EL NINO DATA ANALYSIS

Akhil A. Naik

Contents

1 Introduction

Every few years the El Nino phenomenon kicks into life in the Pacific Ocean around the equator and can affect the weather around the world, changing the odds of floods, drought,heatwaves and cold seasons for different regions, even raising the global temperatures (*El Nino - What Is It?* 2014). The vast stretch of tropical pacific ocean normally sees consistent winds called *trade winds* that blows from east to west. These winds push warm water in their direction of travel thus making the warm water pile up on the western side of the ocean around Asia and Australasia. On the other side of the ocean, around South and Central America, as warm water gets pushed away from the coast, it is replaced by cold water which is pulled up from deeper down the ocean, a process called *upwelling* thereby creating a temperature difference across the tropical pacific with warmer water piled up in the west and cooler water in the east. The warm water adds extra heat to the air which causes the air to rise with more vigor and rising air is the main cause of unsettled weather. The rising air in the west sets up atmospheric circulation with warm moist air rising on one side of the ocean and cooler dryer air descending on the other (*El Nino - What Is It?* 2014).

El Nino is a climate pattern that describes the unusual warming of surface waters in the eastern tropical Pacific Ocean (Jeannie Evers 2022). El Nino can set off a chain of events which can weaken or reverse the usual flow of trade winds, thereby less push of warm surface water to the western side of the ocean and less upwelling of the cold water on the eastern side. This allows the usually colder part of the ocean to be warm and unsettles the climate changing the rainfall pattern across the equatorial pacific as well as large scale wind patterns. This affects the temperature and rainfall across the world. This can have direct socio economic effect of lives on people thus rises a need for predicting the weather and study these climatic variations so that we can prepare for the worst weather. The El Nino/Southern Oscillation (ENSO) cycle of 1982-1983, the strongest of the century, created many problems throughout the world. Parts of the world such as Peru and the Unites States experienced destructive flooding from increased rainfalls while the western Pacific areas experienced drought and devastating brush fires (Dr Di Cook 1999). Since this cycle was neither predicted nor detected until it was near its peak, this highlighted the need for an ocean observing system called Tropical Atmosphere Ocean (TAO) array to support studies of large scale ocean-atmosphere interactions on seasonal-to-interannual time scales (Dr Di Cook 1999).

In the sections ahead we analyse El Nino Data Set from UCI Machine Learning Repository (Dr Di Cook 1999) which contains oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific collected with the TAO array which was developed by the international Tropical Ocean Global Atmosphere (TOGA) program and which provides real time data. The TAO array consists of nearly 70 moored buoys, developed by National Oceanic and Atmospheric Administration's (NOAA) Pacific Marine Environmental Laboratory (PMEL), spanning the equatorial Pacific, each of which measures air temperature, relative humidity, surface winds, sea surface temperatures and subsurface temperatures down to a depth of 500 meters and a few of the buoys measure currents, rainfall and solar radiation and other oceanographic and surface meteorological variables critical for improved detection, understanding and prediction of seasonal-to-interannual climate variations originating in the tropics, most notably those related to the ENSO cycles (Dr Di Cook 1999). The objective is to check key features of the air and sea temperature variations and know how these variations depend on position.

2 Exploratory Data Analysis

The El Nino data set contains 178080 instances/observations with 12 attributes as discussed ahead taken from the buoys from as early as 1980 from some locations.

2.1 Attribute Information

The attributes/variables consists of date when the observation was recorded, latitude and longitude of buoys, zonal winds (west<0, east>0), meridional winds (south<0, north>0), relative humidity, air temperature and sea surface temperature. Note that the date has been split into year, month and day variables separately as well as entire string.

Attributes:

1. obs: observation number
2. year
3. month
4. day
5. date
6. latitude
7. longitude
8. zon.winds: zonal winds (east-west latitudinal component)
9. mer.winds: meridional winds (south-north longitudinal component)
10. humidity
11. air temp.: air temperature
12. s.s.temp.: sea surface temperature

2.2 Data Manipulation And Summary Statistics

Our first step is to import the data set and prepare it for further analysis. The data is available in the *tao-all2.dat* file, with space seperated values and some null values as *(.)*. Note that columns/attribute names are not present in the same file and must be imported/added from a different *tao-all2.col* file. All the explanatory variables are continuous numerical type as expected (except date and obs number). We convert the string date to date format. Following table shows first few observations of the data set after combining date and removing irrelevant features:

Table 1: El Nino Data

date	latitude	longitude	zon.winds	mer.winds	humidity	air.temp.	s.s.temp.
1980-03-07	-0.02	-109.46	-6.8	0.7	NA	26.14	26.24
1980-03-08	-0.02	-109.46	-4.9	1.1	NA	25.66	25.97
1980-03-09	-0.02	-109.46	-4.5	2.2	NA	25.69	25.28

Our next step is to explore the data for missing values as missing values can significantly affect our analysis. We also check the statistical parameters like mean, median, standard deviation etc. to get an overview of the data in terms of its central tendency and spread.

Table 2: Statistical Summary

	latitude	longitude	zon.winds	mer.winds	humidity	air.temp.	s.s.temp.
nbr.val	178080.00	178080.00	152917.00	152918.00	112319.00	159843.00	161073.00
nbr.null	7054.00	0.00	679.00	1755.00	0.00	0.00	0.00
nbr.na	0.00	0.00	25163.00	25162.00	65761.00	18237.00	17007.00
min	-8.81	-180.00	-12.40	-11.60	45.40	17.05	17.35
max	9.05	171.08	14.30	13.00	99.90	31.66	31.26
range	17.86	351.08	26.70	24.60	54.50	14.61	13.91
sum	84343.23	-9620813.55	-505355.80	38193.10	9124405.20	4297789.35	4464187.94
median	0.01	-111.26	-4.00	0.30	81.20	27.34	28.29
mean	0.47	-54.03	-3.30	0.25	81.24	26.89	27.72
SE.mean	0.01	0.32	0.01	0.01	0.02	0.00	0.01
CI.mean.0.95	0.02	0.63	0.02	0.02	0.03	0.01	0.01
var	21.00	18323.41	11.38	9.00	28.23	3.30	4.23
std.dev	4.58	135.36	3.37	3.00	5.31	1.82	2.06
coef.var	9.68	-2.51	-1.02	12.01	0.07	0.07	0.07

As we can see from the above table, the features are not on the same scale. For example humidity has a mean of *81.24* where as meridional winds have mean of *0.25*, although we show note that its due to the properties of the attributes themselves. Also we can find plenty of NA/missing values with highest of *65761* found in humidity variable. Ideally we can replace these by relevant techniques or neglect these missing obervations but we have to consider that not all buoys are capable of measuring all the attributes values and hence these values are missing dependent on the individual buoy so must be considered during our analysis.

Another key step is to check the latitude and longitude values (axis) which indicate the position of the buoys across the ocean. The available longitude values on a scale on -180 to 180 where longitude>0 is Indian Ocean whereas longitude<0 represents the pacific ocean.

As seen below (Fig 1(1)), the buoys positions are cramped at either ends of the longitudinal scale. Hence for a better view of these position we can convert the value range from [-180,180] to [0,360] for better displaying the positions of the buoys across the globe in a single 2d figure (Fig 1(1)). The latitude values above 0 represent the northern hemisphere while the ones below 0 being the southern hemisphere.

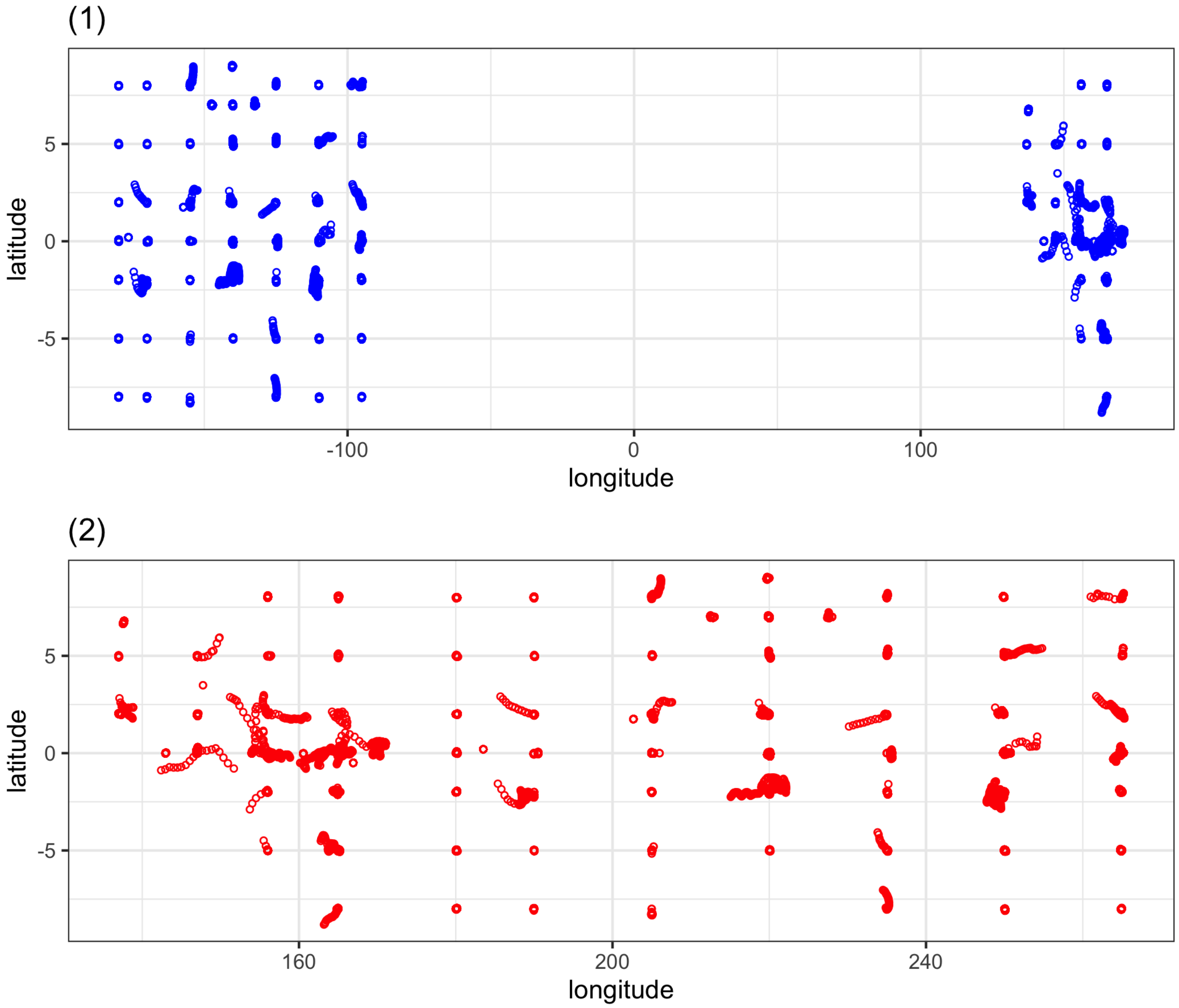

Figure 1: World distribution of Buoys (1)Longitude values before change, (2)Longitude values after change

2.3 Visualisations

In this section we check if the data distributions are normal, check for outliers, explore relationships between the features and study the trends in the data via visualizations. We already got a high level overview from the statistical summary which indicated some features are not normal. From the following pair plot (Fig 2) we find that skewness is observed in some features like zonal winds and humidity. The zonal and meridional winds fluctuated between -10 m/s and 10 m/s. The plot of the two wind variables showed no linear relationship. Also, the plots of each wind variable against the other three meteorolgical data showed no linear relationships. However one key finding as expected is both the air temperature and the sea surface temperature variables shows a positive linear relationship and both values fluctuated between 20 and 30 degrees Celcius. Plots of the other meteorological variables against the temperature variables showed no linear relationship.

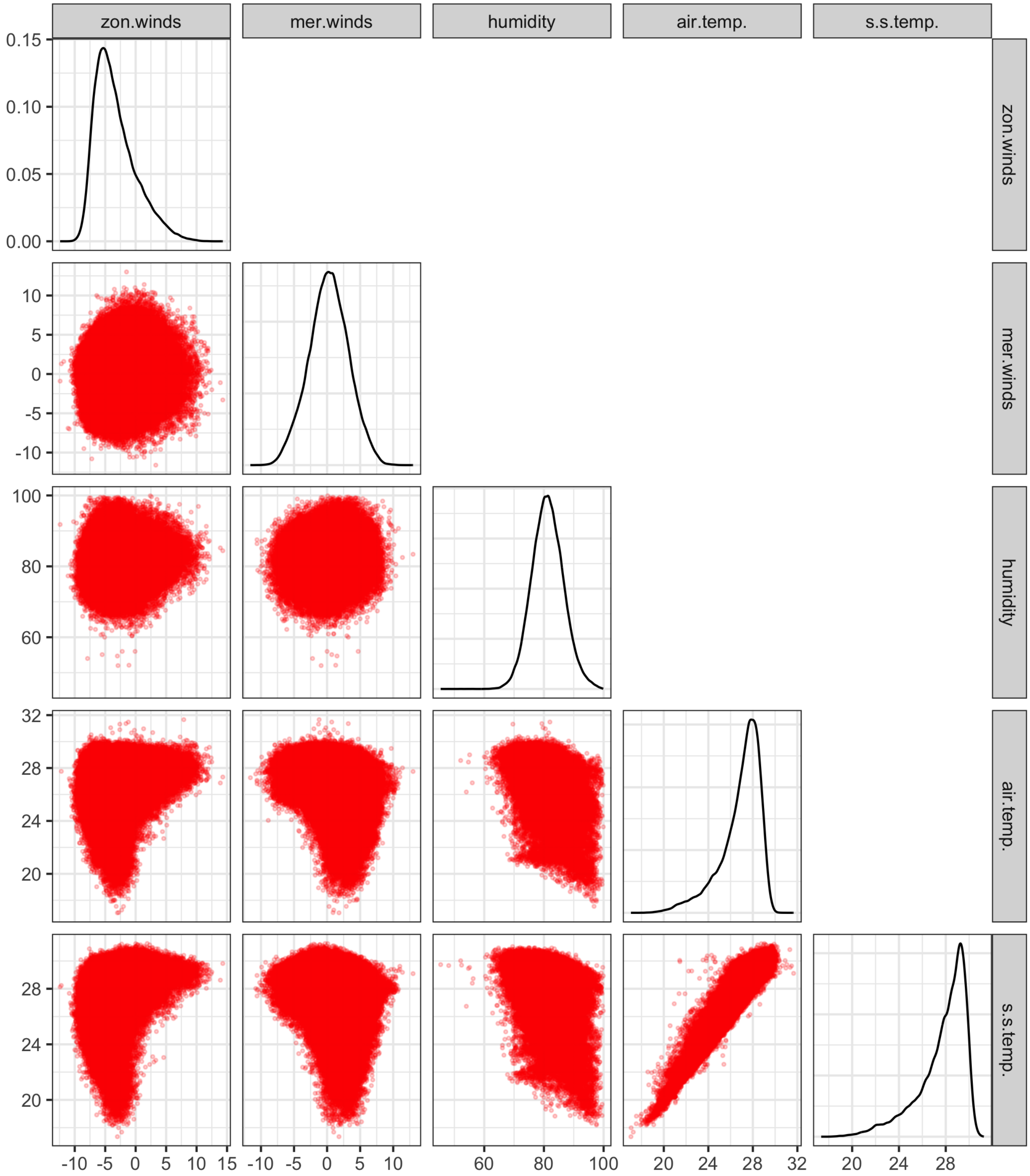

Figure 2: Pair Plots of each continuous variables

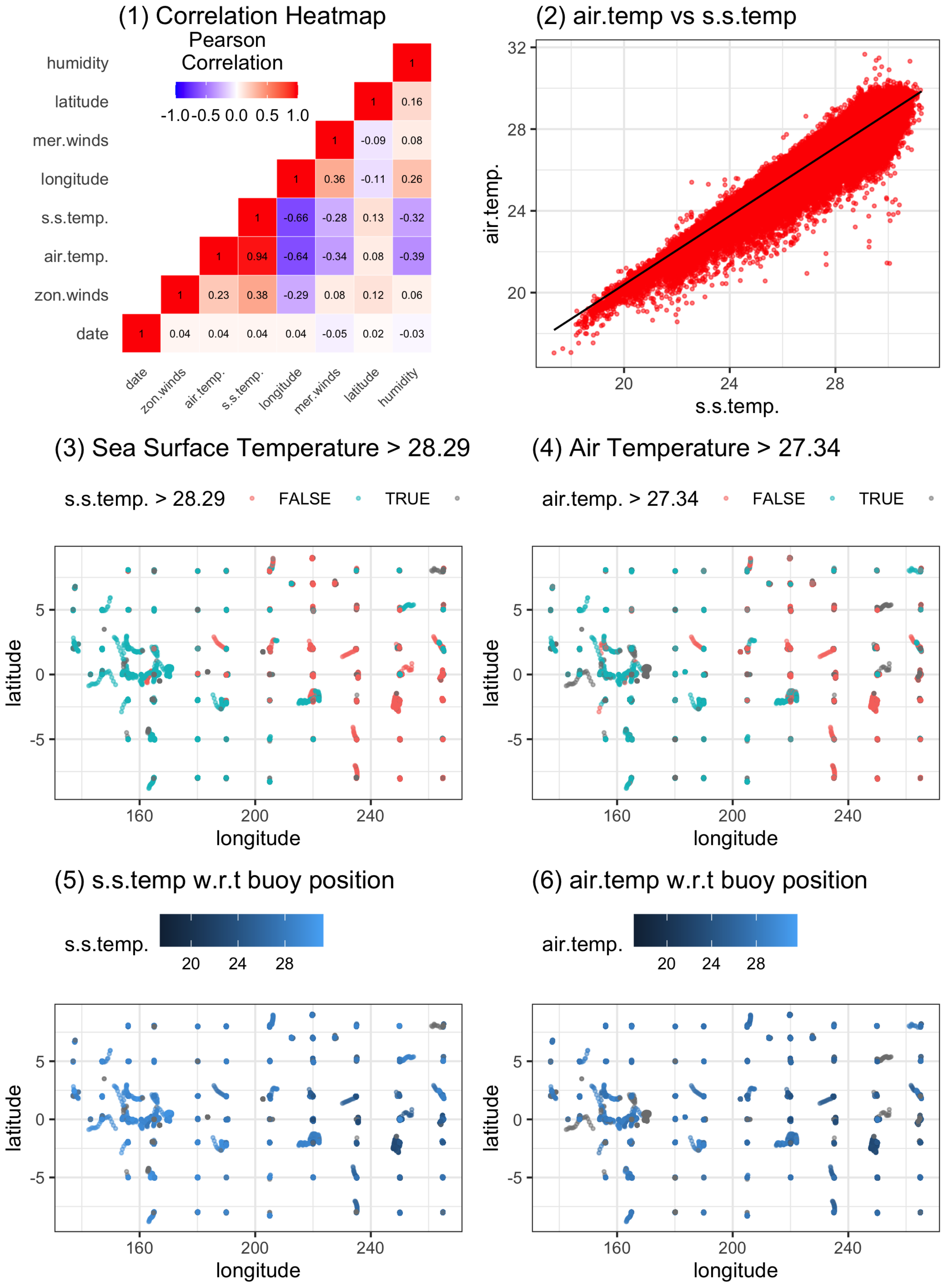

Figure 3: (1)Correlation Heatmap, (2)Relation Between Air temperature and Sea Surface temperature, (3,4,5,6) Variation in the sea surface and air temperatures with respect to position of buoys across the globe

From the above figure (Fig 3 (1),(2)) we confirmed the strong positive linear relation between air temperature and sea surface temperature which has the highest correlation of *0.94*. In our statistical summary (Table 2), we found the median air temperature (*27.34*) and the median sea surface temperature (*28.29*). When these temperature variations of greater than the median values were visualized with respect to the position of the buoys (Fig 3 (3),(4)), we are likely to find low temperatures, both air and sea surface, lower than the median values, in the pacific ocean (updated longitudes: 180-360). We can see (Fig 3 (5),(6)) the pacific region is more cold with low temperatures as compared to the other region (longitude: 0-180). Also the below figure (Fig 4) shows variations in the temperatures with respect to the winds. When median temperature for both air and sea surface, were plotted with respect to the zonal and meridional winds, we find that both temperatures are likely to be lower then the median temperatures when zonal wind is low (below 0) and meridional wind is high (above 0), although it should be noted that there is significant overlap among the observations.

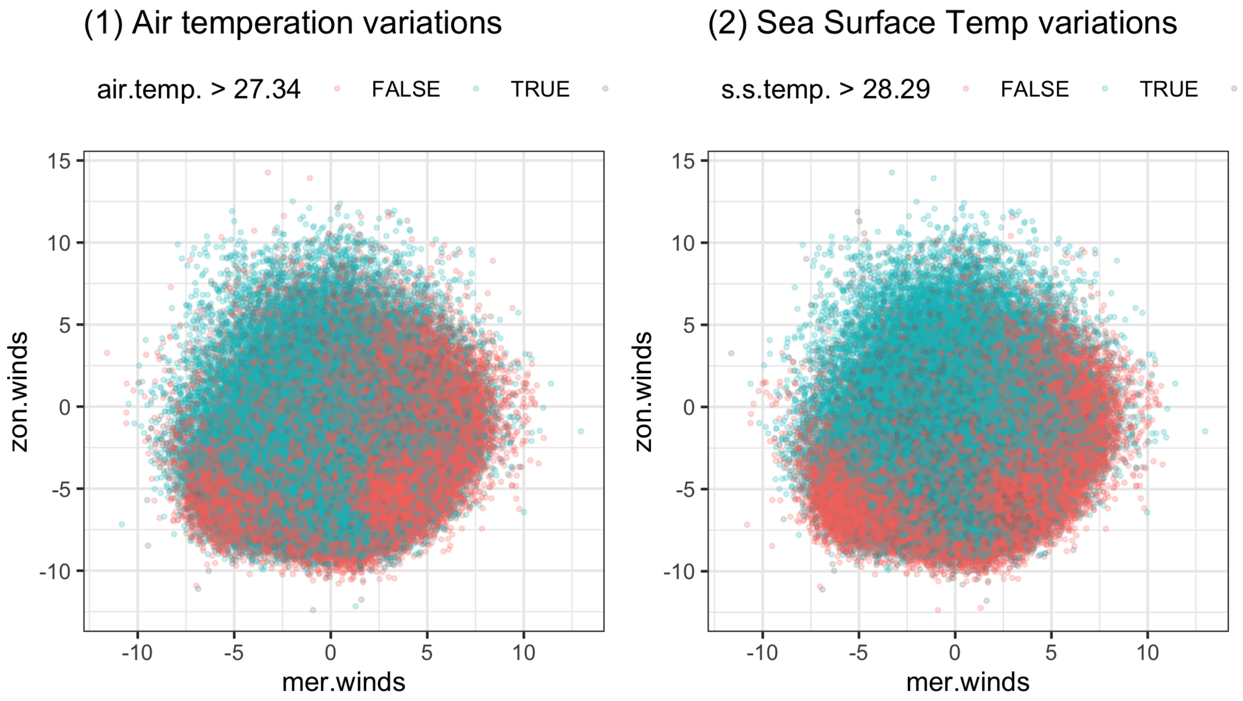

Figure 4: Variations of Air temperature (1) and Sea Surface temperature (2) with zonal and meridional wind

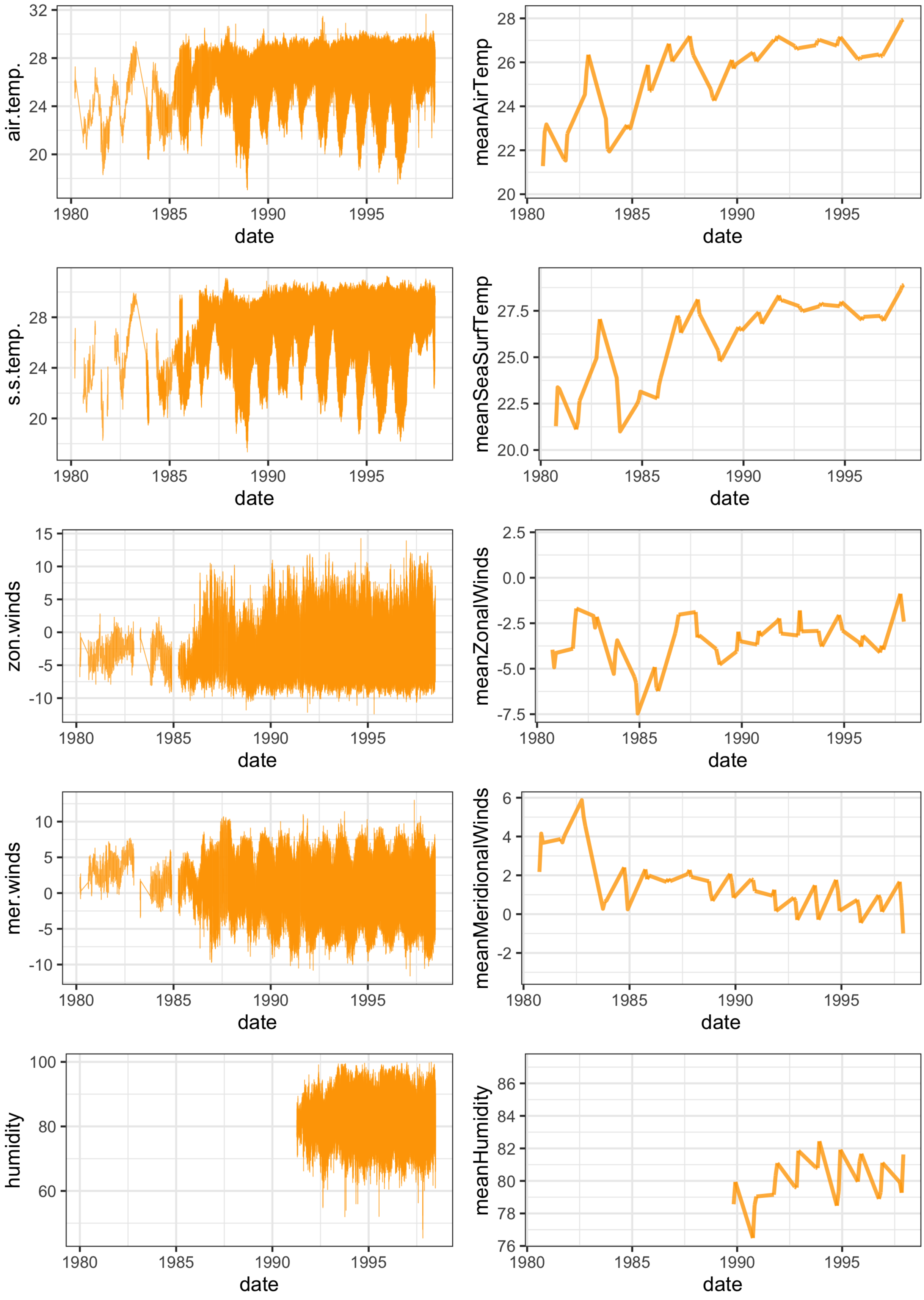

Figure 5: Variations of all the features with respect to time (Values for each day on left and monthly average on right)

The figure above (Fig 5) shows the variations of all the measured features with respect to the time period. Also the month average values for these features across the time period is observed. We find the average air temperature and sea surface temperature has be majorly above 26 degree Celsius since the year 1990. Also we can notice that buoys were not capable of measuring the humidity before 1990s. We also notice that the average meridional winds lie above 0 m/s where as average zonal winds lied below 0 m/s for most part of the time period. However we have to note that the mean values can be affected by outliers hence we can see the difference between their respective overall value plots across entire time period (plots on left) where values range between -10 to 10.

3 Conclusion

We found air temperature and sea surface temperature to be significantly correlated with linear positive relation between the two. Also we found that buoys in the pacific recorded low temperatures compared to the other region. We also found humidity was recorded after 1990 and this can well be because of introduction of new buoys/system for exiting buoys capable of recording the humidity. We can say although all the feature are significant, when trying to predict the weather and me might need additional features along with a high level machine learning algorithm for weather prediction with better accuracy for the El Nino phenomena thereby preparing ourselves for the worst weathers.

4 Code Snippets

```r
# Reading data
columns = c('obs','year','month','day','date','latitude','longitude',
            'zon.winds','mer.winds','humidity','air temp.','s.s.temp.')
el_nino_data = read.table('tao-all2.dat',sep=' ',na.strings = '.', col.names = columns)
# Changing date format
el_nino_data[,'date'] = as.Date(as.character(el_nino_data[,'date']), format= "%y%m%d")
# Changing longitutde scale
el_nino_data[,'longitude'] = el_nino_data[,'longitude'] %% 360
```

```r
# Relation between s.s.temp and air.temp
pr = ggplot(df, aes(y=air.temp., x=s.s.temp.))+
  geom_point(size=0.6, alpha=0.5, col="red") +
  stat_smooth(method = "lm", col = "black", size=0.5) +
  ggtitle("(2) air.temp vs s.s.temp")+ theme_bw()
# median sstemp = 28.29
p11 = ggplot(df, aes(y=latitude, x=longitude))+
   geom_point(size=0.6, alpha=0.5,mapping=aes(col=s.s.temp.>28.29),
position='jitter') + ggtitle("(3) Sea Surface Temperature > 28.29")+ theme_bw()+
theme(legend.position='top', legend.justification='left',
      legend.direction='horizontal')
```

```
# median airtemp = 27.34
p21 = ggplot(df, aes(y=latitude, x=longitude))+
   geom_point(size=0.6, alpha=0.5,mapping=aes(col=air.temp.>27.34),
position='jitter') + ggtitle("(4) Air Temperature > 27.34")+ theme_bw()+
theme(legend.position='top', legend.justification='left',legend.direction='horizontal')
```

```
# feature Variations w.r.t. time
dfNew = el_nino_data %>%
  group_by(year,month) %>%
  summarise(date = as.Date(paste(as.character(year),as.character(month),"1",sep=""),
                           format= "%y%m%d"),
            meanAirTemp = mean(air.temp., na.rm = TRUE),
            meanSeaSurfTemp = mean(s.s.temp., na.rm = TRUE),
            meanZonalWinds = mean(zon.winds, na.rm = TRUE),
            meanMeridionalWinds = mean(mer.winds, na.rm = TRUE),
            meanHumidity = mean(humidity, na.rm = TRUE))

p12 = ggplot(df, aes(x=date,y=air.temp.)) +
  geom_line(col="orange", size=0.2, alpha=0.8) +theme_bw()
p22 = ggplot(dfNew, aes(x=date,meanAirTemp)) +
  geom_line(col="orange", size=1, alpha=0.8) + theme_bw()
#...
```

References

Dr Di Cook, Department of Statistics. 1999. *El Nino Data Set*. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/El+Nino.

El Nino - What Is It? 2014. *Met Office - UK Weather*. https://www.youtube.com/watch?v=WPA-KpldDVc&ab_channel=MetOffice-UKWeather.

Jeannie Evers, Emdash Editing. 2022. *El Niño*. *National Geographic Society*. https://education.nationalgeographic.org/resource/el-nino.