Mapping urban temperature using crowd-sensing data and machine learning

Originally published in:
Urban Climate 35, https://doi.org/10.1016/j.uclim.2020.100739

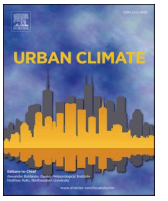
Funding acknowledgement:
167215 - Combining theory with Big Data? The case of uncertainty in prediction of trends in extreme weather and impacts (SNF)

Urban Climate 35 (2021) 100739

Contents lists available at ScienceDirect

Urban Climate

journal homepage: www.elsevier.com/locate/uclim



Mapping urban temperature using crowd-sensing data and machine learning

Marius Zumwald [a,b,*], Benedikt Knüsel [a,b], David N. Bresch [a,c], Reto Knutti [b]

[a] Institute for Environmental Decisions, ETH Zurich, Switzerland
[b] Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[c] Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland

ARTICLE INFO

Keywords:
Random forest
Urban heat
Low-cost sensors
Crowd-sensing
Machine learning

ABSTRACT

Understanding the patterns of urban temperature a high spatial and temporal resolution is of large importance for urban heat adaptation and mitigation. Machine learning offers promising tools for high-resolution modeling of urban heat, but it requires large amounts of data. Measurements from official weather stations are too sparse but could be complemented by crowd-sensed measurements from citizen weather stations (CWS). Here we present an approach to model urban temperature using the quantile regression forest algorithm and CWS, open government and remote sensing data. The analysis is based on data from 691 sensors in the city of Zurich (Switzerland) during a heat wave using data from for 25-30th June 2019. We trained the model using hourly data from for 25-29th June ($n = 71{,}837$) and evaluate the model using data from June 30th ($n = 14{,}105$). Based on the model, spatiotemporal temperature maps of 10×10 m resolution were produced. We demonstrate that our approach can accurately map urban heat at high spatial and temporal resolution without additional measurement infrastructure. We furthermore critically discuss and spatially map estimated prediction and extrapolation uncertainty. Our approach is able to inform highly localized urban policy and decision-making.

1. Introduction

Extreme heat has a range of adverse impacts on humans, e.g., by affecting cognitive and physical capacities (Kjellstrom et al. 2016), mental health (Obradovich et al. 2018) and sleep quality (Obradovich et al. 2017). It further increases mortality rates (Fouillet et al. 2008), the risk of accidents (Rameezdeen and Elmualim 2017) and disruptions in transport, information and communication technology and energy infrastructure (Chapman et al. 2013). An increase in mean and extreme temperature can already be observed due to climate change, and this trend will most likely continue into the future (Lorenz et al. 2019). Cities and urban areas are especially vulnerable to extreme heat because of the urban heat island (UHI) effect (Oke 1982) and are probably disproportionately affected in a changing climate (Tewari et al. 2019). Since increasing parts of the world's population live in cities, understanding the temperature distribution in urban areas is important for domains ranging from architecture and city planning to public health.

While it is known that urban temperatures are higher than rural temperatures on average, the spatial distribution of temperatures is highly complex and exhibits large within-city variation (Fenner et al. 2014; Voelkel and Shandas 2017). Urban heat mitigation

* Corresponding author at: Institute for Environmental Decisions, ETH Zurich, Universitätstrasse 16 8092 Zürich Switzerland
E-mail addresses: marius.zumwald@usys.ethz.ch (M. Zumwald), benedikt.knuesel@usys.ethz.ch (B. Knüsel), dbresch@ethz.ch (D.N. Bresch), reto.knutti@env.ethz.ch (R. Knutti).

https://doi.org/10.1016/j.uclim.2020.100739
Received 3 April 2020; Received in revised form 23 October 2020; Accepted 10 November 2020
Available online 8 December 2020

measures would benefit from a better understanding of how urban and architectural design affects temperature (Martilli et al. 2020). However, obtaining such information first requires knowledge of actual spatial temperature distributions in cities. This is especially important to better assess the vulnerability of groups like construction workers who cannot choose their location freely and can thus be involuntarily exposed to potentially harmful conditions. Furthermore, knowledge of the cooler spots can inform adaptation decisions at the individual level, for instance for leisure activities.

There are essentially three methods to map urban temperature distribution, namely in-situ measurements (Dejmal et al. 2019; Fenner et al. 2019) and subsequent statistical interpolation using additional explanatory factors and statistical prediction models (Johnson et al. 2020; Schatz and Kucharik 2014; Shi et al. 2018), remote-sensed land surface temperature (LST) (Good 2016; Voogt and Oke 2003), and numerical urban climate simulations, e.g., based on computational fluid dynamic (CFD) models (see Toparlar et al. 2017). However, each of these sources faces limitations for high-resolution information about urban temperatures. In-situ measurements from national weather providers are sparse in urban areas because the conditions required by the World Meteorological Organization (WMO) are difficult to meet in many urban environments. Moreover, the high costs to maintain such stations puts practical limits on the spatial resolution of such networks. Remote sensing can be used to measure LST at a high spatial resolution (Parastatidis et al. 2017). However, LST is not the most suitable parameter to estimate heat stress, since air temperature and LST deviate especially in situations of extreme heat (Good 2016). Furthermore, satellites have observational intervals of days to weeks, which makes a continuous observation difficult. Finally, CFD models represent physical processes in urban environments and thus create surrogate observational data. While such models can provide temperature information at high spatiotemporal resolution, they require large computational resources and often lack systematic evaluation (Toparlar et al. 2017).

The increasing digitalization has led to the development and standardization of communication protocols such as LoRaWAN (Long Range Wide Area Network) and to the production and deployment of affordable sensors for smart home applications. Such low-cost sensors have become widespread for example in the form of citizen weather stations (CWS) that are increasingly bought by individuals to measure meteorological parameters such as temperature, humidity and precipitation. The measurements are crowd-sensed by commercial or non-profit organizations. Using information from such low-cost sensors has a great potential to increase the spatial coverage of temperature measurements in urban areas. Hence, thanks to CWS data, the challenges of in-situ measurements and statistical interpolation for urban heat maps might be overcome.

Temperature maps based on in-situ measurements are subject to three sources of uncertainties. First, *measurement uncertainty* arises because of unknown accuracy of the CWS measurement. Second, *contextual uncertainty* arises from unknown biases due to the unknown exact sensor position (e.g., the height above ground). Third, *prediction uncertainty* arises because of the interpolation method used to estimate temperature.

In this study we present a case study of the city of Zurich (Switzerland) to show how CWS data and machine learning can be used to create and evaluate high-resolution maps of urban air temperature. To the best of our knowledge, this is the first study that uses low-cost CWS sensors to create temperature maps with a specific emphasis on careful model evaluation and investigation of uncertainties. The approach presented in this paper has a large potential to inform local urban climate change adaptation and identify vulnerabilities to extreme heat.

In Section 2, we present the construction of the dataset we used as input for machine learning and explain the quantile regression forest algorithm. In Section 3, we evaluate the model, present the uncertainty analysis and the modelled heatmaps for the city of Zurich. In Section 4, we discuss the modeling approach and the specific results for the city of Zurich and we give an outlook on how to

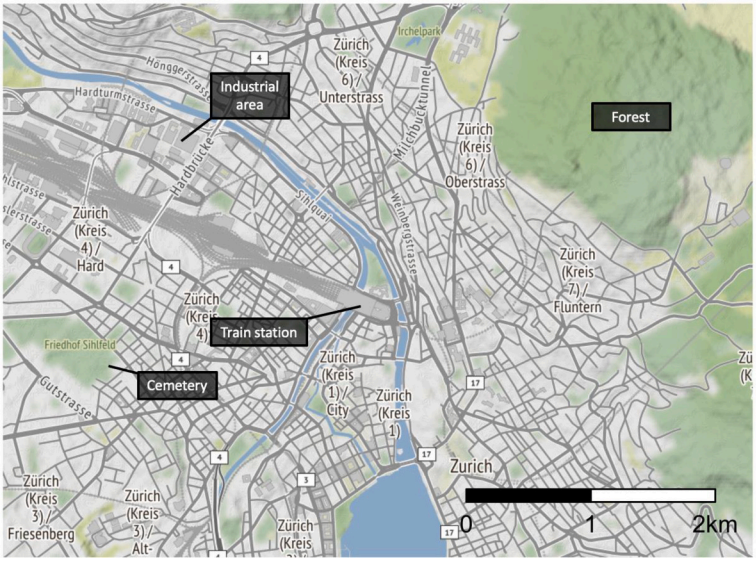


Fig. 1. Schematic map of the city of Zurich, lake level at 406 m a.s.l. The variety of different man-made and natural surfaces is shown. The forest shown is at a higher altitude of about 670 m a.s.l. Stamen map based on open street map data BSD −3 licensed.

further improve our approach. We conclude and discuss implications in Section 5.

2. Material and methods

The heatwave at the end of June 2019 broke many heat records across Europe. In Switzerland, June 2019 was the second warmest June after 2003 ever recorded, with temperature records available since the year 1864. Also, in Zurich, the largest city in Switzerland with a population of about 500′000 inhabitants, many daily and monthly heat records were measured (MeteoSwiss 2019). The approach presented in this paper is based on a case study in Zurich using data of June 2019.

Fig. 1 provides an overview of the city of Zurich. The variation in topography and land cover and use types makes Zurich an ideal case study for our approach. The study area includes surface water, rail tracks, urban greenspaces such as a large cemetery, and industry areas with different building stock characteristics than in residential areas, as well as forests at higher altitude.

2.1. Sensor data

We combined temperature sensors from three different sources (see Table 1). First, we used data from 634 personal weather stations from *Netatmo*, a commercial manufacturer and data aggregator of CWS. The CWS data was obtained using an application program interface (API). To account for uncertainties due to false set-up and irrelevant measurement contexts of the CWS, i.e., the first two sources of uncertainty discussed in the introduction, we applied a statistical quality control scheme developed by Meier et al. (2017) and Napoly et al. (2018) to filter the data based on the CrowdQC package for the statistical software R (Grassmann et al. 2018) which also has been applied successfully in other studies (Feichtinger et al. 2020). Here, we applied five quality control levels for manipulating and filtering inappropriate values. This resulted in 43% of the data lost (see Table A.1), which is around 8% more than in Feichtinger et al. (2020) which after applying the same QC scheme lost 35% of the data.

Second, we used data from 14 reference stations which are operated by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss), which corresponds to WMO standards, and air quality observation networks. Measurements and derived variables from the SMA reference station, which is part of the MeteoSwiss network, were not used to train or validate the model, but rather as predictor variables in our model. This is to account for the temporal dimension of the model. We do not use any direct temporal information such as time stamps as input, rather we use only meteorological variables from the SMA station. This allows to build, in principle, a much more flexible model that can also be used for nowcasting or short-term predictions under different weather conditions. However, applications under different weather conditions would require a prior testing and possibly retraining of the model for those specific weather conditions. Third, we used temperature data from the cantonal environmental agency Zurich (AWEL) which installed 43 low-cost sensors which use the LoRaWAN and could be accessed using an API from the manufacturer.

Fig. 2 shows the spatial distribution of the sensors used in training. The reference stations from official governmental meteorological and air hygiene measurement networks are depicted in green. Red dots denote the sensors deployed by the AWEL. They sample conditions in the urban space better than the reference stations. The CWS are depicted in blue.

2.2. Spatial and meteorological predictors

We generated a training dataset containing hourly temperature values by combining the data from all 691 sensors depicted in Fig. 2. Furthermore, the dataset contains 43 spatial and 35 meteorological predictors. In total we have 85,942 observations. We used observations for 25-29th June 2019 for training (71,837 observations) and the data of the 30th June for testing our model (14,105 observations). Our analysis focuses on extreme heat. All days from which data was taken, including the day used for evaluation, have similar cloud free weather in a persisting high-pressure system without any rain. Compared to random exclusion of certain time periods this is a more honest ways of evaluating our model; especially to predict future conditions.

The physical principles causing the spatial urban temperature variability are known from theory (Oke 1982; Oke et al. 2017). Because of buildings, urban areas exhibit larger surface areas that are more exposed to direct, diffuse and reflected solar radiation than natural environments. Also, the distance between buildings has an influence on temperature, as close-by buildings reduce the radiative night cooling by radiation blocking of neighboring structures. Concrete and asphalt have different thermal bulk and surface properties such as thermal conductivity, albedo and emissivity compared to natural environments. Furthermore, impervious surfaces lead to a lack of evaporative cooling in urban areas. As the city of Zurich is surrounded by hills, temperature is also affected by altitude. We used a combination of open government and freely available satellite data to account for all of the influences mentioned above. All spatial data used has a resolution of 10×10 m, which allows to nearly continuously predict urban temperature distribution (see Table A.2.1).

Specifically, the considered predictor variables were the altitude and a simplified classification of land use based on five classes (buildings, rail and roads, urban green spaces, surface water and forest). Furthermore, for built infrastructure we additionally considered the building volume as continuous predictor. Also, in addition to the categorical classes forest and urban green spaces we used the continuous normalized difference vegetation index (NDVI) (see Appendix A.2) from Sentinel-2 mission data. The NDVI exhibits values from -1 to 1. Values below the threshold of 0.1 were excluded (Weier and Herring 2000), since they do not represent vegetation but surface water, streets and buildings. However, those previously mentioned eight predictors are not sufficient for training a reliable model since temperature is also affected by the surroundings on a larger spatial scales (see Konarska et al. 2016). For example, not only the buildings at the measurement station location are important, also how many buildings and what kind of vegetation surround the measurement station. To account for this, we derived, with the exception of altitude, additional predictors in addition to the eight original ones. Specifically, five additional features were engineered using convolutional kernels of five different

Table 1
Overview of different sensor types.

Type	Amount	Description	Access
CWS	634	CWS deployed by individuals. Uncertainty about the exact measurement conditions, which potentially leads to contextual biases.	https://dev.netatmo.com/en-US/resources/technical/reference/weatherapi
Reference stations	14	Reference stations including WMO conform stations by MeteoSwiss and sensors from the national (NABEL) and cantonal (Ostluft) air quality networks.	https://gate.meteoswiss.ch/idaweb/
AWEL sensors	43	Temporarily deployed sensors by cantonal environmental agency AWEL. Deployed at 2-3 m height above ground.	https://docs.decentlab.com/data-access-guide/v5/index.html

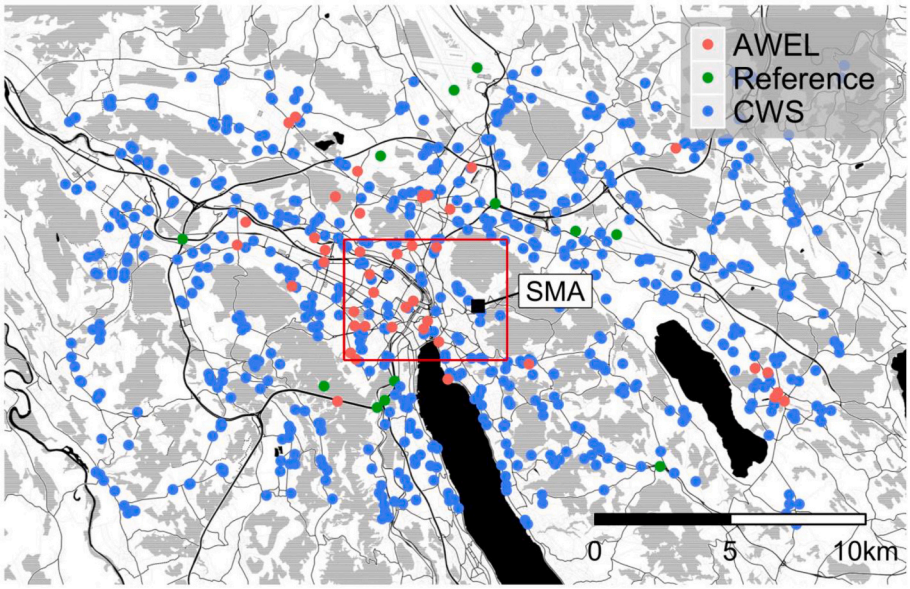


Fig. 2. Spatial distribution of all temperature sensors used. Three different sensor types are depicted in red, green and blue. Meteorological parameters from the SMA reference stations (black square) is used as input in the model. Red frame delineates the area used for the spatial prediction, hence to generate high resolution temperature maps. Stamen map based on open street map data BSD −3 licensed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sizes to account for the influence of the surrounding environment in 10, 30, 100, 200 and 500 m.[1] We used two types of kernels. For the discrete predictors the sum within the kernel size of the aforementioned distances is used, which corresponds to a kernel that only has the value 1 as entry for every instance. For the continuous variables NDVI and building volume a Gaussian blur kernel[2] was applied, which reduces noise in the raster values. This resulted in total 43 spatial predictors (see Figure A.6 for examples).

Finally, to account for temporal variation 36 meteorological predictors from the SMA reference station (see Fig. 2) were used as predictors. The predictors vary hourly and daily. Different meteorological parameters such as air temperature, radiation, wind and humidity and derivates were used. For a detailed overview, see Table A.2.2.

2.3. *Quantile regression forest*

To predict temperature in an urban setting non-linear modeling approaches are needed, since linear models are not able to explain a large part of the spatial variance in temperature (Voelkel and Shandas 2017). Neural network and deep learning approaches allow for large flexibility and predictive power but are harder to interpret than ensemble-based approaches which allow for the required flexibility, while still providing insights into the algorithms' inner workings, e.g. via variable importance and prediction uncertainty estimation. Hence, in this study we use the quantile regression forest (QRF) to predict temperature. QRF is an extension of the random forest algorithm (RF) (Breiman 2001), which itself is based on regression trees. Regression trees divide the predictor space into high-dimensional boxes such that the residual sum of squares is minimized. In order to prevent overfitting a cost complexity pruning punishes trees with too many leaves. However, growing individual regression trees is path dependent, since the algorithm only optimizes one variable selection and splitting point at a time. The RF algorithm uses bootstrap sampling and only a randomized subset of

[1] Given a 10 m grid size this corresponds to 3 × 3, 7 × 7, 21 × 21, 41 × 41 and 101 × 101 kernels.
[2] We used 3 σ at kernel size to generate the kernel matrices.

variables to choose the best variable and split point. These randomizations allow to grow an ensemble of deviating, decorrelated regression trees. As a rule of thumb, often a subset of one third of the variables is used at each split (Hastie et al. 2009). However, here, this parameter was estimated before training the final model by considering the out-of-bag error. RF as described by Breiman (2001) averages the ensemble of regression trees for more robust and better predictions and has been used for spatial prediction problems (Gudmundsson and Seneviratne 2015; Kerckhoffs et al. 2019; Li et al. 2011). QRF uses the RF ensemble to derive information about the conditional quantiles, hence it can be used to estimate prediction uncertainty (Meinshausen 2006). The method applied here deviates from the RF as proposed by Breiman (2001) and QRF by Meinshausen (2006) following more recent ideas (Athey et al. 2016; Athey and Wager 2019). For instance, when drawing a bootstrap sample to learn one single decision tree, we introduce an additional constraint during this learning process. Namely we cluster the training sample by the specific hour and day used. A cluster contains all measurements that were taken e.g. at 12:00 on the 29th of June. Measurements from this cluster are either in the bootstrap sample or discarded completely but cannot be in both. The reason is that we assume heterogeneity between the data of different clusters that cannot be fully captured by the variables included in our model. Accounting for that allows for better generalization of the trained model. Formally, this can be seen as a "non-parametric random effects modeling" (Athey and Wager 2019, p. 4).

Since we only used data from a limited number of days to train our model, this accounts for specific configurations that are not represented by our predictor variables. For this study we used the QRF implementation of the grf R-package (Tibshirani et al. 2019) following the method described by Meinshausen (2006). See Table A.4 for all model parameters.

3. Results

Here, we present the model evaluation of our approach for the specific case of modeling spatiotemporal temperature distribution for the city of Zurich and present an in-depth uncertainty analysis. Finally, we show the generated maps for Zurich.

3.1. Model evaluation

We evaluate the model by using the root mean square error (RMSE) metric and the error. For a model trained on all the data, the RMSE for the reference sensors is 1.43 °C, 1.35 °C for the AWEL sensors and 1.71 °C for the CWS sensors. For all station types, there is a systematic bias of underestimating high temperatures and overestimating low temperatures (Fig. 3).

The RMSE is highest in the early hours from 03:00 until 05:00 (Fig. 4 panel B). It is lowest at 07:00 to 08:00 and then follows some sort of diurnal cycle, with an outlier at 15:00. Generally, the RMSE does not follow a clear pattern over time. In case of the error, the situation looks less noisy (Fig. 4 panel A). Temperature predictions are overestimated more strongly during night starting from 03:00 and underestimated from 07:00 and again around 18:00. The width of the distribution of errors does not follow a clear diurnal pattern but whether the median under- or overestimates the temperature does. The diurnal pattern for the AWEL and reference sensors is very similar and somewhat different for the CWS (see Figure A.5). This might be because the radiation-induced biases could not be completely removed by the applied quality control procedure. However, the different characteristics could also be a result of different measurement locations (e.g. close to building walls). This seems a probable hypothesis as for the reference and AWEL there is a strong drop in bias around sunset but not so for CWS sensors where the temperature error stays constant (as in Fig. 4 panel A), which might be due to strong back radiation from man-built materials.

So far, we evaluated the model that has been trained on all available sensors. To understand how the proposed modeling approach performs using fewer sensors, we performed a sensitivity analysis concerning the training data. The model was re-trained on six different subsets of data. The performance of the model on each of these subsets was then compared with the performance of the model trained on the full dataset, which was introduced above (see Table 2). Two subsets (*temporal subset I* and *temporal subset II*) are used to investigate the effect of less training data. These subsets were created by excluding one and two days of training data from the original set, respectively. A third subset (*sensor subset I*) was created by reducing the dataset to the AWEL and reference data. A fourth one (*sensor subset II*) consists of the AWEL data, only, and a fifth one (*sensor subset III*) consists of the reference data, only. The sixth subset only used CWS data for training (*sensor subset IV*). The model performance is compared with the same evaluation dataset for each of these six models. When excluding data from 25th of June, the RMSE increases marginally. When data from the 26th of June is additionally excluded, the RMSE increases more substantially. This indicates that a few days of data with similar weather conditions is needed and more data helps to improve the model's performance. When excluding the CWS stations ($n = 634$) for training and only training on the AWEL ($n = 43$) and reference sensor ($n = 14$) data, the model performance improves for these sensors while worsening for the CWS station. This indicates that the data in the reference and AWEL stations is insufficient to create accurate heat maps for locations for which no such standardized station is available. Interestingly, training the model on AWEL sensors is leading to worse performance on CWS than only training with the reference sensors. A possible explanation is that the different measurement station types do not sample the relevant features similarly well. For instance, the reference stations do sample, according to our feature importance metric (see Section 3.2), the most important feature altitude better than the AWEL stations. The sensitivity analysis reveals the additional CWS data indeed improves the model substantially and that a certain number of different days representing similar condition is needed. Hence, this analysis confirms the overall modeling approach chosen here. In the subsequent sections, we only use the baseline model trained with all of the data.

In addition we also performed a leave-p-out cross-validation to test the effect of a random exclusion of measurement stations independent of the sensor type, hence a purely spatial validation. Over all data we randomly excluded 10% of the stations which were used as test set and the algorithm was trained on the remaining 90% of data. We repeated this procedure 10 times. The average RMSE on the test sets is 1.86 °C. To test for the sensitivity of using the meteorological predictors of the SMA weather station we trained the

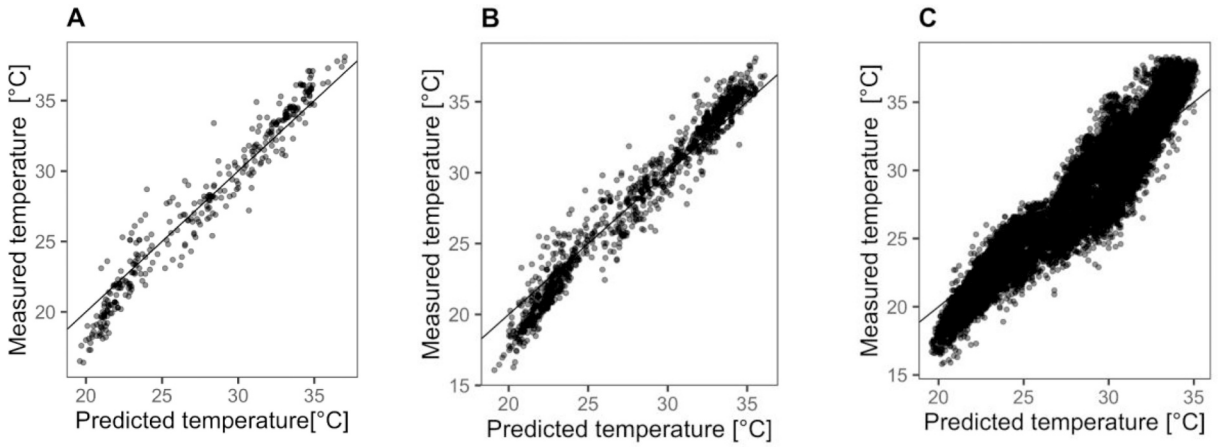


Fig. 3. Comparison of predicted and measured temperature data. A) depicts the reference data resulting in RMSE of 1.43 °C and R-squared of 0.95. B) shows the AWEL sensors with RMSE of 1.35 °C and R-squared of 0.96. C) shows the CWS sensors with RMSE of 1.71 °C and R-squared of 0.92.

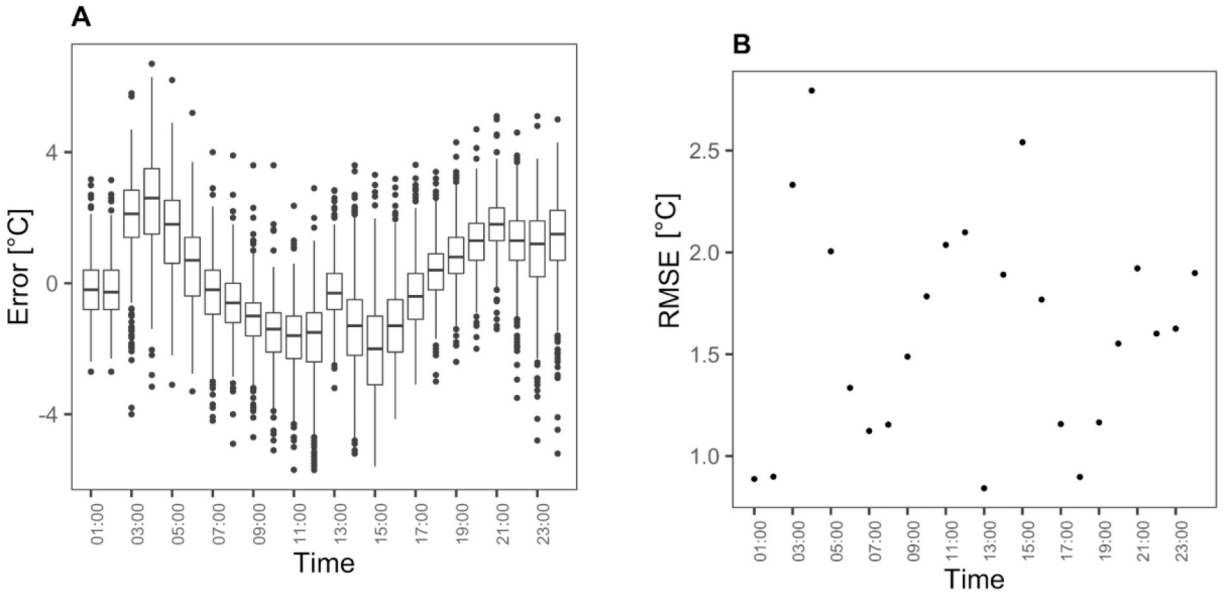


Fig. 4. A) distribution of the error over the day across all hours of the 30th of June. B) RMSE over the day across all hours of the 30th of June. The error shows the directionality of the error, however since errors can cancel the RMSE is a better metric to characterize the magnitude of an error.

Table 2
RMSE comparing evaluation metric on data from 30th June based on using different spatial and temporal subsets for training. The baseline model uses the whole training set ($n = 71{,}837$) for training.

	Baseline	Temporal subset I	Temporal subset II	Sensor subset I (Reference, AWEL)	Sensor subset II (Reference)	Sensor subset III (AWEL)	Sensor subset IV (CWS)
All [RMSE ° C]	1.69	1.83	2.21	2.37	2.29	2.42	1.82
CWS [RMSE ° C]	1.71	1.86	2.23	2.47	2.36	2.50	1.75
Reference [RMSE ° C]	1.43	1.59	2.19	1.30	1.27	2.02	2.67
AWEL [RMSE ° C]	1.35	1.45	2.00	1.14	1.57	1.13	2.38
Size training set	71,837	57,294	42,774	6830	1678	5152	65,007

baseline model also with another weather stations from the MeteoSwiss network located at the airport (KLO). In this case the RMSE for the CWS is 1.75 °C (compared to 1.71 °C for SMA), for the reference stations 1.45 °C (compared to 1.43 °C for SMA) and 1.37 °C for AWEL (compared to 1.35 °C for SMA). Hence, no sensitivity regarding to the choice of the reference station could be detected.

3.2. Variable importance

The variable importance gives a measure of importance for each predictor based on the split frequency of each predictor, hence, summed over all predictors the importance equals to one (see Table A.3 for details). While such a measure gives a good indication of which variables are important for the performance of the model, the outcome is not robust to changes in the number of variables used in the model. Due to the aim of building a spatiotemporal prediction model we have two classes, namely spatial and meteorological predictors. In total 96% of the variance is explained by the meteorological predictors and the remaining 4% by the spatial predictors. The most important meteorological predictors are the ambient air temperature 2 m (33%) and 5 cm (21%) above ground and the relative humidity 2 m above ground (13%). The next most important predictors are global radiation (8%) and sunshine duration (5%). Also, wind gust maximum of the last 2 h (3%) and 6 h (2%) are comparatively important. However, we are mainly concerned with spatial patterns; hence the most important spatial predictors are of large relevance for the subsequent analysis. The most important spatial features are altitude (0.7%), buildings 500 m (0.5%), ndvi500 m (0.5%), rail&road 500 m (0.4%). All predictors concerning water and forests are only of minor relevance.

The reason why the meteorological predictors explain the major part of the variance is that we build a spatiotemporal model and the temperature differences are larger between day and night than within space at the same time. Yet, as our focus is on the spatial patterns one could also only train a model for the same time of the day. Then the spatial variables would likely have a higher importance. However, the modeling approach of this study aims to explicitly track spatiotemporal pattern. Thus, the low importance of spatial predictors is no reason for concern. Furthermore, the QRF approach is well suited to separate the meteorological signal from the spatial signal because it is a non-linear approach and de-correlates the different features used.

3.3. Uncertainty analysis

When generating high-resolution temperature maps based on in-situ observations, the values for the grid points between the measurement sites need to be interpolated. Such temperature maps can be uncertain for three general reasons. First, *measurement uncertainty* is the uncertainty about the accuracy of a measurement. Second, *contextual uncertainties* arise due to unknown or uncorrected biases in the measurement sample (Zumwald et al., submitted). Third, *prediction uncertainty* arises from the use of an interpolation model. The *measurement uncertainty* can be quantified in a straightforward way for the present study, and its magnitude is comparatively low. For warmer temperatures, the reported accuracy of ±0.3 °C for CWS devices has been confirmed by climate chamber experiments (Meier et al. 2017). The AWEL sensors have a specified uncertainty of ±0.1 °C[3] and the measurement uncertainty of the reference stations is probably even lower. The *contextual uncertainty* is low, too, because many known *contextual biases* have been corrected by applying a data quality scheme (Meier et al. 2017; Napoly et al. 2018). The model validation has furthermore shown that the difference between the sensor types is $\Delta_{RMSE} = 0.26$ °C. Hence, the difference between the sensor types is an order of magnitude smaller than the absolute RMSE, indicating that the different sensor types are not affected by substantial unaccounted biases.

The *prediction uncertainty* is the uncertainty of the model output. Assuming that there is no omitted variable bias, there are two types of *prediction uncertainty:* first, *estimated* prediction uncertainty and second, *extrapolation* uncertainty. U*ncertainty* can arise if a similar configuration of predictors maps to different target values in the training data, which is being *estimated* by the prediction algorithm. However, uncertainty can also arise due to *extrapolation* if the model is used to predict temperature based on predictor configurations that have not been observed during training. As a non-parametric method, QRF is highly adaptive to data, but it does not generalize well in case of extrapolation. QRF gives an *estimate* of the prediction uncertainty in order to approximate the real prediction uncertainty. This estimate is specific to the QRF algorithm. The estimated uncertainty by QRF is not or only weakly related to the density of sampling locations in the predictor space but rather to the distribution of the predictor variables (Fouedjio and Klump 2019). Hence, the QRF does not capture uncertainty due to extrapolation well, especially when the predictors are assigned low variable importance. This means that for locations in space where extrapolation uncertainty is large, it is more difficult to estimate the total prediction uncertainty quantitatively. This means in locations where the extrapolation uncertainty is large, one should have less confidence in the estimated prediction uncertainty of the QRF algorithm.

3.3.1. Estimated prediction uncertainty

QRF allows an estimate of the prediction uncertainty of the model by estimating quantiles of the response variable based on the individual trees of the random forest approach. Fig. 5 shows the difference (ΔT) between the estimated 10th and the 90th quantile shows in a spatially explicit way. We additionally stippled areas which are above the 95th percentile of ΔT, regarding the shown area for the respective time (i.e., 15:00 or 04:00). This is to highlight areas of very high estimated prediction uncertainty.

We generally see that the spread of the estimated prediction uncertainty for the situation during daytime (around 4 to nearly 9 °C, panel B) is higher than during the night (around 4.5 to 6.5 °C, panel A). Generally, the estimated prediction uncertainty is largest in

[3] https://docs.decentlab.com/data-access-guide/v5/index.html.

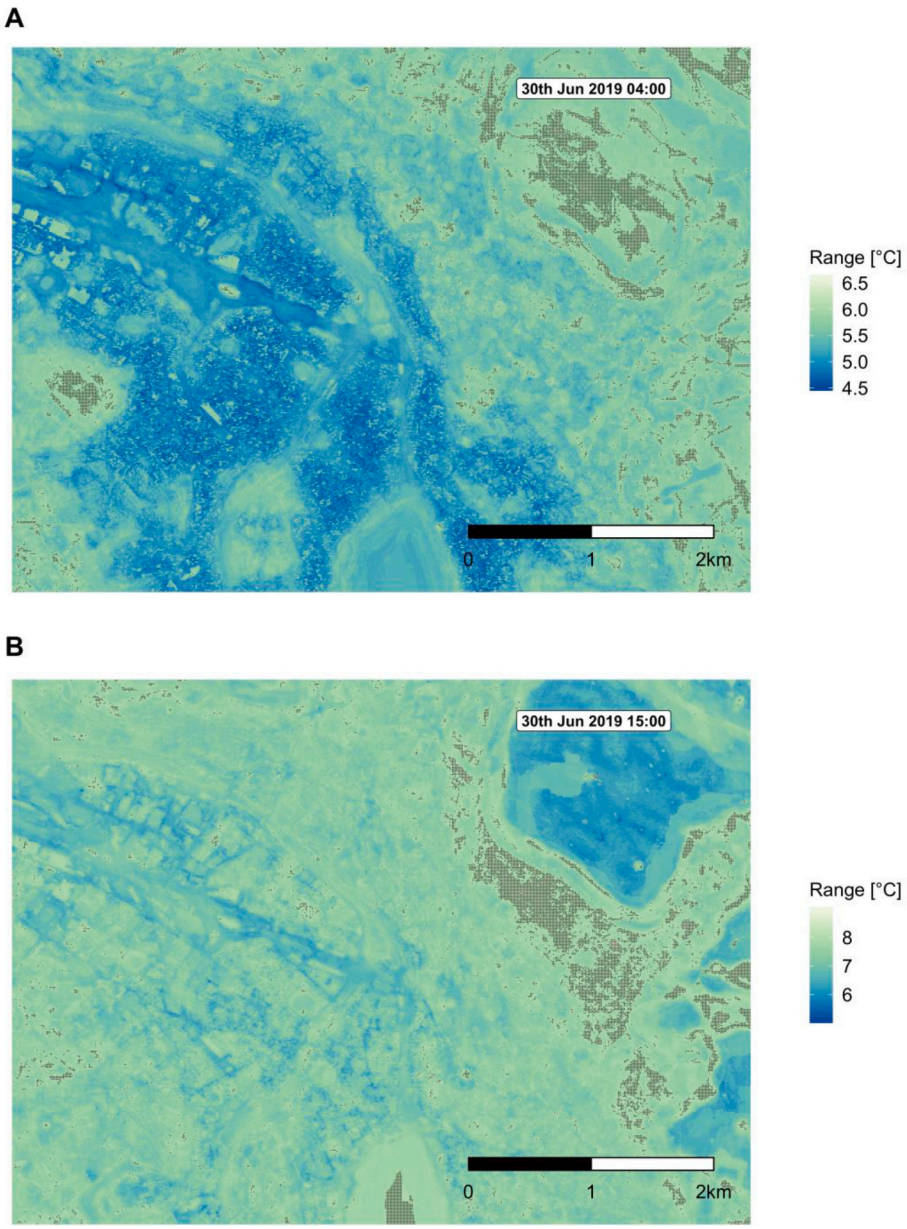


Fig. 5. Prediction uncertainty estimated as the difference in predicted temperature for the 10th quantile and the 90th quantile model. A) the prediction uncertainty for 30th June at 04:00. B), prediction uncertainty for 29th June 2019 15:00. Area depicted within the red frame as of Fig. 2. The colour scale differs for the night and day situation. The reason for this is that it is optimized to show patterns within a situation rather than being directly comparable across different situations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

proximity to forests. During the day, the uncertainties are rather homogenously distributed in the city center. In contrast, during the night, the patterns are more heterogeneous. The estimated uncertainty is generally low in areas with high building density, but this pattern is interrupted by parks and other green areas.

3.3.2. Spatial extrapolation uncertainty

If the model extrapolates to certain predictor values that have not been used in model training, then the estimated prediction uncertainty may not be reliable. This is especially important for RF algorithm which is a non-parametric method that fits on data and is known for performing poorly in extrapolation problems.

To learn about this source of uncertainty, Fig. 6 shows areas in which five spatial predictors exhibit configurations that are not sampled by any measurement station in the dataset. In these geographical areas, the model has to extrapolate, thus the model

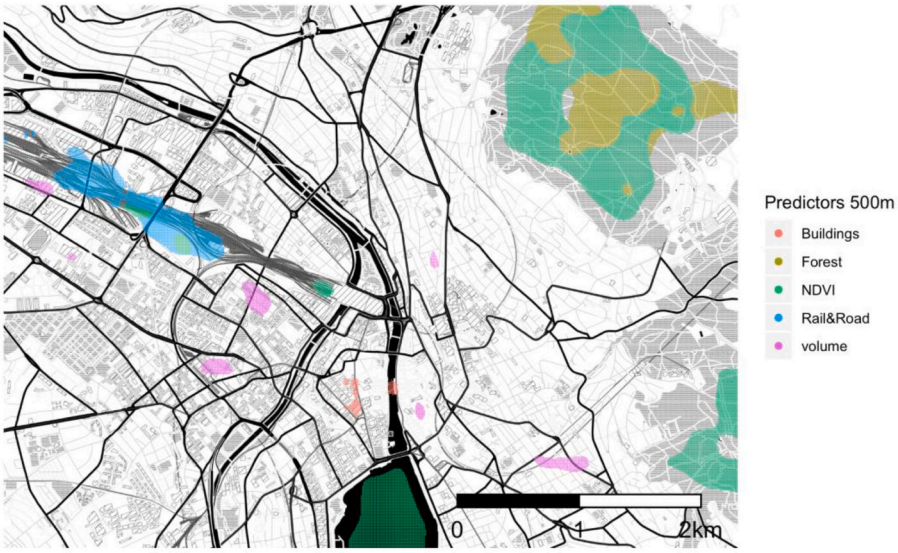


Fig. 6. Areas of extrapolation for the spatial predictors buildings, forest, NDVI, Rail and road, volume. Area depicted within the red frame as of Fig. 2. Stamen map based on open street map data BSD −3 licensed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predictions and the estimated prediction uncertainty might be less reliable.

3.4. Mapping spatiotemporal urban temperature and uncertainty

The trained model can be used to estimate the spatial temperature distribution on a 10x10m grid for current, past or hypothetical meteorological situations. Fig. 7 shows the estimated temperature distribution calculated with the meteorological data for 29th June at 15:00 and 04:00. In order to highlight areas of high *prediction uncertainty*, gray stippling is applied to areas in which the *estimated prediction uncertainty* (90th – 10th percentile) is higher than in 95% of the study area (see Fig. 5), and/or in which the *extrapolation uncertainty* is large (see Fig. 6). This allows to understand where the predicted values are rather uncertain and the result of our model should be interpreted with more caution.

The median prediction for the situation in the afternoon at 15:00 (Fig. 7 panel B) exhibits several patterns. First, on the rail tracks the temperature is warmer. The coldest spot is at higher altitudes in the forest. Furthermore, in built-up areas, there is a temperature difference of about 2–3 °C between the hottest and the coldest spots. Large green areas such as the large cemetery in Southwest Zurich are slightly cooler; however, the difference is less noticeable for smaller parks. Generally, the map allows to identify different hotter areas e.g. at crossings and industrial areas. In contrast, the patterns are much more accentuated during the night. The temperature difference between impervious surfaces such as parks and sealed surfaces is clearly visible. While the spatial predictors applying a convolution filter of the size of 500 m are most important during daytime, this seems not to be the case at night, as then the temperature is also influenced to a relevant degree on the spatial scale of 100 and 30 m. Also, the rail tracks, the hottest area during the day, is colder than the surroundings during night. However, for certain parts of the rail track the model is extrapolating and hence, more uncertain. Given the uncertainty analysis provided above (see Fig. 5), confidence in the maps is lower for the forest areas and the areas at the forest boundary. All temperature maps covering the data for the whole day of June 30th for Zürich are published as a dataset (doi: https://doi.org/10.3929/ethz-b-000442556).

4. Discussion

We present an approach to make use of CWS for data-driven modeling to create spatiotemporal high-resolution temperature maps. An in-depth model evaluation and uncertainty analysis allows to understand where uncertainties are higher in a spatially explicit manner. In this section we discuss the results, the evaluation and uncertainty analysis and for what kind of applications such a modeling approach might be suitable.

4.1. Results

Compared to the baseline RMSE of 1.69 °C, the *Temporal subset I* has a higher RMSE of 1.83 °C. The performance drop is larger for the *Temporal subset II* with an RMSE of 2.21 °C. Hence, the validation experiments using temporal subsets shows that increasing data increases the performance, although there is a saturation effect with increasing amounts of data. The sensor subset experiments show, that excluding certain sensor types from training, as expected, worsens the performance of this sensor type. Finally, the p-leave-out

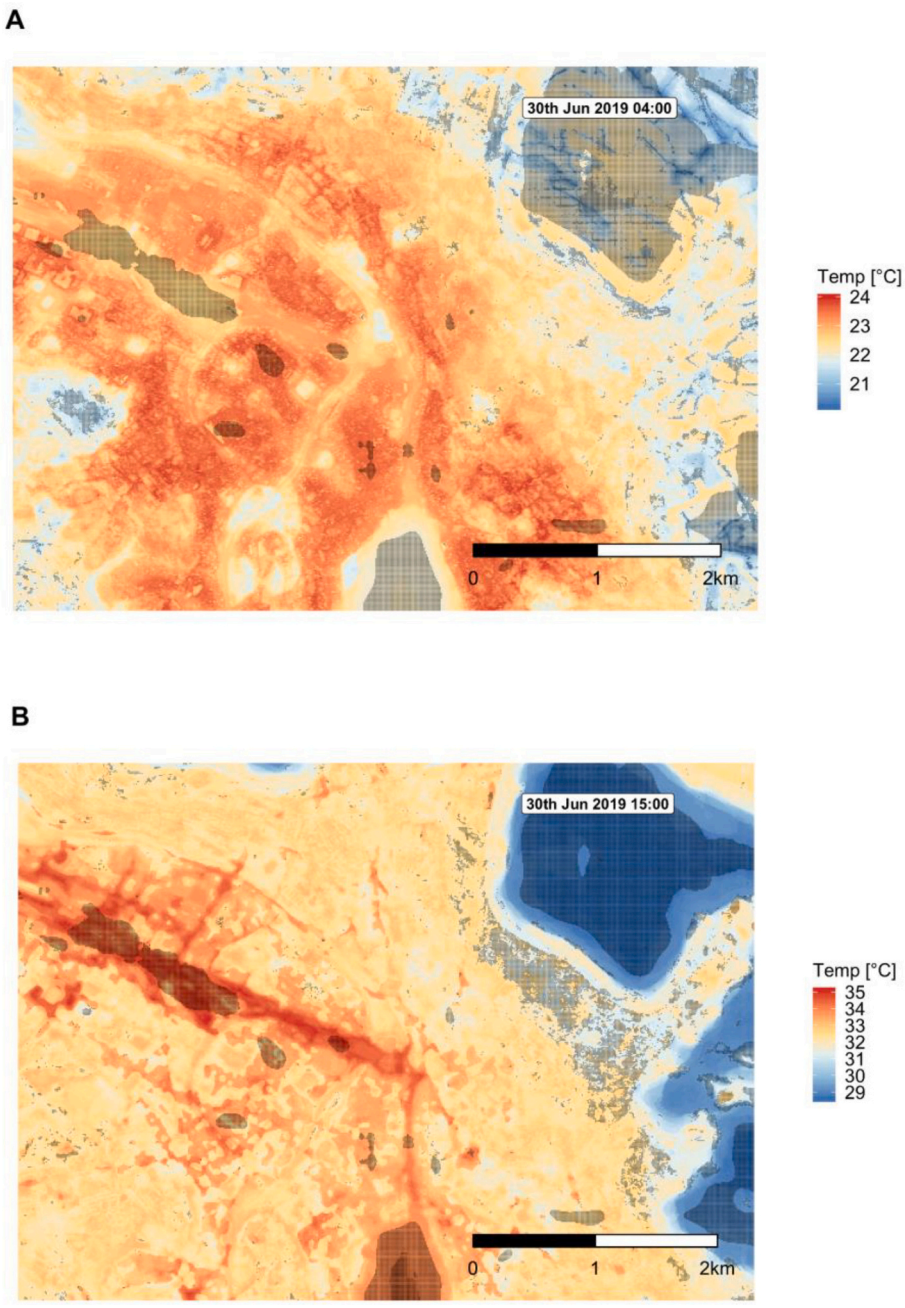


Fig. 7. High resolution heat map for weather situation at June 30th at 04:00 (Panel A) and 15:00 (Panel B). Showing the median estimate of the QRF and the prediction and extrapolation uncertainty. Area depicted within the red frame as of Fig. 2. The colour scale differs for the night and day situation. The reason for this is that it is optimized to show patterns within a situation rather than being directly comparable across different situations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

validation with random exclusion shows that the model can predict locations that have not been observed during training with an RMSE of 1.86 °C. Generally, the evaluation shows, that the model works reasonably well when evaluated on unseen test data, in the spatial and temporal dimension, and it proves adequate for the purpose of understanding especially spatial temperature patterns under conditions of extreme heat.

The model delivers coherent temperature maps at high spatio-temporal resolution but shows a systematic bias when evaluated with data from either sensor type (see Fig. 3), namely it underestimates high-temperatures and overestimates low temperatures. This is a typical problem of RF used for regression problems (see Zhang and Lu 2012). Furthermore, the evaluation shows a time dependency of the RMSE and mean error, confirming the bias observed in Fig. 3, where warmer temperature (day) are underestimated and colder

temperatures (night) are overestimated. During day and night different physical processes are dominant. E.g., back radiation of long-wave radiation is strong during the night, while exposure to direct shortwave radiation from the sun is largest around noon and in the early afternoon. Results shown in Fig. 4 suggest that our model is less well equipped to represent long-wave back radiation during night and direct sun irradiation during the day. It performs best in the early morning when neither process is dominant. Hence, to improve model performance, these processes should be better represented by including additional predictors as discussed in the coming sections.

4.2. Measurement and prediction uncertainty

The model RMSE is at least an order of magnitude larger than the measurement error of all used sensors indicated by the manufacturers. However, the model RMSE is larger for the CWS sensors than for the reference station, with the difference being $\Delta_{RMSE} = 0.26$ °C. Some of this difference might be due to the larger measurement error of the CWS sensors. Due to the lower level of standardization, the CWS sensors are also subject to larger contextual uncertainties. The reference and AWEL sensors are mostly deployed at height of two meters. It is unclear whether the CWS are placed at the same height above ground. Furthermore, CWS sensors are likely closer to building walls. These factors may be responsible for the larger RMSE for CWS data. For instance, temperature in a street canyon can easily vary by 2–3 °C degrees in the vertical direction (Sugawara et al. 2008). However, since the absolute values of the RMSE are one order of magnitude larger than the Δ_{RMSE}, other factors are a likely further source of uncertainty, most notably the predictive uncertainty of the model. At the same time, the fact that the measurement accuracy and contextual biases of the CWS sensors do not seem to be a dominant source of uncertainty indicates that both, the use CWS data and the application of the pre-processing procedures is justified.

4.3. Generalizability

The presented model is fitted to the conditions present in Zurich. However, since CWS are available in many locations in a sufficient number to train data-driven models, the presented approach can be easily transferred to other cities if the model is re-trained with site-specific predictors. Furthermore, transfer learning could be applied in cities without sufficiently many CWS sensors. Transfer learning is a machine learning technique that adapts a trained model in a domain with sufficient data (e.g. Zurich) to a domain with sparse data (Tan et al. 2018). Most examples are used in the field of deep learning where trained models trained on images have for instance been retrained to predict different sound events from spectrograms (Shen et al. 2019). Similar approaches exist for random forest, the algorithm used in this study (Segev et al. 2017).

In the presented case we only trained the model on cloudless hot summer days, hence a prediction is reliable for similar weather conditions as trained and does not generalize to unseen weather conditions such as rainy and cold weather. The model developed here does learn the temporal dimension solely using weather parameters and can be, in principle, be trained for additional weather situations. For this, additional meteorological parameters (e.g. from the SMA reference station) representing factors such as cloud cover, fog or precipitation need to be considered. Furthermore, sensor measurements that represents those different weather conditions all year round are needed to train the model.

4.4. Potential applications

The heat maps of our study can be used as input for vulnerability and risk assessments, e.g., to assess how different vulnerable groups are affected by heat stress or to assess labor productivity loss on high spatiotemporal resolution (Kjellstrom et al. 2009). As the CWS sensors sample areas where people live better than reference stations, this information would not be available without CWS. However, there are better indices to model heat stress than air temperature. Indices such as Wet Bulb Globe Temperature (Budd 2008) or Universal Thermal Climate Index (Jendritzky et al. 2012) are better suited to represent the heat stress on individuals. Most CWS stations do also have humidity sensors which potentially can be used to e.g. calculate a heat index source that is a combination of temperature and humidity. Using further information about radiation and information about shading, more complex indices such as WBGT could be calculated. They could also be approximated using standard meteorological measurements (Liljegren et al. 2008).

The presented approach can be used to inform urban planning and urban design, since, for such applied tasks, planners are more interested in patterns and relative differences are more important than accurate absolute temperature values. In this context, it is noticeable that the patterns and local hotspots are estimated in a consistent manner independent of which estimate of the QRF is used. For example when comparing the median estimate (Fig. 7) and the 10th and 90th percentile estimates (see Figure A.7).

To detect a climate change signal, current observation networks are using standardized measurement conditions and apply homogenization of time series in order to account for potential changes in measurement conditions. Due to lack of knowledge this is not straightforward for CWS, since there is for instance a lack of information about the height above ground for the CWS, or changes in instruments, or properties of buildings. As long as the CWS error is random and changes can be captured with explanatory variables of the models detecting a climate change signal should be possible. Estimating long-term trends would likely involve a more complex model. The approach presented is most straightforwardly used to nowcast and predict the spatial distribution of temperature given the current or estimated meteorological conditions from a weather forecast. Such predictions could subsequently be used for highly localized warning systems. We have furthermore investigated where the model spatially extrapolates. However, for certain applications those regions are of special interest, e.g. as region over the train tracks. The extensive uncertainty analysis provided here can hence allow to guide the placement of sensor networks similar to the AWEL.

4.5. Outlook

To better understand the differences between the different sensor types, an analysis focusing on the measurement outcome contingent on their measurement location in the predictor space would potentially yield valuable insights. Such an analysis could investigate for example questions like: Are CWS more similar to AWEL stations at high altitude or when surrounded by buildings? Such an analysis would lead to additional insights regarding potential biases in the measurements. Also, understanding the spatial importance of features for different times of the day could lead to important insights, which, requires to retrain the model for certain hours of the day.

The accuracy of the model could probably be improved incorporating additional predictors. To engineer predictors that represent the absorption and radiation of energy more accurately, the widely used and applied sky view factor could be used (Jänicke et al. 2016; Miao et al. 2019). Also, information about the albedo of surfaces such as rooftops might potentially improve model performance. Furthermore, also surface and building orientation could be included. While we have evaluated the model with measurements not used during training, the model evaluation could be enhanced by comparing the model performance of different modeling and measurement approaches. For instance, model outputs could be evaluated against land surface temperatures e.g. from the Landsat 8 satellite mission or compared against outputs of CFD simulations.

The focus of our approach lies on providing an accurate prediction of the highly resolved spatial distribution of air temperature. Testing hypothetical scenarios (e.g. by planting trees along streets or open spaces or even where there are buildings currently) would be another important application, especially for policy analysis and decision making. Physics-based models have the great advantage of allowing for direct counterfactual analysis of such hypothetical scenarios. However, also using RF for counterfactual explanations is a promising research direction. Methods developed in causal machine learning and especially causal forests (Athey et al. 2016; Athey and Wager 2019; Chernozhukov et al. 2018; Wager and Athey 2018) allow to model counterfactual scenarios. Hence, hypothetical inputs of used predictors such as NDVI can be used to model the effect of city wide greening scenarios on temperature.

While machine learning models often come with the connotation of being black boxes only useful for prediction, this is not entirely correct. There exists a variety of model specific variable importance metrics (Breiman 2001), but also model agnostic methods such as individual conditional expectation plots (Goldstein et al. 2014) or accumulated local effects (Apley and Zhu 2019) that can be used to gain insights into the model's behavior. Hence, a more in-depth model analysis might help to increase system understanding.

Lastly, approaches that combine data-driven modeling with physical modeling could lead to an approach which is more robust, reliable and trustful. For instance, one could apply a lapse rate correction before training the model and not use altitude as predictor anymore. Because the altitude was the most important spatial predictor, additional signals might be detected if we model the effect of altitude based on physical understanding of the system. To later predict the lapse rate correction can be reversed for all predicted raster grids.

5. Conclusion

Generating high resolution temperature maps requires large datasets, which can be challenging to obtain, specifically because official in-situ measurement stations are sparse in urban areas. CWS sensors are increasingly deployed in many cities and the data is collected and made available by governmental and for-profit companies. In this study, we present a strategy to make use of CWS measurements to create high resolution temperature maps. While the use of this data leads to some uncertainties, it allows to create high-resolution maps.

A careful evaluation of the models is key for making good use of such low-cost sensors. Our approach can be used with no additional measurements as long as some reference stations for model validation are available. Thus, CWS provide a cost-efficient means to gain insights into spatiotemporal temperature distribution in urban areas. Climate adaptation and specifically urban heat mitigation requires local information since the spatial temperature distribution is highly specific to the characteristics of a city. Due to the low costs, our approach is highly adaptable to local contexts and might be an alternative to CFD models especially in the context of urban decision and policy making. However, to build such models the insight gained from process-based models is crucial, e.g. for variable selection and feature engineering. Hence, different modeling approaches are best used in a complementary setting.

Low-cost sensors are currently used to measure various environmental parameters such as precipitation, wind speed and direction various air quality variables. Hence, we conclude that the abundance of low-cost sensors enables data-driven modeling of complex environmental systems in a very efficient manner, given a careful model evaluation and adequate use of reference data.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.uclim.2020.100739.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Gian-Marco Alt, Jörg Sintermann and Felix Baum from AWEL for providing access to their measurement data. Urs Beyerle provided technical assistance. We acknowledge the helpful thoughts and comments by Christoph Baumberger, Lukas Gudmundsson, Gertrude Hirsch Hadorn, Fred Meier, Nicolai Meinshausen, Andreas Scheidegger, Dieter Scherrer, Konrad Schindler and Sebastian

Sippel. This study was funded by the Swiss National Science Foundation, National Research Programme 75 Big Data, project number 167215.

References

Apley, D.W., Zhu, J., 2019. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. ArXiv:1612.08468 [Stat]. http://arxiv.org/abs/1612.08468.

Athey, S., Wager, S., 2019. Estimating Treatment Effects with Causal Forests: An Application. ArXiv:1902.07409 [Stat]. http://arxiv.org/abs/1902.07409.

Athey, S., Tibshirani, J., Wager, S., 2016. Generalized Random Forests. ArXiv:1610.01271 [Econ, Stat]. http://arxiv.org/abs/1610.01271.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Budd, G.M., 2008. Wet-bulb globe temperature (WBGT)—its history and its limitations. J. Sci. Med. Sport 11 (1), 20–32. https://doi.org/10.1016/j.jsams.2007.07.003.

Chapman, L., Azevedo, J.A., Prieto-Lopez, T., 2013. Urban heat & critical infrastructure networks: a viewpoint. Urban Clim. 3, 7–12. https://doi.org/10.1016/j.uclim.2013.04.001.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econ. J. 21 (1), C1–C68. https://doi.org/10.1111/ectj.12097.

Dejmal, K., Kolar, P., Novotny, J., Roubalova, A., 2019. The potential of utilizing air temperature datasets from non-professional meteorological stations in Brno and surrounding area. Sensors 19 (19), 4172. https://doi.org/10.3390/s19194172.

Feichtinger, M., de Wit, R., Goldenits, G., Kolejka, T., Hollósi, B., Žuvela-Aloise, M., Feigl, J., 2020. Case-study of neighborhood-scale summertime urban air temperature for the City of Vienna using crowd-sourced data. Urban Clim. 32, 100597. https://doi.org/10.1016/j.uclim.2020.100597.

Fenner, D., Meier, F., Scherer, D., Polze, A., 2014. Spatial and temporal air temperature variability in Berlin, Germany, during the years 2001–2010. Urban Clim. 10, 308–331. https://doi.org/10.1016/j.uclim.2014.02.004.

Fenner, D., Holtmann, A., Meier, F., Langer, I., Scherer, D., 2019. Contrasting changes of urban heat island intensity during hot weather episodes. Environ. Res. Lett. https://doi.org/10.1088/1748-9326/ab506b.

Fouedjio, F., Klump, J., 2019. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. Environ. Earth Sci. 78 (1), 38. https://doi.org/10.1007/s12665-018-8032-z.

Fouillet, A., Rey, G., Wagner, V., Laaidi, K., Empereur-Bissonnet, P., Le Tertre, A., Frayssinet, P., Bessemoulin, P., Laurent, F., De Crouy-Chanel, P., Jougla, E., Hémon, D., 2008. Has the impact of heat waves on mortality changed in France since the European heat wave of summer 2003? A study of the 2006 heat wave. Int. J. Epidemiol. 37 (2), 309–317. https://doi.org/10.1093/ije/dym253.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2014. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. ArXiv:1309.6392 [Stat]. http://arxiv.org/abs/1309.6392.

Good, E.J., 2016. An in situ-based analysis of the relationship between land surface "skin" and screen-level air temperatures. J. Geophys. Res.-Atmos. 121 (15), 8801–8819. https://doi.org/10.1002/2016JD025318.

Grassmann, T., Napoly, A., Meier, F., Fenner, D., 2018. Quality Control for Crowdsourced Data from CWS. https://doi.org/10.14279/depositonce-6740.3.

Gudmundsson, L., Seneviratne, S.I., 2015. Towards observation-based gridded runoff estimates for Europe. Hydrol. Earth Syst. Sci. 19 (6), 2859–2879. https://doi.org/10.5194/hess-19-2859-2015.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Jänicke, B., Meier, F., Lindberg, F., Schubert, S., Scherer, D., 2016. Towards city-wide, building-resolving analysis of mean radiant temperature. Urban Clim. 15, 83–98. https://doi.org/10.1016/j.uclim.2015.11.003.

Jendritzky, G., de Dear, R., Havenith, G., 2012. UTCI—why another thermal index? Int. J. Biometeorol. 56 (3), 421–428. https://doi.org/10.1007/s00484-011-0513-7.

Johnson, S., Ross, Z., Kheirbek, I., Ito, K., 2020. Characterization of intra-urban spatial variation in observed summer ambient temperature from the new York City Community air survey. Urban Clim. 31, 100583. https://doi.org/10.1016/j.uclim.2020.100583.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ. Sci. Technol. 53 (3), 1413–1421. https://doi.org/10.1021/acs.est.8b06038.

Kjellstrom, T., Kovats, R.S., Lloyd, S.J., Holt, T., Tol, R.S.J., 2009. The direct impact of climate change on regional labor productivity. Arch. Environ. Occup. Health 64 (4), 217–227. https://doi.org/10.1080/19338240903352776.

Kjellstrom, T., Briggs, D., Freyberg, C., Lemke, B., Otto, M., Hyatt, O., 2016. Heat, human performance, and occupational health: a key issue for the assessment of global climate change impacts. Annu. Rev. Public Health 37 (1), 97–112. https://doi.org/10.1146/annurev-publhealth-032315-021740.

Konarska, J., Holmer, B., Lindberg, F., Thorsson, S., 2016. Influence of vegetation and building geometry on the spatial variations of air temperature and cooling rates in a high-latitude city. Int. J. Climatol. 36 (5), 2379–2395. https://doi.org/10.1002/joc.4502.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26 (12), 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004.

Liljegren, J.C., Carhart, R.A., Lawday, P., Tschopp, S., Sharp, R., 2008. Modeling the wet bulb globe temperature using standard meteorological measurements. J. Occup. Environ. Hyg. 5 (10), 645–655. https://doi.org/10.1080/15459620802310770.

Lorenz, R., Stalhandske, Z., Fischer, E.M., 2019. Detection of a climate change signal in extreme heat, heat stress, and cold in Europe from observations. Geophys. Res. Lett. 46 (14), 8363–8374. https://doi.org/10.1029/2019GL082062.

Martilli, A., Krayenhoff, E.S., Nazarian, N., 2020. Is the urban heat Island intensity relevant for heat mitigation studies? Urban Clim. 31, 100541. https://doi.org/10.1016/j.uclim.2019.100541.

Meier, F., Fenner, D., Grassmann, T., Otto, M., Scherer, D., 2017. Crowdsourcing air temperature from citizen weather stations for urban climate research. Urban Clim. 19, 170–191. https://doi.org/10.1016/j.uclim.2017.01.006.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7 (Jun), 983–999.

MeteoSwiss, 2019. Klimabulletin Juni 2019 (Zürich).

Miao, C., Yu, S., Hu, Y., Zhang, H., He, X., Chen, W., 2019. Review of methods used to estimate the sky view factor in urban street canyons. Build. Environ. 106497. https://doi.org/10.1016/j.buildenv.2019.106497.

Napoly, A., Grassmann, T., Meier, F., Fenner, D., 2018. Development and application of a statistically-based quality control for crowdsourced air temperature data. Front. Earth Sci. 6. https://doi.org/10.3389/feart.2018.00118.

Obradovich, N., Migliorini, R., Mednick, S.C., Fowler, J.H., 2017. Nighttime temperature and human sleep loss in a changing climate. Sci. Adv. 3 (5), e1601555.

Obradovich, N., Migliorini, R., Paulus, M.P., Rahwan, I., 2018. Empirical evidence of mental health risks posed by climate change. Proc. Natl. Acad. Sci. 115 (43), 10953–10958. https://doi.org/10.1073/pnas.1801528115.

Oke, T.R., 1982. The energetic basis of the urban heat island. Q. J. R. Meteorol. Soc. 108 (455), 1–24. https://doi.org/10.1002/qj.49710845502.

Oke, T.R., Mills, G., Christen, A., Voogt, J.A., 2017. Urban climates. Cambridge University Press.

Parastatidis, D., Mitraka, Z., Chrysoulakis, N., Abrams, M., 2017. Online global land surface temperature estimation from Landsat. Remote Sens. 9 (12), 1208. https://doi.org/10.3390/rs9121208.

Rameezdeen, R., Elmualim, A., 2017. The impact of heat waves on occurrence and severity of construction accidents. Int. J. Environ. Res. Public Health 14 (1), 70. https://doi.org/10.3390/ijerph14010070.

Schatz, J., Kucharik, C.J., 2014. Seasonality of the urban Heat Island effect in Madison, Wisconsin. J. Appl. Meteorol. Climatol. 53 (10), 2371–2386. https://doi.org/10.1175/JAMC-D-14-0107.1.

Segev, N., Harel, M., Mannor, S., Crammer, K., El-Yaniv, R., 2017. Learn on source, refine on target: a model transfer learning framework with random forests. IEEE Trans. Pattern Anal. Mach. Intell. 39 (9), 1811–1824. https://doi.org/10.1109/TPAMI.2016.2618118.

Shen, Y., Cao, J., Wang, J., Yang, Z., 2019. Urban acoustic classification based on deep feature transfer learning. J. Frankl. Inst. https://doi.org/10.1016/j.jfranklin.2019.10.014.

Shi, Y., Katzschner, L., Ng, E., 2018. Modelling the fine-scale spatiotemporal pattern of urban heat island effect using land use regression approach in a megacity. Sci. Total Environ. 618, 891–904. https://doi.org/10.1016/j.scitotenv.2017.08.252.

Sugawara, H., Hagishima, A., Narita, K., Ogawa, H., Yamano, M., 2008. Temperature and wind distribution in an E-W-oriented urban street canyon. SOLA 4, 53–56. https://doi.org/10.2151/sola.2008-014.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2018. Springer International Publishing, pp. 270–279. https://doi.org/10.1007/978-3-030-01424-7_27.

Tewari, M., Yang, J., Kusaka, H., Salamanca, F., Watson, C., Treinish, L., 2019. Interaction of urban heat islands and heat waves under current and future climate conditions and their mitigation using green and cool roofs in new York City and Phoenix, Arizona. Environ. Res. Lett. 14 (3), 034002 https://doi.org/10.1088/1748-9326/aaf431.

Tibshirani, J., Athey, S., Friedberg, S., Hadad, V., Miner, L., Wager, S., Wright, M., 2019. grf: Generalized Random Forests (R Package Version 0.10.3) [Computer Software]. https://CRAN.R-project.org/package=grf.

Toparlar, Y., Blocken, B., Maiheu, B., van Heijst, G.J.F., 2017. A review on the CFD analysis of urban microclimate. Renew. Sust. Energ. Rev. 80, 1613–1640. https://doi.org/10.1016/j.rser.2017.05.248.

Voelkel, J., Shandas, V., 2017. Towards systematic prediction of urban heat Islands: grounding measurements, assessing modeling techniques. Climate 5 (2), 41. https://doi.org/10.3390/cli5020041.

Voogt, J.A., Oke, T.R., 2003. Thermal remote sensing of urban climates. Remote Sens. Environ. 86 (3), 370–384.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. 113 (523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839.

Weier, J., Herring, D., 2000. Measuring Vegetation (NDVI & EVI). NASA Earth Observatory. https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_1.php.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Stat. 39 (1), 151–160. https://doi.org/10.1080/02664763.2011.578621.