Wildfire Prediction

Ryan Chang*
University of Waterloo
r28chang@uwaterloo.ca

Aniruddh Rao*
University of Waterloo
a36rao@uwaterloo.ca

Aryan Sharma*
University of Waterloo
a98sharm@uwaterloo.ca

Tyler Zeng
University of Waterloo
t4zeng@uwaterloo.ca

Paige Kobzar
University of Waterloo
pkobzar@uwaterloo.ca

Talha Yildirim
University of Waterloo
talha.yildirim720@gmail.com

* *Equal contributions*

***Abstract*—This paper presents a study aimed at the development of a wildfire prediction model tailored to the Kelowna region in British Columbia, Canada, spanning the years 2017-2022. Our research focuses on harnessing past weather parameters to forecast the hectares of area burned across 36 specific zones in Kelowna on a quarterly basis. Employing the XGBoost Regression model, we have achieved commendable accuracy in predicting wildfire outcomes for the majority of cases. However, accurately anticipating outliers, particularly those associated with significant fires in the region, remains a persistent challenge. We then discuss the limitations encountered during our investigation and propose prospective avenues for refining and enhancing the efficacy of our study.**

I. Introduction

Wildfires pose a significant threat to the environment, economy, and communities. To prevent or minimize damage caused by wildfires, the model project aims to develop an advanced wildfire prediction algorithm targeted towards the Vancouver region. Through accurate prediction, communities and local officials can allocate resources to prevent, detect and suppress wildfires. Predicting wildfires accurately is crucial for early warning systems and minimizing adverse effects.

A. Motivation

The core motivation behind focusing on wildfire prediction in British Columbia (BC) arises from the need to enhance the public safety and protection of natural resources. In 2023, the Vancouver region saw 2,245 wildfires, the largest and most destructive year in the history of BC, as reported by the BC Government (2023) [BCGov, 2023], highlighting the urgent need for improved predictive models. By leveraging machine learning and artificial intelligence, this project aims to take on the challenge of developing an accurate and timely prediction model to help create informed decisions.

B. Problem Definition

The primary objective of this project is to develop a predictive model capable of accurately forecasting the occurrence and severity of wildfires in the BC area. This involves using historic burn data (hectares burnt), and weather data (precipitation, temperature, relative humidity, wind speed, wind direction). The challenge in this project lies in parsing and analyzing such a diverse dataset to predict wildfires.

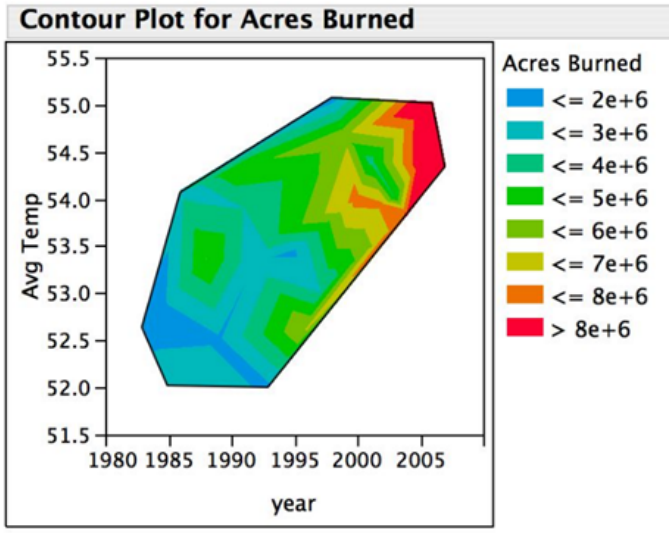

Fig. 1. Average temperature per year compared to acres burned. This figure was obtained from [Leone, 2019]

II. Related Work

A major motivating factor for project, was to examine the extent to which climate change may increase the frequency and size of wildfires in BC. Figure 1 is from a report by the US EPA, Leone et al. (2019) [Leone, 2019] that displays how rising temperatures have impacted wildfires in the western United States. In this paper, they were able to draw a direct correlation between the rise in global temperatures, and an increase in the size of wildfires.

Our project approach is closely aligned with the methodology proposed by Gholami et al [Gholami, 2021]. (2021), who developed predictive models for wildfires in India using historical climate data. In this report, the concept of taking two datasets, one with historic climate data, the other with wildfire burn data of the same area of interest, was introduced. Upon combining the two datasets, this research team was able

to develop a predictive machine learning model for wildfires in the area. Our team's goal for the project was to follow their approach, this time with data from a region of interest in Canada, to see if we could obtain similar results.

III. Methodology

A. Dataset

Two datasets have been examined as part of our study, with the first being a historical numeric weather dataset capturing weather station recordings in BC. This dataset is curated by the Pacific Climate Impacts Consortium, a regional climate service center affiliated with the University of Victoria. The Consortium specializes in conducting quantitative studies to assess the impacts of climate change and climate variability.

Our choice of this dataset is grounded in its open-source nature and ongoing maintenance, making it a reliable resource for weather-related analyses in BC. The dataset offers a diverse range of weather parameters, enabling comprehensive studies and facilitating projections for future years.

The second dataset we examined was a historical wildfire dataset maintained by the BC government. The dataset is a collection of Google Earth satellite data points that contains the size of the fire in hectares, the longitude and latitude coordinates, time of recording, and various metadata.

One of the major advantages to working with satellite data is in its precision and efficiency, as opposed to the satellite imagery used in the previous study by Microsoft [Gholami, 2021]. Satellite data gives a clean set of coordinates along with other parameters defined for each instance of a fire. Comparatively, satellite imagery might derive this information from processing multiple images; thus less consistent. Furthermore, comparing the file size of satellite data and imagery, satellite data takes considerably less memory to store and resources to process.

Both the weather and historical wildfire datasets were studied over a 6 year period, 2017-2022. This time period was chosen since prior to 2017, all wildfire recordings were labelled by year rather than date. We wanted to be able to predict wildfires on a finer granularity than a yearly time step, as such, 2017 and onward was studied.

B. Defining Region of Interest

The first objective of our experiment was to define an area of interest in BC to study. One resource that was used to define this area of interest was the historical wildfire dataset previously mentioned. We then used ArcGIS software to develop a heat map showing which regions in BC had the most wildfires in the past. Figure 1 shows areas of interest with a substantial number of wildfires in the southern central part of BC. The Kelowna region was observed to have particularly intense fires, as such, it was chosen as our region of study.

Kelowna was then spatially divided into a grid consisting of 36 smaller regions using latitude and longitude splits as illustrated in Figure 3. Each grid cell's size is determined by the number of divisions and contains a unique subset of weather stations. This allows our model to leverage the distinct

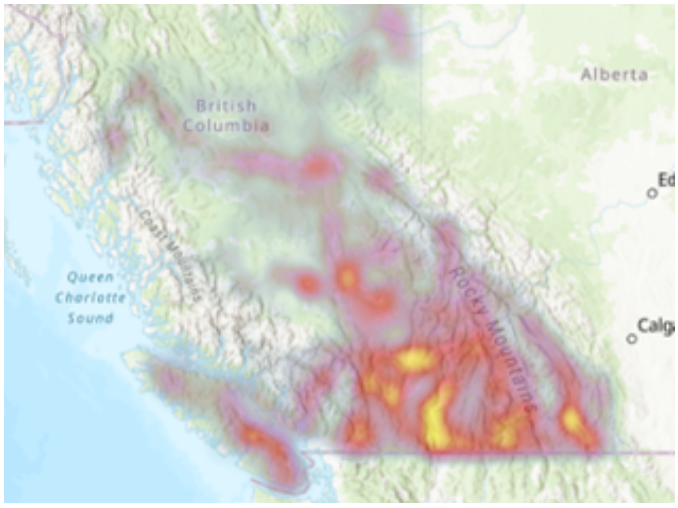

Fig. 2. Heat map illustrating the intensity of previous fires in the BC. Southern and Central BC, particularly the Kelowna region has historically experienced intense fires.

spatial context of each fine area's weather to predict localized burn areas. The number of grid splits was optimized to achieve a balance between granularity and data completeness. We iteratively tested grid configurations, where n was varied and for each n, a dataset M_n was constructed which represents the metadata of stations within each of the n^2 fine areas. This process aimed to maximize the number of grid cells with weather station data available and the proportion of non-zero burn area entries, with $n = 6$ or a 6-by-6 grid identified as the optimal split.

C. Weather Data Preparation

A pipeline was then developed to determine an adequately small time step for wildfire prediction while preserving complete weather data. First, the metadata was utilized to retrieve the weather features corresponding to the stations within each region. The result is a matrix $\mathbf{W}(t,i) = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{n^2}\}$ for each time step t and each fine area i, where $\mathbf{w}_j$ represents a weather feature vector. The matrices are aggregated along their columns, resulting in a single vector $\mathbf{d}(t,i)$. These vectors are concatenated across all fine areas, forming a dataset $\mathbf{F}(t)$ for each time step t such that $\mathbf{F}(t) = \bigoplus_{i=1}^{N} \mathbf{d}(t,i)$ where N is the total number of fine areas and $\bigoplus$ denotes vector concatenation.

Oftentimes, the $\mathbf{d}(t,i)$ vectors contained differing weather features due to the diversity of station parameters. Thus, upon concatenation, a thresholding technique was applied to exclude columns that exceed a predetermined limit of missing values. The thresholding process also determined that a minimum time step *t* of quarters (every 3 months) was required to have at least 50% of rows for all columns to be non null. The remaining columns that fit in the threshold are precipitation, temperature, relative humidity, wind direction, and wind speed.

For the remaining cells that contained null values, interpolation was used. The interpolation methodology was based

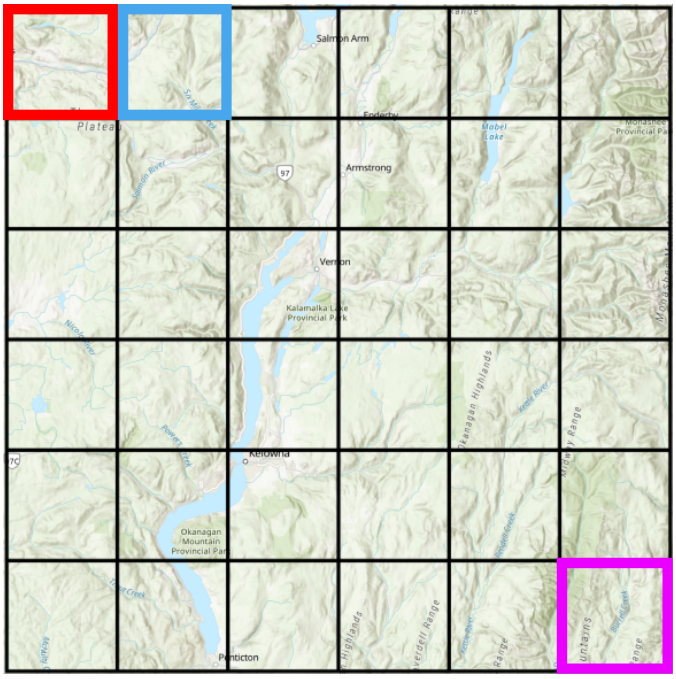

Fig. 3. Kelowna subdivisions into 36 uniform areas. Subdivisions are numbered from left to right, top to bottom. As such, the red square is the 1st fine area, blue is the 2nd, and purple is the 36th

on the presumption that adjacent fine areas within the same time frame would encounter comparable weather conditions. To address the absence of weather data, an average of all the closest neighbouring fine area recordings was calculated, serving as a representative value for the missing information.

D. Wildfire Data Preparation

To process the raw burn data, y, for each cell in A_i, a three-step pipeline was introduced. First, the data was filtered until only the necessary parameters remained: date of ignition, longitude, latitude, and hectares burnt. Next, the data was grouped into quarterly time steps between 2017-2022. Finally, the data was aggregated by applying a latitude-longitude mask and a burn mask to ensure the data points fell within the area of interest. Iterating over each quarter and summing, the total area burned in hectares was flattened into a vector of length 36, the number of fine areas. For any quarters with missing data, a zero vector was used.

E. Dataset Consolidation

Given the weather and burn data consists of points with a specific time and latitude-longitude pairing, the data was encoded in a tabular format. Multiple fine areas of interest were chosen and partitioned into their own fine area grid, denoted by a matrix, $A_i = M_{n \times n}$, where i denotes the i^{th} fine area and n denotes the number of splits. Each row entry is formatted as $[\mathbf{X}, y]_i$, where $\mathbf{X} \in [W]^T$ denotes the chosen weather parameters and $y \in \mathbb{R}^T \geq 0$ denotes the amount of hectares burnt during the time interval T. An example of this data encoding is shown in Figure 4. We created 24 of these tables, each representing a quarter in the 6 year period of study between 2017-2022.

This data format allows a regression to be performed with weather data as the X train and burn data as the y train. In order to accommodate for a regression approach, the model uses the T^{th} time interval to predict the $T + 1^{th}$ time interval. For example, if T was a quarterly interval and the goal was to predict the 2^{nd} quarter's burn data, weather data from the 1^{st} quarter would be used.

The input features are the weather parameters from the 24 encoded tables, denoted as $F(t)$. The corresponding output targets are burn vectors of length 36, one for each fine area. The finalized dataset is distributed into three subsets: 15 examples for the train set, 5 for cross-validation and 4 for the test set.

F. Model Training

In order to predict continuous burn areas, we chose to apply XGBoost Regressor, a machine learning algorithm based on decision trees and ensembles. This model is able to capture the complex, non-linear relationship betweeen climate features and burn area while remaining computationally efficient.

To optimize the XGBoost model, a grid search is performed over a hyperparameter space which includes maximum depth, number of estimators and learning rate. First, the model is sequentially trained on each of the 15 allocated train dataframes. The model then predicts the 36 burn areas for each of the cross-validation dataframes and an average Mean Squared Error loss is computed as such:

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ where n is the number of training exampels, y_i is the ground truth burn, and $\hat{y}_i$ is the predicted burn for the ith observation.

The optimized hyperparameters were found to be a learning rate of 0.005, a max depth of 2 and 25 total estimators. We then run a final training loop on the train set, where the XGBoost model is finetuned on each sequential dataframe. Finally, the 4 test dataframes are used during inference for model evaluation.

G. Evaluation Methods

Three key metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)—were recorded to gauge the accuracy and performance of the model in predicting hectares burned. The utilization of MAE facilitated a straightforward analysis by capturing the average magnitude of errors. Given the stochastic and unpredictable nature of wildfires, particular attention was dedicated to identifying the impact of outliers. This necessitated the inclusion of both MSE and RMSE in the recorded metrics.

To ensure the model's generalization capability and guard against overfitting to the training set, cross-validation was applied. This approach not only assesses the model's accuracy but also addresses its robustness in handling real-world scenarios, where wildfires can exhibit diverse and unpredictable behavior.

			X			y
Fine area	Precipitation (cm)	Temperature (°C)	Relative Humidity	Wind Direction (°)	Wind Speed (km/h)	Hectares Burnt
1	0.05293	-2.00874	81.09615	157.55815	5.58324	2.13253
2	0.04536	3.80863	78.49547	142.21275	4.52923	0.63215
...	...	...	...	...	...	...
35	0.04914	2.85587	79.43904	141.23593	4.57800	0.02234
36	0.08402	0.52527	85.93182	132.64978	5.23957	1.45986

Fig. 4. Example of final data encoding used for training. Each table contains the data for a quarter (3 month period) of study. The data has 36 rows total, each row defining the weather parameters and hectares burnt for the fine area. Refer to Figure 3 for the fine area divisions. Weather features were used as the predictor X and hectares burn was the predicted feature y.

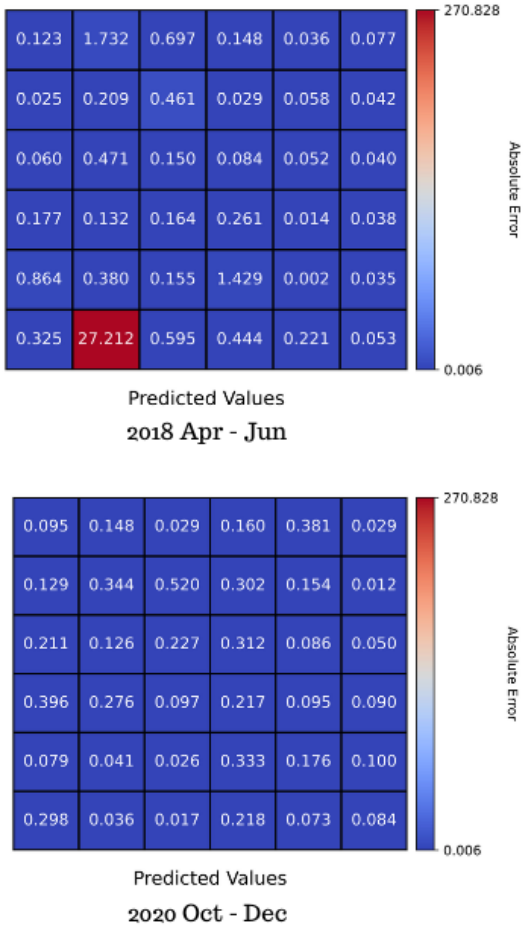

Fig. 5. Predicted burn areas and absolute error for Kelowna fine areas across 2 time steps

IV. Results and Discussion

The predictive results for sample quarters are illustrated in Figure 5, with each grid cell representing a fine area.

The MAE, MSE, and RMSE of the model during training were plotted. The RMSE is seen in Figure 6. All plots demonstrated a close correlation between the model's performance on the training data and its cross-validation counterpart,

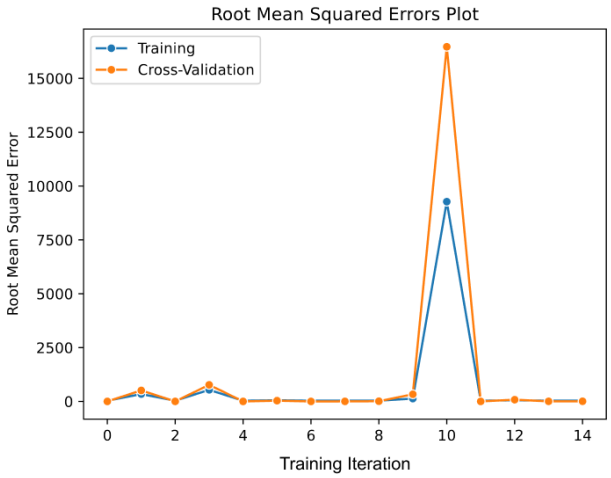

Fig. 6. RMSE vs training iteration. Training (blue) and cross-validation (orange) errors are plotted.

suggesting that overfitting is not a prominent issue. Moreover, a trend is observed as the error decreases over the model's training duration. The final errors are shown in table 1.

TABLE I
AVERAGED ERRORS OVER THE 4 TEST DATAFRAMES

MSE	RMSE	MAE
515.25	22.7	2.629

A point of interest is on the 10th training iteration where the error spikes. Figures 7 and 8 illustrate the distribution of our training set. Examining Figure 7, there is relatively little variation in the weather parameters of each quarter across the years. This however is in stark contrast to the variation of hectares burned in each quarter. As seen in Figure 8, the third quarter of 2018 and 2021 experienced an immense amount of fire compared to other years. This result illustrates the unpredictable nature of wildfires. Despite the weather parameters being relatively consistent, the hectares burned can still experience great variation.

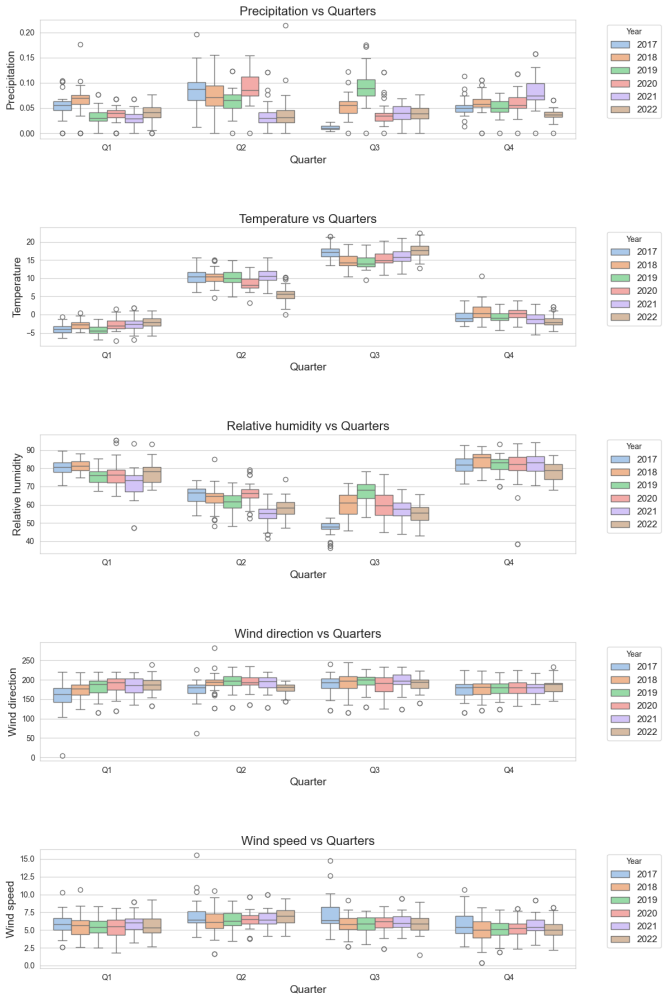

Fig. 7. Weather parameter distribution of each quarter for each year. Weather parameters are relatively consistent with each other.

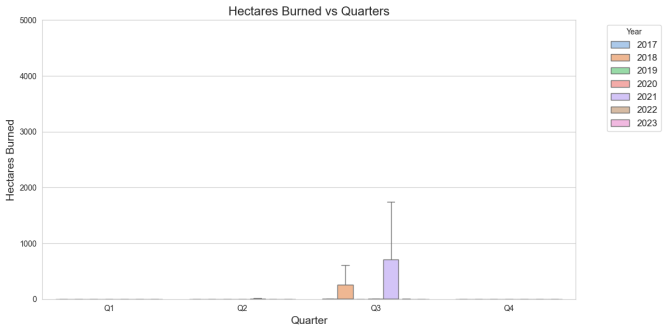

Fig. 8. Hectares burned distribution of each quarter for each year. Hectares burned in the third quarter of 2018 and 2021 experienced immense fires compared to all quarters of all years.

V. Conclusion

This study of wildfires proves to be an extendable exercise in encoding the weather and wildfire data to create a predictive model. From our results, encoding both space and time to perform a regression yields strong results for the majority of quarters in Kelowna.

However, it remains challenging to definitively assert that the five weather parameters under investigation – precipitation, temperature, relative humidity, wind direction, and wind speed – are sufficient for predicting the onset of a major wildfire that can be considered an outlier from the historic data. The nature of wildfires is inherently unpredictable and stochastic, often arising from seemingly random incidents like a lightning strike or human activities. While past research has suggested a correlation between certain weather features and the occurrence of fires [Leone, 2019], the current set of five features alone may not provide the comprehensive basis required for constructing a reliably predictive model.

VI. Limitations

The limitation in our study is reliance on only five weather parameters. This stems from the absence of comprehensive weather recordings in the dataset. While there was an option to augment both time and space steps, extending the analysis to cover a broader time frame, say years instead of quarters, and a larger geographical area, such as the entirety of British Columbia instead of focusing solely on Kelowna, it is important to note that this approach might compromise the practical utility of the predictive model. Wildfires possess the propensity to spread rapidly, necessitating quick containment measures for effective damage mitigation. Therefore, striking a balance between data granularity and the urgency of wildfire response becomes crucial for the model's real-world applicability.

VII. Future Work

Our next crucial step in advancing our study involves augmenting the predictive capabilities of our model. One effective approach is to enhance the richness of our weather data in BC by incorporating information from additional datasets. Furthermore, broadening the spectrum of predictive features beyond weather variables could significantly improve the model's ability to anticipate outliers in the wildfire occurrence, incorporating factors such as proximity to human settlements or infrastructure, and the likelihood of lightning strikes in the region. In addition to these enhancements, it is worth exploring methodologies beyond regression analysis. Deep learning techniques in particular have experienced much success in wildfire and extreme weather prediction.

References

[BCGov, 2023] BCGov (2023). Wildfire season summary.

[Gholami, 2021] Gholami, S., K. N. W. J. . F. J. L. (2021). Where there's smoke, there's fire: Wildfire risk predictive modeling via historical climate data.

[Leone, 2019] Leone, J., P. J. . G. K. (2019). Forecasting wildfires and examining the extent of global climate change.