Science of the Total Environment 928 (2024) 172454

Contents lists available at ScienceDirect

Science of the Total Environment

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

Modelling Europe-wide fine resolution daily ambient temperature for 2003–2020 using machine learning

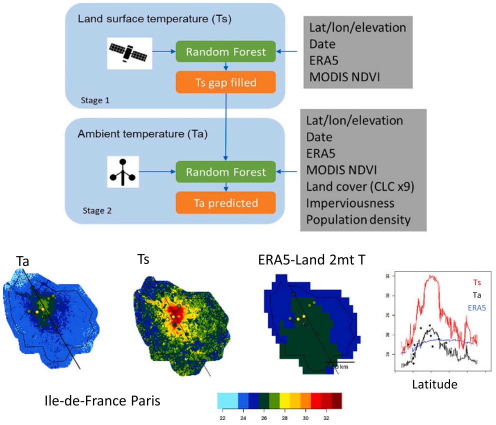
Alonso Bussalleu [a,b,*], Gerard Hoek [c], Itai Kloog [d], Nicole Probst-Hensch [a,b], Martin Röösli [a,b], Kees de Hoogh [a,b]

[a] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Allschwil, Switzerland
[b] University of Basel, Basel, Switzerland
[c] Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands
[d] Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer-Sheva, Israel

HIGHLIGHTS

- Block validation consistently showed lower performance compared to random validation.
- Local performance showed more variability than global performance but a similar mean.
- Performance in areas with fewer weather stations showed more variability and error.
- Models capture the same seasonal patterns as weather stations.
- Models' increased resolution capture within city temperature contrasts.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Kai Zhang

Keywords:
Ambient temperature
Remote sensing
MODIS LST
Random forest
Block cross validation

ABSTRACT

To improve our understanding of the health impacts of high and low temperatures, epidemiological studies require spatiotemporally resolved ambient temperature (Ta) surfaces. Exposure assessment over various European cities for multi-cohort studies requires high resolution and harmonized exposures over larger spatiotemporal extents. Our aim was to develop daily mean, minimum and maximum ambient temperature surfaces with a 1 × 1 km resolution for Europe for the 2003–2020 period. We used a two-stage random forest modelling approach. Random forest was used to (1) impute missing satellite derived Land Surface Temperature (LST) using vegetation and weather variables and to (2) use the gap-filled LST together with land use and meteorological variables to model spatial and temporal variation in Ta measured at weather stations. To assess performance, we validated these models using random and block validation. In addition to global performance, and to assess model stability, we reported model performance at a higher granularity (local). Globally, our models explained on average more than 81 % and 93 % of the variability in the block validation sets for LST and Ta respectively.

[*] Corresponding author at: Swiss Tropical and Public Health Institute, Kreuzstrasse 2, 4123 Allschwil Basel, Switzerland.
E-mail address: Alonso.bussalleu@unibas.ch (A. Bussalleu).

https://doi.org/10.1016/j.scitotenv.2024.172454
Received 20 December 2023; Received in revised form 4 April 2024; Accepted 11 April 2024
Available online 16 April 2024
0048-9697/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Average RMSE was 1.3, 1.9 and 1.7 °C for mean, min and max ambient temperature respectively, indicating a generally good performance. For Ta models, local performance was stable across most of the spatiotemporal extent, but showed lower performance in areas with low observation density. Overall, model stability and performance were lower when using block validation compared to random validation. The presented models will facilitate harmonized high-resolution exposure assignment for multi-cohort studies at a European scale.

1. Introduction

Research investigating associations between temperature and health primarily focuses on time series studies linking daily ambient temperature (Ta) to citywide daily mortality counts or hospital admissions (Baccini et al., 2008; Gasparrini et al., 2015; Wu et al., 2022a; Wu et al., 2022b; Zhao et al., 2021). These studies commonly assign exposure at the city level using data of one or more weather stations. Although weather stations can accurately measure temperature over time, they are sparse, unevenly distributed and often located outside the major cities. Monitoring data thus lack the spatial representativeness to capture small scale contrasts and typically underestimate the variation in ambient temperature across urban areas during warm periods (Kloog, 2019). Furthermore, when studying long-term temperature effects on health, temporal changes in within city contrast are important and require resolved high-resolution temperature surfaces (Ganzleben and Kazmierczak, 2020; Hart and Sailor, 2008; Heaviside et al., 2016; Macintyre et al., 2018).

Climate reanalysis models offer a standardized approach to assign exposure over large areas or in areas without weather stations and are able to capture similar citywide temperature-mortality associations as station data (Masselot et al., 2023; Mistry et al., 2022). However, currently these global models, like for example ECMWF ERA5-land, are too coarse (resolution 10 × 10 km) to capture within city contrasts (Masselot et al., 2023). Remotely sensed surface temperature has been used directly as an alternative to characterize exposure (Chakraborty et al., 2020; Hsu et al., 2021), but it tends to overestimate the UHI effect (Azevedo et al., 2016; Venter et al., 2021) and Ta-LST differences can be as large as 20 °C and are modulated by elevation, hour of the day, seasonality and land cover characteristics (Pepin et al., 2016). Surface temperature data can however be extremely useful when combined with other land-use data and air temperature observations.

Statistical models are commonly used in environmental epidemiological studies to create high-resolution and spatiotemporally resolved environmental exposure maps (Hoek, 2017; Kloog, 2019; Kloog and Zhang, 2023). These empirical models combine environmental measurements with predictor variables such as large-scale meteorological data, land use inventories and satellite observations measured at different resolutions (Kloog, 2019).

The Land Surface Temperature (LST) product from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument onboard the AQUA and TERRA satellites has been used to inform empirical temperature models due to its global coverage, frequent overpass time (twice daily) and fine spatial resolution (1 × 1 km) (Flückiger et al., 2022; Hough et al., 2020; Kloog et al., 2017; Rosenfeld et al., 2017; Zhou et al., 2020). MODIS LST is affected by weather conditions (i.e. cloudiness), leaving gaps on its record. Some recent national Ta models in Israel (Zhou et al., 2020), Switzerland (Flückiger et al., 2022) and Sweden (Jin et al., 2022) have used a 2-stage approach where missing LST is imputed using weather reanalysis predictors and other land use variables to produce a gap-filled LST surface. The gap-filled LST is then used to calibrate the Ta ~ LST association at all weather station locations. Temperature models that include satellite predictors have successfully being used in studies investigating associations between temperature and mortality (Lee et al., 2016; Shi et al., 2015), birth outcomes (Kloog et al., 2015) and to calibrate health warning systems (Ragettli et al., 2023).

While many gridded ambient temperature products exist (De Ridder et al., 2015; Fick and Hijmans, 2017; Haylock et al., 2008; Kilibarda et al., 2014; Verdin et al., 2020: Zhang et al., 2022a, Zhang et al., 2022b), to our knowledge none of the available products cover Europe at a daily 1 × 1 km resolution for the 2003–2020 period using a two-stage hybrid modelling approach.

This study addresses the need for high spatiotemporal continental temperature models to facilitate harmonized exposure assessment in large multi-cohort studies investigating associations of short- and long-term exposure to temperature and health across Europe. Here we aim to develop Europe-wide daily 1 × 1 km resolution models for mean, minimum and maximum ambient temperature for the 2003–2020 period. This work was performed in the framework of the EXPANSE (EXposome Powered tools for healthy living in urbAN SEttings) project (Vlaanderen et al., 2021). EXPANSE aims to evaluate the association of the urban exposome with cardiometabolic and pulmonary health for more than 55 million Europeans.

2. Methods

We estimated daily ambient temperatures (Ta) across Europe at a 1 × 1 km spatial resolution for 18 years between 2003 and 2020 using a two-stage approach. In stage 1, we created daily gap-filled LST surfaces using a range of meteorological predictor variables and in stage 2 we modelled daily mean, minimum and maximum ambient air temperature from weather stations using LST from stage 1 plus additional meteorological and land use variables.

In both stages, we used a random forest regression (RF) model and tested the robustness of the models using customized validation approaches. We evaluated performance for the model in the full spatiotemporal domain (global) and for smaller extents (local).

2.1. Study area

The study area includes the European Union countries plus Iceland, United Kingdom, Switzerland, Norway and the Balkan countries (43 countries in total, Fig. A.1). Excluding the water masses, this area extends for approximately 5′037′854 km^2.

Europe can be divided into subtropical, temperate, cold and circumpolar climate groups which are further influenced by the proximity to the coastlines, latitude and longitude (maritime, transitional/intermediate, continental and polar/subpolar) (Pinborg and Larsson, 2002). Areas with a high degree of urbanization are Belgium, the Netherlands and West Germany (Rhine-Ruhr) and the cities of Paris, Madrid, Berlin and Milano (eurostat, 2022).

2.2. Ambient temperature

We obtained daily mean, minimum and maximum temperature measurements collected by weather station networks from 2 publicly available data repositories: The European Climate Assessment & Dataset (ECA&D) (Klein Tank, 2002) and the Global Surface Summary of the Day (GSOD) (NOAA, 2021). ECAD compiles and harmonizes data from 85 partner institutions in 65 countries across Europe while GSOD uses worldwide data collected by the United States commerce and defense departments. We also included the observations from two national monitoring networks: The Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) (MeteoSwiss, 2021); and The Czech Hydrometeorological Institute (CHMI) (CHMI, 2021), to complement the data in these countries. After quality control and harmonization daily Ta observations were linked to a 1 × 1 km grid cell at the European Terrestrial Reference System 1989 projection (EPSG:3035, here EU GRID). More information about each dataset and the steps carried out to harmonize the four datasets can be found in the supplementary information (appendix B).

2.3. Spatiotemporal predictor variables

2.3.1. Land surface temperature (LST)

We used the day and night LST observations from the MODIS instrument onboard the AQUA (MYD11A1v061) and TERRA (MOD11A1v061) satellites (*myd_day*, *myd_night*, *mod_day*, *mod_night* respectively, Table A.2). Both satellites have a circular sun-synchronous polar orbit completed every 99 min (14 orbits per day). During each orbit, MODIS gathers data from a 2300 km wide swath. Terra crosses the equator at 10:30 am/pm with a descending orbit, while Aqua at 1:30 am/pm with an ascending orbit, thus compared to Aqua, Terra typically measures warmer temperatures during the night and colder temperatures during the day. MYD11A1v061 and MOD11A1v061 have calibration changes and polarization corrections that improve validation compared to previous versions. We filtered the granules within the area of interest for the selected period using EARTHDATA SEARCH (https://search.earthdata.nasa.gov/search) and downloaded a list with the links to all selected granules, which were downloaded using R. We considered as valid observations all grid cells that were produced with an average LST error ≤2 K as reported on the quality indicator bitmask following Hough (Hough et al., 2020). MODIS products have a sinusoidal coordinate system (SR-ORG:6842).

Following Kloog et al. (2017), Zhou et al. (2020) and Flückiger et al. (2022), we included the following predictor variables (Table A.2), which have shown to help explain spatial and temporal variations of LST and Ta:

2.3.2. Normalized Difference Vegetation Index (NDVI)

We use monthly 1 × 1 km NDVI from the TERRA satellite (MOD13A3v061) as a predictor variable for the stage 1 and stage 2 models. We filtered grid cells that fell in land, coastlines & shorelines and shallow inland water bodies. The filtering was based on the product's quality indicator bitmask. While NDVI data availability was almost complete for the model's extent (the average depth of observations per grid cell and the mean daily spatial coverage of the study area were 99 % and 98 % respectively, Table A.3), there was a clear seasonal and latitudinal pattern as all grid cells located at higher latitudes had missing information and on average less valid observations, during winter months (Figs. A.1 & A.4). Missing NDVI values occur because of misclassification of certain land areas as water and sensor malfunction (saturation in ice/snow covered areas). Most gaps occur in winter, at higher latitudes or in mountainous regions. Higher latitudes presented one single large gap that covered the north of Norway, Sweden and Finland every year during January. Gaps in mountainous regions were smaller and appear randomly. The extent of these gaps changed yearly.

2.3.3. Weather data (ERA5)

We extracted Boundary layer height, Total cloud cover, Skin temperature, 2-meter air temperature, Soil temperature level 1, wind speed (calculated form U & V 10 m wind components), Total precipitation and Surface pressure from the "ERA5 hourly data on single levels from 1940 to present" (ERA5) dataset. ERA5 is the global climate reanalysis model from the European Centre from Medium-Range Weather Forecasts (ECMWF) with a horizontal resolution of 0.25 × 0.25° (~27 × 27 km). Hourly data was first aggregated into daily averages and later resampled to the LST modelling reference grid (SR-ORG:6842) using bilinear interpolation. Due to the lack of certain variables (boundary layer height and total cloud cover) and missing coverage in coastal regions, we decided to use the coarser, but complete ERA5 product instead of ERA5-land higher resolution (~9 × 9 km) products. We considered coastal areas important as a large portion of the European population lives in coastal regions.

2.3.4. Land use variables

We represented land use using the Corine Land Cover inventory (CLC), the imperviousness product (IMD) from the Copernicus programme, and population density (POP) by the Gridded Population of World model version4 (GPWv4) (Table A.2). For CLC we aggregated land use classes into urban fabric (I), industrial infrastructure (II), barren areas (III), urban vegetation (IV), agricultural areas (V), natural areas (VI), snow & ice (VII), wetlands (VIII), and water (IX). CLC and GPWv4 datasets are available at Google Earth Engine (GEE) and IMD images were downloaded from the Copernicus Land Monitoring service (https://land.copernicus.eu/).

2.3.5. Elevation (ELV)

We extracted and processed elevation (ELV) from the Global 30 Arc-Second Elevation product (GTOPO30) available in GEE. To match the analysis resolution, this product was resampled to the LST model reference grid using bilinear interpolation.

2.3.6. Linking LST & Ta models reference grids

We extracted LST, NDVI, weather data, and elevation and produced the gap-filled LST surfaces using the MODIS sinusoidal projection (SR-ORG:6842). To align these surfaces with the EU GRID used in stage 2 (Ta modelling), we linked the cells from both reference grids based on proximity using nearest neighbor interpolation.

A complete table with information about predictor variables can be found in the supplementary information (Table A.2).

2.4. Statistical methods

We followed a two-stage modelling approach based on random forest (RF) similar to Zhou et al. (2020) and Flückiger et al. (2022). The first stage aims at creating gap-free daily surfaces for four MODIS LST products, whereas the second stage aims at creating daily *Tmean*, *Tmin* and *Tmax* surfaces. We first discuss the common statistical issues for stage 1 and stage 2 modelling (Algorithm, tuning and variable selection), then discuss stage 1 and stage 2 analyses in more detail and finally we present the model validation strategy.

2.4.1. Random forest algorithm

RF (Breiman, 2001) is able to handle non-linearity and interactions and can accommodate correlated predictors without compromising model performance, making it an efficient algorithm (Belgiu and Drăguţ, 2016). RF has also showed good predictive performance compare to other methods (Chen et al., 2019; Li et al., 2011; Liu et al., 2022; Noi et al., 2017).

RF requires hyperparameter tuning as performance is affected by the number of variables use to split a tree node (mtry) and the number of trees to grow (ntree) (Genuer et al., 2010). As the number of (noisy) predictor variables increases, a higher mtry is required, whereas ntree should increase until results are stable. Including correlated predictors limits model interpretability as the signal is diluted between the correlated predictors and their individual variable importance decreases (Genuer et al., 2010). Furthermore, by increasing the size of the dataset, the number of trees or the number of predictors has exponential penalties on computational time and memory requirements. For these reasons, we avoided including highly correlated predictors in the LST and TA models. Moreover, in case of spatiotemporal models, the sample should also be balanced across the model's spatial and temporal extent (unbiased) and the training set should be representative of the area of interest. We used R 'ranger' package (Wright and Ziegler, 2017) to perform the RF regressions in both stages.

We test the effect of the number of trees and the number of variables per split for stage 1 (ntrees: 100, 200, 300; mtry: 4, 5, 6, 7) and stage 2 (ntrees: 100, 300, 500; mtry: 4, 7, 12, 16).

For stage 1 and due to the large volume of data, we trained individual RF models for smaller temporal extents following Zhou et al., 2020, who trained daily LST models for every month in Israel for the same reason. In addition, we used only a fraction of the available data to train the models. To balance data density across the extent and avoid overfitting models to data rich areas, we followed a sampling scheme stratified by date and block. We tested multiple sampling proportions (1 %, 5 %, 25 % and 50 % of land grid cells by block/date) and modelling extents (in days: 10, 30, 90, 180, 365).

For stage 2 we tested different modelling extents (14 months starting in December, one calendar year or 12 months starting in July), the effect of representing date as Julian day or a pair of sines and cosines and the effect of only including day or night LST products when calculating LSTmean and LSTvar (see Section 2.4.4).

Table C.5 and Figs. C.6–C.8, and Table C.9 and Figs. C.10–C.12 show tested configurations, and their performance for stage 1 and stage 2 respectively.

2.4.2. Variable selection

To exclude highly correlated variables from model training we used the variance inflation factor (VIF, see appendix C.2. for formula and calculation). VIF is a measure of collinearity as it calculates how much of the variance of a given predictor variable can be explained by the other predictors assuming a linear relationship. We sequentially filter out the predictor with the highest VIF until VIF $\leq$ 10 for all predictors. Specifically, in stage 1 we excluded soil and 2 m temperature due to its high correlation with skin temperature (Fig. C.3); and in stage 2 we excluded all three temperature variables from ERA5 due to their high correlation with the much higher resolution gapfilled LST from stage 1 (Fig. C.4).

2.4.3. Stage 1: LST modelling

Cloud cover limits satellite retrieval, creating gaps in the daily LST surfaces. The size and location of these gaps is influenced by seasonality, local weather patterns and local satellite overpass time. In order to impute the missing values and create gap-free surfaces, we use RF regression to model the relationship between LST and location (latitude, longitude and elevation), date (Julian date), NDVI and weather variables from ERA5 (temperature, cloud cover, boundary layer height, precipitation and wind speed). We trained models separately for each of the four LST products. These models were later used to predict missing LST values.

Final model configuration was monthly models trained with 1 % of the possible pixels with 100 trees and 4 variables per split. Stage 1 models for each of the four LST products are represented with Eq. (1):

$$\mathrm{LST_{ij}} \sim \mathrm{RF}\left(\mathrm{LAT_i}, \mathrm{LON_i}, \mathrm{YDAY_j}, \mathrm{ELV_i}, \mathrm{NDVI_{ij}}, \mathrm{ERA5_{CC\,ij}}, \mathrm{ERA5_{blh_{ij}}}, \mathrm{ERA5_{SKT_{ij}}}, \mathrm{ERA5_{WS\,ij}}, \mathrm{ERA5_{PR_{ij}}}, \mathrm{ERA5_{SP_{ij}}}\right) \tag{1}$$

where LST_{ij} represents the LST value in degrees Kelvin for grid cell i and day j of one of the four LST products, LAT_i, LON_i, ELV_i represent the latitude, longitude and elevation (in meters) of grid cell i centroid and $YDAY_j$ represents the Julian date for date j. Additional predictors include, monthly NDVI ($NDVI_{ij}$) for grid cell i and date j and ERA5 cloud cover ($ERA5_{cc\,ij}$, in percentages), boundary layer heights ($ERA5_{BLH\,ij}$, in meters), skin temperature ($ERA5_{SKT\,ij}$, in degrees Kelvin, which represents the temperature of earth's surface), wind speed ($ERA5_{WS\,ij}$, in m/s), precipitation ($ERA5_{PR\,ij}$, in meters) and surface pressure ($ERA5_{sp_{ij}}$, in Pa) for grid cell i and date j.

Final models to fill LST gaps were trained using the sampled pixels from all the blocks, thus covering the whole extent.

Grid cells with missing NDVI values (only affecting on average less than 0.1 % of grid cells) could not be imputed using our modelling approach and were filled using focal averages from all surrounding grid cells in a 5 km buffer (120 cells, window size = 11). We were unable to fill the larger gap on the northernmost latitudes without increasing window size, thus the modelling extent becomes smaller during some winter months (Fig. A.1).

2.4.4. Stage 2: Ta modelling

Ta weather stations are unevenly distributed in space, showing a higher density of stations in for example Germany (Fig. A.1). The number of stations increased over time and not all stations were active during the whole period (Table B.5). Moreover, weather station's location within cities is not random, with airports and parks being common areas for sensor deployment (Kloog, 2019). In stage 2 our aim therefore was to explain the variation in daily Tmax, Tmean and Tmin observations from weather station networks using the predictor variables LSTmean, LSTvar, NDVI, latitude, longitude Julian date and weather variables. The developed yearly RF models were used to predict ambient temperature across our whole study area particularly in grids without observations.

To improve stability, better capture daily variability, limit the effect of local observation time during satellite overpass and avoid using highly correlated predictors, we reduced the four gap-free daily MODIS LST surfaces produced in stage 1 into daily average and a daily variance surfaces (LSTmean, LSTvar). We hypothesize that together; LSTmean and LSTvar can provide additional information about the daily variability in temperature observed in the four LST products and help calibrate the LST-Ta relationship.

Final models configurations used the observations from the same calendar year, 100 trees and 7 variables per split.

After removing highly correlated predictors, the Stage 2 models are represented with Eq. (2):

$$\begin{aligned}\mathrm{TA}(min, \mathrm{mean}, max)_{\mathrm{ij}} \sim \mathrm{RF}\big(&\mathrm{LAT_i}, \mathrm{LON_i}, \mathrm{ELV_i}, \mathrm{YDAY_j}, \mathrm{NDVI_{ij}}, \mathrm{LSTmean_{ij}}, \mathrm{LSTvar_{ij}}, \mathrm{ERA5cc_{ij}}, \mathrm{ERA5blh_{ij}}, \mathrm{ERA5ws_{ij}}, \mathrm{ERA5pr_{ij}}, \mathrm{ERA5sp_{ij}}, \mathrm{IMD_{ij}}, \\ &\mathrm{CLCurban_{fabric_{ij}}}, \mathrm{CLCindustrial_{ij}}, \mathrm{CLCbarren_{ij}}, \mathrm{CLCurban_{green_{ij}}}, \mathrm{CLCagriculture_{ij}}, \mathrm{CLCsnow_{ice_{ij}}}, \mathrm{CLCwetlands_{ij}}, \mathrm{CLCwater_{ij}}, \mathrm{POP_{ij}}\big)\end{aligned} \tag{2}$$

where $TA(min, mean, max)_{ij}$ represents the observed average Ta value in degrees Celsius within grid cell i and date j for each temperature set, LAT_i, LON_i, ELV_i the latitude longitude and mean elevation (meters) of grid cell *i*, $YDAY_j$ the Julian date for *day j*, $NDVI_{ij}, LSTmean_{ij}, LSTvar_{ij}$ the monthly NDVI and daily LSTmean and LSTvar for grid cell *i* and date *j*. ERA5 variables were the same as in stage 1 (excluding skin temperature due to a high correlation with LSTmean). Yearly land use for each grid cell *i* and date *j* was represented by impervious surface density (IMD_{ij})

(0–100 %), population density (POP_{ij}) and the area (as proportion 0–1) covered by urban fabric ($CLCurban\_fabric_{ij}$), industrial infrastructure ($CLCindustrial_{ij}$), barren areas ($CLCbarren_{ij}$), urban vegetation ($CLCurban\_green_{ij}$), agriculture fields ($CLCagriculture_{ij}$), snow & ice ($CLCsnow_{ij}$), wetlands ($CLCwetlands\_ij$) and water ($CLCwater_{ij}$).

Final models to predict daily 1 × 1 km Tmean, Tmin and Tmax were trained using all the observations from the available weather stations.

2.4.5. Model validation

We used two validation approaches, block validation and random validation, likely resulting in underestimation and overestimation of performance respectively.

Model validation requires training and validation sets to be independent, which can be attained by using different observations for model training and validation as long as sampling is at random, balanced and representative (Wadoux et al., 2021). Validation methods for spatiotemporal models should also account for the data autocorrelation structure as close observations might not be statistically independent (Belgiu and Drăguţ, 2016; Hengl et al., 2018).

To meet validation requirements and better understand the effect of data density and distribution on model performance, we used random and block cross validation (random-CV and block-CV respectively). Random-CV splits data in equally sized folds that are sequentially left out from model training and are used for model validation. Random-CV assumes that validations sets are independent and ideally distributed and it does not account for the data autocorrelation structure. Observations are grouped into folds at random, thus observations used by model training can be near to observations used for model validation and high information density areas will also be overrepresented in the training and validation sets (overfitting).

In contrast, for block-CV observations are grouped based on spatial and temporal proximity across the model extent into blocks. Similar to random-CV, spatiotemporal blocks are aggregated into folds for model validation and training. In block-CV observations within validation blocks are distant from those used for training, thus validation blocks are spatiotemporally “independent” from the training set. For this reason, block-CV can provide more realistic performance indicators for models trained with uneven data distributions. Following Roberts (Roberts et al., 2017), our target for the spatiotemporal extent of CV blocks was to mimic the size of real data gaps (stage 1) or the extent of the autocorrelation structure (stage 2). By using spatiotemporal blocks for block-CV, the training set will include observations from the entire spatiotemporal extent.

To define the spatial and temporal extent of CV blocks, we analyzed the data dependence structure by visually inspecting variograms (following Roberts et al., 2017). More information on the calculation can be found in Appendix C.1.

For stage 1, we evaluated model performance using a combination of external validation (similar to random-CV) and 5-fold block-CV. After the stratified sampling, sampled pixels were aggregated into blocks, and blocks were randomly assigned into five folds, each with approximately 20 % of the blocks. For external validation (random), we sampled random pixels within the blocks used for model training, so that for each block and date we had two equally sized groups of different pixels, one for training and one for validation. For block-CV, we validated the trained model with the fold left out for validation which only had observations in blocks that were not sampled for model training (spatially and temporally independent). Thus, pixel samples from each fold and from each block were used four times for model training and external (random) validation, and one time for block-CV.

For stage 2, we evaluated model performance using 10-fold random-CV and 10-fold block-CV as explained above, with each fold for block-CV and for random-CV holding 10 % of the blocks and 10 % of the Ta observations respectively. Due to the number and clustered distribution of grid cells with valid Ta observations, we increased folds to 10 to maximize the number of observations for model training and avoid extrapolation following (Roberts et al., 2017).

2.4.6. Global & local performance

Validation results were presented at two scales: *local* performance was obtained by calculating the performance for each spatiotemporal block fold wise and averaging fold results within blocks; *global* performance was obtained by calculating performance fold wise and taking the mean (traditional approach). Local performance can provide discrete information about model stability across its spatiotemporal extent, and when averaged, we can compare mean local performance (with each block weighting the same independently of the number of observations or area) to global model performance (unweighted).

Performance statistics were reported as the mean absolute error (MAE), the root mean square error (RMSE) and the coefficient of determination (R^2) for the whole model (global) and aggregated at the spatiotemporal block scale (local) for both Random-CV and Block-CV.

We calculated R^2 using the following formula:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (X_i - Xm_i)^2}{\sum (X_i - \widehat{X})^2} \tag{3}$$

where the residual sum of squares (RSS) is calculated as the squared difference between observed values (X_i) and modelled values (Xm_i), and the total sum of squares (TSS) is calculated as the squared difference between observed values and the mean of the observations within the validation set ($\widehat{X}$ represents the average of observations within a fold for global performance, and the average of the observations within a spatiotemporal block and fold for local performance). The comparison of global and local performance is primarily based upon RMSE and MAE and less on R^2. The reason is that there is lower variance in measurements in local areas compared to the global domain.

Due to the clustered distribution of LST and Ta observations, and to improve interpretation while limiting the effect of reporting local performance in blocks with very few observations, we avoid calculating local performance in uninformative areas using information sufficiency thresholds. For stage 1, we only evaluated local performance in blocks with more than 0.5 % of all land pixels or more than 100 pixels available for validation. For stage 2 we filter out all local fold values calculated with less than 10 sampled observations, we assigned NA to all local R^2 folds showing infinite values (no variability within observations), and the value of −1 to all local R^2 values showing values bellow −1 (2 * TSS < RSS), and only calculated local performance in blocks with more than 28 observations.

3. Results

3.1. LST and weather station data availability, temperature distribution and definition of spatiotemporal blocks

Valid LST measurements were not evenly distributed across space or time, or between day and night products (Table A.3, Figs. A.4–A.5). However, within day and night products, data availability showed the same trends. Across the spatial extent, each grid cell had on average 40 % and 20 % of valid daily information for day LST and night LST products respectively (Table A.3 & Fig. A.5). However, on average, only 20 % of the grid cells across the spatial extent had valid LST measurements for each day (Table A.3 & Fig. A.4). During some days, there were no valid LST measurements across the whole extent for day and night products from both satellites (Fig. A.4).

Overall, the mean daily LST range was of approximately 51 °C for day products and 42 °C for night products (Table A.6, Fig. A.7).

Figs. C.18 and C.19 show semivariograms of respectively the temporal and spatial extend of the dependence structure in the LST data set. As the variance stabilizes around 10 day difference and 500 km distance, we use this information to create 500 × 500 km blocks with a temporal depth of 10 days (Fig. A.1). In total more than 33,300 spatiotemporal blocks were used to cover the whole spatiotemporal extent (50 spatial blocks * 37 temporal blocks per year), of which more than 99 % contained valid observations. On average spatiotemporal blocks had approximately 20 % of valid observations, but valid observations for any given spatiotemporal block could be as low as 0.01 % (Fig. A.8).

The number of Ta stations and, thus the number of grid cells with valid Ta measurements increased gradually over the years by approximately 1000 stations between 2003 and 2020, with an average of 4626, 4474 and 4634 per year for Tmean, Tmin and Tmax respectively (Table B.5). Station density was highest in the center of the modelling extent (Germany, Switzerland, Czech Republic and Austria; Figs. A.1, E.1). The number of years that stations remain active also showed variability as some stations stop recording and other were added (Fig. E.1).

Figs. C.20 and C.21 show semivariograms of respectively the temporal and spatial extend of the dependence structure in the weather station data set. Spatial and temporal variance would increase with increasing distances, but the steeper increases were found between the pairs of stations located within 50 to 100 km apart and during the first lag days. Based on these findings and considering the clustered Ta station distribution, we created 125 × 125 km blocks with a temporal depth of 30 days. In total 120,510 spatiotemporal blocks were used to cover the entire spatiotemporal extent (515 spatial blocks * 13 temporal blocks per year). Although very few blocks did not include any valid Ta observations, the distribution of stations (and of observations) was strongly clustered with 10 % of blocks having only one station, 50 % having 5 or less, 90 % having 21 or less and the richer 10 % of blocks having from 21 up to 184 stations (Fig. E.1).

Ta daily range was on average 40.7 °C, 39.04 °C and 40.33 °C for Tmean, Tmin and Tmax respectively (Table B.5). The lowest and highest recorded temperatures were −50 °C and 49.9 °C. However, these were very extreme temperatures as the minimum 1 % and maximum 99 % percentiles were −24.06 °C and 35 °C respectively (Table B.5). Mean year temperature was on average 8.5 °C, 4.4 °C and 12.7 °C for Tmean, Tmin and Tmax respectively (Table B.5).

3.2. *Stage 1 - LST modelling*

We modelled daily LST across Europe for the period 2003–2020 for Aqua and Terra satellites day and night products. These models imputed missing LST data, thus creating gap filled surfaces. However, imputed values represent LST values under cloud-free conditions and validation only occurred in cloud-free pixels. Thus, we report the validation accuracy of estimated LST of clear-sky pixels only.

Table 1 shows aggregated global and local performance indicators for each validation strategy and LST product. The average R^2 for all four products at the global resolution was higher compared to that at the local resolution for block-CV (global block-CV mean 0.847, local block-CV mean 0.498) and at a lesser extent for random (external) validation (0.954 v 0.834). For RMSE and MAE this trend was not observed, suggesting similar performance of the model at global and local scale and the overall spatiotemporal stability of models.

No clear differences were observed in mean R^2s between night and day products (0.779 v 0.787). On the contrary, on average, a lower RMSE (2.057 °C v 2.505 °C) and MAE (1.487 °C v 1.858 °C) were found when comparing night and day products.

As expected, block-CV (mean overall R^2 0.672; RMSE 2.923 °C; MAE 2.213 °C) consistently showed less optimistic results and more variability than external random validation (mean overall R^2 0.894; RMSE 1.639 °C; MAE 1.132 °C).

Local performance informed about the models' stability across its spatiotemporal extent (Fig. 1) and showed a wider distribution in all cases, especially in the block-CV R^2, where local performance for some blocks was lower and even negative (Table 1, Figs. D.2 & D.3). Despite the magnitude differences in RMSE and MAE between validation strategies, the spatiotemporal patterns in local performance remained similar between validation strategies for all LST products (Figs. 1 & D.1).

Seasonality, but not annual differences played an important role at understanding performance variability at the local and global resolution (Figs. D.5–D.6). Performance remain stable across years independently of the resolution, validation strategy and the performance indicator (Figs. D5–D6). During warmer periods, RMSE increased for daily LST products independently of the resolution and validation strategy (Figs. 1, D.2 & D.5). For night products, random RMSE remain relatively stable between months, but block-CV RMSE decreased during warmer months (Figs. D.4 & D.5). In contrast, R^2 remained relatively stable for night products across validations strategies and resolutions (Fig. D.6) decreasing only for random-CV for daily products during warmer months (Figs. D2, D.3 & D.6).

High latitude blocks, especially those encompassing Iceland, Norway and northern Britain and Ireland showed lower performance in all models independently of the product compared to blocks in continental Europe (excluding the Alpine region). The Alpine region showed higher error compared to the surrounding blocks in night product models, especially during colder months. Southern blocks' performance was lower than that of central Europe blocks for day products, especially during warmer months (Figs. 1 & D.1–D.4).

Table 1

Model performance by product, validation strategy and resolution for Land Surface Temperature models (LST - stage 1) for three common performance indicators (MAE = mean absolute error, R^2 = coefficient of determination (variance explained), RMSE = root mean square error). Local performance reports the average of all spatiotemporal block values and its distribution (p10-p50-p90). Global performance reports the average between all monthly (models' temporal extent) model values and its distribution. Spatiotemporal blocks that did not meet the minimum information requirements (0.5 % available data and 100+ observations) were not included in the calculations for local performance. MAE and RMSE are expressed in degrees Celsius.

Product	Validation strategy	Resolution	MAE (°C)	R^2	RMSE (°C)
mod_day	Block-CV	Local	2.4 (1.8–2.3–3.1)	0.505 (0.244–0.533–0.736)	3.1 (2.3–3.0–4.0)
		Global	2.3 (2.1–2.3–2.5)	0.866 (0.819–0.874–0.90)	3.1 (2.8–3.1–3.3)
	External (random)	Local	1.3 (0.7–1.3–1.8)	0.829 (0.709–0.843–0.932)	1.8 (1.1–1.8–2.5)
		Global	1.2 (0.9–1.2–1.6)	0.956 (0.925–0.961–0.977)	1.7 (1.4–1.7–2.2)
mod_night	Block-CV	Local	2.0 (1.3–1.9–2.9)	0.505 (0.229–0.53–0.757)	2.6 (1.7–2.5–3.8)
		Global	2.0 (1.7–1.9–2.4)	0.833 (0.793–0.837–0.866)	2.7 (2.3–2.6–3.2)
	External (random)	Local	1.0 (0.6–0.8–1.5)	0.855 (0.746–0.876–0.938)	1.4 (0.9–1.3–2.1)
		Global	0.9 (0.8–0.9–0.9)	0.957 (0.947–0.957–0.968)	1.4 (1.3–1.3–1.5)
myd_day	Block-CV	Local	2.5 (1.8–2.4–3.2)	0.506 (0.243–0.534–0.74)	3.2 (2.4–3.2–4.2)
		Global	2.4 (2.1–2.4–2.7)	0.872 (0.823–0.879–0.906)	3.2 (2.9–3.2–3.5)
	External (random)	Local	1.4 (0.8–1.4–2.0)	0.812 (0.671–0.828–0.93)	2.0 (1.1–2.0–2.7)
		Global	1.3 (0.9–1.4–1.8)	0.953 (0.918–0.959–0.978)	1.9 (1.4–1.9–2.4)
myd_night	Block-CV	Local	2.1 (1.4–2.0–2.9)	0.474 (0.216–0.495–0.714)	2.7 (1.9–2.6–3.8)
		Global	2.0 (1.7–2.0–2.5)	0.817 (0.777–0.821–0.849)	2.7 (2.3–2.7–3.3)
	External (random)	Local	1.0 (0.7–0.9–1.5)	0.838 (0.727–0.856–0.927)	1.5 (1.0–1.4–2.2)
		Global	0.9 (0.9–0.9–1.0)	0.951 (0.939–0.951–0.963)	1.4 (1.3–1.4–1.5)

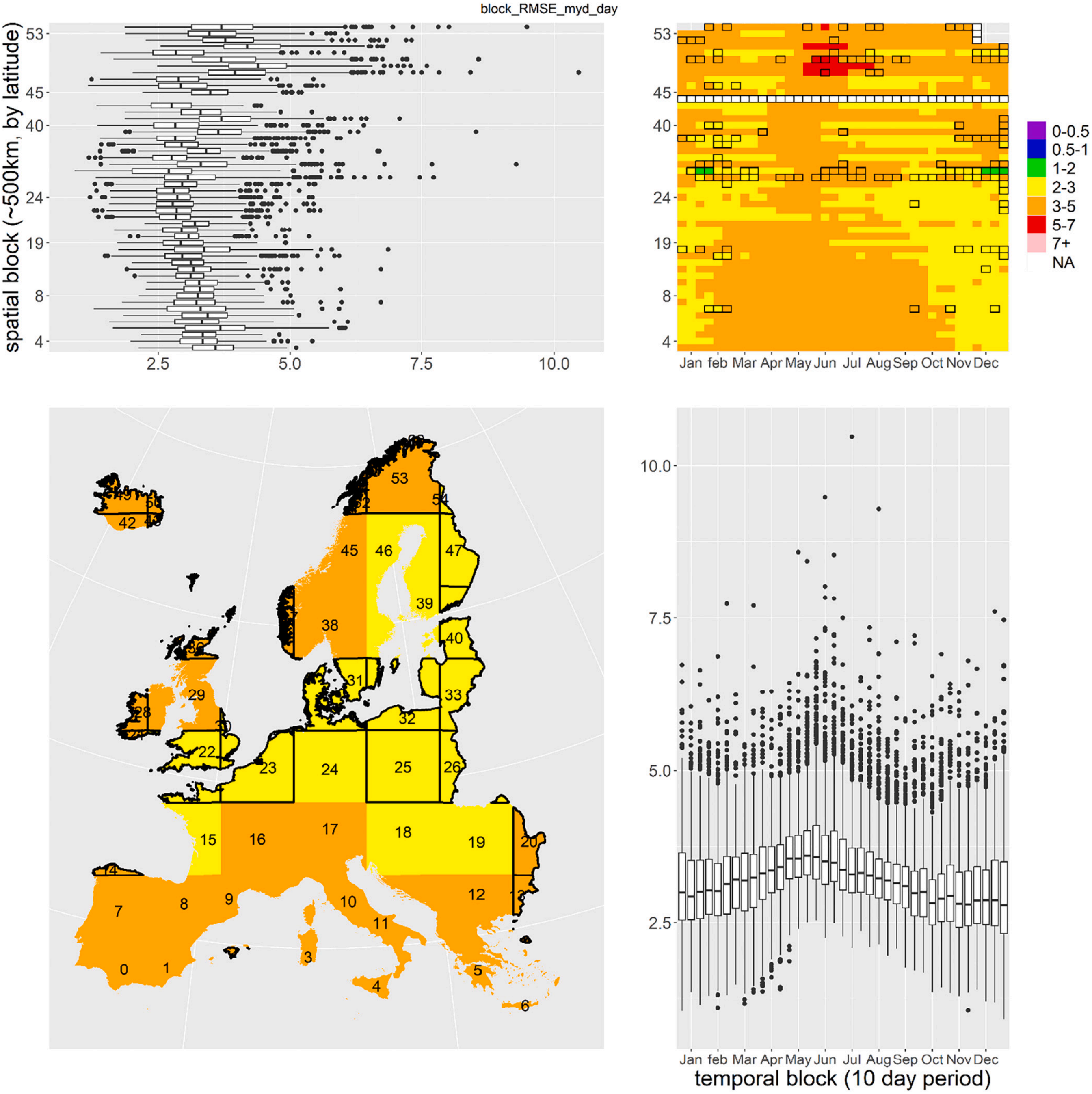


Fig. 1. Local block-CV RMSE for the AQUA satellite day product (myd_day). Upper left and bottom right panels shows the RMSE distribution by spatial block (~500 kmˆ2) and temporal block (~10 day) respectively, allowing visualization of average trends and extreme values. The upper right panel shows the complex patterns in RMSE by aggregating yearly values by spatiotemporal block by taking the average. The bottom left panel represents mean performance by spatial block. Highlighted blocks and areas represent areas where the minimum information requirements were not met at least once (0.5 % available data and at least 100 observations). These values were excluded from the calculations in all panels. Similar panels for other the other products and performance indicators can be found in the supplementary material (Figs. D.1–D3).

Variable importance showed that ERA5 Skin temperature, latitude, NDVI and ERA5 Cloud coverage were on average the most important predictors of LST across all products with some seasonal variability (Fig. D.7).

The resulting four daily 1 × 1 km gap-filled MODIS-LST products that were included in the Ta modelling stage as LSTmean and LSTvar.

3.3. Stage 2 – TA modelling

Table 2 shows the mean MAE, R^2 and RMSE and the distribution of these local and global performance indicators for each validation strategy and product. Even with the strictest validation strategy (local block cross validation), models indicated an overall good performance (Table 2). Overall Tmean models performed the best followed by Tmax and Tmin. Compared to LST, we found more variability in performance between years (Fig. E.2) than between months (Fig. E.3). As with stage 1, performance indicators showed on average a lower overall R^2 locally ($R^2 = 0.7$) compared to the overall global R^2 estimate ($R^2 = 0.96$). Block-CV showed the higher mean MAE and RMSE performance and lower explained variability independently of the product or the resolution

Table 2
Model performance by product, validation strategy and resolution for ambient temperature models (TA - stage 2) for three common performance indicators. Local performance reports the average of all spatiotemporal block values and its distribution (p10-p50-p90). Global performance reports the average between all yearly model values and its distribution. Following Eq. (1) R^2 can take negative values when model performance is lower than the local average. To limit the effect of very small sample sizes on local R^2 and improve interpretability, we filter out all local fold values calculated with less than 10 sampled observations, we assigned NA to all local R^2 folds showing infinite values, the value of −1 to all local R^2 values showing values bellow −1, and only calculated local performance in blocks with more than 28 observations. MAE and RMSE are expressed in degrees Celsius.

Product	Validation strategy	Resolution	MAE (°C)	R^2	RMSE (°C)
Tmean	Block-CV	Global	1.1 (1.1–1.1–1.1)	0.971 (0.966–0.971–0.974)	1.5 (1.4–1.5–1.5)
		Local	1.1 (0.7–1–1.5)	0.736 (0.49–0.799–0.917)	1.4 (0.9–1.3–1.9)
	Random-CV	Global	0.9 (0.9–0.9–0.9)	0.979 (0.976–0.979–0.982)	1.2 (1.2–1.2–1.3)
		Local	0.9 (0.6–0.9–1.3)	0.792 (0.591–0.844–0.938)	1.2 (0.8–1.1–1.6)
Tmin	Block-CV	Global	1.5 (1.5–1.5–1.6)	0.936 (0.927–0.937–0.944)	2 (2–2–2.1)
		Local	1.6 (1.1–1.5–2.1)	0.594 (0.307–0.651–0.83)	2 (1.4–1.9–2.7)
	Random-CV	Global	1.3 (1.3–1.3–1.4)	0.952 (0.946–0.953–0.959)	1.8 (1.7–1.8–1.8)
		Local	1.4 (1–1.3–1.8)	0.666 (0.415–0.717–0.87)	1.7 (1.2–1.7–2.3)
Tmax	Block-CV	Global	1.4 (1.4–1.4–1.5)	0.96 (0.957–0.961–0.964)	1.9 (1.9–1.9–2)
		Local	1.4 (0.9–1.3–2)	0.678 (0.391–0.747–0.896)	1.8 (1.2–1.7–2.5)
	Random-CV	Global	1.2 (1.2–1.2–1.2)	0.972 (0.969–0.972–0.975)	1.6 (1.6–1.6–1.7)
		Local	1.2 (0.8–1.2–1.7)	0.747 (0.508–0.806–0.923)	1.6 (1–1.5–2.2)

(Table 2). On average, RMSE and MAE were similar for local versus global validation, with smaller differences between validation strategies compared to stage1 (Tables 1 & 2), thus suggesting that overall models were stable across their spatiotemporal extent. Global R^2 showed a modestly higher random-CV performance compared that of block-CV.

Local block-CV RMSE showed strong spatial patterns that follow European geographic features (Fig. 2). Specifically, average spatial performance was consistently lower in coastal and mountain regions in all Ta models (Figs. 2, E.4–E.10).

Overall, LSTmean was the strongest predictor of Ta in all models followed by Julian date, NDVI and LSTvar and latitude (Fig. E.11). Land use and weather variables showed a small contribution to variable importance for all models (Fig. E.11).

To illustrate the models' predictions, Fig. 3 shows local daily variability in mean temperature for five European cities during a warm day (2017-06-19). Due to its higher spatial resolution, the Ta and LST models developed are better at capturing within city and urban-rural Ta contrasts compared to ERA5-land 2mt products and weather station data (Fig. 3). This increased resolution allow researchers to map UHI and investigate its effects on health, while also capturing regional patterns in Ta (Fig. 4).

Fig. 4 shows regional and city wide contrasts in mean temperature and temporal contrasts in daily Tmean for summer 2017. Modelled area weighted daily Tmean captures the same temporal variability than weather stations (Fig. 4 central panels), but can also inform about long-term temperature contrasts over urban areas. Tmean models clearly show prominent geographical features such as mountain ranges and large river valleys, but also urban heat islands and local contrasts within city limits (Fig. 4 left and right panels).

4. Discussion

This paper describes the model development of Europe-wide daily mean, minimum and maximum temperature at 1 × 1 km resolution for the 2003–2020 period. We used random forest in a 2 stage modelling framework to firstly gap fill missing data in satellite derived LST and secondly to model ground based Ta combining LST with weather station data and other spatiotemporal predictor variables. Overall, the models showed a good performance at both the global and local scale.

4.1. *Comparison with previous temperature models*

Our LST imputation models (stage 1) show a similar performance to LST models used in other studies using RF (Noi et al., 2017; Xiao et al., 2021) or other approaches (Shiff et al., 2021; Zhang et al., 2022b).

Our stage 2 models showed similar performance as studies using comparable modelling techniques to estimate daily Ta. Several studies developed models to estimate mean Ta in France, Northeastern USA, Southeastern USA, Israel and Sweden with reported random-CV RMSE ranging between 1.52 °C and 2.16 °C (Kloog et al., 2017; Kloog et al., 2014; Shi et al., 2016; Zhou et al., 2020; Jin et al., 2022; Rosenfeld et al., 2017). Hough et al. (2020), Flückiger et al. (2022) and Nikolaou et al. (2023) included Tmin and Tmax models in addition to Tmean for France, Switzerland and Germany respectively; with a random-CV RMSE ranging between 1.03 °C and 1.89 °C.

Some models estimate Ta using three stages (Kloog et al., 2014; Shi et al., 2016; Kloog et al., 2017; Rosenfeld et al., 2017; Hough et al., 2020; Nikolaou et al., 2023). The first stage calibrates the Ta ~ LST relationship in grid cells where both weather station and satellite observations are available. The second stage predicts Ta were LST is available. In the third stage, grid cells without Ta or LST observations are filled using the relationship between the predicted Ta grid cells and the surrounding Ta data (from weather stations within a buffer or using inverse distance weighting) and additional covariates. Even though the LST record is incomplete and the available data is clustered, the major advantage of remotely sensed MODIS LST is its coverage and depth (up to four observations per day). Our results also showed the overall good performance of our four LST models (Table 1), the ability of LST to capture small scale temperature contrasts (Figs. 3 & 4). Gap-filling LST using weather reanalysis data strengthens Ta modelling as it leverages the available data directly by allowing to keep more Ta records for calibrating the Ta ~ LST relationship and for model validation (I); and indirectly by using the relative abundance of LST observations to create resolved and informative Ta predictors independent of the weather station distribution (II). Furthermore it reduces the spatiotemporal bias towards clear sky conditions when calibrating the Ta ~ LST relationship. While stage 1 was able to produce gap-free LST surfaces, LST models were biased towards clear-sky conditions. Thus, gap-filled LST values might represent surface temperature less accurately than LST observations.

Due to its high temporal coverage (18 years) and spatial (1 × 1 km) resolution, our models and the resulting temperature surfaces allow the investigation of long-term health effects of temperature while taking into account within city variation. Temperature surfaces can also be aggregated over space and/or time to match health data at different resolutions and can provided improved population weighted exposure estimates at the city level for investigating the effect of heatwaves on health.

By increasing the spatial resolution, from ~9x9km (ERA5-Land) to 1 × 1 km (our models), we were able to capture within city temperature variability and thus the UHI effect (Fig. 3). At the city scale, LSTmean showed more variability and larger temperature contrasts compared to

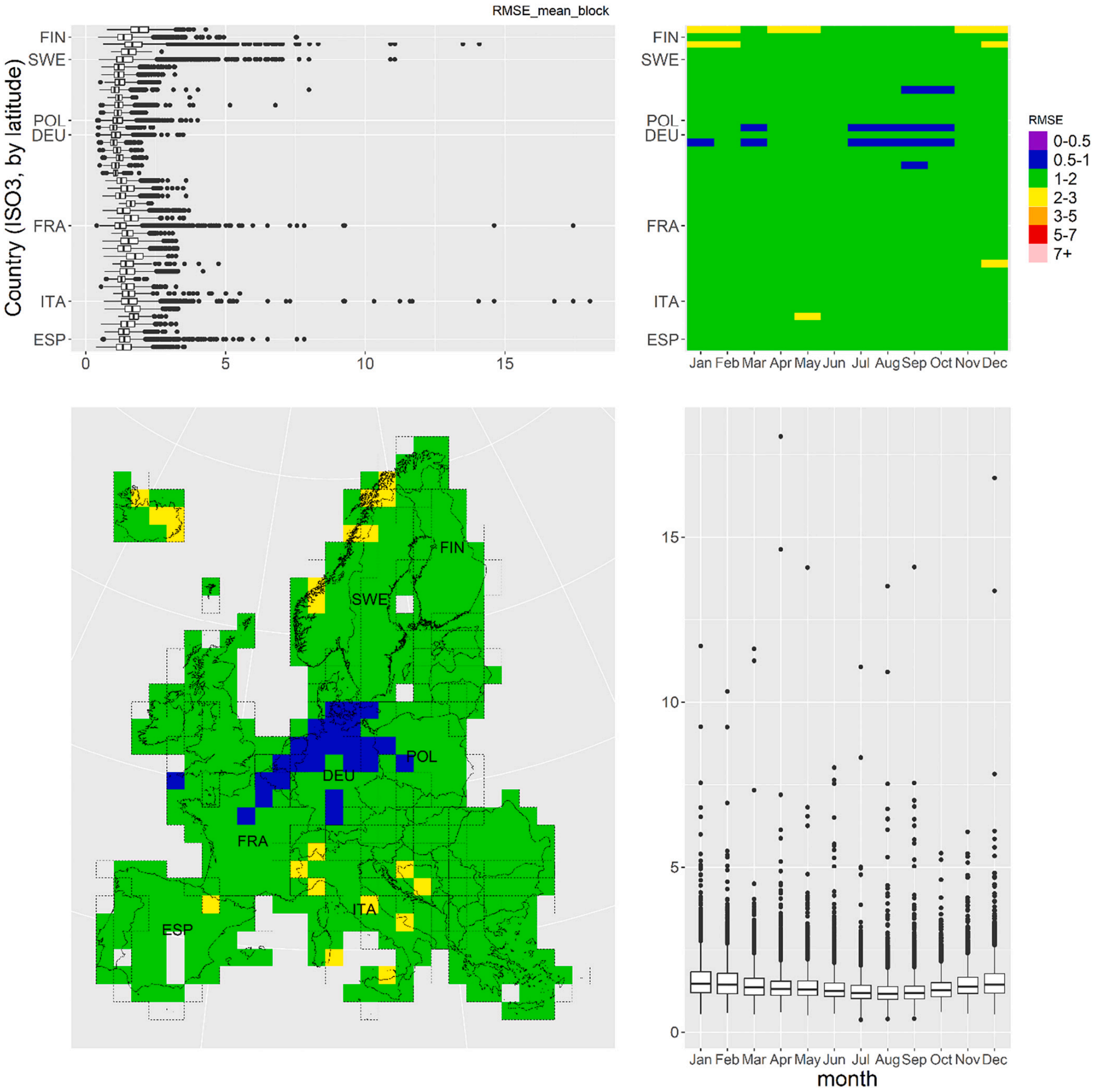


Fig. 2. Local block-CV RMSE for the average temperature models (Tmean). Upper left and bottom right panels shows the RMSE distribution by country and month (~ temporal block) respectively, being able to visualize average trends and extreme values. The upper right panel shows the complex patterns in RMSE by aggregating yearly values by country and month by taking the average. The bottom left panel represents mean RMSE by spatial block (~125 kmˆ2). Shared blocks between countries (dashed lines) were included for each country average. Similar panels for other the other products and performance indicators can be found in the supplementary material.

Tmean (Fig. 3). LST has been found to overestimate the UHI effect (Azevedo et al., 2016; Venter et al., 2021), thus agreeing with our findings and justifying our approach to use LST only as one of the predictor variables in our stage 2 Ta model.

4.2. Model structure

Variable importance showed that ERA5 skin temperature is a good predictor in the stage 1 models and that LSTmean and LSTvar were important predictors of Ta in stage 2. Although Flückiger et al. (2022) used LSTmean to model Ta for Switzerland, we argue that adding LSTvar provides a more complete representation of daily temperatures. Our models showed that land use variables ranked low in importance for all Ta models while variables that represent seasonal and regional patterns like NDVI, date and latitude had the highest importance after including LST. Furthermore, land use variables are stable over time and do not explain any seasonal variation. As the modelled spatiotemporal extent is large and RF relies on minimizing the variance at each split (node) for model fitting and calculating variable importance, predictors that affect regional contrast in the response will drive model calibration. In contrast, there is more room for interventions at the local scale and information regarding the drivers of local contrast can have an important

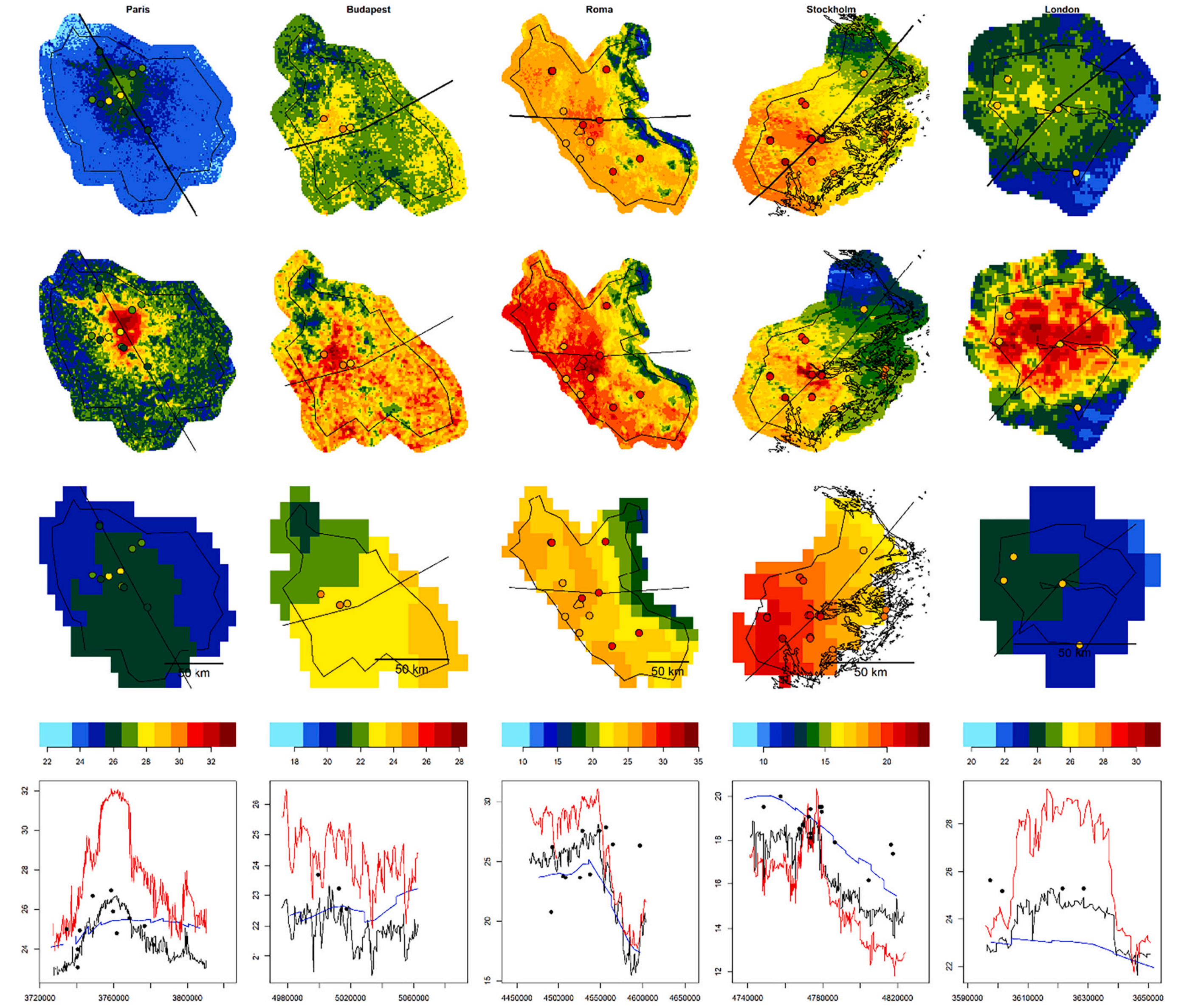


Fig. 3. Spatial Contrast in daily Temperature for five European cities (from left to right: Paris, Budapest, Roma, Stockholm, London) for June 19th 2017. First row: modelled Tmean; second row: Average of the four modelled LST surfaces (LST_mean); third row: ERA5 land 2 m Temperature, fourth row: Transect values for Tmean (black), LST_mean (red), ERA5 land 2mt (blue) and weather station values (black). X axis represents longitude. Points within maps represent the location of weather stations and the recorded average ambient temperature.

effect on policy. While LSTmean and LSTvar are likely to incorporate some information regarding land use characteristics, we argue that in spite of its limited global importance, land use predictors should be kept in Ta models as they represent the mechanisms that drive temperature pattern in urban areas. Due to the clustered distribution of Ta stations very few areas had the station density to inform about local Ta contrast related to land use characteristics. While high density station clusters can provide an accurate representation of local contrast in Ta and of predictor space, these differences are often modulated by regional weather patterns and might not be extrapolated. Future research should provide more accurate descriptors of variable importance at the local scale and make a better use of high-density station clusters by an additional layer of local models. As models increase the spatial resolution to better capture small-scale temperature contrast within urban environments, discussion should also focus on whether the used weather station networks accurately cover the model's extent at such resolution.

4.3. Validation strategies

With an increasing spatiotemporal extent of the study area, finding datasets that accurately represent the area and period of interest becomes more difficult. Moreover, traditional random-CV assumes that the training and validation sets are independent and representative and provides biased performance indicators otherwise, specially under cluster distributions (de Bruin et al., 2022; Wadoux et al., 2021). Evidence suggest that due to the data autocorrelation structure close observations might not be independent (Roberts et al., 2017), and Meyer and Pebesma (2022) showed that a probable cause of overoptimistic performance when using random-CV is that samples used for training are clustered and very distant from the prediction areas. Ta weather station networks are designed to calibrate meteorological models and/or to provide local weather information at for example airports leading to an uneven spatial distribution. Remote sensed LST is biased towards clear sky conditions and clear sky conditions are linked to weather patterns that also affect temperature towards both extremes (heat during

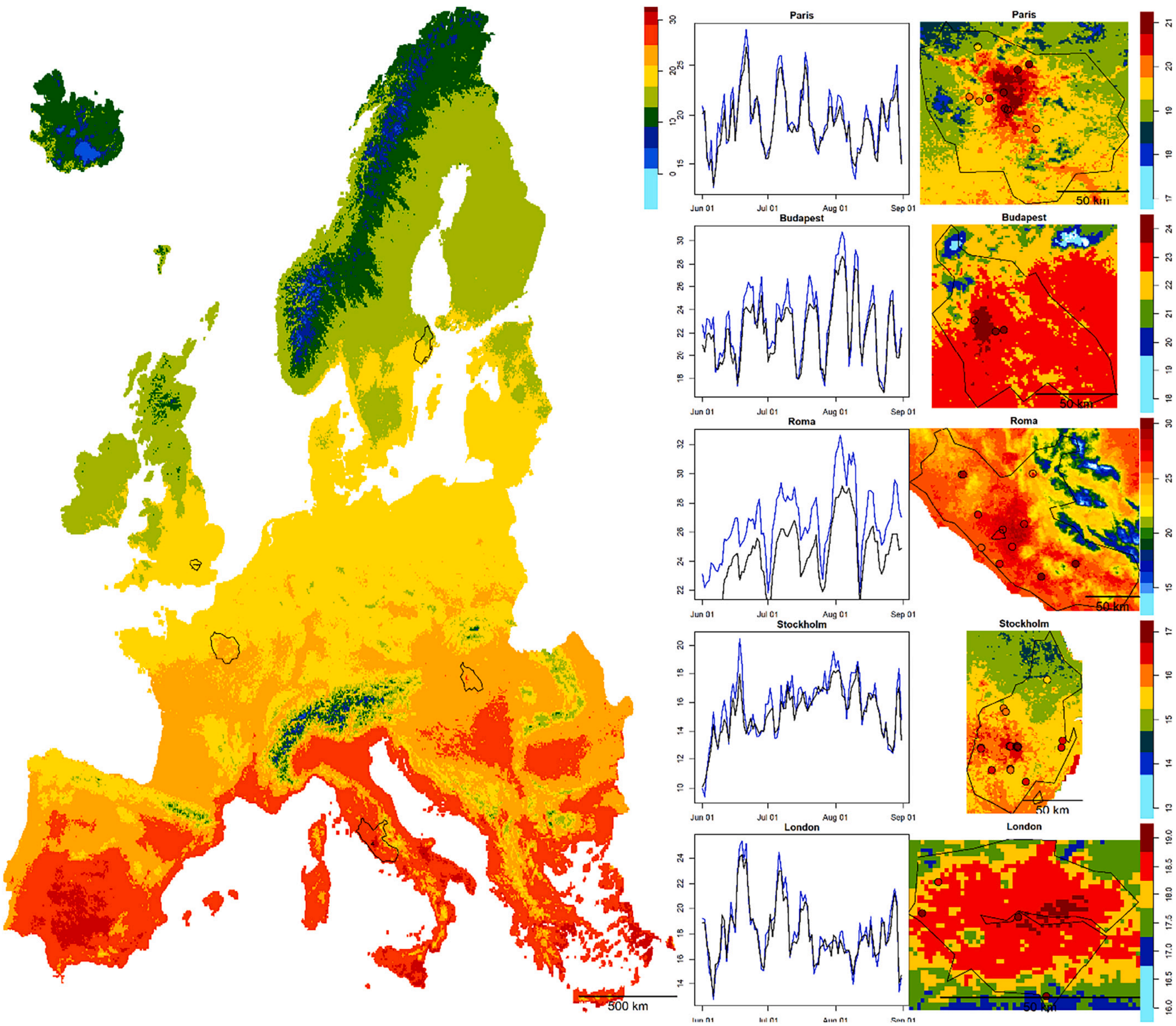


Fig. 4. Summer temperature 2017. Left panel: Average Summer Temperature (JJA) from Tmean surfaces for year 2017; Center panel: Daily temperature for five European cities (from top to bottom: Paris, Budapest, Roma, Stockholm, London) for the period June–August 2017. Daily area-aggregated average from Tmean models (black) and weather stations daily averages (blue); Right panel: Average Summer Temperature from daily Tmean surfaces for year 2017 in 5 European cities.

day and minimum temperature in the night). In our case, both initial LST and Ta datasets showed a strongly clustered distribution (Figs. A.5 & E.1). Improving training set representativeness for large scale Ta models could be achieved by improving the accessibility to national and private weather monitoring networks, using crowdsourced meteorological stations (Fenner et al., 2021; Venter et al., 2021) and using alternative remote sensing products (Zhu et al., 2017). However, for the time being, researchers can still provide less biased performance estimators by balancing observation density through sampling or applying weights, or use alternative validation approaches (Meyer and Pebesma, 2021; Meyer and Pebesma, 2022; Sarafian et al., 2021; de Bruin et al., 2022; Roberts et al., 2017). De Bruin et al. (2022) and Wadoux et al. (2021) showed that, although block-CV overestimates error under balanced and representative sample distributions, under clustered and gapped sample distributions random-CV underestimates error and Block-CV is the least bias validation approach. We argue that random-CV might reflect model performance better in areas with high data density whereas block-CV better reflects performance in areas with little or no information, and that using block and random CV together provides additional information to manage model expectations while also allowing for model comparisons.

4.4. *Global versus local model performance*

Global R^2 values are sensitive to the spatiotemporal extent of a model. As the spatiotemporal extent increases the central tendency is less representative of particular regions and provides less valuable information on how the model is performing locally (Meyer and Pebesma, 2021; Meyer and Pebesma, 2022). Local performance consistently showed less optimistic values than global performance independently of the validation strategy in both stages. We also found that block-CV shows higher variability in local performance and a lower central tendency compared to random-CV (Table 2). Meyer and Pebesma (2022) report striking differences between global and local performance and Wadoux et al. (2021) and de Bruin et al. (2022) report the overall negative effects of clustering on modelled map accuracy. While we report important differences between mean local R2 and global R2, and differences in local RMSE and MAE across the spatiotemporal extent, it is important to highlight that average local RMSE and MAE were very close to those reported globally (Tables 1 & 2). These findings indicate that, on average, models perform well at the local level. While local block-CV showed increased variability and decreased performance compared to global validation and random-CV, it also shows that models were stable across most of the spatiotemporal extent with a local block-CV RMSE lower than 3 °C for approximately 96 % of spatiotemporal blocks included in the calculations, and a local block-CV R^2 higher than 0.5 for approximately 82 % (mean = 89.525, min = 75.079, max = 83.746) of the spatiotemporal blocks included in the calculations. Local performance was able to clearly point areas and periods were the model had limited applicability, such as Poland and Iceland for Tmax models or Iceland for Tmin models (Figs. E.4–E.10). Thus, the use of multiple validation strategies and of local performance estimators improved model understanding while providing realistic expectations and explicitly addressed issues in accuracy assessment that can damage scientific credibility if left unchecked (Roberts et al., 2017; Meyer and Pebesma, 2022). Furthermore, we did not detect an increase in local block-CV performance with increasing local data density or sampled proportion in stage 1 (Fig. D.8). We believe that this difference is due to the stratified sampling of LST data in stage 1. We decided not to do this in stage 2 as the Ta station distribution was more clustered with very few blocks having enough stations to inform about local contrasts. Increased measurement error and bias in exposure-response estimates has been linked to decreasing sample sizes in simulation studies (Basagana et al., 2013). We found a similar trend between local performance and information density (Fig. E.12), suggesting better exposure estimates in data rich areas. Sensitivity of local and global model performance to local information density and clustering could also inform sampling strategies and help at visualizing overfitting (Crosetto et al., 2000). Here, areas with a high degree of urbanization were also data rich areas and had high local performance (Figs. 2, E.1, E.4–E.10), thus suggesting accurate exposure estimates for these densely populated areas (eurostat, 2022). Further research is required on how differences in local performance and model stability across the spatiotemporal extent affect exposure-response relationships over larger areas.

Wang et al. (2020) recently provided a theoretical framework for addressing spatial stratified heterogeneity, spatial autocorrelation and the distinction between population and sample in spatial modelling referred to as the spatial statistics trinity. Our sample is clearly not a random sample of the population, as some countries have a much larger sample density. Temperature data also exhibited spatial autocorrelation. We applied these concepts in the block-validation approach and showed that performance statistics were indeed modestly worse compared to random validation, ignoring these features. Spatial stratified heterogeneity is what we aimed at predicting with our GIS and satellite derived predictor variables. Future temperature modelling might benefit from a more explicit application of this framework as another method to characterize possible bias or uncertainty in the temperature estimates as a result of the uneven spatial distribution of weather station data, the dependent variable in our models.

4.5. Limitations

Due to missing predictors (boundary layer height and cloud cover) and uncompleted records near coastal regions in the higher resolution ERA5-land dataset (~9 km) we decided to use the coarser ERA5 dataset (~27 km). Being important predictors of LST, stage 1 modelling could potentially benefit from a complete record of weather predictors at a higher resolution and future research should consider using ERA5-Land data to gap fill LST. Exploratory analysis found that for 2 m temperature from both datasets (after projection to the 1 × 1 EU reference grid using bilinear interpolation), the distribution of daily average pixel wise differences was 0.64, 0.77 and 0.89 for the median and 75 and 95 percentiles respectively, with higher differences in colder periods. When aggregating daily differences by pixel, the distribution of average pixel-wise differences was 0.51, 0.66 and 1.28 for the median and the 75 and 95 percentiles respectively.

The shape and position of blocks were arbitrarily selected. For this reason, blocks located in coastal areas are smaller than those fully covered by land. Area differences further increases the variability in the number and density of observations per block. In addition, within a block, observations can be clustered in a smaller area following meteorological or land use patterns, affecting local representativeness. Moreover, block-CV can be influenced by the specific combination of blocks, used for model training. Future research using block-CV and local performance should consider running multiple replicas with different points of origin for drawing blocks, thus limiting the effect of specific training sets on block performance and decreasing the effect of block shape, area and position. Further improvements could include using clustering algorithms (i.e. k-means clustering) to create more organic spatiotemporal blocks and creating folds so that they are equally good representations of multivariate predictor space (Roberts et al., 2017).

Local R^2 can inform about spatiotemporal patterns in performance, but it was also sensitive to the data distribution, local data density and the size of the local validation set. We observed that the variability of local random-CV performance estimators between folds increased with decreasing local data density and spatiotemporal blocks with very small validation sets also showed unrealistically low R^2 values. Future research could consider increasing the number and density of weather stations, observation weighting or using areas of applicability following Sarafian et al. (2021) or Meyer and Pebesma (2021).

Finally, here we selected RF as our modelling algorithm a priori and focused on developing a customized and model agnostic validation strategy and aggregating performance at a higher resolution. Future research could apply the developed workflow to compare performance between different modelling algorithms.

5. Conclusion

This work provides Europe wide daily temperature models at 1 × 1 km resolution that will allow harmonizing exposure assessment for multi-cohort studies. Ta models showed overall good performance, even under strict validation strategies. Due to its improved spatial resolution compared to ERA5-land our Ta models capture local contrast and decrease exposure misclassification. Furthermore, mean, minimum and maximum ambient temperature surfaces were developed to provide different exposure estimates that can be related to health outcomes. Tmean surfaces had the highest performance and Tmin the lowest. While all Ta models showed spatiotemporal differences in performance, all models were relatively stable. We found that local performance in areas with low information density showed more variability and higher error. While denser Ta monitoring networks across larger areas would improve model accuracy and stability, our findings also support the need for local performance indicators and improved validation strategies, especially when modelling at large geographical extents.

CRediT authorship contribution statement

Alonso Bussalleu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Gerard Hoek:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Itai Kloog:** Writing – review & editing, Methodology, Conceptualization. **Nicole Probst-Hensch:** Writing – review & editing, Conceptualization. **Martin Röösli:** Writing – review & editing, Conceptualization. **Kees de Hoogh:** Writing – review & editing, Supervision, Methodology, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by the EXPANSE project. The EXPANSE project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 874627. The content of this article is not officially endorsed by the European Union. The authors declare no competing financial interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.172454.

References

Klein Tank, A.M.G., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment (ECA&D). https://www.ecad.eu.

Azevedo, J., Chapman, L., Muller, C., 2016. Quantifying the daytime and night-time urban heat island in Birmingham, UK: a comparison of satellite derived land surface temperature and high resolution air temperature observations. Remote Sens. (Basel) 8.

Baccini, M., Biggeri, A., Accetta, G., Kosatsky, T., Katsouyanni, K., Analitis, A., Anderson, H.R., Bisanti, L., D'Ippoliti, D., Danova, J., Forsberg, B., Medina, S., Paldy, A., Rabczenko, D., Schindler, C., Michelozzi, P., 2008. Heat effects on mortality in 15 European cities. Epidemiology 19, 711–719.

Basagana, X., Aguilera, I., Rivera, M., Agis, D., Foraster, M., Marrugat, J., Elosua, R., Kunzli, N., 2013. Measurement error in epidemiologic studies of air pollution based on land-use regression models. Am. J. Epidemiol. 178, 1342–1346.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31.

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

Chakraborty, T., Hsu, A., Manya, D., Sheriff, G., 2020. A spatially explicit surface urban heat island database for the United States: characterization, uncertainties, and possible applications. ISPRS J. Photogrammetry Remote Sensing 168, 74–88.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934.

CHMI, 2021. Daily Average, Mean and Max Air Temperatures [Online] [accessed. 12. 06. 2021]. Available from. https://www.chmi.cz/historicka-data/pocasi/zakladni-informace.

Crosetto, M., Tarantola, S., Saltelli, A., 2000. Sensitivity and uncertainty analysis in spatial modelling based on GIS. Agric. Ecosyst. Environ. 81, 71–79.

de Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A. M.J.C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Eco. Inform. 69.

De Ridder, K., Lauwaet, D., Maiheu, B., 2015. UrbClim – a fast urban boundary layer climate model. Urban Clim. 12, 21–48.

eurostat, 2022. Urban-rural Europe - introduction. In: Eurostat Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Urban-rural_Europe_-_introduction (Accessed 20 Jul 2023).

Fenner, D., Bechtel, B., Demuzere, M., Kittner, J., Meier, F., 2021. CrowdQC+—a quality-control for crowdsourced air-temperature observations enabling world-wide Urban climate applications. Front. Environ. Sci. 9.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315.

Flückiger, B., Kloog, I., Ragettli, M.S., Eeftens, M., Röösli, M., de Hoogh, K., 2022. Modelling daily air temperature at a fine spatial resolution dealing with challenging meteorological phenomena and topography in Switzerland. Int. J. Climatol. 42, 6413–6428.

Ganzleben, C., Kazmierczak, A., 2020. Leaving no one behind - understanding environmental inequality in Europe. Environ. Health 19, 57.

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., Forsberg, B., Leone, M., De Sario, M., Bell, M.L., Guo, Y.-L.L., Wu, C.-F., Kan, H., Yi, S.-M., de Sousa Zanotti Stagliorio Coelho, M., Saldiva, P.H.N., Honda, Y., Kim, H., Armstrong, B., 2015. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. Lancet 386, 369–375.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31, 2225–2236.

Hart, M.A., Sailor, D.J., 2008. Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. Theor. Appl. Climatol. 95, 397–406.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res. Atmos. 113.

Heaviside, C., Vardoulakis, S., Cai, X.M., 2016. Attribution of mortality to the urban heat island during heatwaves in the West Midlands, UK. Environ. Health 15 (Suppl. 1), 27.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Graler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Hoek, G., 2017. Methods for assessing long-term exposures to outdoor air pollutants. Curr. Environ. Health Rep. 4, 450–462.

Hough, I., Just, A.C., Zhou, B., Dorman, M., Lepeule, J., Kloog, I., 2020. A multi-resolution air temperature model for France from MODIS and Landsat thermal data. Environ. Res. 183, 109244.

Hsu, A., Sheriff, G., Chakraborty, T., Manya, D., 2021. Disproportionate exposure to urban heat island intensity across major US cities. Nat. Commun. 12, 2721.

Jin, Z., Ma, Y., Chu, L., Liu, Y., Dubrow, R., Chen, K., 2022. Predicting spatiotemporally-resolved mean air temperature over Sweden from satellite data using an ensemble model. Environ. Res. 204, 111960.

Kilibarda, M., Hengl, T., Heuvelink, G.B.M., Gräler, B., Pebesma, E., Perčec Tadić, M., Bajat, B., 2014. Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. J. Geophys. Res. Atmos. 119, 2294–2313.

Kloog, I., 2019. Use of earth observations for temperature exposure assessment in epidemiological studies. Curr. Opin. Pediatr. 31, 244–250.

Kloog, I., Zhang, X., 2023. Methods to advance climate science in respiratory health. Immunol. Allergy Clin. North Am. 44, 97–107.

Kloog, I., Nordio, F., Coull, B.A., Schwartz, J., 2014. Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. Remote Sens. Environ. 150, 132–139.

Kloog, I., Melly, S.J., Coull, B.A., Nordio, F., Schwartz, J.D., 2015. Using satellite-based spatiotemporal resolved air temperature exposure to study the association between ambient air temperature and birth outcomes in Massachusetts. Environ. Health Perspect. 123, 1053–1058.

Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., Schwartz, J., 2017. Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data. Int. J. Climatol. 37, 296–304.

Lee, M., Shi, L., Zanobetti, A., Schwartz, J.D., 2016. Study on the association between ambient temperature and mortality using spatially resolved exposure data. Environ. Res. 151, 610–617.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Software 26, 1647–1659.

Liu, T.-L., Flückiger, B., de Hoogh, K., 2022. A comparison of statistical and machine-learning approaches for spatiotemporal modeling of nitrogen dioxide across Switzerland. Atmos. Pollut. Res. 13.

Macintyre, H.L., Heaviside, C., Taylor, J., Picetti, R., Symonds, P., Cai, X.M., Vardoulakis, S., 2018. Assessing urban population vulnerability and environmental risks across an urban area during heatwaves - implications for health protection. Sci. Total Environ. 610-611, 678–690.

Masselot, P., Mistry, M., Vanoli, J., Schneider, R., Iungman, T., Garcia-Leon, D., Ciscar, J. C., Feyen, L., Orru, H., Urban, A., Breitner, S., Huber, V., Schneider, A., Samoli, E., Stafoggia, M., de'Donato, F., Rao, S., Armstrong, B., Nieuwenhuijsen, M., Vicedo-Cabrera, A.M., Gasparrini, A., Network, M.C.C.C.R., Project, E., 2023. Excess mortality attributed to heat and cold: a health impact assessment study in 854 cities in Europe. Lancet Planet Health 7, e271–e281.

MeteoSwiss, 2021. Air temperature 2 m above ground (Federal Office of Meteorology and Climatology MeteoSwiss). https://gate.meteoswiss.ch/idaweb/login.do.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633.

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. 13, 2208.

Mistry, M.N., Schneider, R., Masselot, P., Roye, D., Armstrong, B., Kysely, J., Orru, H., Sera, F., Tong, S., Lavigne, E., Urban, A., Madureira, J., Garcia-Leon, D., Ibarreta, D., Ciscar, J.C., Feyen, L., de Schrijver, E., de Sousa Zanotti Stagliorio Coelho, M., Pascal, M., Tobias, A., Guo, Y., Vicedo-Cabrera, A.M., Gasparrini, A., Multi-Country Multi-City Collaborative Research, N, 2022. Comparison of weather station and climate reanalysis data for modelling temperature-related mortality. Sci. Rep. 12, 5178.

Nikolaou, N., Dallavalle, M., Stafoggia, M., Bouwer, L.M., Peters, A., Chen, K., Wolf, K., Schneider, A., 2023. High-resolution spatiotemporal modeling of daily near-surface air temperature in Germany over the period 2000–2020. Environ. Res. 219, 115062.

NOAA, 2021. Global Surface Summary of the Day - GSOD. 1.0. 2003–2020. NOAA National Centers for Environmental Information (Accessed July 2021. gov.noaa.ncdc:C00516).

Noi, P., Degener, J., Kappas, M., 2017. Comparison of multiple linear regression, cubist regression, and random Forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. Remote Sens. (Basel) 9.

Pepin, N.C., Maeda, E.E., Williams, R., 2016. Use of remotely sensed land surface temperature as a proxy for air temperatures at high elevations: findings from a 5000 m elevational transect across Kilimanjaro. J. Geophys. Res. Atmos. 121.

Pinborg, U., Larsson, T.-B., 2002. Europe's Biodiversity - Biogeographical Regions and Seas. European Environment Agency (EEA). https://www.eea.europa.eu/publications/report_2002_0524_154909.

Ragettli, M.S., Saucy, A., Fluckiger, B., Vienneau, D., de Hoogh, K., Vicedo-Cabrera, A. M., Schindler, C., Roosli, M., 2023. Explorative assessment of the temperature-mortality association to support health-based heat-warning thresholds: a national case-crossover study in Switzerland. Int. J. Environ. Res. Public Health 20.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Rosenfeld, A., Dorman, M., Schwartz, J., Novack, V., Just, A.C., Kloog, I., 2017. Estimating daily minimum, maximum, and mean near surface air temperature using hybrid satellite models across Israel. Environ. Res. 159, 297–312.

Sarafian, R., Kloog, I., Sarafian, E., Hough, I., Rosenblatt, J.D., 2021. A domain adaptation approach for performance estimation of spatial predictions. IEEE Trans. Geosci. Remote Sens. 59, 5197–5205.

Shi, L., Kloog, I., Zanobetti, A., Liu, P., Schwartz, J.D., 2015. Impacts of temperature and its variability on mortality in New England. Nat. Clim. Chang. 5, 988–991.

Shi, L., Liu, P., Kloog, I., Lee, M., Kosheleva, A., Schwartz, J., 2016. Estimating daily air temperature across the southeastern United States using high-resolution satellite data: a statistical modeling study. Environ. Res. 146, 51–58.

Shiff, S., Helman, D., Lensky, I.M., 2021. Worldwide continuous gap-filled MODIS land surface temperature dataset. Sci. Data 8, 74.

Venter, Z.S., Chakraborty, T., Lee, X., 2021. Crowdsourced air temperatures contrast satellite measures of the urban heat island and its mechanisms. Sci. Adv. 7.

Verdin, A., Funk, C., Peterson, P., Landsfeld, M., Tuholske, C., Grace, K., 2020. Development and validation of the CHIRTS-daily quasi-global high-resolution daily temperature data set. Sci. Data 7, 303.

Vlaanderen, J., de Hoogh, K., Hoek, G., Peters, A., Probst-Hensch, N., Scalbert, A., Melen, E., Tonne, C., de Wit, G.A., Chadeau-Hyam, M., Katsouyanni, K., Esko, T., Jongsma, K.R., Vermeulen, R., 2021. Developing the building blocks to elucidate the impact of the urban exposome on cardiometabolic-pulmonary disease: the EU EXPANSE project. Environ. Epidemiol. 5, e162.

Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457.

Wang, J., Gao, B., Stein, A., 2020. The spatial statistic trinity: a generic framework for spatial sampling and inference. Environ. Model. Software 134. https://doi.org/10.1016/j.envsoft.2020.104835.

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1), 1–17. https://doi.org/10.18637/jss.v077.i01.

Wu, Y., Li, S., Zhao, Q., Wen, B., Gasparrini, A., Tong, S., Overcenco, A., Urban, A., Schneider, A., Entezari, A., Vicedo-Cabrera, A.M., Zanobetti, A., Analitis, A., Zeka, A., Tobias, A., Nunes, B., Alahmad, B., Armstrong, B., Forsberg, B., Pan, S.C., Iniguez, C., Ameling, C., De la Cruz Valencia, C., Astrom, C., Houthuijs, D., Van Dung, D., Roye, D., Indermitte, E., Lavigne, E., Mayvaneh, F., Acquaotta, F., de'Donato, F., Rao, S., Sera, F., Carrasco-Escobar, G., Kan, H., Orru, H., Kim, H., Holobaca, I.H., Kysely, J., Madureira, J., Schwartz, J., Jaakkola, J.J.K., Katsouyanni, K., Hurtado Diaz, M., Ragettli, M.S., Hashizume, M., Pascal, M., de Sousa Zanotti Stagliorio Coelho, M., Ortega, N.V., Ryti, N., Scovronick, N., Michelozzi, P., Correa, P.M., Goodman, P., Nascimento Saldiva, P.H., Abrutzky, R., Osorio, S., Dang, T.N., Colistro, V., Huber, V., Lee, W., Seposo, X., Honda, Y., Guo, Y. L., Bell, M.L., Guo, Y., 2022a. Global, regional, and national burden of mortality associated with short-term temperature variability from 2000-19: a three-stage modelling study. Lancet Planet Health 6, e410–e421.

Wu, Y., Wen, B., Li, S., Gasparrini, A., Tong, S., Overcenco, A., Urban, A., Schneider, A., Entezari, A., Vicedo-Cabrera, A.M., Zanobetti, A., Analitis, A., Zeka, A., Tobias, A., Alahmad, B., Armstrong, B., Forsberg, B., Iniguez, C., Ameling, C., De la Cruz Valencia, C., Astrom, C., Houthuijs, D., Van Dung, D., Roye, D., Indermitte, E., Lavigne, E., Mayvaneh, F., Acquaotta, F., de'Donato, F., Sera, F., Carrasco-Escobar, G., Kan, H., Orru, H., Kim, H., Holobaca, I.H., Kysely, J., Madureira, J., Schwartz, J., Katsouyanni, K., Hurtado-Diaz, M., Ragettli, M.S., Hashizume, M., Pascal, M., de Sousa Zanotti Stagliorio Coelho, M., Scovronick, N., Michelozzi, P., Goodman, P., Nascimento Saldiva, P.H., Abrutzky, R., Osorio, S., Dang, T.N., Colistro, V., Huber, V., Lee, W., Seposo, X., Honda, Y., Bell, M.L., Guo, Y., 2022b. Fluctuating temperature modifies heat-mortality association around the globe. Innovation (Camb) 3, 100225.

Xiao, Y., Zhao, W., Ma, M., He, K., 2021. Gap-free LST generation for MODIS/Terra LST product using a random Forest-based reconstruction method. Remote Sens. (Basel) 13.

Zhang, T., Zhou, Y., Zhao, K., Zhu, Z., Chen, G., Hu, J., Wang, L., 2022a. A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003−2020). Earth Syst. Sci. Data 14, 5637–5649.

Zhang, T., Zhou, Y., Zhu, Z., Li, X., Asrar, G.R., 2022b. A global seamless 1 km resolution daily land surface temperature dataset (2003–2020). Earth Syst. Sci. Data 14, 651–664.

Zhao, Q., Guo, Y., Ye, T., Gasparrini, A., Tong, S., Overcenco, A., Urban, A., Schneider, A., Entezari, A., Vicedo-Cabrera, A.M., Zanobetti, A., Analitis, A., Zeka, A., Tobias, A., Nunes, B., Alahmad, B., Armstrong, B., Forsberg, B., Pan, S.C., Iniguez, C., Ameling, C., De La Cruz Valencia, C., Astrom, C., Houthuijs, D., Dung, D. V., Roye, D., Indermitte, E., Lavigne, E., Mayvaneh, F., Acquaotta, F., de'Donato, F., Di Ruscio, F., Sera, F., Carrasco-Escobar, G., Kan, H., Orru, H., Kim, H., Holobaca, I. H., Kysely, J., Madureira, J., Schwartz, J., Jaakkola, J.J.K., Katsouyanni, K., Hurtado Diaz, M., Ragettli, M.S., Hashizume, M., Pascal, M., de Sousa Zanotti Stagliorio Coelho, M., Valdes Ortega, N., Ryti, N., Scovronick, N., Michelozzi, P., Matus Correa, P., Goodman, P., Nascimento Saldiva, P.H., Abrutzky, R., Osorio, S., Rao, S., Fratianni, S., Dang, T.N., Colistro, V., Huber, V., Lee, W., Seposo, X., Honda, Y., Guo, Y.L., Bell, M.L., Li, S., 2021. Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: a three-stage modelling study. Lancet Planet Health 5, e415–e425.

Zhou, B., Erell, E., Hough, I., Rosenblatt, J., Just, A.C., Novack, V., Kloog, I., 2020. Estimating near-surface air temperature across Israel using a machine learning based hybrid approach. Int. J. Climatol. 40, 6106–6121.

Zhu, W., Lű, A., Jia, S., Yan, J., Mahmood, R., 2017. Retrievals of all-weather daytime air temperature from MODIS products. Remote Sens. Environ. 189, 152–163.