Predicting hourly, 1-kilometer relative humidity over Connecticut, 2019-2023

Kevin Chan, Statistics and Data Science, Yale University
Advised by Dr. Kai Chen and Dr. Lingzhi Chu, Yale School of Public Health

Abstract

Relative humidity is a meteorological variable commonly controlled for and used in environmental epidemiology — a growing field seeking to understand the relationships between environmental factors and human health. One common limitation of these epidemiologic studies on exposure-response relationships is the misclassification of personal exposure due to the use of monitor-level exposure observations. To mitigate such concern, we created an hourly, 1-kilometer resolution dataset of relative humidity over the state of Connecticut from 2019-2023. We used random forests to first predict the daily maximum and minimum relative humidity using satellite and weather station data. We then used the predicted daily maximum and minimum relative humidities as inputs to 24 linear fixed effects models, with each one predicting the relative humidity of a different hour of the day. Our daily maximum relative humidity predictions achieved an R^2 and RMSE of 0.54 and 8.4%, respectively. Our daily minimum relative humidity predictions achieved an R^2 and RMSE of 0.68 and 11.12% respectively. Regarding the hourly predictions, we achieved a total R^2 of 0.22 and an RMSE of 19.6%. Our approach allowed us to obtain spatially and temporally highly-resolved measurements, which will facilitate further epidemiological studies on temperature and humidity.

1. Introduction

Relative humidity is a percent measurement of the water vapor content in the air relative to the amount that would be present at full saturation. According to the National Oceanic and Atmospheric Administration, relative humidity is used to determine how hot and humid the air "feels" to us based on the combined effect of air temperature and humidity[1]. Relative humidity is also a key measurement for numerous fields such as hydrology, climatology, and agriculture. In agriculture specifically, relative humidity directly influences the water relations of plants and indirectly affects leaf growth, photosynthesis, and pollination.

In environmental epidemiology, relative humidity consistently serves as a controlled variable in time-series models when assessing how environmental variables of interest (e.g., ambient temperature) affect human health outcomes. Chen et al. controlled for relative humidity when investigating how heat exposure affected the triggering of heart attacks[2]. Qiu et al. similarly controlled for relative humidity when estimating the health effects of particulate matter on emergency hospital admissions for respiratory diseases in Hong Kong[3].

The most common limitation in such exposure response studies is exposure measurement error, which is defined as an unwanted bias or error in the measurement of a population's exposure to an environmental variable. Exposure measurement error is a reoccurring limitation stemming from a lack of high-resolution data[4]. Relative humidity data are typically obtained from weather monitors with locations irregularly distributed over space, most commonly in rural areas, thus leading to a higher chance of exposure measurement error. Small errors in an exposure assessment can have dramatic impacts on the final estimation of effects, consequently compromising the reliability of an exposure-response relationship[5]. Non-differential exposure misclassification, or when the probability of misclassification is equal amongst all groups in a study, biases the measures of effect towards the null value. Differential misclassification, or when the probability of misclassification differs between distinct groups in a study, can bias towards or away from the null value.

There is a current gap in highly resolved relative humidity measurements for multi-year timespans. A notable study that contributed to the research gap was published in 2023, where Nikolaou et al. predicted German-wide 1-kilometer by 1-kilometer daily mean relative humidity between the years 2000 to 2021[6]. The researchers were able to achieve a high accuracy of R^2 = 0.83 and RMSE = 5.07% through the use of random forests. Li et al. leveraged random forests to map relative humidity in the summer over China and achieved an R^2 = 0.70 and RMSE = 7.40%

---

[1] National Oceanic and Atmospheric Administration, "Discussion on humidity," 2024.
[2] Kai Chen, "Triggering of myocardial infarction by heat exposure is modified by medication intake," 2022.
[3] Hong Qiu, "Effects of coarse particulate matter on emergency hospital admissions for respiratory diseases: a time-series analysis in Hong Kong," 2012.
[4] EA Spencer, "Misclassification bias," 2018.
[5] National Research Council Committee on Environmental Epidemiology, "Environmental epidemiology: use of the gray literature and other data in environmental epidemiology," 1997.
[6] Nikolaos Nikolaou, "Improved daily estimates of relative humidity at high resolution across Germany: A random forest approach," 2023.

in 2018[7]. While these studies exhibit high performance with their methodologies, the temporal resolutions of their studies were only on the daily level and did not predict over the continental United States.

This study aimed to estimate relative humidity measurements on the hourly, 1-kilometer by 1-kilometer resolution across the state of Connecticut between the years 2019 and 2023. We used satellite land surface temperature, land cover, vegetation indices, elevation, and Connecticut weather station data on relative humidity. We first used random forests to generate the daily maximum and minimum relative humidities and then separately created 24 distinct fixed effects models to estimate the hourly pattern of relative humidity based on the minimum and maximum daily values. We then combined the random forest-predicted daily maximum and minimum values with our intra-day models to generate the hourly, 1-kilometer by 1-kilometer resolution data. The final product is a high-resolution dataset of relative humidity over the state of Connecticut for more accurate environmental epidemiology studies and other environmental research purposes. Additionally, our research methodology built on the current methodologies of meteorological variable estimation and serves as a proof-of-concept for our novel procedure of predicting on the hourly scale.

[7] Long Li, “Mapping relative humidity, average, and extreme temperature in hot summer over China,” 2018.

2. Methods

2.1 Study Domain

Connecticut comprises an area of 14,360 km^2 with an elevation range of -3 meters to 780 meters. The state's location on the eastern coast of North America exposes it to the moistening influence of the Atlantic Ocean and the hot and cold air masses from the interior of the continent. Connecticut's climate is characterized by cold, snowy winters and warm, humid summers. The state experiences highly variable weather patterns along with high amounts of precipitation due to its proximity to the jet stream[8]. We divided the land into 16,706 grid cells of 1-kilometer by 1-kilometer resolution with adherence to the NAD93 coordinate reference system (CRS), the most commonly used CRS by U.S. federal agencies.

2.2 Data Collection

MODIS Sensor

The Moderate Resolution Imaging Spectroradiometer (MODIS) is a satellite-based sensor used to gather earth and climate measurements to track changes in the landscape over time. There are two MODIS sensors currently in orbit around the Earth: one on the Terra (EOS AM) satellite launched by NASA on December 18th, 1999 and one on the Aqua (EOS PM) satellite launched on May 4th, 2002. The Terra and Aqua satellites began collecting data on February 24th, 2000 and May 12th, 2002 respectively[9,10]. MODIS observes the earth every 1-2 days in 36 discrete spectral bands to record high quality temperature, land cover, cloud cover, and aerosol concentration data. Furthermore, MODIS data is open-source and free for the public. In the following sections, we will discuss the MODIS products we specifically used for our analysis[11].

MOD11A1 Land Surface Temperature Data

The MODIS Terra Land Surface Temperature/Emissivity Daily (MOD11A1) Version 6.1 provides daily land surface temperature measurements in Kelvin with 1-kilometer spatial resolution over the Earth. The temporal range of MOD11A1 is from February 24th, 2000 to the present and the dataset records the meteorological conditions at 10:30 AM and 10:30 PM to provide a daytime and nighttime land surface temperature measurement. The presence of both variables was useful for our modeling purposes as there is significant variation between daytime and nighttime temperatures[12].

MYD11A1 Land Surface Temperature Data

The MODIS Aqua Land Surface Temperature/Emissivity Daily (MYD11A1) Version 6.1 is an alternative source of daily daytime and nighttime land surface temperature with 1-kilometer spatial resolution over the Earth. The temporal range of MYD11A1 is from July 4th, 2002 to the

---

[8] Jennifer Runkle, "State Climate Summaries 2022," 2023.
[9] John Maurer, "Overview of NASA's Terra Satellite," 2001.
[L]azaros Oreopoulos, "Timeline of Aqua On-Orbit Progress," 2024.
[11] NASA, "MODIS, Moderate Resolution Imaging Spectroradiometer," 2024.
[12] Zhengming Wan, "MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061," 2021.

present and the data is recorded at 1:30 PM and 1:30 AM. Like MOD11A1, MYD11A1 provides the temperature measurements in Kelvin[13].

MOD13A3 Vegetation Index Data

The MODIS Terra Vegetation Indices Monthly (MOD13A3) Version 6.1 data provides information on the vegetation conditions of earth with a 1-kilometer spatial resolution. The temporal resolution of the dataset is on a monthly level with a range from February 1st, 2000 to the present. Each measurement is a weighted temporal average through the month and is appropriate for our study as vegetation does not considerably change during the timespan of a single month. The dataset quantifies vegetation conditions using the Normalized Difference Vegetation Index (NDVI)[14]. NDVI ranges from -1 to +1, with values closer to +1 indicating dense green leaves and values closer to -1 indicating water. An NDVI value close to zero suggests an urbanized area. The formula for NDVI is as follows:

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

where *NIR* is the near-infrared light in nanometers and *Red* is the red band in nanometers. Since healthy vegetation reflects more NIR and absorbs more red light, NDVI is a robust way of measuring the quantity of vegetation over an area of land[15].

MCD12Q1 Land Cover Data

The MODIS Terra and Aqua combined Land Cover Type (MCD12Q1) Version 6.1 provides information on the global land cover types at yearly iterations on the 500 m resolution. Land cover describes the types of land surfaces on the earth, ranging from forests to urban areas to water bodies. For the purposes of our research, we use the Type 1 Annual International Geosphere-Biosphere Programme classification with a categorical variable range of 1 to 17, with 1 representing evergreen needleleaf forests and 17 representing water bodies (Table S1)[16]. We combined similar land cover types by reclassifying the 17 labels into five types: vegetation, water, urban, barren, and snow and ice[17].

The temporal extent of MCD12Q1 begins on 01/01/2001 and ends on 12/31/2022. Since the MODIS team is currently working on the release of the 2023 land cover data, we have supplemented the lack of data for that year by using 2022's data for our modeling purposes. When the 2023 data is released, we will incorporate the new data into our modeling for more accurate results.

---

[13] Zhengming Wan, "MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061," 2021.
[14] Kamel Didan, "MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V061," 2021.
[15] Kamel Didan, "MODIS Collection 6.1 (C61) VegetationIndex Product UserGuide," 2019.
[16] Mark Friedl, "MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061," 2022.
[17] Zhihao Jin, "Predicting Spatiotemporally-Resolved Mean Air Temperature over Sweden from Satellite Data Using an Ensemble Model," 2022.

ERA5 Dry Bulb and Dewpoint Temperature Data

ERA5-Land is a dataset that provides dry bulb temperature and dewpoint temperature data on the hourly scale with a 10-kilometer spatial resolution. The temporal coverage of the dataset ranges from January 1950 to the present. ERA5, by nature, is a reanalysis dataset that combines model data with observations from across the world into a globally complete and consistent dataset using physical laws. Dry bulb and dewpoint temperature are both reported in Kelvin[18].

ASTGTM Elevation Data

The Terra Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) Version 3 provides a global dataset of the Earth's elevation at a spatial resolution of approximately 30 meters. While the time range of the dataset is from 03/01/2000 to 11/30/2013, elevation is not an environmental variable that drastically changes over ten years. Thus, for our study period of 2019 to 2023, we used the elevation data from 2013. The elevation measurements are recorded in meters[19].

U.S. Local Climatological Relative Humidity Data

The Local Climatological Data (LCD) are summaries of the climatological conditions from the different weather stations managed by the National Weather Service (NWS), Federal Aviation Administration (FAA), and Department of Defense (DOD). We used the hourly relative humidity measurements from 10 distinct weather stations[20].

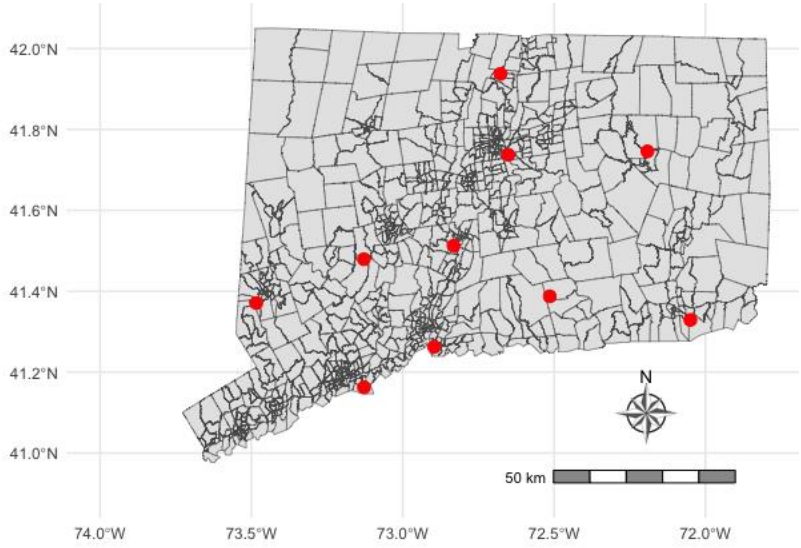

Figure 1: Geographic Locations of the 10 Weather Stations

The data includes a number of quality control checks and records other meteorological variables at the same temporal resolution. The summary statistics of the relative humidity measurements are shown in **Table 1**.

[18] Muñoz Sabater, "ERA5-Land Hourly Data from 1950 to Present," 2019.
[19] U.S./Japan ASTER Science Team, "ASTER Global Digital Elevation," 2019.
[20] National Oceanic and Atmospheric Administration, "U.S. Local Climatological Data (LCD)," 2024.

Table 1: Summary statistics of hourly relative humidity from weather stations, 2019-2023

Variables	Mean	SD	Min	P25	P50	P75	Max
Hourly Relative Humidity	71.50	20.49	8.00	56.00	76.00	89.00	100.00
Daytime Relative Humidity	64.34	21.09	8.00	47.00	65.00	83.00	100.00
Nighttime Relative Humidity	78.69	16.76	13.00	68.00	84.00	93.00	100.00

Notes: We defined daytime as 8 AM to 8 PM and defined nighttime as 8 PM to 8 AM, and the locations of the monitors are Chester Airport, Meriden Markham Municipal Airport, Oxford Airport, Igor I. Sikorsky Memorial Airport, New Haven Tweed Airport, Groton New London Airport, Hartford Bradley International Airport, Willimantic Windham Airport, Danbury Municipal Airport, and Hartford Brainard Airport.

To explore whether the daily maximum and minimum relative humidities could be modeled, we created weighted histograms to view the distributions of the hours of the day when the daily minimum and maximum relative humidities occurred (**Figures 2-3**).

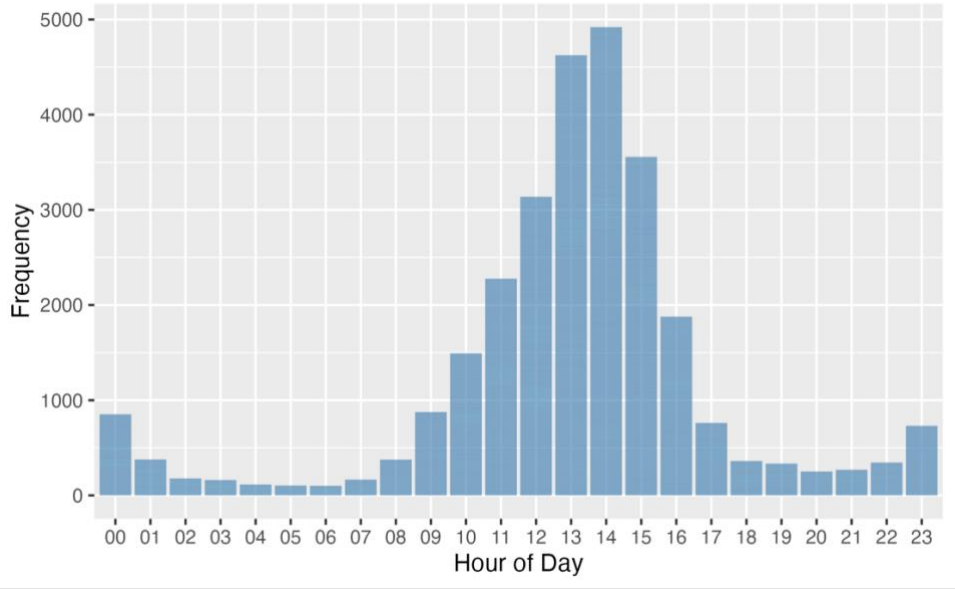

Figure 2: Weighted Histogram of Daily Minimum Relative Humidity Occurrences

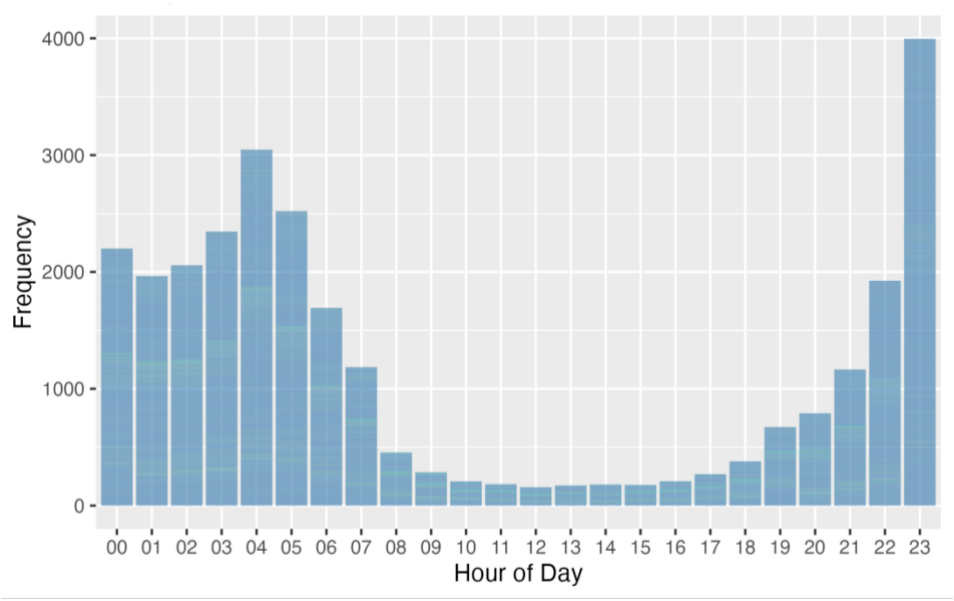

Figure 3: Weighted Histogram of Daily Maximum Relative Humidity Occurrences

We used a weighted histogram since multiple hours of the day can share the minimum or maximum relative humidity. From our figures, we can see that the daily minimum relative humidity tends to occur in the middle of the day while the daily maximum relative humidity tends to occur during the night. Since these histograms resemble a unimodal distribution, we knew that the daily minimum and maximum relative humidities can be modeled.

2.3 Data Cleaning and Processing

Resampling Step

The resampling process standardizes dissimilar geographical raster objects to ensure consistent spatial resolutions. Since our datasets are not uniform in resolution (e.g., MODIS is 1-km and ASTGTM is 30-m), we must perform the resampling procedure with respect to MOD11A1 since the dataset has a spatial resolution of 1-kilometer by default. We have two methods of resampling:

- If the spatial resolution of the dataset we want to resample is of higher resolution than the 1-kilometer MOD11A1 data, we perform resampling with the bilinear method where the data points that overlap with the 1-kilometer data are averaged together into one value.
- If the spatial resolution of the dataset we want to resample is of lower resolution than MOD11A1, then we perform resampling with the nearest neighbors method where for the instances of overlap with the lower resolution data set, the higher resolution data values are assigned the value of the closest lower resolution dataset area.

Replacing Missing Values in MOD11A1 and MYD11A1

We have 61.72% and 66.27% of the values missing in daily MOD11A1 daytime and nighttime respectively. Similarly, we have 82.32% and 83.12% of the values missing in the daily MYD11A1 daytime and nighttime measurements. Considering the large proportion of missing values for land surface temperature, we must devise a method to fill in those values.

To address the missing values in MOD11A1 and MYD11A1, we took the resampled ERA-5 air temperature data, assumed random effects in the relationship between the ERA5 air temperature and the satellite land surface temperature variable by day of year, and fit the following linear mixed effects models:

$$LST_{Night} \sim ERA_{NightMean} + \left(ERA_{NightMean} \middle| doy\right) \quad (2)$$

$$LST_{Day} \sim ERA_{DayMean} + \left(ERA_{DayMean} \middle| doy\right) \quad (3)$$

where LST_{Night} represents the nighttime land surface temperature measurements and $ERA_{NightMean}$ is the mean temperature over the hours of 11 PM to 7 AM in the ERA5 dataset. Similarly, LST_{Day} is the daytime land surface temperature measurement and $ERA_{DayMean}$ is the mean temperature over the hours of 7 AM to 11 PM in the ERA5 dataset. Thus, we are assuming spatial-invariant relationships between the land surface temperature and the ERA5 temperature[21].

Table 2: Summary statistics of environmental variables for Connecticut, 2019-2023

Variables	Mean	SD	Min	P25	P50	P75	Max
MOD11A1 Day Temp (°C)	15.68	10.41	-17.59	6.87	16.97	23.91	50.83
MOD11A1 Night Temp (°C)	6.78	9.19	-27.15	-1.11	6.58	14.96	28.03
MYD11A1 Day Temp (°C)	16.73	10.89	-19.35	7.83	18.19	25.10	56.13
MYD11A1 Night Temp (°C)	6.09	8.75	-21.96	-0.97	5.97	13.83	27.87
MOD13A3 NDVI	0.49	0.16	0.01	0.37	0.47	0.62	0.88
MCD12Q1 LC Vegetation	36.82	37.17	0	0	25	50	100
MCD12Q1 LC Water	7.57	16.07	0	0	0	0	50
MCD12Q1 LC Urban	55.61	36.53	0	25	50	100	100
MCD12Q1 LC Barren	0	0	0	0	0	0	0
MCD12Q1 LC Snow	0	0	0	0	0	0	0
ASTGTM Elevation (meters)	68.42	66.52	4.79	8.67	44.82	117.64	210.07

Notes: The temperature summary statistics are calculated after replacing the missing values. The land coverage variables do not take the year 2023 into account, so they are calculated over 2019 to 2022. P25 = the 25th percentile, P50 = the 50th percentile, P75 = the 75th percentile. All values were rounded to the hundredths place.

[21] Douglas Bates, "Fitting Linear Mixed-Effects Models Using Lme4," 2015.

2.4 Statistical Analysis

Random Forest

Random forest is a supervised ensemble machine learning algorithm capable of solving classification and regression tasks by utilizing the bagging principle. In the case of regression, random sub-samples of the dataset are selected with replacement. The algorithm then constructs a decision set of decision trees, one for each sub-sample, with each tree generating an output of the target variable. The outputs of each tree can then be averaged to retrieve the main model's output. Random forests tend to work especially well with large amounts of data and in cases of multicollinearity and overfitting. The algorithm is also considerably robust against outliers[22].

For our study, we trained two Breiman randoms forests models, one to model the daily minimum relative humidity and another to model the daily maximum relative humidity, with both on the 1-kilometer level. Our response variables were the observed daily maximum and minimum relative humidities at each weather stations. The robustness of the random forest algorithm alleviates the need to intensely tune the hyperparameters so we ran our models using trial and error with respect to variable selection. The predictors of our models were the MOD11A1 and MYD11A1 land surface temperature in Celsius, MOD13A3 NDVI, ASTGTM elevation in meters, and MCD12Q1 land cover categorical values. We incorporated the geographical information of each 1-kilometer square by including longitude and latitude in degrees to account for spatial variations not fully covered by other features in the model. We included the day of the year as well to capture the daily variations in the response-predictor variables' relationship. Similarly, we included the year to capture the yearly variations in the relationship.

Random Forests Ten-Fold Cross-Validation

We used ten-fold cross-validation to assess the performances of our random forests models by dividing the set of weather station monitors into a training and a testing set. Due to the low number of weather stations, to estimate daily maximum and minimum relative humidity, we created 10 different models for each. Each model utilized the data from 9 weather stations as the training set and the data from 1 weather station as the testing set.

At each training iteration, we calculated total R^2, spatial R^2, temporal R^2, total root mean-squared error (RMSE), spatial RMSE, and temporal RMSE and then averaged the metrics across the models at the end of training for final values.

R^2 is calculated as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (4)$$

[22] Leo Breiman, "Random Forests," 2001.

and we calculated RMSE with:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

where *y* is the observed daily maximum or minimum relative humidity, $\hat{y}_i$ is the predicted value for the dataset entry, $\bar{y}$ is the mean of all our observed values, and *n* is the number of data points.

To calculate temporal R^2 and RMSE, we first started with our complete set of observation data for both daily minimum and maximum relative humidity. Taking daily minimum relative humidity as an example, we first grouped the values by their grid ID and year, so for a single grid's daily minimum relative humidity value ***RH***, we have the following set:

$$\boldsymbol{RH}_{i,2019}, \boldsymbol{RH}_{i,2020}, \boldsymbol{RH}_{i,2021}, \boldsymbol{RH}_{i,2022}, \boldsymbol{RH}_{i,2023} \quad (6)$$

where *i* represents the grid ID, a unique classifier for each 1-kilometer by 1-kilometer square in our data set, and *2019,…,2023* represent the years in our study period. With the five groupings by year per grid id, we then calculated the mean within each grouping to obtain:

$$\overline{\boldsymbol{RH}}_{i,2019}, \overline{\boldsymbol{RH}}_{i,2020}, \overline{\boldsymbol{RH}}_{i,2021}, \overline{\boldsymbol{RH}}_{i,2022}, \overline{\boldsymbol{RH}}_{i,2023} \quad (7)$$

for each grid id *i*. We then subtracted each element in $\boldsymbol{RH}_{i,year}$ with the corresponding $\overline{\boldsymbol{RH}}_{i,year}$. We thus have:

$$\begin{gathered} \boldsymbol{RH}_{i,2019} - \overline{\boldsymbol{RH}}_{i,2019}, \\ \boldsymbol{RH}_{i,2020} - \overline{\boldsymbol{RH}}_{i,2020}, \\ \boldsymbol{RH}_{i,2021} - \overline{\boldsymbol{RH}}_{i,2021}, \quad (8) \\ \boldsymbol{RH}_{i,2022} - \overline{\boldsymbol{RH}}_{i,2022}, \\ \boldsymbol{RH}_{i,2023} - \overline{\boldsymbol{RH}}_{i,2023} \end{gathered}$$

or

$$\boldsymbol{RH}_{i,year} - \overline{\boldsymbol{RH}}_{i,year} = \Delta \boldsymbol{t}_{i,year} \quad (9)$$

and we performed the same operation for the predicted values:

$$\widehat{\boldsymbol{RH}}_{i,year} - \overline{\widehat{\boldsymbol{RH}}}_{i,year} = \Delta \hat{\boldsymbol{t}}_{i,year} \quad (10)$$

We then performed the R^2 and RMSE calculations with $\Delta \boldsymbol{t}_{i,year}$ and $\Delta \hat{\boldsymbol{t}}_{i,year}$ across all of the values in the data set to obtain the temporal versions of the evaluation metrics. We calculated

the spatial versions of the metrics by grouping the observations and predictions by grid id and year, and then calculating the average of each year. We took both:

$$\overline{\boldsymbol{RH}}_{i,2019}, \overline{\boldsymbol{RH}}_{i,2020}, \overline{\boldsymbol{RH}}_{i,2021}, \overline{\boldsymbol{RH}}_{i,2022}, \overline{\boldsymbol{RH}}_{i,2023} \quad (11)$$

and

$$\overline{\widehat{\boldsymbol{RH}}}_{i,2019}, \overline{\widehat{\boldsymbol{RH}}}_{i,2020}, \overline{\widehat{\boldsymbol{RH}}}_{i,2021}, \overline{\widehat{\boldsymbol{RH}}}_{i,2022}, \overline{\widehat{\boldsymbol{RH}}}_{i,2023} \quad (12)$$

and calculated the R^2 and RMSE between the grid id and year groupings to obtain the spatial versions of the evaluation metrics.

Calculating Hourly ERA5 Relative Humidity

We used the hourly ERA5 dry bulb and dewpoint temperatures to calculate the hourly relative humidity for each 10-kilometer unit with the following formula:

$$RH = 100 \times \left[\frac{e^{\frac{17.625 \times D_p}{243.04 + D_p}}}{e^{\frac{17.625 \times T}{243.04 + T}}} \right] \quad (13)$$

where *RH* is the relative humidity in percent (%), D_p is the dewpoint temperature in Celsius (°C), and *T* is the dry bulb temperature in Celsius (°C)[23].

Fixed Effects Models and Random Effects Models

A fixed effects model is a statistical model that represents the observed quantities in terms of explanatory variables that are treated as if the quantities were non-random. Similarly, a random effects model treats all of the explanatory variables as if they arise from random processes[24]. In order to approximate each hour of the day's relative humidity, we assumed that the differences in day of year, year, and location all either have non-random or random effects on the daily maximum and minimum relative humidities. We separately used both fixed effects and random effects models to estimate the hourly relative humidity and evaluated the performance of both model types.

We modeled the hourly relative humidity calculated from the ERA5 dataset using fixed effects models with respect to the daily maximum and minimum relative humidity within each 10-kilometer unit with the following formula:

$$RH_{Hour} \sim RH_{DailyMax} + RH_{DailyMin} + RH_{DailyMax} * doy + RH_{DailyMin} * doy + RH_{DailyMax} * year + RH_{DailyMin} * year + RH_{DailyMax} * ERA_{Grid} + RH_{DailyMin} * ERA_{Grid} \quad (14)$$

[23] Purnima Singh, "Relative Humidity Calculator," 2024.

[24] Lu Liu, "Effect of Problem-Based Learning in Pharmacology Education: A Meta-Analysis," 2019.

where RH_{Hour} represents the relative humidity at a specific hour of the day ranging from 12 AM to 11 PM, $RH_{DailyMax}$ and $RH_{DailyMin}$ represent the 10-kilometer space's daily maximum and minimum relative humidity, *doy* is a categorical variable representing day of year ranging from 1 to 366 due to the leap year in 2020, *year* is a categorical variable representing the year with a range of 2019-2023, and ERA_{Grid} is a categorical variable distinguishing each 10-kilometer grid in the ERA5 dataset. In the model above, we treated the daily maximum and minimum relative humidities as singular variables and we then assumed the presence of interactions with those variables based on the day of year, year, and the grid location. To model the 24 hours of the day, we created 24 separate models for each hour.

We additionally trained 24 random effects models to estimate the hourly relative humidities with the following equation:

$$RH_{Hour} \sim RH_{DailyMax} + RH_{DailyMin} + (RH_{DailyMax}|doy) + (RH_{DailyMin}|doy) + (RH_{DailyMax}|year) + (RH_{DailyMin}|year) + (RH_{DailyMax}|ERA_{Grid}) + (RH_{DailyMin}|ERA_{Grid}) \quad (15)$$

where we assumed that the year, day of year, and grid location factors have random effects on the daily maximum and minimum relative humidities.

Combining Random Forests Output and Hourly Modeling

We combined our random forests-generated daily maximum and minimum relative humidities on the 1-kilometer scale with the 24-hourly relative humidity models trained on 10-kilometer ERA5 data to predict the hourly relative humidity on the 1-kilometer scale. Here, we assumed that the relationship between the hourly relative humidity and the daily minimum and maximum relative humidities do not change based on spatial resolution.

3. Results

Our best random forest models utilized the following variables for each weather station:

$$MinRH \sim RF(MOD_{Day}, MOD_{Night}, MYD_{Day}, MYD_{Night}, NDVI, LC_{Veg}, LC_{Water}, LC_{Urban}, LC_{barren}, LC_{Snow}, Elev, Lon, Lat, doy, Year) \quad (16)$$

$$MaxRH \sim RF(MOD_{Day}, MOD_{Night}, MYD_{Day}, MYD_{Night}, NDVI, LC_{Veg}, LC_{Water}, LC_{Urban}, LC_{barren}, LC_{Snow}, Elev, Lon, Lat, doy, Year) \quad (17)$$

where MOD_{Day} and MOD_{Night} are the daily day and night temperatures from MOD11A1 respectively, MYD_{Day} and MYD_{Night} are the daily day and night temperatures from MYD11A1 respectively, $NDVI$ is the normalized difference vegetation index, and $LC_{Veg}, LC_{Water}, LC_{Urban}, LC_{barren}, LC_{Snow}$ denote the presence of vegetation, water, urban, barren, and snow land covers, respectively, over the weather stations. $Elev$ is the elevation, Lon and Lat are the longitude and latitude values, doy is the day of the year, and $Year$ is the year. Our final model used 500 trees and 5 randomly sampled variables as candidates at every split.

Table 3: Random forest metrics for predicting daily minimum and maximum relative humidity

Predictor	Total R^2	Spatial R^2	Temporal R^2	Total RMSE	Spatial RMSE	Temporal RMSE
Min RH	0.68	0.65	0.68	11.12	3.59	10.34
Max RH	0.54	0.33	0.55	8.4	1.93	8.13

Our fixed effects models and random effects models returned the following results:

Table 4: Final fixed and random effects model metrics for predicting hourly relative humidity

Model Type	Total R^2	Slope Estimate	Intercept Estimate (%)	Total RMSE
Fixed effects	0.222	0.596	28.187	19.597
Random effects	0.046	0.092	65.881	48.773

We obtained the slope and intercept estimates by fitting a simple linear regression:

$$lm(HourlyRelativeHumidity \sim PredictedHourlyRelativeHumidity) \quad (18)$$

and extracting the slope and intercept.

The data cleaning and analysis were performed entirely in R with the Fast Fixed-Effects Estimations (fixest) and the Linear Mixed-Effects Models using 'Eigen' and S4 (lme4) packages[25,26].

[25] Douglas Bates, "Fitting Linear Mixed-Effects Models Using Lme4," 2024.

[26] Laurent Bergé, "Efficient Estimation of Maximum Likelihood Models with Multiple Fixed-Effects: The R Package FENmlm," 2024.

4. Discussion

Our study proposed a novel approach for finer-resolution spatial and temporal modeling of relative humidity by using random forests to predict the daily maximum and minimum relative humidity and then using 24 fixed effects models to predict the relative humidity for every hour based on the daily values. Our methodology leveraged several satellite data sources to perform more complex modeling than conventional interpolation methods. We produced a high-resolution dataset of relative humidity over the state of Connecticut on the 1-kilometer by 1-kilometer scale with hourly temporal resolution. While our fixed effects models performed the best, they achieved objectively lower performance (R^2 = 0.22 and RMSE = 19.6%) when compared to previous work in the field.

Nikolaou et al.'s 2023 study leveraged random forests and weather station data to predict German-wide 1-kilometer by 1-kilometer daily mean relative humidity between 2000 to 2021[27]. Their methodology achieved high accuracy (R^2 = 0.83, RMSE = 5.07%) via ten-fold cross-validation and a comparison of their predictions with monitoring stations in Augsburg confirmed the high performance ($R^2 \geq 0.86$, RMSE $\leq$ 5.45%). Kloub et al.'s 2022 study utilized an autoencoder-based residual neural network to predict daily relative humidity using data from 824 monitoring stations across mainland China in 2015[28]. The highest performing model for their predictive task was one that incorporated latitude, longitude, elevation, day of year, and nearest neighbor value to obtain an R^2 of 0.86 and an RMSE of 7.41%. We note that the primary inputs to our fixed and random effects models were daily maximum and minimum relative humidity, which came from random forests that achieved R^2 scores of 0.54 and 0.68 and RMSE scores of 8.4% and 11.12% respectively. Since the performance of our random forest models were not initially high, underperformance is natural for our hourly modeling step. Additionally, relative humidity sees great spatial and temporal variability, making prediction of the weather variable more difficult and data-intensive than dewpoint temperature and dry-bulb temperature[29].

The most prominent limitation of our study was the lack of weather monitoring stations within the state of Connecticut. Our LCD data is limited due to the existence of only 10 weather stations which track relative humidity in the state and we note that the majority of these stations are located at airports. Nikalaou et al.'s study had access to 406 weather monitoring stations and Kloub et al. had access to 824 monitoring stations[30]. The research groups further incorporated variables such as wind speed, precipitation, and color band; factors which we plan to incorporate into our data in the near future to improve random forest performance. However, while we had a lack of monitoring stations, our dataset was by no means small. We have the hourly relative humidity data for 1826 days per weather station, resulting in a robust

---

[27] Nikolaos Nikolaou, "Improved Daily Estimates of Relative Humidity at High Resolution across Germany: A Random Forest Approach," 2023.

[28] Rami Sameer Ahmad Al Kloub, "An Optimal Method for High-Resolution Population Geo-Spatial Data," 2022.

[29] Md. Abdul Fattah, "Spatiotemporal Characterization of Relative Humidity Trends and Influence of Climatic Factors in Bangladesh," 2023.

[30] Nikolaos Nikolaou, "Improved Daily Estimates of Relative Humidity at High Resolution across Germany: A Random Forest Approach," 2023.

20,086 unique measurements. The addition of more weather stations would allow us to more accurately capture the geographical variations of relative humidity in Connecticut.

Another possible limitation lies with the resampling procedure. MODIS predictor data with finer resolution than 1-kilometer are subject to resampling errors and data with lower resolution, such as the ERA5 dataset, are susceptible to spatial interpolation errors. Despite these possible flaws in the satellite data, there is extensive literature supporting the robustness and quality of these datasets. The 1-kilometer by 1-kilometer spatial resolution in our final dataset can also be too coarse for local and small-scale analyses, but our procedure provides a valid representation of relative humidity's spatial variation and improves upon other methodologies in the temporal resolution dimension.

Despite our low performance, our study contributes a new approach for obtaining finer temporal meteorological measurements to address exposure measurement error. While Kloub et al. and Nikalaou et al. achieved higher accuracy, their methodologies only predicted daily average relative humidity whereas we aimed to predict hourly relative humidity. Our study can also be easily adapted to other geographical locations and time periods thanks to the global and temporal extent of the MODIS, ASTGTM, and ERA5 datasets. As long as the study location contains numerous weather stations spanning the timeframe, our methodology can be applied.

Relative humidity is a common control for exposure-response studies, however the meteorological variable's association with human health has not been well-studied. Most studies combine the effects of relative humidity and temperature when assessing the association with human health. Lin et al.'s 2009 paper investigated the effects of temperature and humidity on daily cardiovascular and respiratory hospitalization counts during the summers between 1991-2004[31]. Davis et al.'s 2003 paper investigated the association of extreme temperature and humidity on mortality in 28 major metropolitan areas in the United States from 1964 to 1998[32]. The researchers found evidence of extreme temperature and humidity increasing mortality rates in the 1960s and 1970s, but no evidence of excess mortality in the 1980s and 1990s. They attributed the mass desensitization of high heat and humidity to technological and biophysical adaptations, most notably the increased availability of air conditioning. Through our study, we hope to contribute higher resolution data for the purposes of conducting more accurate exposure-response research.

[31] Shao Lin, "Extreme High Temperatures and Hospital Admissions for Respiratory and Cardiovascular Diseases," 2009.
[32] Robert E. Davis, "Changing Heat-Related Mortality in the United States," 2003.

Future Work

We plan to incorporate satellite data on color band and existing reanalysis data on precipitation, wind speed, and solar radiation. We further plan to incorporate the year 2024's measurements to create hourly relative humidity predictions for the year. Through the Morse College Senior Thesis Mellon Grant, we received funding to buy two HOBO Temperature/Relative Humidity data loggers. The loggers record hourly relative humidity and we have placed one at Morse College and another at an associate's home in Hamden. The monitors have been collecting data between the months of March and April and will serve as external validation for our 2024 relative humidity predictions.

We further recommend testing other machine learning methodologies, such as using eXtreme Gradient Boosting or a neural network, when predicting the daily maximum and minimum relative humidities. While random forests are the most common method in the field to estimate meteorological factors, there has been extensive work utilizing other machine learning techniques for prediction[33]. Another avenue of further research is expanding the geographical or temporal ranges of the study to encompass more of the United States or more historic years. Due to time and computational constraints, we were only able to obtain the relative humidity data for Connecticut between 2019-2023.

5. Conclusion

We showed how environmental satellite and weather station data can be used to estimate relative humidity on a high spatial and temporal resolution. Once we achieve higher accuracy, our dataset will be useful towards reducing exposure misclassification in future epidemiological, environmental, and ecological studies over the state of Connecticut and open the door to higher-resolution studies. Additionally, our methodology is highly applicable to geographic locations within and outside of the US thanks to the spatial range of the satellite data and is applicable to predicting other environmental variables.

[33] Kloub, "An Optimal Method for High-Resolution Population Geo-Spatial Data," 2022.

References (In Order of Reference):

1) National Oceanic and Atmospheric Administration. “Discussion on humidity.” Accessed April 25, 2024. Available from: https://www.weather.gov/lmk/humidity
2) Chen et al. “Triggering of myocardial infarction by heat exposure is modified by medication intake.” *Nature Cardiovascular Research* 1, (2022): 727-731. https://www.nature.com/articles/s44161-022-00102-z
3) Qiu et al. “Effects of coarse particulate matter on emergency hospital admissions for respiratory diseases: a time-series analysis in Hong Kong.” *Environmental Health Perspectives* 120, 4 (2012): 572-576. https://ehp.niehs.nih.gov/doi/10.1289/ehp.1104002
4) Spencer et al. “Misclassification bias.” *Catalogue of Bias* 2018. https://catalogofbias.org/biases/misclassification-bias/
5) National Research Council (US) Committee on Environmental Epidemiology. “Environmental epidemiology: volume 2: use of the gray literature and other data in environmental epidemiology.” *National Academies Press* 1997. Available from: https://www.ncbi.nlm.nih.gov/books/NBK233635/
6) Nikolaou et al. “Improved daily estimates of relative humidity at high resolution across Germany: A random forest approach.” *Environmental Research* 238, 2 (2023). https://doi.org/10.1016/j.envres.2023.117173
7) Li, Long and Zha, Yong. “Mapping relative humidity, average, and extreme temperature in hot summer over China.” *Science of The Total Environment* 615, (2018): 875-881. https://doi.org/10.1016/j.scitotenv.2017.10.022
8) Runkle et al. “State climate summaries 2022: Connecticut.” *NOAA National Centers For Environmental Information* 2023. Accessed April 24, 2024. Available from: https://statesummaries.ncics.org/chapter/ct/#:~:text=Key%20Message%203,rises%20are%20possible%20for%20Connecticut
9) Maurer, John. “Overview of NASA’s Terra satellite.” *University of Hawai’i at Mãnoa* 2001. Accessed May 2, 2024. Available from: https://www2.hawaii.edu/~jmaurer/terra/
10) Oreopoulos, Lazaros. “Timeline of Aqua on-orbit progress.” *NASA*. Accessed May 5, 2024. Available from: https://aqua.nasa.gov/timeline-aqua-orbit-progress
11) National Aeronautics and Space Administration. “MODIS: Moderate resolution imaging spectroradiometer.” *TERRA The EOS Flagship*. Accessed May 2, 2024. Available from: https://terra.nasa.gov/about/terra-instruments/modis
12) Wan et al. “MODIS/Terra land surface temperature/emissivity daily L3 global 1km SIN grid V061.” *NASA EOSDIS Land Processes Distributed Active Archive Center*, 2021. Accessed 04/15/2024. Available from: https://doi.org/10.5067/MODIS/MOD11A1.061
13) Wan et al. “MODIS/Aqua land surface temperature/emissivity daily L3 global 1km SIN grid V061.” *NASA EOSDIS Land Processes Distributed Active Archive Center*, 2021. Accessed 04/15/2024. Available from: https://doi.org/10.5067/MODIS/MYD11A1.061
14) Didan et al. “MODIS/Terra vegetation indices monthly L3 global 1km SIN grid V061.” *NASA EOSDIS Land Processes Distributed Active Archive Center*, 2021. Accessed 04/15/2024. Available from: https://doi.org/10.5067/MODIS/MOD13A3.061
15) Didan, Kamel and Munoz, Armando Barreto. “MODIS vegetation index user’s guide.” *MOD13 Series*, 2019. Accessed 04/26/2024. Available from: https://lpdaac.usgs.gov/documents/621/MOD13_User_Guide_V61.pdf

16) Friedl et al. “MODIS/Terra+Aqua land cover type yearly L3 global 500m SIN Grid V061.” *NASA EOSDIS Land Processes Distributed Active Archive Center*, 2022. Accessed 04/15/2024. Available from: https://doi.org/10.5067/MODIS/MCD12Q1.061
17) Jin et al. “Predicting spatiotemporally-resolved mean air temperature over Sweden from satellite data using an ensemble model.” *Environmental Research* 204, A (2022). https://doi.org/10.1016/j.envres.2021.111960
18) Muñoz Sabater, J. “ERA5-Land hourly data from 1950 to present.” *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 2019. Accessed 04/15/2024. Available from: https://doi.org/10.24381/cds.e2161bac
19) NASA/METI/AIST/Japan Spacesystems and U.S/Japan ASTER Science Team. “ASTER Global digital elevation model V003.” *NASA EOSDIS Land Processes Distributed Active Archive Center*, 2019. Accessed 04/15/2024. Available from: https://doi.org/10.5067/ASTER/ASTGTM.003
20) National Centers for Environmental Information et al. “U.S. local climatological data (LCD).” *National Oceanic and Atmospheric Administration*, 1931. Accessed 04/15/2024. Available from: https://www.ncei.noaa.gov/access/search/data-search/local-climatological-data
21) Bates et al. “Fitting linear mixed-effects models using lme4.” *Journal of Statistical Software* 67, 1 (2015): 1-48. https://doi.org/10.18637/jss.v067.i01
22) Breiman, Leo. “Random Forests.” *Machine Learning* 45, (2001): 5-32. https://doi.org/10.1023/A:1010933404324
23) Singh, P. “Relative humidity calculator.” Accessed 04/26/2024. Available from: https://www.omnicalculator.com/physics/relative-humidity
24) Liu et al. “Effect of problem-based learning in pharmacology education: A meta-analysis.” *Studies in Educational Evaluation* 60, (2019): 43-58. https://doi.org/10.1016/j.stueduc.2018.11.004
25) Kloub, Rami Sameer Ahmad. “An optimal method for high-resolution population geo-spatial data.” *Computers, Materials, and Continua* 73, 2 (2022): 2801-2820. https://doi.org/10.32604/cmc.2022.027847
26) Fattah et al. “Spatiotemporal characterization of relative humidity trends and influence of climatic factors in Bangladesh.” *Heliyon* 9, 9 (2023). https://doi.org/10.1016/j.heliyon.2023.e19991
27) Lin et al. “Extreme high temperatures and hospital admissions for respiratory and cardiovascular diseases.” *Epidemiology* 20, 5 (2009): 738-746. https://doi.org/10.1097/EDE.0b013e3181ad5522
28) Davis et al. “Changing heat-related mortality in the United States.” *Environmental Health Perspectives* 111, 14 (2003): 1712-1718. https://doi.org/10.1289/ehp.6336

Supplemental Table 1 (Table S1): MCD12Q1 International Geosphere-Biosphere Programme Land Cover Legend and Class Descriptions

Name	Value	Description
Evergreen Needleaf Forests	1	Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.
Evergreen Broadleaf Forests	2	Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.
Deciduous Needleaf Forests	3	Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%.
Deciduous Broadleaf Forests	4	Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.
Mixed Forests	5	Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60%.
Closed Shrublands	6	Dominated by woody perennials (1-2m height) >60% cover.
Open Shrublands	7	Dominated by woody perennials (1-2m height) 10-60% cover.
Woody Savannas	8	Tree cover 30-60% (canopy >2m).
Savannas	9	Tree cover 10-30% (canopy >2m).
Grasslands	10	Dominated by herbaceous annuals (<2m).
Permanent Wetlands	11	Permanently inundated lands with 30-60% water cover and >10% vegetated cover.
Croplands	12	At least 60% of area is cultivated cropland.
Urban and Built-up Lands	13	At least 30% impervious surface area including building materials, asphalt, and vehicles.
Cropland/Natural Vegetation Mosaics	14	Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation.
Permanent Snow and Ice	15	At least 60% of area is covered by snow and ice for at least 10 months of the year.
Barren	16	At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation.
Water Bodies	17	At least 60% of area is covered by permanent water bodies.
Unclassified	255	Has not received a map label because of missing inputs.

Supplementary Table 2 (Table S2): Reclassification of the International Geosphere-Biosphere Programme Land Cover Type

IGBP Class Name	Reclassified Category
Evergreen Needleleaf Forests Evergreen Broadleaf Forests Deciduous Needleleaf Forests Deciduous Broadleaf Forests Mixed Forests Closed Shrublands Open Shrublands Woody Savannas Savannas Grasslands Croplands Cropland/Natural Vegetation Mosaics	Vegetation
Urban and Built-up Lands	Urban
Permanent Snow and Ice	Snow and Ice
Barren	Barren
Permanent Wetlands Water Bodies	Water