Environmental Research 238 (2023) 117173

Contents lists available at ScienceDirect

Environmental Research

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

Improved daily estimates of relative humidity at high resolution across Germany: A random forest approach

Nikolaos Nikolaou [a,b,*], Laurens M. Bouwer [c], Marco Dallavalle [a,b], Mahyar Valizadeh [a], Massimo Stafoggia [d], Annette Peters [a,b], Kathrin Wolf [a,1], Alexandra Schneider [a,1]

[a] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[b] Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Pettenkofer School of Public Health, Munich, Germany
[c] Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany
[d] Department of Epidemiology, Lazio Regional Health Service – ASL Roma 1, Rome, Italy

ARTICLE INFO

Handling Editor: Jose L Domingo

Keywords:
Relative humidity
Spatiotemporal modeling
Machine learning
External validation
Exposure assessment
Environmental epidemiology

ABSTRACT

The lack of readily available methods for estimating high-resolution near-surface relative humidity (RH) and the incapability of weather stations to fully capture the spatiotemporal variability can lead to exposure misclassification in studies of environmental epidemiology. We therefore aimed to predict German-wide 1 × 1 km daily mean RH during 2000–2021. RH observations, longitude and latitude, modelled air temperature, precipitation and wind speed as well as remote sensing information on topographic elevation, vegetation, and the true color band composite were incorporated in a Random Forest (RF) model, in addition to date for capturing the temporal variations of the response-explanatory variables relationship. The model achieved high accuracy ($R^2 = 0.83$) and low errors (Root Mean Square Error (RMSE) of 5.07%, Mean Absolute Percentage Error (MAPE) of 5.19% and Mean Percentage Error (MPE) of - 0.53%), calculated via ten-fold cross-validation. A comparison of our RH predictions with measurements from a dense monitoring network in the city of Augsburg, South Germany confirmed the good performance ($R^2 \geq 0.86$, RMSE $\leq$ 5.45%, MAPE $\leq$ 5.59%, MPE $\leq$ 3.11%). The model displayed high German-wide RH (22y-average of 79.00%) and high spatial variability across the country, exceeding 12% on yearly averages. Our findings indicate that the proposed RF model is suitable for estimating RH for a whole country in high-resolution and provide a reliable RH dataset for epidemiological analyses and other environmental research purposes.

1. Introduction

Relative humidity (RH) refers to the water vapor content of air and quantifies how far the atmosphere is from its saturation point. RH is a key parameter for many fields such as agriculture (Zhang et al., 2015), hydrology (Forootan, 2019) and climatology (Sherwood et al., 2010) as it contributes among others to the soil moisture, the hydrological cycle and the weather and climate conditions. Thus, RH plays an important role in plant and animal life (Xiong et al., 2017) as well as in human comfort and well-being (Davis et al., 2016; Yang et al., 2018).

RH has mostly been used as a confounder or effect modifier in studies focusing on air temperature (T_{air}) (Armstrong, 2006; Zeng et al., 2017), or as part of an index, e.g., apparent temperature (Analitis et al., 2008). Nevertheless, there is evidence that RH is potentially an independent risk factor for mortality (Ou et al., 2014) and morbidity (Luo et al., 2020). In epidemiology, RH data are usually retrieved from weather monitors. But their locations are irregularly distributed over space, usually in rural areas, and their number is limited. Hence, weather stations are inadequate to fully represent the spatiotemporal RH variations in complex geo-climatic urban and rural landscapes, and by using their observations, error is introduced in the exposure assessment of study participants leading to estimates biased towards the null (Zeger

* Corresponding author. Address: Ingolstädter Landstr. 1, D-85764, Neuherberg, Germany.
E-mail addresses: nikolaos.nikolaou@helmholtz-munich.de (N. Nikolaou), laurens.bouwer@hereon.de (L.M. Bouwer), marco.dallavalle@helmholtz-munich.de (M. Dallavalle), mahyar.valizadeh@helmholtz-munich.de (M. Valizadeh), m.stafoggia@deplazio.it (M. Stafoggia), annette.peters@helmholtz-munich.de (A. Peters), kathrin.wolf@helmholtz-munich.de (K. Wolf), alexandra.schneider@helmholtz-munich.de (A. Schneider).
[1] Shared last authorship.

https://doi.org/10.1016/j.envres.2023.117173
Received 8 February 2023; Received in revised form 1 September 2023; Accepted 17 September 2023
Available online 20 September 2023

et al., 2000). Climate reanalysis data could be an alternative source for environmental health research (Mistry et al., 2022), but the resolution is usually coarser than 9 km and the data fail to capture the city-level exposure variability effectively. We therefore suggest to extend the methods and datasets in order to improve the predictions of RH exposure for people participating in epidemiological studies, such as prospective cohorts with data on the residential addresses of the participants.

There is a clear methodological gap in RH modeling, especially for high spatiotemporally-resolved RH predictions and for timespans up to multiple years. Li et al. (2014) mapped RH every 3 h at 1 km by using a two-step interpolation procedure of re-analysis data based on a partial thin-plate spline (TPS) and simple kriging (Root Mean Square Error (RMSE) = 11.06%). The traditional interpolation techniques have limited efficiency when mapping meteorological exposures in spatially highly heterogeneous areas, and are characterized by neighbouring effects on exposures predictions, without being capable of capturing small-scale and intra-city variations. Li and Zha (2018) used a Random Forest (RF) model and satellite data, to estimate RH during the summer of 2009 (R^2 = 0.70, RMSE = 7.4%). Spatiotemporal predictors which could explain a large amount of the remaining RH variance, e.g., T_{air}, were not included. Longer periods and more predictors need to be tested to capture the full annual and inter-annual RH variability. For China, the RF model had better results than TPS and kriging, but improvements are needed for better RH variability representation, higher prediction accuracy and further temporal extension to the annual level.

Remote sensing data are progressively used in environmental exposures modeling (Rosenfeld et al., 2017; Yao et al., 2022) being publicly available in high spatiotemporal resolution. There is also a growing body of machine learning (ML) methods applied in the field (Jin et al., 2022; Silibello et al., 2021; Stafoggia et al., 2019).

The specific objectives of this study were (a) to estimate highly spatiotemporal resolved RH for Germany based on T_{air} and other observation, remote sensing and modelled data by using a RF model, (b) to evaluate the model's performance and (c) to produce a reliable German-wide RH dataset for subsequent epidemiological analyses and various research purposes. Thereby, we aimed to extend the current literature and provide a generalizable method for other countries to produce highly resolved RH datasets.

2. Materials and methods

2.1. Study domain

Germany extends in an area of 357,021 km^2, having a strongly diverse landscape and a high elevation range (−3.54 to 2962 m). In the south-eastern regions, the climate is classified as warm summer humid continental, while in north-western regions it is characterized as temperate oceanic (Beck et al., 2018b). We divided Germany's land area into 366,536 grid cells of 1 × 1 km resolution, following the European INSPIRE (Infrastructure for Spatial Information in the European Community) standard for gridded datasets and using the Lambert Azimuthal Equal-Area (LAEA) projection, EPSG: 3035 (©GeoBasis-DE/BKG (2021)).

2.2. Input data

Large amounts of input data were incorporated in the RF modeling process. We used meteorological observations, remote sensing and spatiotemporally resolved modelled data, all retrieved from 2000 to 2021 across the study area.

2.2.1. RH data

We used daily mean RH observations (DWD, 2022a) from 406 weather stations of the German Meteorological Service (DWD) https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/daily/kl/historical/ (Figure S1). The RH data has been quality controlled by the DWD and all the needed information such as station location as well as relocations was included in their metadata files.

2.2.2. T_{air} data

In our previous work (Nikolaou et al., 2022), we estimated daily mean T_{air} in high-resolution (1 × 1 km) across Germany using a regression-based method incorporating two linear mixed models. In brief, we predicted T_{air} by calibrating the strong relationship between the weather stations' T_{air} observations and the satellite-based land surface temperature (LST) also adjusting for various spatial predictors. We also applied a TPS interpolation in T_{air} data in order to achieve a full German-wide coverage. Extensive validation showed high performance ($R^2 \geq 0.96$) and low errors (RMSE ≤ 1.41 °C).

2.2.3. Elevation data

We downloaded elevation data at 30-arc-second spatial resolution https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30, provided by the U.S. Geological Survey's Earth Resources Observation Systems (EROS) Data Center (Gesch et al., 1999). We aggregated these data to a 1 × 1 km grid, including the land borders and the shorelines in the North and Baltic Seas to match our intended spatial resolution (Figure S2).

2.2.4. Greenness data

The normalized difference vegetation index (NDVI) is a proxy of vegetation greenness on the Earth surface, quantifying the vegetation cover and quality over space. We retrieved NDVI data of 1 × 1 km from the TERRA MODIS product MOD13A3v006 https://lpdaac.usgs.gov/products/mod13a3v006/ (Didan, 2015). These are monthly data - weighted temporal average values through the month, which is sufficient, as vegetation does not change considerably during a month.

2.2.5. True color band composite data

The visible red, green and blue light bands demonstrate how we see Earth's surface from space. We retrieved the daily true color band composite, i.e. the surface spectral reflectance for the red (band 1), blue (band 3) and green (band 4) bands at 500 m spatial resolution from the TERRA MODIS product MOD09Gav006 https://lpdaac.usgs.gov/products/mod09gav006/, corrected for atmospheric conditions (Vermote, 2015). We aggregated the data to a 1 × 1 km grid, to suit the output's spatial resolution.

2.2.6. Precipitation data

We used daily precipitation data of 1 × 1 km developed by the REGNIE (Regionalisierte Niederschlagshöhen) method which are publicly available from the DWD Climate Data Center https://opendata.dwd.de/climate_environment/CDC/grids_germany/daily/regnie/ (DWD, 2022b). REGNIE is based on the interpolated DWD weather station precipitation measurements, using a combination of multiple linear regressions and Inverse Distance Weighting (IDW), with orographic conditions considered (Rauthe et al., 2013). In a recent update, the REGNIE dataset has been substituted with HYRAS-DE-PRE (DWD, 2023), which shares the same methodology and references the identical paper by Rauthe et al. (2013).

2.2.7. Wind speed data

We retrieved daily mean wind speed (DWD, 2022a) of the same 406 weather stations as for the RH data https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/daily/kl/historical/ (Figure S1). We interpolated this dataset to 1 × 1 km spatial resolution using TPS, since studies have suggested that TPS outperformed other interpolation methods such as kriging or IDW for mapping climate variables (Wu et al., 2013, 2015). Details regarding the spatiotemporal distribution and the assessment of wind speed interpolation are available in the Supplementary material (Figure S3 and Table S1).

2.3. Modeling

RF (Breiman, 2001) is a well-known and powerful supervised ensemble ML algorithm, utilized for solving both classification and regression tasks - based on the bagging principle. For regression, random sub-samples of the given dataset (i.e., the training set in most applications) are selected (with replacement). Then, the algorithm constructs decision trees - one for every sub-sample, also including a subset of the specified features (i.e., the model predictors). Each decision tree will generate an output/prediction of the target variable. The main model's output is calculated by averaging all the outputs of the individual decision trees.

The RF algorithm copes greatly with big data, with potentially correlated predictors and their non-linearity, and with overfitting. RF is also considered as a robust method against outliers.

In our study, we trained the RF model, trying to evaluate its efficiency in reproducing the observed RH values measured by the weather stations, i.e., the ground-based truth. As RF inherent robustness alleviates the need for complex hyperparameter tuning, we did not proceed with highly sophisticated methodologies for hyperparameters tuning but rather some trial and error by deviating from the default settings. We did not observe any strong differences to the model performance by testing different sets of hyperparameters. Eventually, we used 500 trees and 8 randomly sampled variables as candidates at every split (num. trees = 500, mtry = 8), training the model for each year separately to capture annual variations. The daily observed mean RH (%) at the DWD stations was the response variable. The predictors were our previously modelled daily mean T_{air} (Celsius), the daily red, green and blue bands (dimensionless), the daily mean precipitation height (mm) and the daily mean wind speed (m/s) as well as elevation (meters) and monthly NDVI (dimensionless). We also integrated the geographical coding information [i.e., longitude (°) and latitude (°)] to account for spatial variations that might not be fully represented by other spatial features in the model, and we included the day of the year (1–365|366) in order to capture daily variations in the response-predictor variables relationship.

2.3.1. Model performance

Ten-fold cross-validation (CV) was used to assess the model performance by randomly dividing the set of the DWD weather monitors to a training and a testing set (90:10) ten times. Each time, the model was refitted using the training set and then the RH was predicted in the respective testing set. Our aim was to estimate a full time series of RH in locations without weather stations and therefore in grid cells where the RF model was not previously trained, and consequently to simulate the prediction step of the modeling procedure. Regressing the observed mean RH vs. the predicted mean RH by the RF model's testing set, we calculated the corresponding R^2, RMSE, Mean Absolute Percentage Error (MAPE) and Mean Percentage Error (MPE) (formulations are written in the Supplementary), each of them ten times and then we took their average to represent each year's CV-R^2, CV-RMSE, CV-MAPE and CV-MPE.

In the prediction step, we applied the RF model to all grid cells and days combinations without available RH measurements of DWD weather stations in order to obtain a complete RH dataset for entire Germany.

2.3.2. Validation with external data

An additional validation was conducted by comparing our daily mean RH predictions with measurements of an independent dense monitoring network during 2015–2019. The network included RH measurements of 4 min temporal resolution from 82 HOBO-Logger devices (ONSET, Type Pro v2), which were located in the city of Augsburg and in two adjacent counties (Augsburg county and Aichach-Friedberg) (Figure S4). Detailed information for the monitoring network and the measurements' quality assurance can be found in the corresponding paper (Beck et al., 2018a). For our comparison, we aggregated the 4-min RH values to daily means and then 7-day averages. We generated the corresponding R^2, RMSE, MAPE and MPE as derived from linearly regressing the predicted RH from the model against the observed RH from the HOBO-Logger monitors.

The majority of the HOBO-Logger stations were located in the city center of Augsburg or close to it, where no DWD measurements were available in the training step of the RF model (closest stations were approx. 10 and 18 km apart from the city center, see Figure S4). Thus, we investigated the performance of the model in an area without prior information but of great epidemiological interest since highly populated implicating that more people are exposed here.

2.4. Descriptive analyses and case study

Descriptive statistics [mean, standard deviation (SD), minimum (min), first quartile (Q1), median, third quartile (Q3) and maximum (max)] were calculated from our German-wide RH predictions and from the DWD observations. We also investigated the spatiotemporal RH patterns over the last 2 decades, overall and by season.

To demonstrate the improvement in our exposure assessment, we compared the spatial distributions of the daily mean RH predictions from the RF model and the daily mean RH measurements from the DWD stations in an urban location for the two last decades. The city of Regensburg covers an area of 80.76 km^2 with about 150,000 inhabitants, and, as one of the study sites of the German National Cohort (NAKO) (German National Cohort Consortium, 2014), has also an epidemiological research interest.

We performed our analysis in R, v. 4.2.2 (R Core Team, 2022). The RF model was developed with the R package "ranger" (Wright and Ziegler, 2017).

3. Results

Figure S5 shows the Spearman correlation coefficients for the models' variables. Briefly, RH was found to be highly and positively associated with the true color band composite ($r \approx 0.5$) while there was a strong negative correlation with T_{air} ($r \approx -0.5$). In Figure S6, we demonstrate the variable importance plot findings. Date played a very important role. We also observed that T_{air} and the blue band were the most important spatiotemporal predictors of the RF model for estimating RH. They were followed by precipitation, green band, wind speed and longitude, and then elevation, latitude, NDVI and red band. The order of the predictors was slightly different through the years, but there were main trends as described.

3.1. Model performance

The model achieved high accuracy [22-year average R^2 = 0.83 (range: 0.77–0.88)] and small errors [22-year average RMSE = 5.07% (range: 4.44%–6.27%), MAPE = 5.19% (range: 4.45%–6.93%) and MPE = - 0.53% (range: - 0.35% - - 0.89%), Table 1]. We observed an increase of the model performance (increase of R^2 and decrease of errors), together with an increase of the total number of available weather station data over the years. Scatterplots depicting the example years with the lowest and highest fitting scores, specifically 2001 and 2020, have been included in the Supplementary material (Figure S7). Autumn months (September–November) had the lowest RMSE = 4.65% (range: 3.89%–5.83%) while spring months (March–May) had the highest RMSE = 5.32% (range: 4.60%–6.44%) (Fig. 1). We also observed that predictions belonging to the lower 10% of the dataset gave higher errors [RMSE = 7.85% (range: 6.86%–9.28%)] compared to the predictions of the upper 10% of the dataset [RMSE = 5.38% (range: 4.47%–6.79%] (Fig. 1). The corresponding results for MAPE and MPE are available in the Supplementary (Figure S8 and S9).

Table 1
Prediction accuracy for the RF model: 10-fold CV results for the daily mean RH predictions over Germany during 2000–2021.

Year	R^2	RMSE (%)	MAPE (%)	MPE (%)	Sample size (number of cell-days)
2000	0.78	5.71	5.88	−0.64	100,699
2001	0.78	5.53	5.50	−0.52	121,225
2002	0.77	5.69	5.77	−0.59	123,946
2003	0.81	6.27	6.93	−0.89	123,364
2004	0.78	5.64	5.74	−0.61	126,604
2005	0.81	5.21	5.26	−0.51	134,386
2006	0.82	5.28	5.37	−0.65	135,600
2007	0.84	4.81	4.89	−0.48	139,482
2008	0.83	5.00	5.14	−0.52	140,135
2009	0.82	5.06	5.15	−0.49	142,295
2010	0.86	4.72	4.73	−0.39	142,629
2011	0.86	4.91	5.04	−0.56	141,781
2012	0.84	4.74	4.81	−0.48	141,820
2013	0.84	4.80	4.72	−0.44	140,928
2014	0.85	4.55	4.47	−0.38	142,641
2015	0.85	4.91	5.03	−0.52	142,908
2016	0.83	4.72	4.65	−0.41	139,491
2017	0.83	4.69	4.64	−0.41	143,206
2018	0.87	4.94	5.32	−0.57	143,026
2019	0.85	5.09	5.37	−0.62	140,866
2020	0.88	4.87	5.26	−0.53	116,670
2021	0.85	4.44	4.45	−0.35	116,544
Overall	**0.83**	**5.07**	**5.19**	**−0.53**	**133,648**

3.2. Validation with external data

We found a strong correspondence between our RH model predictions and the external HOBO-Logger network measurements with a 5-year average R^2 of 0.86 (range: 0.82–0.89) and a 5-year average RMSE of 5.45% (range: 5.14%–6.16%), MAPE of 5.59% (range: 5.19%–6.42%) and MPE of 2.98% (range: 1.82%–4.47%) for the daily average RH exposure (Table 2). For the 7-day average RH exposure, as expected, the accuracy was even higher [$R^2 = 0.87$ (range: 0.84–0.92)] and the errors lower [RMSE = 4.49% (range: 4.06%–5.29%), MAPE = 4.59% (range: 4.08%–5.51%), MPE = 3.11% (range: 1.77%–4.81%)]. Density scatterplots confirmed the good correlation (Figure S10).

3.3. Case study - Regensburg

In Fig. 2, we display the average spatial RH patterns for the region of Regensburg for the period 2000–2021. The city area showed up to 4.5% lower RH values than the surrounding rather rural county area. However, the variability of the daily values which will be also considered in subsequent epidemiological analysis is much larger than the 22-year average - e.g., up to 9% (randomly selected example day in Figure S11). Yet, the rural region was characterized by variations even in neighbouring tiles. The average RH exposure in Regensburg measured by the available DWD weather station of the region was far below the Q1 of the RH predictions of the RF model for the region (Fig. 3).

3.4. Spatiotemporal RH patterns

Table 3 shows descriptive statistics of measured and modelled RH across Germany for 2000–2021. Germany was characterized by high RH values with Q1 of both DWD stations' and model's RH distribution to be 71% and 71.91%, respectively. The observed and predicted 22-year average RH derived by the DWD stations and the RF model were 79.05% (SD = 12.38%) and 79.00% (SD = 10.46%), respectively.

Fig. 4 displays the 22-year averaged predicted RH output map of Germany (plot 1) which indicates spatial RH patterns, including urbanization, mountainous regions, rivers, forests and coastlines. Metropolitan areas such as those of Berlin, Hamburg and Munich and the extended and other dense urban cores (e.g., from Karlsruhe to Frankfurt) had much lower RH values compared to the neighbouring rural settings.

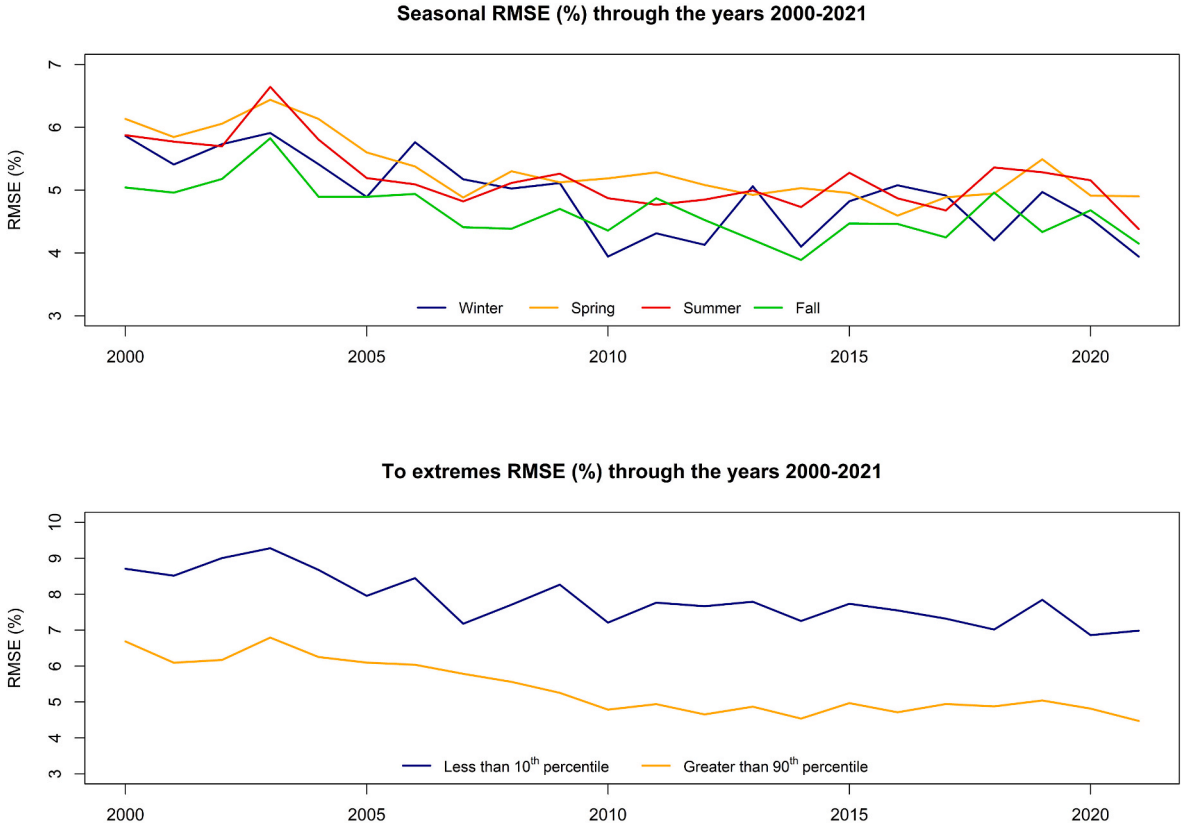

Fig. 1. Seasonal RMSE and RMSE to extremes for the model's RH predictions in Germany during 2000–2021.

Table 2
Accuracy results from the validation with external data using the HOBO-Logger daily mean RH observations and 7-day averages over the Augsburg region during 2015–2019.

					7-day average			
Year	R^2	RMSE (%)	MAPE (%)	MPE (%)	R^2	RMSE (%)	MAPE (%)	MPE (%)
2015	0.87	5.14	5.22	2.07	0.89	4.06	4.08	2.15
2016	0.82	5.23	5.19	2.48	0.84	4.14	4.10	2.58
2017	0.84	5.15	5.43	1.82	0.84	4.30	4.34	1.77
2018	0.89	5.58	5.67	4.07	0.92	4.68	4.93	4.25
2019	0.86	6.16	6.42	4.47	0.87	5.29	5.51	4.81
Overall	**0.86**	**5.45**	**5.59**	**2.98**	**0.87**	**4.49**	**4.59**	**3.11**

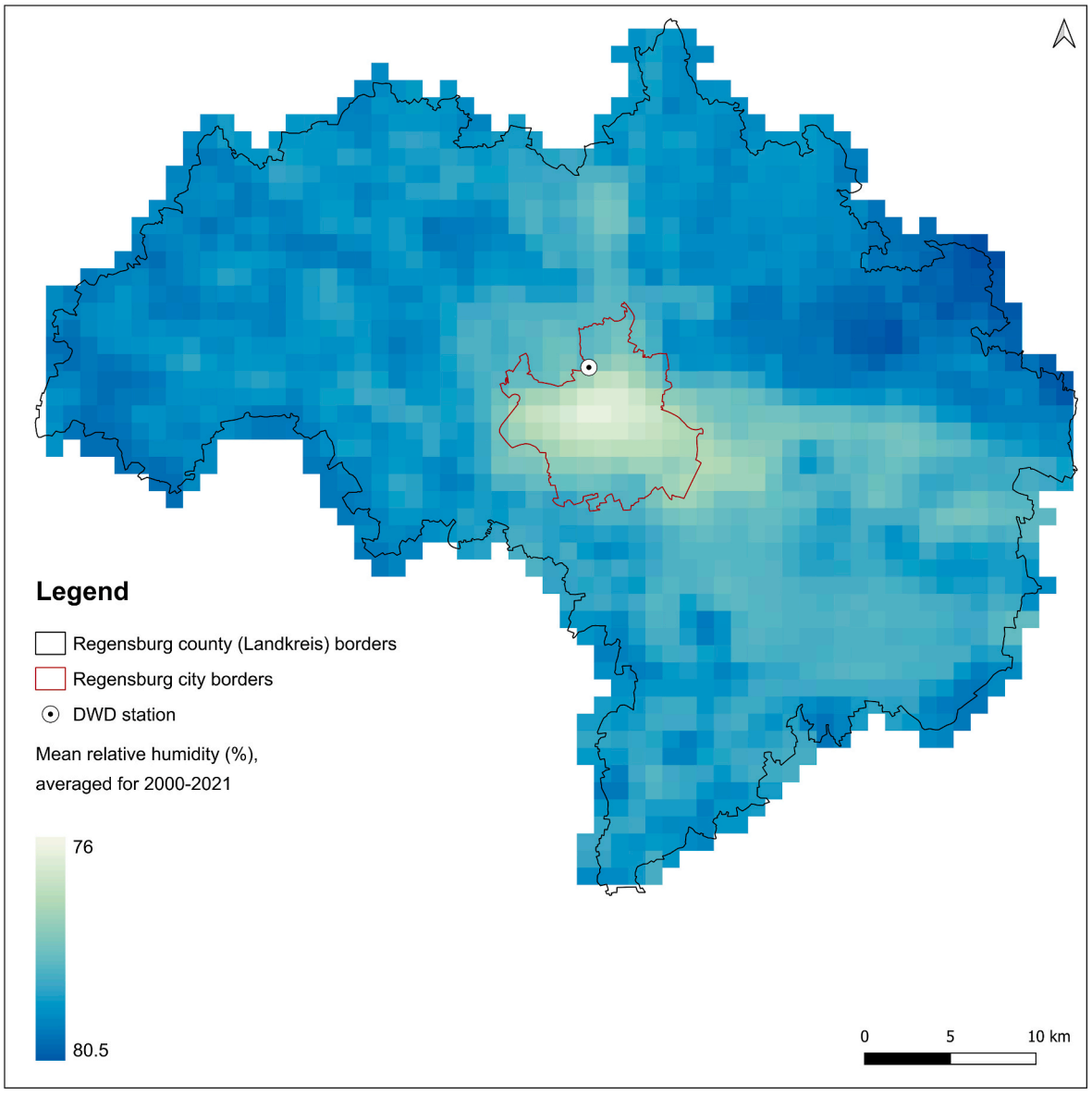

Fig. 2. Spatial pattern of the averaged predicted RH in Regensburg during 2000–2021.

In Figure S12, we zoomed in the Augsburg region, which consists of the city center and two adjacent counties, to give an example of the high spatial difference between a city center and its neighbouring but less urbanized areas. Additionally, dense mountainous regions such as the Alps and Harz, coastlines as the North Sea coast and rivers as Elbe in a large part of it, had the highest RH values country-wide (Fig. 4). Furthermore, we included the spatial distribution map exhibiting the interannual change of RH (Figure S13) to ensure comprehensive coverage. Significant interannual spatial variations were not discernible and the spatial variability which remained mostly constant through the years, aligned with the patterns observed and described in the averaged map (Fig. 4, plot 1). Also, the temporal RH variability in Germany is presented for 2001–2021, by exhibiting the differences between the predicted RH yearly averages and the 21-year average (Fig. 4, plot 2). We excluded the year 2000 because the model predictions are only available from late February of that year due to the missing T_{air} values

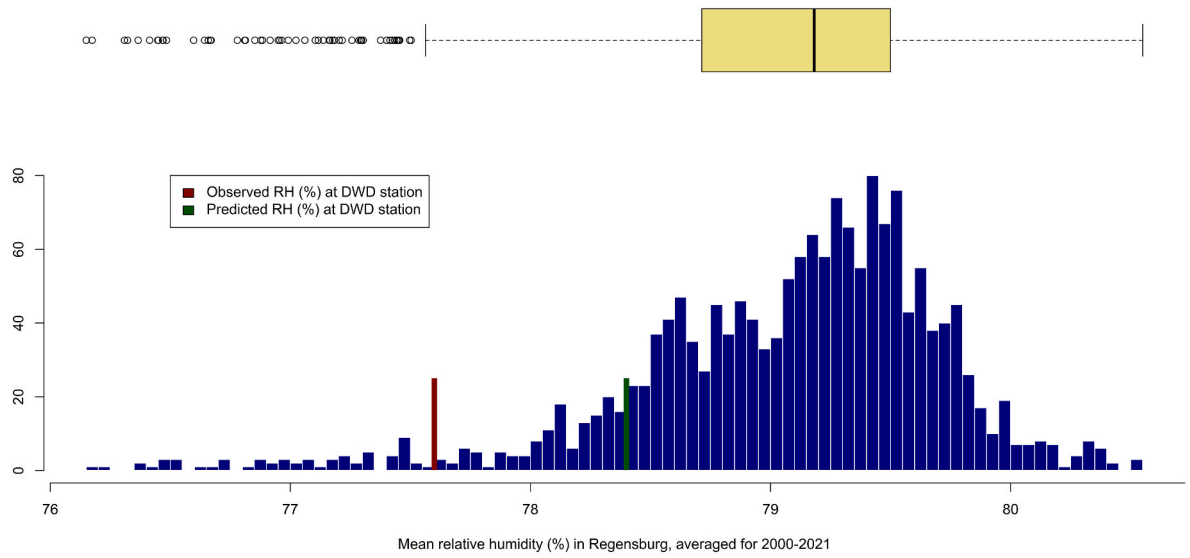

Fig. 3. Distribution of predicted RH in the Regensburg region for 2000–2021 (histogram in blue and corresponding boxplot above).

Table 3
Observed and predicted mean RH (%) over Germany during 2000–2021.

Source	Mean	SD	Min	Q1	Median	Q3	Max
DWD stations (n = 406)	79.05	12.38	3.00	71.00	81.00	88.75	100.00
RF model (n = 366,536 cells)	79.00	10.46	13.70	71.91	80.56	87.44	100.00

until then. There were some fluctuations over the years but without indication of an increasing or decreasing trend. The most humid years were 2001 (81.30%), 2014 (81.20%) and 2013 (80.94%) while the most arid were 2003 (75.31%), 2020 (75.53%) and 2018 (75.52%), which are known hot and dry years from the recent climatological record.

Mapping the 22-year average RH by season (Figure S14) identified winter and fall as the most humid seasons. High spatial RH variability was also observed within each season.

4. Discussion

In this paper, we introduced an approach for spatial and temporal modeling of RH using RF, a popular ML method for prediction tasks. The approach goes beyond the conventional interpolation of meteorological observations and uses several other data sources. We produced a reliable spatiotemporally-resolved RH dataset at 1 × 1 km spatial resolution across Germany for the period 2000–2021. The RF model achieved good performance with high predictive accuracy and low errors, validated with both internal data using cross-validation (R^2 = 0.83, RMSE = 5.07%, MAPE = 5.19%, MPE = - 0.53%), and with independent observational data ($0.86 \leq R^2 \leq 0.87$, $4.49\% \leq$ RMSE $\leq 5.45\%$, $4.59\% \leq$ MAPE $\leq 5.59\%$, $2.98\% \leq$ MPE $\leq 3.11\%$). A case study for the city of Regensburg shows that our dataset is capable of capturing the full range of spatial variability of RH compared to the standard use of meteorological observations. These DWD station observations could not represent the high RH values of the peripheral areas in Regensburg, but also not the very low RH values of the city center. This clearly demonstrates the added value of our approach and how the use of additional data sources supplementing the conventional use of meteorological observations improved the RH prediction. It is especially important to capture the RH spatial variability for assessing differences in human's individual exposure in epidemiological studies. We also presented an analysis of the spatiotemporal RH patterns in Germany during 2000–2021.

The RH-health relevance has not been clarified adequately (Bind et al., 2014). RH adverse effects on human health could be partially explained by its interplay with the excessive heat stress and the body dehydration, as described in Davis et al. (2016). During extended and excessive heat events such as heatwaves, the human body struggles against heat-driven physiological responses and a key mechanism for its temperature regulation is evaporation. However, when RH is high and therefore air contains a lot of moisture, it is difficult for the sweat to be relieved and thus cooling becomes insufficient. Hence, the body core temperature increases while this increase is associated with a variety of detrimental health effects (Schneider et al., 2017). Additionally, low RH can affect the human skin sensitivity to mechanical stress (Engebretsen et al., 2016). RH is also associated with the transition of vector-borne diseases e.g., from mosquitos and ticks (Davis et al., 2016) as well as with the development and stability of microorganisms in aerosols, facilitating airborne diseases (Božič and Kanduč, 2021).

So far there is a literature gap in the investigation of the RH exposure's direct effects on human health and the accompanying underlying mechanisms. Further and more detailed research is needed. Hence, it is critically important for epidemiologists to have access to high-resolution and reliable RH datasets.

Most epidemiological studies retrieve the participants' exposure information, in this case RH, from available meteorological stations that do not capture the full variability of RH, especially at the city scale. In the Regensburg area, an epidemiological study would usually assign RH measurements from the station most closely located to each participant's residential address but fails to account for the spatial variability of RH that is actually occurring. Therefore, some measurement error would be introduced and the variability would be lost. Focusing on the city area, participants who live there would be assigned with a higher RH value than their actual one. At the same time, those living outside the city center would be assigned with RH values that are too low. This clearly demonstrates the urgent need for high spatiotemporal RH datasets for health studies for less biased exposure estimates.

Compared to other studies that use interpolation techniques such as TPS or kriging, our RF model is capable of reducing errors by half. Li et al. (2014) introduced a two-step procedure to map RH every 3 h at 1 km resolution over China during 1958–2010. They fitted a partial TPS interpolation to reanalysis data, location and elevation as predictors, to estimate a trend surface, and then a simple kriging was applied to the residuals for trend surface correction. They reported a RMSE of 11.06% whereas our model showed a RMSE of 5.07%. More recently, Li and Zha (2018) also used an RF model, combining station and satellite data, to

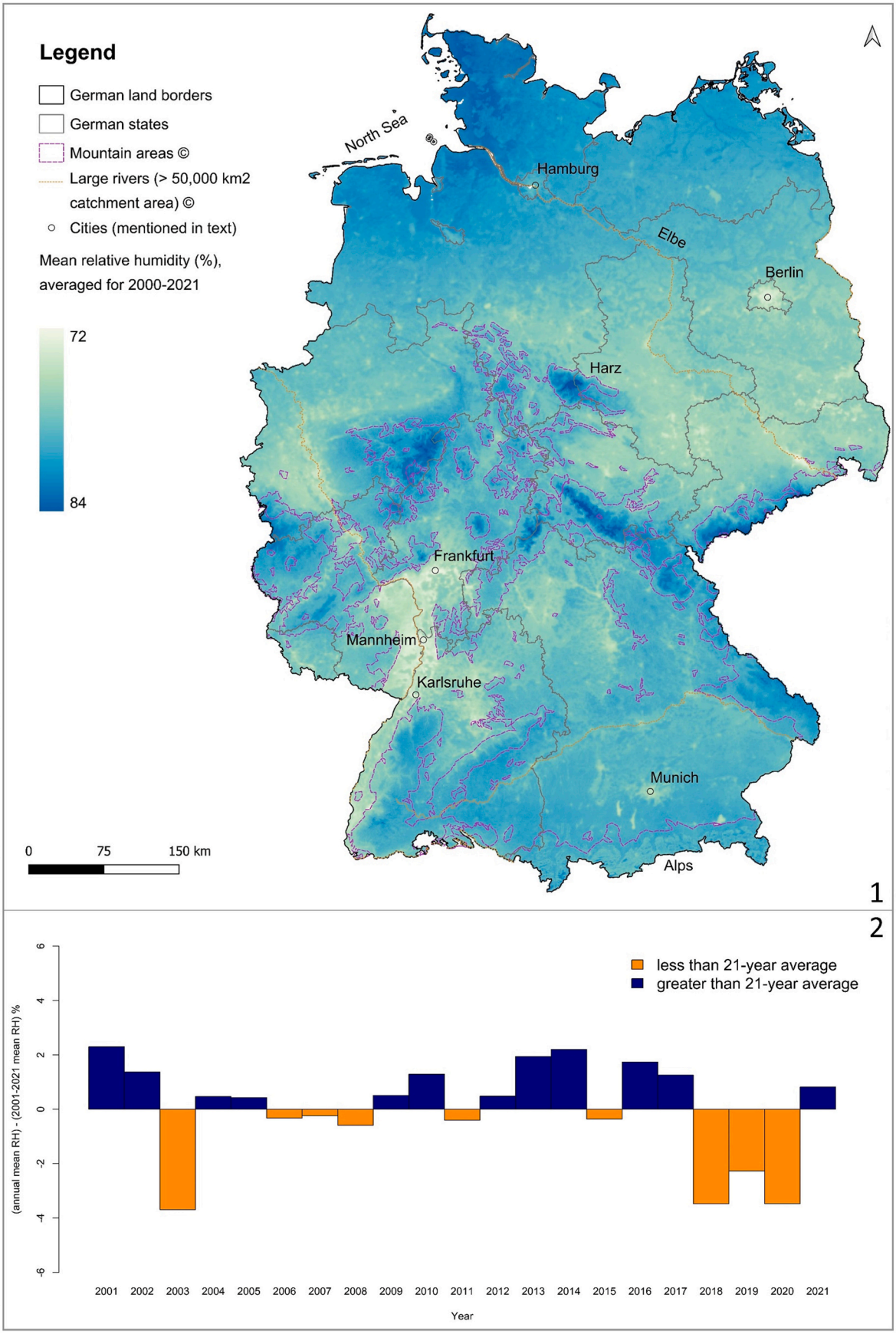

Fig. 4. Spatiotemporal RH patterns in Germany during 2000–2021. Plot 1: Spatial patterns of the predicted RH in Germany, averaged for 2000–2021. Plot 2: Difference between the predicted RH yearly averages and the predicted RH 21-year average (2001–2021), German-wide.

estimate RH during the hot summer of 2009 over China. Elevation and vegetation were found to be the most important predictors for RH. Comparing our model with their work, it seems that our additional inclusion of T_{air}, date information, precipitation and wind speed data in the modeling process, significantly improved the model's performance. Li and Zha (2018) reported a $R^2 = 0.70$ and RMSE = 7.4%, whereas our model could improve the R^2 to 0.83 and lower the errors to RMSE = 5.07%. In addition, our RF model allowed us to model RH for entire years and not only for one season. Lately, Kloub (2022) used an auto-encoder residual neural network, incorporating monitor, re-analysis and satellite data, to estimate various meteorological factors including RH for China in 2015. The accuracy of their RH model considerably improved from 0.77 to 0.86 upon including the monthly index, highlighting the importance of diverse temporal variables for RH models (we also incorporated the day of the year index). Kloub (2022) achieved a fairly good model performance of $R^2 = 0.86$, Mean Absolute Error (MAE) = 5.58% and RMSE = 7.41%, whereas our model yielded an R^2 of 0.83, MAPE of 5.19%, and notably lower RMSE of 5.07%. It is important to consider that Kloub (2022) predicted RH only for a single year, while our model covered a 22-year period. For instance, we also reported an R^2 of 0.88 for the year 2020. Significantly, the confidence for our model's performance benefited from the conducted external validation using a dense and independent monitoring network in Augsburg, a distinctive advantage not present in other studies.

This study was also subject to limitations. Satellite-derived predictors like NDVI and the true color band composite may encounter resampling errors, whereas precipitation and wind speed data involve spatial interpolation errors. Nevertheless, these datasets maintain a high standard of quality and are extensively employed in existing literature. Furthermore, the external validation set was not representative of the whole Germany. The HOBO-Logger monitoring network was placed in Augsburg, South Germany. However, we used the Augsburg's greater region which consists of a dense city center and two adjacent rural settings and therefore the validation area was characterized by high spatial RH variability. Additionally, we were already able to measure the model's predictive accuracy country-wide due to our monitor-based split in the applied CV scheme (2.3.1 Model performance). The 1 × 1 km spatial resolution could be too coarse for some studies, especially for local and small-scale analyses. However, as we demonstrated in the case study of the city of Regensburg, the RF model of 1 × 1 km provided a valid representation of the RH spatiotemporal variation at the city scale. For future analyses, we could consider downscaling methods especially for cities (Hough et al., 2020).

For future applications, there is a potential to enhance the predictive capabilities of a RH model by augmenting its array of predictors to include re-analysis data or wind direction for instance, which were absent in our study due to the lack of appropriate data for Germany. This could be advantageous if these variables achieve higher spatial resolutions in upcoming developments. However, we do not expect considerable improvements as humidity is predominantly governed by temperature and by the vertical/horizontal mixing of wind which have already been integrated into our model. Additionally, other ML methodologies, such as eXtreme Gradient Boosting (XGBoost) or Neural Networks, could be explored if they align more effectively with distinct spatial contexts and the datasets at hand. These methodologies have been examined in the literature for various exposure scenarios (Ma et al., 2020; Tian et al., 2022).

5. Conclusion

We showed how observation, remote sensing and modelled data can be combined under a RF modeling scheme to reliably estimate RH in high temporal and spatial resolution across a country. Our product contributes substantially to reduce exposure errors for subsequent epidemiological studies, by better representing the spatiotemporal RH variability. For cohort studies using geocoded participant address information for exposure assessment, the investigation of changes over time and space is considerably improved by such a spatiotemporal model compared to relying solely on data from measurement stations. We provide a reliable RH dataset for Germany and a well-founded and generalizable approach for RH prediction for other study domains and countries.

Author contributions

Nikolaos Nikolaou: Conceptualization, Data curation, Methodology, Analysis, Visualization, Writing - original draft, Writing - review & editing. **Laurens M. Bouwer:** Conceptualization, Data curation, Methodology, Writing - review & editing. **Marco Dallavalle:** Data curation, Writing - review & editing. **Mahyar Valizadeh:** Methodology, Writing - review & editing. **Massimo Stafoggia:** Methodology, Writing - review & editing. **Annette Peters:** Conceptualization, Writing - review & editing, Supervision. **Kathrin Wolf**: Conceptualization, Data curation, Methodology, Writing - review & editing, Supervision. **Alexandra Schneider**: Conceptualization, Methodology, Writing - review & editing, Supervision.

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.

Acknowledgements

This work was supported by the Helmholtz Climate Initiative (HICAM) project, which is funded by the Helmholtz Association's Initiative and Networking Fund, the Helmholtz Information & Data Science Academy (HIDA), financially supported by the HIDA Trainee Network program, and by the Digital Earth project, supported by the Helmholtz Association's Initiative and Networking Fund (funding code ZT-0025). The authors are responsible for the content of this publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2023.117173.

References

Analitis, A., Katsouyanni, K., Biggeri, A., Baccini, M., Forsberg, B., Bisanti, L., Kirchmayer, U., Ballester, F., Cadum, E., Goodman, P.G., Hojs, A., Sunyer, J., Tiittanen, P., Michelozzi, P., 2008. Effects of cold weather on mortality: results from 15 European cities within the PHEWE project. Am. J. Epidemiol. 168 (12), 1397–1408. https://doi.org/10.1093/aje/kwn266.

Armstrong, B., 2006. Models for the relationship between ambient temperature and daily mortality. Epidemiology 624–631. http://www.jstor.org/stable/20486290.

Božič, A., Kanduč, M., 2021. Relative humidity in droplet and airborne transmission of disease. J. Biol. Phys. 47 (1), 1–29. https://doi.org/10.1007/s10867-020-09562-5.

Beck, C., Straub, A., Breitner, S., Cyrys, J., Philipp, A., Rathmann, J., Schneider, A., Wolf, K., Jacobeit, J., 2018a. Air temperature characteristics of local climate zones in the Augsburg urban area (Bavaria, southern Germany) under varying synoptic conditions. Urban Clim. 25, 152–166. https://doi.org/10.1016/j.uclim.2018.04.007.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018b. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5, 1–12. https://doi.org/10.1038/sdata.2018.214.

Bind, M.A., Zanobetti, A., Gasparrini, A., Peters, A., Coull, B., Baccarelli, A., Tarantini, L., Koutrakis, P., Vokonas, P., Schwartz, J., 2014. Effects of temperature and relative humidity on DNA methylation. Epidemiology 25 (4), 561. https://doi.org/10.1097/EDE.0000000000000120.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Davis, R.E., McGregor, G.R., Enfield, K.B., 2016. Humidity: a review and primer on atmospheric moisture and human health. Environ. Res. 144, 106–116. https://doi.org/10.1016/j.envres.2015.10.014.

Didan, K., 2015. MOD13A3 MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V006 [Data Set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13A3.006, 2022-03-21 from.

DWD, 2022a. DWD Climate Data Center (CDC): Historical daily station observations (temperature, pressure, precipitation, sunshine duration, etc.) for Germany version v21.3, 2021.

DWD, 2022b. DWD Climate Data Center (CDC): REGNIE Grids of Daily Precipitation. (Accessed 21 March 2022).

DWD, 2023. DWD Climate Data Center (CDC): Raster Data Set of Daily Sums of Precipitation in Mm for Germany - HYRAS-DE-PRE. Version v5.0.

Engebretsen, K.A., Johansen, J.D., Kezic, S., Linneberg, A., Thyssen, J.P., 2016. The effect of environmental humidity and temperature on skin barrier function and dermatitis. J. Eur. Acad. Dermatol. Venereol. 30 (2), 223–249. https://doi.org/10.1111/jdv.13301.

Forootan, E., 2019. Analysis of trends of hydrologic and climatic variables. Soil Water Res. 14 (3), 163–171. https://doi.org/10.17221/154/2018-SWR.

German National Cohort (GNC) Consortium, 2014. The German National Cohort: aims, study design and organization. Eur. J. Epidemiol. 29 (5), 371–382. https://doi.org/10.1007/s10654-014-9890-7.

Gesch, D.B., Verdin, K.L., Greenlee, S.K., 1999. New land surface digital elevation model covers the Earth. EOS 80 (6), 69–70. https://doi.org/10.1029/99EO00050.

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 https://doi.org/10.1016/j.envres.2020.109244.

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 https://doi.org/10.1016/j.envres.2021.111960.

Kloub, R.S.A.A., 2022. An optimal method for high-resolution population geo-spatial data. Comput. Mater. Contin. 73 (2) https://doi.org/10.32604/cmc.2022.027847.

Li, T., Zheng, X., Dai, Y., Yang, C., Chen, Z., Zhang, S., Wu, G., Wang, Z., Huang, C., Shen, Y., Liao, R., 2014. Mapping near-surface air temperature, pressure, relative humidity and wind speed over Mainland China with high spatiotemporal resolution. Adv. Atmos. Sci. 31 (5), 1127–1135. https://doi.org/10.1007/s00376-014-3190-8.

Li, L., Zha, Y., 2018. Mapping relative humidity, average and extreme temperature in hot summer over China. Sci. Total Environ. 615, 875–881. https://doi.org/10.1016/j.scitotenv.2017.10.022.

Luo, C., Ma, Y., Liu, Y., Lv, Q., Yin, F., 2020. The burden of childhood hand-foot-mouth disease morbidity attributable to relative humidity: a multicity study in the Sichuan Basin. China. Sci. Rep. 10 (1), 1–10. https://doi.org/10.1038/s41598-020-76421-7.

Ma, J., Yu, Z., Qu, Y., Xu, J., Cao, Y., 2020. Application of the XGBoost machine learning method in PM2. 5 prediction: a case study of Shanghai. Aerosol Air Qual. Res. 20 (1), 128–138. https://doi.org/10.4209/aaqr.2019.08.0408.

Mistry, M.N., Schneider, R., Masselot, P., Royé, D., Armstrong, B., Kyselý, J., Orru, H., Sera, F., Tong, S., Lavigne, É., Urban, A., Madureira, J., García-León, D., Ibarreta, D., Ciscar, J.-C., Feyen, L., DeSchrijver, E., Coelho, M.S.Z.S., Pascal, M., Tobias, A., , Multi-Country Multi-City (MCC) Collaborative Research Network, Guo, Y., Vicedo-Cabrera, A.M., Gasparrini, A., 2022. Comparison of weather station and climate reanalysis data for modelling temperature-related mortality. Sci. Rep. 12 (1), 1–14. https://doi.org/10.1038/s41598-022-09049-4.

Nikolaou, N., Dallavalle, M., Stafoggia, M., Bouwer, L.M., Peters, A., Chen, K., Wolf, K., Schneider, A., 2022. High-resolution spatiotemporal modeling of daily near-surface air temperature in Germany over the period 2000–2020. Environ. Res. 115062 https://doi.org/10.1016/j.envres.2022.115062.

Ou, C.Q., Yang, J., Ou, Q.Q., Liu, H.Z., Lin, G.Z., Chen, P.Y., Qian, J., Guo, Y.M., 2014. The impact of relative humidity and atmospheric pressure on mortality in Guangzhou, China. BES (Biomed. Environ. Sci.) 27 (12), 917–925. https://doi.org/10.3967/bes2014.132.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from. https://www.R-project.org/.

Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., Gratzki, A., 2013. A Central European precipitation climatology - Part I: generation and validation of a high-resolution gridded daily data set (HYRAS). Meteorol. Z. 22 (3), 235–256. https://doi.org/10.1127/0941-2948/2013/0436.

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. https://doi.org/10.1016/j.envres.2017.08.017.

Schneider, A., Rückerl, R., Breitner, S., Wolf, K., Peters, A., 2017. Thermal control, weather, and aging. Curr. Environ. Health Rep. 4 (1), 21–29. https://doi.org/10.1007/s40572-017-0129-0.

Sherwood, S.C., Ingram, W., Tsushima, Y., Satoh, M., Roberts, M., Vidale, P.L., O'Gorman, P.A., 2010. Relative humidity changes in a warmer climate. J. Geophys. Res. Atmos. 115 (D9) https://doi.org/10.1029/2009JD012585.

Silibello, C., Carlino, G., Stafoggia, M., Gariazzo, C., Finardi, S., Pepe, N., Radice, P., Forastiere, F., Viegi, G., 2021. Spatial-temporal prediction of ambient nitrogen dioxide and ozone levels over Italy using a Random Forest model for population exposure assessment. Air Qual. Atmos. Health 14 (6), 817–829. https://doi.org/10.1007/s11869-021-00981-4.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., De Hoogh, K., De'Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., Schwartz, J., 2019. Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Env. Int. 124, 170–179. https://doi.org/10.1016/j.envint.2019.01.016.

Tian, J., Liu, Y., Zheng, W., Yin, L., 2022. Smog prediction based on the deep belief-BP neural network model (DBN-BP). Urban Clim. 41, 101078 https://doi.org/10.1016/j.uclim.2021.101078.

Vermote, E.W.R., 2015. MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1kmand 500m SIN Grid V006 [Data Set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD09GA.006, 2022-03-21 from.

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

Wu, W., Tang, X.-P., Yang, C., Guo, N.-J., Liu, H.-B., 2013. Spatial estimation of monthly mean daily sunshine hours and solar radiation across mainland China. RES 57, 546–553. https://doi.org/10.1016/j.renene.2013.02.027.

Wu, W., Xu, A.-D., Liu, H.-B., 2015. High-resolution spatial databases of monthly climate variables (1961–2010) over a complex terrain region in southwestern China. Theor. Appl. Climatol. 119 (1), 353–362. https://doi.org/10.1007/s00704-014-1123-1.

Xiong, Y., Meng, Q.-S., Jie, G., Tang, X.-F., Zhang, H.-F., 2017. Effects of relative humidity on animal health and welfare. J. Integr. Agric. 16 (8), 1653–1658. https://doi.org/10.1016/S2095-3119(16)61532-0.

Yang, Y., You, E., Wu, J., Zhang, W., Jin, J., Zhou, M., Jiang, C., Huang, F., 2018. Effects of relative humidity on childhood hand, foot, and mouth disease reinfection in Hefei, China. Sci. Total Environ. 630, 820–826. https://doi.org/10.1016/j.scitotenv.2018.02.262.

Yao, R., Wang, L., Huang, X., Cao, Q., Peng, Y., 2022. A method for improving the estimation of extreme air temperature by satellite. Sci. Total Environ. 155887 https://doi.org/10.1016/j.scitotenv.2022.155887.

Zeger, S.L., Thomas, D., Dominici, F., Samet, J.M., Schwartz, J., Dockery, D., Cohen, A., 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. EHP 108 (5), 419–426. https://doi.org/10.1289/ehp.00108419.

Zeng, J., Zhang, X., Yang, J., Bao, J., Xiang, H., Dear, K., Liu, Q., Lin, S., Lawrence, W.R., Lin, A., Huang, C., 2017. Humidity may modify the relationship between temperature and cardiovascular mortality in Zhejiang Province, China. IJERPH 14 (11), 1383. https://doi.org/10.3390/ijerph14111383.

Zhang, P., Zhang, J., Chen, M., 2015. Available at: SSRN 2598810 Economic impacts of climate change on Chinese agriculture: the importance of relative humidity and other climatic variables. https://doi.org/10.2139/ssrn.2598810.