Final Project Stor 565: COLLEGE DROPOUTS

Global Temperature Change Predictions

Kayla Casey, Josh Carlson, Claudia Dare, Cora McAnulty

TABLE OF CONTENTS

Temperature Change on Land

FAOSTAT (Food and Agriculture Organization of the United Nations): Worldwide temperature change data

- Years Collected: 1961-2023
  - Data collected on a monthly basis
- Areas collected based on Countries, Region and "Special Groups"
  - 198 countries and 39 territories
  - 28 regions
  - 9 special groups
- Temperatures recorded in Celsius, with change being recorded as change compared to a baseline climatology from 1951-1980

All datasets offered by FAO are freely available, from GISTEMP data, and Global Surface Temperature Change data from NASA

Ecological Footprint

NFA (National Footprints Accounts from the UN):

- Years Collected: 1961-2014
  - Data collected on a yearly basis
- Ecological footprint is a measure of how much “biologically active” land is required to support the population and activities of a region
- Can be separated into specific types of lands: crop, grazing, forest, fishing, and built-up land
- Each country’s output measured looking at area, looking at biocapacity and ecological footprint

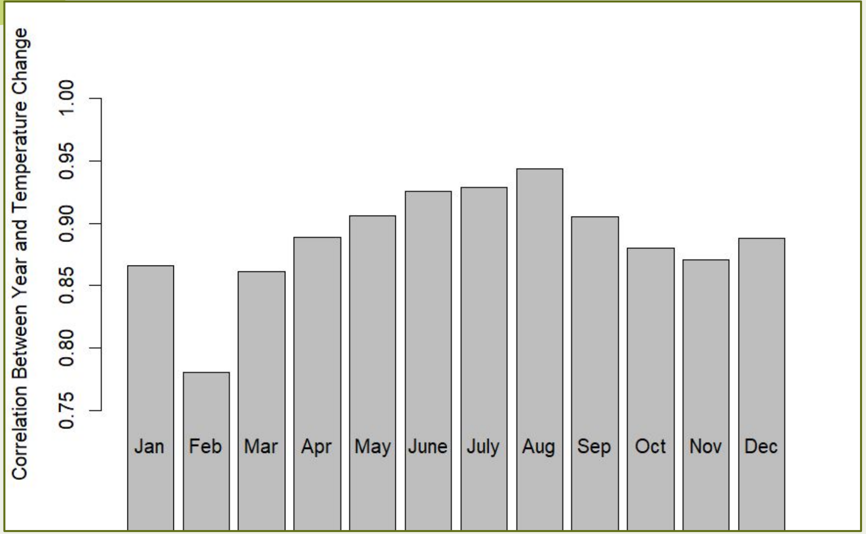
Data owned by Global Footwork Network and freely available on data.world

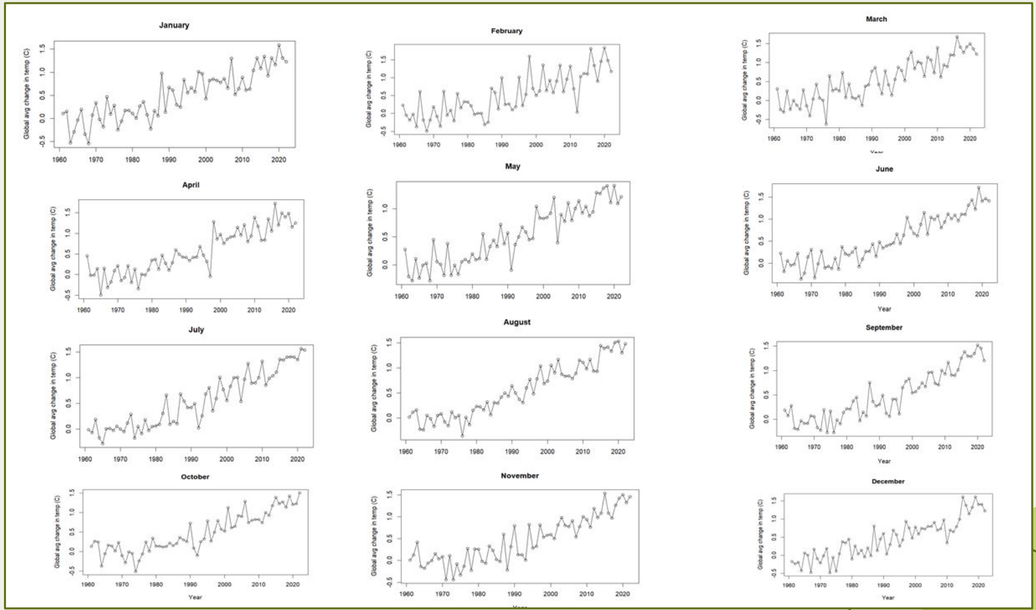
Goal

Accurately predict worldwide temperature change using various time series, ensemble, and neural network models, comparing results and performance across models to determine which model is “best”.

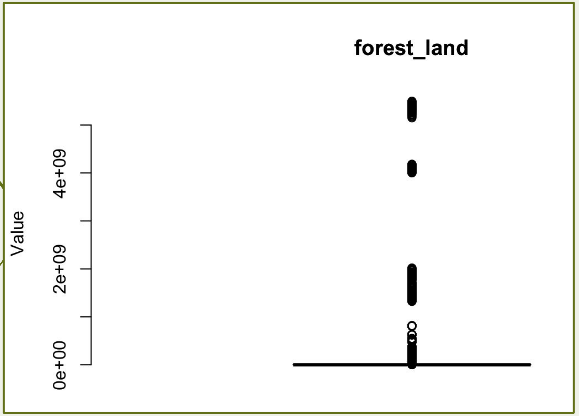
Create said models according to worldwide, geographic proximity, and groups determined using ecological footprint data.

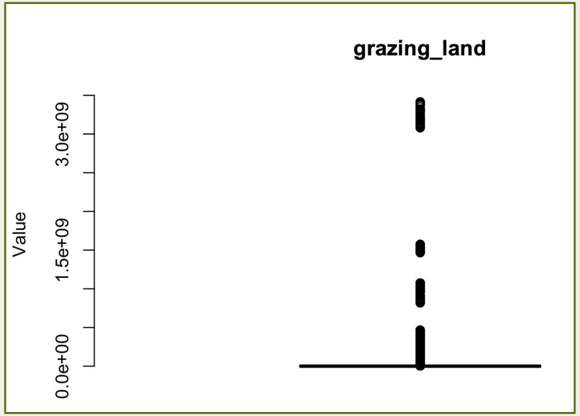
Initial EDA: Visualizing Data

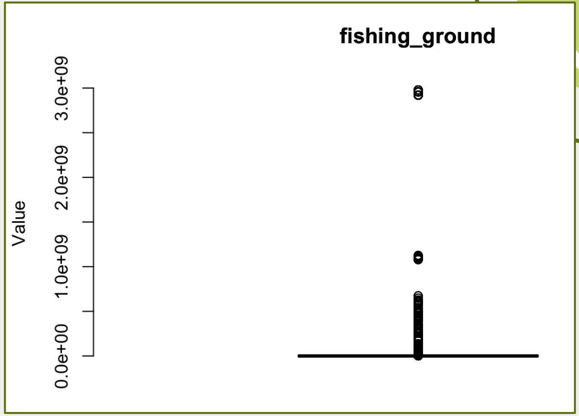




Initial EDA: NFA Dataset



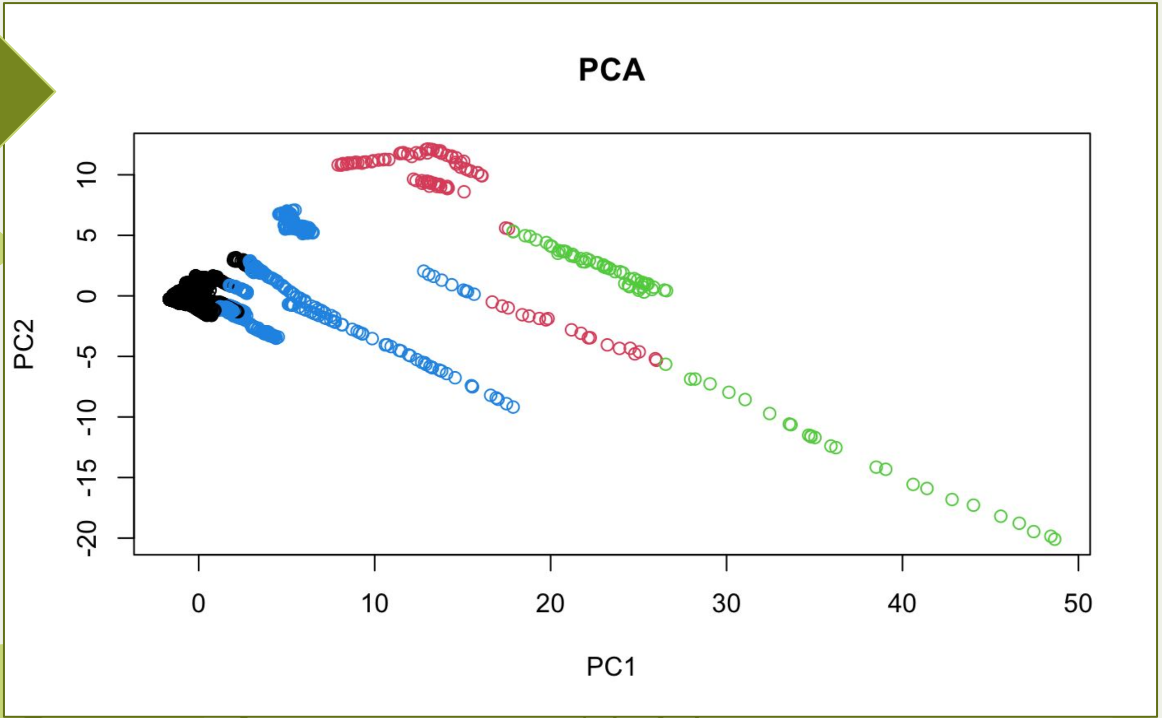




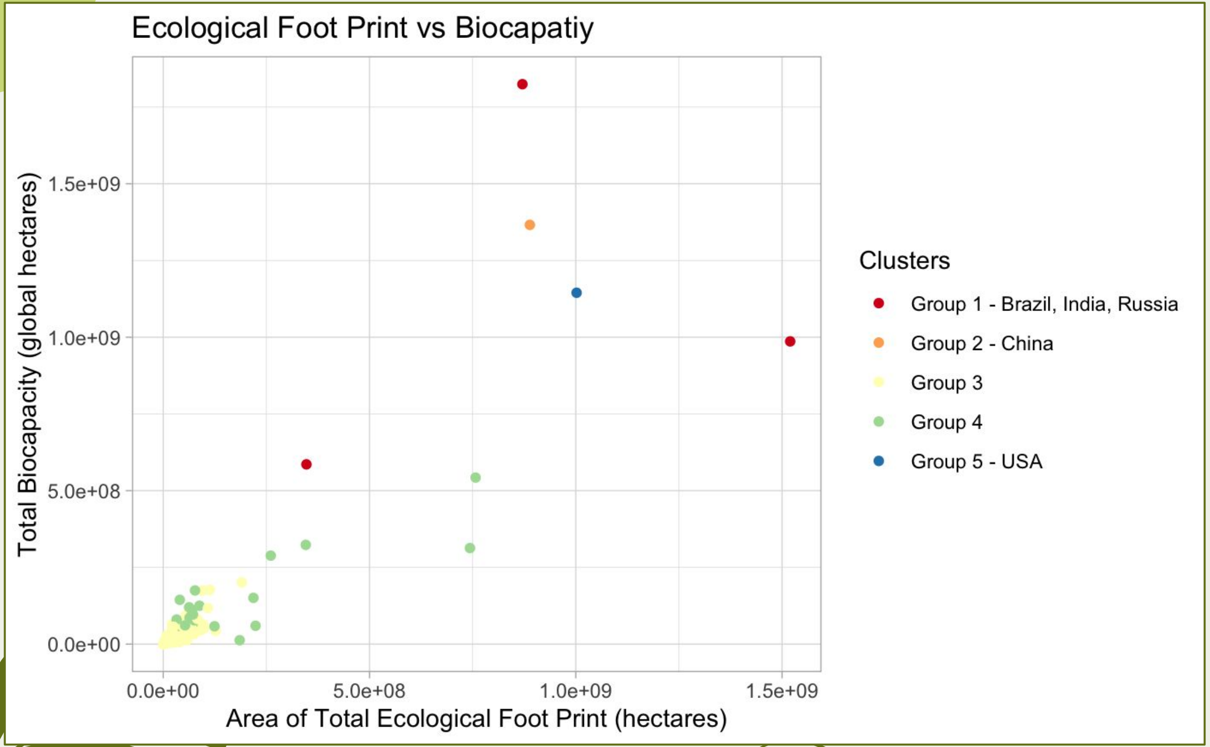
Boxplots used to visualize the spread of numerical variables in the NFA dataset. This initial EDA helped us see that this data set might not provide the same amount of information for every country so it should be used in a more supplementary way. Moving forward we thought it would be a good idea to use this data set to create clusters since we could group countries based on similarity in their land use.

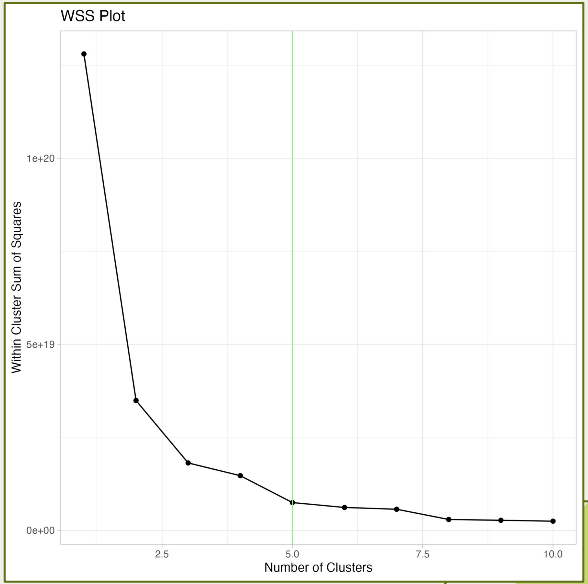
Initial EDA: Clustering

NFA Data Set, 53 Variables...



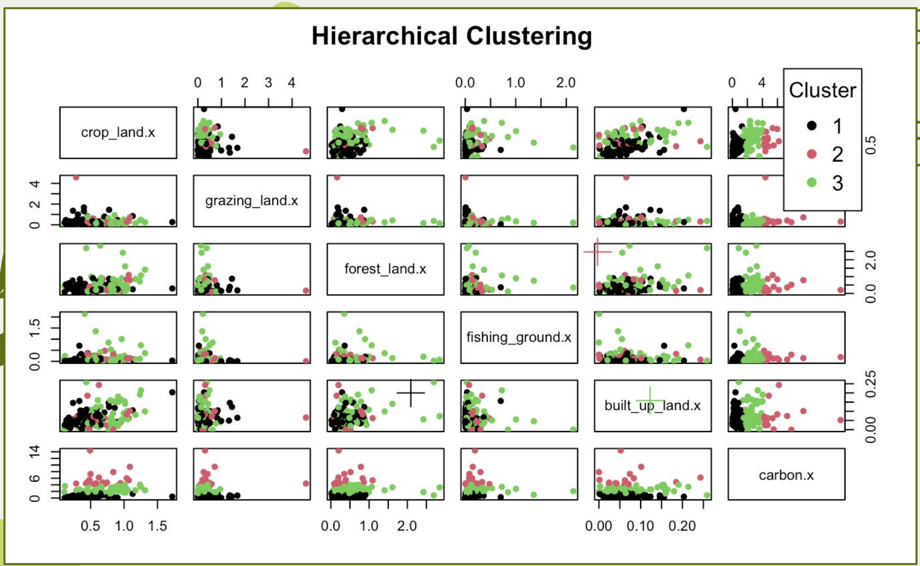
K-Means Clustering



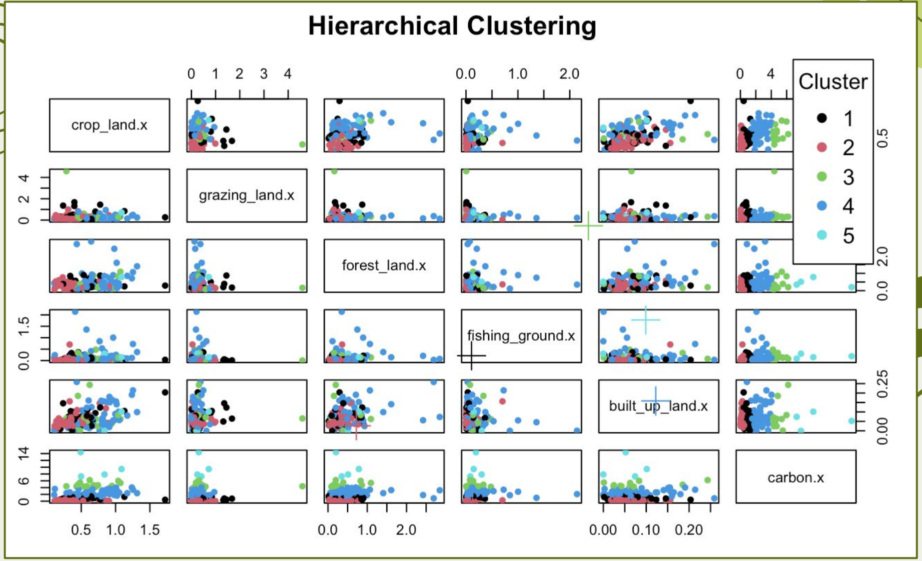


Hierarchical Clustering

3 Clusters



5 Clusters



- Distance Matrix created using Euclidean method
- Clusters used the NFA dataset for year = 2014

Final Clusters

K-Means, 5 Clusters

- Group 1 - Brazil, India, Russia
- Group 2 - China
- Group 3 - All Other Countries
- Group 4 - Argentina, Australia, Canada, France, Germany, Indonesia, Iran, Italy, Japan, Kazakhstan, Korea, Mexico, Nigeria, Poland, Saudi Arabia, South Africa, Spain, Thailand, Turkey, Ukraine, UK, Vietnam
- Group 5 - United States of America

Hierarchical, 3 Clusters

- Group 1 - Armenia, Afghanistan, Albania, Algeria, Angola, Argentina, Bangladesh, Bolivia, Brazil, Myanmar, Burundi, Cameroon, Central African Republic, Sri Lanka, Chad, Colombia, Congo, Costa Rica, Cuba, Azerbaijan, Benin, Dominican Republic, Ecuador, El Salvador, Djibouti, Georgia, Gabon, Gambia, Ghana, Guatemala, Guinea, Guyana, Haiti, India, Indonesia, Iraq, Cote d'Ivoire, Jamaica, Jordan, Kyrgyzstan, Kenya, Lao People's Democratic Republic, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mexico, Morocco, Mozambique, Moldova, Namibia, Nepal, Nicaragua, Niger, Nigeria, Pakistan, Panama, Papua New Guinea, Paraguay, Peru. Philippines, Guinea-Bissau, Timor-Leste, Eritrea, Zimbabwe, Rwanda, Saint Lucia, Senegal, Sierra Leone, Somalia, Tajikistan, Swaziland, Syrian Arab Republic, United Republic of Tanzania, Thailand, Togo, Tunisia, Uganda, Burkina Faso, Uruguay, Uzbekistan, Viet Nam, Ethiopia, Yemen, Democratic Republic of Congo, Zambia, Sudan, South Sudan
- Group 2 - Australia, Bahrain, Brunei Darussalam, Canada, Kazakhstan, Republic of Korea, Kuwait, Mongolia, Qatar, Saudi Arabia, Singapore, Turkmenistan, Trinidad and Tobago, Oman, United Arab Emirates, United States of America, Luxembourg
- Group 3 - Austria, Bahamas, Barbados, Bhutan, Botswana, Belize, Bulgaria, Chile, Denmark, Belarus, Equatorial Guinea, Estonia, Fiji, Finland, France, French Guiana, French Polynesia, Germany, Bosnia and Herzegovina, Greece, Guadeloupe, Hungary, Croatia, Iran, Islamic, Ireland, Israel, Italy, Japan, Democratic People's Republic of Korea, Latvia, Lebanon, Libyan Arab Jamahiriya, Lithuania, Malaysia, Malta, Netherlands, Macedonia TFYR, Norway, Czech Republic, Poland, Portugal, Romania, Russian Federation, Slovenia, Slovakia, South Africa, Spain, Suriname, Sweden, Switzerland, United Kingdom Turkey, Ukraine, Bolivarian Republic of Venezuela, Belgium, Serbia, Montenegro, China

Supervised Models

After our EDA, we decided to create 3 types of models:

- Time Series
- Ensemble when appropriate
- Recurrent Neural Networks

Within these types of models, we created multiple models of each for:

- Worldwide data
- Geographical Region Data (e.g. continent)
- Clusters & Hierarchical Data

Starting with Time Series models, we followed these steps for each model:

1. Remove trend and seasonality
2. Check for iid noise and stationarity
3. Fit to appropriate model with coefficients selected due to AIC, ACF, and check the Ljung-Box test
4. Test model on data 5 years in the past, report on RMSE
5. Forecast into the future

Why we chose ARMA(p,q) models

$$\hat{X}_t \approx \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{k=1}^{q} \theta_k Z_{t-k} + \hat{Z}_t$$

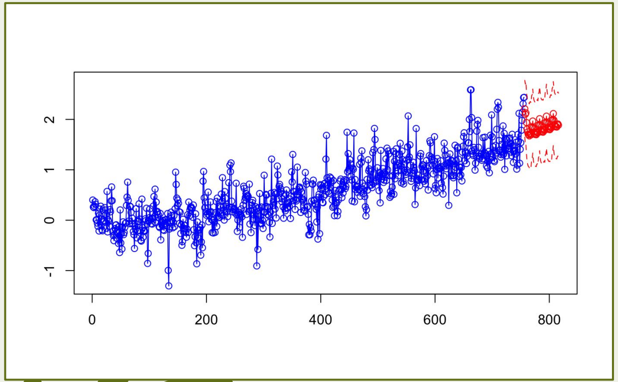
White Noise Term

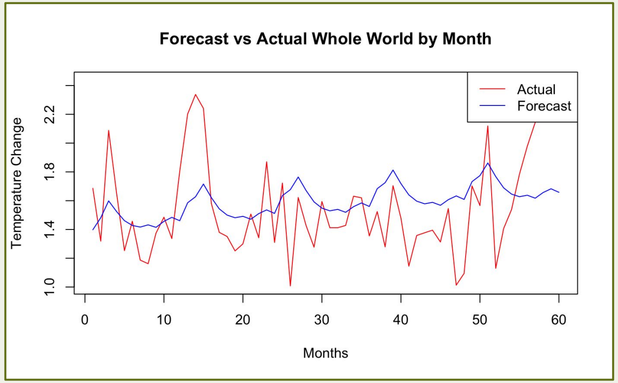
AutoRegressive Moving Average model:

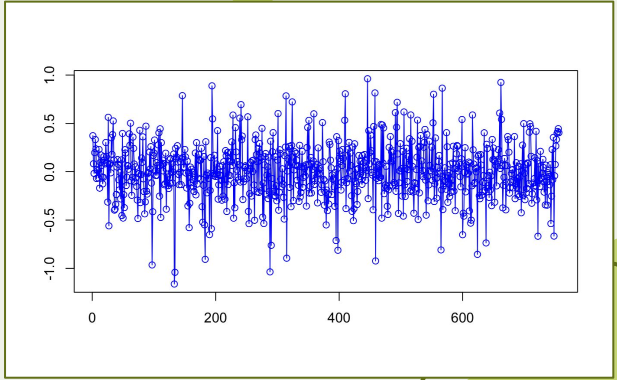
- Combination of two linear models = one linear model:
    - AR(p): captures linear relationship b/w observation and past values (lagged observations) which is useful for calculating temporal dependencies like trend and seasonality
    - MA(q): captures relationship between observations and a linear combination of past error terms which is useful in finding short term fluctuations in the data
- Data exhibits iid noise and stationary behavior, making it well suited for ARMA modelling
- ARMA and ARIMA models are both used in climate change forecasting

Why we didn't use monthly data for time series models

- Data is very noisy
- After removing seasonality, the time series models give good confidence intervals but poor forecasting results
- RMSE values are high, hard to predict monthly deviations as they are due to many factors





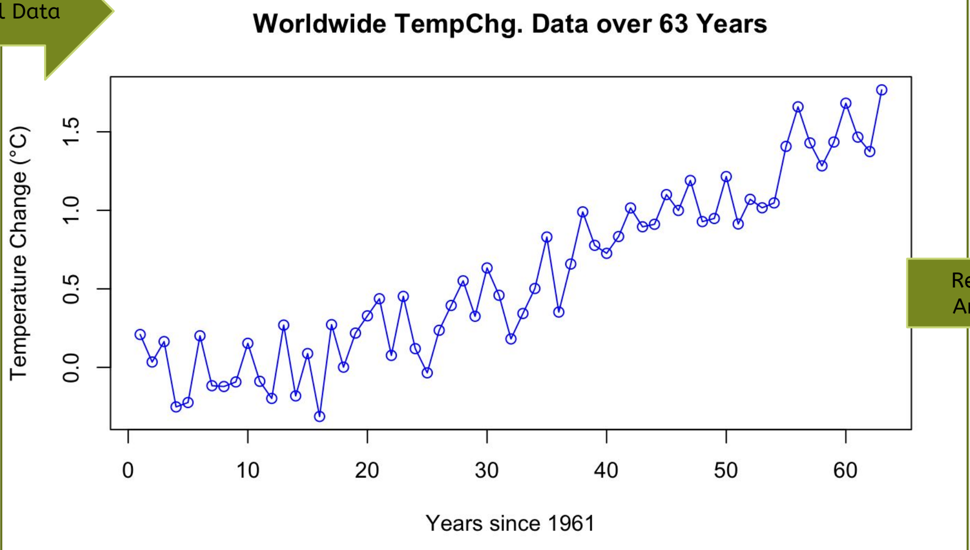


Mean Absolute Error (MAE): 0.2690869
Root Mean Squared Error (RMSE): 0.3389958

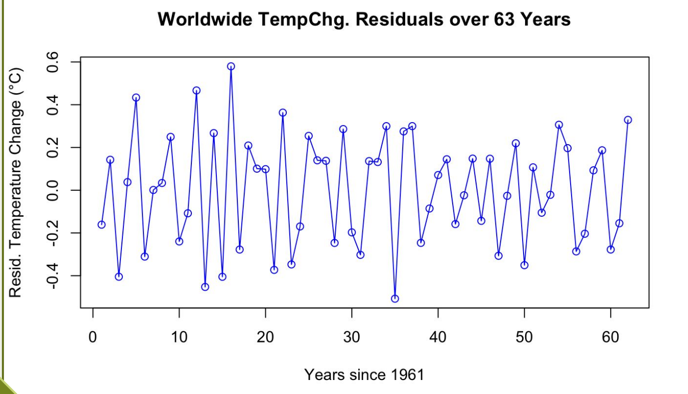
Data_model = c("season", 1,"trend", 2)

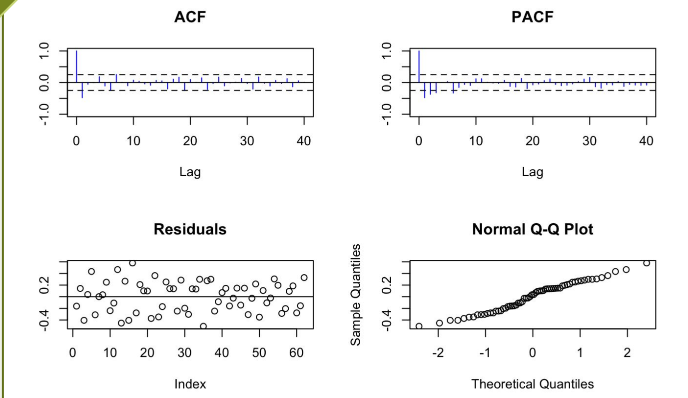
Worldwide TS Model

Original Data



Residual Analysis





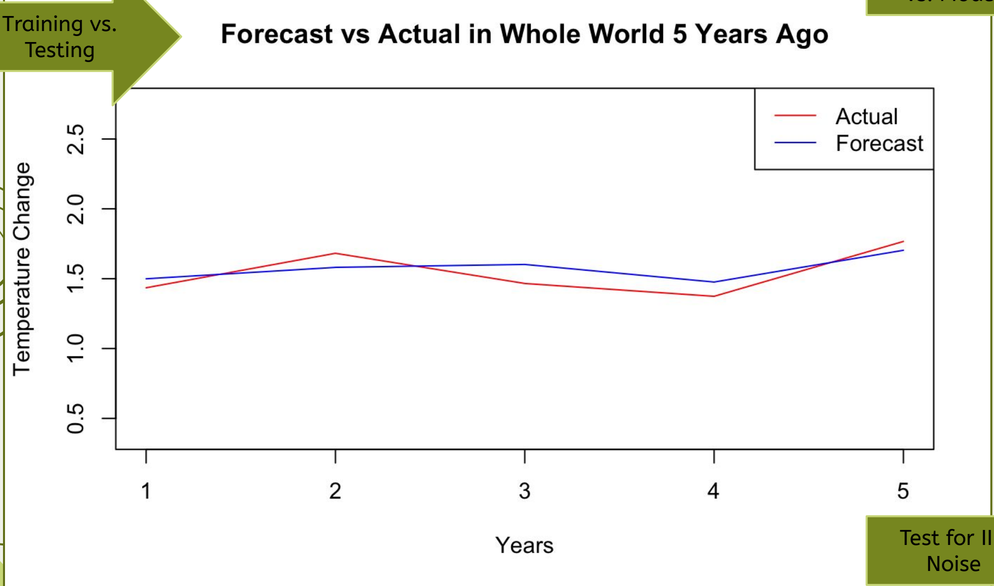
Test for Stationary

Augmented Dickey-Fuller Test

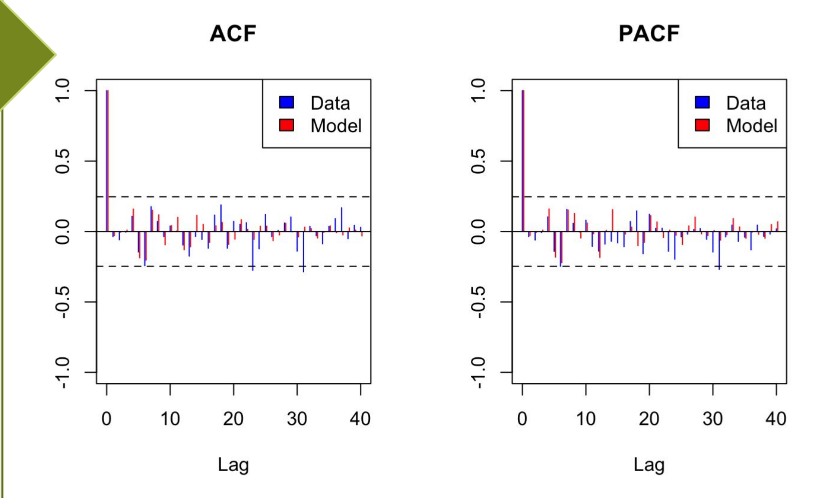
data: world_residuals
Dickey-Fuller = -3.5639, Lag order = 3, p-value = 0.04357
alternative hypothesis: stationary

Worldwide TS Model

Training vs. Testing



ACF in Data vs. Model



Testing RMSE

Mean Absolute Error (MAE): 0.09337891
Root Mean Squared Error (RMSE): 0.09725503

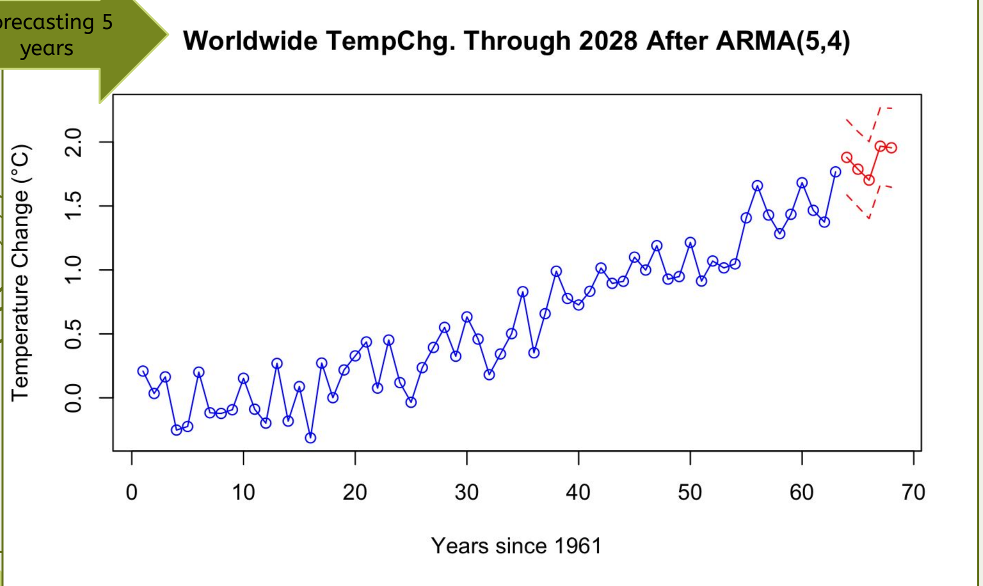
Test for IID Noise

<u>Null hypothesis: Residuals are iid noise.</u>

Test	Distribution	Statistic	p-value
Ljung-Box Q	Q ~ chisq(20)	17.13	0.6442
McLeod-Li Q	Q ~ chisq(20)	12.35	0.9036
Turning points T	(T-40)/3.3 ~ N(0,1)	44	0.2214
Diff signs S	(S-30.5)/2.3 ~ N(0,1)	29	0.5127
Rank P	(P-945.5)/82.3 ~ N(0,1)	869	0.3527

Worldwide TS Model

Forecasting 5 years



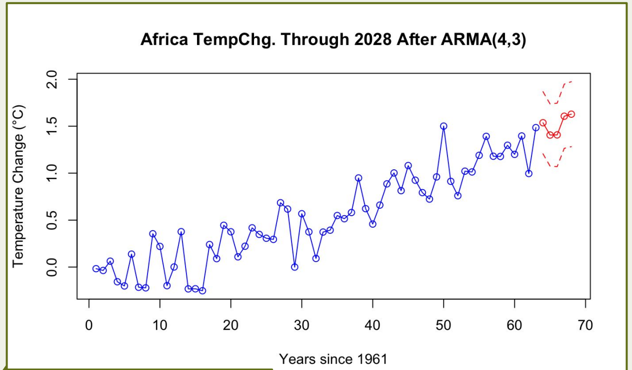
Step	Prediction	sqrt(MSE)	Lower Bound	Upper Bound
1	1.880086	0.1497866	1.586509	2.173662
2	1.787547	0.14985	1.493846	2.081247
3	1.702291	0.1535982	1.401244	2.003338
4	1.967495	0.1536404	1.666365	2.268625
5	1.955082	0.157521	1.646347	2.263818

Model Coefficients

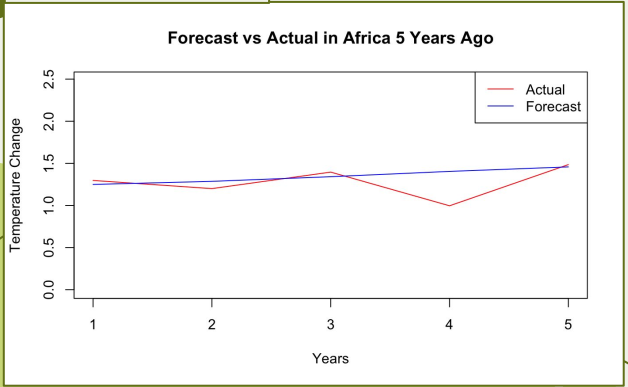
```
$phi
0.3149807 -0.9808362  0.2205324 -0.5331324 -0.2420449

$theta
-0.3440749  1.2151546 -0.3440217  0.9999483

$sigma2
0.02243603

$aicc
-27.55233

$se.phi
0.1488220 0.1326408 0.1860778 0.1343397 0.1440557

$se.theta
0.1066784 0.1249735 0.1410481 0.1769436
```

Geographical Region TS Models

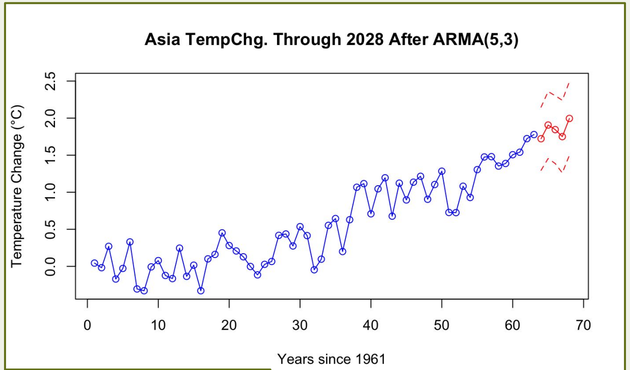
Africa



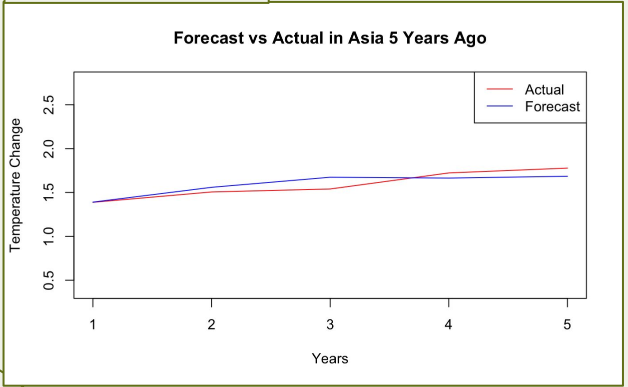
RMSE: 0.190



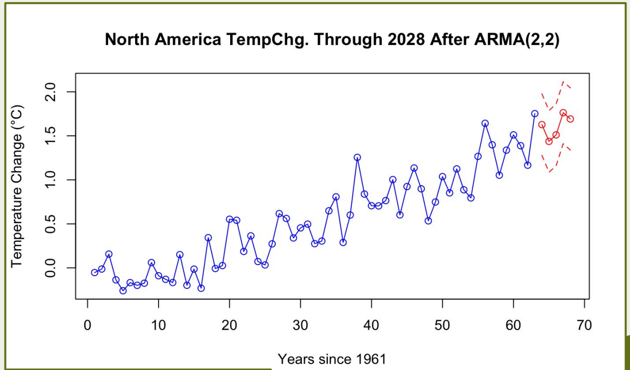
Asia



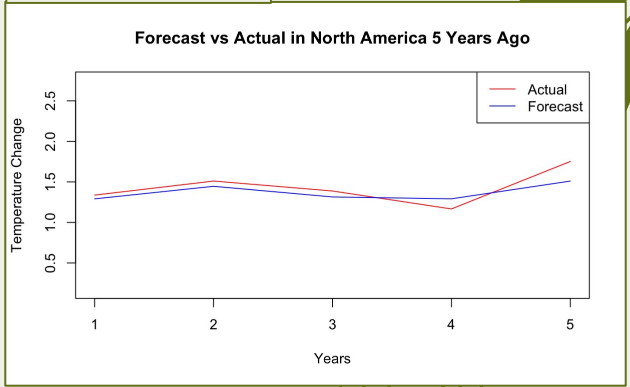
RMSE: 0.081



North America

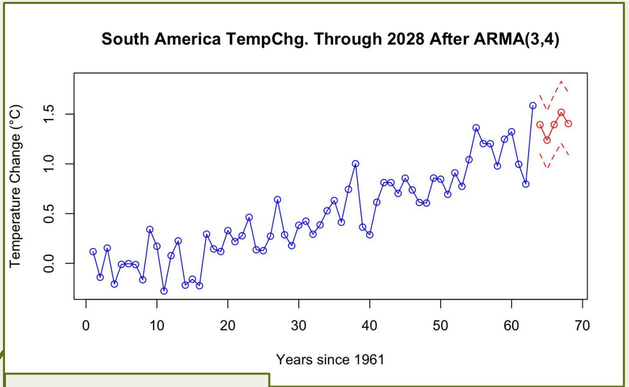


RMSE: 0.131

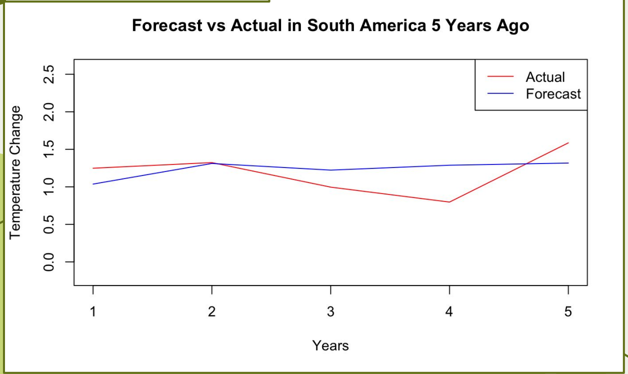


Geographical Region TS Models

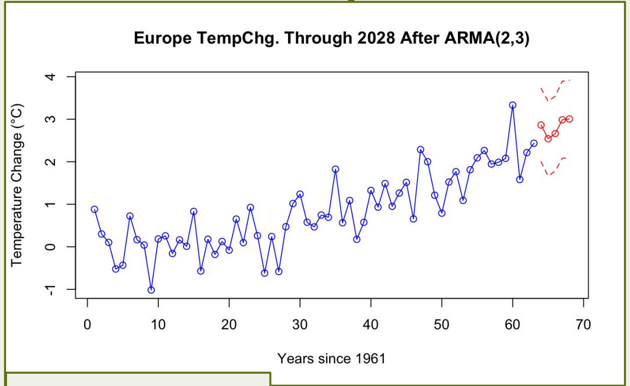
South America



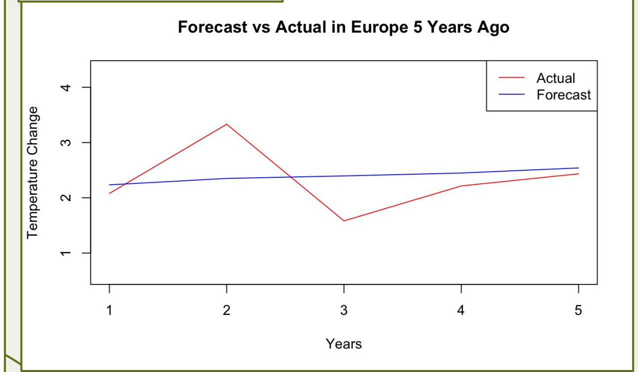
RMSE: 0.286



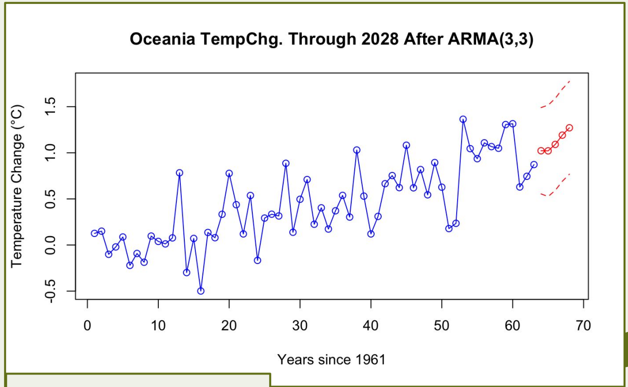
Europe



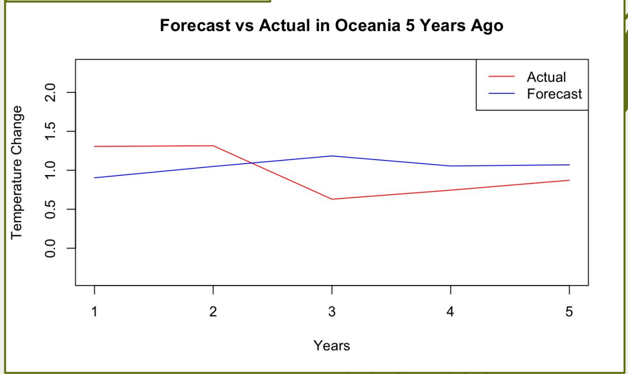
RMSE: 0.585



Oceania

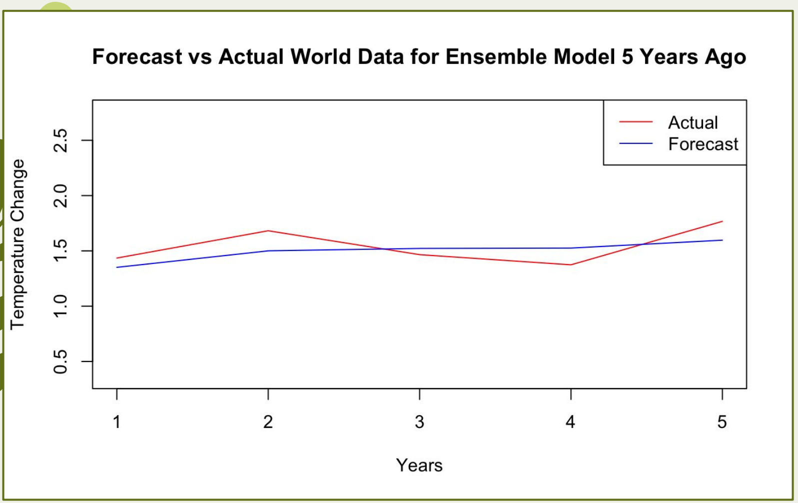


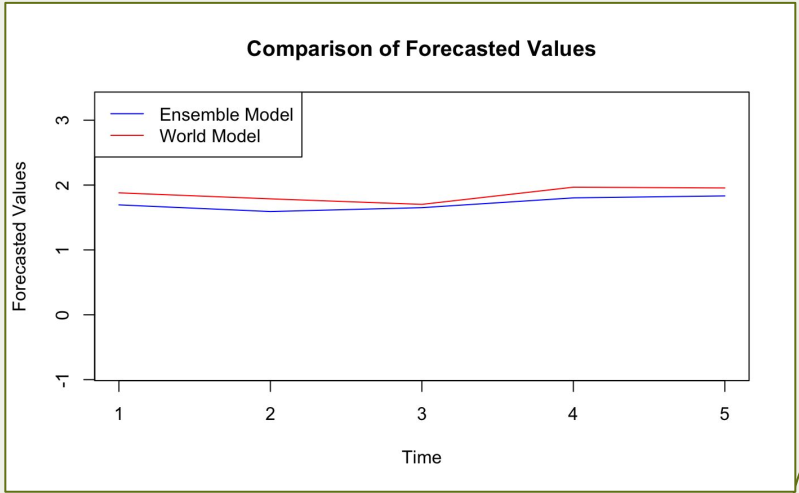
RMSE: 0.368



Ensemble Model of Regions

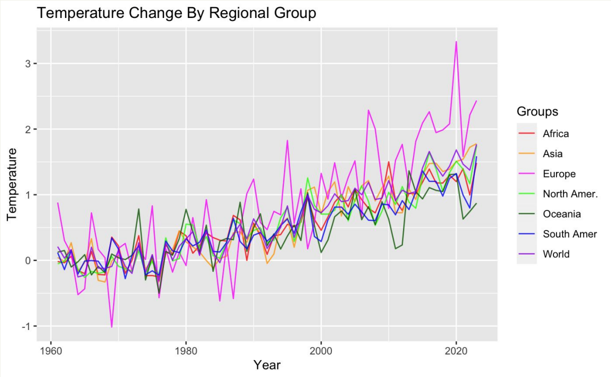
- Calculated by weighing each of the 6 regional models the same and combining them
- RMSE found by comparing this model's results to the worldwide data

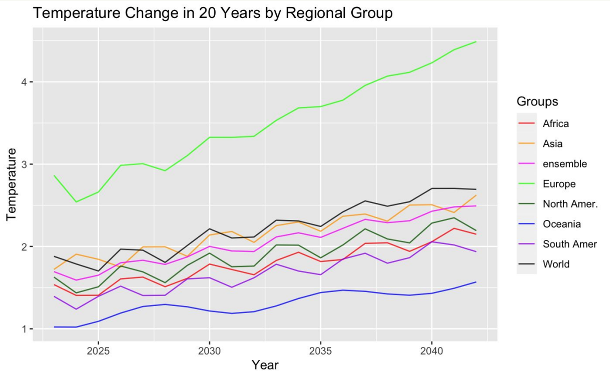


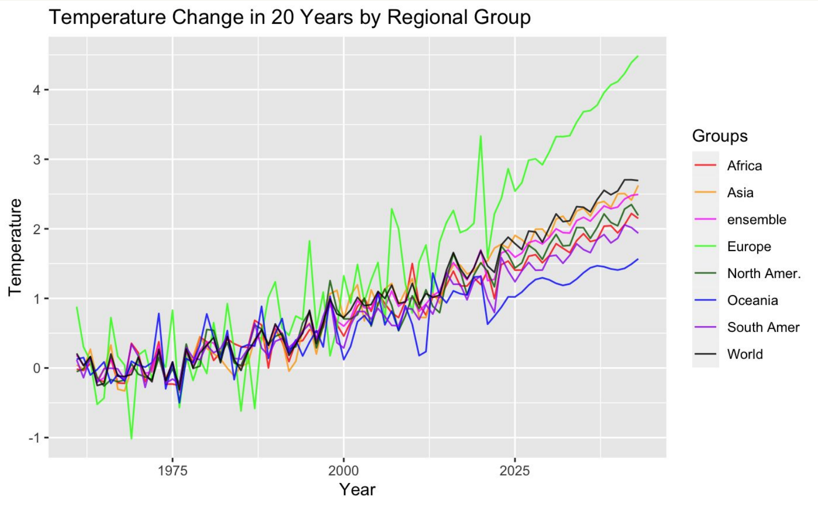


Mean Absolute Error (MAE): 0.1286367
Root Mean Squared Error (RMSE): 0.1378332

Geographic Models vs. World

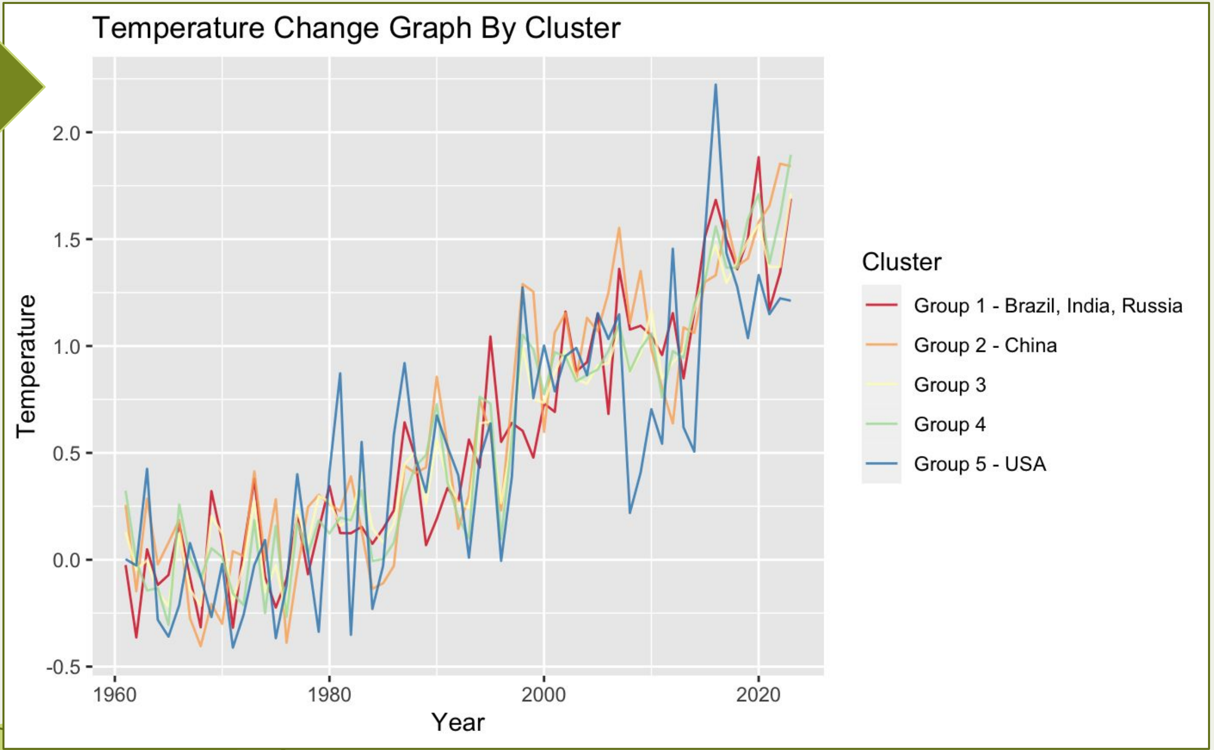






Time Series Models for Clustered Data

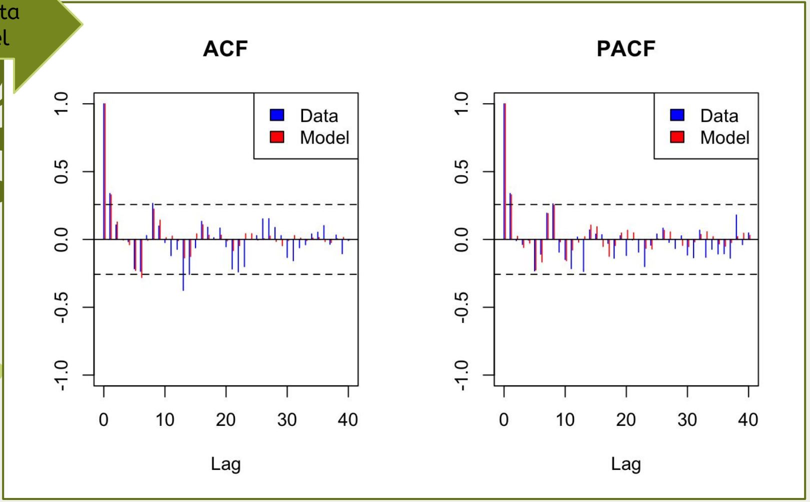
All Clusters data...



TS Models by Cluster

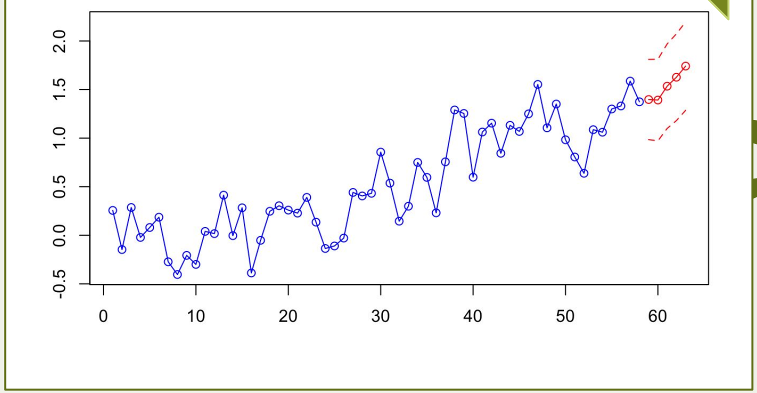
TS model for Group 2 – China

ACF in Data vs. Model



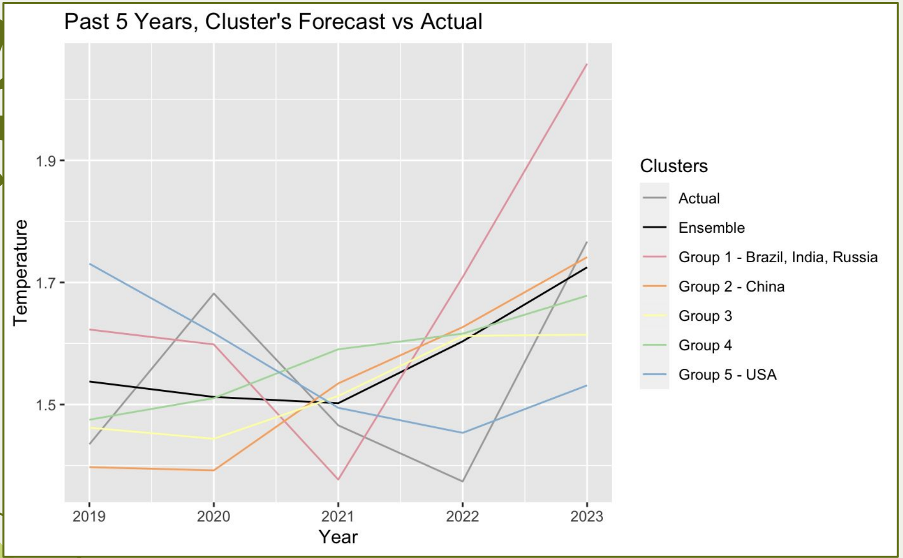
Eliminated Quadratic Trend
Arma(p=5,q=4) model

Forecasting 5 years



Clustered Models - Ensemble

Ensemble Predictions



Group 1 - Brazil, India, Russia RMSE
0.222378

Group 2 - China RMSE
0.1759161

Group 3 RMSE
0.1672246

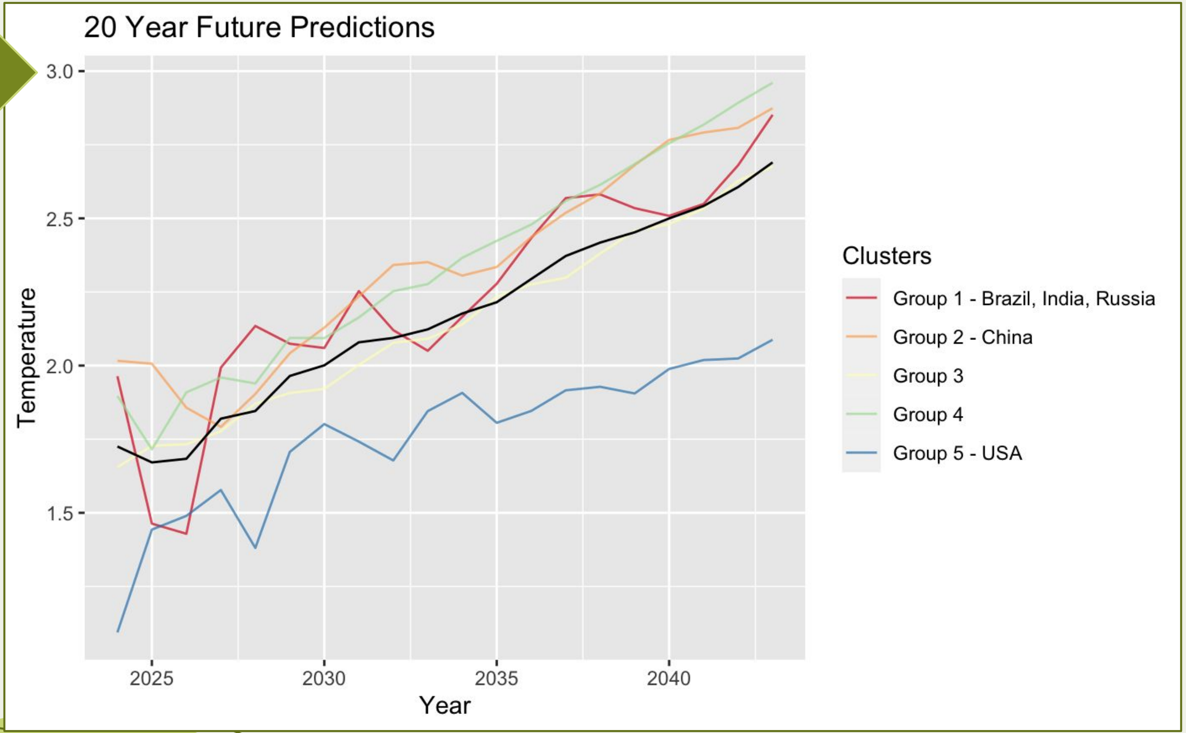
Group 4
0.1503529

Group 5 - USA RMSE
0.175717

Ensemble RMSE
0.1379182

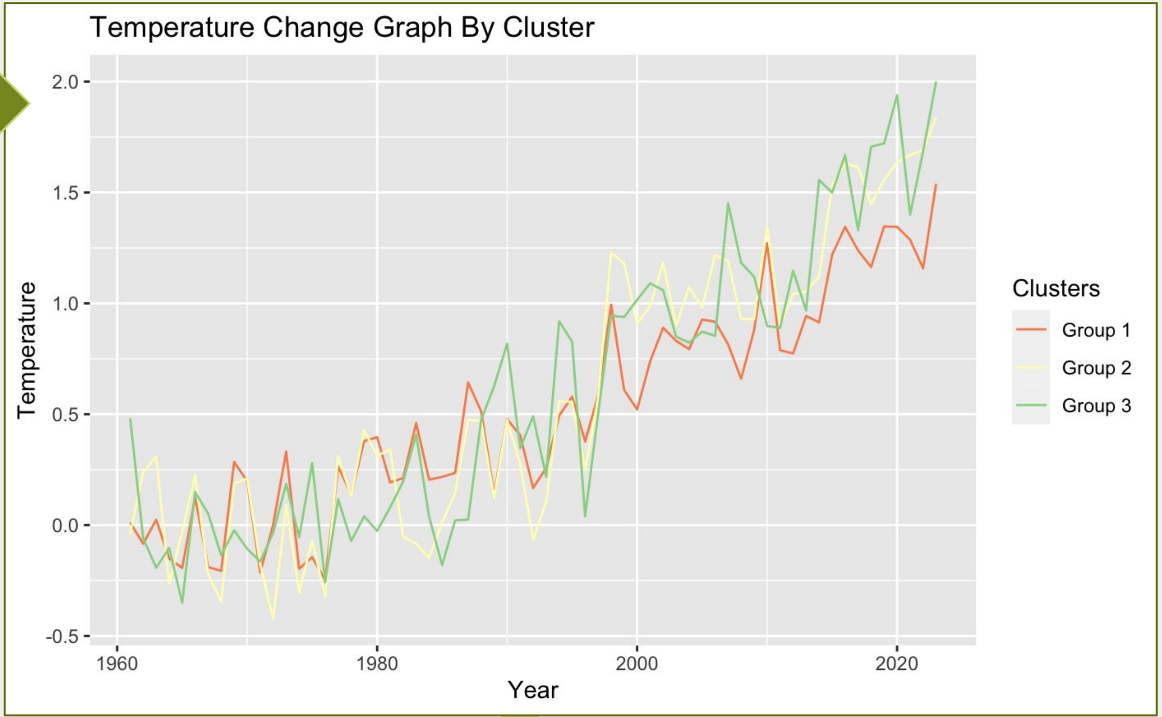
Clustered Models – Future Predictions

Black Line = Ensemble

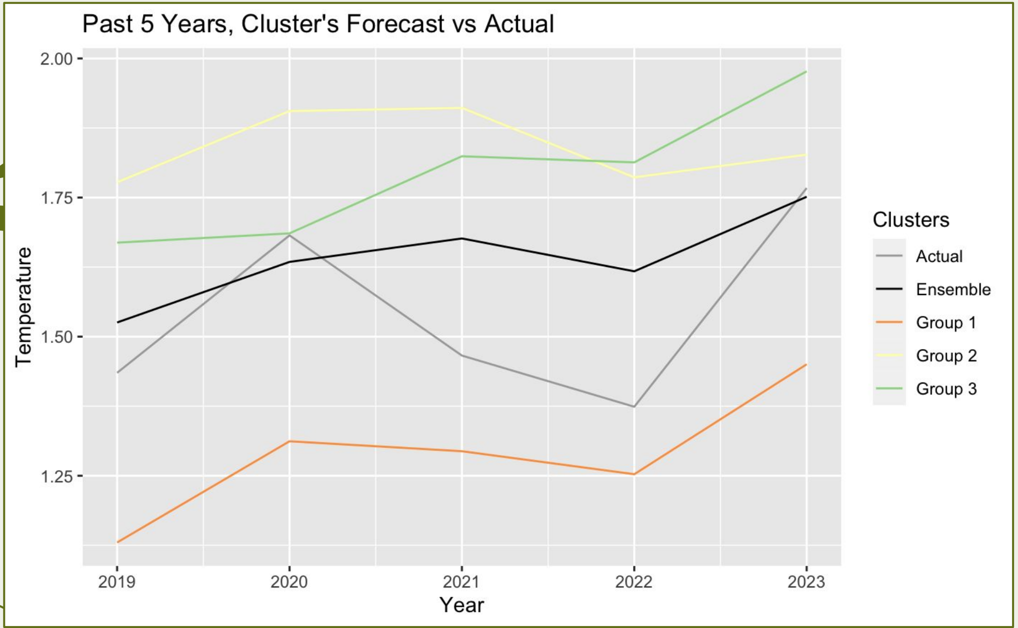


TS Models for Hierarchical Clustered Data

All Clusters data...



Hierarchical TS Models - Ensemble



Group 1 RMSE
0.2736701

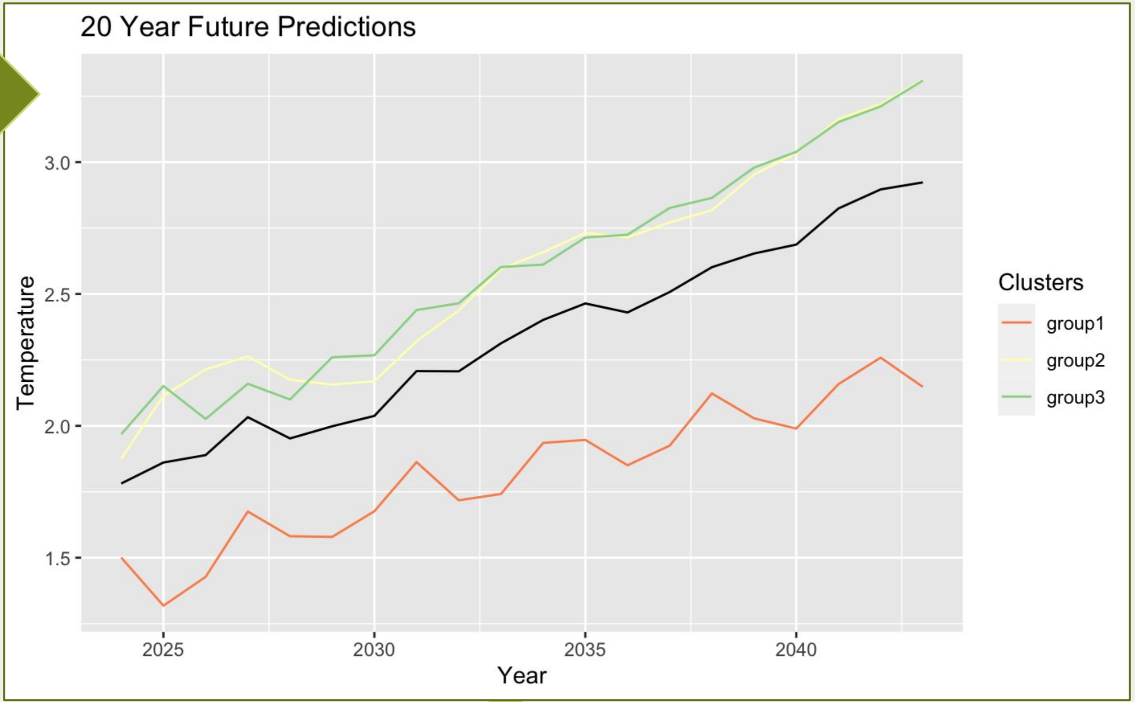
Group 2 RMSE
0.3283115

Group 3 RMSE
0.2897866

Ensemble RMSE
0.1511616

Hierarchical TS Models - Future Predictions

Black Line = Ensemble

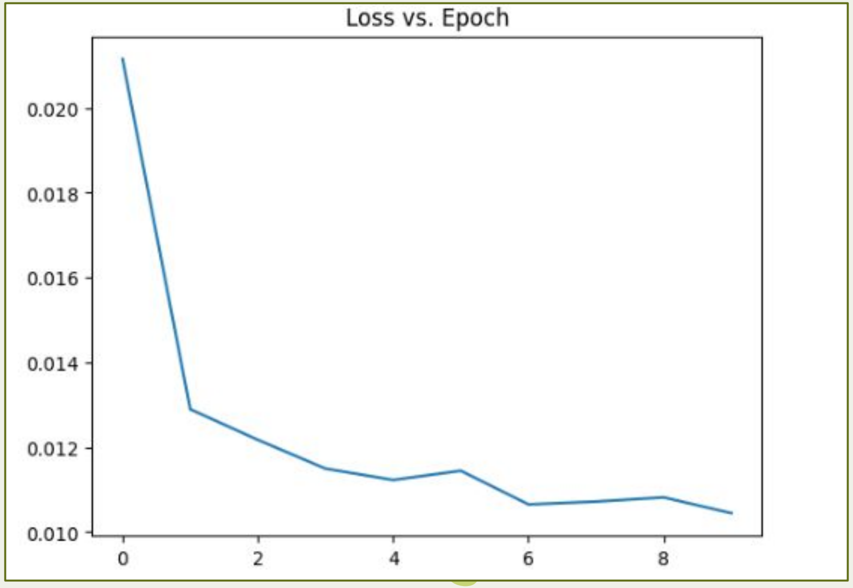


Recurrent Neural Networks (RNN)

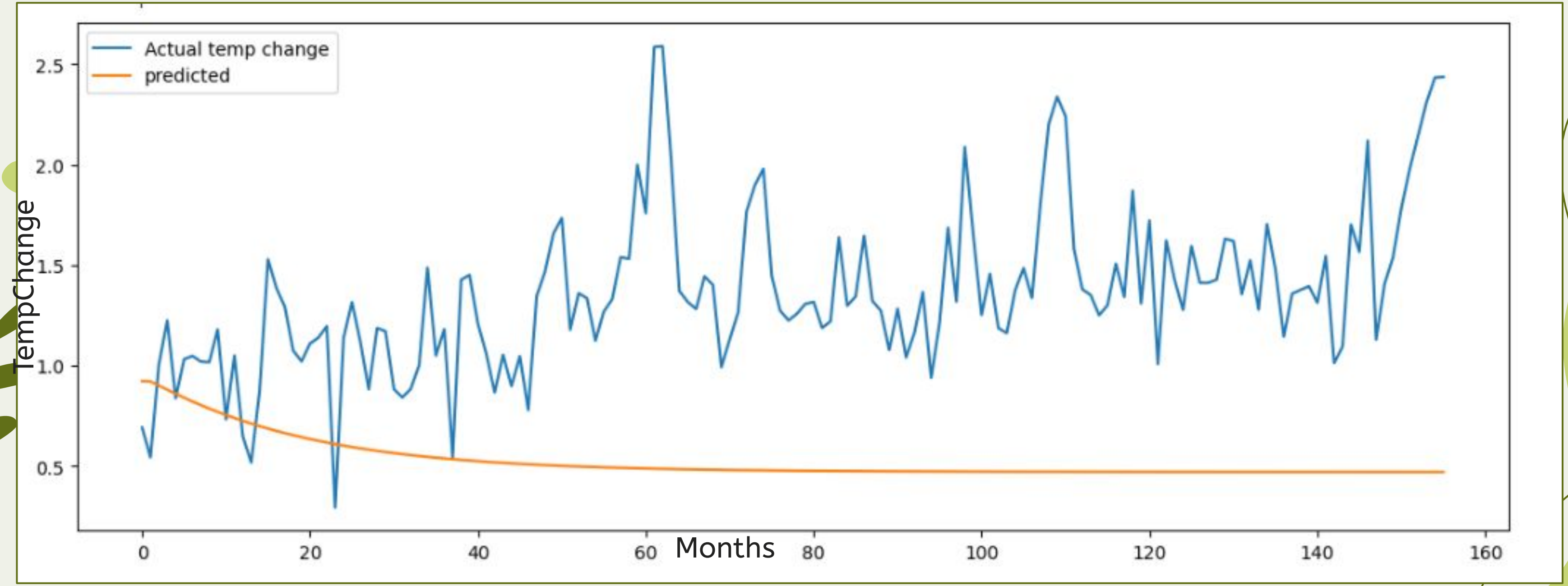
- Typically made of several types of layers
  - Long Short-Term Memory (LSTM): Used to capture long-term dependencies in sequential data
  - Dropout: Randomly selects neurons for deletion to prevent overfitting
  - Bidirectional: Analyzes sequential data forwards and backwards to detect more patterns
  - Dense: Fully connected layers used to detect complex patterns
- Useful for detecting overall trends in noisy data and predicting locations of future spikes
- Can be arbitrarily complexified to capture more attributes of the data

Naive Attempt

- Used just one LSTM layer and one Dense layer
- RELU activation function for nonlinearity
- 10 epochs



Naive Result

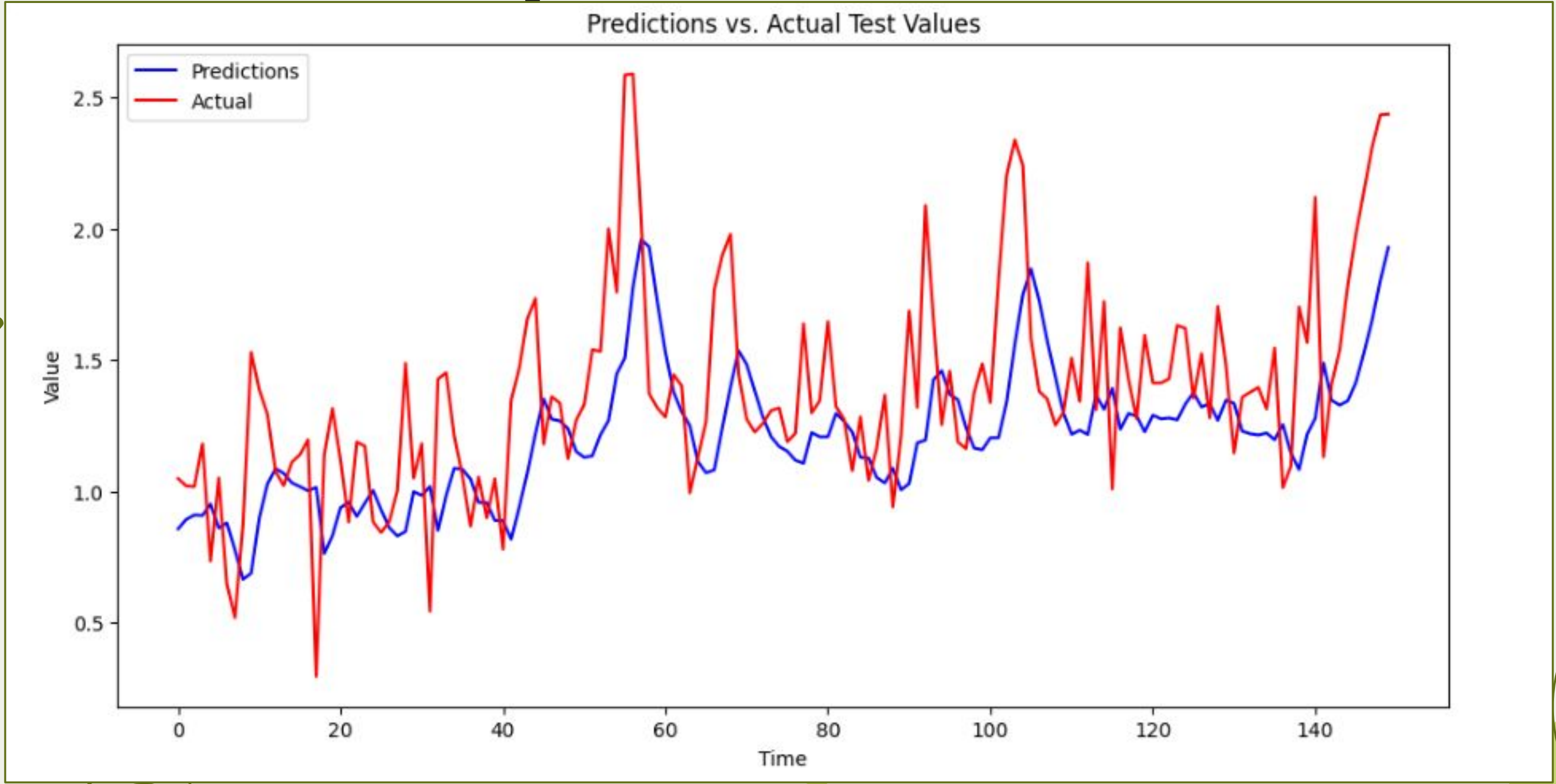


Improved Model

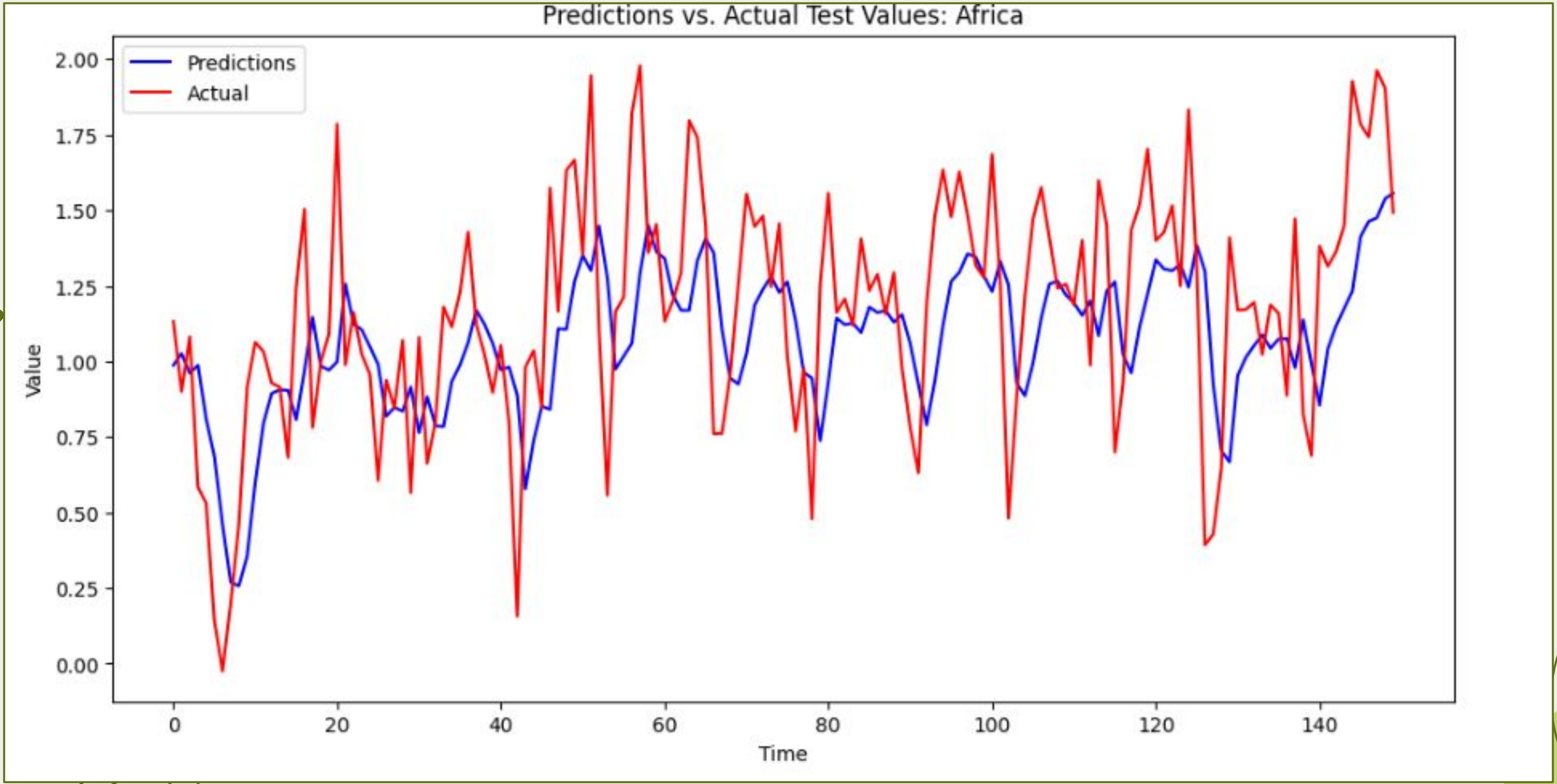
- Bidirectional
- LSTM
- Dropout
- LSTM
- Dropout
- LSTM
- Dropout
- Dense
- Dense (out)

- Many more layers
- Incorporates bidirectional layer
- Utilizes dropout
- Smaller sample size
- Different loss function

Improved Results



Africa



Predictions vs. Actual Test Values: Africa
Predictions
Actual
Value
Time
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
100
120
140

RNN Conclusions

- Numeric metrics do not tell the whole story
- RNN model is suitable for general, long-term trend prediction
- Model is not suitable for month-by-month predictions
- Predicts oscillations and placements of steep jumps and spikes
  - However, it often cannot predict their amplitudes
- Clustered predictions
  - Groupings based on geography and ecological data can lead to better results, but must be used with caution

Comparison of Models- Predicting Next Year

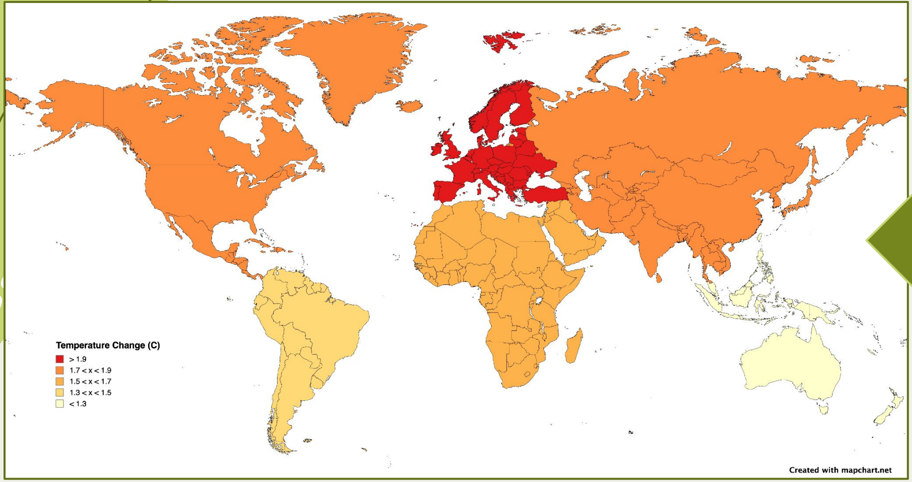
		ARMA(p,q)		Ensemble		RNN	
		Pred.	RMSE	Pred.	RMSE	Pred.	RMSE
World		1.880	0.097			1.378	0.393
Geographical Region	Africa	1.537	0.190			1.732	0.354
	Asia	1.722	0.081			1.230	0.705
	Europe	2.864	0.585			0.348	1.727
	North Amer.	1.628	0.131			0.994	0.539
	South Amer.	1.394	0.286			1.406	0.388
	Oceania	1.022	0.368			0.666	0.709
	Total			1.694	0.138	1.108	
K Means				1.725	0.138	1.016	
Hierarchical				1.781	0.151	1.237	

Terminology

Pred. -> Forecasting for model one year into future in °C

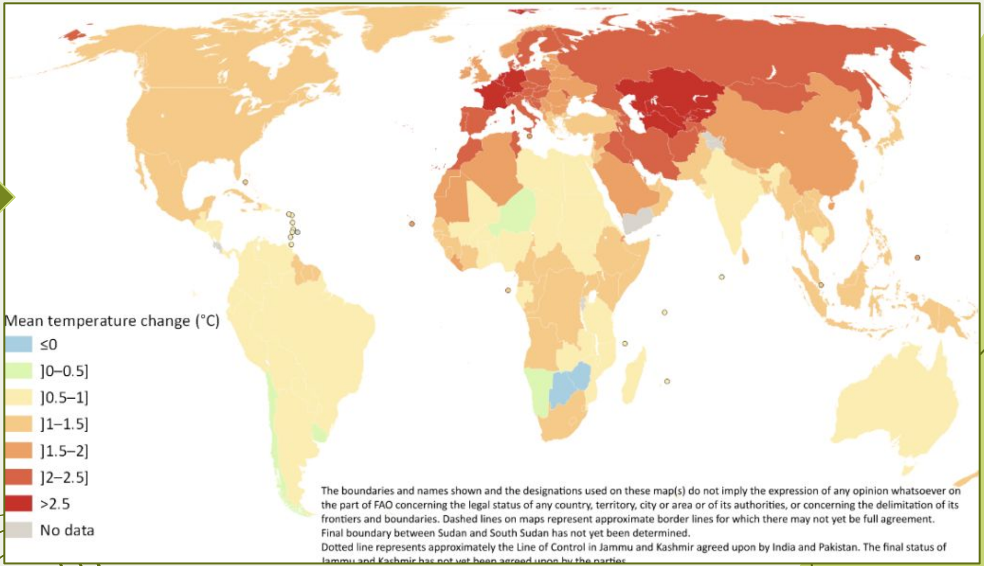
RMSE -> The testing RMSE for the model

Total -> Ensemble Model that combines Geographical Models



Our estimate by region

FAOSTAT 2022 Results



Conclusions

- Splitting the data by geographical regions gave interesting insight into the temperature change we might expect across different parts of the world; splitting the data into groups based on clustering and hierarchies gave insight into which countries share similar footprint data and thus might exhibit similar temperature change results
- The time series models are effective at predicting short term future temperature change forecasts, most of them performing with very confident data that aligns with the FAOSTAT data
- The ensemble models tend to perform better than their components but not necessarily better than the world model
- RNN (LSTM specifically) performed worse than the time series models with one feature available in the data

Future of this Project

- Explore how we can better fit the monthly data to time series models since this could provide more accurate results into forecasting
- Explore how adding more features to the data can improve upon LSTM models seeing that it has been scientifically concluded in the past that ARIMA models perform better with one feature than LSTM
- Create individual models for each country and then ensemble models of them in groups
- Thoroughly compare our results to the results of other climatologists who are reporting on temperature change

THANKS!

Do you have any questions?