Data Driven ENSO Prediction

Contents

- *690010160*
- *690034411*
- *670025305*
- *690042718*
- *690065386*
- *680065574*
- *690040375*

Abstract

El Niño and La Niña are highly effective events that occur irregularly in and around the Pacific Ocean. Otherwise known as ENSO, this climate phenomenon needs to be understood and predicted to help minimise the consequences of extreme weather events that occur when El Niño or La Niña are active. the events, for example, could be hurricanes, monsoons, droughts and much more, all of which have a devastating effect on the places that experience them. Many researchers have studied ENSO and have found irregular 'patterns' of weather and the reasonings behind them. There are multiple forms of measurements taken around the pacific gathering data to be able to help predict ENSO events, and hence mathematicians can create models to predict future events. These models could revolve around different things being influential on ENSO, such as the seasons, or they have the possibility of greater accuracy, non-linear vs linear. Many models lead to useless outcomes, but some better ones give insightful information about future weather events.

Introduction

Every 3 to 7 years the regions of, and around, the Pacific Ocean experience a climate phenomenon called El Niño. This is an event that occurs due to the eastern Pacific Ocean warming up a possible 10°C higher than normal and could last from a few months up to a year. It is also referred to as the El Niño Southern Oscillation, or ENSO, which also includes an event called La Niña, which is a period much like El Niño but instead its cools the Pacific Ocean. ENSO is one of the most significant climate phenomena in the world. There are many factors that lead to an El Niño event being declared, such as the wind, currents, and overall temperature of the Pacific Ocean. This warming effect over the Pacific Ocean is measured by the southern oscillation index (SOI), which is comprised of measurements of sea level pressure of the atmosphere over different regions of the ocean. There are also many effects of ENSO, both positive and negative. The goal of our project was to prediction ENSO by using past data which involved many steps, from understanding the basics of El Niño, to being able to code precise and in-depth models of ENSO.

How/when does ENSO happen? (NOAA, 2016)

- The surface winds over the Pacific Ocean close to the equator tend to blow from east to west. For unknown reasons, sometimes this wind gets stronger or weaker for periods of a couple of weeks up to months.
- The weaker winds allow for the surface temperature in the eastern pacific to rise/build up. The atmosphere above the Pacific sometimes responds to the hotter surface temperatures with 'increased rising air motion' and 'above average rainfall.' This difference in sea surface temperatures and the atmosphere is what tiggers and El Niño event. Furthermore, the warmer the water gets the weaker the winds get, and so the ocean gets warmer, meaning it continues in a cycle, and the El Niño event gets more and more dramatic. La Niña often follows the following year.
- The signs of La Niña start to show when the winds get stronger, which push surface ocean water from the East Pacific to the west. This means that cooler water from deeper down in the Pacific Ocean rises in the Eastern Pacific. If this continues it can affect the rising air movement and rainfall which is the start of a La Niña event. The same things happens now as it does in El Niño with a cycle of stronger winds leading to cooler waters, leading to stronger winds, leading to cooler waters etc.

How is ENSO measured? (NOAA, updated 2021)

The Southern Oscillation Index (SOI) is a data set that involves measurements of sea level pressure (SLP) taken at Tahiti and Darwin, that are then used to calculate measurements of SOI. There is also data of sea surface temperature, and the measurements are taken in distinct regions of the Pacific Ocean. Each of these regions are separate indices of sea surface temperature (SST), and this data – along with the SOI – is what we'll use to model ENSO.

The following equation is how the southern oscillation index is calculated:

$$\text{SOI} = \frac{\text{Standardised Tahiti - Standadarsied Darwin}}{\text{Monthly Standard Deviation}}$$

Where standardised Tahiti is calculated using the measured value of sea level pressure, subtracting the mean SLP and dividing the result by the Tahiti SLP standard deviation. The same calculations are used to calculate Darwin SOI measurements. These are shown below:

$$\text{Standardised Tahiti} = \frac{\text{Actual Tahiti SLP-Mean Tahiti SLP}}{\text{Standard Deviation Tahiti}}$$

and,

$$\text{Standardised Darwin} = \frac{\text{Actual Darwin SLP-Mean Darwin SLP}}{\text{Standard Deviation Darwin}}$$

When the SOI measurement is negative, this suggests warmer climate, and when its positive, it's cooler.

There are 4 regions that are measured for SST called Niño 1, Niño 2, Niño 3, and Niño 4. There is also a region called Niño 3.4, which is in between Niño 3 and 4. These 3 (3, 3.4, 4) are the three most important indices for measuring ENSO. Niño 1 and 2 can be interesting if looking at a specific area, but when looking at ENSO as a whole - and trying to predict it - they are unnecessary.

What are the effects of ENSO? (NOAA, 2014, updated 2021)

As ENSO is such a significant climate pattern, there are many places (and hence people) that will be affected. El Niño and La Niña can increase the likelihood of extreme weather events. These extreme weather events can massively disrupt workers, destroy local farms etc, and impact the economy. ENSO tends to greatly affect rainfall patterns all over the world, as well as temperature. The greater effects of El Niño and La Niña will be discussed in depth later, from globally to locally, and from the changes in climate to the impact on people.

What is our project?

The aim of our project is to research ENSO and develop models to be able to predict future El Niño and La Niña events. This report will summarise all the work we have done for the project, from the basics of ENSO to the much more complex modelling and analysis. We intend to make it accessible and easy to understand, whilst maintaining a level of detail and complexity. In our group, we naturally gravitated towards different elements of the project. So far, we have discussed the foundations – both mathematical and scientific – of ENSO, any important definitions that are introduced later will be explained then. The next stages of this report are as follows; Research, Models and Selection (our methodology), Analysis and Predictions (our findings), and Conclusion.

Research

The workings of ENSO:

ENSO is the second most major cause of weather fluctuation, accounting for a significant portion of short-term climate variance globally, closely following the seasonal cycle. ENSO accounts for over half of the fluctuation in local weather in certain areas. ENSO has the greatest direct impact on regional climatic patterns in the area closest to the tropical Pacific. On the other hand, signals of its influence may change the worldwide seasonal temperature and precipitation patterns (Climate, 2020). ENSO is also used to detect temperature differences on the surface areas of the Pacific, Atlantic and Indian ocean.

During the summer in the southern hemisphere, winds blow from east to west under normal conditions. This raises sea levels in the western Pacific, preventing warmer water from moving toward South America. Instead, it is heated and remains mostly in place. During an ENSO year, wind velocity

is drastically lowered. This permits the western Pacific's warmer water to move eastward, raising South America's coast temperatures. The impacts of this increase in ocean temperature are carried into the air, affecting the whole West Coast of the Americas (National Drought Mitigation Centre, 2018). The fact that average years are mentioned indicates that ENSO is not expected to occur often. One may try to explain this behaviour only in statistical terms, but this would provide no insight into the phenomenon's causes.

On the other hand, if a mathematical model can be created that describes the qualitative traits of ENSO, then at least part of the physics underlying it can be understood. Using a simple differential equations model, research published in science sought to explain the qualitative properties of ENSO. This exercise is focused on this model. Consider the equatorial Pacific Ocean a fluid-filled big box (although a fairly huge one) (Finotti, 2020). The water in this part of the Pacific is generally separated into two zones in the vertical direction, which is one of its distinguishing features. There is a surface body of water where the temperature changes and currents circulate and a deeper body of water where the temperature stays relatively constant. The thermocline is a surface that separates the two bodies. T- stands for the constant temperature below the thermocline. One must also be worried about a second equilibrium (Finotti, 2020). The sun shines on the water every day, warming it. Simultaneously, the ocean radiates and transfers heat into the atmosphere, away from itself. In this method, the ocean surface retains some temperature, which is labelled T. *If the temperature of the ocean climbs over T*, it cools according to Newton's Law of Cooling: $T' = -A(T - T^*)$

The origins of ENSO:

The atmospheric component of this phenomenon, the Southern Oscillation, was first introduced in 1924 by Sir Gilbert Walker. Due to his strong mathematical qualifications, the English physicist was appointed as Director-General of Observatories in the Indian Meteorological Department. From 1904 to 1924, Walker's duty was to forecast Indian monsoon rains whose failure could result in famine (Davis, 2001). To achieve this, Walker implemented statistical methods to the problem of monsoons making him a "pioneer in the use of correlation in meteorology" (Normand, 1953). More precisely, he developed systems of regression equations through correlation analysis that would depend on various predictor variables to predict Indian monsoon rainfalls. One of them is central in the understanding of ENSO: The atmospheric pressure at Centres of action (i.e.: at locations central in the world's weather). In fact, while studying the Indian monsoons, Walker discovered a correlation between those and the pressure oscillations at key long-distance locations and developed what he called the Southern Oscillation that he described like so: "In general terms, when pressure is high in the Pacific Ocean, it tends to be low in the Indian Ocean from Africa to Australia; these conditions are associated with low temperatures in both areas, and rainfall varies in the opposite direction to pressure. Conditions are related differently in winter and summer…" (Bliss, Walker, 1932). Furthermore, the English physicist was an ardent defender of a world's weather and believed that the Southern Oscillation would be of great use to forecast the world's weather (Walker, 1932).

The impacts of ENSO:

The El Niño and La Niña oscillations impact temperature, seasonal winds, and precipitation, as well as agriculture, fisheries and water supply, causing major economic, environmental and social consequences in all continents. First of all, "ENSO especially impacts winter rainfall and temperature distribution" (Knox, et al., 2015) in areas neighbouring the Pacific Ocean but also "stretch far beyond the region through interactions called 'teleconnections." (Knox, et al., 2015), implying that the Pacific oceanic and atmospheric conditions impact seasonal precipitations and temperatures in all territories, regardless of their location or distance from the origin of the oscillations (IRI, 2021). In fact, "Both El Niño and La Niña influence where Atlantic hurricanes tend to form" (Climate.Gov, 2021) even if they originate in the Pacific, reason why the impacts are on a global scale and there are serious concerns from organizations like NOAA or FAO who declared that: "Fourteen countries in Africa, the South Pacific, Asia and Central America are being specifically targeted due their increased risk to extreme weather and a subsequent negative effect on vulnerable people. Another 19 countries are classed as facing moderate risk." (FAO, 2021). In order to contain or even avoid these "substantial socio-economic impacts" (Davey, 2014) it is crucial to estimate the possible consequences of the ENSO oscillation, that in some regions "is responsible for as much as 50 percent of year-to-year climate variability" (NASA, 2017), providing a probable analysis of the risk.

El Niño oceanic conditions typically foresee a general increase in ocean temperature along the east and centre of the equatorial Pacific as well as "A deeper than average oceanic thermocline across the east-central equatorial Pacific, with depths typically ranging from 150-175 m." (NOAA, 2021). Accompanied by atmospheric conditions such as an increase in convective rainfalls and below average air pressure in the eastern half of the equatorial Pacific compensated by a decrease in convective rainfalls and above average air pressure in the western equatorial Pacific, regions like Indonesia and northern Australia (NOAA, 2021). Furthermore, in the eastern equatorial Pacific, easterly trade winds are noted to be weaker than average alongside with westerly winds encountered across the western equatorial Pacific (NOAA, 2021). Teleconnections associated with El Niño appear in the period from August to October, where westerly winds located in the upper atmosphere cause vertical wind shears above average causing less hurricanes in the tropical North Atlantic, while vertical wind shears lower than average generate an increase in hurricane activity across the eastern tropical North Pacific (NOAA, 2021). Overall, "Major El Niño events—such as 1972-73, 1982-83, 1997-98, and 2015-16—have provoked some of the great floods, droughts, forest fires, and coral bleaching events of the past half-century." (NASA, 2017)

La Niña oceanic condition, on the other hand, typically entail a below average oceanic temperature along the east and central equatorial Pacific as well as "A shallower than average oceanic thermocline across the east-central equatorial Pacific, with depths typically ranging from 50-100 m." (NOAA, 2021). Differently from El Niño, the ocean temperatures are followed by atmospheric conditions such as a decrease in convective rainfalls and above average air pressure in the eastern half of the equatorial Pacific compensated by an increase in convective rainfalls and below average air pressure in the western equatorial Pacific (NOAA, 2021). In addition, easterly winds are expected to be stronger than average in the entire equatorial Pacific. Teleconnections associated with La Niña appear in the same period as for El Niño, where easterly winds located in the upper atmosphere cause vertical wind shears below average causing more hurricanes in the tropical North Atlantic, while vertical wind shears higher than average generate a decrease in hurricane activity across the eastern tropical North Pacific (NOAA, 2021).

Looking more deeply into the specific regional impacts, it is important to evaluate the direct and indirect consequences on neighbouring regions of the Pacific: The American continent on the western part, Australia, and Asia on the eastern part, as well as the repercussions on the African continent and Europe. These periodical changes impact food and water supply as well as the entire population from a socio-economic point of view.

In order to analyse at best the impacts on the west side of the pacific, we will look at North America, USA and Canada, followed by Central America and finally South America.

In an El Niño year, temperatures as well as precipitations impact the entire American continent. In fact, the Met Office estimates a warmer tendency, "approximately a greater than 50% chance that the temperature will be in the top third of temperatures observed" (Met Office, 2021) in that specific period, in the Alaska/West Canada area from January to May, in the North East, region characterized by the big lakes, from December to March, and almost the entire Latin America, from Guatemala to the far Chilean edge, all year long, but in Bolivia and Paraguay especially from July to March and in the far south from June to August (Met Office, 2021). Whereas a colder season is expected in the area going from Massachusetts to Oklahoma from June to September and in the region around Texas and Mexico from December to April. Concerning precipitations, on the other hand, predictions are not as accurate since "data are more limited and probability estimates are more uncertain" (Met Office, 2021). A drier tendency is expected on the US east coast from December to May, while a wetter tendency is expected in the south of the US from September to February. In the islands of Central America, we can expect wetter tendencies from January to April and drier tendencies from June to October, identified as a mixed tendency, reversing with season. From Guatemala to the north Brazilian region, we expect warmer tendencies all year long, but in Central America specifically from June to October and on the east Brazilian coast from February to May. Finally, other wetter tendencies have been common in Ecuador from September to November and on the east Argentinian coast and Uruguay from September to May; while a drier tendency is expected in Peru from February to March (Met Office, 2021).

In the US, El Niño has specific consequences on hurricanes and storms: "El Niño does increase the chances for a wet and stormy winter and early spring overall across the southern tier of the United States" (Climate.Gov, 2021), in fact "The likelihood of tornadoes and severe weather increases in the Florida peninsula." (Knox, et al., 2015) which may delay the "Harvests of summer crops such as corn, peanuts, and cotton" (Knox, et al., 2015). Moreover, while "the continental United States and Caribbean Islands have a substantially decreased chance of experiencing a hurricane during El Niño" (Climate.Gov, 2021), eastern and central Pacific experience more hurricanes, while El Niño contributes to "fewer Atlantic hurricanes." (Climate.Gov, 2021). Finally, "During El Niño, the Aleutian low-pressure centre over the North Pacific deepens, high pressure develops over western North America, and low pressure prevails over the south-eastern US." (McPhaden, 2002), which leads "to the subtropical jet stream moving into Florida, southern Georgia, and Alabama, steering cloudy, rain-bearing systems into the region in winter." (Knox, et al., 2015), this heavily affects agriculture: "Wheat yields in southern AL and GA are generally higher than average" (Knox, et al., 2015), although corn yields are typically lower than average, as well as "Yields of winter vegetables such as tomatoes, bell peppers, sweet corn, and snap beans" (Knox, 2015). In the contrary, in Central America, "El Niño is typically associated with below-average rainfall" (FAO, 2021), especially in the zone of the Dry Corridor, where dry conditions could worsen and cause "delays in plantings and vegetative development of basic grains of the postrera season with planting starting in September and harvest in November." (FAO, 2021). Hence El Niño indirectly impacts the economy as well as on "households, communities and institutions" (FAO, 2021) which is why FAO is supporting the area of the Dry Corridor, in order to "prevent and address disaster risks that affect agriculture and food and nutrition security in a timely and efficient manner." (FAO, 2021). El Niño also majorly impacts fisheries "a particularly striking example of which was the collapse of the Peruvian anchoveta fishery following the 1972–1973 El Niño" (McPhaden, 2002), due to intense fishing pressure as well as higher than average mortality rates that lasted for at least 10 years, dramatically affecting Peru's economy. More recently, other fisheries were affected by El Niño in the Galapagos Islands, Panama and Australia, because of a "Massive and widespread coral bleaching occurred during 1998 ... in response to the exceptionally strong 1997–1998 El Niño." (McPhaden, 2002), since "lethal bleaching is that it threatens the vitality of coral reef ecosystems, which support local fisheries and provide tourist income" (McPhaden, 2002). The latest example is April 2016, when "nearly 8,000 tons of sardines died and washed up along the coast of Chile, likely the result of El Niño related changes in the ocean." (NASA, 2017)

In a La Niña year, temperatures and precipitations tend to be opposite from El Niño events. In fact, colder tendencies have been identified in the Alaska/West Canada region from February to November and from Guatemala to central Argentina, excluding Brazil, all year long. On the contrary, the south states experience warmer tendencies from October to April as well as the east coast from July to October. With regard to precipitation impacts, are expected to be drier in the southern states and Mexico as well as Cuba and Dominican Republic from December to February, and in central Argentina from October to April. On the contrary, wetter tendencies have been registered in the whole north part of South America, in the period from February to March in Ecuador, from August to February in Venezuela and from February to June in the Brazilian coast (Met Office, 2021).

Since La Niña is the counterpart of El Niño, its impacts are practically causing opposite effects. In fact, "La Niña contributes to fewer eastern and central Pacific hurricanes and more Atlantic hurricanes." (Climate.Gov, 2021), while "The likelihood of tornadoes and severe weather increases in Alabama and Georgia." (Knox, et al., 2015), due to the fact that "Pressure shifts cause the subtropical jet stream in the U.S. to shift north, moving the storm track to northern Georgia and Alabama and leaving Florida sunnier and drier than usual." (Knox, et al., 2015). As well as unusually cold waters in the eastern Pacific Ocean off the coasts of Peru and Ecuador "causing contrasting shifts in local weather patterns as well as in the global climate." (Knox, et al., 2015). This has, again, major impacts on agriculture, like generally higher irrigated corn yields but generally lower pasture crop yields, and specifically the "Flowering of subtropical fruits, such as mango, lychee, and longan, may decrease or be eliminated, but production of tropical fruits such as banana, guava and papaya may increase." (Knox, et al., 2015).

Furthermore, the oceanic and atmospheric oscillations heavily impact the eastern part of the Pacific.

In an El Niño year, in Asia and the west central Pacific region, we expect colder weather in New Zealand from April to November while a warmer tendency is registered for the south and east coast of Australia. In Papua New Guinea and Solomon Islands we experience a colder tendency from June to September while from Myanmar to Indonesia a warmer tendency has been registered from October to June in Thailand and from November to March in Malaysia, along with India experiencing the same tendency from May to December. Finally, Japan is impacted by seasonal reversed tendencies: warmer in November and December and colder from June to September. With regards to precipitations, the majority of the west pacific experience's drier tendencies: in India from May to September, Thailand from March to June, in the Philippines from September to June, in the southwest Australian coast in January and February, from August to December in the southeast Australian coast and from May to January in Papua New Guinea. Only Taiwan and southeast China experience wetter tendencies from October to February (Met Office, 2021).

El Niño was responsible, from 1997 to 1998, for extensive crop failures and livestock losses due to droughts and floods in many regions. For example, Papua New Guinea experienced "Life-threatening food and drinking water short- ages" (McPhaden, 2002); or also Indonesia where droughts and strong fires "produced deadly smog that covered an area one half the size of the continental US, causing widespread respiratory ailments, and contributing to air- line and shipping disasters." (McPhaden, 2002). El Niño also indirectly impacts the spread and development of diseases and bacteria, causing increasing problems to agriculture and food supply: "Flood-contaminated water supplies in some regions con- tributed to outbreaks of cholera and dysentery" (McPhaden, 2002) as well as "breeding grounds for mosquitos and other insects that spread infectious diseases like malaria and dengue fever." (McPhaden, 2002). Finally, El Niño also creates consequences in the socio-economic sector: "Many businesses lost income as a result of weather-related closures, commodity shortages or shifts in market demand for consumer goods." (McPhaden, 2002) underlining even more the importance in the analysis, estimation, and prediction of its variations.

On the other hand, in a La Niña year, precipitations are mainly wetter: in north India from July to September and from January to March in the south, from September to may in Thailand and Vietnam, all year long in Indonesia, from April to January in Papua New Guinea, from September to January in the west Australian coast and finally from November to March on the east Australian coast. The only region experiencing drier tendencies is the area including Taiwan and the eastern Chinese coast from May to August. Temperature is less homogeneous, with colder tendencies in India from June to December and in west Australia from November to February. In contrast, warmer tendencies are experienced from Thailand – from March to December – to Indonesia from November to March, in Japan from July to September, in New Zealand from April until January and in north Australia from July to October (Met Office, 2021).

The drier than average conditions in Southwest Asia "affected winter crops in the area" (FAO, 2021), for example Afghanistan, in 2018, declared a drought emergency that continued for months and "caused substantial cumulative rainfall deficits throughout the country, resulting in the lowest wheat production since 2011" (FAO, 2021). La Niña also brings about "excessive rains and can also increase the risk of flooding, particularly in low-lying agricultural lands" in East Asia, resulting in "extensive damage to standing crops and an increase of pest and disease outbreaks. It can also heighten the potential for landslides." (FAO, 2021). For example, in Papua New Guinea, La Niña triggers wetter conditions over most of the island, as well as "an increased likelihood of floods, landslides and cyclones, with a potential to cause damage to home gardens and standing crops and trees, which are crucial to sustain about 85 percent of the population's livelihoods, with an increased risk of loss of human lives and infrastructure." (FAO, 2021), which also results in increasing damage to crops as well as higher mortality in pigs (FAO, 2021).

Finally, because of teleconnections, the African and European continent present some impacts in temperature and precipitation leading to various consequences of El Niño and La Niña, mainly concerning agriculture, fisheries and food and water supply.

El Niño has an impact on precipitation on Africa as it turns the central tropical African line drier than usual from July to September and the south-eastern African sector from November to April. Conversely, El Niño presents wetter tendencies in Kenya from September to January as well as in Europe, specifically in Spain and Portugal, from August to November. Temperatures are mainly warmer across Africa, specifically in the south from December to April and around Liberia and Ghana from February to April; and across southern Europe in October and November while in January and February a colder tendency has been registered in the Scandinavian region (Met Office, 2021).

El Niño mainly impacts African production as crop and livestock is weakened by lowered rains and increased temperatures (FAO, 2015). As a consequence, food prices will rise, generating even more socio-economic issues to a continent already struggling with its production structure, as well as the concerns relating to food security in the region. Africa is especially impacted by these temperature and precipitation oscillations since its food supply mainly depends on agriculture, hence "The region's small-scale farmers are almost entirely dependent on rain, rendering their output highly susceptible to its variations" (FAO, 2015). Europe is not a particularly affected region; in fact, El Niño only causes "cold winter anomalies over Northern Europe." (Scaife, 2010).

La Niña, on the other hand, doesn't have a very strong impact on the temperatures since it just creates a colder tendency in Congo and Kenya from April to May and in South Africa from November until March, and another colder wave from Morocco (January until May) to Cote d'Ivoire from January until March, and doesn't present any direct impact in Europe. Precipitations are estimated to be wetter in South Africa from November until April, in the African equatorial line from July to September; while drier tendencies are expected in the Iberic Peninsula from September until November (Met Office, 2021).

La Niña greatly affected Africa over the years and had particularly strong effects as recently as 2020/2021. From 2010 to 2011, it caused one of the worst droughts in the region of the Greater Horn of Africa, which resulted in "conflict and humanitarian access constraints" as well as "it pushed almost 10 million people into humanitarian emergency and led to a famine declaration in south-central Somalia." (FAO, 2021). Moreover, less rains in the primary cropping season of 2010 followed by even less rains in the second cropping season (October – December) resulted in "crop failure, reduced labour demand, poor livestock body conditions and excessive animal mortality." (FAO, 2021). On the contrary, the same region benefited from great rains from March to June 2020, "which have resulted in improved crop and livestock production in most areas" (FAO, 2021). Europe is, again, not a particularly affected region, and during a La Niña season the effects are mainly opposite to El Niño, hence implying warm winter anomalies over the Northern European region (Scaife, 2010).

Global warming:

Two main reasons can show there is a correlation between ENSO and climate change. First are the carbon cycles, when ENSO-driven the global air-sea CO2 flix and marine productivity being unusual. Also, the surface and subsurface waters temperature has a remarkable change, it can verify ENSO correlates with the carbon cycle (Keller, Joos, Lehner, & Raible, 2015). Furthermore, ENSO will lead to dry conditions, which easy to trigger bushfires, in the west Pacific like Amazon. Bushfire will emission carbon dioxide and reduce the three, which can absorb carbon dioxide (Bastos et al., 2018). Hence, there is a correlation between ENSO with climate change, because during the ENSO event, there is more CO2 that affect the carbon cycles, and the carbon cycles are the main character of climate change. Secondly, the vegetation can absorb CO2, and the vegetation productivity will affect global warming directly. Sea surface temperatures can affect vegetation productivity dramatically. Vegetation productivity and carbon balance are sensitive to the warm and cool West Pacific War Pool index pattern (Huang et al., 2019). So, the vegetation productivity can facilitate climate change, depending on the sea surface temperature, and when the ENSO event, the sea surface temperature is unusual.

Why is it important:

Why ENSO is important because it not only will bring the weather impact, but also will bring some side effects. Here are some side effects from ecology, economics, and public health.

For ecological, ENSO will reduce the upwelling of cold nutrition-rich water from the bottom of the ocean. It affects marine life and the birds' population. For example, In the Gulf of Panama coral mortality has a dramatic higher after ENSO, the result present that ENSO sea warming events will affect the development of coral (Glynn & D'Croz, 1990). Also, forest fire frequency has inclined during ENSO events, and the wildfire will affect the species living in the forest. For instance, in Sumatra, Indonesia, there are wildfires frequently. It has broken bird community composition and led to a decline in diversity (Adeney, Ginsberg, Russell, & Kinnaird, 2006). Thus, ENSO can lead to a reduction in biodiversity.

In the economic aspect, the ENSO event may damage the infrastructure. For example, in South America, there is a dramatic effect on the efficient fish industries, because upwelling affects the fish population. Moreover, the drought affects the exports of agriculture. Since ENSO event frequency and strength, the major agricultural products such as wheat, corn, cotton, and sorghum production can drop 80% (Anyamba et al., 2014). there are annual damages in the 3 to 4 hundred million US dollar range in agriculture (Chen, McCarl, & Adams, 2001). Next, infrastructure is the core element of logistics. However, brushfire and flooding occur during when ENSO event period. For example, in some island nations inundation events will remarkably damage infrastructure (Hoeke et al., 2013). As a result, the ENSO event affects a countries fish industry, agriculture, and logistics.

In public health, ENSO will create extreme weather in some developing countries such as East Africa and Southern Africa. And these countries do not have a health system. So, the disease will be out of control. Anomalously wet conditions or dry conditions, both can be associated with ENSO, can lead to ecological conditions favouring the emergence of pandemic diseases, for instance, malaria, dengue virus, and cholera (Anyamba et al., 2014). Furthermore, the drought affects agriculture. For example, in Indonesia, most of the citizens affected by the drought are forced to significantly reduce expenditures for food (Keil, Zeller, Wida, Sanim, & Birner, 2007). Also, ENSO can change the condition, which is easy to trigger a bushfire. Also, the smog produced by the bushfire. For instance, in 1997 there is the most hazardous smoke problem, the air pollution index values over in Malaysia (Khandekar, Murty, Scott, & Baird, 2000). Hence, extreme weather has changed the condition of the environment, which can bring public health issues to the world.

Overall, the inference of ENSO is not only on climate change but also have side effects. These side effects can be a disaster to the world. Hence, that is how important are ENSO because it will affect the world by ecology, economics, and public health.

How do we detect? What technology?

ENSO events need many elements to support them, so here are some indices and how can we detect the data. Most of the data is provided by NASA and the National Oceanic and Atmospheric Administration (NOAA). Firstly, underwater temperature and water masses mean the water under 300meters in-depth, so to observe the moored and floating instruments will be satellite-tracked drifting and mooring. Secondly, sea surface temperature was measured by the instruments on ships and mooring. With ever-increasing space technology, the sea surface temperature can be measured from space by radiometers, for example, the Advanced Very High-Resolution Radiometer use microwave emission to measure the temperature. Thirdly, sea surface height is measured by NASA's jet propulsion with the theory that warm water will expand bigger volumes. Fourthly, surface winds can be spotted the RapidScat instrument by NASA. Finally, cloudiness and precipitation are also measured by NASA's MODI instrument in space (Carlowicz & Schollaert Uz, 2017).

Data

First, the Southern Oscillation Index (SOI) is a standardized atmospheric component of El Niño. This traditional ENSO component index is an oscillation in surface air pressure between the tropical eastern and western Pacific Ocean waters. The magnitude of the Southern Oscillation is measured by the SOI. This is computed from changes in the surface air pressure difference between Tahiti (in the pacific) and Darwin, Australia (on the Indian Ocean). El Niño episodes have negative SOI as there is a lower pressure over Tahiti and higher pressure in Darwin. Whereas, La Niña have a positive SOI, as there is a higher pressure in Tahiti and lower in Darwin. SOI strongly correlates with SST anomaly indices. El Niño episodes tend to occur over warm water and El Niña pressure occurs over cold water in part because of deep convection over the warm water. El Niño episodes are defined as sustained warming of the central and eastern Pacific Ocean, thus resulting in a decrease in the magnitude of Pacific trade winds, and a reduction in rainfall over eastern and northern Australia. The opposite is true for La Niña which involves sustained cooling of the central and eastern tropical Pacific Ocean, thus resulting in an increase in the strength of the Pacific trade winds, and increased risk locally of flooding and cyclones in Australia. The weather phenomenon La Niña can contribute to "once in a century" rains battering parts of Australia. The last episode was recorded in 2020 and one is potentially going to strike again this year however The World Meteorological Organization (WMO) is yet to declare but has warned it may re-emerge. Although, the SOI spans back to the 1800s, its reliability is limited due to the location being far south of the equator, resulting in surface air pressure not being directly related to ENSO. To overcome this a new index has been generated, the Equatorial Southern Oscillation Index (EQSOI) centred on the equator, however data only goes back to 1949.

Second, the Niño1/2/3/4/3.4 indices (area-averaged sea surface temperature). The numbers correspond with the labels assigned to the ship tracks that crossed these regions. Data from these tracks enabled the historic records of El Niño to be carried back in time to 1949, as discussed in a classic study by Rasmusson and Carpenter (1982). Niño 1+2 (0-10S, 90W-80W): The Niño 1+2 region is the smallest and most eastern of the El Niño regions and corresponds with the region of coastal South America where El Niño was first recognized by the local populations. This index tends to have the largest variance of the Niño SST indices. Niño 3 (5N-5S, 150W-90W): This region was once the primary focus for monitoring and predicting El Niño, but researchers later learned that the key region for ENSO interactions lies further west (Trenberth 1997). Hence, the Niño 3.4 and ONI became favoured. Niño 3.4 (5N-5S, 170W-120W): The Niño 3.4 anomalies may be thought of as representing the average equatorial SSTs across the pacific from about the dateline to the South American coast. The Niño 3.4 index usually uses a 5-month running mean and El Niño or La Niña events are defined when the Niño 3.4 SSTs exceed +/- 0.4C for a period of sex months or more. ONI (5N-5S, 170W-120W): The ONI uses the same region as the Niño 3.4 index. The ONI uses a 3-month running mean, and to be classified as a full-fledged El Niño or La Niña, the anomalies must exceed +0.5C or -0.5Cfor at least 5 consecutive months. This is the operational definition used by NOAA. Niño 4 (5N-5S, 160E-150W): The Niño 4 index captures SST anomalies in the central equatorial pacific. This region tends to have less variance than the other Niño regions

Mathematical Concepts

Autoregression

Autoregression models predict future values in a time series data set using a linear combination of previous values in the time series. This would be denoted as $AR(P)$ and written as:

$$x_{n+1} = c + \sum_{i=1}^{p} \beta_i x_{n-i} + \mu_i$$

The first unknown parameter in the model is p which would be the number of historical data points used in the model.For example an $AR(2)$ model would be written as $x_{n+1} = c + \beta_1 x_n + \beta_2 x_{n-1} + \mu_i$ has the two previous values in the time series as predictors. The value of p is determined by the partial autocorrelation function (PACF). The mathematical background of the PACF is beyond the scope of the paper but is summarized as: *"The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags"* (Andrew V. Metcalfe 2009). After the PACF is computed for each lag p is chosen to be the point where all further lags the $PACF = 0$. This is often done with a plot.

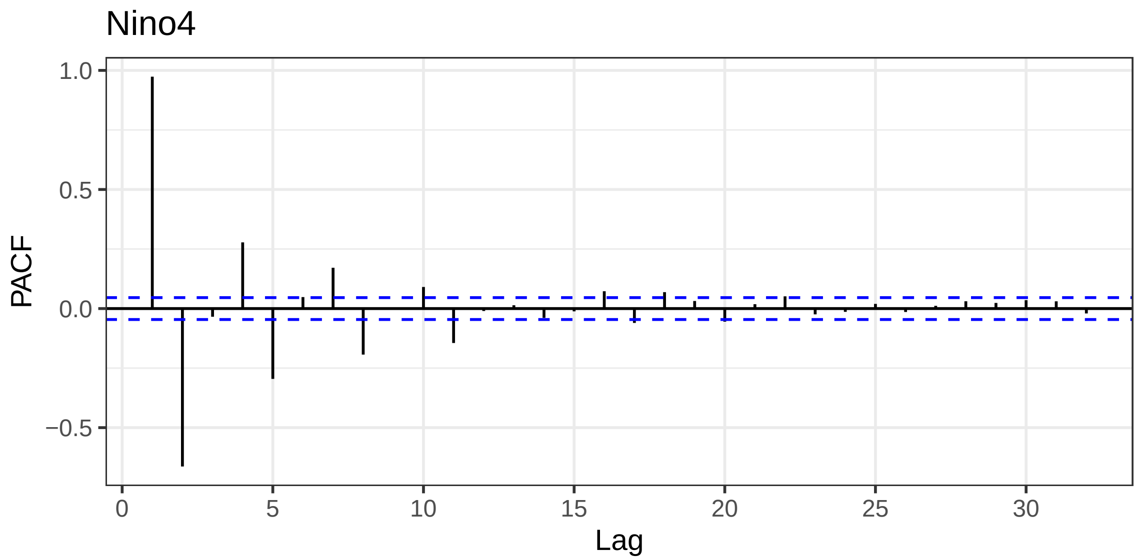

Figure 1: Nino 4 PACF

After the value of p is found the model parameters are estimated using Ordinary Least Squares Regression (OLS). Consider the system:

$$y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \text{ for i=(1,2...n)}$$

Can be written in matrix form:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} \implies \mathbf{Y} = \mathbf{X}\beta + \mu$$

Hence, $\mathbf{Y}$ is the response vector,$\mathbf{X}$ is the design matrix, β is vector of coefficients and μ is the residual vector. Given $\mu_i \sim N(0, \sigma^2) \implies Y_i \sim N(\mathbf{x_i}\beta, \boldsymbol{\sigma^2})$, the log likelihood $l(\beta, \sigma^2)$ is maximized when the residual sum of squares $= (\mathbf{y} = \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$. Minimzing this value gives:

$$\hat{\beta} = (\mathbf{X'X})^{-\mathbf{1}}\mathbf{Xy}$$

Which is the least squares estimate of β

There are two options to increase the complexity of the standard $AR(P)$ process. First, the use of *Non-Linear Predictors*. For example $x_{n+1} = c + \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^3$. The next extension is the use of *Seasonal Predictors*. Such a model would be defined as:

$$x_{n+1} = c + \beta_1 x_n + \beta_2 x_n Cos(\frac{2\pi k_n}{12}) + \beta_3 x_n Sin(\frac{2\pi k_n}{12})$$

Where k_n is the month of the year i.e $k = 1$ for January. Both models are fitted in the same way as the standard $AR(P)$ process but the value of p doesn't need to be determined using the PACF.

The main advantage of autoregression is that is easy to perform. First of all only time series data is required for prediction and the use of OLS is simple in standard statistical packages. Furthermore the use of OLS allows the use of classical inference techniques on models such as parameter hypothesis testing ("Vector Autoregression" 2021). On the contrary, (Hai 2005) such models may be poor at forecasting "turning points" in data. This could have an impact on ESNO prediction but only if the data were at extreme points i.e a global high or low. Fortunately, this is not the current state, however a 3 month moving average is taken to mitigate the effects as this smooths out the data somewhat. It is also a weakness that models are not suited to long term prediction. The impact of this will be mitigated through the use of the skill score which helps decide when the model no longer becomes accurate. However, it is worth noting that even highly complex models (Nan Chen 2021) are skillfil for $\approx$ 12 months.

Vector Autoregression

Vector Autoregression is an extension of the standard $AR(P)$ process in which multiple time series are used as predictor terms. In matrix form a VAR(1) process is:

$$\begin{bmatrix} x_{1,n+1} \\ x_{2,n+1} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{bmatrix} \begin{bmatrix} x_{1,n} \\ x_{2,n} \end{bmatrix} + \begin{bmatrix} \mu_{1,n} \\ \mu_{2,n} \end{bmatrix}$$

Hence, this simply yields two linear equations to be solved but now there are predictors from both time series. As with the Autoregressive model p is to be determined. There are many ways to do this including: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The mathematical background of such metrics is beyond the scope of the paper. The models are fitted as before using OLS.

In the case of ENSO prediction it is likely that $VAR(P)$ modeling will be used on the Nino3 and the Nino4 indices.

The advantages and disadvantages of VAR models are similar to that of AR models but with the added advantage of being able to capture the relationship between two time series.

K-Fold Cross Validation

When fitting models it is important to determine how good each of the models fit is. To do this k-fold cross validation can be used. The process is as follows:

1. Divide data into k equal length sets known as folds
2. Create training data set with k-1 of the folds
3. Create testing set with the remaining fold
4. Fit model onto the training set
5. Predict new using the testing set
6. Asses the predictions using error metrics
7. Repeat until each fold has been the testing set

The main error metric used in cross validation is the Root Mean Squared Error (RMSE). For predicted values $\hat{x_n}$ and actual values x_n. The RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x_n} - \hat{x_n})^2}{n}}$$

In this report cross validation will only be used on AR(P) models.

There are a few advantages to cross validation and virtually no drawbacks, firstly it reduces the chance of over fitting (Gavin C. Cawley 2010) and secondly parameter tuning will help give the the best model for the data. Consequently, increasing the accuracy and validity of model results. It can be computational intensive to perform k-fold cross validation however modern statistical packages make it efficient meaning this is likely not to be an issue.

Reference Forecasts

Reference forecasts are baseline forecasts that help assess how good a RMSE is as well as helping to assess the skill of models. They work by making an arbitrary assessment of the value of x_{n+1} and then calculating the standard deviation of the error of this forecast over the whole data set. *Climatology* assumes that the x_{n+1} value is simply equal to the mean:

$$x_{n+1} = \frac{1}{n}\sum_{i=1}^{n} x_i \implies Error_{Climatology} = SD(X)$$

Where SD represents the standard deviation. *Persistence* assumes that x_{n+1} is equal to x_n

$$x_{n+1} = x_n \implies Error_{Persistence} = SD(x_i - x_{i-1})$$

The values of *Persistence* and *Climatology* are then compared the the RMSE of the model to value if the model was more skillful than the references. A model would be rejected if its RMSE didn't beat either of the references, therefore increasing the efficiency of the modelling process.

Skill Score

Although by definition both VAR(P) and AR(P) simply predict the next value in the time series this does not mean they cannot predict more than one step ahead. The predicted values are simply used in the linear combination. To calculate the number of steps ahead that should be predicted a *skill score* needs to be calculated for each prediction in to the future. To do this the model chosen predicts a certain number of steps ahead (In the case of ENSO no more than 12 the equivalent of a year) across a large sample of the data set which allows the standard deviation of the error to be calculated at each step ahead predicted. Once this has been calculated a simple skill score ("Forecast Skill" 2021) using the reference forecasts can be calculated at for each step ahead:

$$Skill = 1 - \frac{SD_{forecast}}{SD_{reference}}$$

Therefore when the skill score drops below zero this shows that the prediction is no more skillful than the reference forecast at that step ahead. Hence, predictions beyond where the skill score is zero are rejected meaning the accuracy of the predictions made is greater and therefore conclusions drawn from them are more valid. Furthermore, it makes for simple comparison between models drawn from other work.

Probabilstic Modelling

The AR(P) and VAR(P) models are deterministic but it may be useful in the context of predicting ENSO to have probabilistic forecasts. For example, what is the probability that we will be in La Nina state in 3 months time? To get probabilistic forecasts the value of the random noise term in the models is considered. That is $\mu_i \sim N(0, \sigma^2)$ where the value of σ is estimated during the OLS fit. This means at each step ahead we can add uncertainty to the prediction by adding random terms sampled from this distribution. At each step ahead another set of random variation is added. Therefore if there are n random noise elements:

$$\Pr(\text{La Nina}) = \frac{1}{n}\sum_{1}^{n} I(x_i), \text{ where } I(x) = \begin{cases} 1, & \text{if } x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Furthermore, confidence intervals can also be constructed by simply taking quantiles of the samples.

Code/Method

The modelling, analysis and presentation of the data was performed in R. The vector autoregression modelling was performed using the *"vars"* package. However, for the autoregression no package existed that performed: seasonal/non-linear predictors, k-fold cross validation, probabilistic forecasting/confidence intervals, skill score calculations in an efficient way. Hence these elements were all coded manually as follows (Appendix *code 1*):

First, the predictor values are defined in a single string to avoid the manually entering them each step. Given that every predictor value is lagged the data becomes filled with "NA" values so these are removed. Next a linear model is defined using the predictor values and then 10-fold cross validation is performed using the *caret* package. The model fit metrics and the final fit are then extracted and put into a table for assessment. Using the final model values are then predicted and samples from the $N(0, \sigma^2)$ are drawn which in turn allows for the creation of the confidence intervals and the state probabilities. As a result the prediction(with 90% confidence intervals) and state probability plots are then created. Finally the skill score is calculated by predicting from every single value in the time series and calculating the standard deviation of the error at each prediction forward, this piece of code is very time consuming as $\approx$ 1800 values are found for each prediction. Similar results were were obtained effectively using smaller sample sizes of 250. However, the skill scores will be less accurate thus reducing validity of conclusions. This is an element of the code that needs further optimization. The skill score plot is then created and all plots are combined into one single output.

Data Plots

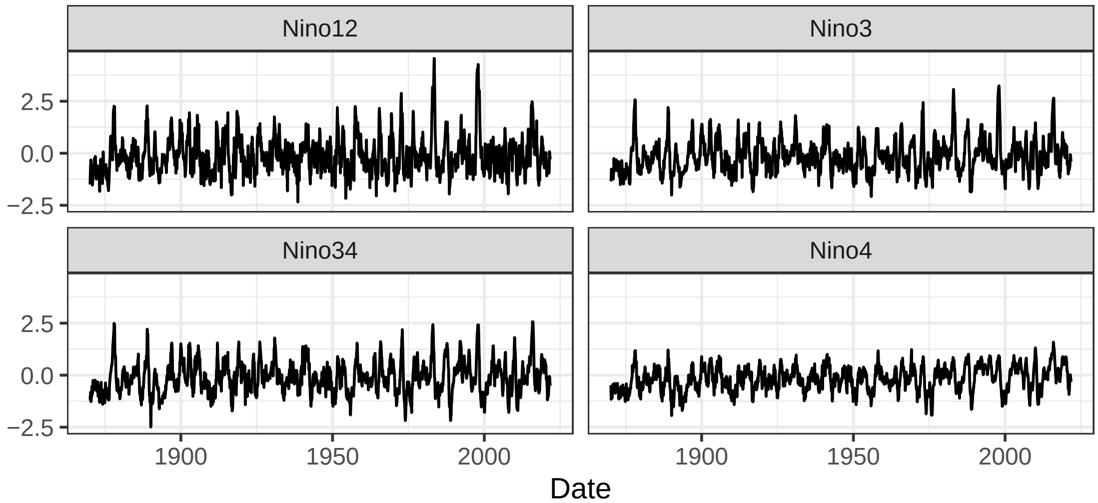

Figure 2: Sea Surface Temperature Anomolies - All data

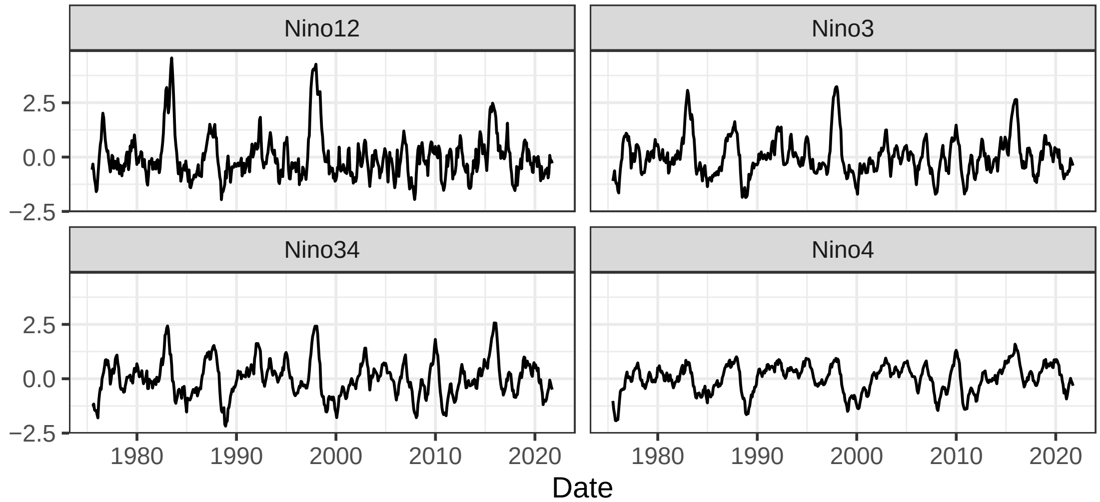

Figure 3: Sea Surface Temperature Anomolies - Recent data

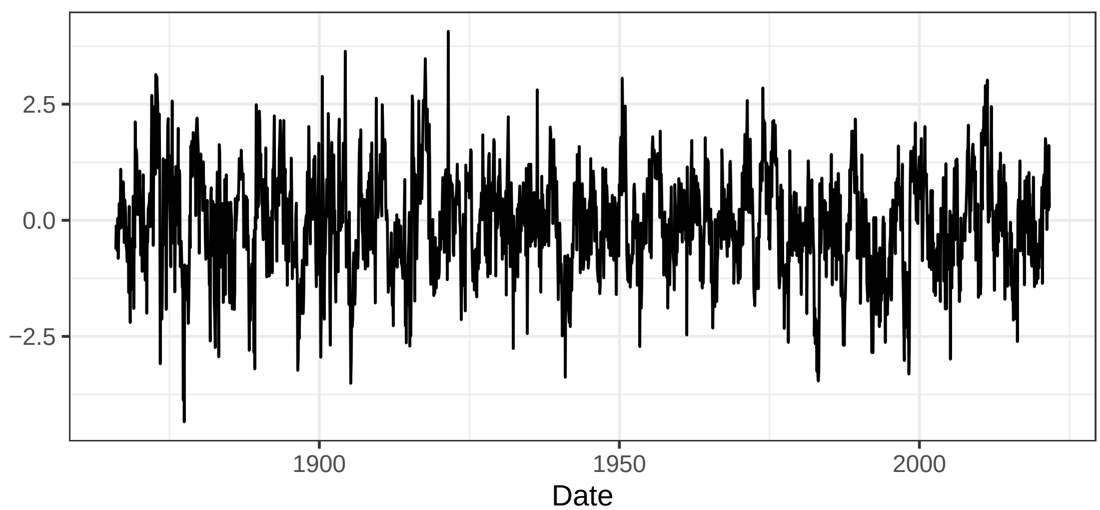

Figure 4: Southern Oscillation Index - All data

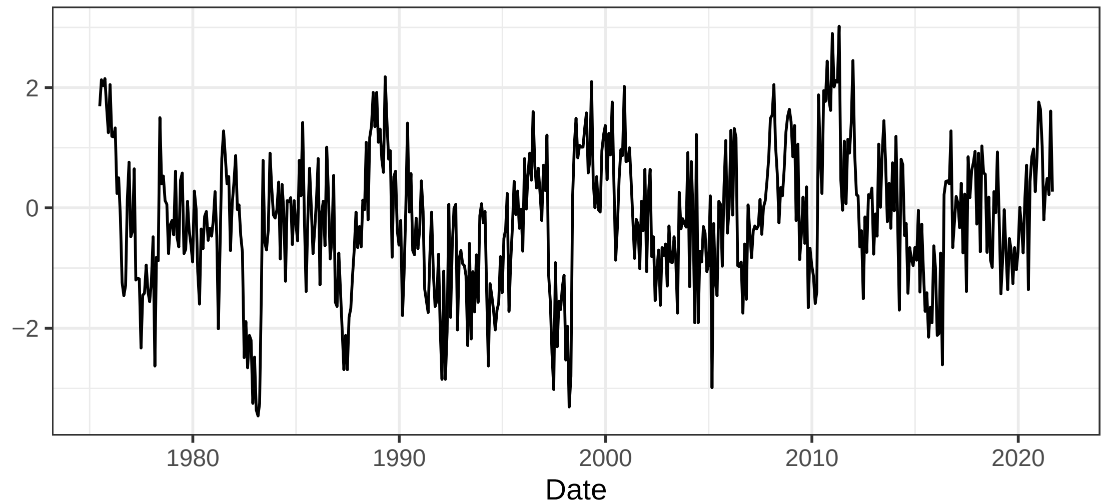

Figure 5: Southern Oscillation Index - Recent data

As described in the introduction there are 4 Sea Surface Temperature (SST) Anomaly time series: Nino12, Nino3, Nino34 and Nino 4. They are anomolies as the data has had its mean removed. Furthermore, the Southern Oscillation index (SOI) which are pressure differences between two locations and therefore are not "anomolies." In both cases a 3 month moving average (MA) is taken and this is what will be used as the predictors in the models.

Analysis

Standard AR(P)

Starting with using standard autoregression, the values of P are calculated with the following PACF plots.

Fitting the models using the specified p yields the following results.(Note full results only shown for Nino34 and SOI models, the others are just summaries)

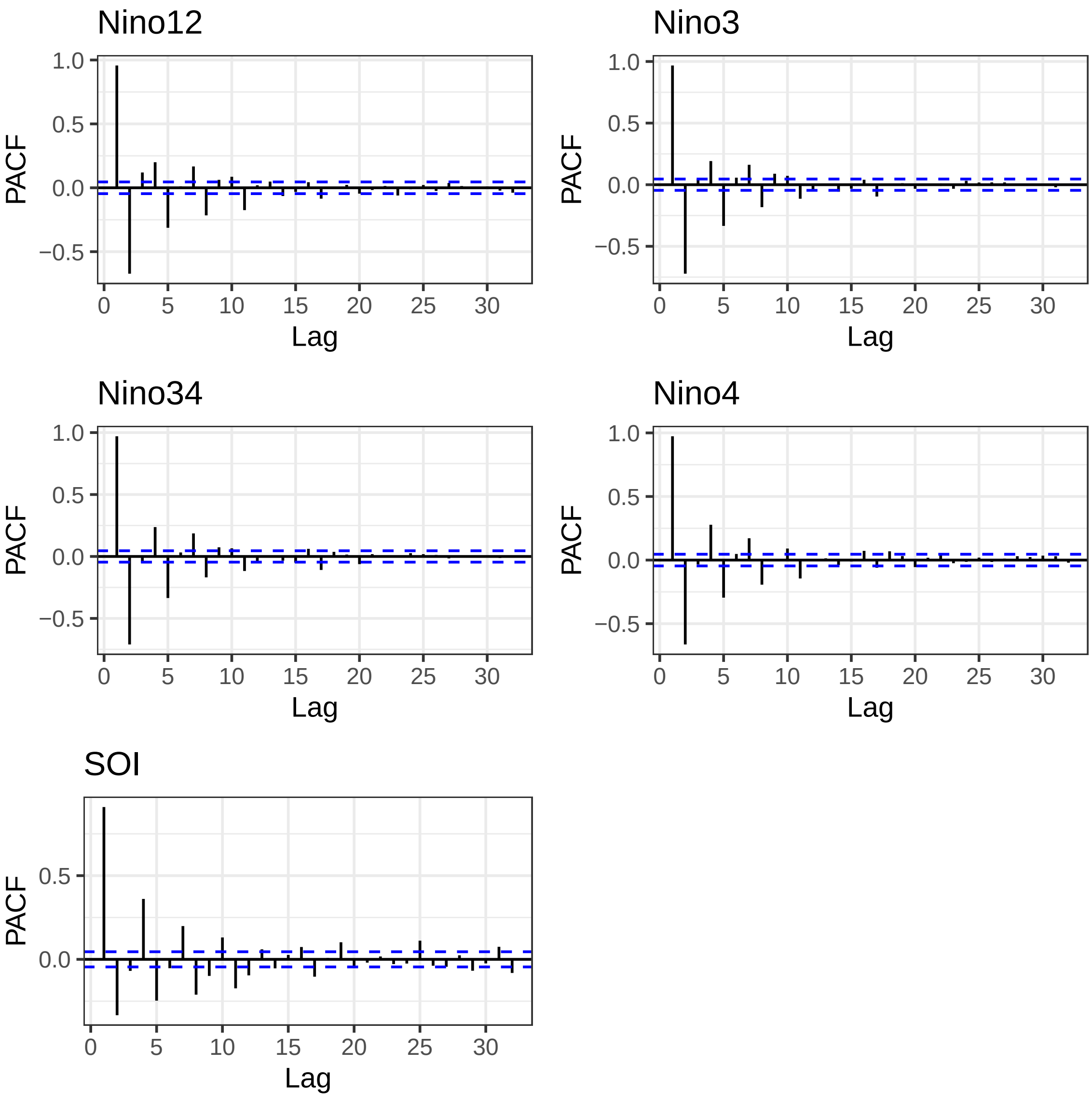

Figure 6: PACF plots - All Data

	RMSE	RMSESD	Climatology	Persistence
1	0.302098	0.0193505	0.913278	0.955456

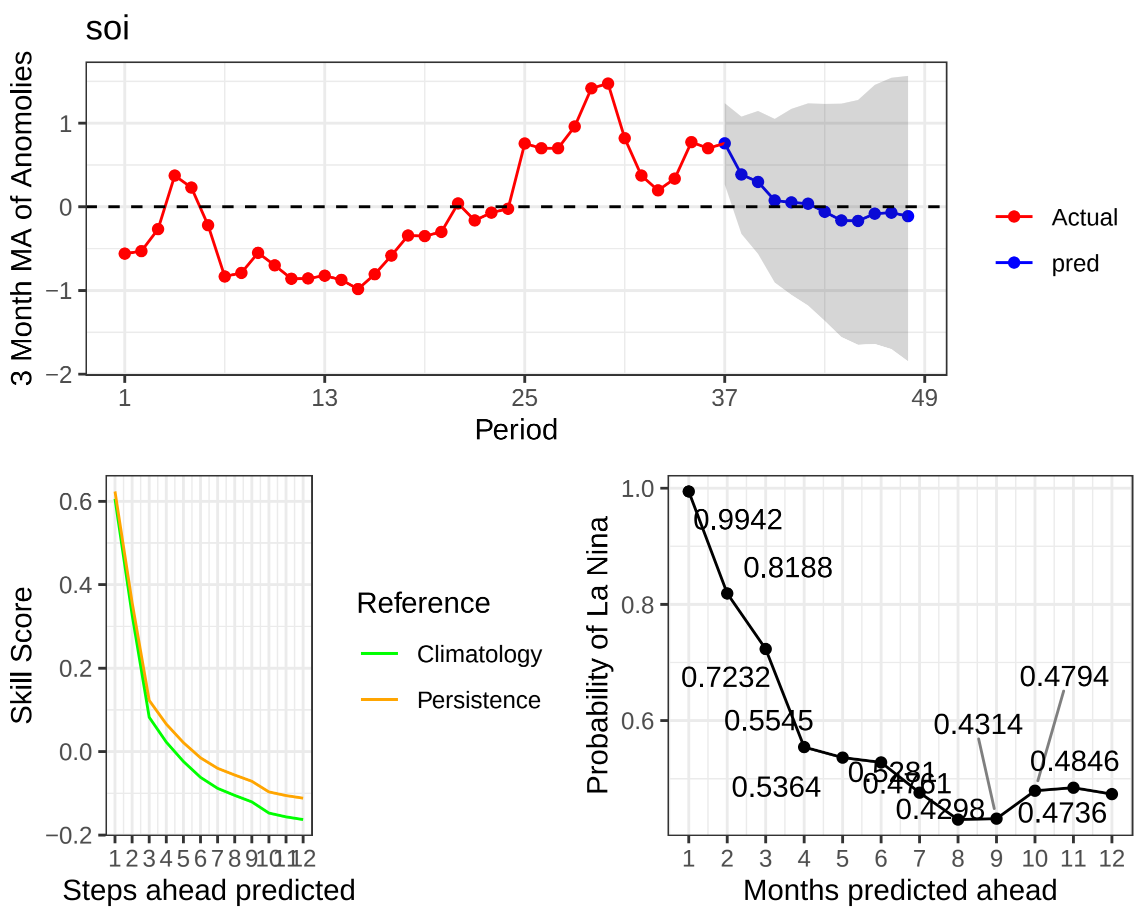

Figure 7: SOI AR(32) Model - Full Results

	RMSE	RMSESD	Climatology	Persistence
1	0.100363	0.00438106	0.553659	0.744275

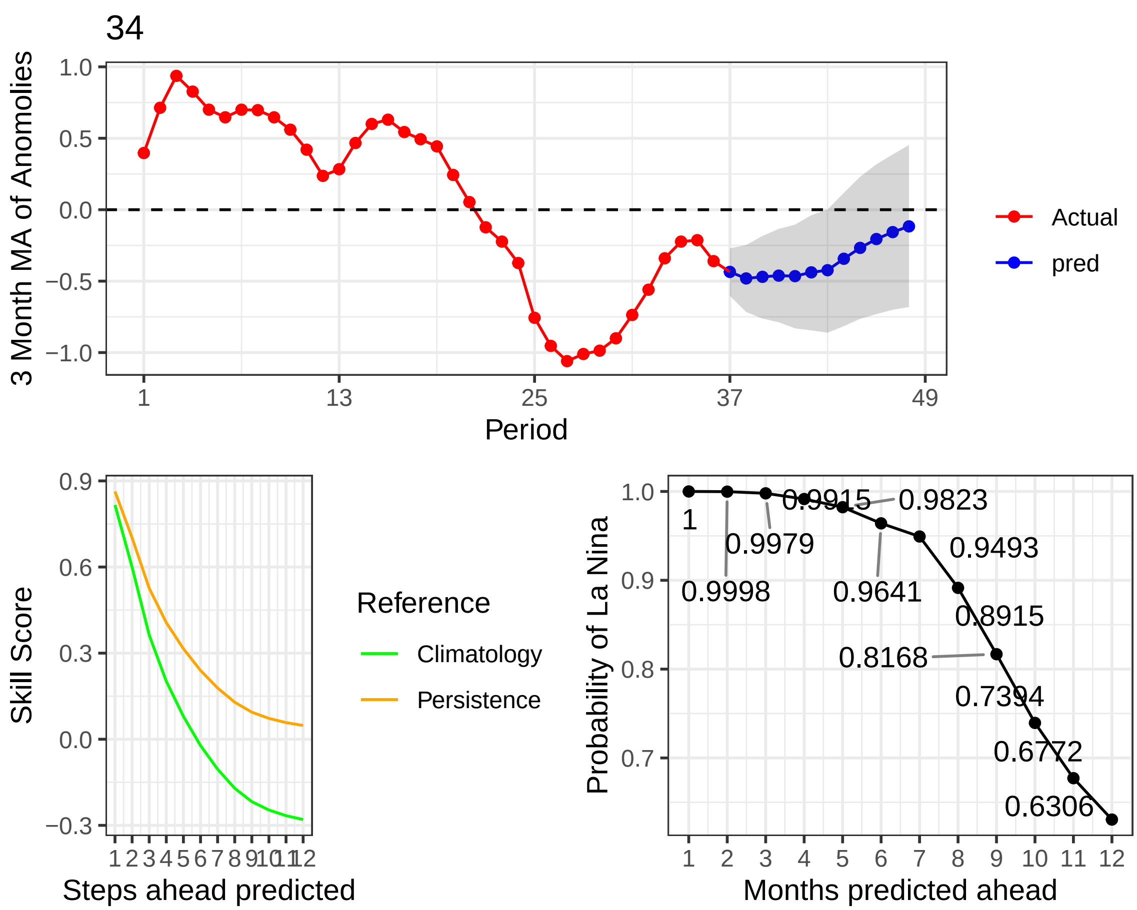

Figure 8: Nino34 AR(20) Model - Full Results

	RMSE	RMSESD	Climatology	Persistence
1	0.150406	0.00531384	0.727257	0.852673

Figure 9: Nino12 AR(23) Model - Model Results

	RMSE	RMSESD	Climatology	Persistence
1	0.10704	0.0074327	0.584007	0.764378

Figure 10: Nino3 AR(17) Model - Model Results

	RMSE	RMSESD	Climatology	Persistence
1	0.0781702	0.00398317	0.313199	0.559673

Figure 11: Nino4 AR(22) Model - Model Results

The results of the fits look good with all the models beating both *Climatology* and *Persistence* by a significant amount which means we do not reject any of the fits. The plots of the predictions for the SOI and Nino34 model show approximately the same trend - a continuation of an La Nina state for the next 3-4 of months but after a switch to La Nina becomes more probable. Although the model of the Nino34 is predicting with much more certainty that La Nina will remain. For example, at month 6 the SOI predicts a 52% chance compared to a 97% chance. The SOI model is concordant with the predictions set out by the National Climate Prediction Centre (Centre 2021), which states "*La Niña is likely to continue through the Northern Hemisphere winter 2021-22 (~90% chance) and into spring 2022 (~50% chance during March-May).*" Perhaps it would be plausible to reject the Nino34 model based on its large discrepancy between consensus forecasts. However, Nino34 specific forecasts (School 2021) are extremely similar to that of Nino34 model, with their next 3 future predictions being $0.99, 0.99, 0.92$.

It appears that the Nino34 model is more accurate for short term predictions than the SOI model with the RMSE being $\approx 15\%$ of the average reference forecast compared with $\approx 30\%$ for the SOI model. In addition, the skill scores show that the Nino34 model remains skillful (Skill Score > 0) for a lot longer, up to about 6 iterations forward compared to 3 for the SOI model. Furthermore, we see that the skill of the Nino34 model starts off at ≈ 0.75 whereas the SOI model was ≈ 0.68. This suggests that the use of the SOI vs Nino34 may be based on whether predictions are needed for the short term (3-4 months) or the longer term (4-10 months). The skill of the Nino34 model is not atypical (Jieshun Zhu1 and Towers 2015) and (Timothy N. Stockdale 2011) with skill scores generally less than 6-7 months. Furthermore, (Xu and Storch 1990) report an ARIMA(1,7,1) (Autoregression integrated moving average) with a skill of 12 iterations for the SOI, suggesting the SOI model may be lacking in long term skill. This could be attributed to larger variation in the data . This further supports the notion of using the two indices for predictions over a different time periods.

Looking at the results of the other fits none of them stand out as being better than the other. The results of the RMSE relative to the reference forecasts seem to mimic that of the Nino34 model. Furthermore, analysis of the the skill score, prediction and probability plots (Not Shown) show almost identical results. This would be expected as Nino34 is simply a mixture of Nino3 and Nino4 and therefore is highly correlated. The Nino12 has a very small area of measurement and therefore is not often used on its own for prediction. Furthermore, (Kevin E. Trenberth 2001) finds that the Nino34 index hold the most predictive power. Indicating that it is probably futile in continuing to predict on the other indices(12,3,4) in further models.

Seasonal Predictors

Fitting a standard seasonal model to both the SOI and Nino34, $x_{n+1} = c + \beta_1 x_n + \beta_2 x_n Cos(\frac{2\pi k_n}{12}) + \beta_3 x_n Sin(\frac{2\pi k_n}{12})$ yields the following summaries:

	RMSE	RMSESD	Climatology	Persistence
1	0.393495	0.0234387	0.903834	0.95091

Figure 12: SOI Simple Sesonal Model - Model Results

	RMSE	RMSESD	Climatology	Persistence
1	0.160046	0.00915863	0.552228	0.743059

Figure 13: Nino34 Simple Sesonal Model - Model Results

Neither models outperforms that of the standard autoregressive ones as we see an increase in RMSE and RMSESD in both cases. Despite this, it may not be evidence to reject that seasonality may exist as both models to outperform the reference metrics. To further investigate, adding the seasonal terms onto the end of the existing AR(P) model yields the following models.

$$SOIx_{n+1} = c + \sum_{i=1}^{32} \beta_i x_{n-i} + \beta_{33} x_n Cos(\frac{2\pi k_n}{12}) + \beta_{34} x_n Sin(\frac{2\pi k_n}{12})$$

$$Nino34x_{n+1} = c + \sum_{i=1}^{20} \beta_i x_{n-i} + \beta_{21} x_n Cos(\frac{2\pi k_n}{12}) + \beta_{22} x_n Sin(\frac{2\pi k_n}{12})$$

	RMSE	RMSESD	Climatology	Persistence
1	0.301028	0.0233139	0.913278	0.955456

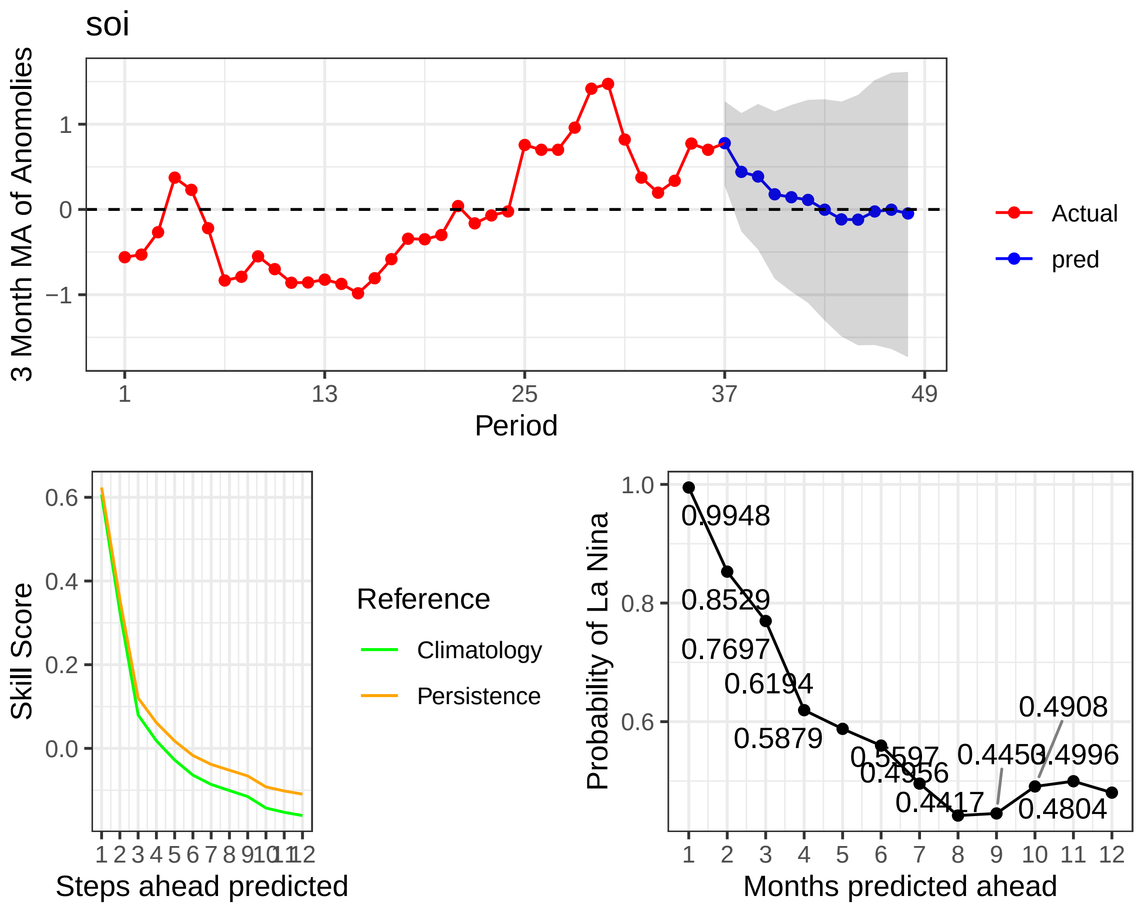

Figure 14: SOI AR(32)+Seasonal predictors Model - Full Results

	RMSE	RMSESD	Climatology	Persistence
1	0.0964884	0.00501118	0.553659	0.744275

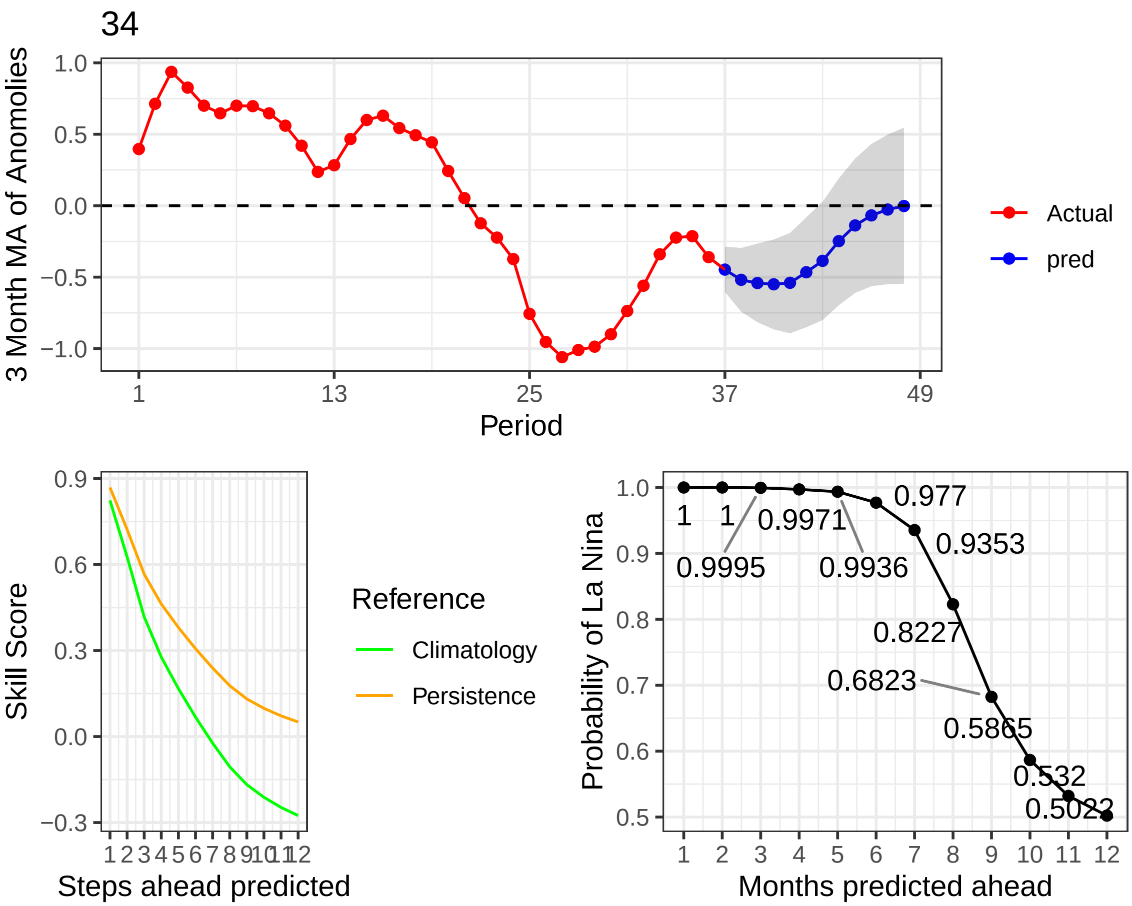

Figure 15: Nino34 AR(20)+Seasonal predictors Model - Full Results

Immediately, it is obvious that the SOI model is virtually unchanged. There is no obvious improvement in RMSE or skill nor vastly different predictions. This would suggest that there is little evidence of a seasonal cycle in the index. On the contrary, both (Ping Chang 1994) and ("MEI.v2" 2021) suggest that ENSO is seasonal perhaps implying that the model is poor. However, it is worth noting that these models are vastly more complex (e.g coupled ocean-atmosphere model) as well as using different data. For example, they include predictors such as surface winds and outgoing long-wave radiation (OLR). It is likely to be the case that the seasonality is simply not evident in the simple SOI data which would explain why the model doesn't improve.

The results from the Nino34 model seems to have improved. The RMSE has increased slightly along with Skill by $\approx$ 1 lead. As a result the predictions become more certain which can be seen in the probability of La Nina predictions which hug 1 for the first 6 predictions. Indicating that very few of the 10,000 predictions are greater than 0. The probability results appear very high on inspection but there is no reason to reject the model given the improvement in the metrics. However, an increase in sample size of the random numbers may give lower probabilities. The Skill Score does remain $\approx$ 3 months more than the SOI model, providing further evidence that longer term predictions may be more effective with the Nino34. An interesting discussion arises as to why seasonality is visible in Nino34 but not the SOI when they are describing the same phenomenon? Overall, the updated Nino34 model has better prediction metrics and is concordant with the literature that ENSO has seasonal properties but may not consider future variance sufficiently.

Non-Linear Predictors

The search for non-linear predictors is a manual one. Using a looped version of code 1 models containing combinations of linear predictors(Appendix Non-Linear Models) with quadratic and cubic terms were fitted to each data set. **All** models were rejected as they didn't beat the the reference forecasts. Visually this is obvious as there is no non-linear relationship between predictors. Although, autoregressive conditional heteroscedasticity (ARCH) models also known as non-linear stochastic models have been fitted successfully to the SOI (J. H. Ahn 2005), the evidence is quite clear that using non-linear predictors is ineffective in the case of standard linear models.

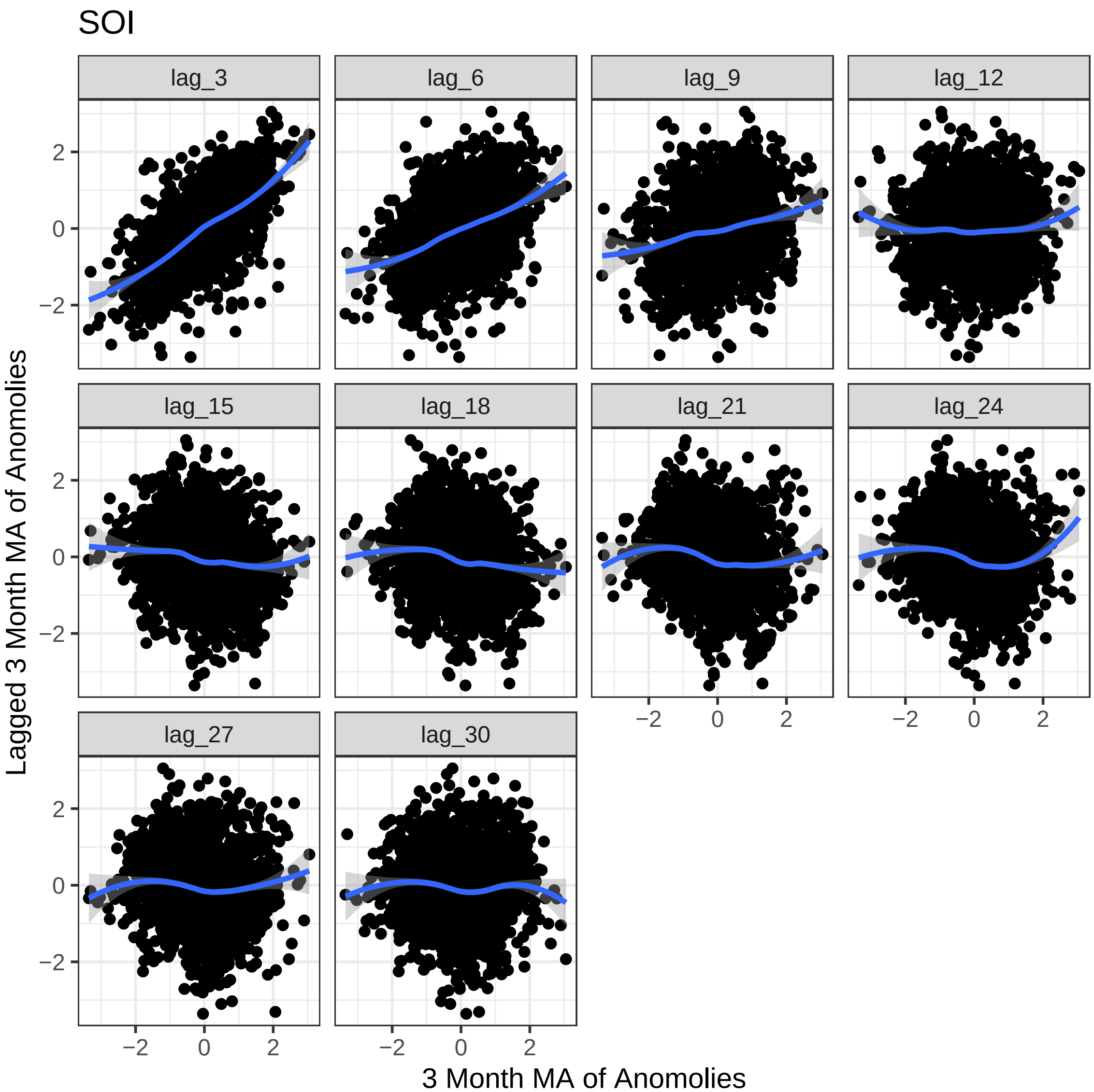

Figure 16: SOI Lagged Values

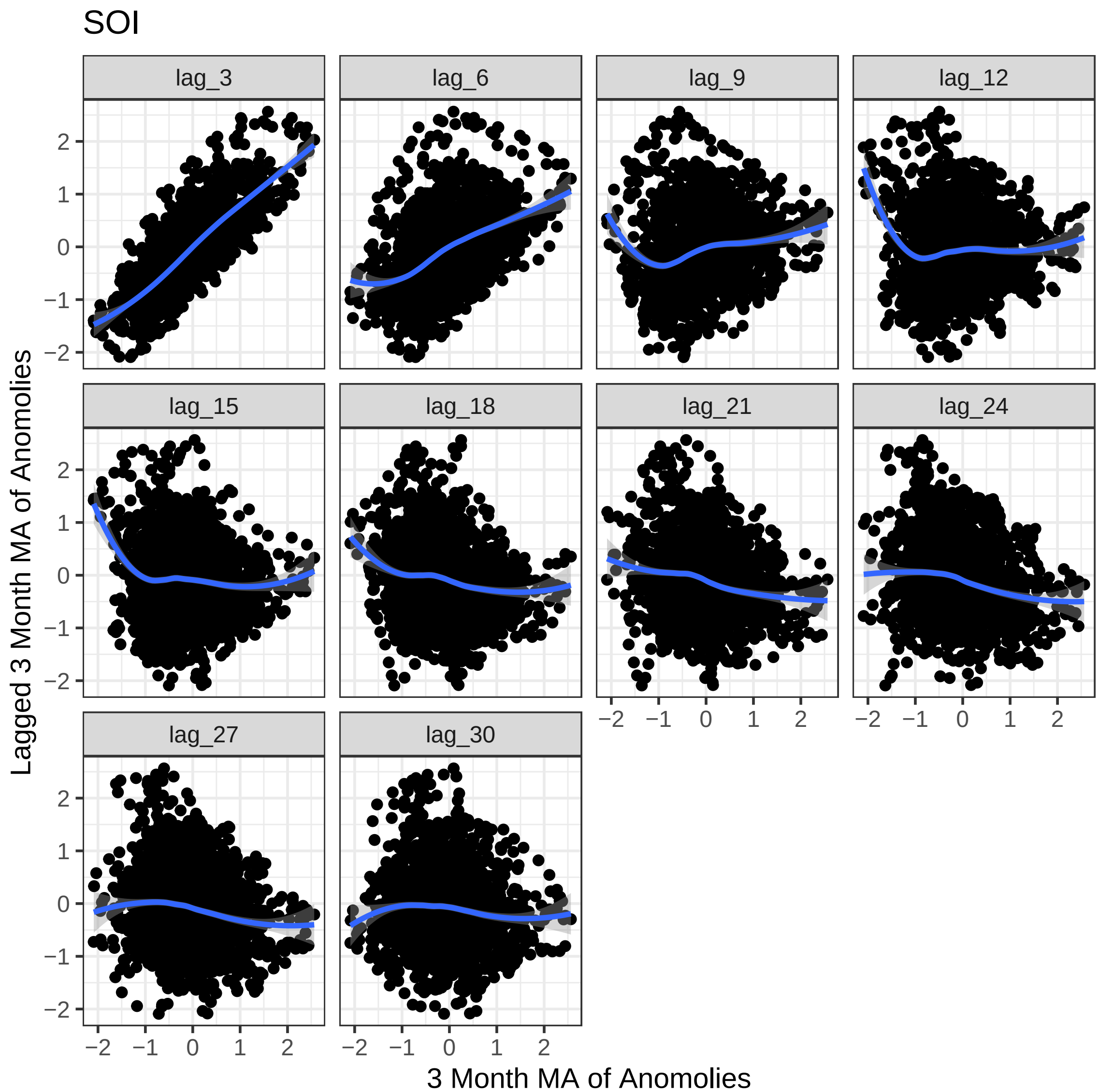

Figure 17: Nino34 Lagged Values

Vector Autoregression

Work using vector autoregression for the use of predicting SST anomalies has been widely published much more so than that of standard autoregression (DAVID CHAPMAN 2015),(Miftahuddin 2019). For that reason an in depth discussion of it is not productive. However,a prediction of the Nino3 and Nino4 indices via autoregression may help validate the result achieved by autoregression on the Nino34. Using the *vars* package in R it is found the both the AIC and BIC for the multivariate time series is 10. A standard fit (No cross validation and skill score) yields the following prediction.

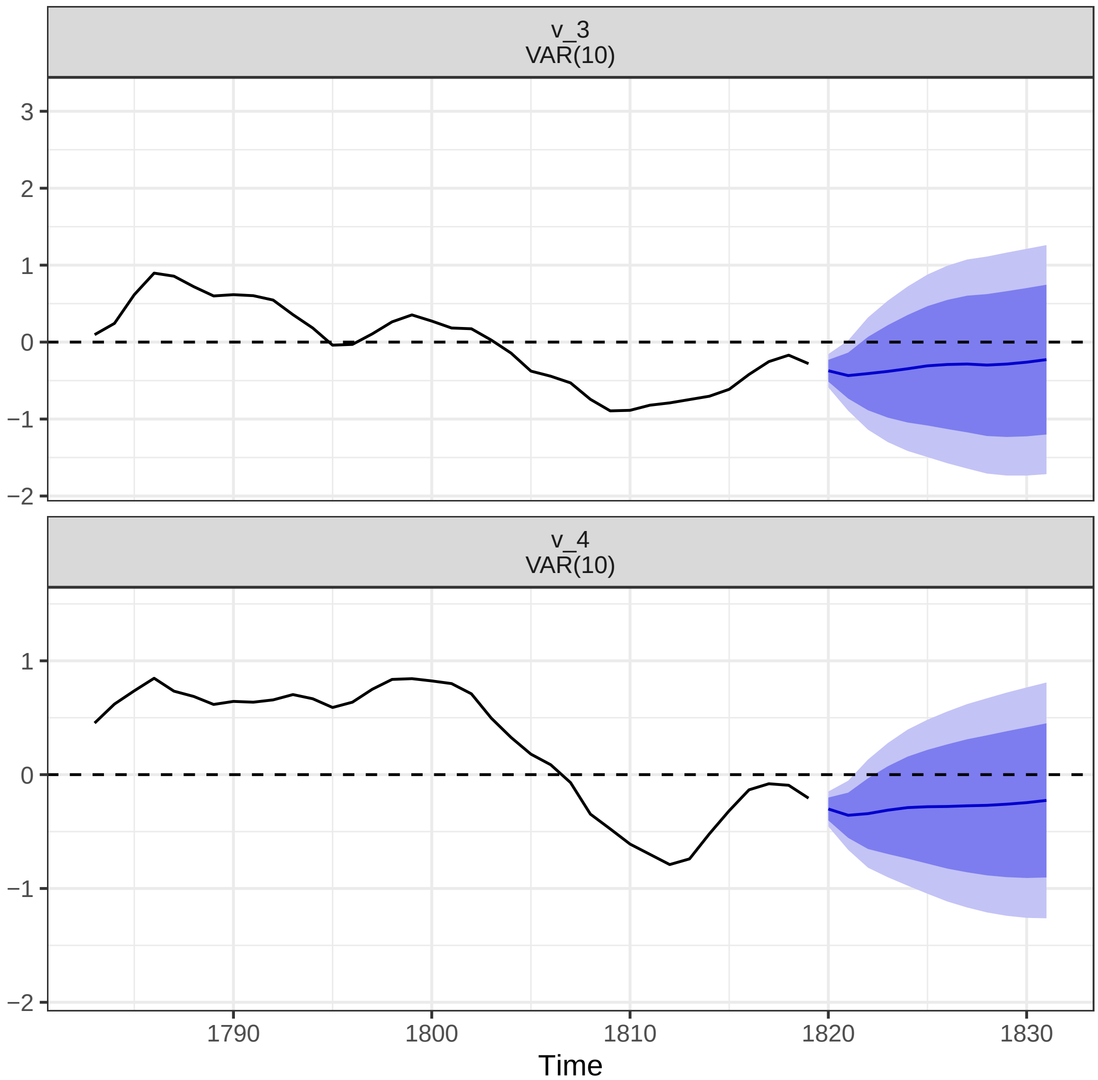

Figure 18: Nino34 AR(20)+Seasonal predictors Model - Full Results

The results are promsing with the predicted values matching that of the Nino34 seasonal model. There is great certainty that the La Nina state will remain in the next 3 months but after that a steady trend towards the neutral/El Nino phase. The question arises of whether the inclusion of seasonal terms in such a model will have the same impact as it did on the single Nino34 model.

Conclusions

The standard autoregression models in both cases yields very good error metrics relative to the references. Furthermore, the skill scores are in line with that of other models. Although, some more modern bayesian models may offer double the skill (Nan Chen 2021) however, this is not a reasonable comparison. The predictions of the SOI are concordant with that of the *Climate Prediction Centre*(Centre 2021), however the Nino34 model predictions may be a little high. There is no evidence of a seasonal cycle in the SOI however a seasonal model on Nino34 increases the skill score and decreases the RMSE therefore showing some evidence of seasonality. The predictions of the model may also be too high but are roughly in line with that of the *Columbia Climate School*(School 2021). The difference in skill between the models of the two indices was apparent in both models supporting the notion that longer term predictions should be made using the Nino34. A search for non-linear predictors was not successful and was expected given the success of the standard autoregression models. Furthermore, VAR made predictions of the Nino3 and Nino4 indices separately which were concordant with AR predictions on the Nino34.

Using the SOI AR(32) model and the Nino34 AR(20) with seasonal terms the following predictions are made:

- La Nina state is likely (70%<) to continue for the next 3 months on the SOI, in 6 months 50% chance of being in La Nina state
- La Nina State is highly likely (95%) to continue for the next 6 months on the Nino34, decreasing to 50% at 12 months

Appendix

Code 1

```r
# Model Fitting and evaluation
data_list <- list(nino_12_model_data, nino_3_model_data, nino_34_model_data,
    nino_4_model_data, soi_model_data)
name_list <- c("12", "3", "34", "4", "soi")

# Write mutation for variables

expr <- map(c(x1 = "lag(y_ma3)", s1 = "lag(y_ma3*cos(2*pi*Month/12))",
    s2 = "lag(y_ma3*sin(2*pi*Month/12))"), rlang::parse_expr)


i <- 1
mod_data <- data_list[[i]]
name <- name_list[[i]]

# Applying Data transformation, write *formula* in here
mod_data <- mod_data %>%
    mutate(!!!expr)
```

```r
# Completing Data Set by removal of NA's
mod_data <- mod_data[complete.cases(mod_data), ]


# Reference forecasts
climatology <- var(mod_data$y_ma3)
persistence <- sd(mod_data$y_ma3[1] - mod_data$y_ma3[-1])


# Model fitting
train <- trainControl(method = "cv", number = 10)
# Add *variables* created into formula
model_fit <- train(y_ma3 ~ x1 + s1 + s2, data = mod_data, method = "lm",
    trControl = train)
# Storing Model
assign(paste0("model", sep = "_", name), model_fit)

# Skill Scores
rmse <- model_fit$results["RMSE"]



# Creation of data frame
df <- data.frame(Data = name, `Model Formula` = as.character(model_fit$finalModel)[1],
    RMSE = model_fit$results["RMSE"], Rsqaured = model_fit$results["Rsquared"],
    MAE = model_fit$results["MAE"], RMSESD = model_fit$results["RMSESD"],
    RsquaredSD = model_fit$results["RsquaredSD"], MAESD = model_fit$results["MAESD"],
    Climatology = climatology, Persistence = persistence)


mod_results <- df %>%
    dplyr::select(RMSE, RMSESD, Climatology, Persistence)
mod_results <- mod_results %>%
    mutate(RMSE = signif(RMSE, 6), RMSESD = signif(RMSESD, 6),
        Climatology = signif(Climatology, 6), Persistence = signif(Persistence,
            6))



# Prediction Extract residual standard error from model fit
sigma2 <- sqrt(deviance(model_fit$finalModel)/df.residual(model_fit$finalModel))

# Applying predictions Prediction Length
pred_length <- 12
for (j in 1:pred_length) {
    # Setting Data Frame for each loop
    ifelse({
```

```
        j == 1
    }, {
        mod_pred_data <- mod_data %>%
            mutate(type = "Actual")
    }, {
        mod_pred_data <- mod_pred_data
    })
    # Adding Predictions Extending data, including months
    # add data mutation applied earlier
    data_extension <- add_row(mod_pred_data) %>%
        mutate(!!!expr, Month = lag(Month)%%12 + 1)
    # Switch out 'var' for the y variable your predicting
    start <- add_predictions(slice(data_extension, nrow(data_extension)),
        model_fit$finalMode, var = "y_ma3")
    start <- start %>%
        mutate(type = "pred")
    mod_pred_data <- rbind(mod_pred_data, start)
}


mod_pred_data <- mod_pred_data %>%
    mutate(data = name)


# Propagation of error term for confidence intervals Number
# of samples
sample_length <- 10000
# Matrix Creation
output <- matrix(ncol = pred_length, nrow = sample_length)
# Adding Random Numbers from residual distribution
for (k in 1:pred_length) {
    output[, k] <- rnorm(sample_length, 0, sigma2)
}
output <- data.frame(output)
# Confidence Interval Defining the quantiles
upper_q <- 0.95
lower_q <- 0.05
preds <- (filter(mod_pred_data, type == "pred"))$y_ma3
# Summing over each iteration
for (l in 1:pred_length) {
    pred <- preds[l]
    output_sample <- output[c(1:l)]
    vec <- rowSums(output_sample)
    upper <- unname(quantile(vec, c(upper_q))) + pred
    lower <- unname(quantile(vec, c(lower_q))) + pred
    n <- l + 36
    type <- "pred"
```

```
    df_r <- data.frame(n, upper, lower, pred)
    ifelse({
        l == 1
    }, {
        conf_df <- df_r
    }, {
        conf_df <- rbind(conf_df, df_r)
    })
}



# Extraction of probabilistic forecast

conf_matrix <- matrix(ncol = pred_length, nrow = sample_length)

for (m in 1:pred_length) {
    pred_2 <- preds[m]
    output_sample_1 <- output[c(1:m)]
    samples_sum <- rowSums(output_sample_1)
    pred_samples <- samples_sum + pred_2
    conf_matrix[, m] <- pred_samples

}
conf_matrix <- data.frame(conf_matrix)



conf_df <- conf_df %>%
    mutate(data = name)




pred_plot_colours <- c(Actual = "red", pred = "blue")
prediction_plot <- ggplot() + geom_point(tail(mod_pred_data,
    pred_length + 36), mapping = aes(x = seq(1, pred_length +
    36, 1), y = y_ma3, colour = type, group = 1)) + geom_line(tail(mod_pred_data,
    pred_length + 36), mapping = aes(x = seq(1, pred_length +
    36, 1), y = y_ma3, colour = type, group = 1)) + geom_ribbon(conf_df,
    mapping = aes(x = n, ymin = lower, ymax = upper), alpha = 0.2) +
    geom_hline(yintercept = 0, colour = "black", lty = 2) + labs(x = "Period",
    title = name, y = "3 Month MA of Anomolies") + scale_x_continuous(breaks = seq(1,
    49, 12)) + scale_colour_manual(values = pred_plot_colours,
    name = "") + theme_bw()
```

```
prob_calc_data <- conf_matrix
for (i in 1:pred_length) {
    v <- prob_calc_data[i]
    p <- 1 - sum(v <= 0)/sample_length
    df <- data.frame(i, p)
    ifelse({
        i == 1
    }, {
        probs_data <- df
    }, {
        probs_data <- rbind(probs_data, df)
    })
}
prob_plot <- ggplot(probs_data, mapping = aes(x = i, y = 1 -
    p, label = 1 - p)) + geom_point() + geom_line() + geom_text_repel(box.padding = 0.35,
    point.padding = 0.5, segment.color = "grey50") + scale_x_continuous(breaks = seq(1,
    pred_length, 1)) + theme_bw() + labs(x = "Months predicted ahead",
    y = "Probability of La Nina")
```

```
# Skill Score

sub_data_length <- 70
data_length <- nrow(mod_data)
# iterations <- data_length-sub_data_length-2
iterations <- 20
for (i in 0:iterations) {
    # Setting start and end points
    start <- i + 1
    end <- sub_data_length + i
    # Slicing main data set
    main_data <- slice(mod_data, c(start:end))
    # Creating predicting and reference data
    predicting <- slice(main_data, c(1:(sub_data_length/2)))
    reference <- slice(main_data, c((sub_data_length/2 + 1):sub_data_length))
    # Actual Data values
    actual <- reference$y_ma3
    # Predicting Prediction Length
    pred_length_2 <- sub_data_length/2
    for (j in 1:(pred_length_2)) {
        # Setting Data Frame for each loop
        ifelse({
            j == 1
```

```
        }, {
            mod_pred_data_2 <- predicting %>%
                mutate(type = "Actual")
        }, {
            mod_pred_data_2 <- mod_pred_data_2
        })
        # Adding Predictions Extending data, including
        # months add data mutation applied earlier
        data_extension <- add_row(mod_pred_data_2) %>%
            mutate(!!!expr, Month = lag(Month)%%12 + 1)
        # Switch out 'var' for the y variable your
        # predicting
        start <- add_predictions(slice(data_extension, nrow(data_extension)),
            model_fit$finalMode, var = "y_ma3")
        start <- start %>%
            mutate(type = "pred")
        mod_pred_data_2 <- rbind(mod_pred_data_2, start)
    }
    # predicted Values
    predicted <- filter(mod_pred_data_2, type == "pred")$y_ma3
    df <- data.frame(predicted, actual, n = seq(1, (sub_data_length/2),
        1))
    ifelse({
        i == 0
    }, {
        skill_df <- df
    }, {
        skill_df <- rbind(skill_df, df)
    })
}

skill_df <- skill_df %>%
    mutate(error = predicted - actual)

skill_df_final <- skill_df %>%
    group_by(n) %>%
    summarise_each(funs(mean, sd, se = sd(.)/sqrt(n())))

skill_df_final <- skill_df_final %>%
    mutate(ss_c = 1 - error_sd/climatology, ss_p = 1 - error_sd/persistence)


ss_plot_colours <- c(Climatology = "green", Persistence = "orange")

skill_score_plot <- ggplot(skill_df_final, mapping = aes(x = n)) +
    geom_line(mapping = aes(y = ss_c, colour = "Climatology")) +
    geom_line(mapping = aes(y = ss_p, colour = "Persistence")) +
```

```
    labs(x = "Steps ahead predicted", y = "Skill Score") + scale_colour_manual(values = ss_plot
    name = "") + theme_bw()


layout_matrix <- rbind(c(1, 1, 1, 1), c(2, 2, 2, 2), c(2, 2,
    2, 2), c(3, 3, 4, 4), c(3, 3, 4, 4))

grid.arrange(tableGrob(mod_results), prediction_plot, skill_score_plot,
    prob_plot, layout_matrix = layout_matrix)
```

Code 2 - Data Import

```
# Nino12
nino12_table <- read.table("~/nino12.data.txt", quote = "\"")
colnames(nino12_table) <- (c("Year", seq(1, 12, 1)))
for (i in 1:nrow(nino12_table)) {
    row <- slice(nino12_table, i)
    year <- row$Year
    val <- row[c(2:13)]
    val <- as.numeric(val)
    df <- data.frame(Year = rep(year, 12), Month = seq(1, 12,
        1), y = val)
    ifelse({
        i == 1
    }, {
        data <- df
    }, {
        data <- rbind(data, df)
    })
}
nino_12_model_data <- filter(data, y != -99.99)  #Removal of missing values
nino_12_model_data <- nino_12_model_data %>%
    mutate(type = "Nino12", y_ma3 = rollmean(y, 3, na.pad = TRUE,
        align = "right"))

# Nino3
nino3_table <- read.table("~/nino3.data.txt", quote = "\"")
colnames(nino3_table) <- (c("Year", seq(1, 12, 1)))
for (i in 1:nrow(nino3_table)) {
    row <- slice(nino3_table, i)
    year <- row$Year
    val <- row[c(2:13)]
    val <- as.numeric(val)
    df <- data.frame(Year = rep(year, 12), Month = seq(1, 12,
        1), y = val)
```

```r
    ifelse({
        i == 1
    }, {
        data <- df
    }, {
        data <- rbind(data, df)
    })
}
nino_3_model_data <- filter(data, y != -99.99)  #Removal of missing values
nino_3_model_data <- nino_3_model_data %>%
    mutate(type = "Nino3", y_ma3 = rollmean(y, 3, na.pad = TRUE,
        align = "right"))

# Nino4
nino4_table <- read.table("~/nino4.data.txt", quote = "\"")
colnames(nino4_table) <- (c("Year", seq(1, 12, 1)))
for (i in 1:nrow(nino4_table)) {
    row <- slice(nino4_table, i)
    year <- row$Year
    val <- row[c(2:13)]
    val <- as.numeric(val)
    df <- data.frame(Year = rep(year, 12), Month = seq(1, 12,
        1), y = val)
    ifelse({
        i == 1
    }, {
        data <- df
    }, {
        data <- rbind(data, df)
    })
}
nino_4_model_data <- filter(data, y != -99.99)  #Removal of missing values
nino_4_model_data <- nino_4_model_data %>%
    mutate(type = "Nino4", y_ma3 = rollmean(y, 3, na.pad = TRUE,
        align = "right"))

# Nino34
nino34_table <- read.table("~/nino34.data.txt", quote = "\"")
colnames(nino34_table) <- (c("Year", seq(1, 12, 1)))
for (i in 1:nrow(nino34_table)) {
    row <- slice(nino34_table, i)
    year <- row$Year
    val <- row[c(2:13)]
    val <- as.numeric(val)
    df <- data.frame(Year = rep(year, 12), Month = seq(1, 12,
        1), y = val)
    ifelse({
```

```
        i == 1
    }, {
        data <- df
    }, {
        data <- rbind(data, df)
    })
}
nino_34_model_data <- filter(data, y != -99.99)  #Removal of missing values
nino_34_model_data <- nino_34_model_data %>%
    mutate(type = "Nino34", y_ma3 = rollmean(y, 3, na.pad = TRUE,
        align = "right"))

# SOI
soi_table <- read.table("~/soi.data.txt", quote = "\"")
colnames(soi_table) <- (c("Year", seq(1, 12, 1)))
for (i in 1:nrow(soi_table)) {
    row <- slice(soi_table, i)
    year <- row$Year
    val <- row[c(2:13)]
    val <- as.numeric(val)
    df <- data.frame(Year = rep(year, 12), Month = seq(1, 12,
        1), y = val)
    ifelse({
        i == 1
    }, {
        data <- df
    }, {
        data <- rbind(data, df)
    })
}
soi_model_data <- filter(data, y != -99.99)  #Removal of missing values
soi_model_data <- soi_model_data %>%
    mutate(type = "SOI", y_ma3 = rollmean(y, 3, na.pad = TRUE,
        align = "right"))
```

Non-Linear Models Attempted

Formula
$x_{n+1} = \beta_1 x_{n-3}^2$
$x_{n+1} = \beta_1 x_{n-3}^3$
$x_{n+1} = \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^2$
$x_{n+1} = \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^3$
$x_{n+1} = \beta_1 x_{n-3}^3 + \beta_2 x_{n-6}^3$
$x_{n+1} = \beta_1 x_{n-3}^3 + \beta_2 x_{n-6}^2$
$x_{n+1} = \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^2 + \beta_3 x_{n-9}^2$

Formula
$x_{n+1} = \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^3 + \beta_3 x_{n-9}^2$
$x_{n+1} = \beta_1 x_{n-3}^3 + \beta_2 x_{n-6}^3 + \beta_3 x_{n-9}^2$
$x_{n+1} = \beta_1 x_{n-3}^3 + \beta_2 x_{n-6}^2 + \beta_3 x_{n-9}^2$
$x_{n+1} = \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^2 + \beta_3 x_{n-9}^3$
$x_{n+1} = \beta_1 x_{n-3}^2 + \beta_2 x_{n-6}^3 + \beta_3 x_{n-9}^3$
$x_{n+1} = \beta_1 x_{n-3}^3 + \beta_2 x_{n-6}^3 + \beta_3 x_{n-9}^3$
$x_{n+1} = \beta_1 x_{n-3}^3 + \beta_2 x_{n-6}^2 + \beta_3 x_{n-9}^3$

References

NOAA, 2014, updated 2021. How ENSO leads to a cascade of global impacts. s.l., s.n.

NOAA, 2016. El Nino and La Nina: Frequently asked questions. s.l., s.n.

NOAA, updated 2021. El Nino/Southern Oscillation (ENSO). s.l., s.n.

Adeney, J. M., Ginsberg, J. R., Russell, G. J., & Kinnaird, M. F. (2006). Effects of an ENSO-related fire on birds of a lowland tropical forest in Sumatra. Animal Conservation, 9(3), 292–301. https://doi.org/10.1111/j.1469-1795.2006.00035.x

Anyamba, A., Small, J. L., Britch, S. C., Tucker, C. J., Pak, E. W., Reynolds, C. A., … Linthicum, K. J. (2014). Recent Weather Extremes and Impacts on Agricultural Production and Vector-Borne Disease Outbreak Patterns. PLoS ONE, 9(3), e92538. https://doi.org/10.1371/journal.pone.0092538

Bastos, A., Friedlingstein, P., Sitch, S., Chen, C., Mialon, A., Wigneron, J.-P., … Melton, J. (2018). Impact of the 2015/2016 El Nino on the terrestrial carbon cycle constrained by bottom-up and top-down approaches.

Carlowicz, M., & Schollaert Uz, S. (2017, February 14). El Nino: Pacific Wind and Current Changes Bring Warm, Wild Weather. Retrieved from Nasa.gov website: https://earthobservatory.nasa.gov/features/ElNino

Chen, C.-C., McCarl, B. A., & Adams, R. M. (2001). Economic Implications of Potential ENSO Frequency and Strength Shifts. Climatic Change, 49(1/2), 147–159. https://doi.org/10.1023/a:1010666107851

Glynn, P. W., & D'Croz, L. (1990). Experimental evidence for high temperature stress as the cause of El Nino-coincident coral mortality. Coral Reefs, 8(4), 181–191. https://doi.org/10.1007/bf00265009

Hoeke, R. K., McInnes, K. L., Kruger, J. C., McNaught, R. J., Hunter, J. R., & Smithers, S. G. (2013). Widespread inundation of Pacific islands triggered by distant-source wind-waves. Global and Planetary Change, 108, 128–138. https://doi.org/10.1016/j.gloplacha.2013.06.006

Huang, M., Wang, Z., Wang, S., Gu, F., Gong, H., Hao, M., & Shao, Y. (2019). Global vegetation productivity responses to the West Pacific Warm Pool. Science of the Total Environment, 655, 641–651. https://doi.org/10.1016/j.scitotenv.2018.11.170

Keil, A., Zeller, M., Wida, A., Sanim, B., & Birner, R. (2007). What determines farmers' resilience towards ENSO-related drought? An empirical assessment in Central Sulawesi, Indonesia. Climatic Change, 86(3-4), 291–307. https://doi.org/10.1007/s10584-007-9326-4

Keller, K. M., Joos, F., Lehner, F., & Raible, C. C. (2015). Detecting changes in marine responses to ENSO from 850 to 2100 C.E.: Insights from the ocean carbon cycle. Geophysical Research Letters, 42(2), 518–525. https://doi.org/10.1002/2014gl062398

Khandekar, M., Murty, T., Scott, D., & Baird, W. (2000). The 1997 EI Nino, Indonesian Forest Fires and the Malaysian Smoke Problem: A Deadly Combination of Natural and Man-Made Hazard . Natural Hazards, 21, 131–144.

Davey, M. K., et al., 2014. The probability of the impact of ENSO on precipitation and near-surface temperature

FAO, 2021

Glimate.Gov NOAA, 2021

IRI, 2021

Knox, P., et al., 2015. El Nino, La Nina and Climate Impacts on Agriculture- Southeastern U.S.

McPhaden, 2002. El Nino and La Nina - Causes and Global Consequences

Met Office, 2021

NASA, 2021

NOAA, 2021

Scaife, A. A., 2010. Impact of ENSO on European Climate

All data sourced from the National Oceanic and Atmospheric Administration : https://psl.noaa.gov/gcos_wgsp/Timeseries/

Andrew V. Metcalfe, Paul S. P. Cowpertwait. 2009. *Introductory Time Series with r (Use r!)*.

Centre, Climate Prediction. 2021. "ENSO Diagnostic Discussion." https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/enso_advisory/ensodisc.shtml.

DAVID CHAPMAN, NAOMI HENDERSON, MARK A. CANE. 2015. "A Vector Autoregressive ENSO Prediction Model."

"Forecast Skill." 2021. https://en.wikipedia.org/wiki/Forecast_skill.

Gavin C. Cawley, Nicola L. C. Talbot. 2010. "On over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation."

Hai, Z. 2005. "Unnamed ARIMA Paper." https://libres.uncg.edu/ir/uncw/f/zhai2005-2.pdf.

J. H. Ahn, H. S. Kim. 2005. "Nonlinear Modeling of El Nino/Southern Oscillation Index."

Jieshun Zhu1, Ben Cash1, Bohua Huang, and Peter Towers. 2015. "ENSO Prediction in Project Minerva: Sensitivity to Atmospheric Horizontal Resolution and Ensemble Size."

Kevin E. Trenberth, David P. Stepaniak. 2001. "Indices of El Niño Evolution."

"MEI.v2." 2021. https://psl.noaa.gov/enso/mei/.

Miftahuddin, Ichsan Setiawan, Eva Maulia. 2019. "Detecting the Phenomena of Sea Surface Temperature Anomaly by Vector Autoregressive."

Nan Chen, John Harlim, Faheem Gilani. 2021. "A Bayesian Machine Learning Algorithm for Predicting ENSO Using Short Observational Time Series."

Ping Chang, Tim Li, Bin Wang. 1994. "Interactions Between the Seasonal Cycle and the Southern Oscillation - Frequency Entrainment and Chaos in a Coupled Ocean-Atmosphere Model."

School, Columbia Climate. 2021. "IRI ENSO Forecast." https://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/.

Timothy N. Stockdale, Magdalena A. Balmaseda, David L. T. Anderson. 2011. "ECMWF Seasonal Forecast System 3 and Its Prediction of Sea Surface Temperature."

"Vector Autoregression." 2021. https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/vector-autoregression.

Xu, Jin-Song, and Hans von Storch. 1990. "Predicting the State of the Southern Oscillation Using Principal Oscillation Pattern Analysis."