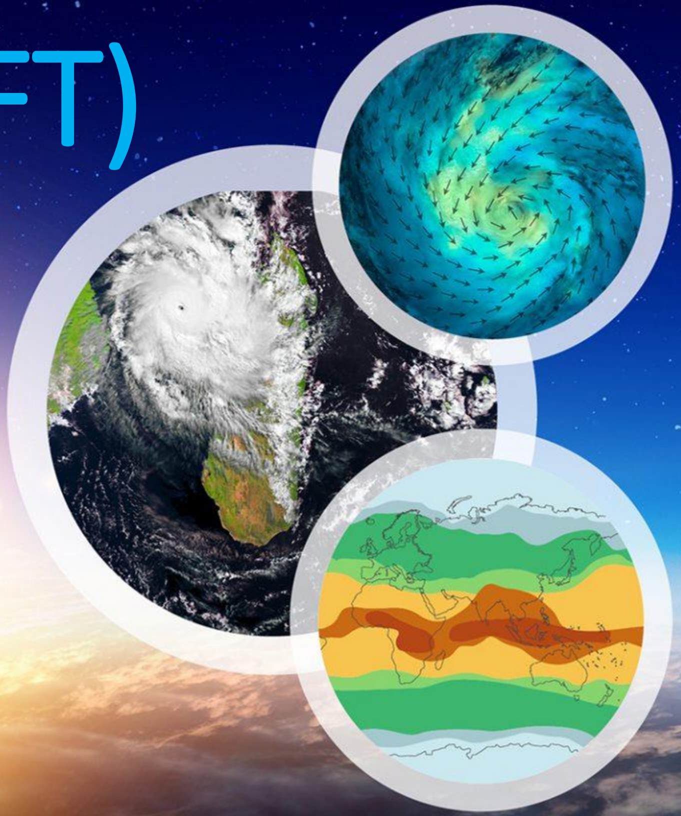
ClimSA
INTRA-ACP CLIMATE SERVICES AND RELATED APPLICATIONS PROGRAMME

Une initiative de l'Organisation des Etats d'Afrique, des Caraïbes et du Pacifique financée par l'Union européenne

STEP 4 of 9

Climate Forecasting Toolbox (CFT)

Zoning, Forecasting and Verification



Thembani MOITLHOBOGI (Independent CFT Developer)
And
Dr. Pierre KAMSU TAMO (Senior Forecaster - ACMAD)
Sunshine M. GAMEDZE (CFT Climate & Climate Expert - ACMAD)
Hubert KABENGELA NYAMABU (Climate Expert - ACMAD)

PRESAGG TRAINING WORKSHOP

14 to 20 February, Online

Introduction

The Climate Forecasting Toolbox (CFT) is a fork from the former SADC Climate Forecasting Tool, and has been further developed with new features and bug fixes.

The major version of the CFT follows the major Python version used (currently v3).

The latest version of CFT is available from Github where the source code is maintained:

https://github.com/taxmanyana/cft/tags

Features (2 / 2)

The CFT v3.1.0 has the following main features:

- ➢ Create homogenous zones for a country/region
- ➢ Forecast based on existing indices (CSV/Text) data, or can detect high correlation areas from gridded data (NetCDF) to use as input
- ➢ Multiple Linear Regression and Artificial Intelligence (MLP) statistical forecasting methods
- ➢ Predictand in NetCDF and CSV format
- ➢ Forecast verification using predictand data in NetCDF format

Features (2 / 2)

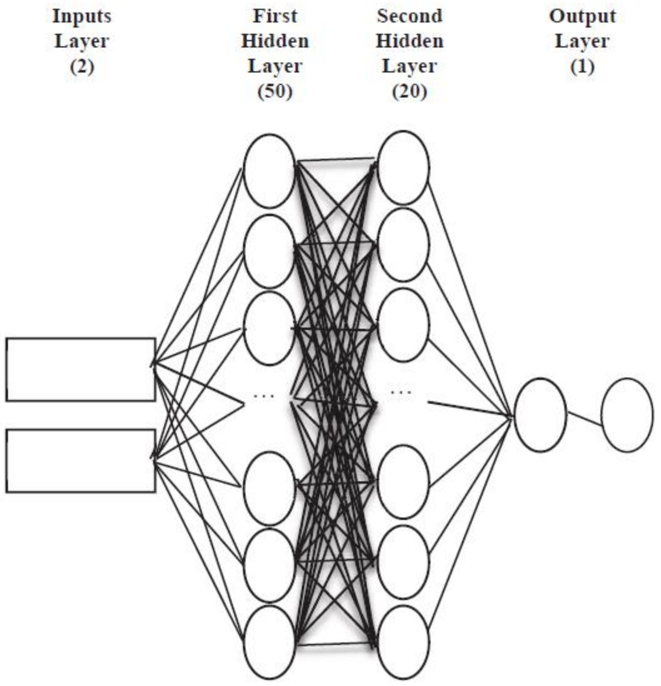
The CFT, through class **MLPRegressor** (from sklearn.neural_network Python module), has a component of Artificial Intelligence in the form of a multi-layer perceptron (MLP) that is a feedforward artificial neural network with backpropagation.

The advantages of MLP are:

- Capability to learn non-linear models
- Capability to learn models in real-time (on-line learning)

The disadvantages of MLP include:

- MLP with hidden layers have a non-convex loss function where there exists more than one local minimum. Therefore different random weight initializations can lead to different validation accuracy
- MLP requires tuning a number of hyper-parameters such as the number of hidden neurons, layers, and iterations
- MLP is sensitive to feature scaling



MLP of **2-50-20-1** architecture

CFT uses the hyperbolic tan function ‘tanh’ as the activation function, ‘lbfgs’ (an optimizer in the family of quasi-Newton methods) as the solver, and maximum iterations set at 700.

Installation (1 / 4)

The installation instructions for different platforms are provided in the **README.md** file which is part of the installation package.

For Windows, CFT can make use of QGIS v3.16 installed in the system to satisfy the required dependencies.

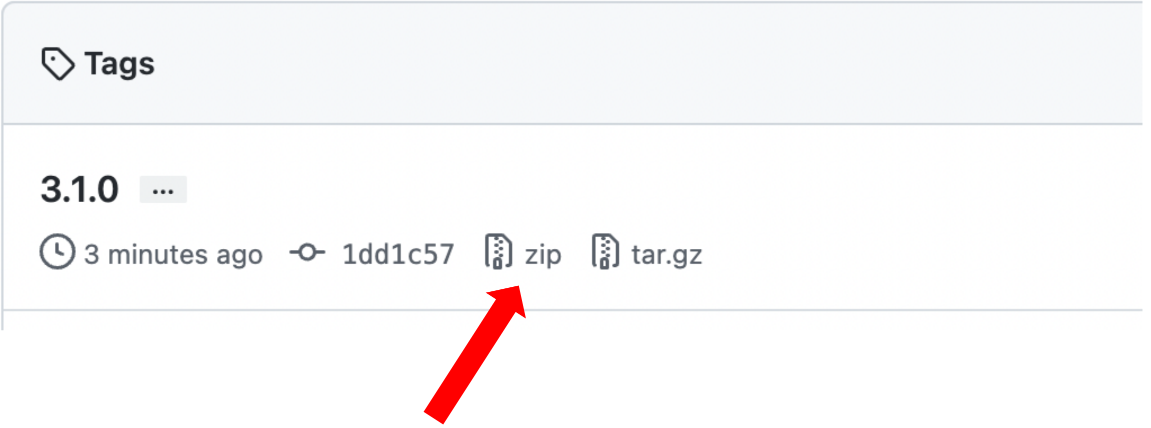
Follow the following steps to install CFT…

Installation (2 / 4)

The installation instructions

Download the latest release from the link below:

https://github.com/taxmanyana/cft/tags



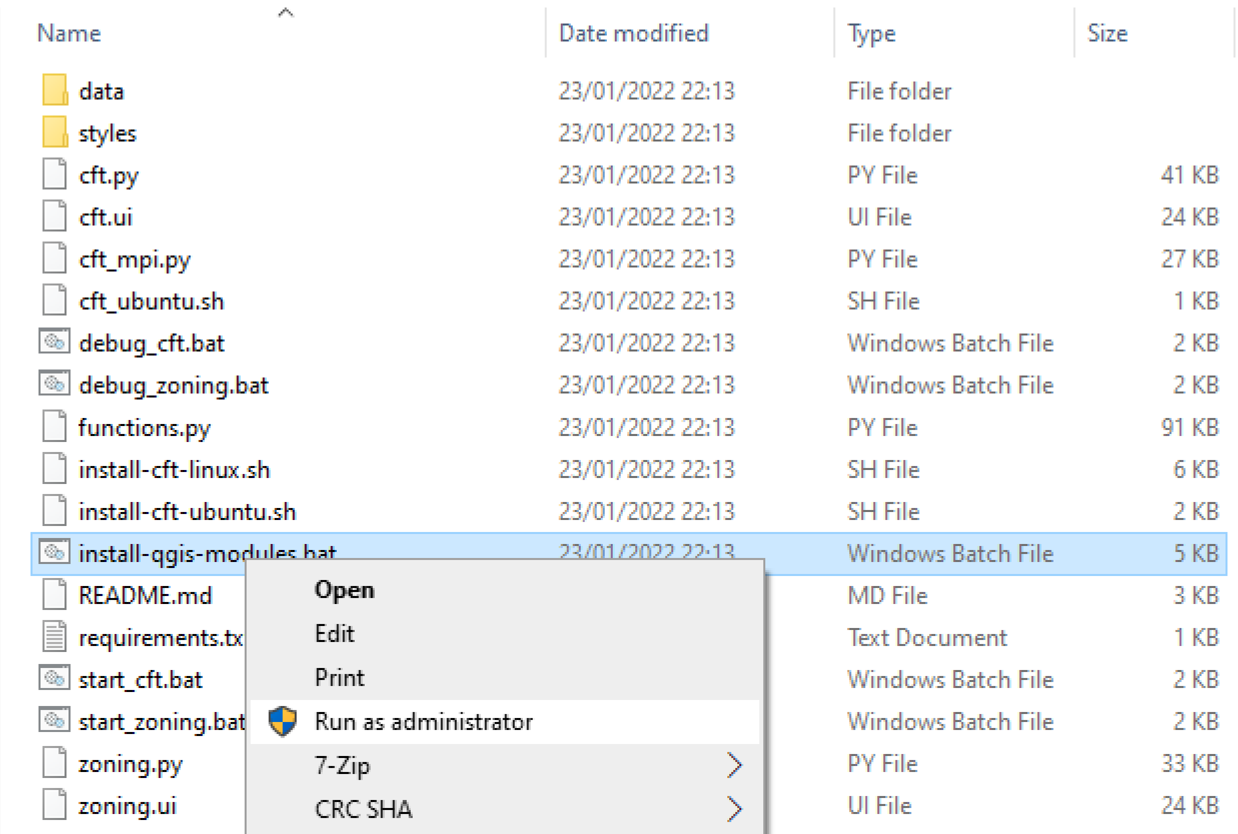
Click on the zip icon to download a zip file

Copy the downloaded zip file (e.g. cft-3.1.0.zip) to an installation directory, and unzip the file.

Installation (3 / 4)

The installation instructions

To update the required Python modules on QGIS, right-click on **install-qgis-modules.bat** and select **Run as administrator**



The script will download and install the needed modules. When completed the terminal will return the message “Press any key to close”

Installation (4 / 4)

Contents of the CFT installation folder:

- data
- icon
- styles
- CFT
- cft
- cft.ui
- cft_mpi
- cft_ubuntu.sh
- debug_cft
- debug_zoning
- functions
- install-cft-linux.sh
- install-cft-ubuntu.sh
- install-qgis-modules
- qgis_env
- README
- requirements
- start.ui
- start_cft
- start_verification
- start_zoning
- startup
- startup
- verification
- verification.ui
- zoning
- zoning.ui

data – contains test data

Icon– contains the desktop icon

styles – contains QGIS color schemes for the CFT outputs

CFT – desktop launcher shortcut (appears only after running the installation script)

qgis_env.bat – environmental variables needed to run the tool (appears only after running the installation script)

install-cft-linux.sh – install CFT in generic Linux environment

install-cft-ubuntu.sh – install CFT in Ubuntu OS

install-qgis-modules.bat – install required Python modules to QGIS

README.md – installation instructions

start_cft.bat – run forecasting tool with debugger terminal

start_zoning.bat – run forecasting tool with debugger terminal

start_verification.bat – run verification tool with debugger terminal

Data Requirements

Predictors

The CFT accepts predictor data in NetCDF and CSV/Text formats. The **data must be in monthly time steps**, and should be one parameter per file.

The preferred source of predictor NetCDF data is the IRI Data Library, some links provided below:

https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.ERSST/.version5/.sst/#expert
SOURCES .NOAA .NCDC .ERSST .version5 .sst
T (Dec 1960-2021) RANGEEDGES
T 12 STEP

https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/.Intrinsic/#expert
SOURCES .NOAA .NCEP-NCAR .CDAS-1 .MONTHLY .Intrinsic .PressureLevel .phi
P (850) VALUES
T (Dec 1960-2021) RANGEEDGES
T 12 STEP

https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/Intrinsic/.MSL/.pressure/#expert
SOURCES .NOAA .NCEP-NCAR .CDAS-1 .MONTHLY
Intrinsic .MSL .pressure
T (Dec 1960-2021) RANGEEDGES
T 12 STEP

Data Requirements

Predictors

The CFT accepts predictor data in NetCDF and CSV/Text formats. The **data must be in monthly time steps**, and should be one parameter per file.

CSV and Text data must be in the format below: The Tokyo Climate Centre and other sources provide indices in the required format:

Text data (.txt extension)

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1950	-0.44	-0.92	-0.47	0.25	-0.03	-1.01	-1.18	-1.14	-0.91	-0.62	-0.50	-0.45
1951	-0.75	-0.01	0.16	-0.04	0.24	-0.93	-1.26	-1.23	-0.40	-0.36	0.07	-0.37
1952	-0.63	-0.34	-0.08	0.56	0.03	-0.82	-1.63	-1.45	-1.01	-0.82	-0.22	-0.26
1953	-0.33	-0.37	0.08	0.06	0.79	-0.57	-1.11	-0.89	-0.38	-0.36	-0.02	-0.30
1954	-1.09	-0.51	-0.18	0.67	0.32	-0.91	-1.67	-1.23	-1.28	-1.06	-0.52	-0.44
1955	-0.97	-0.73	-0.38	0.32	-0.43	-1.14	-1.33	-0.95	-1.23	-0.98	-0.19	-0.11
1956	-0.41	-0.30	-0.39	-0.07	-0.34	-1.18	-1.78	-1.26	-1.15	-0.89	-0.71	-0.57
1957	-1.11	-0.88	-0.28	0.28	0.10	-1.16	-1.55	-1.32	-0.72	-0.25	-0.88	-0.66
1958	-1.04	-0.71	-0.34	-0.23	-0.12	-1.42	-2.05	-1.57	-1.46	-0.98	-0.53	-0.34
1959	-0.16	-0.63	-0.56	0.02	-0.69	-1.52	-1.76	-1.63	-0.97	-0.80	-0.72	-0.90
1960	-0.97	-0.31	-0.42	-0.35	-0.73	-1.19	-1.64	-1.66	-0.95	-1.14	-1.12	-1.04
1961	-0.74	-0.53	0.19	-0.02	-0.34	-0.51	-0.32	-0.16	0.30	1.18	0.35	-0.12
1962	-0.05	0.09	0.12	0.57	-0.23	-0.94	-1.14	-0.97	-0.15	-0.44	-0.34	-0.40

CSV data (.csv extension)

```
Year,JAN,FEB,MAR,APR,MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC
1950,-0.44,-0.92,-0.47,,0.25,-0.03,-1.01,-1.18,-1.14,-0.91,-0.62,-0.50,-0.45
1951,-0.75,-0.01,,0.16,-0.04,,0.24,-0.93,-1.26,-1.23,-0.40,-0.36,,0.07,-0.37
1952,-0.63,-0.34,-0.08,,0.56,,0.03,-0.82,-1.63,-1.45,-1.01,-0.82,-0.22,-0.26
1953,-0.33,-0.37,,0.08,,0.06,,0.79,-0.57,-1.11,-0.89,-0.38,-0.36,-0.02,-0.30
1954,-1.09,-0.51,-0.18,,0.67,,0.32,-0.91,-1.67,-1.23,-1.28,-1.06,-0.52,-0.44
1955,-0.97,-0.73,-0.38,,0.32,-0.43,-1.14,-1.33,-0.95,-1.23,-0.98,-0.19,-0.11
1956,-0.41,-0.30,-0.39,-0.07,-0.34,-1.18,-1.78,-1.26,-1.15,-0.89,-0.71,-0.57
1957,-1.11,-0.88,-0.28,,0.28,,0.10,-1.16,-1.55,-1.32,-0.72,-0.25,-0.88,-0.66
1958,-1.04,-0.71,-0.34,-0.23,-0.12,-1.42,-2.05,-1.57,-1.46,-0.98,-0.53,-0.34
1959,-0.16,-0.63,-0.56,,0.02,-0.69,-1.52,-1.76,-1.63,-0.97,-0.80,-0.72,-0.90
1960,-0.97,-0.31,-0.42,-0.35,-0.73,-1.19,-1.64,-1.66,-0.95,-1.14,-1.12,-1.04
```

Data Requirements

Predictors

The CFT accepts predictor data in NetCDF and CSV/Text formats. The **data must be in monthly time steps**, and should be one parameter per file.

NB:

Predictor data should cover the year before the training period up to the most recent month.

For instance: if training period is 1971-2000 and predictor month being used is July to predict OND 2022, the predictor data should cover July 1970 to July 2022.

Data Requirements

Predictand

Predictand data must be NetCDF or CSV formats, and **<u>must be in monthly time steps.</u>**

NetCDF data should be single parameter (e.g. rainfall) in monthly time steps.

CSV data should contain the name of the point (station), its Latitude, Longitude, and monthly values for each year in the following format.

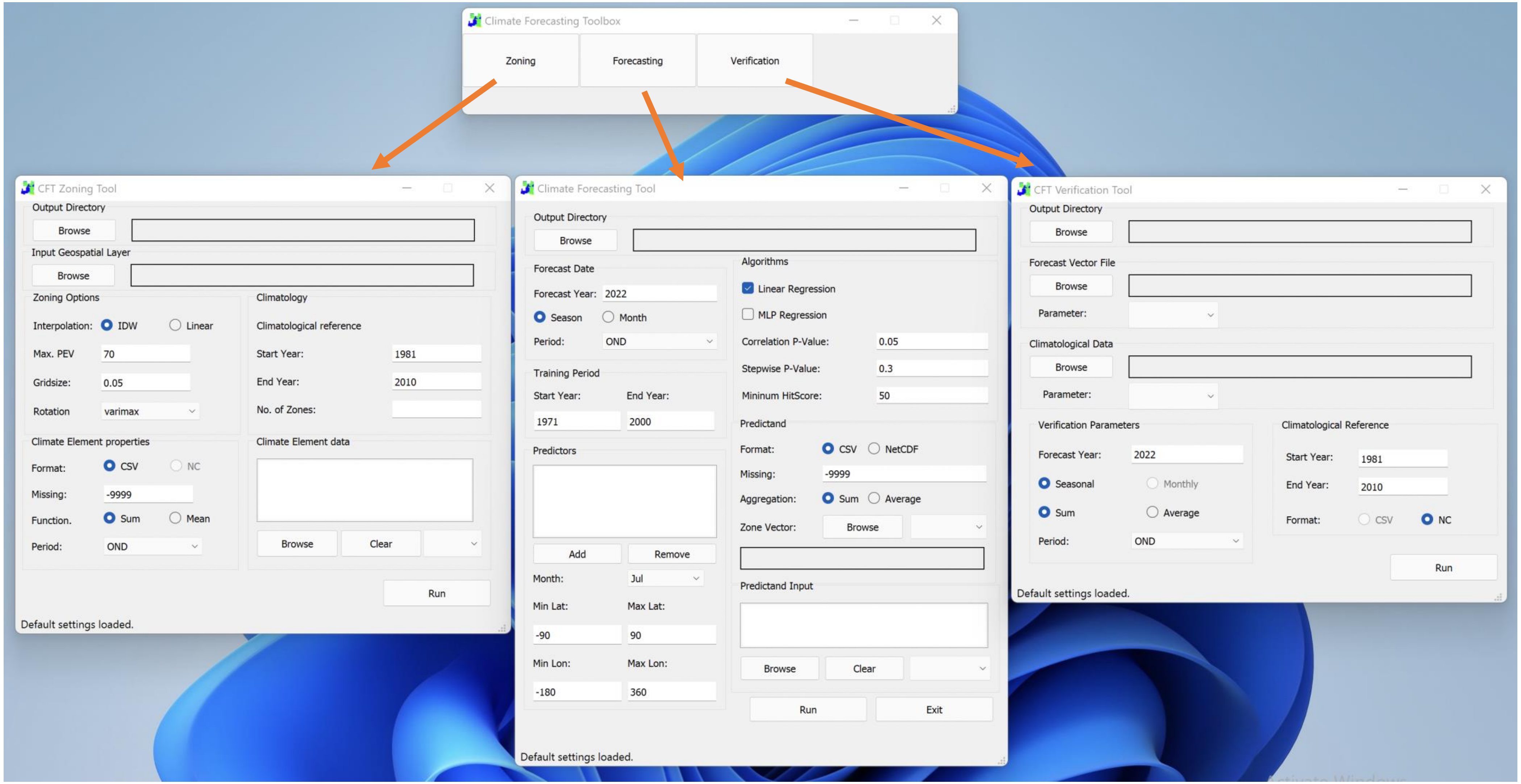
CSV data (.csv extension)

```
ID,Lat,Lon,Year,Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov,Dec
PEMBA,-12.983,40.533,1981,52.1,150.9,95.9,41.5,42.5,19.2,5.1,3.6,1.8,11.0,11.6,109.9
PEMBA,-12.983,40.533,1982,90.2,239.6,111.9,71.3,28.4,7.0,6.9,12.0,1.3,35.4,76.1,173.9
PEMBA,-12.983,40.533,1983,221.9,91.4,209.2,73.6,45.5,14.3,5.2,7.9,1.6,7.9,19.4,80.7
PEMBA,-12.983,40.533,1984,165.1,157.1,152.6,101.2,32.3,26.2,14.5,7.8,1.2,13.5,21.4,223.1
PEMBA,-12.983,40.533,1985,114.0,185.5,104.8,140.0,18.2,11.7,19.3,10.5,1.7,8.0,70.2,99.7
PEMBA,-12.983,40.533,1986,189.3,197.3,130.9,102.6,16.9,10.3,9.3,5.9,1.4,23.2,66.6,224.5
PEMBA,-12.983,40.533,1987,130.0,114.0,146.8,45.4,21.1,7.9,9.0,7.8,1.4,7.9,62.0,11.4
PEMBA,-12.983,40.533,1988,152.8,173.6,66.3,77.0,35.3,6.0,5.3,9.8,1.4,6.4,74.0,68.9
PEMBA,-12.983,40.533,1989,244.7,141.7,259.0,82.4,107.5,13.0,13.4,14.2,1.4,5.8,48.6,175.5
PEMBA,-12.983,40.533,1990,81.1,233.9,97.4,144.0,28.9,5.3,22.8,12.8,1.5,12.0,13.8,102.3
PEMBA,-12.983,40.533,1991,117.9,133.4,392.5,138.1,38.2,8.6,22.8,6.5,1.5,7.1,34.8,223.8
PEMBA,-12.983,40.533,1992,107.1,158.2,137.8,129.2,35.0,12.5,18.6,6.3,1.7,11.0,86.1,20.0
PEMBA,-12.983,40.533,1993,201.0,149.8,311.5,101.8,15.3,9.7,13.7,9.8,1.4,5.2,26.7,75.0
PEMBA,-12.983,40.533,1994,194.2,216.0,104.4,61.7,11.8,6.7,6.0,7.0,1.5,7.0,11.5,111.4
PEMBA,-12.983,40.533,1995,198.4,216.7,184.1,73.0,38.7,5.3,5.2,15.4,1.6,6.4,8.9,118.7
PEMBA,-12.983,40.533,1996,174.9,204.4,76.2,41.5,34.0,7.9,11.1,6.0,1.2,9.6,6.7,87.4
```

Loading CFT

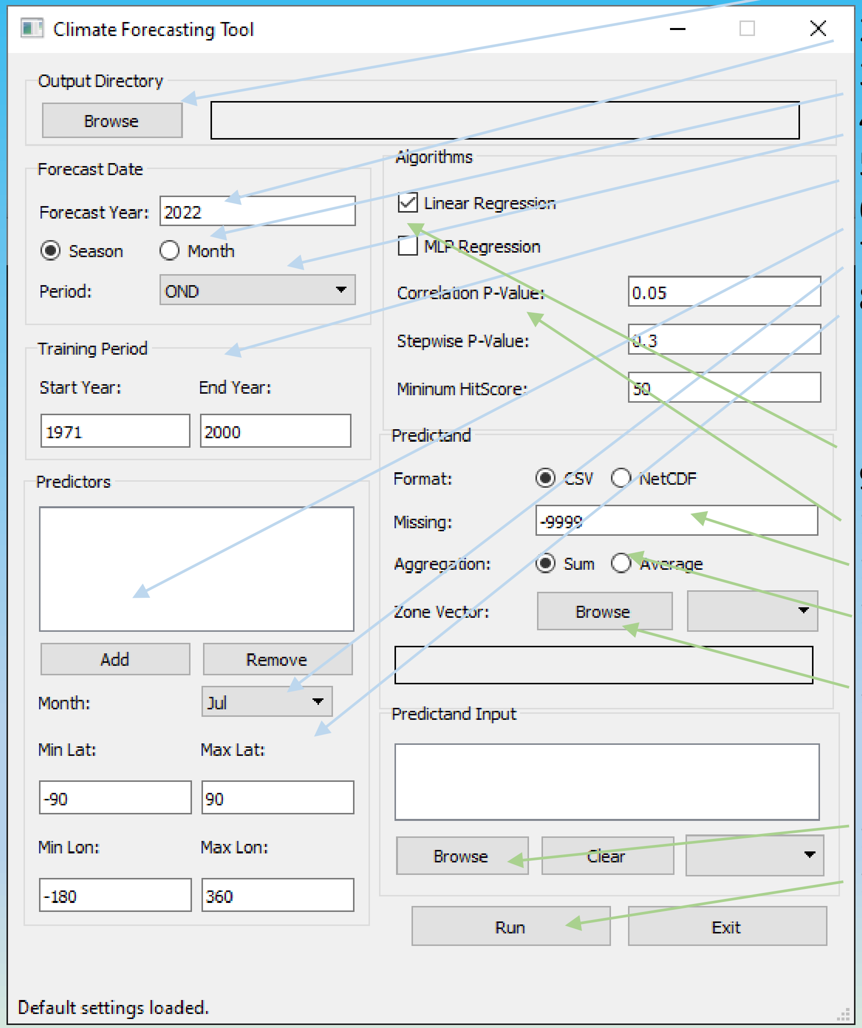
Double/click on the launcher shortcut CFT to open the launcher.

From the launcher, you can click on the button for each tool to open it



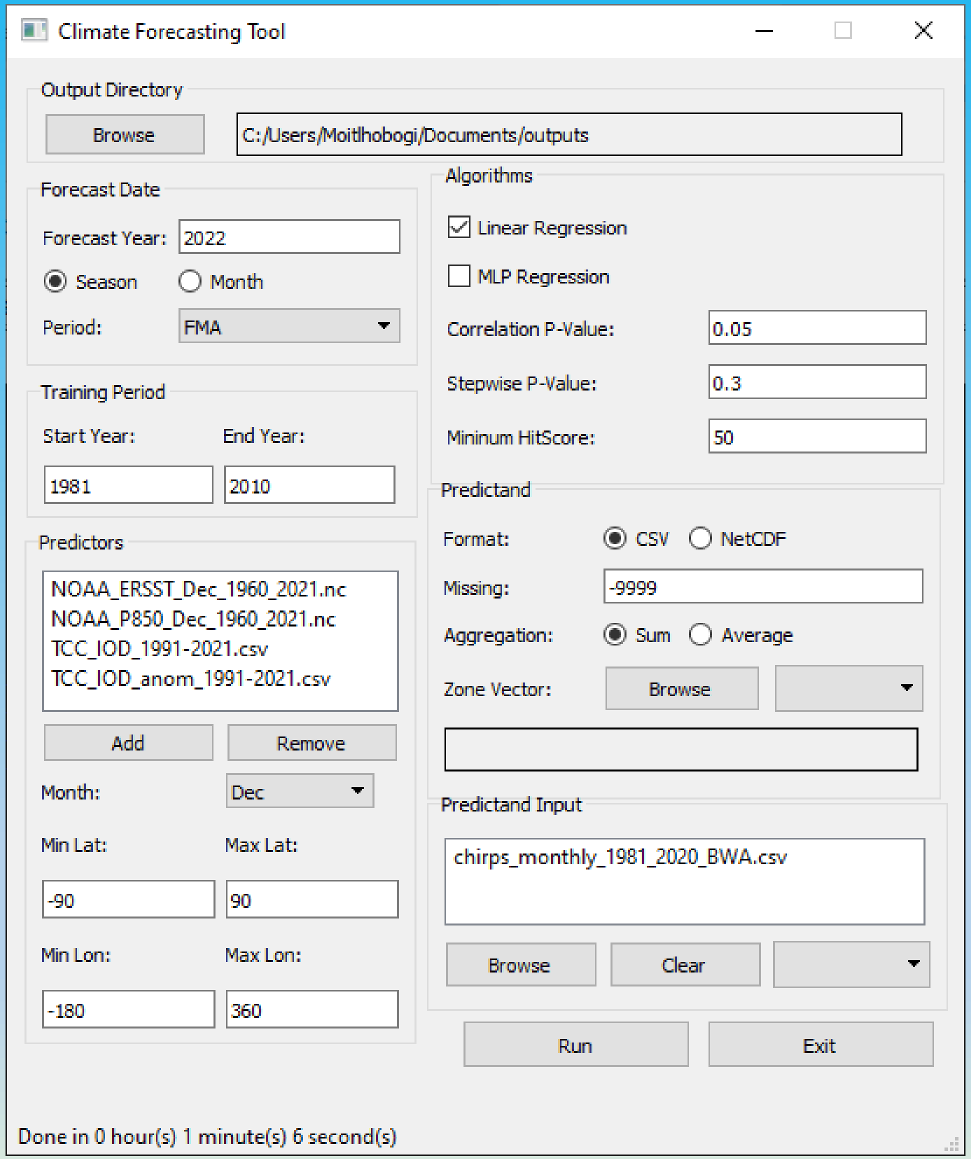
Forecasting

Run the forecasting tool from the launcher (or by double-clicking on the start-cft.bat script)



1. Select output directory where outputs will be delivered
2. Indicate year to be forecasted
3. Select forecast period type (Season or Month)
4. Select forecast period
5. Update the training period if necessary
6. Add NetCDF/CSV/Txt file(s) to be used as predictors
7. Select the month to be used from the predictor files
8. If you would like to focus your predictor on a specific area, you can indicate the extent (min and max Lat/Lon)
9. Select the regression algorithm(s) to use (Linear or MLP or both)
10. Correlation and Stepwise Selection P-Values can be maintained as default
11. Indicate the figure used as missing values in your predictant (station) data
12. Indicate the aggregation function to be used on the predictant, use Cumulation for rainfall and Average for temperature data
13. Browse to the Zone vector file (geojson format) to be used if you want forecast to be classified per zone, and select the attribute which will be used for naming the zones
14. Browse to the CSV files which will be used as predictant (station) data
15. Click on Run to generate the forecast

Forecasting



❖ Once the forecast is completed, a message will appear on the status box indicating "Done in xxxx seconds"

❖ The amount of time taken to generate a forecast depends mainly on 4 factors
  1. Number of predictors
  2. Number of stations
  3. Algorithms selected
  4. CPU count of your computer (CFT using 90% of your CPU)

❖ To estimate the runtime, Linear regression (LR) for one station and one predictor takes about 3 seconds. MLP regression (MLP) for one station and one predictor takes about 57 seconds.

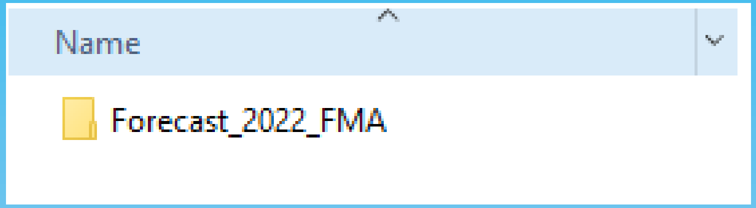
❖ For example: with two predictors to forecast 60 stations using both LR and MLP, it will require about 2 x 60 x (57 + 3) = 7 200s = **2hrs**

If you have an 8 core CPU then 7 cores will be used by CFT to reduce the time to about 7200 / 7 = 1028s = **17 minutes**

Forecasting

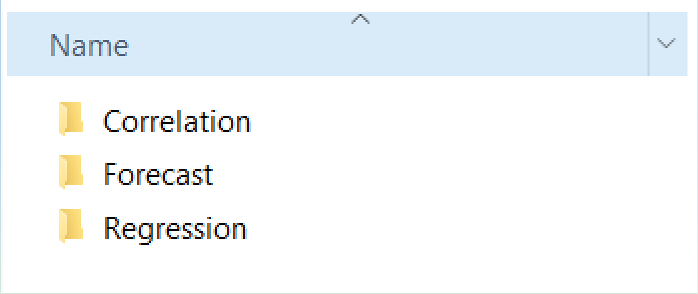
In the output directory a forecast output folder will be created in the following naming structure:

Forecast_<fcstYear>_<fcstPeriod>



In the forecast output directory, the following subfolders will be created:

- ❖ Correlation this is where correlation basin maps and basin summary files will be delivered, these will be used as inputs to the regression
- ❖ Regression this is where regression information files will be delivered
- ❖ Forecast this is where the forecast graphs, maps and summary files will be delivered



Forecasting

Correlation outputs:

The **Correlation** folder contains the following files for each NetCDF predictor:

- correlation-p-values.tif raster map of p-values
- correlation-basins.tif raster map of the basins, pixel value is the basin ID
- correlation-basin-avgs.csv summary of basins detected (where p-values are <= 0.05)
- forward-selection.csv output of forward selection algorithm on eliminating basins with less significance (where p-values are <= 0.03)

Name

NOAA_ERSST_Dec_1960_2021_sst_Dec_1981-2010_FRANCISTOWN_correlation-basin-avgs.csv
NOAA_ERSST_Dec_1960_2021_sst_Dec_1981-2010_FRANCISTOWN_correlation-maps.nc
NOAA_ERSST_Dec_1960_2021_sst_Dec_1981-2010_FRANCISTOWN_forward-selection.csv
NOAA_ERSST_Dec_1960_2021_sst_Dec_1981-2010_GABORONE_correlation-basin-avgs.csv

NB:
CSV/Text predictors are already areas of high correlation, hence no correlation outputs will be computed by CFT for these formats.

Forecasting

Correlation outputs:

The **Regression > LR** folder contains the following files for each predictor:

- correlation-formula.csv linear regression formula used
- correlation-graph.png plot of the linear regression formula (only for single basins)
- forecast_matrix.csv yearly forecast compared to actual values
- score-contingency-table.csv contingency table used to derive the scores
- score-statistics.csv score summary for the station

NB:

The correlation-graph.png is plotted only where the predictor has a single correlation basin.

Forecasting

Correlation outputs:

The **Regression > MLP** folder contains the following files for each predictor:

- mlpsummary.txt summary of the parameters using for MLP
- correlation-graph.png predictor/predictand relationship plot (only for single basins)
- forecast_matrix.csv yearly forecast compared to actual values
- score-contingency-table.csv contingency table used to derive the scores
- score-statistics.csv score summary for the station

TCC_IOD_1991-2020_TCC_IOD_1991-2020_Dec_1971-2000_Diphala_correlation-graph.png
TCC_IOD_1991-2020_TCC_IOD_1991-2020_Dec_1971-2000_Diphala_FMA_forecast_matrix.csv
TCC_IOD_1991-2020_TCC_IOD_1991-2020_Dec_1971-2000_Diphala_FMA_mlpsummary.txt
TCC_IOD_1991-2020_TCC_IOD_1991-2020_Dec_1971-2000_Diphala_score-contingency-table.csv
TCC_IOD_1991-2020_TCC_IOD_1991-2020_Dec_1971-2000_Diphala_score-statistics.csv

NB:

The correlation-graph.png is plotted only where the predictor has a single correlation basin.

Forecasting

The **Forecast** folder contains the following files for each predictor:

- station_members.csv forecasts members (predictor+algorithm) generated for each station
- station_members_selected.csv filtered forecasts members (low skilled members removed)
- forecast.csv list of all forecasts generated for each station, predictor, and regression algorithm
- station-forecast.csv final forecast for each station based on an weighted average consensus
- station-forecast.geojson final forecast for each station (above) in geojson format
- station-forecast.png final forecast for each station (above) in PNG format
- zone_members.csv forecasts members (predictor+algorithm) generated for each station
- zone_members_selected.csv filtered forecasts members (low skilled members removed)
- zone_station-forecast.csv list of all forecasts generated for each station, predictor, and regression algorithm. The zone each station belongs to is also indicated
- zone-forecast.csv final forecast for each zone based on an weighted average consensus
- zone-forecast.geojson final forecast for each zone (above) in geojson format
- zone-forecast.png final forecast for each zone (above) in PNG format
- forecast_graphs_xxx.png forecast graph of each station for each predictor and regression algorithm

Forecasting

MLP selection of best neural architecture (best model):

To select the best neural architecture for the MLP forecast, an iterative process is carried out running from 2 to 21 for the first hidden layer and 0 to 21 for the second hidden layer.

For each iteration the **Hit Score (HS)** and **r^2** are computed, and the iteration resulting in the highest combination of **$(HS+1) * r^2$** is selected for the MLP forecast.

This iterative process to select the best MLP model results in the long processing time for MLP.

Forecasting

Selection criteria for final forecast:

To generate a final forecast for a station (or zone), the following criteria is applied using the testing dataset (data after training period):

1. Only members with significant Probability of Detection (POD) for their forecasted class will be considered, i.e. the highest POD should lie with the forecasted class.

2. Only members who meet a set minimum Hit Score (HS) will be considered (default is 50)

❖ Only Members who meet the two conditions above will be considered for the weighted average to produce a final forecast

Forecasting

Weighted average

	Member 1	Member 2
Class	2	3
HS	65	55

If for a station (or zone) there are two or more members with hit-scores (HS) above 50 as shown above, then we perform a weighted average (HS used as the weight) to determine the final forecast classification:

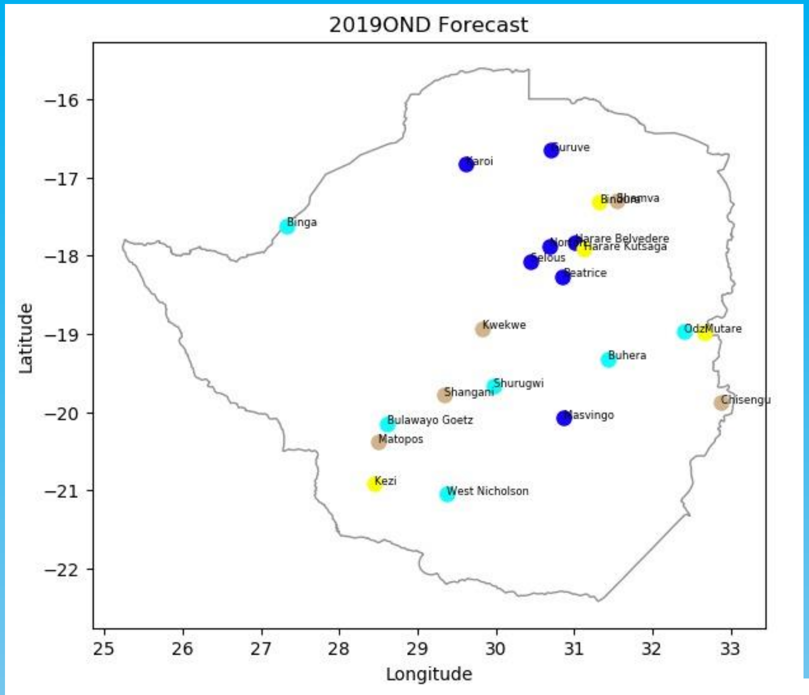
Normal Average = (2 + 3) / 2 = 2.5

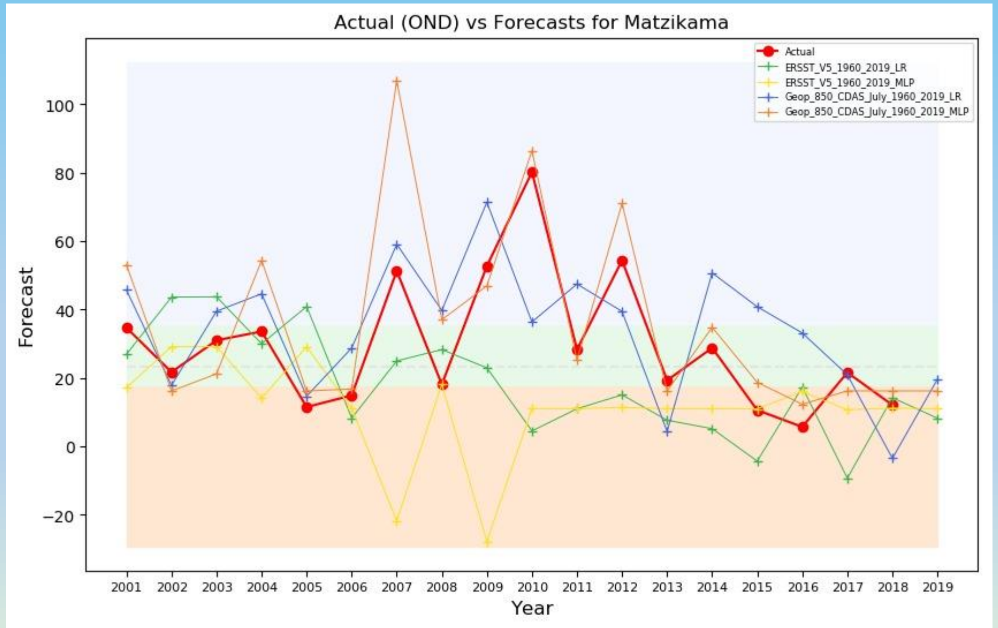
Weighted average = (2*65 + 3*55)/(65 + 55) = 2.46

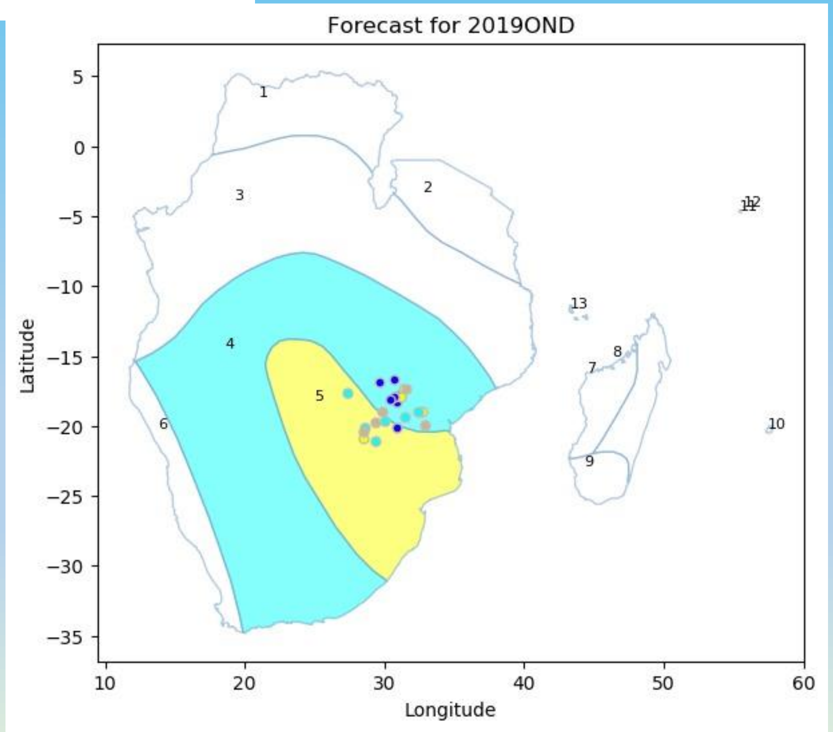
Final forecast classification = 2

Forecasting

Forecast output examples:







THANK YOU

Follow us on:

Twitter:
@AcmadNiamey

Facebook:
https://www.facebook.com/ACMAD-470332183044388

Website:

www.acmad.org