with the best performance during hyper-parameter optimization is prohibitively expensive to use in practice, this needs to be considered during selection of the final model just as much.

In prior work, multiple regional models are created, with interpolation being done between them to infer model outputs. Hashimoto *et al.* [26] use a grid of Random Forests that are trained on surrounding weather stations, whereas Zheng *et al.* [75] derive regions by using coefficients from a geographically weighted regression to split the regions into common sub-regions for which they fit one model per sub-region and suggest that this is necessary to achieve good performance. To test whether these claims hold true for the model adopted in this thesis, two pairs of regions are chosen, namely Europe and Africa and North- and South America. For each pair, a model is first trained on observations from each member separately and then on data from both and the errors are compared. The same test is repeated on the combination of both pairs.

3.1.4 Performance Evaluation of Gridded Data

Now that a method for determining features and models that work well for the application at hand have been defined, a few further error metrics shall be defined on top of existing ones defined in Section 2.2.4. These may be necessary to account for both the geospatial nature of the data as well as to measure more elaborate properties about the model and how it affects the error. As one of the properties of special interest in this thesis is bias and how it is corrected by the model, the correction model skill score, as introduced in Zampieri *et al.* [90] and defined as

$$\mathrm{CMSS} = 1 - \frac{|\mathrm{T_{cor}} - \mathrm{T_{obs}}|}{|\mathrm{T_{org}} - \mathrm{T_{obs}}|},$$

is used, where $\mathrm{T_{cor}}$ represents temperature as corrected by the model, $\mathrm{T_{org}}$ represents the original reanalysis temperature and $\mathrm{T_{obs}}$ refers to independent observations through weather stations. A CMSS of one indicates that the model perfectly removes all bias, $0 < \mathrm{CMSS} < 1$ indicates that the model reduces the bias, $\mathrm{CMSS} = 0$ means that the model has no effect on the bias and $\mathrm{CMSS} < 0$ indicates that the model deteriorates the bias.

While the metrics from Section 2.2.4 allow calculation of errors on a per-station basis, it is often helpful to visualize, aggregate and assess errors at larger scales. However, such assessments require careful consideration as they can quickly be skewed under the influence of non-homogeneous observation distributions. A global average error calculated by some metric that originates from networks experiencing strong clustering is dominated by those dense clusters [70]. In order to get a balanced view on the spatial error distribution and the

global error, irrespective of station density, a block-average error can be calculated. This is done by calculating the errors against observations in the common latitude-longitude grid, but then projecting them into the sinusoidal equal-area grid and aggregating them into $500 \times 500\,\mathrm{km}$ blocks, following Kilibarda *et al.* [70]. Aggregating on an equal-area grid alleviates the issue of a changing aggregation reference frame as it would be the case when using the latitude-longitude grid. This it due to the fact that at higher latitudes, the circumference of an iso-latitudinal cycle in latitude-longitude grids decreases while the number of pixels in raster images remains identical. Therefore, pixels at higher latitudes cover a narrower meridional range, leading to proportionally smaller extents used for aggregating contained points.

3.1.5 Determination of Urban Heat Island Effect

Finally, a way to measure the strength of the Urban Heat Island effect is necessary. As mentioned in Section 2.3, formation of UHIs is driven by a multitude of factors. When looking at the cross-sectional temperature profile of air temperature across a city and its surroundings, the highest temperatures are experienced inside the urban core, referring to the part of the city with the highest building density, whereas rural areas surrounding the city are comparatively colder. When looking at individual cities, a common approach for determining the UHI intensity is selecting two weather stations, one located inside or near the urban core and one outside of it in the rural surroundings of the city. For a global application, however, this approach does not scale, in part due to the lack of available weather stations located in such a geometry for many cities, but also due to ambiguity with respect to which stations to select in the case of multiple rural or city stations. To overcome these issues, UHI can be derived from gridded data sets that are available globally at the same resolution and by using appropriate algorithms for the delineation of urban and rural pixels.

Valmassoi and Keller [80] give a thorough overview of such methods and also propose one, which they call the nearest neighbor approach (referred to as M7 in [80]). M7 appears to be most promising and serves as the basis for the implementation in this work. However, some changes are necessary to adapt the algorithm to the coarser spatial resolution and the consequences that come with this.

In the adapted version of M7, the urban core of a city is derived in a first step. A city here refers to a polygon as contained in the GHS UCDB data set [91], which provides urban extents derived from satellite imagery. Valmassoi and Keller [80] use a threshold based on urban fraction, which works well at fine resolutions, such as $1\,\mathrm{km}$ or finer, where the urban core can be clearly resolved by an urban fraction raster alone. However, when averaging the urban fraction to the $5\,\mathrm{km}$ resolution used in this thesis, especially smaller cities often do not fulfill this threshold criterion anywhere anymore, and as such, no urban core can be derived with the original method. If no cells are found using the threshold-based approach, an additional step is added that selects the cell with the highest urban fraction inside the city extent. Through these steps, an urban core is marked and an exclusion zone

of $8\,\mathrm{km}$ is created around it, prohibiting any cells inside that radius to be used as rural reference pixels in the next step. Following [80], the algorithm now finds the closest five pixels with an urban fraction smaller than $20\,\%$ and a distance of more than $8\,\mathrm{km}$ to the urban core for each pixel contained inside the city extents. To speed up execution, the nearest-neighbor lookup is done using a BallTree [92] from `scikit-learn`. At this point, the algorithm presented here deviates from [80] once more by considering the elevation difference between rural and city pixels. The average cell elevation is used to correct for this difference by applying the same lapse rate approach as done for station measurements previously, using a lapse rate Γ of $-0.0065\,\mathrm{K\,m^{-1}}$, where the city pixel in question is used as the elevation reference and temperature from rural pixels is matched to it through their elevation difference to the reference. A similar correction is done in [93] and it is argued in [94] that, while it is a large adjustment with an empirically-derived lapse rate, it is the best option available when wanting to explicitly adjust for such differences. This step in general is vitally important, as CUHI can exhibit a small signal, where even an elevation difference of a mere $100\,\mathrm{m}$, causing a systematic error of $0.65\,\mathrm{K}$ under the assumed lapse rate, can already overpower the natural signal on some days. Once the temperature of the rural neighbors is adjusted, the five points are averaged and the difference between the two series yields the CUHI of the city pixel. This approach is then repeated for all pixels inside of a city and all cities of interest to yield a gridded timeseries of daily CUHI anomalies per city. These can then be further aggregated, with a common UHI intensity indicator being the summer season average, i.e. the average CUHI across June, July and August on the Northern Hemisphere.

3.2 Results

Having covered all the steps involved in the implementation, it is now time to look at the results produced by the model in more detail. First, a short summary of the findings from model selection is given and the effect of station density on the model predictions is evaluated. Afterwards, a comparison of the raw reanalysis and improved model results will follow and finally, the model's ability to add sub-gridscale details is evaluated, along with a preliminary calculation of the CUHI anomaly.

3.2.1 Results of Model Selection

As mentioned, grid search with cross validation is used to assess different models with respect to their performance on the target regression problem. All training steps were conducted on a machine with 32 cores and $256\,\mathrm{GB}$ of memory. Running through the entire parameter space as documented in Table B.1 for ExtraTrees and Random Forests, in Table B.2 for Gradient Boosting, in Table B.3 for Histogram Gradient Boosting and in Table B.4 for the MLP, the results of the best set of parameters for each model can be found in Table 3.3.

It can be seen in Table 3.3 that ExtraTrees provide the best performance after tuning, but

Table 3.3: Results of various error metrics after model selection. Training was conducted on a spatial and temporal subset, containing observations from Germany and neighboring countries between 2010 and 2014 with all features from Table 3.2 selected during feature selection.

Model	**RMSE** [K]	**MAE** [K]	R^2
Random Forest	1.14	0.79	0.97
ExtraTrees	1.12	0.79	0.95
Gradient Boosting	1.15	0.80	0.96
Histogram Gradient Boosting	1.26	0.88	0.97
Multilayer Perceptron	1.29	0.91	0.98

are closely followed by Random Forests and Gradient Boosting, all yielding similar performance. Histogram Gradient Boosting and Multilayer Perceptrons perform worse, but provide comparable or higher R^2 values with respect to the other models. While one may just use the best-performing model, it is also necessary to understand practical aspects, such as memory necessary to store the trained model and time to train. When training on the global data, Gradient Boosting is aborted after roughly 15 h of training time, indicating a scaling issue at the size of the training data used in this thesis. Training the Random Forest and Extra Trees regressor with the optimal set of parameters found during hyper-parameter optimization takes around 5 h initially. However, the trained models occupy around 80 GB and 82 GB of memory, respectively. This number even doubles when the model is loaded from disk, as the implementation loads the entire contents and then serializes them into a Python object, thus necessitating a machine with more than 160 GB memory just to be able to load the model. Adapting the hyper-parameters slightly by limiting the maximum depth of trees and increasing the minimum number of samples per leaf to produce smaller models, a new Random Forest and Extra Trees model is then trained, this time taking around 4.5 h to train and producing models requiring roughly 15 GB of memory when loaded into memory. However, due to the change of hyper-parameters, the new Random Forest (RF) and ExtraTrees (ET) model yield an RMSE of 1.18 K (1.14 K) and 1.20 K (1.12 K), respectively, on the same set used during hyper-parameter optimization, making them slightly worse than their optimal counterparts indicated by the values in parentheses. Considering that Random Forests yield a competitive error compared to Gradient Boosting, even when using sub-optimal hyper-parameters, and that they are much faster to train, a choice is made to use Random Forests for all subsequent modelling steps. The final set of hyper-parameters used for the Random Forest can be found in Table B.5.

While an optimized model is a necessary pre-condition for achieving a good solution of the problem at hand, it is also important to understand how the model comes to its predictions. In a first step, the influence of individual training features can be analyzed, as not all features contribute equally to the predictions made by the Random Forest. This can be achieved by iterating through all trees in the ensemble and for each tree to accu-

mulate the reduction of the loss function at each tree node for each feature separately. The accumulated number for each feature can then be summed up across the trees and its magnitude describes the importance of that feature in the decision process of all trees, where higher numbers indicate a larger importance. However, this way of determining feature importances suffers from a bias towards high-cardinality features when the RF has capacity to overfit the training data [95]. To counteract this, permutation-based importances can be used that determine feature importance by randomly permuting each feature from the model inputs and comparing the errors produced before and after permutation. The rationale is that when a feature is highly correlated with the target variable, its permutation will noticeably increase the error, compared to irrelevant features that will barely affect the error amplitude during prediction. Figure 3.1 depicts the feature importances derived in both ways for the Random Forest trained on the global data. While the native importance measures based on reduction of the loss function at tree nodes are derived from the training data, the permutation-based importances are derived from the independent holdout set, also giving insight into which variables are truly important when considering unseen data.

As can be seen in Figure 3.1, features relating to air and surface temperature directly, such as lapse-rate-corrected ERA5 daily maximum temperature and the gap-filled MODIS land surface temperature fields, have the highest influence under both importance measures, whereas other reanalysis parameters, like daily mean wind speed and sensible heat flux, do not appear to play as much of a role, but more so when looking at the tree node-based importance than permutation importance. Regarding the relevance of geographical and temporal features, a lot of variability exists, with latitude for example being much more important than longitude during training, but essentially identical during inference. The most visible trend is that ERA5 has a much higher influence as measured by permutation importance compared to tree node-based importance, where the former shows a difference of more than one order of magnitude between ERA5 temperature and the next-important feature, while they have a much more comparable importance in the latter.

Lastly, to get a feel for the sensitivity of the model to different climates, one Random Forest was trained on multiple regions of varying spatial extent and with varying numbers of stations inside each region located in different climatic zones. The goal of this test is to show whether training on data from all latitudes affects the quality of predictions made on high latitudes for example. Table 3.4 shows the errors when training on observations from these differently-sized regions from the training set and evaluating only on parts of the region using data from the test set.

It is clearly visible, when looking at the pairs of Europe and Africa and North- and South America individually, errors remain essentially equal when training on one or both of the members of each pairs and validating against one of them. When training on all four regions, errors very marginally increase, but still remain very close to the errors of the model trained on a smaller region.

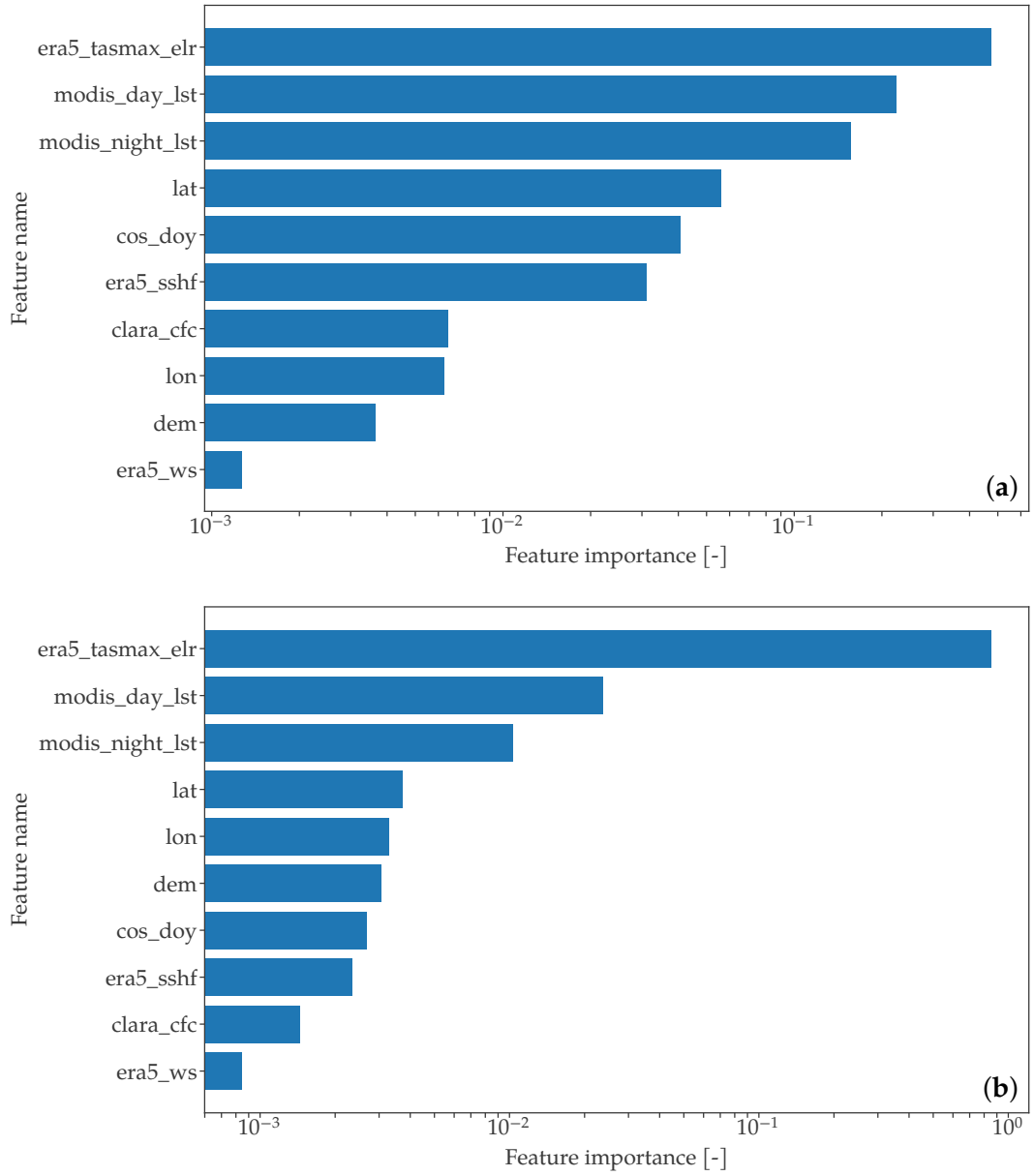

Figure 3.1: **(a)** shows feature importances derived from the Random Forest model trained on global training data across the entire time frame, **(b)** shows feature importances derived from permutation importances instead. Note the logarithmic x axis.

Table 3.4: Comparison of model errors when training on differently-sized regions and validating results against the test set. All points from the training and test set are used inside the respective regions, where Europe refers to the bounding box $[-18.28, 51.86, 37.0, 71.33]$, Africa to $[-18.28, 51.86, -52.38, 37.00]$, North America to $[-143.61, -33.93, 24.00, 57.23]$ and South America to $[-143.61, -33.93, -52.38, 24.00]$ for the year 2010 to shorten training times.

Training region	Validation region	RMSE [K]	MAE [K]	R^2
Europe	Europe	1.95	1.44	0.972
Europe and Africa	Europe	1.92	1.42	0.972
North America	North America	2.75	2.03	0.951
North- and South America	North America	2.74	2.03	0.951
Europe, Africa, North- and South America	Europe	1.97	1.47	0.971
Europe, Africa, North- and South America	North America	2.78	2.06	0.950

3.2.2 Influence of In-Situ Observation Density

With the choice of model made, the effect of station density on the model's ability to improve predictions of daily maximum temperature is investigated now. As previously established, the density of observations around the globe varies greatly by country. Figure 3.2 shows the distribution of stations from the merged GHCNd and GSOD stations used for training and validation as well as the subset of those stations that is operated by the World Meteorological Organization (WMO) for Europe and the United States (US). Density of the WMO stations is crucial for the accuracy of ERA5, as it exclusively assimilates measurements provided by those stations [27].

Figure 3.2 shows that the density of stations from GHCNd and GSOD varies greatly around the globe, with the US showing a very high density, whereas Europe exhibits a more varied distribution depending on the respective country. While Germany has a very dense network, France, Ireland and Spain are covered by a comparatively much sparser network. On the contrary, when looking exclusively at stations operated by the WMO, the situation changes and the density in the US is much lower compared to Europe. When focusing on WMO stations in the US, an east-west gradient exists, with the eastern US having a higher density compared to the topographically more complex western part. Especially around the Rocky Mountains, the WMO network is sparse.

In order to now assess the influence that station density exhibits on the proposed algorithm, a small region containing observations from Germany and parts of its neighboring countries is chosen, offering around 240 stations with valid data in the training set for the year 2014. This region is chosen based on its high density of observations, both from

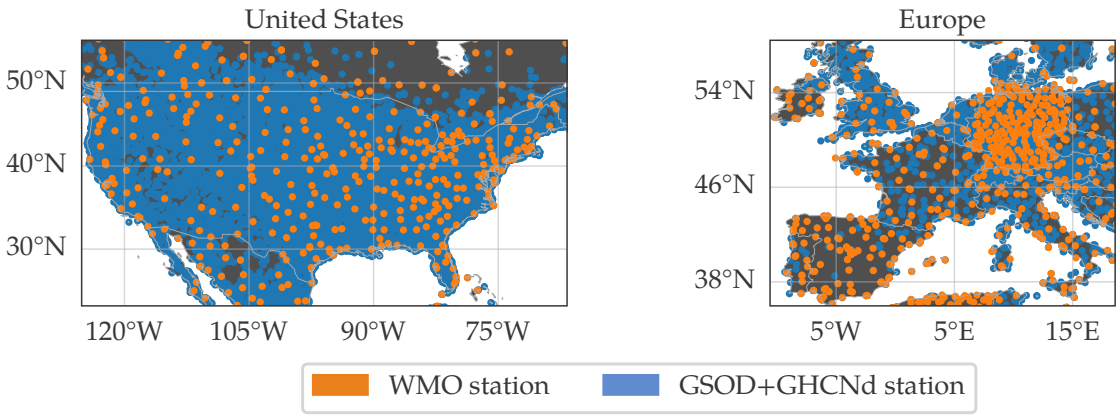

Figure 3.2: Location of observations from GHCNd and GSOD compared to those run by the World Meteorological Organization (WMO) for the United States and Europe. WMO stations are those that are contained in either GHCNd or GSOD, but there may be additional stations that are not contained in either of them around the globe.

GSOD and GHCNd, but also from those operated by the WMO in order to guarantee high accuracy of the ERA5 baseline maps due to a large amount of already assimilated information. All stations contained in that region are then split into a training and validation set, based on the label assigned to them during pre-processing. Then, N stations are randomly picked from the training set, starting with $N = 240$, i.e. all available training stations, and a model is trained on them and validated on all stations in the validation set. This step is repeated, but each time halving the number of stations N, until only a small fraction is left. For each of these subsets, the prediction error is noted and a summary of errors at different station densities can be found in Table 3.5.

As expected, decreasing the number of stations over a fixed training region causes a decline in model predictive performance. While the model trained on 3 stations produces predictions worse than the ERA5 reanalysis baseline as measured by its RMSE, even 7 stations appear to be enough such that model predictions match the baseline error, while increasing the number to 15 stations yields a model that out-performs it. Adding more stations continues this trend, with the model trained on all 240 stations producing bias-free predictions along with a $22.8\,\%$ improvement in RMSE. The bias does not linearly improve when more observations are added, but varies between -0.04 and 0.12 without any visible correlation with the number of stations. Compared to the ERA5 baseline, the model reduces the bias greatly, irrespective of the number of stations used during training.

To further assess the change in bias, a second experiment is conducted where at each halving, not one, but 25 models are trained and their errors are averaged. This way, the random subset of stations that each model is trained on varies, thus alleviating the effect of

Table 3.5: Summary of the influence of stations density on model performance in a fixed training area. The Random Forest is trained on a decreasing number of stations and its errors, as measured by RMSE and Bias, as well as the R^2, are tracked on a static validation set. The last row contains the baseline results from ERA5 at all validation stations.

Number of stations	**RMSE** [K]	**Bias** [K]	R^2
240	1.08	0.00	0.98
120	1.12	−0.04	0.98
60	1.20	0.01	0.98
30	1.24	−0.04	0.97
15	1.31	0.10	0.97
7	1.41	0.12	0.97
3	1.51	0.03	0.96
-	1.40	−0.587	0.97

some stations being more representative than others and therefore influencing the model performance more. At this point, the declining pattern in RMSE remains identical, but bias mostly stabilizes and consistently remains $\leq 0.02\,\mathrm{K}$ for 15 and more stations, while remaining $\leq 0.1\,\mathrm{K}$ at all station densities. R^2 also remains almost identical to Table 3.5. The results from averaging the 25 models can be found in Table C.1.

3.2.3 Performance Comparison of Raw and Improved Data

Finally, this section now presents results of the maps generated in this thesis compared to those from the ERA5 reanalysis baseline and CHIRTS-daily, another data set that provides global daily maximum temperature data at $5\,\mathrm{km}$ horizontal resolution, but uses a much different methodology [96]. In order to quantitatively assess how model maps improve upon the baseline, global summary statistics will be shown first, followed by a more in-depth analysis of the temporal and spatial behavior of error correction as performed by the model. Lastly, the model's ability to correct bias in the baseline data will be assessed. In order to remain consistent with the model methodology and to fairly assess the different data sources, all comparisons between data sets are made based on the lapse-rate-corrected observations. The native elevation data, corresponding to averaged MERIT [97] elevation for fields produced by this thesis's model (further referred to as model), the ERA5 orography for reanalysis fields and GTOPO30 [98] for CHIRTS-daily (further referred to as CHIRTS), are used to calculate an elevation difference between the elevation grid and observation locations. This difference is then used to derive corrected observations for all models, according to Equation (3.1) with a lapse rate of $\Gamma = -0.0065\,\mathrm{K\,m^{-1}}$, which is utilized for all further assessments. The resulting effect on ERA5 time series data can be seen in Figure D.1, depicting a marked improvement in monthly mean RMSE across all months

and years in the data when comparing ERA5 daily maximum temperature against the observations. Subsequent validation is based on observations from the test set, as defined in Section 3.1.1, i.e. those stations that the model never saw during training or model selection.

When simply comparing global error metrics, as defined in Section 2.2.4, calculated from all global in-situ observations from the test set over the entire period of 2003–2020, the RMSEs achieved by ERA5, CHIRTS and the model are 2.62 K, 2.85 K and 2.31 K, with a mean bias of −0.48 K, 0.29 K and 0.04 K, respectively. Figure 3.3 depicts histograms of the mean bias of all global stations from the test set across the entire time span from 2003–2020.

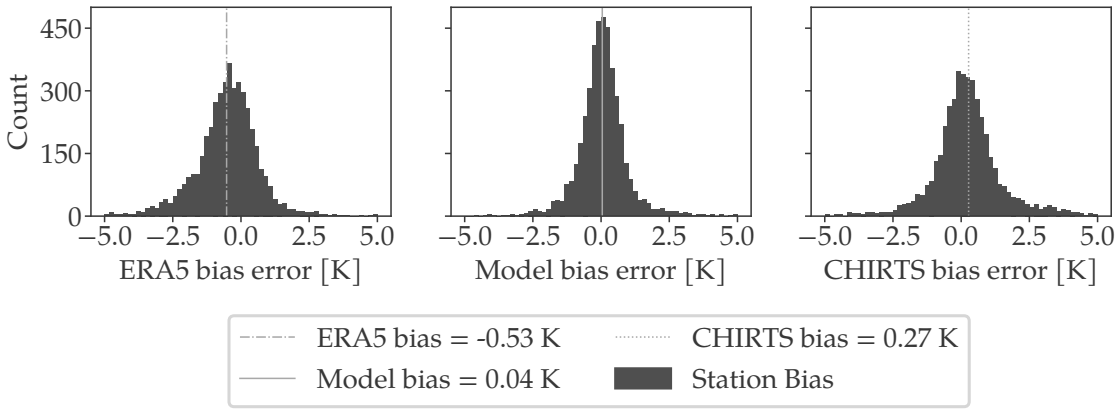

Figure 3.3: Histograms showing average bias per station across 2003–2020 derived from ERA5, the model and CHIRTS. All stations come from the holdout set.

When looking at histograms of average station bias in Figure 3.3, ERA5 exhibits a minor cold-bias of −0.53 K for daily maximum temperature, whereas CHIRTS shows a slight warm-bias of 0.27 K. Meanwhile, the fields produced by the model are essentially bias-free with a residual bias of 0.04 K and also show a tighter spread of the distribution. The histograms produced by the model and CHIRTS exhibit a slight right-skewedness with an empirical skewness of 0.35 and 0.75, while ERA5 is slightly left-skewed with an empirical skewness of −0.41. Especially the tails of the distribution improve when comparing the model bias to ERA5 and CHIRTS, where the number of stations exhibiting an absolute bias between 2.5 and 5.0 K is noticeably reduced.

While unbiasedness is a promising first indicator, the temporal evolution of errors is equally important. Figure 3.4 shows errors from the model and ERA5 against observations, disaggregated into long-term monthly mean RMSEs along with bands of the inter-quartile range (IQR) of the RMSE.

It becomes visible in Figure 3.4 that both ERA5 and the model produce consistently lower errors during the June, July and August (JJA) as compared to the rest of the year. Looking at the IQR of RMSE, it appears to be narrower during JJA in both ERA5 and the

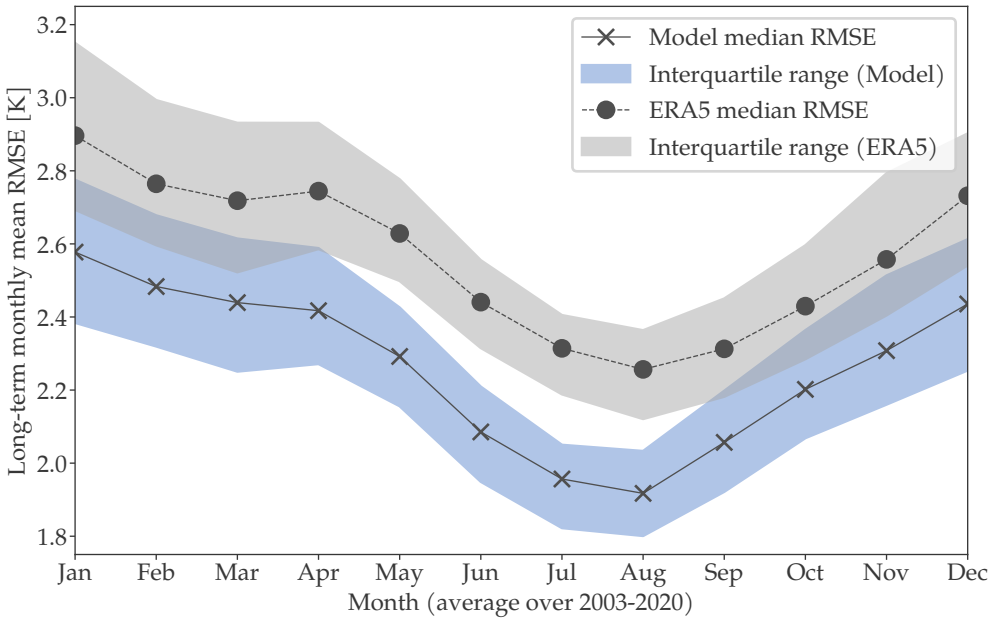

Figure 3.4: Long-term monthly means of RMSE derived from model and ERA5 reanalysis against observations. Bands represent the inter-quartile range of RMSE.

model, while widening during September, October and November (SON) and being the widest during December, January and February (DJF). For easier visual representation, CHIRTS was left out of the graph, but follows a similar temporal pattern over the span of the year.

In Figure 3.5, the evolution of the global mean RMSE over time can be observed. As the number of stations reporting valid measurements varies on a daily basis, the comparison is done on a subset of all test stations that have at most 250 days of measurements missing across the entire time span of 6575 days between 2003 and 2020. This way, variations in the plotted RMSE correspond to a real change in the error of the model instead of different reporting stations that exhibit different errors.

As Figure 3.5 shows, a similar trend to Figure 3.4 emerges, with the JJA season exhibiting lower errors compared to the DJF season. There exists some amount of inter-year variability, where some months in one year show lower errors compared to the same month in another year. The month in which the minimum RMSE occurs varies, but always happens between June and September for both ERA5 and the model. ERA5 exhibits a statistically-significant trend in RMSE change per year at $p < 0.05$ during May and October, with an amplitude of $0.0050\,\mathrm{K\,yr^{-1}}$ and $0.0094\,\mathrm{K\,yr^{-1}}$, respectively. The model only exhibit a statistically significant trend at $p < 0.05$ only during May with a slope of $0.0055\,\mathrm{K\,yr^{-1}}$, corresponding closely to the value from ERA5.

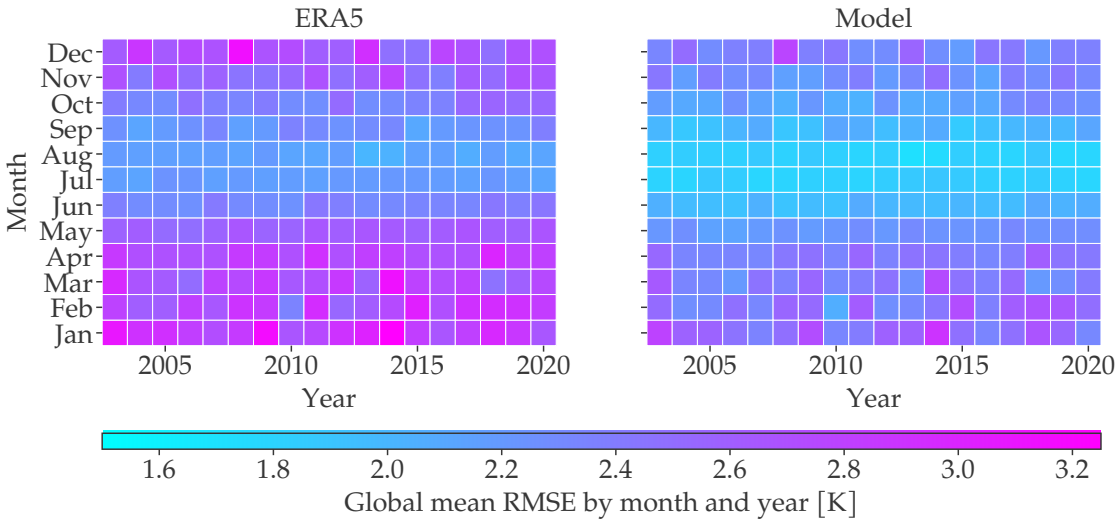

Figure 3.5: Time evolution of the global monthly mean RMSE on the test set. All station RMSEs are averaged for each month and year in the produced data separately.

Besides a temporal disaggregation, the model performance is also assessed with respect to different land use classes. To do so, data from the Copernicus Global Land Cover Layers for the year 2015 [99, 100] is used and aggregated to a $5\,\mathrm{km}$ horizontal resolution through using the majority class. Multiple classes representing different kinds of forests are merged and each in-situ observation is then assigned the land use class of the nearest pixel from the global layer. Figure 3.6 shows RMSE, Bias and number of stations located in a certain land use class for the entire period of 2003–2020 across all different classes.

It can be seen in Figure 3.6 that all land use categories exhibit lowered errors after model application, with the *ocean* class seeing the largest improvement. Reduction in error ranges between $0.2\,\mathrm{K}$ to $0.33\,\mathrm{K}$ and corresponds to a relative improvement of roughly $11\,\%$. Note that the *ocean* class in this case refers to cells located at the coast. This erroneous classification can happen when the coarse-scale aggregation region, used during land use averaging, only slightly overlaps the land mass near the coast and is therefore dominated by pixels classified as being an ocean. As a result of the majority vote, observations near the coast are assigned the *ocean* class, while in reality they are located on land. The same behavior can be seen with the *water* class, where observations near lakes or large rivers are assigned the wrong class, while in reality being located on land. After model application, all classes, except for *wetlands*, exhibit slightly positive bias, whereas all values from ERA5 exhibited a negative bias. Further analysis of the *wetlands* class shows that a single station with a large negative bias of $-3.22\,\mathrm{K}$ exists in the test set that skews the summary statistics. Removing that, the mean bias becomes $-0.078\,\mathrm{K}$, putting the value in line with the remaining classes. From all classes, the model slightly over-corrects cells near lakes and rivers, with the *water* class exhibiting the largest positive bias after correction.

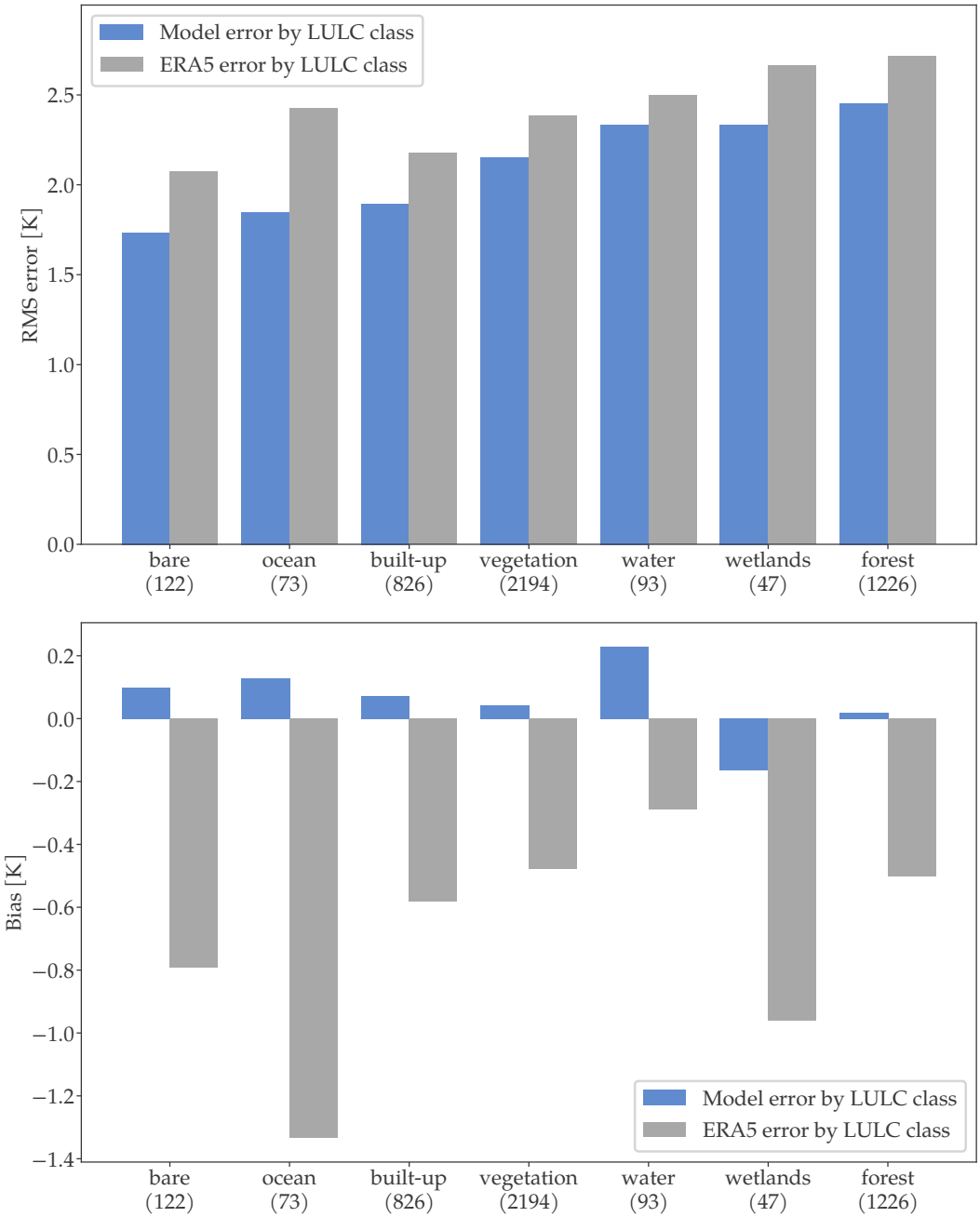

Figure 3.6: Model and ERA5 RMSE and Bias, disaggregated by land use class. Classes are assigned to each station by nearest-neighbor interpolation from Copernicus Global Land Cover data, aggregated to $5\,\mathrm{km}$ resolution. Numbers in parentheses below the label indicate the number of stations with that land use class.

Next, the spatial distribution of errors is assessed. The block-aggregated MAE, following Section 3.1.4, is shown in Figure 3.7. This refers to the spatial aggregation of all mean absolute errors between modelled temperature fields and observations, located inside of an aggregation region.

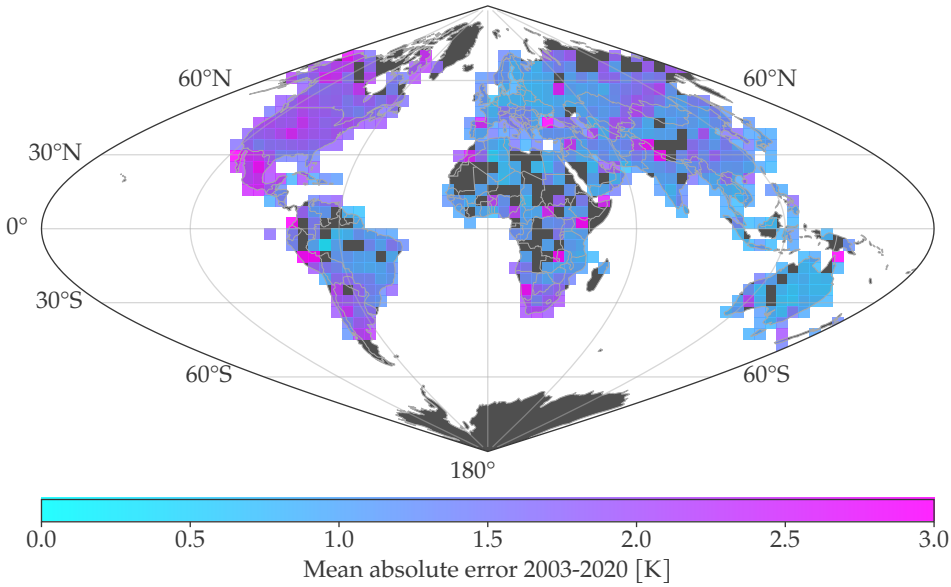

Figure 3.7: Global block-MAE, as defined in Section 3.1.4, aggregated to 500×500 km resolution in the equal-area sinusoidal projection.

Figure 3.7 shows that stations in the continental US exhibit higher average errors, with regions located near the Rocky Mountains exhibiting noticeably larger errors than neighboring cells. Along the west coast of South America, errors tend to be elevated near the Andes and generally increase towards the South of the continent. Eurasia shows a balanced distribution of errors, with pixels around the Tibetan Plateau exhibiting increased errors and generally increasing towards higher latitudes. Africa similarly shows a homogeneous distribution, with a slight increase of error in South Africa. Australia appears to have the lowest errors globally with a slight increase in amplitude around New Zealand. Overall, Europe and Australia show very consistent and smooth error fields, while other regions exhibit choppier fields with more variation among neighboring pixels.

Figure 3.8 depicts the relative improvement in block-aggregated MAE of the model over ERA5, defined as

$$\mathrm{MAE}_{\mathrm{rel}} = \frac{\mathrm{MAE}_{\mathrm{cor}} - \mathrm{MAE}_{\mathrm{org}}}{\mathrm{MAE}_{\mathrm{org}}},$$

where $\mathrm{MAE}_{\mathrm{cor}}$ is the MAE between the model's fields and observations and $\mathrm{MAE}_{\mathrm{org}}$ is the MAE between ERA5 fields and observations.

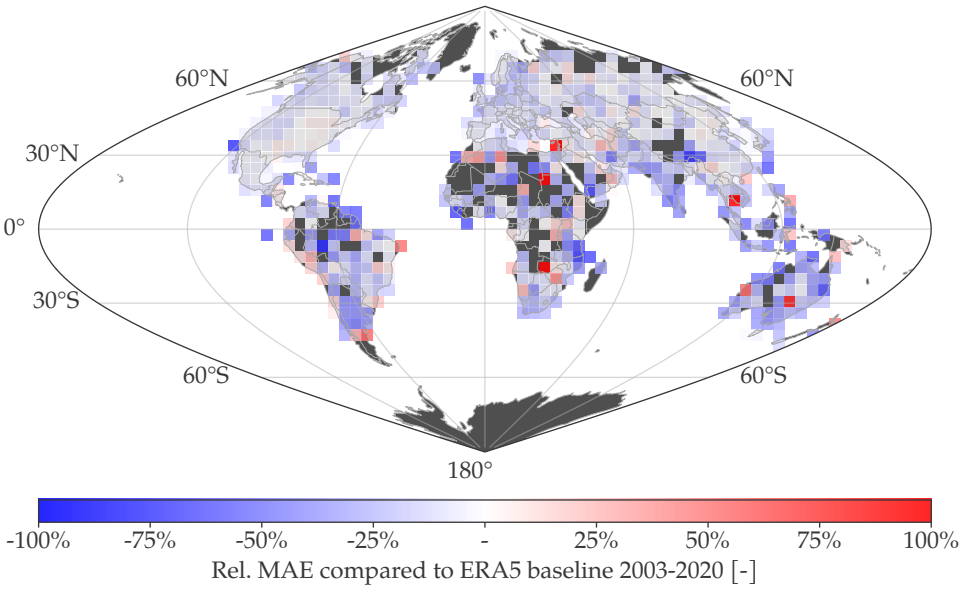

Figure 3.8: Relative change of block-aggregated MAE between model and ERA5, measured against observations.

As can be seen in Figure 3.8, especially Europe and the US do not appear to improve much after model application, whereas the Southern Hemisphere (SH) sees more widespread improvements in many countries. The largest improvement in the Northern Hemisphere (NH) can be found around India and South-East Asia, with the mid-latitudes not seeing much improvement at all. Towards high latitudes, skill increases a bit again. In the SH, the largest improvements can be found in coastal regions around Africa, the South of South America and Australia. There also tend to be sporadic outlier pixels that show a decrease in performance.

In order to investigate the effect of elevation on the performance of the model further, Figure 3.9 shows the RMSE of all stations in the test set against the elevation.

As Figure 3.9 shows, both ERA5 and the model show a clear increase in elevation-binned average RMSE up to about $3500\,\mathrm{m}$ elevation. Above that, errors in ERA5 appear to fluctuate randomly, but stay below the levels found at lower altitudes. A similar pattern can be observed in the model histogram, however, the random fluctuations appear to start at roughly $3000\,\mathrm{m}$ already. Above that, the model exhibits errors of the same magnitude as at low elevations between $0\,\mathrm{m}$ and $400\,\mathrm{m}$.

Finally, looking at the block-aggregated CMSS, the model exhibits mixed performance, as can be seen in Figure 3.10. The map is derived by calculating the CMSS for each station, aggregating it in the sinusoidal equal-area projection and then mapping all values larger than zero to indicate an improvement in bias and the rest to negatively affect the bias. This

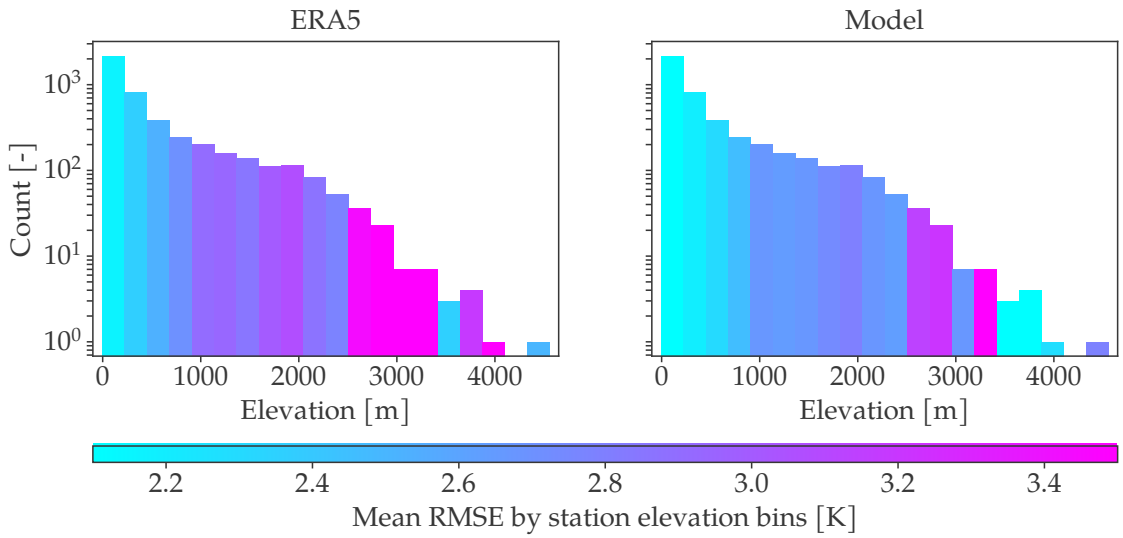

Figure 3.9: Histogram showing the RMSE of all test stations averaged into elevation bins. The shading of individual bars indicates the average station RMSE inside that elevation bin. Results of ERA5 are shown on the left, while the model results are shown on the right. Note the logarithmic y axis to account for a majority of stations being located at lower elevations.

step is done as the CMSS is unbounded to negative infinity, sometimes causing almost perfect predictions by the baseline reanalysis to inflate slightly worse predictions by the model to skew the station-average CMSS.

As with the relative MAE, Figure 3.10 shows that CMSS improvements do not follow any systematic pattern except in the US, where model application almost universally decreases the performance over the baseline. Similarly, decreased performance is shown in many parts of Central Europe, except in Spain and the southern Nordics. For South America, the performance of the east and west coast has worsened, while pixels located in the Amazon rain forest show improvements. The rest of the globe sees mixed improvements, sometimes improving in one block while worsening in a neighboring one. Performance for Australia tends to improve overall, while Tasmania and New Zealand worsen.

To conclude this section, Figure 3.11 shows plots of the mean daily maximum temperature around the Alps, calculated from the entire period of 2003–2020 for ERA5, CHIRTS and the model to illustrate the difference in resolution over the baseline maps and to give an impression of the added sub-gridscale detail when compared to ERA5.

The much improved detail of both CHIRTS and the model over ERA5 is immediately visible in Figure 3.11, owing in large to the much more detailed representation of elevation. When looking at very high elevation, mountain tops appear colder in the model maps compared to CHIRTS, as the model uses raw elevation during every step compared to

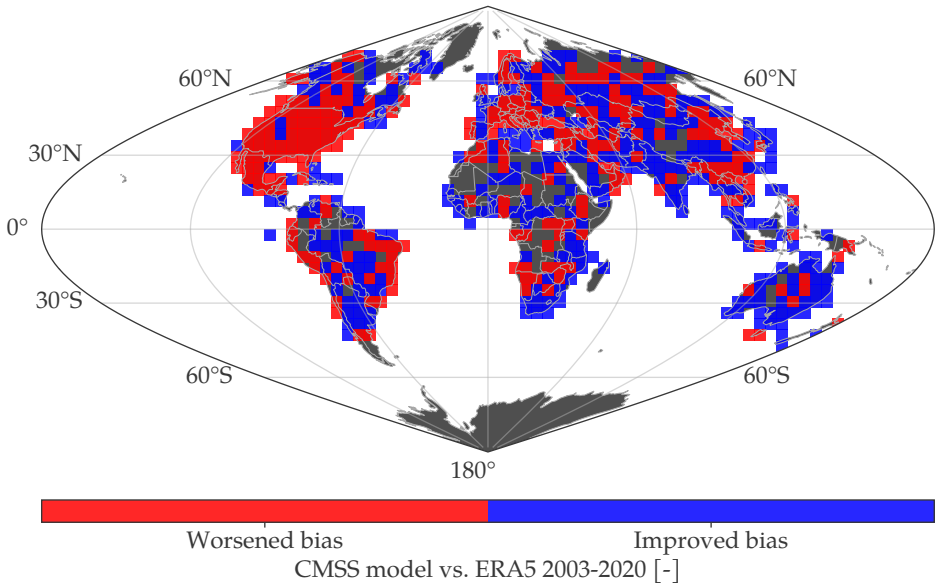

Figure 3.10: Block-aggregated CMSS with binary mapping, indicating blocks with overall improved and worsened bias after model application. Improved means that CMSS ≥ 0, whereas worsened indicates CMSS < 0.

CHIRTS, which uses a square-root transformed elevation in its regression model to counteract unrealistically cold mountain top temperatures without it [101]. Both CHIRTS and the model properly represent small valleys, which are not visible in the ERA5 maps due to its coarse horizontal resolution.

In summary, the model can successfully remove the bias of daily maximum temperature fields compared to ERA5 baseline data, producing a year-round reduction in RMSE, irrespective of the land use class. Spatial behavior tends to vary across regions, with the NH experiencing in general a smaller improvement compared to the SH. However, the improvement of bias, as measured through the CMSS, is not uniform globally, with the US showing a clear decrease in performance after model application.

3.2.4 Ability of Model to add Sub-gridscale Variability

While previous chapters have looked at large-scale evaluation of the model, it is also important to understand the performance at the smallest scales. Since downscaling is one of the two tasks the model is supposed to excel at, a high-resolution numerical simulation, produced by the UrbClim model [102], is used as ground-truth data and the model predictions are compared against it. Tests are conducted for selected cities in Europe and the data used comes from the Copernicus Climate Change Service, which has run UrbClim simulations for a hundred European cities between the years 2008–2017 and produced maps with

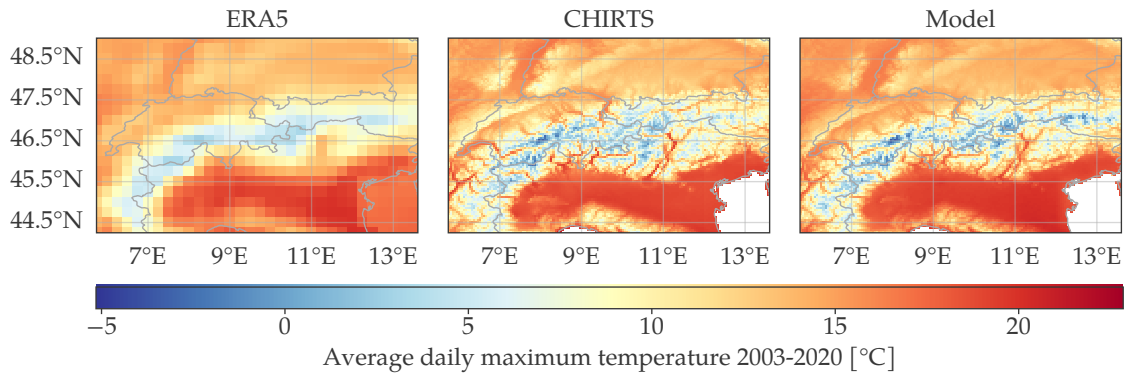

Figure 3.11: Long-term average daily maximum temperature map over the Alps and surrounding countries between 2003–2020, as calculated from ERA5, CHIRTS and the model.

a hourly temporal and 100 m spatial resolution of air temperature, relative and specific humidity and wind speed [103]. To account for the drastically higher horizontal resolution provided by UrbClim, the fields are first averaged to approx. 5 km and then a daily maximum of the simulations is derived. In a subsequent step, temperature fields produced by the model are interpolated to the averaged UrbClim extents from the previous step using first-order conservative regridding. The evaluation of model maps is conducted for Paris, Milan and Barcelona, as those cities show both a large CUHI amplitude [93] as well as being affected by different environmental processes. Paris represents a continental city, while Milan and Barcelona are both affected by complex environmental interactions due to their mountainous and coastal surroundings, respectively. The results of calculating long-term monthly bias errors are shown in Figure 3.12, Figure 3.13 and Figure 3.14.

It can be seen in Figure 3.12 that for Barcelona, the model overestimates the coastal temperature in the south-east at the coast all year around, while showing a slight, but persistent, under-estimation for pixels with higher urban fraction near the center of the image. Underestimation in the central pixels is more pronounced during March, April and May (MAM) and JJA season, whereas during SON and DJF, they slightly overestimate the temperature. In general, fluctuations are more pronounced away from the city center, whereas in the areas with high urban fraction, the bias remains neutral between October and February, while becoming slightly negative during summer. The modelled temperature fields for Barcelona exhibit an R^2 value of 0.94 and their RMSE ranges from 0.79 K during December to 1.47 K during August.

For Paris, as seen in Figure 3.13, the model exhibits some of the same properties as seen in Barcelona. Near the city center, bias is neutral during April to August, while showing a pronounced cold-bias during October until February and transitioning between the two states in September and March, respectively. A warm bias with increasing ampli-

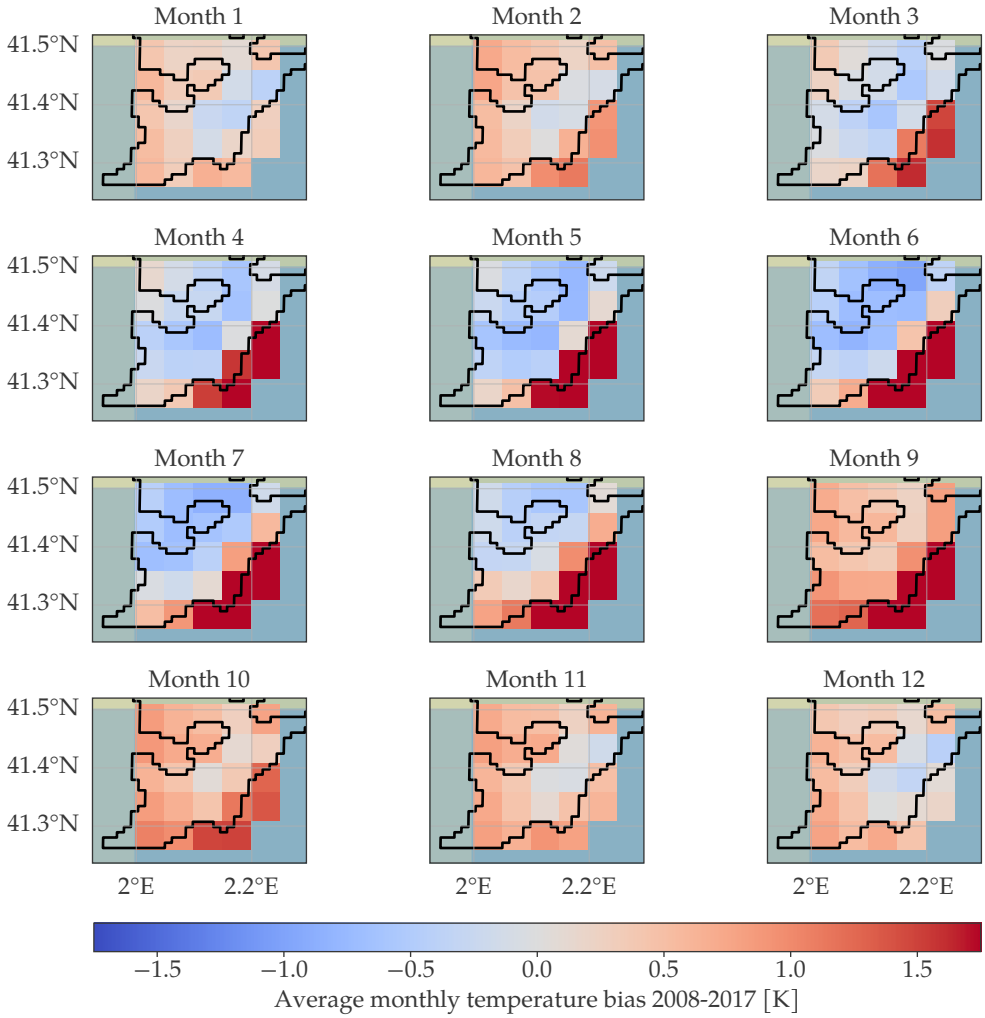

Figure 3.12: Bias of modelled temperature fields versus UrbClim reference for Barcelona. Positive bias indicates the modelled temperature to overestimate true temperature. Black outlines mark the city extents.

tude spreads radially around the city center during April to August, while being less-pronounced / neutral during October to February. In contrast, during October to February, the city center shows a strong cold bias. Interestingly, the bias appears to behaves like a constant offset, where the city center shows neutral bias, but rural surroundings show a warm bias during during April to August, while the city center shows a cold bias during winter months, but the surroundings are neutral. The modelled field exhibits an R^2 value of 0.97 and the RMSE varies between $0.87\,\mathrm{K}$ during October to $1.24\,\mathrm{K}$ during July.

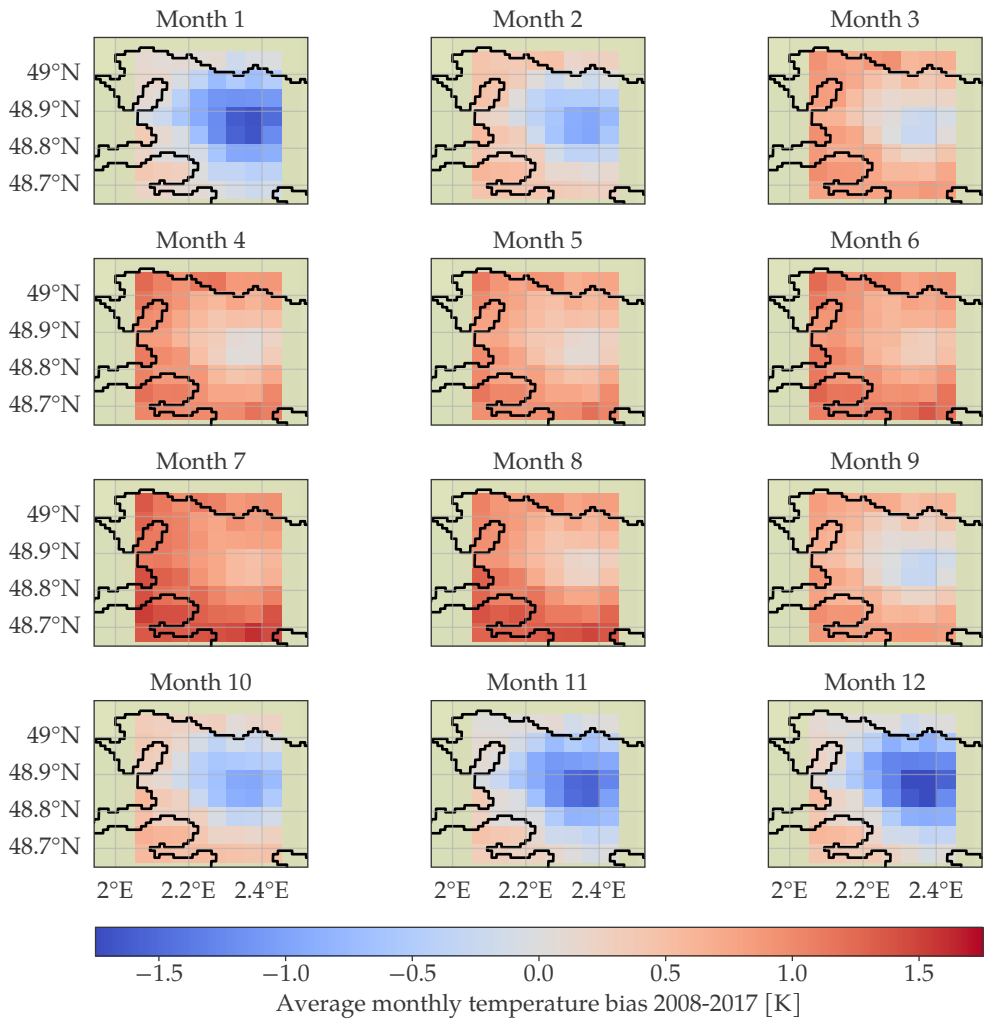

Figure 3.13: Bias of modelled temperature fields versus UrbClim reference for Paris. Positive bias indicates the modelled temperature to overestimate true temperature. Black outlines mark the city extents.

Finally, for Milan in Figure 3.14, the model shows yet different behavior to the prior two cities. This time, only a very slight north-south gradient is visible in the fields, but the warm-bias during MAM and SON persist, while turning into a cold bias during October to February. Modelled temperature fields for the city show an R^2 value of 0.97 with an RMSE error ranging between $0.93\,\mathrm{K}$ during November to $1.49\,\mathrm{K}$ in July.

After looking at these cities, it can be concluded that compared to UrbClim, there does seem to be a city-dependent systematic difference with respect to spatial patterns exhibited by daily maximum air temperature. While the trend behaves like a constant bias

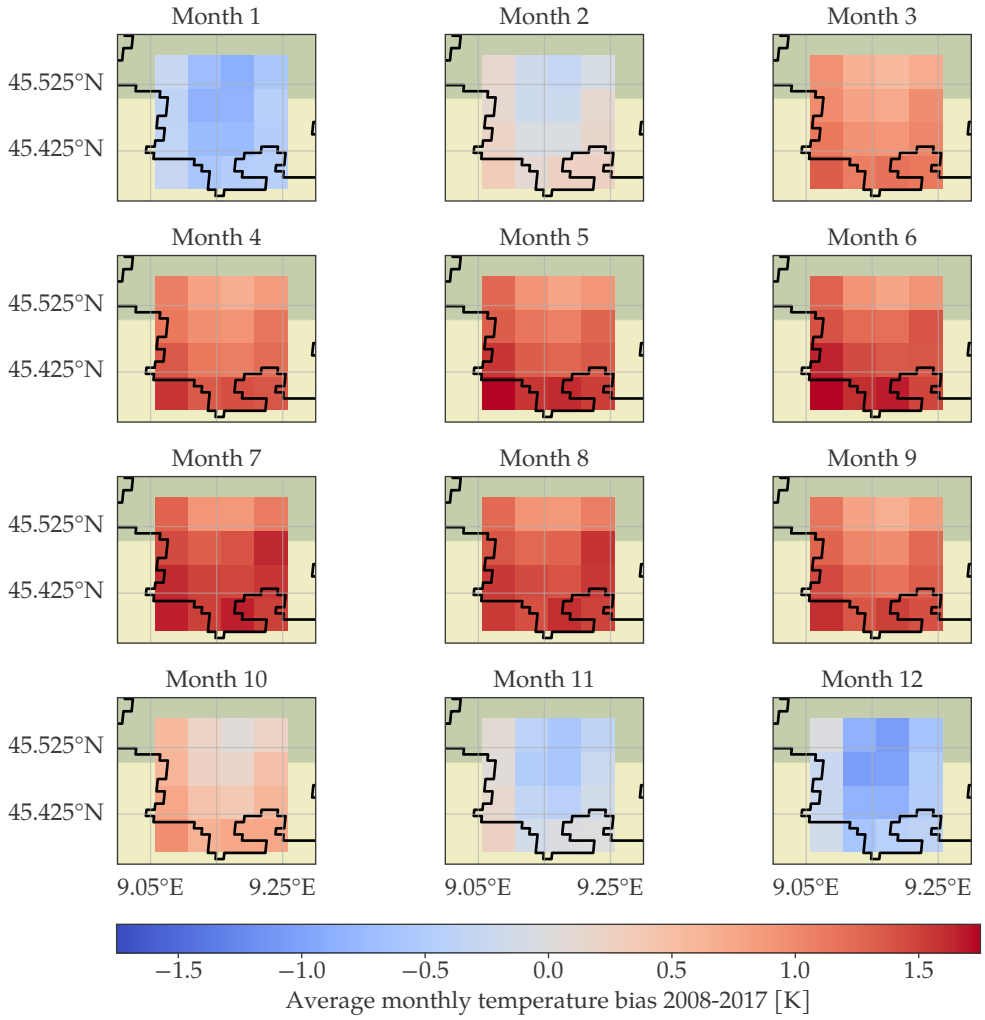

Figure 3.14: Bias of modelled temperature fields versus UrbClim reference for Milan. Positive bias indicates the modelled temperature to overestimate true temperature. Black outlines mark the city extents.

offset in continental cities, such as Paris and Milan, its spatial pattern is much more erratic in Barcelona, which is located right next to the coast. The overestimation during summer and under-estimation during winter appears to be consistent for the continental cities, but inverted in Barcelona, where a visible cold bias exists during summer months away from the coast. Interestingly, the bias has similar amplitude to the RMSE for these cities, indicating that the largest error source is systematic mis-estimation of temperatures instead of random errors.

3.2.5 Assessment of Model Physicality

The final test is an assessment of the physicality of the model. As the machine learning approach chosen here does not enforce any physical constraints during training, it is nevertheless vital to understand whether temperature fields produced by it follow physical patterns or whether they contain spurious, non-physical features. To assess this property of the model, the power spectral density (PSD) of temperature fields is compared with numerical simulations at high resolutions as well as CHIRTS. PSD plots are a tool that has been applied in many publications on downscaling of weather- and climate-related variables [104, 105, 106]. Very generally, PSD describes the power of individual frequencies constituting a physical signal, which can be calculated through the use of the Fourier Transform. For signals, such as time series, that vary only in one parameter, this produces a simple-to-interpret two-dimensional plot, whereas for multi-dimensional signals, such as images, one has to use a two-dimensional Fourier Transform, which does not directly allow a presentation of power versus frequency. Therefore, averaging through the use of the radially-averaged PSD (RAPSD) is applied, which aggregates identical frequencies across dimensions of the input signal to allow for a simple power-versus-frequency representation. In this thesis, the RAPSD implementation of PySteps [107] is used. For a thorough mathematical description, the reader is deferred to [108].

Besides the model developed in this thesis, temperature fields from the Uncertainties in Ensemble of Regional Reanalysis (UERRA) / Copernicus Regional Reanalysis for Europe (CERRA) [109] model are used, available at a $5.5\,\mathrm{km}$ horizontal resolution for a region covering Europe. This presents a NWP model that runs at a roughly similar resolution as the maps produced in this thesis and by CHIRTS. A second source are temperature fields from the ICON-EU and ICON-D2 [110] models, operated by the *Deutsche Wetterdienst*. These represent regional models covering Europe and Germany, Switzerland, Austria and Benelux countries, respectively. Whereas ICON-EU operates at roughly $6.25\,\mathrm{km}$ horizontal resolution (approx. $7\,\mathrm{km}$ when projected to a regular latitude-longitude grid), ICON-D2 operates at $2.2\,\mathrm{km}$ horizontal resolution, thus acting as a very-high resolution model, which will subsequently be used as the ground-truth prediction and a reference of how spectral patterns are expected to look. Finally, daily maximum temperature produced by this model and data from CHIRTS serve as non-NWP-simulated comparison data. Results from calculating normalized RAPSD spectra for all data sets can be found in Figure 3.15. Normalization is applied as the different spectra originate from source images of different resolution, thus yielding different power amplitude, which makes them hard to compare visually.

It is important to note that while data from CERRA, CHIRTS and the model developed in this thesis represents the 16^{th} of July, 2013, data from ICON-EU and ICON-D2 represents the 26^{th} of March, 2024. This discrepancy is due to availability issues and no common time frame being available for all data sets. While this may make any comparison seem moot, it is important to consider that spectral behavior should be largely independent of signal amplitude and largely driven by the underlying mathematical formulation of

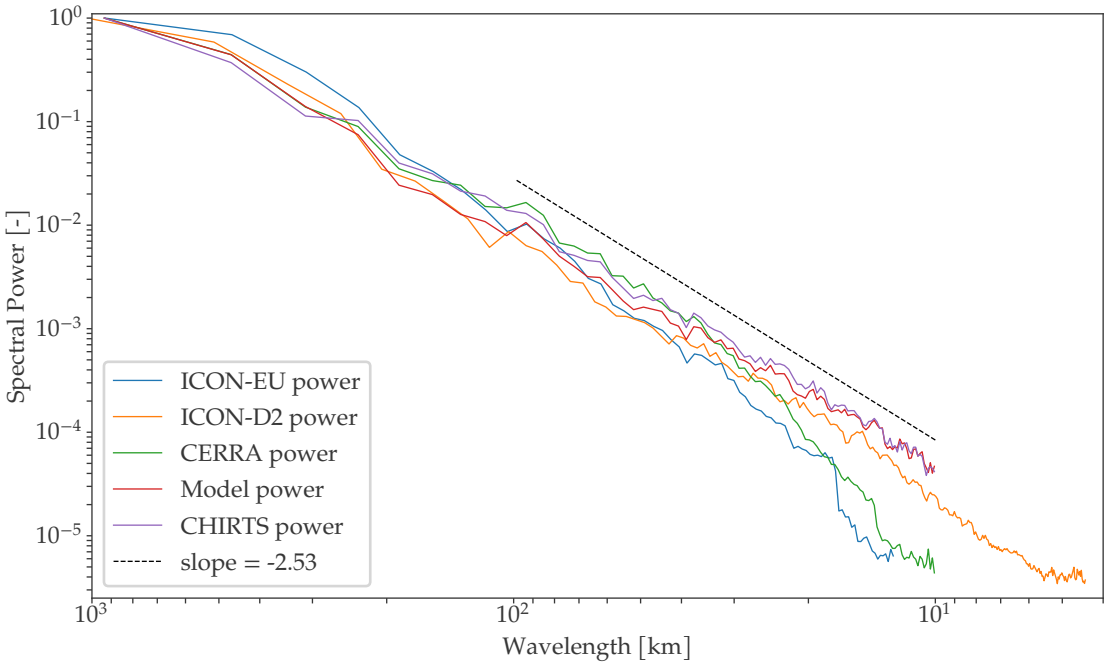

Figure 3.15: Normalized Radially-averaged Power Spectral Density plot for daily maximum temperature, as provided by the model, CHIRTS-daily, CERRA, ICON-EU and ICON-D2.

physical equations, therefore being largely independent of the model state. Even if an exact comparison of how well a model tracks the curve produced by another model is not possible, trends exhibited by the curves are similar as the underlying processes are the same.

It can be seen in Figure 3.15 that for wavelengths larger than$500\,\mathrm{km}$, all models follow a similar trend, but tend to drift apart at wavelengths below $150\,\mathrm{km}$. Especially below wavelengths of $40\,\mathrm{km}$, the model and CHIRTS closely match ICON-D2, whereas CERRA and ICON-EU show a drop in power at wavelengths roughly twice their horizontal resolution. Generally, both CERRA and ICON-EU tend to drift away from ICON-D2 below approx. $30\,\mathrm{km}$. The non-NWP models both show a slightly smaller slope compared to ICON-D2 below wavelengths of roughly $50\,\mathrm{km}$.

3.2.6 Derivation of Canopy Urban Heat Island Anomaly

To show one possible application of the model, CUHI anomalies are derived for selected cities. The calculation follows the logic detailed in Section 3.1.5. Results for Milan and Paris for the European heat wave in August 2003 can be seen in Figure 3.16.

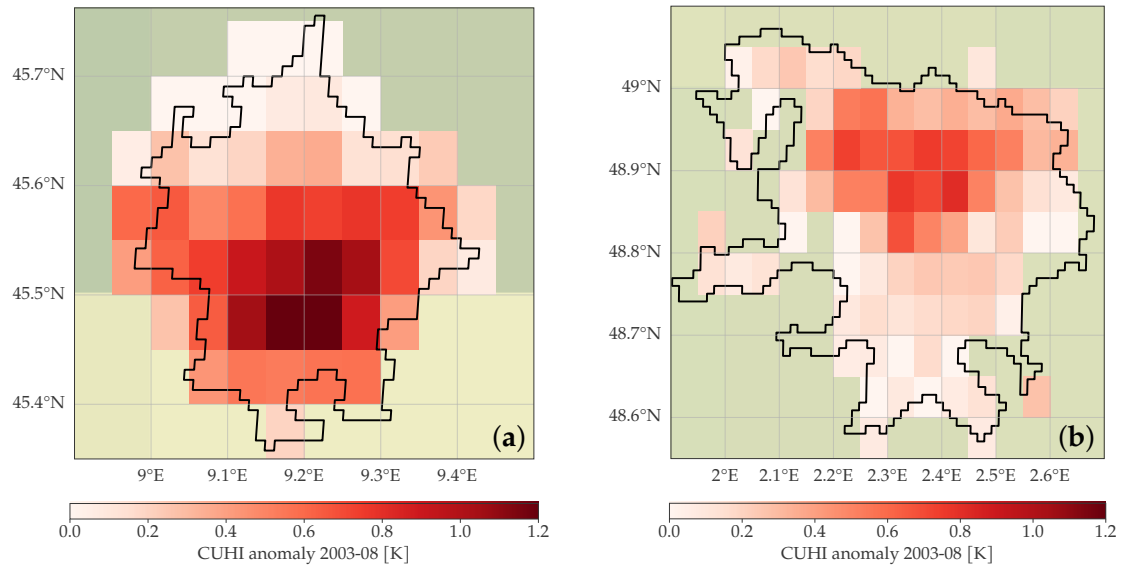

Figure 3.16: Monthly average Canopy Urban Heat Island anomaly derived during the European heat wave in August 2003. Black outlines show city boundaries derived from GHS UCDB. **(a)** shows CUHI anomaly for Milan while **(b)** shows CUHI anomaly for Paris.

Both cities show a clear signal around the city center, which then gradually decreases towards the city boundary. The signal of Milan peaks around $1.2\,\mathrm{K}$, whereas Paris reaches around $0.85\,\mathrm{K}$ for the monthly average CUHI anomaly in August. Due to different locations of the city center within the cities, the location of the maximum differs. The maps show slight variation of the intensity near the location of the maxima, with no clear radial trend.

3.2.7 Model Performance without Remote Sensing Observations

While previous sections evaluated a model based on both remotely-sensed and reanalysis-based fields, the initial prototype consisted of a model purely based on data from reanalysis, following similar works by [111, 112, 113]. This has multiple benefits, such as firstly allowing much longer time horizons to be produced, as the availability of reanalysis fields is the only limiting factor and secondly being able to correct fields into the future. The approach used for that model is identical to the one presented in previous sections, with the only major difference being a different set of training features and that instead of training to predict the corrected temperature directly, the model is trained to predict the error as that showed much-improved feature utilization. A selection of 22 features is used, which can be found in Table E.1. Instead of a Random Forest, a Histogram Gradient Boosting

regressor (HGB) is used, as model selection shows this algorithms performs best on the problem. Land cover is treated as a categorical feature internally inside the HGB, using the implementation provided by `scikit-learn`. Results of assessing the performance achieved by the model through summary statistics for January and June of 2015 can be found in Table 3.6.

Table 3.6: Results of evaluating the model based only on reanalysis features. Error metrics are calculated as monthly means.

Source	Time	RMSE [K]	MAE [K]	Bias [K]	R^2
ERA5	2015/01	3.78	2.82	0.227	0.83
Model	2015/01	3.31	2.51	−0.08	0.87
ERA5	2015/06	2.59	1.97	0.59	0.79
Model	2015/06	2.3	1.71	0.02	0.84

The change in evaluation strategy of testing against unseen stations in an unseen year instead of unseen stations during the training period is due to also wanting to test the model's ability to correct fields at unseen times. Testing only at unseen locations in space, but during the training period, yielded similar results. As can be seen from Table 3.6, the RMSE improvement over reanalysis fields is approximately 12.5 % in January and approx. 11 % in June.

However, this approach exhibits several shortcomings, such as the the inability to reproduce sub-gridscale features necessary to accurately portray cities and unphysical visual patterns when inferencing the model to calculate corrected fields. Both of these effects are shown in Figure 3.17.

As Figure 3.17 shows, no clear pattern of UHI, i.e. an increase of temperature near the city center and a decrease towards the boundary, is visible in the top map, but rather some sort of pattern noise is present. The bottom plot shows the visual artefacts mentioned above. When looking at the white horizontal dashed line, an abrupt jump is visible, where values below the line tend to be slightly larger than those above. For the vertical line, the same pattern is visible where pixels to the left show larger values compared to those on the right. The region shown here is just one example, with other regions exhibiting exactly the same behavior. This further varies on a day-to-day basis, with some days not showing the same effect at the same location.

3.3 Discussion

In this thesis, a machine learning model that downscales and bias-corrects daily maximum air temperature from a coarse horizontal resolution of 25 km, such as produced by reanalyses like ERA5, to a finer grid spacing of 5 km is detailed. This task is achieved by combining various remote-sensing and surface-level observations with the coarse-resolution NWP inputs. The results are assessed with regards to their errors against observations

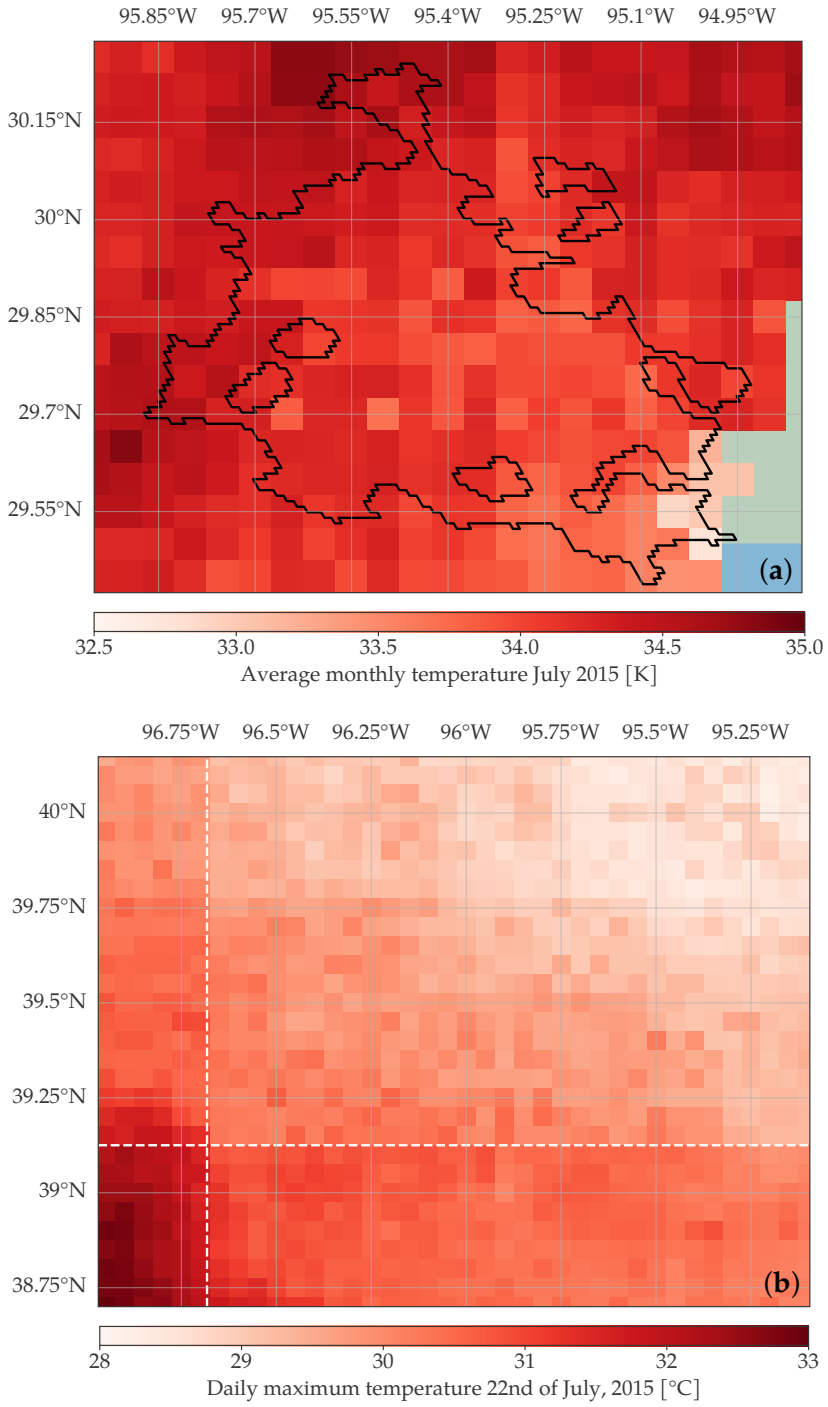

Figure 3.17: **(a)** Monthly average of daily maximum temperature in Houston for July 2015. Black lines indicate the city outline from GHS UCDB, **(b)** Daily maximum temperature for the 22^{nd} of July, 2015 near Kansas City. White dashed lines indicate markers of abrupt jumps in the fields.

and whether they faithfully reproduce physical behavior at finer scales. The methodology first interpolates all training features to the location of in-situ observations while properly considering differences in elevation between the gridded input fields and point observations through an ELR correction in order to reduce artificial bias during training. In a subsequent step, a Random Forest is trained against these in-situ observations as targets and inferenced quasi-globally to produce a time series of daily maximum air temperature fields between 2003 and 2020. It is shown in previous sections that the model is successfully able to remove almost all bias (residual of $0.04\,\mathrm{K}$) from the input ERA5 field while also managing to improve the RMSE by $11.8\,\%$ globally when validated against an independent holdout set of observations. ERA5 daily maximum temperature and gap-filled MODIS land surface temperature are found to be the most important features for its predictions. Further assessment shows that improvements are not homogeneous in space, with the Northern Hemisphere seeing smaller performance increases compared to the Southern Hemisphere after model application and that topographically complex regions show higher errors on average. In a second step, the ability of the model to add sub-gridscale detail to the reanalysis maps is assessed by comparing its results to high-resolution UrbClim simulations. This comparison yielded mixed results, where the fields produced by the model showed a temporally- and spatially-varying bias that differed among the example cities Milan, Barcelona and Paris and indicated systematic error of temperature fields in urban areas. Finally, a new method for deriving Urban Heat Islands is tested on these example cities and showed physically-consistent results.

Due to the quasi-global scale of the produced maps, prior work of similar spatial extent is quite limited. The work by Zheng *et al.* [75] provides the closest reference, as a similar strategy for deriving daily maximum temperature fields for Eurasia is used. When looking at the subset of approx. 800 stations from the holdout set of this thesis, located in the region that [75] produced data for, and using those to derive error metrics without applying any form of ELR correction to remain consistent with [75], the model achieves an RMSE of $1.78\,\mathrm{K}$ ($2.177\,\mathrm{K}$), a Bias of $-0.04\,\mathrm{K}$ ($0.095\,\mathrm{K}$) and an R^2 of 0.99 (0.98), where values in parentheses indicate scores achieved by [75] and are taken from Table 7. While both models produce essentially bias-free results, the comparison favors the model developed in this thesis, as it achieves a lower RMSE, even though it only uses a single Random Forest for the entire globe instead of multiple regional HGB regressors, as done in [75]. However, when repeating the same error assessment above using ERA5 and CHIRTS, lower RMSEs of $2.73\,\mathrm{K}$ ($3.661\,\mathrm{K}$) and $2.33\,\mathrm{K}$ ($2.590\,\mathrm{K}$) are achieved compared to [75] in parentheses. This could either be caused by an adverse set of stations used in their validation, which just generally yields worse errors, but may also be related to the additional QA implemented in this thesis for the GSOD stations that successfully removed outliers, causing their numbers to be inflated. Overall, the findings contradict their statement that multiple regional models are necessary to achieve good performance, but may in part be explained by the additional ELR correction done in this thesis. This greatly reduces artificial bias in the training procedure due to elevation differences between observations and reanalysis cells, in turn allowing the model to better focus on learning the underlying patterns in the data.

The usage of multiple models is evaluated in Table 3.4, but suitability of the methodology to assess whether more than one model benefits prediction quality remains to be discussed. This is because the same hyper-parameters are used for all models, even though the size of the training set varies between regions. While this would be perfectly fine for a MLP that has a fixed number of trainable parameters, for Random Forests, the possibility exists that the number of trained parameters for each RF varies between regions, as one region may produce a RF where the average depth of its Decision Trees is $h/2$, whereas another may produce trees with an average depth of h. Since RFs can have a different numbers of trained parameters, Table 3.4 compares errors produced by entirely different models instead of one fixed model trained on different data.

While the validation results around the globe indicate good model performance, the comparatively worse errors of the model in the US are likely tied to the intricate details of the origination of GSOD and GHCNd data, affecting both the summary statistics as shown in Figure 3.8, but also the bias correction skill, as shown in Figure 3.10. As identified in Section 2.1.2, some in-situ observations aggregated in GSOD and GHCNd are operated by volunteers in the COOP network. Due to its voluntary nature, it does not enforce a strict and standardized time of reporting on measurements, therefore possibly containing shifted time series when the local time of reporting is improperly translated into a global time during merging. This effect was also found in [26] and [27], where the latter implemented a routine that compared the RMSE of station time series against ERA5 at the same time and shifted one day forward or backward and then selected the shift with the lowest error at each station for further computations. Such a step is not done in this thesis, due to both a lack of information about aspects such as the time of recording at many stations around the globe and a worry that this shift may introduce artificial trends, which likely cause decreased performance both during training and validation. The comparatively larger errors in the US are probably further amplified by a lack of surface stations available for assimilation, as found in [27]. As only WMO stations are assimilated in ERA5 and their density in the US is much lower compared to for example Europe, as shown in Figure 3.2, the reanalysis fields are likely to be less well-constrained to observations and thus show inherently larger errors, which also manifest in the resuls after model application.

The analysis of the RAPSDs of various temperature fields produced by different models yield more interesting results. While the comparison is limited by the lack of available data in a common time frame, it is nevertheless the first example that investigates spectral behavior of non-NWP-based temperature maps to the author's knowledge. In general, temperature changes gradually in space and is driven by large-scale atmospheric circulation patterns. Consequently, when looking at the spectral behavior of temperature fields across continental scales, the amplitude of small-scale fluctuations, representing for example a drop in temperature due to fast-changing topography near mountains, is expected to be lower than that of large-scale patterns corresponding to the atmospheric circulation. While temperature fields from both CHIRTS and this thesis generate a spectrum with a roughly constantly-decreasing slope in the RAPSD plot in Figure 3.15, numerical simu-

lations from CERRA and ICON-EU produce a drop at around $30\,\mathrm{km}$ wavelength with a more pronounced decrease. The most likely explanation for this behavior lies in the effective resolution of numerical models being several times their horizontal resolution [114]. Although no studies exist for the CERRA and ICON models, Abdalla *et al.* [115] found that the IFS, which ERA5 is based on, has an approximate atmospheric resolution of $8\Delta x$, whereas Vosper *et al.* [116] showed that the UK Met Office model poorly resolved effects at wavelengths smaller $8\Delta x$ to $10\Delta x$. To reiterate, Δx refers to the horizontal model resolution, which is $5.5\,\mathrm{km}$ for CERRA and $6.25\,\mathrm{km}$ for ICON-EU. These effects mostly originate from numerical discretization causing overly-strong smoothing of fields and sub-gridscale discretizations of properties such as orographic wave drag not accounting for all unresolved scales properly [115]. When looking at the ICON-D2 model, which has a much higher resolution of $\Delta x = 2.2\,\mathrm{km}$, no steep drops at short wavelengths are present, thus suggesting that the behavior seen in the CERRA and ICON-EU models are actually errors due to their numerical discretization instead of physical behavior. Consequently, the observed spectral behavior of CHIRTS and the model developed in this thesis seems reasonable, as they closely match ICON-D2, although showing slightly larger power at small wavelengths. This can be explained when considering that even ICON-D2 does not perfectly resolve all small-scale processes, as it is subject to the same numerical limitations as ICON-EU and the other NWP models, although the onset is much more gradual due to the higher horizontal resolution. Therefore, even ICON-D2 shows a small drop across the spectrum towards smaller wavelengths, consistent with findings by Skamarock, who examined turbulent kinetic energy spectra produced by the WRF model at different horizontal resolutions and found identical behavior [114]. Overall, both CHIRTS and this thesis will exhibit spectral power very close to the spectral properties of the underlying elevation model through its explicit coupling in the downscaling process of the temperature fields. These models therefore produce spectral content down to wavelengths as low as $2\Delta x$, which NWP models tend to not achieve in practice [114].

Lastly, it is important to shortly touch upon the calculation of the CUHI amplitude from gridded data. The maps shown in Figure 3.16 exhibit plausible patterns that match the expectation of a larger CUHI amplitude near the city center, which decays closer to the city boundary. However, considering the results of Section 3.2.4, CUHI anomalies should be considered as a rough approximation of the truth at best, until the underlying model manages to better approximate UrbClim simulations in cities. Still, the application of the algorithm suggested in [80] at a $5\,\mathrm{km}$ horizontal resolution provides valuable information about its applicability at coarser scales. Furthermore, the improvement to explicitly consider elevation differences seem like an important addition due to the comparatively low CUHI amplitude that often only reaches between 1–2K [93]. Considering the standard lapse rate $\Gamma = -0.0065\,\mathrm{K\,m^{-1}}$, a mere $150\,\mathrm{m}$ elevation difference between city and rural cells is enough to introduce an artificial bias of similar amplitude to the underlying signal.

Overall, this thesis is among the first, if not the first, to estimate global daily maximum temperature using a machine learning model that combines reanalysis fields with satellite and in-situ observations to provide high-resolution maps with $5\,\mathrm{km}$ horizontal grid

spacing. Compared to previous literature using machine-learning-based approaches for downscaling, it utilizes a lapse rate correction to account for differences in elevation between reanalaysis fields and point observations, therefore reducing artificial bias during training and improving the model performance. In contrast to previous works that utilized multiple machine learning models for different regions, this thesis also showed that a single model, when combined with the previous pre-processing step and in the case of daily maximum temperature, is sufficient to cover the entire globe, reducing both computational requirements while also improving interpretability, as only a single model has to be characterized, instead of multiple regional ones. Although similar data sets, such as CHIRTS, have existed before, the error assessment shows that the model produces consistently lower errors, allowing for further improvement of impact and risk assessments around the globe.

However, a few shortcomings of the model can be identified. First of all, more work is necessary to improve the representation of urban areas. Whether this may be through the inclusion of additional features that better represent urban micro-climatic conditions, such as shown in [117], a change in temporal and/or spatial resolution or a combination of these aspects is left for further research. As identified in [118], a factor that will have to be accounted for here is the lack of weather stations inside cities. When looking at the training data, only $6.4\,\%$ of measurements are located in areas with an urban fraction larger than $60\,\%$, limiting the ability of the model to recover urban air temperature profiles due to a lack of data. Still, even if the derivation of effects like Urban Heat Islands from model results in their current state may not be advisable, the higher resolution provided by this thesis along with the generally much improved representation of temperature profiles in cities over ERA5 still provides value for various impact assessments and analysis. Secondly, care needs to be taken when using this data for trend analysis. While trends are likely to closely follow the raw ERA5 reanalysis, more investigation is necessary here, as the training targets are not homogenized beforehand and may contain erroneous trends that the model learned and subsequently imposed on its outputs. Even though these effects may be subtle, until further validation is conducted, a similar warning as in [20, 21] is issued to not rely on trends derived from these maps for any climatic analysis. This point is also related to some practical limitations of the model, foremost its limited temporal length from 2003 and 2020 due to the availability of data from [33]. Considering that the current WMO Climatological Reference Period ranges from 1991 and 2020, the lack of data in the first twelve years is unfortunate. While land surface temperature series with much longer time spans exist, such as [119, 120], those pose their own issues, such as being derived from multiple different sensors onboard different satellites that have experienced orbital drift to different degrees, thus possibly causing inconsistencies in the data and furthermore not being available in a gap-filled version. Consequently, an extensive pre-processing step and analysis is necessary if they were to be used as a drop-in replacement in the framework developed in this thesis. However, the general approach remains applicable and when combined with a proper homogenization of stations lays the groundwork for a totally new way of deriving gridded temperature fields. Lastly, a lot of predictive skill in the model comes from

exploitation of the relation between temperature and elevation. While similar relations exist for different variables (compare methods in [63] for examples), the method presented in this thesis in its current form only works for temperature. Adaptation of this thesis's approach to quantities like incoming solar radiation should be rather easy, but especially variables like precipitation and wind speed—although influenced by topography—do not show such simple relationships and therefore likely require more work to achieve good performance under a similar approach.

4 Conclusion

In this thesis, different machine learning algorithms were evaluated with respect to their ability to bias-correct and downscale daily maximum temperature from a native $25\,\mathrm{km}$ horizontal grid, as produced by the ERA5 reanalysis, to a $5\,\mathrm{km}$ grid and whether these models are able to add sub-gridscale detail in a physically-consistent way. After a rigorous feature and model selection procedure, a Random Forest was trained on a combination of global observational and reanalysis data from 2003 and 2020 and used to produce an improved daily maximum temperature field, which was subsequently evaluated against a hold-out set of observations in order to assess its performance. The comparison turned out favorably for the model when measuring errors at a global scale, as it successfully bias-corrected and downscaled input fields and even provided a slight improvement of around $10\,\%$ to $30\,\%$ in RMSE and MAE, depending on the region. At city-scale, however, more work is necessary, as was shown by a comparison with results from the high-resolution city-scale weather model UrbClim. While the daily maximum temperature produced by this thesis exhibited low errors when compared to UrbClim data, it showed considerable bias with both spatial and temporal structure that varies by city, indicating that the model is not yet able to properly reproduce these fine-scale urban temperature distributions. Finally, calculation of Urban Heat Island anomalies, representing one use case that benefits from the increased horizontal resolution, was demonstrated. Although the results for Paris and Milan looked sensible, further validation is required due to the considerable bias contained in the base temperature layer.

While downscaling of reanalysis fields, such as those provided by ERA5, is not a new idea, the approach detailed in this thesis is one of the first, if not the first, to use a machine learning approach at a global scale for such a long time period. One of the great advantages of this method is its ability to learn corrections at point locations, but make predictions at any location around the globe. Furthermore, no interpolation of point measurements is necessary a-priori, but instead, the model is left to figure out a correlation between inputs, relieving practitioners from a lot of work. Compared to previous works, this thesis also highlights the importance of proper pre-processing of the inputs. By applying a lapse rate correction to observations prior to training, artificial bias is removed, allowing the model to learn the underlying error patterns. This makes a single global model sufficient, compared to multiple regional models used in prior works, while achieving lower errors. Even though the approach works well at large scales, this research once more demonstrates that machine learning is not a silver bullet. Processes, such as those governing urban micro-meteorology, that happen way below resolved scales of the input features in both space and time, demand too much of such models and likely require more data to be

resolved accurately. Still, when assessing the city-scale results purely based on their error metrics, results of this thesis are remarkably close to high-resolution simulations, so the model presents a promising first step towards the full recovery of small-scale features in meteorological fields.

These findings pose new and interesting questions and provide multiple directions for future research. First of all, the consequences of training a model on higher-resolution data in space and time require more investigation. Hourly data is available for all variables, even remotely-sensed land surface temperature, although this puts an even tighter limit on the length of the time series that can be produced. Investigating whether just decreasing the time step yields sufficient amounts of new information to improve the reproduction of small-scale features might even allow new insights into what sort of information and training features could be beneficial to coarse-scale models, such as the one proposed in this thesis. At the same time, such a model could also be used as a surrogate in NWP systems, as the physics governing near-surface air temperature rely on many approximations about properties of the surface, which may be better represented by a machine learning model compared to existing parametrizations. A further topic worth exploring is generalizability of this approach with respect to different variables. While only daily maximum temperature was explored in this thesis, one might want to look into radiation variables, of which accurate knowledge is mandatory for solar power assessments for example. Furthermore, properties such as wind and precipitation may be interesting to investigate and also to compare the skill of the methodology proposed here to more elaborate method used generally for these applications.

Bibliography

[1] V. Bjerknes, "Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physik," *Meteorologische Zeitschrift*, vol. 21, pp. 1–7, 1904.

[2] R. A. Pielke Sr, G. E. Liston, J. L. Eastman, L. Lu, and M. Coughenour, "Seasonal weather prediction as an initial value problem," *Journal of Geophysical Research: Atmospheres*, vol. 104, no. D16, pp. 19463–19479, 1999.

[3] J. J. Tribbia and D. P. Baumhefner, "Scale interactions and atmospheric predictability: An updated perspective," *Monthly weather review*, vol. 132, no. 3, pp. 703–713, 2004.

[4] M. Göber, E. Zsoter, and D. S. Richardson, "Could a perfect model ever satisfy a naïve forecaster? on grid box mean versus point verification," *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, vol. 15, no. 3, pp. 359–365, 2008.

[5] T. Janjić, N. Bormann, M. Bocquet, J. Carton, S. E. Cohn, S. L. Dance, S. N. Losa, N. K. Nichols, R. Potthast, J. A. Waller, *et al.*, "On the representation error in data assimilation," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 713, pp. 1257–1278, 2018.

[6] A. McNally and P. Watts, "A cloud detection algorithm for high-spectral-resolution infrared sounders," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 129, no. 595, pp. 3411–3423, 2003.

[7] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[8] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y. Huang, W. Wang, and J. G. Powers, "A Description of the Advanced Research WRF Version 3," *NCAR Technical Note NCAR/TN-475+STR*, vol. 475, p. 1, 2008.

[9] N. P. Wedi, I. Polichtchouk, P. Dueben, V. G. Anantharaj, P. Bauer, S. Boussetta, P. Browne, W. Deconinck, W. Gaudin, I. Hadade, *et al.*, "A baseline for global weather and climate simulations at 1 km resolution," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002192, 2020.

[10] A. M. Amorim, A. B. Gonçalves, L. M. Nunes, and A. J. Sousa, "Optimizing the location of weather monitoring stations using estimation uncertainty," *International Journal of Climatology*, vol. 32, no. 6, pp. 941–952, 2012.

[11] E. Pardo-Igúzquiza and P. A. Dowd, "Multiple indicator cokriging with application to optimal sampling for environmental monitoring," *Computers & Geosciences*, vol. 31, no. 1, pp. 1–13, 2005.

[12] E. Pardo-Igúzquiza, "Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing," *Journal of hydrology*, vol. 210, no. 1-4, pp. 206–220, 1998.

[13] M. J. Menne, I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston, "An overview of the global historical climatology network-daily database," *Journal of atmospheric and oceanic technology*, vol. 29, no. 7, pp. 897–910, 2012.

[14] A. Smith, N. Lott, and R. Vose, "The integrated surface database: Recent developments and partnerships," *Bulletin of the American Meteorological Society*, vol. 92, no. 6, pp. 704–708, 2011.

[15] National Centers for Environmental Information (NCEI), "Global surface summary of the day - gsod." https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc%3AC00516, 2023. [Accessed May 15, 2024].

[16] T. R. Karl and C. N. Williams Jr, "An approach to adjusting climatological time series for discontinuous inhomogeneities," *Journal of Applied Meteorology and Climatology*, vol. 26, no. 12, pp. 1744–1763, 1987.

[17] T. C. Peterson, D. R. Easterling, T. R. Karl, P. Groisman, N. Nicholls, N. Plummer, S. Torok, I. Auer, R. Boehm, D. Gullett, *et al.*, "Homogeneity adjustments of in situ atmospheric climate data: a review," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 18, no. 13, pp. 1493–1517, 1998.

[18] P. Jones, S. Raper, B. Cherry, C. Goodess, and T. Wigley, "A grid point surface air temperature data set for the southern hemisphere, 1851-1984," tech. rep., US Dept of Energy, Carbon Dioxide Research Division, 1986.

[19] J. B. Wijngaard, A. Klein Tank, and G. Können, "Homogeneity of 20th century european daily temperature and precipitation series," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 23, no. 6, pp. 679–692, 2003.

[20] N. Hofstra, M. Haylock, M. New, and P. D. Jones, "Testing e-obs european high-resolution gridded data set of daily precipitation and surface temperature," *Journal of Geophysical Research: Atmospheres*, vol. 114, no. D21, 2009.

[21] R. C. Cornes, G. van der Schrier, E. J. van den Besselaar, and P. D. Jones, "An ensemble version of the e-obs temperature and precipitation data sets," *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 17, pp. 9391–9409, 2018.

[22] C. Daly, "Guidelines for assessing the suitability of spatial climate data sets," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 26, no. 6, pp. 707–721, 2006.

[23] I. Durre, M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose, "Comprehensive automated quality assurance of daily surface observations," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 8, pp. 1615–1633, 2010.

[24] National Research Council and Division on Engineering and Physical Sciences and Commission on Engineering and Technical Systems and National Weather Service Modernization Committee, *Future of the National Weather Service Cooperative Observer Network*. National Academies Press, 1998.

[25] C. A. Davey and R. A. Pielke Sr, "Microclimate exposures of surface-based weather stations: Implications for the assessment of long-term temperature trends," *Bulletin of the American Meteorological Society*, vol. 86, no. 4, pp. 497–504, 2005.

[26] H. Hashimoto, W. Wang, F. S. Melton, A. L. Moreno, S. Ganguly, A. R. Michaelis, and R. R. Nemani, "High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous united states," *International Journal of Climatology*, vol. 39, no. 6, pp. 2964–2983, 2019.

[27] F. M. Lopes, E. Dutra, and S. Boussetta, "Evaluation of daily temperature extremes in the ecmwf operational weather forecasts and era5 reanalysis," *Atmosphere*, vol. 15, no. 1, 2024.

[28] B. Thies and J. Bendix, "Satellite based remote sensing of weather and climate: recent achievements and future perspectives," *Meteorological Applications*, vol. 18, no. 3, pp. 262–295, 2011.

[29] P. Edwards and D. Pawlak, "Metop: The space segment for eumetsat's polar system," *ESA bulletin*, pp. 7–18, 2000.

[30] J. R. Piepmeier, P. Focardi, K. A. Horgan, J. Knuble, N. Ehsan, J. Lucey, C. Brambora, P. R. Brown, P. J. Hoffman, R. T. French, *et al.*, "Smap l-band microwave radiometer: Instrument design and first year on orbit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 4, pp. 1954–1966, 2017.

[31] A. J. Geer, K. Lonitz, P. Weston, M. Kazumori, K. Okamoto, Y. Zhu, E. H. Liu, A. Collard, W. Bell, S. Migliorini, P. Chambon, N. Fourrié, M.-J. Kim, C. Köpken-Watts,

and C. Schraff, "All-sky satellite data assimilation at operational weather forecasting centres," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 713, pp. 1191–1217, 2018.

[32] S. Favrichon, C. Prigent, C. Jimenez, and F. Aires, "Detecting cloud contamination in passive microwave satellite measurements over land," *Atmospheric Measurement Techniques*, vol. 12, no. 3, pp. 1531–1543, 2019.

[33] T. Zhang, Y. Zhou, Z. Zhu, X. Li, and G. R. Asrar, "A global seamless 1 km resolution daily land surface temperature dataset (2003–2020)," *Earth System Science Data Discussions*, vol. 2021, pp. 1–16, 2021.

[34] D. J. Weiss, P. M. Atkinson, S. Bhatt, B. Mappin, S. I. Hay, and P. W. Gething, "An effective approach for gap-filling continental scale remotely sensed time-series," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 106–118, 2014.

[35] C. Zeng, H. Shen, M. Zhong, L. Zhang, and P. Wu, "Reconstructing modis lst based on multitemporal classification and robust regression," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 512–516, 2014.

[36] S. L. Ermida, I. F. Trigo, C. C. DaCamara, C. Jiménez, and C. Prigent, "Quantifying the clear-sky bias of satellite land surface temperature using microwave-based estimates," *Journal of Geophysical Research: Atmospheres*, vol. 124, no. 2, pp. 844–857, 2019.

[37] K. Gallo and P. Krishnan, "Evaluation of the bias in the use of clear-sky compared with all-sky observations of monthly and annual daytime land surface temperature," *Journal of Applied Meteorology and Climatology*, vol. 61, no. 10, pp. 1485–1495, 2022.

[38] A. Jia, S. Liang, D. Wang, L. Ma, Z. Wang, and S. Xu, "Global hourly, 5 km, all-sky land surface temperature data from 2011 to 2021 based on integrating geostationary and polar-orbiting satellite data," *Earth System Science Data Discussions*, vol. 2022, pp. 1–35, 2022.

[39] L. S. Gandin, "Objective analysis of meteorological fields," *Israel program for scientific translations*, vol. 242, 1963.

[40] E. Kalnay, *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2002.

[41] F.-X. Le Dimet and O. Talagrand, "Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects," *Tellus A: Dynamic Meteorology and Oceanography*, vol. 38, no. 2, pp. 97–110, 1986.

[42] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[43] A. J. Simmons and A. Hollingsworth, "Some aspects of the improvement in skill of numerical weather prediction," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 128, no. 580, pp. 647–677, 2002.

[44] P. Bauer, A. J. Geer, P. Lopez, and D. Salmond, "Direct 4d-var assimilation of all-sky radiances. part i: Implementation," *Quarterly Journal of the Royal Meteorological Society*, vol. 136, no. 652, pp. 1868–1885, 2010.

[45] F. Johannsen, S. Ermida, J. P. Martins, I. F. Trigo, M. Nogueira, and E. Dutra, "Cold bias of era5 summertime daily maximum land surface temperature over iberian peninsula," *Remote Sensing*, vol. 11, no. 21, p. 2570, 2019.

[46] D. Choudhury, F. Ji, N. Nishant, and G. Di Virgilio, "Evaluation of era5-simulated temperature and its extremes for australia," *Atmosphere*, vol. 14, no. 6, p. 913, 2023.

[47] K. Velikou, G. Lazoglou, K. Tolika, and C. Anagnostopoulou, "Reliability of the era5 in replicating mean and extreme temperatures across europe," *Water*, vol. 14, no. 4, 2022.

[48] U. Von Luxburg and B. Schölkopf, "Statistical learning theory: Models, concepts, and results," in *Handbook of the History of Logic*, vol. 10, pp. 651–706, Elsevier, 2011.

[49] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and regression trees*. New York: CRC Press, 1984.

[50] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[51] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[52] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[53] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[54] A. M. Rizwan, L. Y. Dennis, and L. Chunho, "A review on the generation, determination and mitigation of urban heat island," *Journal of environmental sciences*, vol. 20, no. 1, pp. 120–128, 2008.

[55] H. Du, W. Zhan, Z. Liu, J. Li, L. Li, J. Lai, S. Miao, F. Huang, C. Wang, C. Wang, *et al.*, "Simultaneous investigation of surface and canopy urban heat islands over global cities," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 67–83, 2021.

[56] C. Heaviside, S. Vardoulakis, and X.-M. Cai, "Attribution of mortality to the urban heat island during heatwaves in the west midlands, uk," *Environmental health*, vol. 15, pp. 49–59, 2016.

[57] W. B. Goggins, E. Y. Chan, E. Ng, C. Ren, and L. Chen, "Effect modification of the association between short-term meteorological factors and mortality by urban heat islands in hong kong," *PLOS ONE*, vol. 7, no. 6, pp. 1–6, 2012.

[58] A. Smargiassi, M. S. Goldberg, C. Plante, M. Fournier, Y. Baudouin, and T. Kosatsky, "Variation of daily warm season mortality as a function of micro-urban heat islands," *Journal of Epidemiology & Community Health*, vol. 63, no. 8, pp. 659–664, 2009.

[59] J. A. Patz, D. Campbell-Lendrum, T. Holloway, and J. A. Foley, "Impact of regional climate change on human health," *Nature*, vol. 438, no. 7066, pp. 310–317, 2005.

[60] D. Li, E. Bou-Zeid, and M. Oppenheimer, "The effectiveness of cool and green roofs as urban heat island mitigation strategies," *Environmental Research Letters*, vol. 9, no. 5, p. 055002, 2014.

[61] T. Susca, S. R. Gaffin, and G. Dell'Osso, "Positive effects of vegetation: Urban heat island and green roofs," *Environmental pollution*, vol. 159, no. 8-9, pp. 2119–2126, 2011.

[62] E. Eumorfopoulou and D. Aravantinos, "The contribution of a planted roof to the thermal protection of buildings in greece," *Energy and Buildings*, vol. 27, no. 1, pp. 29–36, 1998.

[63] D. N. Karger, S. Lange, C. Hari, C. P. Reyer, O. Conrad, N. E. Zimmermann, and K. Frieler, "Chelsa-w5e5: Daily 1 km meteorological forcing data for climate impact studies," *Earth System Science Data Discussions*, vol. 2022, pp. 1–28, 2022.

[64] H. E. Beck, A. I. Van Dijk, P. R. Larraondo, T. R. McVicar, M. Pan, E. Dutra, and D. G. Miralles, "Mswx: global 3-hourly 0.1 bias-corrected meteorological data including near-real-time updates and forecast ensembles," *Bulletin of the American Meteorological Society*, vol. 103, no. 3, pp. E710–E732, 2022.

[65] G. Tang, M. P. Clark, and S. M. Papalexiou, "Em-earth: the ensemble meteorological dataset for planet earth," *Bulletin of the American Meteorological Society*, vol. 103, no. 4, pp. E996–E1018, 2022.

[66] R. Dodson and D. Marks, "Daily air temperature interpolated at high spatial resolution over a large mountainous region," *Climate research*, vol. 8, no. 1, pp. 1–20, 1997.

[67] K. E. Kunkel, "Simple procedures for extrapolation of humidity variables in the mountainous western united states," *Journal of Climate*, vol. 2, no. 7, pp. 656–669, 1989.

[68] I. Harris, T. J. Osborn, P. Jones, and D. Lister, "Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset," *Scientific data*, vol. 7, no. 1, p. 109, 2020.

[69] D. N. Karger, O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, and M. Kessler, "Climatologies at high resolution for the earth's land surface areas," *Scientific data*, vol. 4, no. 1, pp. 1–20, 2017.

[70] M. Kilibarda, T. Hengl, G. B. Heuvelink, B. Gräler, E. Pebesma, M. Perčec Tadić, and B. Bajat, "Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution," *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 5, pp. 2294–2313, 2014.

[71] T. Zhang, Y. Zhou, K. Zhao, Z. Zhu, G. Chen, J. Hu, and L. Wang, "A global dataset of daily near-surface air temperature at 1-km resolution (2003–2020)," *Earth System Science Data Discussions*, vol. 2022, pp. 1–18, 2022.

[72] M. Kim, L. Wang, and Y. Zhou, "Spatially varying coefficient models with sign preservation of the coefficient functions," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 26, pp. 367–386, 2021.

[73] N. Huband and J. Monteith, "Radiative surface temperature and energy balance of a wheat canopy: I. comparison of radiative and aerodynamic canopy temperature," *Boundary-Layer Meteorology*, vol. 36, pp. 1–17, 1986.

[74] D. Mutiibwa, S. Strachan, and T. Albright, "Land surface temperature and surface air temperature in complex terrain," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4762–4774, 2015.

[75] M. Zheng, J. Zhang, J. Wang, S. Yang, J. Han, and T. Hassan, "Reconstruction of 0.05° all-sky daily maximum air temperature across eurasia for 2003–2018 with multi-source satellite data and machine learning models," *Atmospheric Research*, vol. 279, p. 106398, 2022.

[76] Z. B. Bouallègue, F. Cooper, M. Chantry, P. Düben, P. Bechtold, and I. Sandu, "Statistical modeling of 2-m temperature and 10-m wind speed forecast errors," *Monthly Weather Review*, vol. 151, no. 4, pp. 897–911, 2023.

[77] Columbia University, Center for International Earth Science Information Network - CIESIN - Columbia University, "Global urban heat island (uhi) data set, 2013." https://doi.org/10.7927/H4H70CRF, 2016. [Accessed May 15, 2024].

[78] T. Chakraborty and X. Lee, "Simplified urban-extent algorithm to characterize surface urban heat islands on a global scale and examine vegetation control," *International Journal of Applied Earth Observation and Geoinformation*, vol. 74, no. 2019, pp. 269–280, 2019.

[79] T. Chakraborty and X. Lee, "Yale center for earth observation (yceo) surface urban heat islands, version 4, 2003-2018." https://doi.org/10.7927/s5m5-zk14, 2023. [Accessed May 15, 2024].

[80] A. Valmassoi and J. D. Keller, "How to visualize the urban heat island in gridded datasets?," *Advances in Science and Research*, vol. 18, pp. 41–49, 2021.

[81] M. Brunetti, M. Maugeri, T. Nanni, C. Simolo, and J. Spinoni, "High-resolution temperature climatology for italy: interpolation method intercomparison," *International Journal of Climatology*, vol. 34, no. 4, pp. 1278–1296, 2014.

[82] E. Dutra, J. Muñoz-Sabater, S. Boussetta, T. Komori, S. Hirahara, and G. Balsamo, "Environmental lapse rate for high-resolution land surface downscaling: An application to era5," *Earth and Space Science*, vol. 7, no. 5, p. e2019EA000984, 2020.

[83] C. Lussana, I. Seierstad, T. Nipen, and L. Cantarello, "Spatial interpolation of two-metre temperature over norway based on the combination of numerical weather prediction ensembles and in situ observations," *Quarterly Journal of the Royal Meteorological Society*, vol. 145, no. 725, pp. 3626–3643, 2019.

[84] P. de Rosnay, P. Browne, E. de Boisséson, D. Fairbairn, Y. Hirahara, K. Ochi, D. Schepers, P. Weston, H. Zuo, M. Alonso-Balmaseda, *et al.*, "Coupled data assimilation at ecmwf: Current status, challenges and future developments," *Quarterly Journal of the Royal Meteorological Society*, vol. 148, no. 747, pp. 2672–2702, 2022.

[85] B. Brasnett, "A global analysis of snow depth for numerical weather prediction," *Journal of Applied Meteorology and Climatology*, vol. 38, no. 6, pp. 726–740, 1999.

[86] S. Grossman-Clarke, J. A. Zehnder, T. Loridan, and C. S. B. Grimmond, "Contribution of land use changes to near-surface air temperatures during recent summer extreme heat events in the phoenix metropolitan area," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 8, pp. 1649–1664, 2010.

[87] S. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 157–176, 2011.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[89] E. L. Bolager, I. Burak, C. Datar, Q. Sun, and F. Dietrich, "Sampling weights of deep neural networks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[90] L. Zampieri, G. Arduini, M. Holland, S. P. Keeley, K. Mogensen, M. D. Shupe, and S. Tietsche, "A machine learning correction model of the winter clear-sky temperature bias over the arctic sea ice in atmospheric reanalyses," *Monthly Weather Review*, vol. 151, no. 6, pp. 1443–1458, 2023.

[91] A. Florczyk, C. Corbane, M. Schiavina, M. Pesaresi, L. Maffenini, M. Melchiorri, P. Politis, F. Sabo, S. Freire, D. Ehrlich, T. Kemper, P. Tommasi, D. Airaghi, and L. Zanchetta, "Ghs urban centre database 2015, multitemporal and multidimensional attributes, r2019a." https://doi.org/10.2905/53473144-b88c-44bc-b4a3-4583ed1f547e, 2019. [Accessed May 15, 2024].

[92] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 209–226, 1977.

[93] D. Lauwaet, J. Berckmans, H. Hooyberghs, H. Wouters, G. Driesen, F. Lefebre, and K. De Ridder, "High resolution modelling of the urban heat island of 100 european cities," *Urban Climate*, vol. 54, p. 101850, 2024.

[94] L. Mentaschi, G. Duveiller, G. Zulian, C. Corbane, M. Pesaresi, J. Maes, A. Stocchino, and L. Feyen, "Global long-term mapping of surface temperature shows intensified intra-city urban heat island extremes," *Global Environmental Change*, vol. 72, p. 102441, 2022.

[95] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, pp. 1–21, 2007.

[96] A. Verdin, C. Funk, P. Peterson, M. Landsfeld, C. Tuholske, and K. Grace, "Development and validation of the chirts-daily quasi-global high-resolution daily temperature data set," *Scientific Data*, vol. 7, no. 1, p. 303, 2020.

[97] D. Yamazaki, D. Ikeshima, J. C. Neal, F. O'Loughlin, C. C. Sampson, S. Kanae, and P. D. Bates, "Merit dem: A new high-accuracy global digital elevation model and its merit to global hydrodynamic modeling," in *AGU fall meeting abstracts*, vol. 2017, pp. H12C–04, 2017.

[98] D. B. Gesch, K. L. Verdin, and S. K. Greenlee, "New land surface digital elevation model covers the earth," *Eos, Transactions American Geophysical Union*, vol. 80, no. 6, pp. 69–70, 1999.

[99] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets, "Copernicus global land cover layers—collection 2," *Remote Sensing*, vol. 12, no. 6, p. 1044, 2020.

[100] M. Buchhorn, B. Smets, L. Bertels, B. D. Roo, M. Lesiv, N.-E. Tsendbazar, M. Herold, and S. Fritz, "Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2015: Globe." https://doi.org/10.5281/zenodo.3939038, 2020. [Accessed May 15, 2024].

[101] C. Funk, P. Peterson, S. Peterson, S. Shukla, F. Davenport, J. Michaelsen, K. R. Knapp, M. Landsfeld, G. Husak, L. Harrison, *et al.*, "A high-resolution 1983–2016 t max climate data record based on infrared temperatures and stations by the climate hazard center," *Journal of Climate*, vol. 32, no. 17, pp. 5639–5658, 2019.

[102] K. De Ridder, D. Lauwaet, and B. Maiheu, "Urbclim–a fast urban boundary layer climate model," *Urban Climate*, vol. 12, pp. 21–48, 2015.

[103] H. Hooyberghs, J. Berckmans, D. Lauwaet, F. Lefebre, and K. De Ridder, "Climate variables for cities in europe from 2008 to 2017." https://doi.org/10.24381/cds.c6459d3a, 2019. [Accessed May 15, 2024].

[104] M. Mardani, N. D. Brenowitz, Y. Cohen, J. Pathak, C. Chen, C. Liu, A. Vahdat, K. Kashinath, J. Kautz, and M. Pritchard, "Generative residual diffusion modeling for km-scale atmospheric downscaling," *CoRR*, vol. abs/2309.15214, 2023.

[105] I. Price, A. Sanchez-Gonzalez, F. Alet, T. Ewalds, A. El-Kadi, J. Stott, S. Mohamed, P. W. Battaglia, R. Lam, and M. Willson, "Gencast: Diffusion-based ensemble forecasting for medium-range weather," *CoRR*, vol. abs/2312.15796, 2023.

[106] L. Li, R. Carver, I. Lopez-Gomez, F. Sha, and J. Anderson, "Generative emulation of weather forecast ensembles with diffusion models," *Science Advances*, vol. 10, no. 13, p. eadk4489, 2024.

[107] S. Pulkkinen, D. Nerini, A. A. Pérez Hortal, C. Velasco-Forero, A. Seed, U. Germann, and L. Foresti, "Pysteps: an open-source python library for probabilistic precipitation nowcasting (v1.0)," *Geoscientific Model Development*, vol. 12, no. 10, pp. 4185–4219, 2019.

[108] E. Ruzanski and V. Chandrasekar, "Scale filtering for improved nowcasting performance in a high-resolution x-band radar network," *IEEE transactions on geoscience and remote sensing*, vol. 49, no. 6, pp. 2296–2307, 2011.

[109] M. Ridal, S. Schimanke, and S. Hopsch, "Documentation of the rra system: Uerra." https://datastore.copernicus-climate.eu/documents/uerra/D322_Lot1.1.1.2_Documentation_of_the_RRA_system_UERRA_v2.pdf, 2018. [Accessed May 15, 2024].

[110] D. Reinert, F. Prill, H. Frank, M. Denhard, M. Baldauf, C. Schraff, C. Gebhardt, C. Marsigli, and G. Zängl, "Dwd database reference for the global and regional icon and icon-eps forecasting system," *DWD 2023*, 2020. [Accessed May 15, 2024].

[111] S. Rasp and S. Lerch, "Neural networks for postprocessing ensemble weather forecasts," *Monthly Weather Review*, vol. 146, no. 11, pp. 3885–3900, 2018.

[112] P. Grönquist, C. Yao, T. Ben-Nun, N. Dryden, P. Dueben, S. Li, and T. Hoefler, "Deep learning for post-processing ensemble weather forecasts," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200092, 2021.

[113] G. Stachura, Z. Ustrnul, P. Sekuła, B. Bochenek, M. Kolonko, and M. Szczech-Gajewska, "Machine learning based post-processing of model-derived near-surface air temperature–a multimodel approach," *Quarterly Journal of the Royal Meteorological Society*, vol. 150, no. 759, pp. 618–631, 2024.

[114] W. C. Skamarock, "Evaluating mesoscale nwp models using kinetic energy spectra," *Monthly weather review*, vol. 132, no. 12, pp. 3019–3032, 2004.

[115] S. Abdalla, L. Isaksen, P. Janssen, and N. Wedi, "Effective spectral resolution of ecmwf atmospheric forecast models," *ECMWF Newsletter*, vol. 137, pp. 19–22, 2013.

[116] S. Vosper, A. Brown, and S. Webster, "Orographic drag on islands in the nwp mountain grey zone," *Quarterly Journal of the Royal Meteorological Society*, vol. 142, no. 701, pp. 3128–3137, 2016.

[117] S. Koopmans, B. Heusinkveld, and G. Steeneveld, "A standardized physical equivalent temperature urban heat map at 1-m spatial resolution to facilitate climate stress tests in the netherlands," *Building and Environment*, vol. 181, p. 106984, 2020.

[118] B. Flückiger, I. Kloog, M. S. Ragettli, M. Eeftens, M. Röösli, and K. de Hoogh, "Modelling daily air temperature at a fine spatial resolution dealing with challenging meteorological phenomena and topography in switzerland," *International Journal of Climatology*, vol. 42, no. 12, pp. 6413–6428, 2022.

[119] J.-H. Li, Z.-L. Li, X. Liu, and S.-B. Duan, "A global historical twice-daily (daytime and nighttime) land surface temperature dataset produced by advanced very high resolution radiometer observations from 1981 to 2021," *Earth System Science Data*, vol. 15, no. 5, pp. 2189–2212, 2023.

[120] EUMETSAT, "Avhrr fundamental data record - release 1 - multimission." http://doi.org/10.15770/EUM_SEC_CLM_0060, 2023. [Accessed May 15, 2024].

[121] K. E. Kunkel, "Simple procedures for extrapolation of humidity variables in the mountainous western united states," *Journal of Climate*, vol. 2, no. 7, pp. 656–669, 1989.

Appendix

A Environmental Lapse Rate Factors

Table A.1 shows the results of experimenting with different environmental lapse rates Γ to see the effect they have on the accuracy of daily maximum temperature from ERA5 when compared against in-situ observations in Europe. Different values are tested, where a static value of Γ in the **Lapse Rate** column indicates that the same lapse rate is used across all stations, whereas Kunkel (1989) refers to monthly-varying Γ from Table 5 in [121] and Dutra *et al.* (2020) refers to the digitized monthly values for *tasmax* from Figure 2d in [82]. ERA5 refers to dynamical ELRs derived similarly to [82], where the difference in temperature at pressure levels 950, 900, 850, 825 and $800\ \mathrm{hPa}$ from ERA5 is taken for pairs of pressure levels and only negative values are kept, which are then averaged into a dynamical ELR. The table agrees with [82] that no lapse rate yields uniformly better results and when applying the same rates to another region (not shown here), errors change yet again.

Table A.1: Various lapse rates and their effect on ELR correction following Equation (3.1) on ERA5 daily maximum temperature in Europe, based on observations in the validation set.

Lapse Rate	**RMSE [K]**	**MAE [K]**	**Bias [K]**	$\boldsymbol{R^2}$
$-0.0045\,\mathrm{K\,m^{-1}}$	2.02	1.47	0.05	0.965
$-0.0065\,\mathrm{K\,m^{-1}}$	1.96	1.44	0.01	0.967
Kunkel (1996)	1.92	1.42	0.01	0.966
Dutra *et al.* (2020)	1.95	1.43	0.03	0.965
ERA5	1.96	1.44	0.03	0.967
No correction	2.29	1.61	0.14	0.956

B Hyperparameter Optimization

The tables below show the hyper-parameter space explored during hyper-parameter optimization for each model. Table B.1 shows parameters used to tune Random Forest and ExtraTree regressors, Table B.2 those used for Gradient Boosting regressors, Table B.3 those for Histogram Gradient Boosting regressors and finally, Table B.4 shows the parameter space explored during training of the Multilayer Perceptron regressor. All parameter names refer to those used in `scikit-learn`, except for the MLP, which are self-explanatory.

Table B.1: Parameter space explored during hyper-parameter optimization of Random Forest and ExtraTrees regressors.

Parameter	**Value**
max_depth	10, 25, None
min_samples_split	2, 5, 25
min_samples_leaf	1, 5
bootstrap	False, True
max_features	*sqrt*, None, 0.6

Table B.2: Parameter space explored during hyper-parameter optimization of the Gradient Boosting regressor.

Parameter	**Value**
max_iter	50
max_depth	10, 25, None
min_samples_split	2, 5, 25
min_samples_leaf	1, 5
max_features	*sqrt*, None, 0.6
learning_rate	0.05, 0.1, 0.25

Table B.3: Parameter space explored during hyper-parameter optimization of the Histogram Gradient Boosting regressor.

Parameter	**Value**
min_samples_split	2, 5, 25
min_samples_leaf	1, 5
max_features	*sqrt*, None, 0.6
learning_rate	0.05, 0.1, 0.25
l2_regularization	0.0, 0.00001, 0.0001, 0.001

Table B.5 describes the final set of parameters used for the Random Forest, which was then used to produce the daily maximum temperature fields in this thesis. Each parameter there refers to those listed in the documentation of `scikit-learn` for the `RandomForestRegressor`.

Table B.4: Parameter space explored during hyper-parameter optimization of the Multi-layer Perceptron.

Parameter	**Value**
n_layers	2,3,4,5
layer_width	20, 50, 250, 500, 2500
activation_function	*relu*, *tanh*

Table B.5: Parameters represent the hyper-parameters used for the final Random Forest, based on the `RandomForestRegressor` from `scikit-learn` used to produce the daily maximum temperature fields in this thesis.

Parameter	**Value**
n_estimators	50
criterion	*squared_error*
max_depth	25
min_samples_split	25
min_samples_leaf	5
min_weight_fraction	0.0
max_features	*sqrt*
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	False
ccp_alpha	0.0
max_samples	None
monotonic_cst	None

C Additional Results from Model Selection

Table C.1 shows the effect of an artificial decrease in station density on the model performance. The Random Forest is trained on a decreasing number of stations and its errors, as measured by RMSE and Bias, as well as the R^2, are tracked on a static validation set. The last row contains the baseline results from ERA5 at all validation stations. All results, except the baseline, are calculated by training 25 individual models on random subsets of N stations for each station density and averaging their errors. This is done to counteract the effect of some observations being more representative of environmental conditions than others.

Table C.1: Summary of the influence of stations density on model performance in a fixed training area.

Number of stations	**RMSE** [K]	**Bias** [K]	$\boldsymbol{R^2}$
240	1.08	−0.005	0.98
120	1.13	0.007	0.98
60	1.19	0.015	0.98
30	1.24	−0.008	0.97
15	1.33	−0.004	0.97
7	1.42	0.089	0.96
3	1.67	−0.054	0.95
-	1.40	−0.587	0.97

D Additional Results from Temperature Field Evaluation

Figure D.1 shows the effect of applying a lapse rate correction to ERA5 daily maximum temperatures, depicting the global mean RMSE, aggregated to monthly values, both from the raw data and one where observation temperatures have been adjusted using a lapse rate of $\Gamma = -0.0065\,\mathrm{K\,m^{-1}}$.

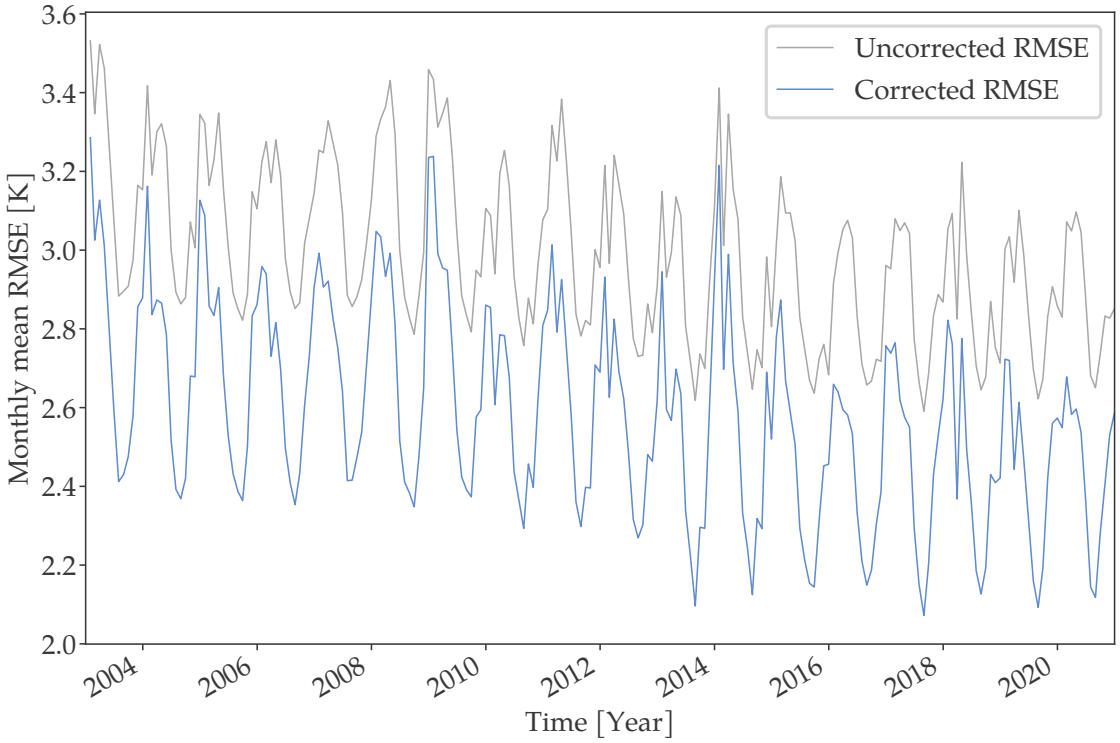

Figure D.1: Curves showing monthly mean RMSE of ERA5 against all test stations. The corrected curve is the result of applying a lapse rate correction with $\Gamma = -0.0065\,\mathrm{K\,m^{-1}}$, following Equation (3.1) and using the station elevation and ERA5 orography, to the temperature observations.

E Reanalysis-only Model

Table E.1 shows all trainig features from reanalysis and other sources used to train the model without any remotely-sensed inputs. It uses a wide variety of fields, both at the surface as well as in higher altitudes and combines them with static indicators, such as distance to the coast and population density.

Table E.1: A list of all input features considered for model training of the purely reanalysis-based model. Spatial and temporal resolution are given by $\boldsymbol{\Delta x}$ and $\boldsymbol{\Delta t}$, respectively. Auxiliary features may not have a spatial or temporal resolution, or they may be constant in time.

Variable	**Description**	$\boldsymbol{\Delta x}$	$\boldsymbol{\Delta t}$
t500	500hPa temperature	$0.25°$	1 day
t850	850hPa temperature	$0.25°$	1 day
z500	500hPa geopotential	$0.25°$	1 day
z850	850hPa geopotential	$0.25°$	1 day
q500	500hPa specific humidity	$0.25°$	1 day
q850	850hPa specific humidity	$0.25°$	1 day
tasmax	daily maximum temperature	$0.25°$	1 day
sp	surface pressure	$0.25°$	1 day
sktmax	daily maximum surface temperature	$0.25°$	1 day
sktmin	daily minimum surface temperature	$0.25°$	1 day
ws	daily mean wind speed	$0.25°$	1 day
ssrd	surface downward shortwave radiation	$0.25°$	1 day
strd	surface downward thermal radiation	$0.25°$	1 day
sde	snow depth equivalent	$0.25°$	1 day
oro	ERA5 orograhpy	$0.25°$	-
dem	MERIT digital elevation	$1\,\mathrm{km}$	-
d2c	Distance to coast line	$1\,\mathrm{km}$	-
lulc	Land cover	$1\,\mathrm{km}$	-
ghsl	Global Human Settlements population density	$0.01°$	-
cos_day	Cosine transform of day of year	-	1 day
lat	Latitude	-	-
lon	Longitude	-	-