How Emissions Will Impact Wildfire Risk

Nathan Brodie
njbrodie@ucsd.edu

Benjamin Xue
bxue@ucsd.edu

Kai Morita
kmorita@ucsd.edu

Duncan Watson-Parris
dwatsonparris@ucsd.edu

Abstract

Climate change has been shown to have a profound effect on the amount of droughts and wildfires. Specifically, climate change can lead to an increase in Vapor Pressure Deficit (VPD), which represents the difference between the level of H_2O present in the atmosphere compared to how much water the atmosphere can hold. With more climate data being available, we were able to use deep learning models to forecast and emulate Vapor Pressure Deficit. This gave us a better understanding of which areas have drier vegetation, and as a result are more at risk of wildfires. With the prevalence of wildfires in various parts of the world and its relation to climate change, finding ways to efficiently model Vapor Pressure Deficit can uncover a lot about how certain climate patterns are correlated with climate change. We developed a series of climate emulators using a Random Forest model, Gaussian Process model, and Convolutional Neural Network to measure Vapor Pressure Deficit.

Website: https://njbrodie.github.io/DSC180B-B03/
Code: https://github.com/njbrodie/DSC180B-B03

1 Introduction 2
2 Methods 3
3 Results 7
4 Discussion 9
5 Conclusion 10
References 10
Appendices A1

1 Introduction

Vapor Pressure Deficit (VPD) represents the difference between the water vapor present in the atmosphere and the maximum amount of water vapor the atmosphere can hold. This can be used to represent how dry the plants within a surface are, which is an indicator of high wildfire chance. Our input climate variables used to calculate VPD are near-surface relative humidity and near-surface air temperature. Using these variables, we aim to predict Vapor Pressure Deficit using our Convolutional Neural Network, Gaussian Process, and Random Forest models. These machine learning methods are useful because they are able to scale our climate variables for efficient training and accurately depict the time dependencies of our data.

Previous work has attempted to predict wildfire likelihood by relying directly on values of temperature, precipitation, carbon emission from fire, and other wildfire related variables. Given the CMIP6 predictions for a given climate pathway (climate scenario), Gallo and Blackett (2023) uses the Canadian Fire Weather Index System (CFWIS), which takes in values for temperature, precipitation, relative humidity, and wind speed to make wildfire predictions, and also provides a method to evaluate these predictions based on different models. Yu (2022) takes a similar approach, but instead gauges wildfires based on carbon emission from fire. In predicting fire behavior, Rodrigues et al. (2024) finds that Vapor Pressure Deficit (VPD) is a better predictor than several other common predictor variables when predicting fire behavior. However, current machine learning approaches which utilize VPD as a main predictor for wildfires are limited in geographical scope and not yet adapted towards climate change predictions (Buch et al. 2023).

Although previous work has provided a strong framework for making wildfire predictions based on existing predictions of key variables for a certain climate pathway, our model makes predictions based solely on the emissions of key climate change pollutants: CO_2, SO_2, CH_4, and BC (black carbon). As in our Quarter 1 project, following the approach of ClimateBench (Watson-Parris 2022), we trained our models using the CESM2 dataset. We tested several machine learning models: a random forest, Gaussian process, and CNN-LSTM. We then evaluated our model predictions and performed model selection using the normalized RMSE as described in Watson-Parris (2022). By training our model in this way, our model can extrapolate to climate pathways outside the 5 Shared Socioeconomic Pathways (SSPs) used in many climate models. This can be done by customizing the levels of CO_2, SO_2, CH_4, and BC emissions over two spatial (latitude, longitude) and one temporal dimension. Thus, our model is able to more flexibly predict future wildfires, whereas previous work is limited to methods that struggle to generalize past the 5 SSPs. We use our model to make predictions of the global VPD in the next century, and accordingly extrapolate wildfire likelihood.

We found that the best data for our project was found in the CESM2 dataset. As we looked into what the most important variables are when it comes to predicting wildfires, we found that vapor pressure deficit is the most predictive. There are multiple different factors that come into play when it comes to what increases the chance of wildfires in an area. One of the data points most indicative of wildfires is vapor pressure deficit. Vapor pressure deficit

essentially is the value representing the difference between the amount of water vapor an area can contain and the amount of water vapor an area can actually contain. The data in the CMIP6 dataset does not explicitly contain vapor pressure deficit as a measured variable, but it does contain variables that can be utilized to calculate vapor pressure deficit. The data points used to calculate vapor pressure deficit in this project are relative humidity and average temperature. We found that the best data for our project was found in the CESM2 dataset. As we looked into what the most important variables are when it comes to predicting wildfires, we found that vapor pressure deficit is the most predictive.

Our training and test data came from the CESM2 dataset, which contain wide variety of scenarios, including scenarios known as SSPs (Shared Socioeconomic Pathways) as described in the UN Climate Synthesis Report (Lee and Romero 2023), along with several historical simulations which emulate past climate change. Our training data consisted of 387 data points (years) of temperature and humidity data that we trained our models on. This comes from combining the climate scenarios ssp126, ssp370, ssp585, and historical simulations. The training data has three dimensional axes: time, latitude, and longitude. The data forms a geographical grid with 144 subdivisions for longitude ranging from 0° to 360°, and 96 subdivisions for latitude ranging from -90° to 90°. For time, the simulations are aggregated into yearly data, with the SSPs ranging from 2015–2100, and the historical simulations ranging from 1850–2100. We created models trained on the data and tested them using the climate scenario ssp245. We took the emission inputs and calculated the Vapor Pressure Deficit from the data.

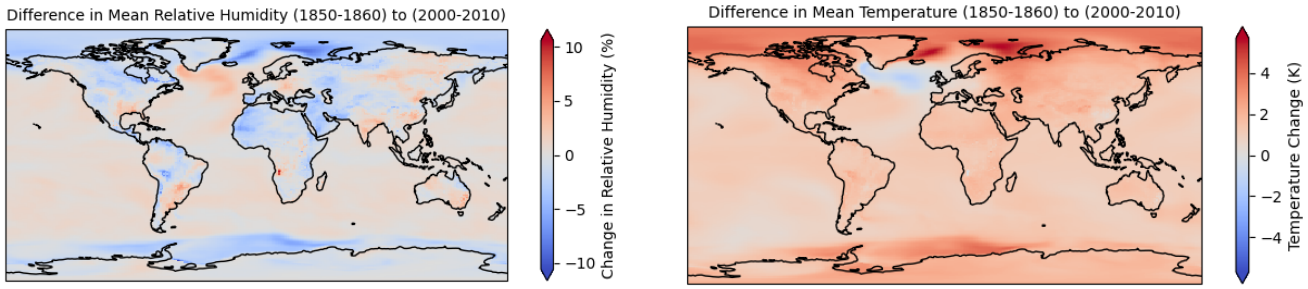


Figure 1: Humidity and Temperature Differences

Humidity and Temperature can be utilized to calculate the actual vapor pressure and the saturation vapor pressure. The saturation vapor pressure is derived from the Clausius-Clapeyron Equation which uses relative humidity and temperature as variables. The formula used to compute the saturation vapor pressure is given in Equation 1.

2 Methods

2.1 Computing Vapor Pressure Deficit

We trained our models on VPD by computing VPD from relative humidity and temperature for each of the climate scenarios we used as training and test data. The formula we

used for computing VPD is given in Bolton (1980). First, we compute the saturation vapor pressure (SVP), which is a measure of the amount of water vapor that air can hold at a given temperature. The SVP can be obtained via the equation

$$\text{SVP} = 0.6112 \exp\left(\frac{17.67T}{T+243.5}\right), \tag{1}$$

where T is the temperature in Celsius, and the pressure is given in kilopascals (kPa) (Bolton 1980). This formula is accurate to 0.3% for temperatures $-35^{\circ}C \leq T \leq 35^{\circ}C$ (Bolton 1980), and nearly all regions we are concerned about fall within that temperature range. It is of note that there exist other similar but slightly different formulas for SVP; the formula we use is the one above, but we expand upon this discussion in subsection A.2. Now that we have the SVP, we can calculate the VPD from the relative humidity as follows:

$$\text{VPD} = \left(1 - \frac{RH}{100}\right) \times \text{SVP}, \tag{2}$$

where RH is relative humidity as a percentage (Bolton 1980).

2.2 Evaluation Metric

In modeling the simulations performed by CESM2, we use four different models: a linear (pattern scaling) model, a Gaussian process model, a CNN-LSTM model, and a random forest model. To evaluate the performance of these models against the true simulation, we employ a modified RMSE metric. Since the model predictions have a time dimension as well as two spatial dimensions, there are different ways to aggregate the data in a way to better capture regional or global errors. The NRMSE defined in ClimateBench rectifies these differences, and is the metric we use to evaluate our model. The ClimateBench paper uses a spatial NRMSE (NRMSE_s) and a global NRMSE (NRMSE_g) into a total NRMSE (NRMSE_t)(Watson-Parris 2022), and defines them as follows:

$$\text{NRMSE}_s = \frac{1}{|\langle y_{i,j}\rangle|_{t,n}} \sqrt{\langle (|x_{i,j,t}|_t - |y_{i,j,t,n}|_{t,n})^2 \rangle}, \tag{3}$$

$$\text{NRMSE}_g = \frac{1}{|\langle y_{i,j}\rangle|_{t,n}} \sqrt{|(\langle x_{i,j,t}\rangle - \langle |y_{i,j,t,n}|_n \rangle)^2\rangle|_t}, \tag{4}$$

$$\text{NRMSE}_t = \text{NRMSE}_s + \alpha \cdot \text{NRMSE}_g. \tag{5}$$

In the above equations, x and y refer to the predictions and targets, while i, j, t, n refer to the vertical (latitude) grid index, horizontal (longitude) grid index, year, and ensemble number, respectively. The value α is set to 5 to weight the spatial and global NRMSEs equally. The $|\cdot|$ refers to a mean taken over time, ensemble members, or both. Meanwhile $\langle\cdot\rangle$ refers to a global mean, weighted to account for smaller grid size towards the poles, and defined as

$$\langle x_{i,j} \rangle = \frac{1}{N_{\text{lat}} N_{\text{lon}}} \sum_{i=1}^{N_{\text{lat}}} \sum_{j=1}^{N_{\text{lon}}} \cos(\text{lat}(i)) x_{i,j}. \tag{6}$$

Here, $N_{\text{lat}} = 96$ and $N_{\text{lat}} = 144$ refer to the number of grid subdivisions for each geographical direction, and $\text{lat}(i)$ refers to the latitude given the vertical grid index.

2.3 Linear

To set a baseline performance standard for the following machine learning models, we created a linear model to make predictions of VPD from mean global temperature. We trained our model on the mean global temperature of the training SSP and historical simulations, averaged over all ensemble members. We used this to fit a regression model to the variable VPD. We then computed the differences between the linear model and the actual model over the years 2080–2100 and plot the differences Figure 2, as well as computed the NRMSE of the model over the years 2080–2100 in Table 3.

2.4 Gaussian Process

Similar to the Climate Bench models we will be utilizing a Gaussian Process model to create predictions on the calculated vapor pressure deficit variable. GP models use kernel functions to define the shape that a prediction function can take and choosing different functions allow for differing levels of smoothness and differentiability of the kernel and thus the predicted function. The kernel used for the GP model in this project was a Matern-1.5 because it "guarantees the GP is a continuous, once differentiable function" (Watson-Parris 2022). The input values used are carbon dioxide, sulfur dioxide, black carbon, and methane, but different kernels are used separately for each variable in addition to a Matern-1.5 kernel.

Additionally, dimensionality reduction was applied to sulfur dioxide and black carbon using principal component analysis in order to lessen the amount of data used, while only sacrificing a small amount of the explained variance within each variable (4 percent of the variance in sulfur dioxide and 2 percent in black carbon). Only the five most significant principal components for these two variables were used, enabling faster calculations for the GP model. The different input variables were also standardized using the training data's mean and standard deviation. Also the input variables of sulfur dioxide and black carbon were also treated with automatic relevance determination kernels. "The GP covariance function is obtained by summing all kernels together" which can account for any relationships between the different input variables (Watson-Parris 2022). To visually demonstrate the predictions of the GP model, we will project its predictions of vapor pressure deficit onto a world map. Because the model previously being used in the Climate Bench paper and in our previous quarter's project, we know that the model performs well when predicting other data. We will compare the predictions of this GP model to the predictions of vapor pressure deficit by the CESM2 dataset in order to test the validity and accuracy of our predictions.

2.5 Random Forest

A random forest is a model which averages decision trees to make predictions. Some strengths of a random forest model include their relative interpretability of other machine learning models, and their ability to fit to nonlinearities within the data (Schonlau and Zou 2020). A weakness of random forest models is their inability to extrapolate outside the training dataset (Schonlau and Zou 2020). However, our training dataset contains ssp126 and ssp585 at the extremes, and any scenario we would like to model will likely fit within that range. This means our random forest model can predict VPD for most realistic climate scenarios.

Similarly to the Gaussian Process model, we performed PCA to obtain the first 5 principal components for SO_2 and BC. We combined this with CO_2 and CH_4 to train our random forest model. We fit a random forest model directly on the VPD variable, and used cross-validation to choose the best hyperparameters. These hyperparameters are listed in below in Table 1. We then computed the differences between the random model and the true CESM2 model over the years 2080-2100 and plotted the differences in Figure 2, as well as computed the NRMSE between the models over the years 2080–2100, as seen in Table 3.

Num trees	Min Samples Split	Min Samples Leaf	Max Features	Max Depth	Bootstrap
600	15	4	All	35	False

Table 1: Random Forest Hyperparameters

2.6 Convolutional Neural Network

We implement a Convolutional Neural Network(CNN) and Long short-term memory(LSTM) network to model the time and spatial climate data. CNNs use filters on input images to learn representations of the features and identify patterns within the data. This is helpful for spatial data that involves our climate prediction tasks. LSTMs have special memory cells that help it retain information across long periods of time. This is helpful for the modeling of our time series data, since our training data is being trained on 10 year chunks at a time, and LSTMs will be able to perform well on the temporal aspect of the data.

Using humidity data from the CESM2 dataset, our model is trained on inputs that are normalized across all observations. Using 10 year time steps in the convolution layer, we use 3x3 filters with ReLU activation to learn the features at each time step. We perform pooling by taking averages across increments of the outputs from the previous layer. This is done to reduce the dimensionality of the data by averaging multiple pixels together and feeding into the LSTM layer. For the LSTM layer, we use ReLU activation and learn the weights for the input features from the pooling layer using 35 memory cells. These memory cells help with improving the model by capturing the data for the time steps. The output of the model is formed by reshaping and adding a linear activation function to be able to predict continuous outputs of the temperature, precipitation, diurnal temperature range, and extreme precipitation.

Epochs	Num Filters	Filter Size	LSTM Units	Activation	Batch Size
30	20	3	35	ReLU	16

Table 2: CNN-LSTM Hyperparameters

3 Results

After running each model on the training data, we predicted the vapor pressure deficit on the input variables (CO_2, CH_4, BC, and SO_2) from the ssp245 scenario of the CESM2 simulation dataset. The average of the vapor pressure deficit predictions was taken over the years 2080–2100 and compared to simulated values for vapor pressure deficit. The difference between these values was plotted for each of the models below.

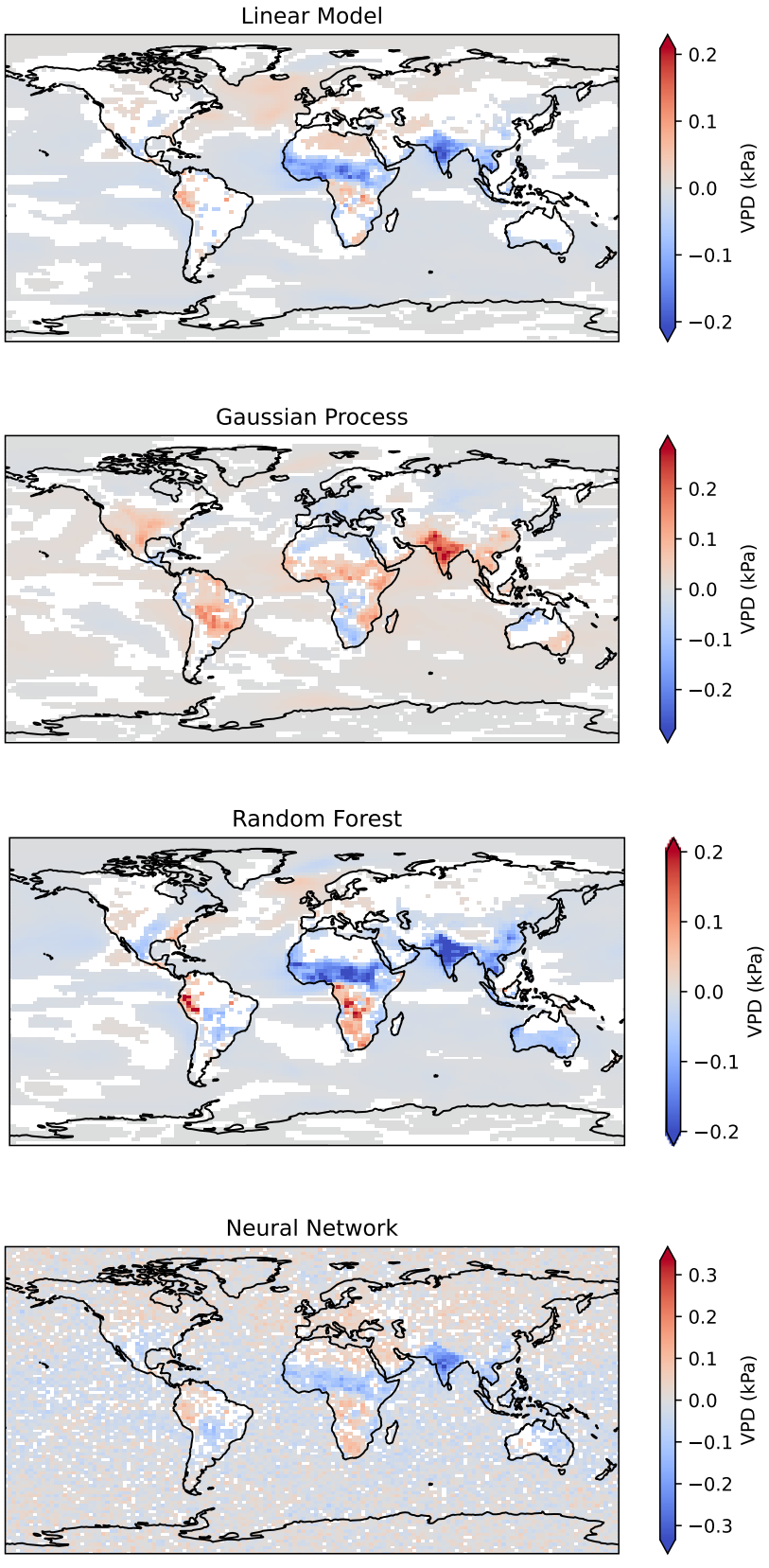


Figure 2: Model Differences Comparisons

The predictions were then compared to the vapor pressure deficit values as calculated from the ssp245 scenario. Then, the spatial, global, and total NRMSE were calculated using the methods described previously for each model. Our results showed that the linear model performed best for all three NRMSEs. The linear model had a spatial NRMSE of 0.036, a global NRMSE of 0.012, and a total NRMSE of 0.096. The model that performed second best in all NRMSE measures was the Gaussian process model, which had a spatial NRMSE of 0.044, a global NRMSE of 0.013, and a total NRMSE of 0.11. The model that performed the worst spatially and totally was the convolutional neural network, which had spatial and total NRMSEs of 0.058 and 0.154, respectively. The models that performed the worst in terms of global NRMSE were the convolutional neural network and the random forest models, with an NRMSE of 0.019.

In the following table, we compare the NRMSE results of each of our models. Global NRMSE represents the global averages across the Earth, while spatial is concerned with differences in individual regions.

Model	**Spatial**	**Global**	**Total**
Linear	**0.036**	**0.012**	**0.096**
CNN	0.058	0.019	0.154
GP	0.044	0.013	0.11
RF	0.051	0.019	0.144

Table 3: NRMSE results of different climate models used

4 Discussion

Interestingly the linear model performed better than the three other models we used. This suggests that the relationship between the input variables of CO_2, SO_2, CH_4, and BC and the output variable of vapor pressure deficit is very linear in nature.

A possible explanation for why the linear model performs so well is because of the relationship between temperature and VPD. The linear model in Watson-Parris (2022) outperforms the machine learning models when predicting temperature. If we assume relative humidity stays fairly consistent over time, then we can view VPD as a function of saturation vapor pressure (SVP), which itself is a function of temperature. We show in subsection A.2 using Equation 8 that the derivative of SVP with respect to temperature is quite small, which means that when we are looking at small changes in temperature, the relationship between VPD and temperature is mostly linear. Thus, we conclude that our linear model for VPD is likely performing around as strongly as the linear model in Watson-Parris (2022) for temperature.

When using our machine learning models to forecast temperature, diurnal temperature range, precipitation, extreme precipitation, the CNN performed very well in emulating ssp245. This does not appear to be the case when it is trained on vapor pressure deficit

data, as the NRMSE for the CNN is significantly worse than the linear baseline. This could be due to

To improve our results further, we can utilize other variables within the dataset to increase accuracy. For example, we could include evapotranspiration, wind direction, precipitation, and other variables that might not have as significant an impact as VPD on wildfires but could still contribute to greater accuracy.

Another approach we can take to improve the real-world implication of our model could be to find where trees and other possible flammable plants are prevalent on Earth. Combining our VPD data with this will allow us to predict where wildfires will occur more accurately. This would also enable our models to look at data more specifically targeted to the areas in which we are interested in the VPD predictions. An additional approach to improving the results of our models would be to remove the data that is over water and only predict the vapor pressure deficit on the area over land. This would make our models' predictions more focused on the locations where wildfires can actually occur. It would also reduce the chance that our models are biased towards predictions of VPD over the ocean. This could occur because the majority of the planet is covered by water. Removing the data over the ocean could result in results that favor our machine learning models over the linear model because there is greater variance in the VPD over land than over the oceans.

5 Conclusion

Machine learning methods can be an effective tool for emulating climate models beyond predicting simple climate metrics such as temperature and precipitation. We studied 3 different models for climate emulation: a random forest model, gaussian process model, and convolutional neural network. Based on the results of our study, linear models are good at predicting VPD data. We have achieved methods for emulating VPD predictions in ssp245 through the use of emissions data and cliamte simulations, which will give us an efficient tool for modeling future VPD. The presence of VPD can be particularly helpful in predicting which areas can be subject to wildfires. This produces many insights about climate change and how emissions will continue to play a role in that moving forward.

References

Bolton, David. 1980. "The Computation of Equivalent Potential Temperature." *Monthly Weather Review* 108(7): 1046 – 1053. [Link]

Buch, J., A. P. Williams, C. S. Juang, W. D. Hansen, and P. Gentine. 2023. "SMLFire1.0: a stochastic machine learning (SML) model for wildfire activity in the western United States." *Geoscientific Model Development* 16(12): 3407–3433. [Link]

Gallo, Eden J. M. Dieppois B. Drobyshev I. Fulé P. Z.-San-Miguel-Ayanz J., C., and M. Blackett. 2023. "Evaluation of CMIP6 model performances in simulating fire weather

spatiotemporal variability on global and regional scales." *Geoscientific Model Development* 16. [Link]

Lee, H., and J. Romero. 2023. "Summary for Policymakers." *Climate Change 2023: Synthesis Report*: 1–34. [Link]

Rodrigues, Marcos, Víctor Resco de Dios, Ângelo Sil, Àngel Cunill Camprubí, and Paulo M. Fernandes. 2024. "VPD-based models of dead fine fuel moisture provide best estimates in a global dataset." *Agricultural and Forest Meteorology* 346, p. 109868. [Link]

Schonlau, Matthias, and Rosie Yuyan Zou. 2020. "The random forest algorithm for statistical learning." *The Stata Journal* 20(1): 3–29. [Link]

Sedano, F, and JThttps Randerson. 2014. "Multi-scale influence of vapor pressure deficit on fire ignition and spread in boreal forest ecosystems." *Biogeosciences* 11(14): 3739–3755

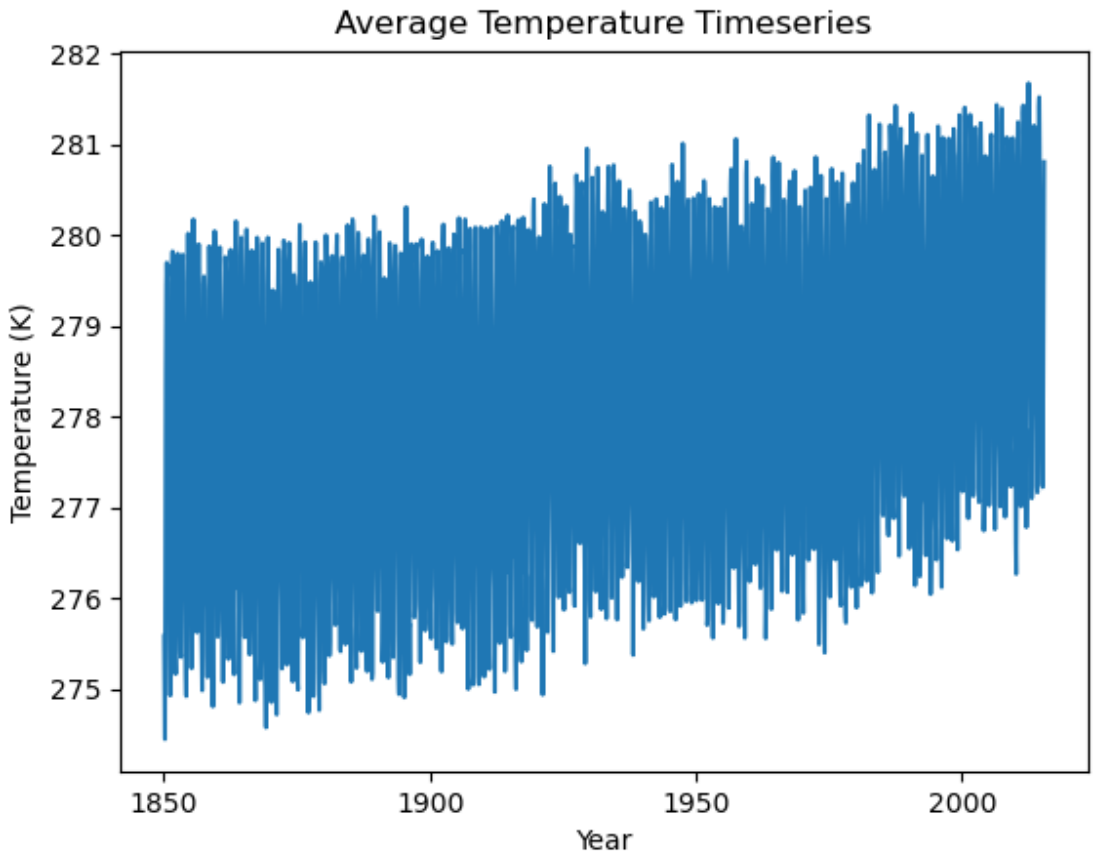
Watson-Parris, Rao Y. Olivié D. Seland Ø. Nowack P. Camps-Valls-G.-et al., D. 2022. "ClimateBench v1.0: A benchmark for data-driven climate projections." *Journal of Advances in Modeling Earth Systems* 14. [Link]

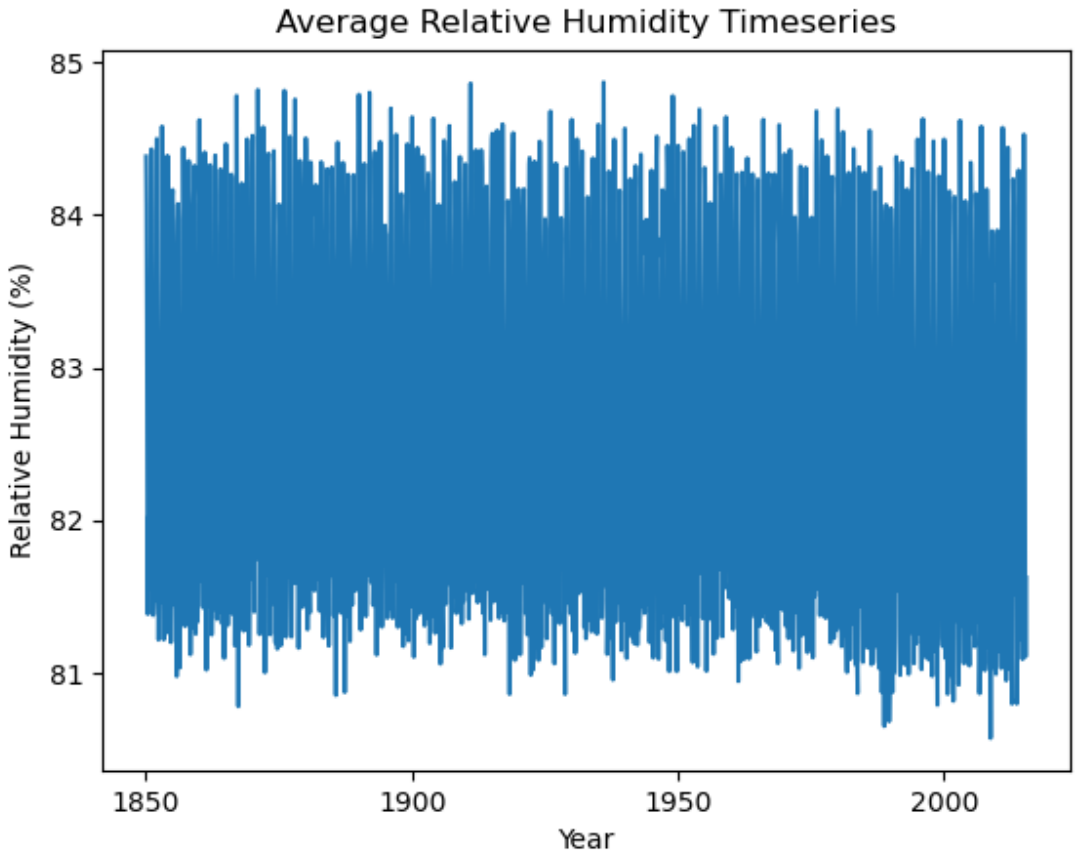
Yu, Mao J. Wullschleger S.D. et al., Y. 2022. "Machine learning–based observation-constrained projections reveal elevated global socioeconomic risks from wildfire." *Nature Communications* 13. [Link]

Appendices

A.1 Additional Figures A1
A.2 Discussion on SVP A2

A.1 Additional Figures





A.2 Discussion on SVP

There exist numerous formulas for SVP; another common formula is given in Sedano and Randerson (2014) as

$$\text{SVP}_2 = 0.6107 \cdot 10\hat{} \left(\frac{7.5T}{T + 237.3} \right). \tag{7}$$

If we denote the formula for SVP Equation 1 as SVP_1, then for a given temperature $-30°C \leq t \leq 40°C$, we have $\text{SVP}_2(t) = \text{SVP}_1(t)^\gamma$, where $0.995 \leq \gamma \leq 1.01$ depends on t. This means our formulas for SVP are roughly the same.

The derivative of SVP is given as

$$\frac{d}{dT}\text{SVP} = \frac{2629.8 \exp\left(\frac{17.67T}{T+243.5}\right)}{(T + 243.5)^2} \tag{8}$$

From this, we compute $0 < \frac{d}{dT}\text{SVP} < 0.19$ for $T < 25°C$, so SVP and temperature are fairly linear for most temperatures. For hotter temperatures, we have a higher derivative which suggest a less linear relationship. This contributes to the underpredictions of the linear

model for VPD in India and parts of the African continent, as depicted in Figure 2.