Section 2. Information technology

https://doi.org/10.29013/EJTNS-22-6-12-19

Katherine Dong,
Senior, highschool at Princeton Highschool

AN AUGMENTATIVE APPROACH TO STUDY CLIMATE CHANGE USING RANDOM FOREST

Abstract. Understanding the best indicators of climate change is essential to predicting the magnitude of climate change in the future. Machine learning models can use features that indicate climate change to determine its impacts. To forecast mean temperature rise, the random forest algorithm is used on a collective dataset containing different indicators of climate change. The indicators include sea levels, temperature anomalies, CO_2 levels, land minus ocean means, and arctic sea ice volumes. The original dataset began with one feature, mean temperature, and several other datasets were augmented to create a larger, more informative dataset. Projecting climate change is modeled as a classification problem with the mean temperature rise as a dependent variable using the random forest model. The features Land Minus Ocean Mean Rise Indicator, Arctic Sea Ice extent, and Mean Temperature Anomalies are the most important variables for predicting temperature change. Running the model yields an R-squared value of 0.58 with mean squared error (MSE) of 0.10, indicating a reasonably effective predicting power of climate change using the identified indicators. This research serves as a guide for effectively curating and augmenting climate change data and the forecasting of climate change and other similar environmental changes using similar temporal and statistics data.

Keywords: Random Forests, Deep learning, Classification, Climate Change.

1. Introduction

Climate change threatens the future of the planet's safety. It creates visible effects such as shrinking glaciers, increasing droughts, stoking wildfires, and intensifying storms. Scientists also predict that in the future, temperatures will continue to rise; there will be stronger droughts, heat waves, and hurricanes; the sea level will continue to rise; and the arctic will likely become ice-free.

In recent years, machine learning (ML) and artificial intelligence (AI) has been broadly applied as a powerful tool to solve problems such as image recognition, auto driving vehicles, and language processing, etc. Similarly, machine learning models such as decision trees and random forest, can be applied to predict climate changes.

There are many examples of such models in the literature. Anderson G J et al applied random forest algorithm, a special machine learning technique, to a multiresolution perturbed parameter

ensemble of the Community Atmosphere Model version 5 (CAM5) [1]. The climatic dependent variables in the research are global annually averaged top-of-the-atmosphere (TOA) energy flux and global annually averaged precipitation. Eleven parameters spanning five physical schemes related to clouds, cloud microphysics, turbulent mixing, and deep and shallow convection are found to be important features to predict these two climatic dependent variables. These parameters' values are varied by three types (resolutions) of perturbed parameter experiments over a total of 906 simulations to create training and testing data to feed the random forest machine learning model. These random forests are able to learn the relationship between resolution changes, parameter perturbations, and model responses from the multiresolution ensemble and are able to make predictions at high resolution while substantially reducing the computational expense.

R. Meenal et al. [15] also applied random forest machine learning models to predict weather [15]. In this research, the global solar radiation (GSR) and wind speed are used as dependent variables to be predicted for Tamil Nadu, India using random forest ML models. The features used for building the model include maximum temperature, minimum temperature, surface pressure, percentage relative humidity (RH), months, latitude and longitude. The random forest ML model is validated with measured wind and solar radiation data collected from IMD, Pune. The prediction results based on the random forest ML model are compared with statistical regression models and support vector machine (SVM) ML models. Overall, the random forest machine learning model has minimum error values of 0.750 MSE and R2 score of 0.97. Compared to regression models and SVM ML models, the prediction results of random forest ML models are more accurate. This research avoids the need for an expensive measuring instrument in all potential locations to acquire the solar radiation and wind speed data.

Random forest has also been successfully applied to species distribution models to landscape applications, and to gradient modeling of conifer species [9; 10].

There remains the need, however, for these tools to be best applied to tackle global climate change directly. To our best knowledge, few of these previous efforts make a concrete attempt to directly model the global climate changes with relevant variables such as CO_2 emission levels and other human production activity related variables which may have contributed to global warming.

With an increasing amount of historical climate data and observations, the machine learning algorithms can be used to model and to predict the Earth's future climate. First, climate related data can be collected from various sources to build a climate related dataset. Then machine learning techniques and algorithms can be applied to the dataset to model climate change. As more data becomes available, this dataset can include more features and variables which can then be used to improve the already-built machine learning models.

In this research, one of the ML techniques, random forest, is applied to climate change, where data is collected, analyzed, and used for modeling climate change. The collected features are calibrated into the random forest model to predict temperature rise.

Most of the papers use a singular dataset for ML/DL analysis. To better model climate change, in this research, multiple datasets from multiple data sources are used to calibrate the machine

learning algorithms. It is necessary to look at climate datasets from similar time frames measured at the same granularity. Specifically, a seed dataset was first collected from the datahub, which contains the monthly dependent variable *Mean Temperature Rise* and monthly features such as *Mean Sea Level, Mean Temperature Anomalies, CO_2 Mole Fraction Mean, Mean Sea Level Rise,* and *CO_2 Emissions Rise*. It was determined that all features collection would focus on the monthly time frame as the dependent variable is in monthly format. These features are all related with the *Mean Temperature Rise* variable and can be potentially used in the ML/DL model calibration.

The feature collection effort continues to other data sources. Two more additional data sources are explored and used to enhance the feature collection. Specifically, the NASA data source and DataWorld data source are also used to collect additional features data. For feature details, see Data Collection section below.

The rest of this paper is organized as follows: Section II describes the methodology of collecting and preparing training data as well as the construction and training of the models. Section III presents the model and discusses its results. Finally, Section IV expands upon the potential uses of this research and discusses applications and future work that could be done in this field.

2. Methodology

The overall methodology of the approach is depicted in Figure 1 below.

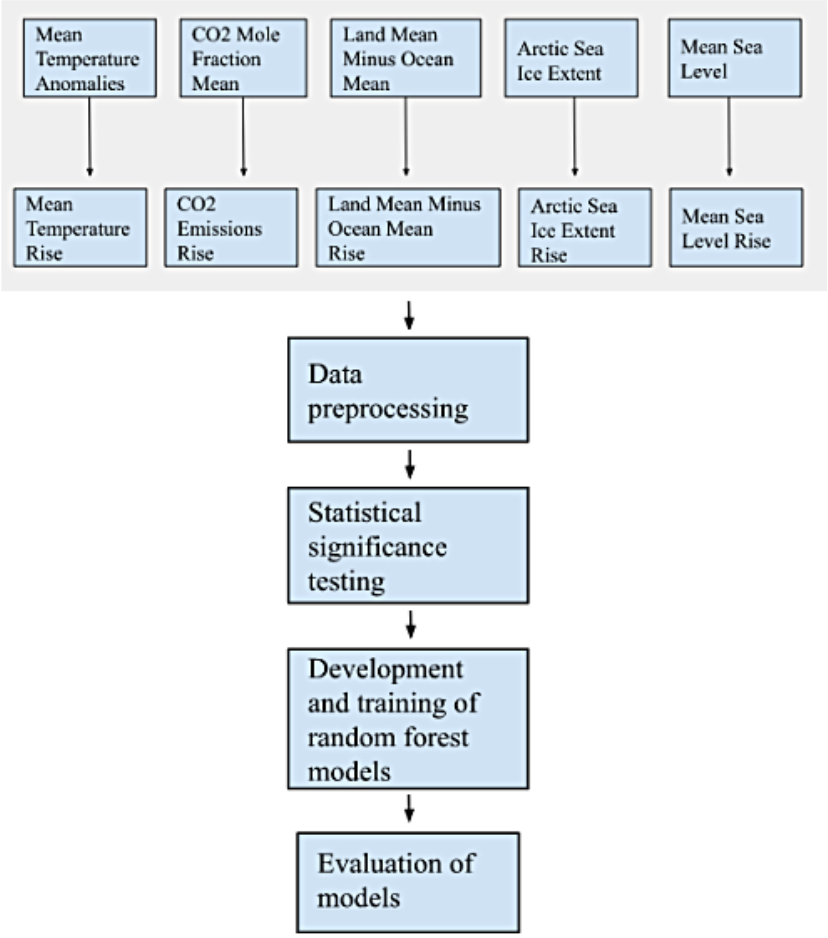

Figure 1. General Methodology

A. Data Collection

The training data for the project was collected and consolidated from a variety of sources. These features are summarized in Table 1.

Table 1.– Summary of Features in the collected data

Feature	Description	Format	Frequency
Mean Sea Level	Average global sea level changes	Numeric	Monthly
Mean Temperature Anomalies	Changes in average global temperatures	Numeric	Monthly
CO_2 Mole Fraction Mean	The number of molecules of carbon dioxide divided by the number of all molecules in air, including CO_2 itself, after water vapor has been removed	Numeric	Monthly
Land Mean Minus Ocean Mean	Sea-surface water temperatures subtracted from land-surface air anomalies	Numeric	Monthly
Arctic Sea Ice Extent	Square kilometers of the amount of arctic sea ice	Numeric	Daily
Mean Sea Level Rise	Whether or not the sea level increases from the previous year	Boolean	Monthly
CO_2 Emissions Rise	Whether or not the CO_2 emissions increases from the previous year	Boolean	Monthly
Land Minus Ocean Mean Rise	Whether or not land minus ocean level increases from the previous year	Boolean	Monthly
Arctic Sea Ice Extent Increase	Whether or not the arctic sea ice extent increases from the previous year	Boolean	Monthly
Mean Temperature Rise	Whether or not the average global temperature increases from the previous year	Boolean	Monthly

B. Software Tools

Python is chosen for this research as the modeling software as it is one of the most popular ML programming languages with many available ML related packages to use. Data preprocessing was performed with Python's NumPy and Pandas libraries. Statistical significance tests were conducted using the SciPy library. Finally, the random forest ML model was created using the sklearn library.

C. Feature Selection

To validate the use of the selected features in the final data set, a series of statistical significance tests were performed on the data.

The student-T test was used to find the statistical significance between each of the features of the data and the mean temperature rise. See Table 2. below for detailed statistics and p-value results for the tests.

Table 2.– Significant Features

	Student-T Test	
Features	**statistic**	**p-value**
1	***2***	***3***
Mean Sea Level	–14.3046	1.43E–41

1	*2*	*3*
Mean Temperature Anomalies	3.372053	0.000781
CO_2 Mole Fraction Mean	–439.468	0
Land Mean Minus Ocean Mean	3.797648	0.000157
Arctic Sea Ice Extent	–69.3466	0
Mean Sea Level Rise	–3.3668	0.000796
CO_2 Emissions Rise	–16.712	3.98E–54
Land Minus Ocean Mean Rise	0.490219	0.624111
Arctic Sea Ice Extent Increase	2.033065	0.04237

These tests indicated that all the features had predictive power; that is, they were all statistically significant and could be used in the dataset. All the features were therefore selected for the final dataset, as they all proved to be relevant.

D. Random Forest Model

Random forest is a supervised machine learning algorithm that is used widely in classification and regression problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems such as climate change.

A random forest algorithm consists of many decision trees. The "forest" generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The random forest algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It also reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages such as scikit-learn.

One of the most important features of the random forest algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables in the case of classification. It yields better results for classification problems. In this research, the mean temperature rise dependent variable has boolean values of either 0, meaning the mean temperature has not risen from the previous year, or 1, meaning the mean temperature has risen from the previous year.

First, the dataset was randomly split into two sub datasets with one sub dataset containing 80% of the original dataset for training the model and the other sub dataset containing the remaining 20% of the original dataset for predicting the model results. The python *train_test_split* function from the *sklearn.model_selection* package was used for this purpose.

3. Results

– Reformat tables

The python RandomForestRegressor from sklearn.ensemble was used for calibration of the model. The random forest model performance is measured by the mean squared error (MSE) and R-squared.

The first random forest model was calibrated using the seed dataset and the other features which came from the same source, including *Mean Temperature Anomalies, Mean Sea Level,* CO_2 *Mole Fraction Mean,* and *Mean Sea Level Rise*. The model performance in terms of the mean

squared error (MSE) and the R-squared for this model is demonstrated in the table below.

Table 3. – Random forest model results using first source

	Mean Square Error (MSE)	R-Squared
Training	0.187	0.247
Prediction	0.211	0.155

The calibrated random forest model included the following features:

1. Mean Temperature Anomalies;
2. CO_2 Mole Fraction Mean;
3. Mean Sea Level.

After running the importance function, it is found that the features Mean Temperature Anomalies, CO_2 Mole Fraction Mean, and Mean Sea Level had the highest predicting powers, from greatest to least. The model prediction MSE and R-squared are larger and smaller than their training counterparts quite significantly. This indicates that the model's prediction power is not as large as should be, and some additional features may need to be added to improve the model performance.

The additional datasets which was later augmented were then used to try to improve the model. The model was re-run with all the data, including the augmented data from the two other sources. The model performance in terms of the mean squared error (MSE) and the R-squared for this model is demonstrated in the table below.

The calibrated random forest model included the following features:

1. "Land Minus Ocean Mean Rise";
2. "Arctic Sea Ice Extent";
3. "Mean Temperature Anomalies".

Table 4. – Random forest model results with augmented dataset

	Mean Square Error (MSE)	R-Squared
Training	0.105	0.577
Prediction	0.157	0.371

Compared with the first model with seed dataset, this random forest model with augmented dataset has smaller MSE and larger R-squared for both training and prediction. This model's MSE values are almost halved the seed dataset model while this model's R-Squared values are almost doubled the seed dataset model. This augmented dataset model results in a much better model compared with the first model, only containing data from the first source. The improvement comes from the benefit of additional sources. For example, the *Land Minus Ocean Mean Rise* and the *Arctic Sea Ice Extent* features which are deemed important come from different data sources. Utilizing more data sources improved the model performance significantly.

In the next model, each feature was added individually, with the random forest model being run after each was added. The purpose of this is to see the impact of augmenting the data together. As shown in the table below, generally, the MSE value decreases while the R-squared value increases as more data is added to the model. Although there are some features that don't make a notable difference in the model's predicting power, the model does improve as a whole after all the data is augmented.

Table 5. – Augmented Random forest model results

Features Added	MSE (Training)	MSE (Prediction)	R-Squared (Training)	R-Squared (Prediction)
1	***2***	***3***	***4***	***5***
Mean Temperature Anomalies	0.208	0.243	0.162	0.026

1	*2*	*3*	*4*	*5*
Mean Sea Level	0.190	0.217	0.237	0.133
CO_2 Mole Fraction Mean	0.187	0.211	0.247	0.155
Land Mean Minus Ocean Mean	0.187	0.213	0.246	0.148
Arctic Sea Ice Extent	0.186	0.217	0.251	0.130
Mean Sea Level Rise	0.186	0.217	0.251	0.130
CO_2 Emissions Rise	0.186	0.218	0.251	0.129
Land Minus Ocean Mean Rise	0.105	0.155	0.579	0.381
Arctic Sea Ice Extent Increase	0.105	0.155	0.579	0.381

Conclusion

In this paper, extensive climate change data was collected, several random forest machine learning models were trained, and their prediction performances were evaluated. The final model has a reasonably high prediction accuracy when modeling the climate change problem as a classification problem.

The software code used in this research can serve as a ready-to-use software tool for analyzing climate changes using the machine learning algorithms. Built on freely available data, the model can easily be extended to make predictions based on more recent or even real time data, while its reasonable high accuracy would provide valuable insights into the future course of climate change.

This research provides an exciting approach for predicting climate change and contributes to important climate change forecasts. The software system developed by this research provides a valuable tool for continued research in this very important research topic.

References:

1. Anderson G. J. and Lucas D. D. Machine learning predictions of a multiresolution climate model ensembleGeophys. Res. Lett. 2018.
2. Andrews T., Gregory J. M., Webb M. J. and Taylor K. E. Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models Geophys. Res. Lett. 2012.
3. Archer K. J., Kimes R. V. Empirical characterization of random forest variable importance measures. Comput Stat Data An – 52. 2008.– P. 2249–2260.
4. Bellard C., Bertelsmeier C., Leadley P., Thuiller W., Courchamp F. Impacts of climate change on the future of biodiversity. Ecol Lett – 15(4). 2012 – P. 365–377.
5. Bradter U., Kunin W. E., Altringham J. D., Thom T. J., Benton T. G. Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods Ecol Evol – 4. 2013.– P. 167–174.
6. Breiman L. Random forests. Mach Learn – 45. 2001.– P. 5–32.
7. Brook B. W., Sodhi N. S., Bradshaw C. J. Synergies among extinction drivers under global change. Trends Ecol Evol – 23. 2008.– P. 453–460.
8. Burns C. E., Johnston K. M., Schmitz O. J. Global climate change and mammalian species diversity in US national parks. Proc Natl Acad Sci USA – 100(20). 2003.– P. 11474–11477.

9. Cushman S. A., Wasserman T. N. Landscape applications of machine learning: comparing random forests and logistic regression in multi-scale optimized predictive modeling of American marten occurrence in northern Idaho, USA. In: Humphries G., Magness D., Huettmann F. (eds) Machine Learning for Ecology and Sustainable Natural Resource Management. Springer, Cham, 2018.– P. 185–203.
10. Evans J. S., Cushman S. A. Gradient modeling of conifer species using random forests. Landscape Ecol – 24. 2009.– P. 673–683.
11. Evans J. S., Murphy M. A., Holden Z. A., Cushman S. A. Modeling species distribution and change using random forest. In: Drew CA (ed) Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer, New York. 2011.
12. Genuer R., Poggi J. M., Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Lett – 31. 2010.– P. 2225–2236.
13. Liaw A., Wiener M. Classification and regression by randomForest. R News – 2(3). 2002.– P. 18–22.
14. Nicodemus K. K., Malley J. D., Strobl C., Ziegler A. The behavior of randomforest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics – 11. 2010.– 110 p.
15. Meenal R., Prawin Angel Michael D. Pamela E. Rajasekaran. Weather prediction using random forest machine learning model. Indonesian Journal of Electrical Engineering and Computer Science – Vol. 22. 2021.
16. Rodriguez-Galiano V.F., Ghimire B., Rogan J., Chica-Olmo M., Rigol-Sanchez J. P. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens – 67. 2012.– P. 93–104.
17. Rogelj J., Meinshausen M., Knutti R. Global warming under old and new scenarios using IPCC climate sensitivity range estimates. Nat Clim Change – 2(4). 2012.– P. 248–253.
18. Sandri M., Zuccolotto P. Variable selection using random forests. Data Analysis, Classification and the Forward Search. Springer,– Berlin, 2005.– P. 263–270.
19. Strobl C., Boulesteix A. L., Zeileis A., Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics – 8. 2007.– 25 p.
20. Thomas C. D., Cameron A., Green R. E., Bakkenes M., Beaumont L. J., Collingham Y. C., Erasmus B. F.N., Ferreira de Siqueira M., Grainger A., Hannah L., Hughes L., Huntley B., van Jarsveld A. S., Midgley G. F., Miles L., Orgeta-Huerta M.A., Peterson A. T., Philips A. L., Williams S. E. Extinction risk from climate change. Nature – 427(6970). 2004.– P. 145–148.
21. Watanabe M. et al. Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. J Climate – 23(23). 2010.– P. 6312–6335.