# Integrating Tree Ring and Weather Data for Combating Illegal Logging using Machine Learning

Avery Iorio, Michael Anderson, S M Rayeed, Aidan Saunders, and Charles Stewart

*Abstract*—Our study pioneers an exciting novel approach to combatting the global issue of illegal logging by seamlessly integrating tree ring and climate data. Utilizing the International Tree-Ring Data Bank and OpenMeteo Historical Weather Data API, we meticulously constructed a dataset featuring 11508 trees across 693 sites, with a focus on 28 pine species in the USA for our initial attempts. Our robust data processing pipeline addresses intricate challenges in parsing diverse tree ring data and efficiently handling large-scale weather data retrieval. Two groundbreaking models are introduced: a Verification Model, achieving an impressive 89.53% accuracy in discerning the likelihood of a log's origin, and a Location Prediction Model that ambitiously predicts tree locations solely based on tree ring measurements, employing Dynamic Time Warping and Haversine distance. The paper acknowledges ethical concerns such as false litigation, invasive data collection, and potential disruptions to supply chains. In essence, our interdisciplinary research not only offers a novel approach to combating illegal logging but also contributes to the broader discourse on the ethical, effective, and innovative use of advanced data analytics in conservation efforts.

*Keywords*—Illegal logging, tree ring data, OpenMeteo Historical Weather Data API, location verification model, location prediction model, Dynamic Time Warping, Haversine distance.

## I. Introduction

The project addresses the critical issue of illegal logging, emphasizing the unwitting involvement of manufacturers and consumers in the illicit timber trade, which is estimated to be worth $50 billion annually. The consequences extend beyond monetary value, as the trade also involves the transportation of untreated firewood, posing threats like spreading non-native pests and diseases to vulnerable trees. The team utilizes tree ring data from the International Tree-Ring Data Bank and historical weather data from OpenMeteo to establish correlations between tree ring widths and climate conditions, enabling location predictions. The comprehensive interdisciplinary research involves data processing challenges, species distribution analysis (focusing on 28 pine species in the USA), and the development of innovative analytics models for both verification and direct location prediction. Ethical considerations, such as false litigation and invasive data collection, are acknowledged, reflecting a responsible approach to conservation efforts. The work culminates in a unique database combining tree ring sequences and weather data for potential applications in combating illegal logging through anti-poaching deployment pipelines and location-based prediction models.

A Iorio, M Anderson, S M Rayeed, and A Saunders are students from the Department of Computer Science, Rensselaer Polytechnic Institute, Troy NY 12180, USA. E-mail: iorioa,anderm18,rayees,saunda4@rpi.edu

Dr. Charles Stewart is a Professor of Computer Science at Rensselaer Polytechnic Institute, Troy NY 12180, USA. E-mail: stewart@rpi.edu

## II. Related Works

The integration of weather and tree ring data to fight illegal logging is a challenging but promising novel approach. Numerous research has investigated the potential benefits of combining climate data with machine learning methods, providing new understandings of the complex interactions between environmental conditions and tree ring development. This section sheds lights on some of the notable works on dendroclimatology (i.e. how climate affects tree rings).

In their research, Jevšenak et. al explores the links between climate and growth in Quercus robur (English Oak), highlighting the importance of vessel lumen area (VLA) and its relationship to the mean temperature in April as well as the temperature at the conclusion of the previous growing season [1]. In order to forecast VLA, machine learning methods are used, with Artificial Neural Networks (ANN) and Random Forests of Regression Trees (RF) being the most effective. The comprehensive review by George et. al. examines the tree-ring width network in the Northern Hemisphere and identifies regional differences in the retention of climatic data [2]. Their study emphasizes how climatic modes such as AMO and ENSO affect tree development, underscoring the need of careful record selection in climate reconstructions. This emphasizes how critical it is to comprehend the standards for accepting or rejecting certain records, an essential component of sound dendroclimatology. In [3], Cook and Peters introduces a novel approach to standardizing forest interior tree-ring width series using smoothing splines which does not make any assumptions about the form of the curve, making it a potential tool for dendroclimatology, particularly in forest interiors where sites similar to the study sites are common. In [4], the authors investigate the distinct climatic information derived from ring width and density measurements in central and northern Alaska. The study reveals variations in annual temperatures over the past century, indicating a peak in the 1940s and near-highest levels for the past three centuries. The divergence in information from ring width and density suggests the complexity of climate responses in boreal forests, highlighting the potential impact of temperature increases and drier years on tree responses. Jevšenak et. al. investigates the use of machine learning in dendroclimatology. Comparing multiple algorithms, including ANN and Bagging of Model Trees (BMT), the research predicts temperatures at a Quercus robur (English Oak) stand in Slovenia using various tree-ring proxies [5]. BMT emerges as the most effective method, underlining the importance of comparing and selecting

optimal ML regression techniques for dendroclimatology applications. Incorporating econometric methods, Visser and Molenaar introduce a model for estimating growth trends in ring widths or basal-area increments. The model's ability to predict future growth and handle missing or unreliable data highlights its potential in dendroecological research [6]. In [7], Kuhl et. al. presents a unique application of machine learning in dendroclimatology, using tree-ring width and density to parameterize classification models for geographical provenance. The use of Extreme Gradient Boosting and density-based models offers a promising approach for assigning historical tree-ring series to specific elevations. Salehnia and Ahn designed regression and artificial intelligence models to understand the relationship between climate variables and tree ring standardized growth index (TRSGI) in northeast South Korea [8]. The study highlights the superior performance of AI approaches, particularly an artificial neural network model, in modeling TRW data compared to traditional regression models. In [9], Rahman et. al. investigates tree growth responses to climate in monsoon South Asia, specifically studying Chukrasia tabularis (Chittagong wood) in Bangladesh. The research highlights the negative correlation between tree growth and temperature, emphasizing the changing climate sensitivity and predicting further decline in tree growth with anticipated temperature increases. Gärtner et. al. introduces a protocol for integrating wood anatomical parameters into dendroecological research, emphasizing the importance of analyzing digital images to support time-series analysis [10]. The study by D'Arrigo et. al. compares temperature information inferred from tree-ring parameters, ring width, and maximum latewood density in white spruce along a forest-tundra transition in northern Canada [11]. The research concludes that both parameters are necessary for a comprehensive understanding of changes in climate and forest dynamics. In [12], the authors present a novel approach linking tree-ring data to drought simulations in dynamic vegetation models. Using a modified forward model of tree-ring width, the study captures the intra-annual variability of drought effects on tree growth.

These works collectively lay a robust foundation for the integration of tree ring and weather data, showcasing the potential of advanced statistical and machine learning techniques in enhancing our understanding of climate-growth relationships and supporting efforts to combat illegal logging.

## III. Tree Ring Data Collection and Processing

This section describes step by step process of tree ring width data collection and processing, which includes several phases - data filtration, data downloading, parsing, normalization and integration, finally followed by finalizing the novel dataset for this work.

### A. Data Source

The tree ring width has been collected from Paleoclimatic data source of the National Source of Environmental Information (NCEI)[1], a reliable data resource under the National Oceanic and Atmospheric Organization (NOAA). NCEI has an enormously large tree database, containing tree data that belongs to more than 400 species from countries all over the world, collected from more than 7300 research studies. It is noteworthy that there are two types of tree data: tree-ring data (denoted as ITRDB or, International Tree Ring Data Bank) and fire data (denoted as IMPD or, International Multiproxy Paleofire Database). Since we are only focused on the ring data from the United States (US) in this study, the first step is to properly filtrate the data before acquiring all from this gigantic reservoir.

### B. Data Filtration

As mentioned above, we are only considering the tree ring data from the US for this study. But that too, for data from trees across species, would be excessively large. So we decided to concise our work to pine trees only (genus name: Pinus). So in the first step, we set a filter to collect only tree ring data of pine trees from the US.

### C. Data Acquisition

Once we filtered the dataset according to our need, we downloaded the data. One notable feature of NCEI paleo data is that each directory downloads aligned with the corresponding studies, ensuring that the data downloaded specifically pertains to the information outlined in the conducted study. Upon collecting all the data, we found more than 750 studies on pine trees in the US over 28 species. After downloading, we organized the directories species-wise and observed the structure of the data. We found that the data comes in the form of .rwl and .txt files without any formatting (necessitates parsing the files), where rows correspond to a particular tree (with a unique tree id) and columns correspond to a particular year. And many cells contain no data for when data was not available for that particular tree on that year. Along with these, the files also contain some metadata, including study id, species name, study site name, site location, latitude, longitude, elevation, most dated and recent data of the study, study publication, doi and many more. Since the ring data in these files was not organized, we had to parse the files individually to extract the required information from each file.

### D. Data Parsing

After collecting the data we needed, we parsed the files in order to convert these unorganized stream of data into a structured csv formatted one. Each study came with a set of .crn files, one .rwl file, and one .txt file, among which we parsed the .txt files only since they contained the ring data and some metadata. The parsing was done using Python, and for each .txt file, a corresponding .csv file was generated, where

[1]Source: https://www.ncei.noaa.gov/products/paleoclimatology/tree-ring

each row represents a tree (with unique tree id) and each column represents year the data was collected, then we added some of the metadata in following columns (Northernmost Latitude, Southernmost Latitude, Easternmost Longitude, Westernmost Longitude, Site, Location, and Species). One notable feature of these .csv files is that they all contained ring data from trees ranging to different years. So, the number of columns in these .csv files varied based on the number of years the study covered. Another feature to mention in these files is, there are many null values which denotes the ring data was not available for that tree on that year.

### E. Data Integration

Since the individual .csv files (belonging to a specific study) varied in number of columns, we had to process it in a uniformed structure to fit the data into our neural network models. Among all the .csv files, the earliest recorded date for a tree ring dates back to 6000 BC. However, in order to limit the amount of the data, we decided to keep ring data only from year 1 CE to 2023 CE. So we merged all the .csv files altogether in such a way that they had ring data from year 1 to year 2023 (and put NaN if any data is missing). With this integration of data, we prepared a dataset that combines all the tree ring information of all pine trees across the US. Summary of the merged dataset is shown in Table I:

TABLE I
TREE RING DATASET STATISTICS

| Dataset Type | Tabular (.csv) |
|---|---|
| Number of Rows/Trees | 28637 |
| Number of Columns | 2031 |
| Number of Studies (ITRDB) | 764 |
| Number of Species (PINE) | 28 |
| Number of Study Sites | 693 |
| Number of Locations (States) | 40 |

For clarification, the 2031 columns are: tree id, 2023 year columns (from 1 CE to 2023 CE), Nethermost Latitude, Southernmost Latitude, Easternmost Longitude, Westernmost Longitude, Site, Location, and Species. The data is available here for further research and studies. But for this work, we had to trim to data further, as we will see in following sections.

### F. Data Distribution and Visualization

For any classification or prediction task, it is important to have a look at the dataset before starting to design a model. A comprehensive understanding of the data distribution helps making informed decisions regarding feature selection, model choice, and potential challenges. Visualization methods are critical in identifying patterns, trends, and outliers in a dataset. Visually exploring the data not only improves interpretability, but also drives the preprocessing stages, ensuring that the chosen model is appropriately aligned with the underlying properties of the data. So in this part, we will have a visual and statistical look of the data at hand.

*1) Species Distribution:* As mentioned earlier, there are 28 species of pine from US found in the integrated database, which is heavily dominated by Ponderosa Pine (known as *Pinus Ponderosa*) and Pinus Edulis (known as *Two-needle Pine*). Table II contains a statistical distribution of most common species found in the dataset :

TABLE II
MOST FREQUENT PINE SPECIES

| Species | Name | Count |
|---|---|---|
| PIPO | Ponderosa Pine | 9735 |
| PIED | Two-needle Pine | 5024 |
| PIPA | Longleaf Pine | 1961 |
| PIFL | Limber Pine | 1882 |
| PIEC | Shortleaf Pine | 1469 |
| PIJE | Jeffrey Pine | 984 |
| PIRE | Red Pine | 921 |
| PIAL | Whitebark Pine | 818 |
| PICO | Lodgepole Pine | 801 |
| PILO | Bristlecone Pine | 728 |

*2) Site and Location Distribution:* We have found tree ring data from 40 states of the US. However, like species distribution, the data was not balanced in terms of locations as well; it was highly biase towards the south-west region of the country, more specifically four states, Colorado, Arizona, New Mexico, and California. Table III enlists the states with most tree data :

TABLE III
STATES WITH MOST TREE DATA

| State | Count | Sites | State | Count | Sites |
|---|---|---|---|---|---|
| Colorado | 4331 | 100 | Nevada | 1038 | 27 |
| Arizona | 3962 | 90 | Utah | 1011 | 27 |
| California | 3697 86 | | Oregon | 899 | 34 |
| New Mexico | 3666 | 80 | Florida | 875 | 13 |
| Wyoming | 1193 | 29 | Montana | 836 | 18 |

For a better graphical representation, we plotted a heatmap of location distribution of ring data, as shown in figure 1. From the figure, we can clearly see that most of the data is from the south-western states of the US.

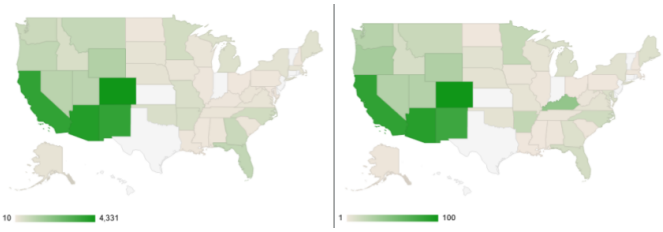


Fig. 1. Location and Site Distribution : Left (according to number of trees), Right (according to number of study sites)

We further sorted the 693 sites according to number of trees in each site, and found consistency with these heatmaps - most of the 'top-sites' are from the south-western part. We used python's folium library to pinpoint the exact coordinates of the top 20 sites and plotted it in Figure 2 :

### G. Trimming the Data

Along with tree ring part, another principle component of our work is the weather data. Since our objective is to

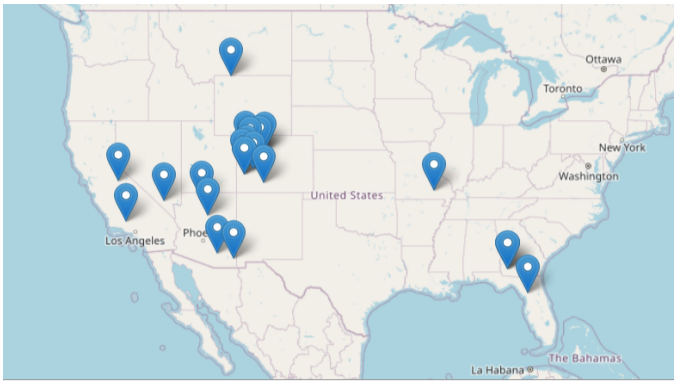


Fig. 2. Pinned coordinates of top 20 sites (according to number of trees) of our dataset. This is in consistence with the heatmap in Figure 1, showing a bias towards the south-western states of Colorado, Arizona, New Mexico, and California. The two prominent sites in the south-eastern coast is JW Jones Ecological Research Center, Georgia and Big Pine Key, Florida.

integrate the tree ring data with key weather data (temperature, precipitation) of that corresponding year, it is essential to extract the latter. Unfortunately, we could only retrieve the weather data since 1941, which has been discussed at length in the following section. On top of that, in the tree ring data, we did not have much information on years later than 1990. So we decided to take a 50-year time window, incorporating both ring and weather data of that timefrmae. So, we had to truncate our integrated tree ring dataset, limiting the included data to the period from 1941 to 1991, and subsequently discard those trees who didn't have enough ring data from that time period. After the data truncation, we had 11508 instances of trees, all merged in a single csv file (which is available here). Table IV shows a sample snapshot of the csv file for a better understanding of how the data is structured.

### *H. Data Collection: Oaks, Maples, Firs*

Following the same data collection approach, we extended our investigation to Oaks, Maples, and Firs. For Oaks, we found into 384 studies, examining 13,395 trees across 14 different species. Turning our focus to Maples, we explored 27 studies worth of data involving 936 trees, with a focus on two distinct species. For Firs, there were 89 studies containing ring data of 1914 different trees.

## IV. Weather Data Collection and Processing

The integration of weather and tree ring data to fight illegal logging is a challenging, due to the large amount of data that must be collected, but promising approach. Research by Mizanur Rahman et al [9] suggests a correlation between tree-ring width, measured precipitation levels, and measured temperature, showing a connection between climate and a tree's rate of growth. Measures by Mizanur Rahman et. al. [9] is not the only study in which this result was found; Nasrin Salehnia and Jinho Ahn [8] were able to reconstruct a tree growth index off of these climate variables with high precision. With this information in mind, we can use climate information for the tree rings we have locations on in order to try to model the correlation between ring widths and weather, as well as collect other weather points from around the United States to compare to with the goal of the model being able to distinguish a correct location for a tree-ring sequence.

The proposed method of gathering weather data has the need for large-scale data retrieval and processing. We've identified OpenMeteo as a good source of this data, as their Weather Archive API is free and hosts weather data for almost all of the United States since the 1940s. From our own research, OpenMeteo seemed to be the cheapest, most efficient source of precipitation and temperature readings for the past 50+ years. As such, the rate at which we read climate data must be greatly accelerated, as procedurally making requests to the API for each tree-ring sequence will not only take a long time to finish, but may also be limited by the rates in which these requests are processed by OpenMeteo's servers. This issue was solved using Parallel Processing. The idea is that due to the large amount of data we need to retrieve at minimum 50 years of weather data for each 50 year sequence. As we increase the model's scale this bottleneck will prove an issue, though parallelism should be able to speed up this process to the point where these wait times become minimal. We can do this by finding how many requests need to be sent, provisioning child processes to run these requests, and executing them all at once.

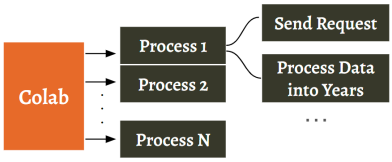


Fig. 3. Parallel Processing OpenMeteo Requests and Post-Processing on Google Colab. Instead of just sending requests in parallel, our setup configures the Pool Manager to process daily data that is returned by OpenMeteo into a yearly statistics dataframe.

The process referenced in Figure 3 points to how we solved the bottleneck with parallelism. A Python Pool is set up to manage child processes. We then wrote a wrapper class to interact with and process requests from OpenMeteo, as OpenMeteo will only allow us to retrieve daily weather data. This appeared to be the trend amongst other weather APIs and data sources we found. The wrapper works by taking the daily data and splitting into years by slicing on the dates January 1, XXXX and December 31, YYYY, with XXXX being the start year and YYYY being the end year. Then, it takes these slices and calculates the yearly average, maximum, minimum, and sum for precipitation and temperature, respectively. In order to make requests, all the wrapper needs is the start year, end year, and coordinates. The parallelized wrapper accepts the same, except for a list of coordinate tuples (longitude, latitude).

With these programs we can easily retrieve all the weather

data for the points in our set that we know. The next step is to generate random coordinate points so our models have incorrect points to compare to. This will help in our Verification Model, as the goal is to prevent it with an incorrect data point and have it verify the location is incorrect, but it will also help with our Prediction Model as the more weather data provided in training could help the model better distinguish the correlation between tree growth by ring sequences and climate. The way we accomplished this was by masking the continental United States in a bounding box and generating coordinates starting from the north-west of the mask, moving 30 miles to the east until ocean was hit, since the United States cannot be perfectly mapped into a box. To do this, we made use of the globe-land-mask Python package, which could tell us when any coordinates generated were over oceanic territories. We would then pass this set to the parallelized wrapper in order to retrieve weather data for all points generated.

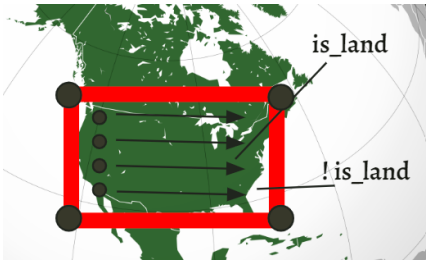


Fig. 4. Masking the United States with a bounding box allows for an easily tunable coordinate generation system, using globe-land-mask we can simply generate coordinates inside the mask, each 30 miles east from each other, and 30 miles north of the next row, while not worrying about if we are in the Great Lakes, Atlantic Ocean, Gulf of Mexico, etc.

When querying OpenMeteo, this resulted in approximately 100,000 individual requests to retrieve this scope of data. While we continuously tried to use our parallel programming solution to solve this issue, it continued to overload OpenMeto's servers. We tried serving them in batches, and this worked with the occasional dropout. However, batching 100,000 requests means a dropout can result in a lot of lost time. We estimated that it would take just over a week to retrieve data in this fashion. Eventually, we settled for another solution: instead of masking the entire United States, we could mask just the areas around where our tree ring data was collected. If we create a large enough possible radius, we could simply generate random points within a certain distance past the radius and still have spoofed weather data that our model could compare to. Figure 5 below provides a visual as to how this method works. We start by generating a radius around a point that we do not want to collect a point from (generally, this was around 100-150 miles from our data point). Then, we would generate $n$ number of points within some radius outside of the prevented collection circle. This way, we are able to fragment our data retrieval and significantly reduce the number of weather points we need. Going forward, this could grow with proper funding and resources to allow us to properly gather more weather data across the United States, though due to the limits in Hardware we are unable to store large amounts of OpenMeteo data, and due to cost we are unable query the API as much as we would have liked. Overall, the data retrieved is nominal for our machine learning model, though we are bound to the areas in which we collect weather data for.

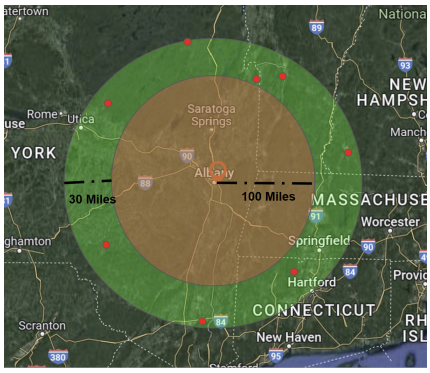


Fig. 5. With the example above around Troy, NY, the red circle shows the area of radius $X$ (in this case, 100 miles) that we assume the weather data looks similar for. The outer green circle is the zone with radius $y+x$, where $y$ is equal to the length (miles) away from $x$ we want to collect inside. in which we suspect is far away enough to collect random weather samples from while not collecting the same weather information. This way, we can scale down the weather points we initially wanted and have enough variation in weather data in our spoofed locations from the data from our tree-ring sample locations. Finally, the red dots represent random weather data points we may generate.

## V. Model Architecture

The fundamental difficulty in learning anything about the location of origin of a tree using measurements of its growth rings is that location by itself has no *direct* impact on growth rings. In other words, a tree grown in a greenhouse in Maryland with the exact same temperature and artificial precipitation as Idaho will *always* trick our model into believing the tree was grown in Idaho. We are not concerned about someone actually doing this as much as it is a point to convey that tree rings are only a surrogate for weather information. Since there is little to no existing literature on this approach we iterated through multiple model designs to try to explore exactly what type of learning was possible. In this section, we will talk about the models we proposed for predicting the location (latitude and longitudes) of a tree solely based on its tree ring pattern (model 0), predicting the location using weather data to determine the loss function (model 2), and finally verifying the claimed location of a tree based on weather and ring data (model 1).

### A. Model 0: Location from Ring Data only

In order to run an experiment for validating the use of weather data in our work, first we tried to predict the location of a tree only using the ring data. This is our first model, consisting of 51 input nodes (for tree ring data from 51 years: 1941-1991), four hidden layers (each having 256 nodes), and finally an output layer with two nodes (one for latitude and one for longitude). In this model, we used the haversine distance as the loss function, which measures the closest difference between the predicted and true latitude and longitude on the

surface of a sphere. For this model, quantifies the overall accuracy of the predicted geographical coordinates. The haversine loss function is defined as:

$r$ – Earth's radius

$y_{\text{lat}}$ – Actual Latitude

$\hat{y}_{\text{lat}}$ – Predicted Latitude

$y_{\text{lon}}$ – Actual Longitude

$\hat{y}_{\text{lon}}$ – Predicted Longitude

$$L_{\text{hav}}(y, \hat{y}) = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{y_{\text{lat}} - \hat{y}_{\text{lat}}}{2}\right) + \cos(y_{\text{lat}})\cos(\hat{y}_{\text{lat}})\sin^2\left(\frac{y_{\text{lon}} - \hat{y}_{\text{lon}}}{2}\right)}\right)$$

Architecture of the model is shown in Figure 6:

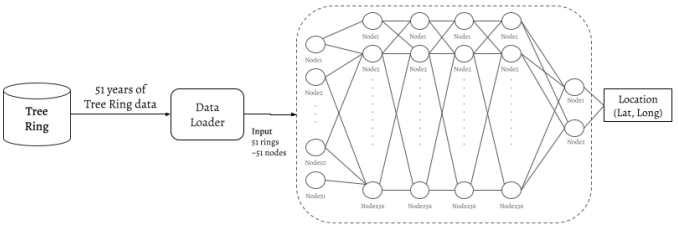


Fig. 6. Proposed model only taking tree ring data as input features, trying to predict the location of the tree. The model has a input layer with 51 nodes, four hidden layers of each having 256 nodes, followed by an output layer.

The model failed badly in its objective, yielding an accuracy of around 50%, but this was expected. The primary purpose of model 0 was to ensure that the model was not powerful enough to simply remember what tree rings looked like and motivate the need for a model that uses weather data.

*1) Relationship Between Weather and Tree Location:* Before we get into the details of the two models that utilize weather information we must first establish why weather data is both helpful and necessary. The assumption is that each location on earth has a *unique* distribution governing its "weather." We argue that this should seem obvious considering each location on earth has a unique combination of latitude and longitude, geological composition, topography, long-term weather systems, and an uncountable number of factors resulting in some distribution for temperature and precipitation on any given day. We can never know these distributions, but every day can be thought of as a sample from this distribution meaning that given long enough we can approximate this distribution. The goal of using weather in our model is that it allows us to understand how accurate the model was in connecting the tree rings to a particular location.

### B. Model 1: Verification Model

The premise behind the verification model is that a lumber mill or conservation authority may be provided with a log and then question the supplier of the log about where it was harvested from. The goal is that if the supplier did not get the log from where they purported to get it from then the model would output that the log did not match that specific location. To train this model we generated true positives and true negatives by randomly selecting a tree, getting its tree ring widths for the previous 51 years, and either attaching to it the weather from its actual location over the past 51 years or giving it the weather data from another location over the same time period. Trees given their actual weather data were labeled 1 and trees with spoofed weather data were labeled 0. We then trained this binary classifier using the 153 inputs with three hidden layers with 1024, 1024, and 512 nodes respectively. We used the binary cross entropy loss and trained the model over 50 epochs achieving a final accuracy around 89.53%.

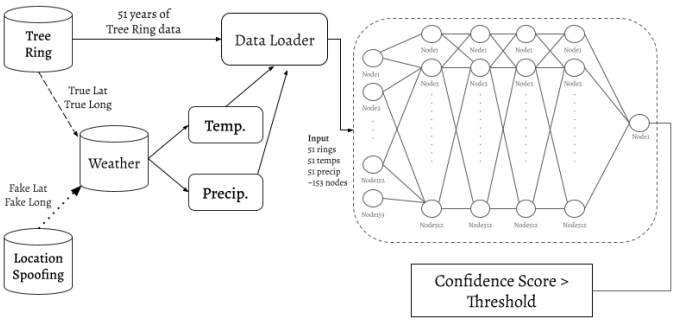


Fig. 7. Proposed model for verifying the location of a tree using ring data and weather data

We agree that since the accuracy standard for any verification task is high, that this is not yet sufficient to be deployed. However, using this verification model multiple times over a large area could serve as a *de facto* way of generating a possible location from just the tree ring data. In other words we would have lay a mesh on top of the desired area and for each point use that weather data in the verification model.

### C. Model 2: Location Prediction

The final and most ambitious model that we constructed sought to once again only take in tree ring data in the input layer and output a location, but this time use both the latitude, longitude, and the weather at the predicted and actual locations in the loss function. Our thought process here was that the reason why model 0 failed is that the relationship between location and tree rings is too non-convex for our model to learn. The primary concern is that the model makes a guess that is actually "better" in the sense that it more closely aligns tree ring widths to the weather at the actual location, but picks a location that is far away geographically from the target. In this case the model will discard the better guess for the one that is geographically closer. The reason why we define "better" as being closer to the weather at the actual location is due to our above assumption about the uniqueness of weather at each point. Our weather informed loss function seeks to see how "good" a predicted location by comparing the time series of the weather at the predicted and actual locations. In our case, we use Dynamic Time Warping (DTW) to align the temperature and precipitation time series data for two locations. The reason for this is that we know areas with similar weather might

follow a particular pattern of temperature. One such example is the impact of El Nino/ La Nina in the southwest which has a distinct influence on precipitation. The idea with using DTW is that elevation of local phenomena might change the actual precipitation values but the *pattern* is more fundamental even though it is not happening at a strict time.

$$L_{\text{weather}}(y, \hat{y}) = (y_{\text{lat}} - \hat{y}_{\text{lat}})^2 + (y_{\text{lon}} - \hat{y}_{\text{lon}})^2 + (DTW(y_{\text{temp}}, \hat{y}_{\text{temp}}) * DTW(y_{\text{precip}}, \hat{y}_{\text{precip}}))$$

Ultimately, model two failed due to what we believe was an insufficient granularity of data (specifically with regard to the weather). This prevented our model from building the necessary profile of what weather looked like over the entire geographic region so that it could learn the complex mapping of tree ring widths to location.

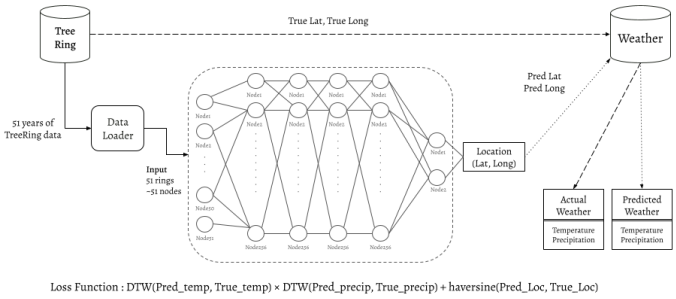


Fig. 8. Architectural diagram of model 2

## VI. Results and Findings

As stated in $Model\ 0$, we did not find success in predicting where the location a tree was harvested based on of its ring width sequence alone. Since the model requires a more diverse distribution of tree ring sequences around the United States to tell of more patterns, it's entirely probable $Model\ 0$ would find more success given more data, though it is more likely this is due to the lack of whether data. For $Model\ 2$, we can safely reference what was discussed about the limitations to retrieving publicly available weather data, as we would need more expansive computing resources in order to achieve a more reliable accuracy in this model. However, it is worth noting that we have retrieved a fair distribution of Tree Ring sequences with the inclusion of Maple, Oak, and Fir tree ring data that could better support these models. In conclusion, the failure of these models can be attributed to the distribution of the Pine data we used being mostly based in Western United States, so anything outside of these areas would be tough to map for a model who's never seen a substantial number of sequences from outside the West, as well as the lack of weather data.

For $Model\ 1$, our verification model, the introduction of a longitude and latitude allowed for the model to see weather data and spoofed weather data so it could recognize trends inside the areas our data was collected as well as outside. This yielded a much higher success rate, with our best model performing at $89.53\% Accuracy$.

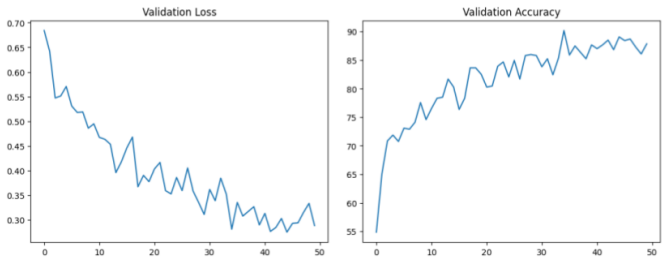


Fig. 9. We saw a steady decrease in Validation Loss with a steady increase in Validation accuracy. Our best model performed with an 89.53% accuracy

Our model seemed to be a little less accurate when it came to predicting trees outside of the zones we trained and tested on, this is most likely because while a pine from Colorado and New York might be similar, the lack of weather data could be making the model think a New York zone its never seen before could look sort of like a Colorado zone, and make an improper prediction. Figure 11 displays the frequency of predictions based on pines that we gave our model from the Validation set.

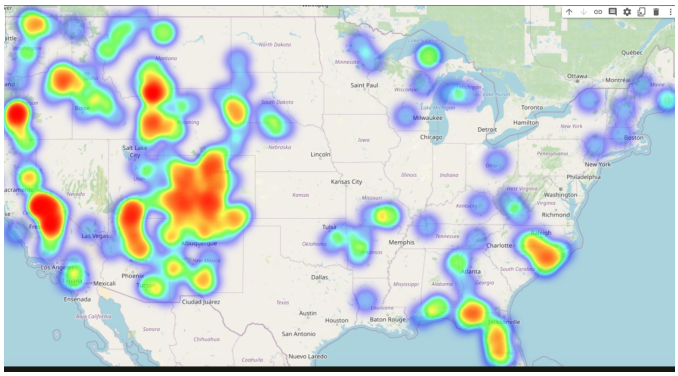

Fig. 10. Our model saw great success in verifying Pines from the West Coast. While the Northeast did have proper predictions as well, some of these predictions were made with less precision than those made on the Western side of the USA. With the introduction of Maple, Fir, and Oak data, we hope to change this.

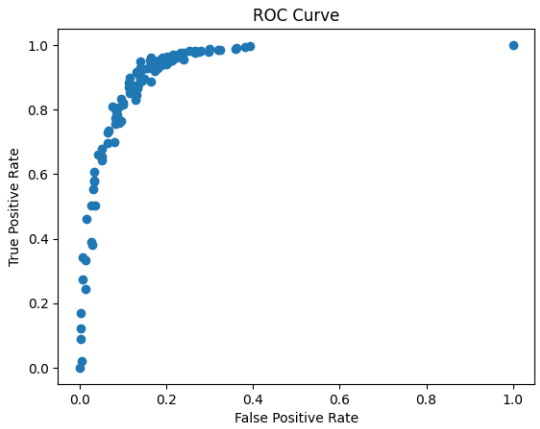


Fig. 11. Model 1's Receiver Operating Characteristic (ROC) curve shows a quick rise to towards the top left, indicating a high true positive rate for a low false positive rate.

Overall, we believe $Model\ 1$ can reliably be deployed in the

logging industry to check whether or not a supplier is being truthful of the location of harvest when raw lumber is being sold to distributors. We found that the better distributions of Trees around the USA could potentially yield better accuracy for our models, and we are excited to discuss further works on this project in order to better refine these models!

## VII. Ethical Concerns

*1) False Litigation:* When trying to stop illegal wood poaching, there is a substantial risk of false litigation. If any individual, company or enterprise is wrongly accused of such charges, it can result in legal troubles, leading to delays, stress, and financial losses for the accused. These issues not only disrupt legal processes but also impact the overall environmental protection initiative. It is crucial to exercise caution, consider both perspectives when addressing this concern.

*2) Invasive Data Collection:* The methodology for measuring tree rings often involves invasive techniques, including drilling into and, at times, cutting down trees. One might argue that the alternative process of data collection, which is measuring the tree ring widths of the fallen trees only - often results in potential inaccuracies and limitations in the acquired data. Depending upon how long ago the tree has fallen, the ring measurements change, significantly impacts the accuracy of ring measurements, introducing complexities in data processing.

*3) Disruption of Major Supply Chains and Socioeconomic Impacts:* The identification and intervention against illegal logging, though crucial for conservation, carry substantial social and economic consequences. The impact reverberates through major supply chains, particularly affecting vulnerable and marginalized communities. The sudden detection of illegal activities can disrupt these supply chains, leading to increased prices for building materials and exerting immense pressure on local economies that depend on the harvesting of illegal lumber. The ethical dilemma lies in balancing the need for conservation with the unintentional hardships imposed on communities already dealing with socioeconomic challenges.

## VIII. Conclusion and Future Work

In conclusion, the integration of tree ring data and weather data marks a significant milestone in the field, addressing critical issues such as illegal logging, and poached lumber. The model proposed for verifying the claimed location of a tree log has demonstrated an impressive accuracy, providing a powerful tool for combating illegal logging activities. With 28 distinct species examined in 764 studies, our study on pine trees in the United States demonstrates the scope and depth of the field's research. The International Tree-Ring Data Bank (ITRDB) and OpenMeteo Historical Weather Data are used in the data preparation and filtering process, which demonstrates the difficulty of managing big datasets from a variety of sources and formats. The ethical concerns associated with false litigation, invasive data collection methods, and potential disruptions to major supply chains have been duly acknowledged. It is crucial to approach these concerns with caution, understanding the implications and ensuring responsible and ethical use.

Looking ahead, the research paves the way for future endeavors. The successful integration of tree ring sequences with weather data for pine trees prompts exploration into other tree species such as oaks, maples, and firs. The ongoing works on these species, with 500 studies and more than 16000 trees, suggest a comprehensive expansion of the methodology. As the research progresses, refining the model for location prediction and addressing challenges in mapping tree rings to location will be crucial. The exploration of dynamic time warping (DTW) for comparing weather data and the consideration of Kullback-Leibler Divergence (KL Divergence) for measuring underlying weather distributions would help advancing the precision and accuracy of location predictions. In summary, the integrated dataset of tree ring sequences and weather data represents a pioneering achievement. The verification model's success lays a foundation for broader applications and inspires confidence in the potential of combining ecological and climatic data for conservation purposes. The ongoing and future work on diverse tree species and the exploration of advanced modeling techniques position this research at the forefront of scientific innovation in environmental conservation.

## Appendix A
## Mapping Tree Rings to Location

The primary challenge lies in establishing a logical framework for assessing the efficacy of location predictions within the context of tree-ring data. Our model's training process spans multiple epochs, allowing it to gradually discern the intricate relationship between tree rings and geographical coordinates. The underlying assumption is that specific latitudes and longitudes are linked to distinct patterns within the tree-ring data. However, achieving this understanding demands weather data at a sufficiently granular level. Our model's ability to learn hinges on comprehending the non-convex function that maps a given location to its corresponding weather conditions. Complicating matters further is the reliance solely on tree-ring data as a proxy for weather. The absence of direct meteorological measurements necessitates a robust approach, as our model strives to uncover the nuanced connections between tree rings and the environmental factors that shape them. In navigating these challenges, Appendix A serves as a comprehensive exploration of the methodology employed in mapping tree rings to precise geographical locations. By providing insights into the intricacies of the relationship between tree-ring data and weather conditions, this section enhances the transparency and understanding of our modeling approach for the conference audience.

## Appendix B
## Comparing Weather

In this section, we break down how we ensure our model's accuracy in predicting locations. Beyond just examining

weather conditions, we consider both the actual and predicted locations, with a penalty for their geographical separation. Our model's core assumption is that weather disparities arise from differences in measurements at specific times. To address this, we employ Dynamic Time Warping to adjust time measurements, accommodating minor weather variations in nearby areas. Rather than focusing on daily conditions, we propose that weather similarity is better understood through the broader patterns influencing daily samples, shaped by geographical and temporal factors. This shift leads us to evaluate weather similarity based on the likeness of underlying distributions, introducing Kullback-Leibler Divergence as a metric that emphasizes information over the specific time of observation. Appendix B unveils the layers of our refined approach, showcasing how KL Divergence aids in discerning subtle distinctions in weather patterns, offering conference attendees a straightforward insight into the sophisticated methodologies applied in our research.

## ACKNOWLEDGMENT

We extend our utmost gratitude to Professor Chuck Stewart for his support, guidance and direction throughout this work.

## REFERENCES

[1] Jernej Jevšenak, Sašo Džeroski, Tom Levanič, and others, *Predicting the vessel lumen area tree-ring parameter of Quercus robur with linear and nonlinear machine learning algorithms*, *Geochronometria*, Volume 45, Number 1, Pages 211–222, Year 2018, Publisher: Sciendo.

[2] Scott St George, *An overview of tree-ring width records across the Northern Hemisphere*, *Quaternary Science Reviews*, Volume 95, Pages 132–150, Year 2014, Publisher: Elsevier.

[3] Stan Dardizing Forest Interior Tree-Ring Series Width, *The Smoothing Spline: A New Approach to*, *with the cooperation of The Laboratory of Tree-Ring Research University of Arizona*, Volume 41, Page 45, Year 1981.

[4] Gordon C Jacoby and Rosanne D D'Arrigo, *Tree ring width and density evidence of climatic and potential forest change in Alaska*, *Global Biogeochemical Cycles*, Volume 9, Number 2, Pages 227–234, Year 1995, Publisher: Wiley Online Library.

[5] Jernej Jevšenak, Sašo Džeroski, Saša Zavadlav, Tom Levanič, *A machine learning approach to analyzing the relationship between temperatures and multi-proxy tree-ring records*, *Tree-ring Research*, Volume 74, Number 2, Pages 210–224, Year 2018, Publisher: Tree-Ring Society Laboratory of Tree-Ring Research, Building 58, University.

[6] Hans Visser and Jaap Molenaar, *Estimating trends in tree-ring data*, *Forest Science*, Volume 36, Number 1, Pages 87–100, Year 1990, Publisher: Oxford University Press.

[7] Eileen Kuhl, Christian Zang, Jan Esper, Dana FC Riechelmann, Ulf Büntgen, Martin Briesch, Frederick Reinig, Philipp Römer, Oliver Konter, Martin Schmidhalter, and others, *Using machine learning on tree-ring data to determine the geographical provenance of historical construction timbers*, *Ecosphere*, Volume 14, Number 3, Page e4453, Year 2023, Publisher: Wiley Online Library.

[8] Nasrin Salehnia and Jinho Ahn, *Modelling and reconstructing tree ring growth index with climate variables through artificial intelligence and statistical methods*, *Ecological Indicators*, Volume 134, Page 108496, Year 2022, Publisher: Elsevier.

[9] Mizanur Rahman, Mahmuda Islam, Jakob Wernicke, Achim Bräuning, *Changes in sensitivity of tree-ring widths to climate in a tropical moist forest tree in Bangladesh*, *Forests*, Volume 9, Number 12, Page 761, Year 2018, Publisher: MDPI.

[10] Holger Gärtner, Paolo Cherubini, Patrick Fonti, Georg von Arx, Loïc Schneider, Daniel Nievergelt, Anne Verstege, Alexander Bast, Fritz H Schweingruber, Ulf Büntgen, *A technical perspective in modern tree-ring research-how to overcome dendroecological and wood anatomical challenges*, *JoVE (Journal of Visualized Experiments)*, Number 97, Page e52337, Year 2015.

[11] Rosanne D D'Arrigo, Gordon C Jacoby, Rosemary M Free, *Tree-ring width and maximum latewood density at the North American tree line: parameters of climatic change*, *Canadian Journal of Forest Research*, Volume 22, Number 9, Pages 1290–1296, Year 1992, Publisher: NRC Research Press Ottawa, Canada.

[12] Marco Mina, Dario Martin-Benito, Harald Bugmann, Maxime Cailleret, *Forward modeling of tree-ring width improves simulation of forest growth responses to drought*, *Agricultural and Forest Meteorology*, Volume 221, Pages 13–33, Year 2016, Publisher: Elsevier.

TABLE IV
SAMPLE FILE STRUCTURE

| Tree ID | 1941 | 1942 | 1943 | ... | 1990 | 1991 | Site | Loc | N | S | E | W | Species |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TH-007 | 1.31 | 1.8 | 1.61 | ... | 0.99 | 0.93 | Telephone Hill | Arizona | 36.603 | 36.603 | -112.156 | -112.156 | PIPO |
| TH-008 | 1.26 | 1.8 | 1.41 | ... | 0.79 | 0.81 | Telephone Hill | Arizona | 36.603 | 36.603 | -112.156 | -112.156 | PIPO |
| TH-009 | 1.39 | 1.82 | 1.58 | ... | 1.02 | 0.91 | Telephone Hill | Arizona | 36.601 | 36.601 | -112.153 | -112.153 | PIPO |
| TH-010 | 1.42 | 1.75 | 1.55 | ... | 1.00 | 0.94 | Telephone Hill | Arizona | 36.606 | 36.606 | -112.155 | -112.155 | PIPO |
| TH-011 | 1.48 | 1.79 | 1.59 | ... | 1.03 | 0.92 | Telephone Hill | Arizona | 36.604 | 36.604 | -112.158 | -112.158 | PIPO |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| DRY01AAU | 0.038 | 0.037 | 0.021 | ... | 0.31 | 0.03 | Dry Park | Colorado | 38.26 | 38.26 | -108.33 | -108.33 | PIED |
| DRY02AAU | 0.035 | 0.038 | 0.019 | ... | 0.33 | 0.032 | Dry Park | Colorado | 38.25 | 38.25 | -108.34 | -108.34 | PIED |
| DRY03AAU | 0.037 | 0.036 | 0.023 | ... | 0.29 | 0.028 | Dry Park | Colorado | 38.27 | 38.27 | -108.32 | -108.32 | PIED |
| DRY04AAU | 0.040 | 0.035 | 0.020 | ... | 0.30 | 0.031 | Dry Park | Colorado | 38.28 | 38.28 | -108.31 | -108.31 | PIED |
| DRY05AAU | 0.036 | 0.039 | 0.022 | ... | 0.32 | 0.029 | Dry Park | Colorado | 38.29 | 38.29 | -108.30 | -108.30 | PIED |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| BLM01B3 | 2.16 | 3.69 | 2.21 | ... | 0.68 | 1.17 | Bally Mountain | Oregon | 45.28 | 45.28 | -118.57 | -118.57 | PIPO |
| BLM02B3 | 2.11 | 3.71 | 2.18 | ... | 0.72 | 1.14 | Bally Mountain | Oregon | 45.27 | 45.27 | -118.58 | -118.58 | PIPO |
| BLM03B3 | 2.09 | 3.67 | 2.25 | ... | 0.70 | 1.20 | Bally Mountain | Oregon | 45.29 | 45.29 | -118.56 | -118.56 | PIPO |
| BLM04B3 | 2.13 | 3.65 | 2.20 | ... | 0.66 | 1.19 | Bally Mountain | Oregon | 45.30 | 45.30 | -118.55 | -118.55 | PIPO |
| BLM05B3 | 2.14 | 3.70 | 2.17 | ... | 0.69 | 1.18 | Bally Mountain | Oregon | 45.26 | 45.26 | -118.59 | -118.59 | PIPO |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |