Can ML Models Predict What Areas of the Planet Will Have the Biggest Temperature Changes?

Data 200 Foundations of Data Analytics

Dalit Hendel

Thursday December 16th, 2021

Abstract:

The aim of this analysis is to compare the accuracy of a Random Forest and a Neural Net model both trained on WorldClim (historic and future) data. The models use training and test data comprised of the variables: longitude, latitude, elevation, historic mean temperature, historic precipitation, and a binary variable representing extreme weather changes. The model is trained to identify land surface areas that have had an absolute value temperature change of more than 1.5°C. It was hypothesized that the Random Forest model would have the greatest accuracy but after training the models with repeated cross-validation including 10 folds and 5 repeats, I found that both models performed exceptionally well with the Random Forest model accuracy at 97% and the Neural Net model accuracy at 98%.

Introduction:

Understanding and mitigating the effects of climate change is the most pressing issue of the 21st century. Global-scale concerns that should have been addressed decades ago linger and are swept under the rug for future generations to tackle. Lack of data is not the cause of the lack of climate change mitigating action. There is in fact a plethora of data available. Using this previously collected data for meaningful analysis is paramount to better understanding past and future climate trends and preventing/mitigating their adverse effects.

In this paper, I will be utilizing H*istorical climate change* data from WorldClim version 2.0 as well as *Future climate data* from the same source. The historical data contains ground surface data of mean temperature, precipitation, and elevation. The data were collected from land surface sensors at a spatial resolution of 10 minutes (~340 km2) and have been averaged over the years 1970-2000. For instance, with the mean temperature data, this means there is **one** data point for every ~340 km2 of land surface representing the mean temperature for that point averaged across all months for all years 1970-2000 [see appendix Image 1 for plot]. The future data is comprised of the maximum projected land surface temperature for every 10 minutes (~340 km2) [see appendix image 2 for plot]. The projections are averaged over the years 2021-2040. These future temperature predictions are a product of the RCP2.6 emission and concentration pathway modeled to show potential climactic outcomes of successful climate mitigation for which the global average temperature does not rise more than 2°C (van Vuuren et al. 2011). This model is on the optimistic end of future climate forecasting and shows outcomes that can only be achieved with joint global cooperation. For this scenario to be realized there would need to be a 70% decrease in cumulative greenhouse gas emissions from the years 2010-2100 (van Vuuren et al. 2011).

I will be training two machine learning models with the historic climate change data (elevation, precipitation, and average temperature) and the future climate maximum temperature data in order to see which model best predicts the areas that will change in temperature by more than 1.5°C. It is well-established that one major effect of climate change is 'Global Weirding', for which weather-related changes will be extremes in all directions (Held & Soden, 2006). Due to the inevitability of 'Global Weirding', having a machine learning model that can successfully predict which areas will have the largest temperature change is essential to preparing vulnerable areas for tumultuous ecological disruptions. Due to the use of structured data for a binary classification task, I hypothesize that the Random Forest model will have a higher prediction accuracy than the Neural Net model.

The goal of this study is to not only gain insights into predicting possible temperature and climate trends but to use this gained knowledge to influence actionable policy change and intervention on the government level. Several such urgent changes are: the necessary transition to clean energy sources, the implementation of a carbon cap-and-trade system, and increasing the prices of luxury goods with high emissions costs such as animal livestock byproducts.

Literature Review:

The research article 'Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity' (1967) by Manabe and Wetherald is recognized today as one of the most influential research papers on climate change. In the paper, Manabe and Wetherald detail their findings on the radiative convective equilibrium of the atmosphere with a given distribution of relative humidity. In simpler terms comparing relative vs absolute humidity outcomes. The paper was the first published computer simulation of many elements of the earth's climate in order to project what a doubling of CO2 in the atmosphere would do to global temperature. Their model concluded that doubling the CO2 level in the earth's atmosphere would result in a 2°C increase in the earth's global temperature.

Nearly a decade later, Keeling et al. reported in their paper 'Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii' on real word CO2 increases in the atmosphere. They found that in the years 1959 – 1971 the annual average CO2 concentration levels in the atmosphere increased by 3.4%. This paper was among the first to document the increases of CO2 levels in the atmosphere. The paper was also among the first to point to a link between the increase in industrial CO2 being emitted and the trend of increasing CO2 levels in the atmosphere. Keeling's meticulous work monitoring and recording CO2 levels in the atmosphere birthed the Keeling Curve which is maintained to this day.

In 2006, Held and Soden published the paper 'Robust Responses of the Hydrological Cycle to Global Warming' which examined the effects of climate change on the global quantity and distribution of rainfall. Held and Soden were among the first to point out a phenomenon that is now sometimes referred to the 'Global Weirding'. Although the term was established after the paper was published, Held and Soden's findings support that climate change is and will cause weather related extremes in all directions. For example, the drier climates will get drier as the wetter ones get wetter. There is not one trend that will hit all areas of the globe equally.

Data Description and Visualization:

Both datasets used in this study were downloaded from the WorldClim organization's website. The datasets are from two different time periods. The Historical Climate data is from the years 1970-2000 and the Future Climate data is from the years 2021-2040 and is based on RCP2.6 emission and concentration pathway (van Vuuren et al. 2011). The data frequency for both data sets is at the spatial resolution of 10 minutes (~340 km2) and was downloaded as GeoTiff (.tif) files. The variables included are temperature is measured in °C, precipitation measured in millimeters [see appendix Image 3 for plot], and elevation data which did not specify the unit of measurement [see appendix Image 4 for plot]. Each .tif file contained one layer of information which was extracted in RStudio using the package *raster*. After extracting longitudinal and latitudinal information, each dataset contained 2,332,800 rows. After dropping empty rows (longitude and latitude associated with non-land surface areas of the planet) the data frame contained 80,5693 rows. Due to the

computational power required to run this quantity of data and a lack of access to highspeed computers I had to slice the data and run the machine learning models on a single slice of the data. The final data frame size used in the model was 5,000 rows x 6 columns (the largest slice that would run on my computer). The methods used for slicing left the dataset with 4,868 target values where the temperature change > 1.5°C and 132 target values where the temperature change < 1.5°C. The targets are imbalanced with the majority of the target variables being a temperature change > 1.5°C. Before slicing the data the target variable distribution was 766,758:38,935 respectively. The main concern with this model is a potential bias and underfitting the model with too much focus towards temperature change > 1.5°C values. Repeated cross-validation for the model training includes 10 folds and 5 repeats. Test size of 30% and training size was 70%.

Model:

Gini Impurity: $Gini = 1 - \Sigma_{i=1} (p_i)^2$

- Pure: all data belongs to same class
- Impure: data belongs to a mixture of classes
- Gini Impurity: likelihood of misclassifying new values

Expected information gain: $IG = H(T) - H(T \mid a)$

- Entropy before – entropy after a decision "a"
- Entropy: measures degree of randomness
- Information Gain: used to test which element is most relevant in a decision tree

Decision Tree

- Infer class labels based on if/else sequences on features

Random Forest

- Ensemble of many decision trees operating simultaneously on random subsets of the data thus final prediction is robust

Neural Net : $Z = Bias + W_1X_1 + W_2X_2 + \ldots + W_nX_n$

- W = weights for coefficients
- X's are independent variables
- Bias is the intercept

Accuracy = number of correct predictions / total predictions

- Measures how well your model identifies relationships and patterns between variables in the data based on the input and training data

X = mean temperature (1970-2000), precipitation (1970-2000), elevation

y = Δ in temperature

- Δ = max temperature (2021-2040) - mean temperature (1970-2000)
- 0 if Δ < 1.5°C & 1 if Δ > 1.5°C
- 1.5°C was chosen as it is the threshold due to its status as the marker between normality and vulnerable ecosystem collapse (Climate Reality Project, 2021). It is also the agreed upon goal under the 2015 Paris Agreement (Unfccc.int, 2021).

Empirical Analysis:

This analysis shows that both Random Forest and Neural Net machine learning models predicted temperature change outcomes to an accurate degree [for accuracy scores see Image 5 in appendix]. The accuracy of the Random Forest model was 97% with **30** values being misidentified resulting in **type 1 error** (false positive) and **10** values being miss identified resulting in **type 2 error** (false negative) [for RF confusion matrix see image 7 in appendix]. The accuracy of the Neural Net model was 97% with **23** values being misidentified resulting in **type 1 error** (false positive) and **14** values being miss identified resulting in **type 2 error** (false negative) [for NN confusion matrix see image 6 in appendix]. Most of the errors for both datasets were type 1 suggesting that the model may have been underfit and is biased towards Δ > 1.5°C. Due to the use of structured data for a binary classification task, it was hypothesized that the Random Forest model would outperform the Neural Net model. The performance of the Random Forest model is consistent with the hypothesis, while the results of the Neural Net model exceeded expectations in surpassing the accuracy of the Random Forest model.

Conclusion:

The aim of this analysis is to use available environmental data to test several machine learning models and their abilities to predict which areas will be vulnerable to the greatest temperature changes. While the action required to combat climate change is on the global level, the greatest effects of climate change will be felt on the local level in the most vulnerable areas. The necessity of gaining these insights is paramount to instigating actionable policy change and interventions to the regions that will be the most at risk. This model is run on some of the most optimistic climactic outcomes for which the global average temperature does not rise more than 2°C. These models' predictive relevancy to the real world are dependent on the global average temperature not increasing by more than 2°C. For this to be possible immediate action should be taken to reduce greenhouse gas emissions on a global and local scale.

Reference:

The Climate Reality Project. (2021, January 25). Why is 1.5 degrees the danger line for global warming? Climate Reality. Retrieved December 15, 2021, from https://www.climaterealityproject.org/blog/why-15-degrees-danger-line global-warming

Gleckler, Durack, P. J., Stouffer, R. J., Johnson, G. C., & Forest, C. E. (2016). Industrial-era lobal ocean heat uptake doubles in recent decades. Nature Climate Change, 6(4), 394 398.https://doi.org/10.1038/nclimate2915

Held, & Soden, B. J. (2006). Robust Responses of the Hydrological Cycle to Global Warming. Journal of Climate, 19(21), 5686–5699. https://doi.org/10.1175/JCLI3990.1

Keeling, Bacastow, R. B., Bainbridge, A. E., Ekdahl Jr, C. A., Guenther, P. R., Waterman, L. S., & Chin, J. F. S. (1976). Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii. Tellus, 28(6), 538–551. https://doi.org/10.3402/tellusa.v28i6.11322

Manabe, S., & Wetherald, R. T. (1967). Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity, Journal of Atmospheric Sciences, 24(3), 241-259. https://doi.org/10.1175/1520-0469(1967)024<0241:TEOTAW>2.0.CO;2

Neukom, Barboza, L. A., Erb, M. P., Shi, F., Emile-Geay, J., Evans, M. N., Franke, J., Kaufman, D.S., Lucke, L., Rehfeld, K., Schurer, A., Zhu, F., Bronnimann, S., Hakim, G. J., Henley, B. J., Ljungqvist, F. C., McKay, N., Valler, V., & von Gunten, L. (2019). Consistent multidecadalvariability in global temperature reconstructions and simulations over the Common Era. Nature Geoscience, 12(8), 643–649. https://doi.org/10.1038/s41561-019 0400-0

Unfccc.int. (n.d.). Retrieved December 15, 2021, from https://unfccc.int/process-and meetings/the-paris-agreement/the-paris-agreement

van Vuuren, D.P., Stehfest, E., den Elzen, M.G.J. et al. RCP2.6: exploring the possibility to keep global mean temperature increase below 2°C. Climatic Change 109, 95 (2011). Https://doi.org/10.1007/s10584-011-0152-3

WorldClim (2020). *Historical climate data, 1970-2000.* Version 2.1. https://www.worldclim.org/data/worldclim21.html Web. 12 Dec 2021.

WorldClim (2020). *Future climate data, 1970-2000.* Version 2.1. https://www.worldclim.org/data/cmip6/cmip6climate.html Web. 12 Dec 2021.

Statistics were done using R 3.5.0 (R Core Team, 2018), the dplyr (v1.0.7; Wickham, 2021), the raters (v2.0.1; Giardiello, 2014) packages, the nnet (v7.3; Ripley and Venables, 2021) packages, and the randomForest (v4.6; Liaw and Wiener, 2018) packages.

Appendix:

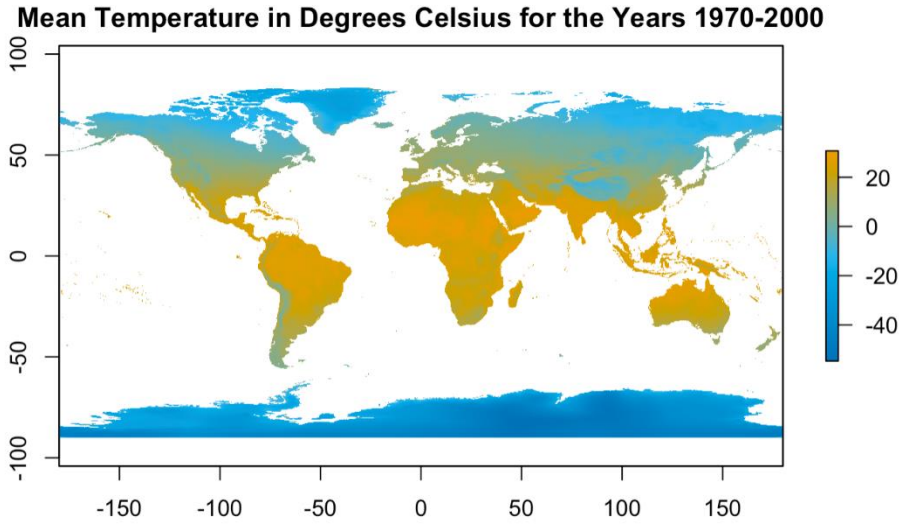


Image 1: A plot of the mean temperature made from the GEOTIF file of the temperature averaged over 30 years at 10 minutes (~340 km2) spatial resolution. Orange high is warmest temperatures and blue is coldest.

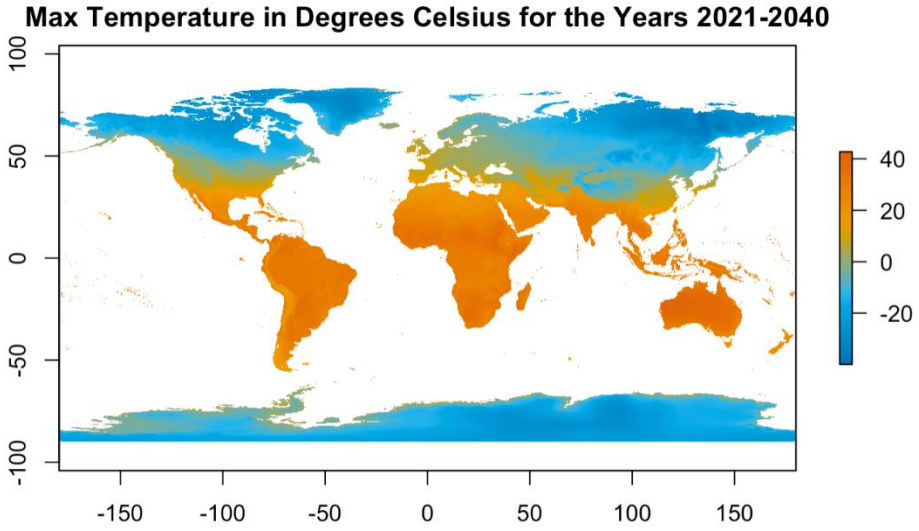


Image 2: A plot of the max predicted temperature made from the GEOTIF file of the temperature averaged over 30 years at 10 minutes (~340 km2) spatial resolution. Orange high is warmest temperatures and blue is coldest.

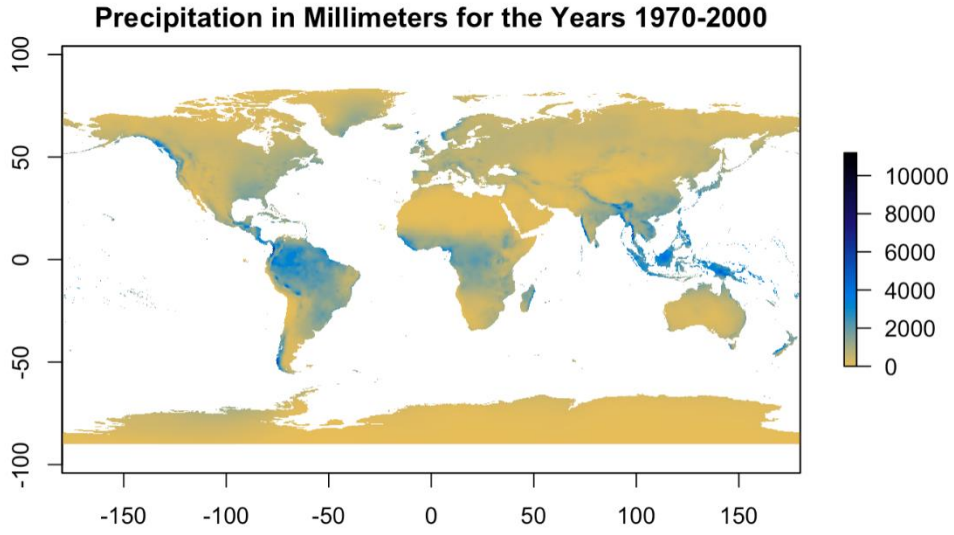


Image 3: A plot of the annual precipitation from the GEOTIF file of the precipitation averaged over 30 years at 10 minutes (~340 km2) spatial resolution. Tan is little to no rainfall, blue is high rainfall, and black is the maximum rainfall recorded

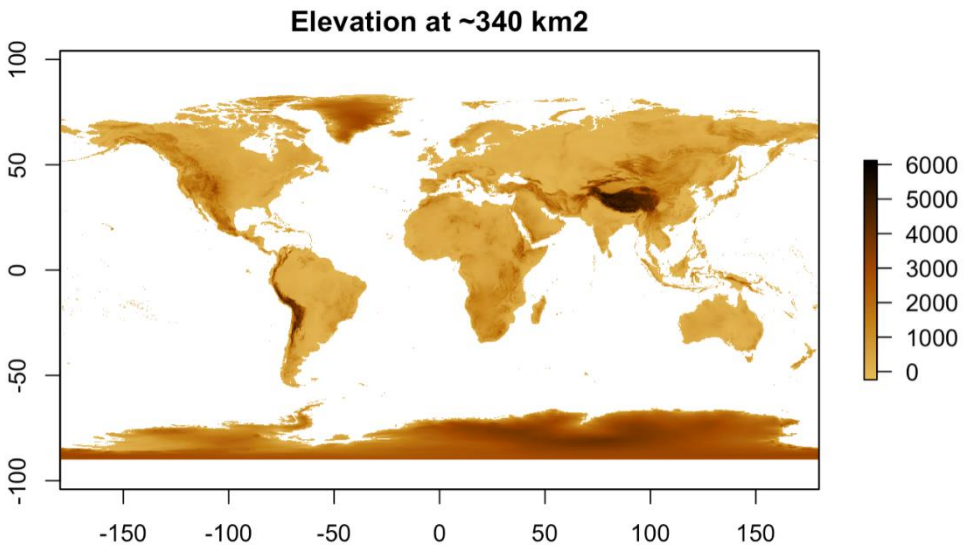


Image 4: A plot of the elevation from a GEOTIF file taken at 10 minutes (~340 km2) spatial resolution. Tan is the lowest elevation; back is the highest elevation recorded.

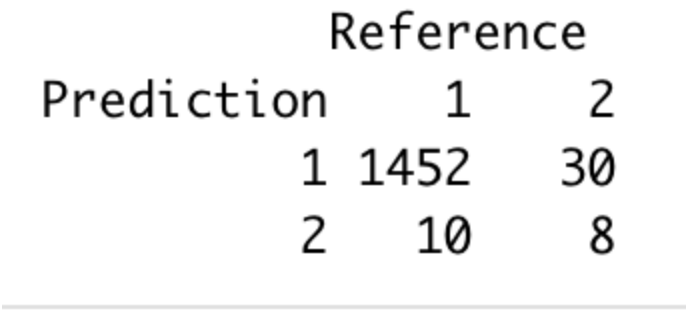
```
[1] "  Neural net accuracy:  0.98"
[1] "  Random Forest accuracy:  0.97"
```

Image 5: Accuracy scores for both models

```
            Reference
Prediction    1    2
         1 1448   23
         2   14   15
```

Image 6: Confusion Matrix for Nural Net model 2 is $\Delta < 1.5$ °C and 1 is $\Delta > 1.5$ °C.

1460 values were correctly identified as being a high Δ. 15 values were correctly identified as being low Δ. 23 values were miss identified as high Δ when they were actually low Δ. And 14 values were miss identified as low Δ when they were actually high Δ.



Prediction	Reference 1	2
1	1452	30
2	10	8

Image 7: Confusion Matrix for Random Forest model 2 is $\Delta < 1.5$ °C and 1 is $\Delta > 1.5$ °C.

1452 values were correctly identified as being a high Δ. 8 values were correctly identified as being low Δ. 30 values were miss identified as high Δ when they were actually low Δ. And 10 values were miss identified as low Δ when they were actually high Δ.

file:///Users/dalithendel/Library/CloudStorage/Box-Box/Data_200_lab/PROJECT/SHORT_final.html

Link to HTML of RMD used for data analysis in this paper.