ENVIRONMENTAL DRIVERS OF CARBON AND WATER FLUXES OVER AUSTRALIAN SAVANNAS: STATISTICAL AND MACHINE LEARNING TECHNIQUES.

By Mahrita Harahap

Doctor of Philosophy: Science
Master of Statistics
Bachelor of Mathematics and Finance (Hons)

Faculty of Science,
University of Technology Sydney

June 2019

Submitted in fulfillment of the requirements of the degree of Doctor of Philosophy: Science
Certificate of Original Authorship

I, Mahrita Harahap declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Science at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. I confirm that:

- This work was done mainly while in candidature for a research degree at the University.
- This document has not been submitted for qualifications at any other academic institution.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

Signature of Student,

Mahrita Harahap

28th June 2019
Acknowledgements

This work was supported by an Australian Research Council grant awarded to Dr. Jason Beringer and Prof. Qiang Yu ('Australian savanna landscapes: past, present and future', DP130101566). A portion of the material present in this thesis is based upon work supported in part by OzFlux within the Terrestrial Ecology Research Network (TERN) and the Ecological Modelling Group at UTS.

I have been privileged to complete a PhD on applying statistical methods on ecological processes to understand how ecosystems respond to climate change. Thanks to my supervisors Prof. Qiang Yu, Dr. James Cleverly and Dr. Jason Beringer for this opportunity, for embracing my early career as a professional scientist, for all the exceptional guidance and support, for your patience, for the many questions that helped me understand the environmental processes, and endless advice to make this thesis possible, I will always be grateful. I would like to thank my co-supervisor Dr. James Cleverley for his insightful comments, for helping me to comprehend OzFlux data and help me with data issues. I do not know how I would of done this without your help. I am also thankful to Dr. Jason Beringer for supporting my research in Australian Savannas and providing supervision at the start of my candidature. I appreciate all my supervisors helpful suggestions, reviews and comments to improve my PhD thesis.

This thesis would also not have been accomplished without my family support. To my parents and friends who were patient and always there for me in the most difficult situations. I would like to give a special thanks to Hao Shi, Liang He, Wei Zhuang, Longhui Li, Lindsay Hutley, Rhys Whitley and Caitlyn Moore and the
many more people in my field that I met throughout my candidature. Also to my friends at UTS, Bojana Manojlovic, Marco Giardina, Alex Thomson and Tonantzin Tarin.
Abstract

In light of recent concerns of climate change, it is important to conserve natural ecosystems with vast spatial extent, such as savannas, that contribute to saving the increasing additional carbon dioxide (CO$_2$) in our atmosphere due to anthropogenic activities, into the future. They account for approximately 25% global productivity (GPP), this makes them a key terrestrial biome that provides us a critical service in helping slow the rise of CO$_2$. Therefore, there is a current need to improve the understanding of these ecosystem functions and land surface models that quantify carbon capture and storage capacity to provide an improvement to savanna productivity estimates. These models of primary production in terrestrial ecosystems cannot rely on a single limiting factor but instead must consider multiple potentially limiting coupled processes on multiple time and spatial scales. The lack of flux data and the failure of models to capture and account for the spatial, temporal and multiple timescale effects of this land-surface variability on atmospheric exchanges are recognised as a major source of uncertainty in the current land surface models.

We use multivariate statistical methods and machine learning techniques to extract the signal in the eddy covariance flux data in the presence of noise and to find out what the data show us in the midst of their apparent chaos. Empirical statistical data analysis is needed to help identify the patterns (potential explanations) and key eco-hydrological or physiological processes of savanna vegetation (which is influenced by the carbon and water fluxes) that alters savanna productivity, which some are currently neglected in present land surface models. Modern machine learning techniques is needed to help predict future amounts of carbon storage capabilities. These methods will allow the detection of possibly unanticipated patterns in the
This work will contribute to improve our understanding of the savanna ecosystem function which will indirectly improve the model’s capability in simulating and predicting savanna productivity estimates based on the interactive effects of climate change and CO₂ increase over fine spatial and temporal resolutions at the Northern Australia Territory Transect. Hopefully the models will provide robust future estimates of carbon and water that will aid our management of savannas to ensure viable water resources and carbon sequestration, to provide tools for land managers to account for the consequences of climate change and elevated atmospheric CO₂ concentrations on ecosystem viability and hopefully inform the policy decision makers on the importance on conserving the savannas natural ecosystem. As this project deals with the environmental sustainability of Australian ecosystems, it will be of considerable social benefit for Australia.
# Contents

Certificate of Original Authorship \( i \)

Acknowledgements \( ii \)

Abstract \( iv \)

Table of Contents \( vi \)

List of Tables \( ix \)

List of Figures \( x \)

Chapter 1  **General Introduction** \( 1 \)

1.1 **Introduction** \( 1 \)

1.1.1 Terrestrial ecosystem carbon and water fluxes \( 1 \)

1.1.2 Australian Savannas \( 5 \)

1.1.3 The global carbon cycle in savannas \( 8 \)

1.1.4 Modelling of Primary Production \( 11 \)

1.1.5 Background of Land Surface Models \( 11 \)

1.1.6 Assessing the Reliability of Complex Models \( 14 \)

1.1.7 Multivariate Statistical Methods \( 15 \)

1.1.8 The seven sites on the Northern Australian Territory Transect \( 16 \)

1.2 **Objectives** \( 21 \)

1.3 **Significance of the Research** \( 23 \)

1.4 **Scope and structure of the thesis** \( 26 \)
Chapter 2  
Machine learning tree-based methods for predicting carbon and water fluxes over Australian Savannas  
2.1 Introduction .............................................. 28  
2.2 Methods ................................................. 28  
  2.2.1 Site descriptions .................................... 28  
  2.2.2 Variables used in models ....................... 29  
  2.2.3 Data processing ..................................... 30  
  2.2.4 Statistical analyses .............................. 30  
    2.2.4.1 Regression Trees ......................... 30  
    2.2.4.2 Bagged Trees .............................. 30  
    2.2.4.3 Random Forests ......................... 31  
    2.2.4.4 Gradient Boosting Machines .......... 31  
  2.2.5 Evaluation Metrics ............................. 31  
    2.2.5.1 Adjusted R$^2$ .......................... 31  
    2.2.5.2 Root Mean Squared Error .............. 32  
2.3 Results ................................................ 32  
  2.3.1 Alice Springs ...................................... 33  
  2.3.2 Howard Springs .................................. 33  
2.4 Discussion ............................................ 33  
2.5 Conclusions .......................................... 33

Chapter 3  
Identifying the group of driving factors of Carbon and Water fluxes at Alice Springs using Regression Trees  
3.1 Introduction .............................................. 34  
3.2 Methods ................................................. 34  
3.3 Results ................................................ 35  
3.4 Discussion ............................................ 38  
3.5 Conclusion ............................................. 38

Chapter 4  
Comparison of Carbon and Water Fluxes between two contrasting semi arid sites using Random Forests: Alice Springs and Howard Springs  
3.1 Introduction .............................................. 34  
3.2 Methods ................................................. 34  
3.3 Results ................................................ 35  
3.4 Discussion ............................................ 38  
3.5 Conclusion ............................................. 38
Chapter 5  Identifying environmental drivers in carbon and water fluxes over Australian Savannas using Principal Component Analysis vs. wavelet based Principal Component Analysis  

5.1 Introduction  

5.2 Methods  

5.2.1 Multivariate Analysis  

5.2.1.1 Principal Component Analysis  

5.3 Results  

5.4 Discussion  

5.5 Conclusions  

Chapter 6  Discussion and Conclusion  

6.1 Thesis Aim  

6.2 Implications for vegetation modelling  

6.2.1 Disturbance Effects on the Carbon Cycle  

6.2.2 Constraints in Vegetation  

6.2.3 Coupling Interactions of Complex Models with Multiple Time Scales  

6.3 Concluding remarks  

References
List of Tables

1.1 Properties and locations of flux tower sites on the Northern Australian Territory Transect (NATT) (Source: Information from OzFlux site) ................................................................. 20
List of Figures

1.1 Atmospheric CO$_2$ over time (Source: ?). ........................... 1
1.2 Photosynthesis function through a pore of a leaf called a Stomata, CO$_2$ enters a stomata and H$_2$O leaves a stomata (Source: ?). ... 2
1.3 This graph compares atmospheric levels of carbon dioxide since the 1960s with the levels that would have occurred as a result of the accumulation of anthropogenic CO$_2$ emissions in the absence of carbon sinks. The amount of CO$_2$ that has accumulated in sinks is represented by the difference between the two curves (Source: Ingeborg Levin, 2012). ................................................. 3
1.4 Global distribution of the major savanna biomes (Source: http://www.nrel.colostate.edu) ................... 4
1.5 Savannas Distribution in Australia (Source: http://www.nrel.colostate.edu/projects/srs/australia.html) 4
1.6 Feedback of terrestrial ecosystems and the changing climate. .... 5
1.7 modelled estimates of global mean annual gross primary productivity (GPP) (g C m$^{-2}$ yr$^{-1}$) (Source: Field et al., 1998) ................. 9
1.8 Components of Productivity and the fluxes of carbon C within an ecosystem (modified source on the left: Weathers et al., 2012 and modified source on the right: Chapin et al., 2002) ............... 10
1.9 Australian Savanna Carbon and Water Balance (Source: Beringer et al., 2014) .................................................. 10
1.10 Schematic flow diagram of obtaining simulated flux responses .... 15
1.11 The international FLUXNET Network Distribution and it’s major regional networks (Source: FLUXNET website) ................. 17
1.12 Flux tower sites and type of land cover based on MODIS IGBP Classification (Source: FLUXNET website) ................... 17
1.13 Site distribution of the flux towers in the OzFlux Network (Source: OzFlux website) .................................................. 18
1.14 The flux tower sites on the Northern Territory Transect NATT (modified Source: Ma et al. 2013) ......................... 19
1.15 The Howard Springs site and what the eddy covariance flux towers look like. Picture on the left is the main tower over the canopy and the tower on the right is the mini tower for the understory (Taken on the site in April 2014) ........................................ 20
1.17 Long term drivers of an ecosystem function of a savanna .... 24
2.1 Ecosystem at Alice Springs and Howard Springs ............. 29
2.2 Ensemble tree methods .................................................. 31
3.1 Regression Tree with cp=0.001 for Alice Springs Wet Season with NEE (Net Ecosystem Exchange as proxy for carbon flux) as the response variable. The first split is partitioned by Radiation at 142. Figure 3.2 tells us NEE is first partitioned into day time and night time/early morning time dynamics. ..................... 35
3.2 First split at Radiation=142. Tells us NEE is first partitioned into day time and night time/early morning time dynamics. NEE below 0 represents a carbon sink and above 0 means carbon source. ... 36
3.3 Second split at LAI=0.52. Tells us NEE is secondly partitioned into understory dominance and overstorey dominance. .......... 37
3.4 Driving factors for NEE .................................................. 38
4.1 Howard Springs NEE predictions where training set was 2012 to 2016 and test set is 2016 to 2017 ........................ 40
4.2 Alice Springs ET predictions where training set was 2012 to 2016 and test set is 2016 to 2017 .............................. 41
Chapter 1

General Introduction

1.1 Introduction

1.1.1 Terrestrial ecosystem carbon and water fluxes

Global mean temperatures have been increasing over time and this has a direct and indirect relationship to the rising of atmospheric carbon dioxide (CO$_2$) (See Figure 1.1). In light of recent concerns of climate change, we are interested in what are causing these trends. Studying specific parts of the global carbon and water cycle helps us achieve this.

The land and ocean act as natural carbon sinks - reservoirs that accumulates and stores carbon for an indefinite period, thereby slowing the rise of human induced atmospheric CO$_2$ concentrations. It achieves this through the absorption of carbon

![Atmospheric CO$_2$ at Mauna Loa Observatory](image)

Figure 1.1: Atmospheric CO$_2$ over time (Source: ?).
Figure 1.2: Photosynthesis function through a pore of a leaf called a Stomata, CO₂ enters a stomata and H₂O leaves a stomata (Source: ?).

by the growing vegetation via photosynthesis - which is influenced by carbon and water fluxes (gas exchanges). Figure 1.2 shows the productivity of vegetation in terms of carbon sequestration, is measured by looking at the fluxes of CO₂ entering the pores of the leaf, called stomata, and H₂O leaving the pores of the leaf. A flux is the rate of flow of particles across a given area.

Figure 1.3 tells us why it is important to enhance this natural capacity of carbon sinks (Trumper et al., 2009) by preservation, protection and conserving our natural ecosystems that contribute to binding the increasing carbon dioxide in our atmosphere due to anthropogenic activities (burning of fossil fuels, clearing natural vegetation and forest fires - this therefore leads to the reduced ability of natural ecosystems to bind carbon through photosynthesis).

Due to the vast spatial extent of semi-arid ecosystems such as savannas, which account for approximately 20% global terrestrial surface (see Figure 1.4), Beer et al.(2010) estimated savanna productivity as 31.3 Pg C yr⁻¹ which accounts for about 25% of the global production productivity (Scholes and Archer, 1997), in terms of carbon sequestration, this makes them a key terrestrial biome. Savannas are grassy plains in subtropical regions and are significant for grazing, the habitat
Figure 1.3: This graph compares atmospheric levels of carbon dioxide since the 1960s with the levels that would have occurred as a result of the accumulation of anthropogenic CO$_2$ emissions in the absence of carbon sinks. The amount of CO$_2$ that has accumulated in sinks is represented by the difference between the two curves (Source: Ingeborg Levin, 2012).

of our native wildlife, cultural heritage (significant landscapes, Indigenous land and terrestrial biodiversity) and wilderness value (uncultivated - not prepared for growing crops). Australian savannas cover 25% of the continent (see Figure 1.5) (Fox et al., 2001) and host much of the grazing industry in Australia, although conservation and Indigenous land covers a significant fraction of this area and this region has been described as one of the world’s last great wilderness areas (Woinarski et al., 2007).

With highlighting the importance of savannas, they provide important ecosystem services (carbon cycle, water balance, biodiversity) but are highly sensitive to climate change and variability, especially precipitation, and pose a large vulnerability to the global carbon cycle (Grace et al., 2006). Climate change and anthropogenic activities is likely to provide significant feedback to the structure
Figure 1.4: Global distribution of the major savanna biomes (Source: http://www.nrel.colostate.edu/projects/srs/)

Figure 1.5: Savannas Distribution in Australia (Source: http://www.nrel.colostate.edu/projects/srs/australia.html)
and function of savanna and pose a threat to this ecosystem (See Figure 1.6).

There is currently limited research surrounding savanna vegetation structure, function and dynamics, how they interact with the atmosphere and how they respond to climate change. Quantifying and modelling the carbon and storage capacity of savannas accurately, is vital in enhancing the natural capacity of these ecosystems.

1.1.2 Australian Savannas

Savannas are a type of biome that is found in tropical and sub-tropical parts of Africa, South America and Australia (see Figure 1.4), generally within 30°degrees north and south of the equator (Bond, 2008) and are characterised by an extensive understorey of tall grasseslands (C4 photosynthetic pathway) spotted with dominating overstorey of drought and fire resistant thorny eucalypts trees (C3 photosynthetic pathway) that senesce in the dry season (Beringer et al., 2011). Increasing atmospheric CO$_2$ will have differential effects on wood plants vs. grasses which
compete for water and being dominated by their co-existence makes the estimation of savanna productivity complex (Donohue et al., 2009). Accounting for the distribution and composition of tree and grass components is critical but is currently inadequate in models since most do not deal with varying tree and grass composition (Beringer et al., 2011). Trees store carbon in dense hardwoods, which may take many centuries to die and release the carbon, whereas much of the CO$_2$ absorbed by plants in semi-arid regions are short-lived grasses and shrubs (high carbon turnover). This is just one of the reasons why it is difficult to simulate savanna fluxes.

Savannas are highly sensitive to climate change and variability, especially precipitation. Links between precipitation and the land CO$_2$ sink in semi-arid ecosystems are currently missing from many major climate models (Poulter et al., 2014). Typically, a period of heavy rainfall is followed by a prolonged drought, which results in a very seasonally structured ecosystem. At a global scale, savannas occur in regions with strongly alternating wet and dry seasons and with rainfall ranging from 300 to 2000 mm.yr$^{-1}$ (Mistry, 2000). The wet season contribute most to the carbon and water fluxes. The dry season processes are typically harder to characterise (Ma et al., 2013).

At local scales, vegetation productivity in savannas are modulated by soil-water content, nutrient availability, disturbance regimes and land-use change (Scholes, 1997). The partitioning of soil moisture depth to counter for soil-water content based on nutrients is important to incorporate in the models. This is because the life and productivity of vegetation is dependent on the length of their roots for plant water uptake from the soil. Li et al. (2012) has compared modelled and observed fluxes and concluded that accounting for the differential rooting distribution of savanna vegetation, namely shallow rooted grasses and deeper rooted woody vegetation (O'Grady et al., 2000), represent a valuable improvement for land surface modelling for studying carbon and water dynamics where rainfall varies seasonally
or interannually.

Fire is a common disturbance of the savanna. Beringer and Hutley et al., (2007) found evidence that these fire disturbances significantly impact the productivity in savannas and therefore they should be stochastically modelled and incorporated in the land surface models.

Many land surface models fail to capture the seasonality of fluxes because they cannot simulate the complex interactions between rainfall, soil moisture, root distribution and plant water uptake (Pitman, 2003) which leads to uncertainties in local and global climate simulations. Given the poor understanding of the savanna ecosystem and its vulnerability to change, further scientific research of the region is needed to sustainably manage these ecosystems.

Savannas are closely coupled with the atmosphere through both biophysical processes (surface water and heat balance which influences the regional climate) and biogeochemical cycles (carbon cycle and other nutrients). This section gives a background about how the carbon cycle and the surface energy balance regulates savanna productivity.

New research by Poulter et al. (2014) suggests that semi-arid areas are the key carbon sinks in the future which may alter global greenhouse gas levels. This is due to the unusual above average plant growth in the semi-arid areas of South America, Australia and Africa. The la nina events caused 60% of the extra new plant growth was in semi-arid areas of Australia. Whereas the mean sink is dominated by highly productive lands (mainly tropical forests) the trend and interannual variability of the sink are dominated by semi-arid ecosystems whose carbon balance is strongly associated with circulation-driven variations in both precipitation and temperature (Ahlstrom, A. et al. 2015). Poulter et al. (2014) suggested the reasons may be that different vegetation types may have different responses and sensitivities to meteorological variables and different types of ecosystems lock away absorbed
CO₂ for different lengths of time. Another reason may be is that very little information exists about vegetation in semi-arid ecosystems compared to other regions and therefore the current land surface models that predict carbon and water fluxes may not be matching with the observations accurately for these ecosystems. So we need to somehow counter for these effects and the missing processes in land surface models.

So to see how these “carbon sinks” might change in the future, we have land surface models to estimate this. The problem is that trends are poorly understood and these ecosystems remain insufficiently studied statistically.

1.1.3 The global carbon cycle in savannas

Primary production is the storage of energy (which is essential for all life) through the formation of organic matter from inorganic carbon compounds (e.g., Photosynthesis) and is carried out by autotrophic organisms (plants and algae), which is then consumed by herbivores and predators that depend on it directly or indirectly for their energy supply (Weathers et al., 2012).

Primary production begins with the fixation of CO₂ into organic matter, which is represented by Gross Primary Production (GPP). Net primary production (NPP) is the difference between GPP and autotrophic respiration (Rₐ).

\[ NPP = GPP - Rₐ \]

Ecosystem respiration (Rₑ) is the sum of Rₐ and heterotrophic respiration Rₕ.

\[ Rₑ = Rₐ + Rₕ \]

Organic carbon that is not respired by plants or animals either accumulates as carbon in biomass or detritus, is exported from an ecosystem or is lost through fire (See Figure 2.2) is called the Net Ecosystem Production (NEP) or Net Ecosystem
Exchange (NEE).

\[ NEE = NEP = GPP - R_e = GPP - R_a - R_h = NPP - R_h \]

Organic carbon accumulation in ecosystems sequestered over long time periods provides a sink for atmospheric CO₂ and is very important to those studying global carbon budgets (Weathers et al., 2012), especially when ecosystems like savannas account for 25% of the global primary production (Beer et al., 2010) because it indicates they are an important source of carbon storage within ecosystems (Grace et al., 2006).

Figure 1.9 shows the current carbon and water budget in Australian savanna ecosystems.

Measurement of grassland primary production is done by harvesting (clipping and weighing) plant material (otherwise known as biomass) produced over a given time interval and this is often approximately correlated with NPP and is represented as mass per area independent of time (Weathers et al., 2012). These gas
Figure 1.8: Components of Productivity and the fluxes of carbon C within an ecosystem (modified source on the left: Weathers et al., 2012 and modified source on the right: Chapin et al., 2002)

Figure 1.9: Australian Savanna Carbon and Water Balance (Source: Beringer et al., 2014)
measurements are typically representative of large spatial areas. Eddy covariance is a gas exchange measurement approach in terrestrial ecosystems where the term flux is a measurement of gas exchange through a unit area per unit time (Luo et al., 2012). A fast-response CO$_2$ sensor is paired with a multidimensional wind speed sensor on these eddy covariance flux towers extending above a vegetation canopy to measure the CO$_2$ fluxes into and out of the canopy on these air currents which are calculated using algorithms programmed into the computer (Burba and Anderson 2010). The difference, integrated over time, is termed the net ecosystem exchange (NEE). NEE is an instantaneous measurement of NEP which is a good estimation of the organic carbon accumulation rate in the ecosystem (Weathers et al., 2012). NPP cannot be readily calculated because it is difficult to partition R$_e$ into R$_a$ and R$_h$.

"High uncertainties in quantifying ecosystem processes mean that the global terrestrial carbon sink is often estimated as the residual between emissions from the combustion of fossil fuels, cement production and net land-use change, and sinks combining accumulation in the atmosphere and uptake by the ocean" By Poulter 2014.

1.1.4 Modelling of Primary Production

1.1.5 Background of Land Surface Models

Aim of ecosystem modelling is to improve insight and understanding of the complex interactions within an ecosystem. Helps to generate future scenarios and forecasts under future variations in climate. Land surface models have drivers and processes that go into the model which then give outputs of simulated fluxes that carry a degree of uncertainty. However, there is a need to improve the current land surface models that quantify carbon capture and storage capacity to provide an improvement to savanna productivity estimates. The land surface models of primary production must consider multiple potentially limiting coupled processes on multiple time and spatial scales and this affects the land-surface variability on atmospheric exchanges and are recognised as a major source of uncertainty in the current land surface models. Which is why statistical data analysis is needed to identify the
major drivers of the fluxes through observing the patterns in the data.

In order to examine past variability and in turn to better understand the sensitivity to future change, we require land surface models that can accurately simulate savanna fluxes. These process-based land surface models are needed to make projections about how ecosystems and their interactions with the atmosphere may change under a changing climate.

Most climate models used in the IPCC Fourth Assessment Report (Randall et al., 2007) treat all savannas as a single vegetation/biome type and therefore do not adequately capture the variability of savanna structure, composition and function. Models that have been applied in Australian savannas are CABLE, SPA, BIOS2, BESS, VOM, VPJ-GUESS but we will only describe the first two in the next section. As said earlier, many of these models fail to capture the seasonality of fluxes because they cannot simulate the complex interactions between rainfall, soil moisture, root distribution and plant water uptake (Pitman, 2003). Savannas are typically of seasonally water-limited ecosystems and our current models do not perform well in these environments. Some new research has shown current land surface models under-simulate how much carbon is stored by plants and in consequence overestimate how much carbon goes into the atmosphere (reference?). Therefore any advances in simulating savanna dynamics and fluxes will be beneficial for all arid and water-limited ecosystems.

Possible reason why it is difficult to simulate savanna fluxes is that the models have simplified important eco-hydrological and physiological processes of plant and soil. In the context of the carbon cycle, many land surface models have been developed to scale up photosynthesis from the leaf to canopy level based to estimate regional/global production productivity. Canopies can slow down wind speed and decrease boundary layer conductance (Hastings et al., 2012). These models are based on the canopy structure and how the gradients of the environmental factors are treated. The models can be categorised as:
• big-leaf (single-layer) models - regard the whole canopy like one "big leaf" by assuming all the leaves in the canopy are the same and have the same water conditions. This model tends to overestimate photosynthesis rate and transpiration (Luo et al., 2012) but are appropriate when canopy structure and its climate conditions can be ignored such as deserts. Current big leaf-models do not account well for the heterogeneous nature of savanna canopies.

• two-leaf models - separate leaves into shaded and sunlit leaves, based on the structure of the canopy and the angles of solar radiation, then simulate photosynthesis in the two classes of leaves individually. Can be comparable as the following multiple layer model but are much simpler. They are widely used in current Earth system models (Luo et al., 2012).

• multiple layer models - separate a canopy into many layers and calculate carbon and water fluxes at each layer according to its physiological properties and climatic conditions (Luo et al., 2012).

Each model has a different philosophy and algorithms. As the scope of this work, we will only describe the first two models from the previous list for this report:

• The CABLE model is an example of a two-leaf model and

• The SPA model is an example of a big-leaf multiple canopy layer model.

CABLE: CSIRO and the Bureau of Meteorology partnered together to develop the Australian Community Climate and Earth System Simulator (ACCESS) through The Centre of Australian Weather and Climate Research by scientists in international agencies and Australian universities. ACCESS is a fully coupled earth system model that seamlessly link together models of the oceans, atmosphere, sea-ice, land surface, global carbon cycle and chemistry, and aerosols, to give improvements in predictive capabilities to assist coastal communities, emergency services, primary producers, energy and resource management, public health and other ecosystem services (CSIRO, 2013).
ACCESS, who is developed by scientists, assist with the development and testing of the Community Atmosphere - Biosphere Land Exchange Model (CABLE) to ensure it represents terrestrial processes well (CSIRO, 2013).

Over the past decade, land surface models have developed from simple to more complex representations of soil-vegetation-atmosphere interactions (Wang et al., 2006). This evolution has been facilitated by advances in plant phenology and the availability of global land surface parameters obtained from remote sensing (Wang et al., 2006). The CSIRO Atmosphere Biosphere Land Exchange (CABLE) model is a two-leaf model that calculates carbon, water and heat exchanges between the land surface and atmosphere (Wang et al., 2010).

SPA: The Soil-Plant-Atmosphere model (SPA) (Williams et al., 1996) is a process-based big-leaf model that simulates carbon and water fluxes at spatiotemporal scales (multiple canopy and soil layers). The multi-layer canopy allows simulation of tree and grass components.

This model have been designed to provide a tool for scaling up leaf level processes to canopy scales. More details can be found in Williams et. al., 1996 and Williams et. al., 2001.

The current problem with the SPA model is it works well at a local scale but not at a regional scale.

1.1.6 Assessing the Reliability of Complex Models

- **Uncertainty Quantification:** How do the various sources of error feed into uncertainty in the model-based prediction of the quantities of interest (NRC, 2012)? We will look at the simulated fluxes obtained from the models against the observations observed from flux tower observations then compare $R^2$ and the sources of error Root Mean Squared Error (RMSE). This will be discussed in the next chapter.
The mathematical sciences foundations assessing the reliability of models will help researchers identify the inconsistencies and the driving factors of the gaps in the present models.

1.1.7 Multivariate Statistical Methods

We use multivariate statistical and machine learning methods to extract the signal in the eddy covariance flux data in the presence of noise to find out what the data show us in the midst of their apparent chaos. Empirical statistical data analysis is needed to help identify the patterns (potential explanations) and key eco-hydrological or physiological processes of savanna vegetation (which is influenced by the carbon and water fluxes) that alters savanna productivity, which some are currently neglected in present land surface models. These methods will allow the detection of possibly unanticipated patterns in the data, opening up a wide range of competing explanations to contribute to better savanna data inputs into the capable models that focus on estimating global GPP values.

The methods that this PhD thesis will use are Regression Trees, Bagged Trees, Random Forests, Gradient Boosting Machines, Principal Component Analysis and
Wavelet Analysis. The thesis will also take a brief look into other statistical and machine learning methods applied in this field in previous literature such as Path Analysis, Factor Analysis and Cluster Analysis. The methods can be categorised into the following areas (although some methods can be argued to be in several areas):

- 1. Exploratory multivariate statistical methods: Principal Component Analysis, Factor Analysis, Cluster Analysis
- 2. Spectral analysis: wavelet analysis
- 3. Machine learning methods: regression trees, bagged trees, random forests and gradient boosting machines, Artificial Neural Networks
- 4. Extension of regression methods: Path Analysis

1.1.8 *The seven sites on the Northern Australian Territory Transect*

FLUXNET is an international network of regional networks such as OzFlux, AmeriFlux, AsiaFlux, CarboAfrica, CarboEuro-Flux, ChinaFlux, Fluxnet-Canada and comprises other unaffiliated or non network sites (see Figure 3.1 and 3.2).

We will look at flux tower sites based in the Northern Territory and obtain the data sets from OZFlux. OzFlux is a network of micrometeorological flux stations located at various sites within Australia and New Zealand (see Figure 3.3) and is part of the global network, FLUXNET, of 683 flux tower sites that use eddy covariance methods to measure the exchange of carbon dioxide, water vapour, and energy between terrestrial ecosystem and atmosphere which are measured continuously over long periods. The eddy covariance technique is a complex mathematical atmospheric measurement technique to measure vertical turbulent fluxes within atmospheric boundary layers by analysing high-frequency wind to determine gas exchange rates of trace gases over natural ecosystems and to quantify gas emission rates from other land and water areas (Burba and Anderson, 2010).

We will use flux tower measurements at all seven sites along the spatial subcontinental scale rainfall gradient of more than 1100km in length, known as the North Australian Tropical Transect (NATT) shown on Figure 3.4 (CSIRO, 2013), to provide a unique data set (over a span of 5-10 years) for this research. The list
Figure 1.11: The international FLUXNET Network Distribution and it’s major regional networks (Source: FLUXNET website)

Figure 1.12: Flux tower sites and type of land cover based on MODIS IGBP Classification (Source: FLUXNET website)
of flux towers on the NATT along with their properties and locations are shown in Table 3.1. It includes Howard Springs, Adelaide Rivers (which is no longer active), Daly River Uncleared, Dry River, Sturt Plains, Ti Tree East, Alice Springs.

The Australian Transect Network (established under the International Geosphere-Biosphere Program (IGBP)) comprises four major subcontinental transects that traverse major biomes from the coast to inland to facilitate the study of ecological structure and processes over the major biomes, and one of them is the North Australian Tropical Transect (TERN, 2013). These transects can be used to help develop and calibrate ecological models for an enhanced understanding of controls on ecosystem functions in the context of climate change (TERN, 2013).

The NATT provides a framework for modelling the dynamics of Australia’s tropical savannas in relation to variation in soils and disturbance, and it covers a range of ecosystem types, from monsoonal tall-grass systems near the coast, to dry grassland near the inland arid zone (CSIRO, 2013).
At each site half-hourly fluxes of carbon, water (ET) and energy (NEE) are measured using the tower based eddy covariance technique (Burba and Anderson, 2010). We will use these flux measurements to quantify savanna productivity. Data collection and processing procedures are described in Beringer et al. (2007).

During April 2014, my co-supervisor Prof. Jason Beringer, took me to the Howard Spring site in the Northern Territory to see how these eddy covariance flux towers function and how the data is collected (see Figure 3.5).

The Eddy Covariance flux tower measurements are of course never perfect, because of assumptions, physical phenomena, instrumental problems, and specifics of the particular terrain or setup. As a result, there are a number of potential flux errors (Burba and Anderson, 2010). There are time response errors (occur because instruments may not be fast enough to catch all the rapid changes that result from
Table 1.1: Properties and locations of flux tower sites on the Northern Australian Territory Transect (NATT) (Source: Information from OzFlux site)

<table>
<thead>
<tr>
<th>Flux Tower Sites on the NATT</th>
<th>Landcover</th>
<th>Annual Rainfall</th>
<th>Temp Range</th>
<th>Location</th>
<th>Group</th>
<th>Status</th>
<th>Longitude and Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Howard Springs</td>
<td>Tropical savanna (wet)</td>
<td>1700mm</td>
<td>20 - 33</td>
<td>E of Darwin, NT 12° 49.42'S 131° 15.23'E</td>
<td>Monash University (Melbourne)</td>
<td>Running since August 2001</td>
<td>12°29.732' S, 131°09.003'E</td>
</tr>
<tr>
<td>Adelaide Rivers</td>
<td>Savanna</td>
<td>1730mm</td>
<td>16 - 36</td>
<td>approx. 10.5km south east of Bachelor, Northern Territory 13° 4' 36.84'' S 131° 7' 4.00'' E</td>
<td>Charles Darwin University (Melbourne)</td>
<td>Ran from November 2007. No longer active</td>
<td>13° 4' 36.84'' S, 131° 7' 4.08'' E</td>
</tr>
<tr>
<td>Daly River Uncleared</td>
<td>Woodland savanna</td>
<td>1170mm</td>
<td>12 - 37</td>
<td>Douglas River Daly River Esplanade Conservation area, approx. 89km south west of Pine Creek, Northern Territory 14°09'33.12''S 131°23'17.16''E</td>
<td>Charles Darwin University</td>
<td>Running since February 2007</td>
<td>14°09'33.12''S, 131°23'17.16''E</td>
</tr>
<tr>
<td>Dry River</td>
<td>Open forest savanna</td>
<td>895mm</td>
<td>14 - 37</td>
<td>approx. 280km north of Tennant Creek, Northern Territory 15°10'31.62''S 132°22'14.04''E</td>
<td>Charles Darwin University (Melbourne)</td>
<td>Running since September 2008</td>
<td>15°10'31.62''S, 132°22'14.04''E</td>
</tr>
<tr>
<td>Sturt Plains</td>
<td>Low lying plain dominated by Mitchell Grass</td>
<td>640mm</td>
<td>11 - 39</td>
<td>approx. 171km north of Tennant Creek, Northern Territory 17° 9' 2.46'' S 133° 21' 0.79'' E</td>
<td>Charles Darwin University (Melbourne)</td>
<td>Running since August 2008</td>
<td>17° 9' 2.46'' S, 133° 21' 0.79'' E</td>
</tr>
<tr>
<td>Ti Tree East</td>
<td>Grassy muga woodland and Eucalyptus Fordii savanna</td>
<td>305mm</td>
<td>-4 - 46</td>
<td>Pine Hill cattle station, Northern Territory 22.267 °E</td>
<td>Charles Darwin University (Melbourne)</td>
<td>Running since July 2012</td>
<td>22.287 °S, 133.640 °E</td>
</tr>
<tr>
<td>Alice Springs</td>
<td>Semi-arid muga (Acacia aneura) ecosystem</td>
<td>306mm</td>
<td>-4 - 46</td>
<td>133 640 °E Pine Hill cattle station, near Alice Springs, Northern Territory</td>
<td>University of Technology Sydney</td>
<td>Running since September 2010</td>
<td>22.283 °S, 133.249 °E</td>
</tr>
</tbody>
</table>

Figure 1.15: The Howard Springs site and what the eddy covariance flux towers look like. Picture on the left is the main tower over the canopy and the tower on the right is the mini tower for the understory (Taken on the site in April 2014)
the eddy transport), sensor separation error (happens because of physical separation between the places where wind speed and concentration are measured), path averaging error (caused by the fact that the sensor path is not a point measurement, but rather integration over some distance; therefore it can average out some of the changes caused by the eddy transport) and other errors (Burba and Anderson, 2010).

The field trip to Howard Springs flux tower site in April and August showed us how there could be instrumental problems. For example a green ants nest could be formed in the cabinet where the data is stored, or memory gets overheated or rain water gets into the memory and data gets lost. When there are missing observations, gaps of shorter than 2 hours duration can be filled with linear interpolation methods, gaps greater than 2 hours, the data user selects priority for source of alternate met file such as the Bureau of Meteorology etc. then uses it to generate values in the gaps. Berlinger et. al., (2011) explains this process more in detail (see Figure 5.1). Once the meteorological driving variables with missing observations are gap filled, the fluxes are gap filled.

Gap filled data sets is only a problem if the percentage of data that is gap-filled is more than 5-10%. We want to work with less synthetically generated datasets as possible.

1.2 Objectives

The aims of this research will address three important scientific questions:

1. What are the most important eco-hydrological or physiological processes of savanna vegetation that alters savanna productivity? Objective here is do data analysis and use explanatory tools to identify the patterns and potential drivers of the interactive impact of climatic variables on savannas and decompose the fluxes down to their processes.
2. What controls savanna phenology spatiotemporally and in time? Objective here is to .....to identify the biases in multiple spatial and time scales performance and diagnose improvements.

3. How do we improve quantification and modelling of the carbon sequestration and water use of savannas to reduce model uncertainty? Objective here is to undertake some alternative approaches to traditional methods of modelling in this field like statistical methods and machine learning techniques to predict the carbon sink.

As a summary of this section, this research focusus on improving modelling on these two key ecosystem indicators;

1. savanna productivity, which is defined as the Net Ecosystem Exchange (NEE) and is the key input to the carbon cycle, and
2. savanna evapotranspiration (ET otherwise known as latent heat flux), which is the flux of water via evaporation from the land surface and transpiration
by vegetation. ET is a key part of the hydrological cycle and water budget, with precipitation being the key input.

1.3 Significance of the Research

The efficiency of savannas is estimated by land surface models, such as Australia’s CSIRO Atmosphere Biosphere Land Exchange model (CABLE) and the Soil-Plant-Atmosphere (SPA) model (will be explained further in the next chapter), which include many factors regulating productivity and these models of primary production in terrestrial ecosystems cannot rely on a single limiting factor but instead must consider multiple potentially limiting coupled processes on multiple spatial and scales.

Spatial heterogeneity assures that eco-hydrological or physiological processes of savanna vegetation and gradients exhibit a complex dynamical three-dimensional structure. Meteorological drivers and landscape dynamics operate on a variety of time scales from diurnal to seasonal to interannual but instrumental measurements do not capture the full range of time scale variability and so the models have limitations in the climate representations in the data (TERENO, 2014).

Capable complex models are needed to factor all this in and project the long term consequences of changing climatic variability. There is a lack of studies using statistical methods in this area. Observation data can inform modelers and statistical methods are in great need to show how field observations can help to identify patterns and the key processes which are neglected in models. This justifies the direction and scope outlined in this research, as this project will help identify the gaps in present models used to predict fluxes.

Pitman (2003) has identified some gaps in regards to land surface models (LSM), such as uncertainties in parameterising soil-root interactions, neglecting hydrological processes and biogeochemical processes and spatial heterogeneity. There are multiple drivers of savanna composition, structure and function which results in a
mosaic of patches in differing functional states and capturing this spatial and seasonal variation is complex.

It is evident from the list of references that there is a lot of research in this area based on physical based processes and assessing and evaluating reliability of the land surface models. There is a lack of research in data driven empirical research and perhaps this is why the simulation of savannas is not very accurate. Statistical methods and big data analytics are in a great need in this field. Present field observations can help identify the key processes which are neglected in models. The first research paper I will be working on is "Pattern identification of interactive impacts of climatic variables". Here I will use multivariate explanatory data analytics tools such as cluster analysis and factor analysis to identify the patterns and potential explanations to formulate the right hypothesis questions.
The research will addresses the National Research Priority of An Environmentally Sustainable Australia through three priority goals:

1. **Water:** Assessing the role of climate variability on water fluxes and the impact on site water balance in savannas. This will increase our understanding of sustainable water management in the region;
2. **Reducing and capturing emissions in transport and energy generation:** Determining the role of Australian savannas in sequestering carbon and their likely response to climate change and;
3. **Responding to climate change and variability:** Addressing the response of ecosystems to change and the impact on water and carbon resources.

We will address the national priorities through advanced research using an integrated observational and modelling approach that will hopefully improve the models that will ultimately reduce the uncertainty of our predictions (Beringer et al., 2012).

The anticipated outcomes of this research are therefore:

1. A greatly improved ability to simulate land-surface-atmosphere exchange in the Australian land surface model, thereby improving numerical climate forecasting and land and water resource management.
2. An improved mechanistic understanding about the interactive impact of climate change and CO₂ increase on Australian vegetation function, especially carbon and water fluxes.

Hopefully this research will contribute to providing robust future estimates of carbon and water that will aid our management of savannas, to provide tools for land managers to account for the consequences of climate change and hopefully inform the policy decision makers on the importance on conserving the savannas. As this project deals with the environmental sustainability of Australian ecosystems, it will be of considerable social benefit for Australia and a contribution to science.

Lit Review: To date, process-based land surface models (LSMs) have been challenged in attempting to predict observed fluxes, especially in savanna ecosystems (Beringer et al., 2011). Wang et al. (2011), has compared modelled and observed
fluxes at two forested flux station sites and diagnosed errors in the CABLE model using wavelet analysis to identify the frequencies where model errors are relatively large and then analysed the sensitivities of model errors at those frequencies to selected model parameters, yielding improvements in several measures of model performance. Wang et al. (2011), concluded that analysing both mean and variance of model errors at a range of time/frequency scales is useful for identifying and reducing errors of a land surface model. Li et al. (2012) and O’Grady et al. (2000) has also compared modelled and observed fluxes and concluded that root functions represent a valuable improvement for land surface modelling and should be implemented into CABLE and other land surface models for studying carbon and water dynamics where rainfall varies seasonally or interannually.

Kanniah et al., 2009 indicated a need to adequately capture the extremes of seasonally available moisture and rapid changes in phenology and leaf area index (LAI) observed in savanna vegetation during transition phases from dry to wet seasons and vice versa.

1.4 Scope and structure of the thesis

This thesis encompasses six chapters including this Introduction chapter, which provides the general background of the terminologies and their methodologies used in subsequent chapters. Each chapter embodies different methodologies applied to similar sites, thus there is a degree of redundancy in the introduction and methodologies when referring to the sites, data and sampling methods.

The focus of Chapter 2 is to review four different types of machine learning tree based methods, such as Regression Trees, Bagged Trees, Random Forests and Gradient Boosting Machines, in predicting NEE (Carbon Flux) and QLE (Water Flux). This study compared the four tree-based methods on the Alice Springs and Howard Springs site.
The focus of Chapter 3 was to evaluate the pairwise relationships amongst NEE (Carbon Flux), QLE (Water Flux) and meteorological variables through the machine learning method of Regression Trees also known as Recursive Partitioning in statistics. This study explored these relationships across different temporal scales using multiple years of eddy covariance data and satellite observations of two different kind of savanna ecosystems situated on the two ends of the Northern Australian Territory Transect (NATT).

The focus on Chapter 4 is to mainly use the Random Forest method to predict NEE and QLE and compare the results between two sites, Howard Springs and Alice Springs.

The focus on Chapter 5 is to compare results on Principal Component Analysis and wavelet PCA between multiple sites on the NATT. This study looked at Howard Springs, Adelaide River, Sturt Plain, Ti Tree and Alice Springs.
CHAPTER 2

Machine learning tree-based methods for predicting carbon and water fluxes over Australian Savannas

2.1 Introduction

Advancement of technology and greater speed of computers to compute algorithms with availability and affordability of software, over the last 5 decades have played a significant role in helping machine learning techniques such as tree-based methods, gain popularity and acceptance in the scientific community. There have been a number of reviews (e.g. Loh 2016 and Murthy 1998) of tree-based methods and some introductions intended for non-statistics audiences (e.g., De’ath & Fabricius 2000; Harper, 2003 and Lemon et al., 2005). In this paper, the objective is to review some of the majorly popular tree-based methods with application to eddy covariance and meteorology data to predict carbon and water flux in our atmosphere.

"write about why we are interested in predicting carbon and water fluxes over Australian savannas"

2.2 Methods

2.2.1 Site descriptions

We decided to choose to conduct the analyses of this study on two savanna sites: Howard Springs and Alice Springs. The two sites encompass contrasting savanna types. Howard Springs is a more wet savanna type with rainfall ranging from .... whereas Alice Springs is a much more drier site where rainfall ranges from .... These two sites on the two ends of the Northern Australian Territory Transect have eddy covariance (EC) towers installed and are part of the OzFlux Network (http://ozflux.org.au)
and the Terrestrial Ecosystem Research Network (TERN). Howard Springs is more of a tropical savanna and Alice Springs is more of a semi arid savanna. Several related statistical studies have been conducted in these ecosystems (give references...).

![Figure 2.1: Ecosystem at Alice Springs and Howard Springs](image)

**2.2.2 Variables used in models**

For both sites, we use $F_c$ variable from the datasets to represent Carbon Flux (NEE), and $F_e$ variable to represent Water Flux (QLE).

Alice Springs: Vapor Pressure Deficit (VPD) is measured at 1167cm height. Temperature ($T_a$), Net radiation ($R_n$), Soil Water Content (below understory). LAI was extracted from MOD13QI (EVI, NDVI) 16 day interval 250 pixels and MOD15A2H (LAI, FPAR), 8 day interval, 500m pixels. Because remote sensing LAI measurements were given per 9 day intervals, we used linear interpolation in R to gapfill to match the 30min interval eddy covariance and meterological data.

Howard Springs: 18% of the wet season data is missing and 21% of the dry season data is missing. There could be a number of reasons why this is. This
means we have 47% complete cases in wet season data and 51% complete cases in dry season data. Because it is such a big proportion of missing data, the analyses will be done for both ungapfilled and gapfilled data for Howard Springs to see if the results change drastically. Prof. Jason Beringer has gapfilled the Howard Springs data using the DINGO model which uses the Artificial Neural Network method to fill the missing data. For Howard Springs we use: Vapor Pressure Deficit (VPD) is measured at 1167cm height. Temperature (Ta), Net radiation (Rn), Soil Water Content (below understory). LAI was extracted from MOD13QI (EVI, NDVI) 16 day interval 250 pixels and MOD15A2H (LAI, FPAR), 8 day interval, 500m pixels

2.2.3 Data processing

We divide the data for each site as 80% training set, which builds the model, 10% validation set, which the purpose of the validation set is to tune the hyperparameters and 10% is the test set, which evaluates the generalisability of the model.

2.2.4 Statistical analyses

2.2.4.1 Regression Trees

An alternative approach to linear regression models is recursive partitioning, where the resulting models are called regression trees. The first tree-based method published in literature was a regression tree algorithm called Automatic Interaction Detection (AID) by Morgan & Sonquist (1963) then eventually developed into the popular Classification And Regression Trees (CART) method by Breiman et al. 1985. This method was originally invented to deal with possible nonlinear relationships between covariates and responses. Unlike linear regression models, it does not allow linear and additive effects of predictors but instead allows a more complex dependency of the response variable on these predictor variables. The basic idea is to partition the covariate space and to compute simple statistics of the dependent variable inside each cell. The value used for splitting is determined by testing every value for every variable. The one which minimises the residual sum of squares (RSS) is chosen.

2.2.4.2 Bagged Trees

Bagged trees are a tree ensemble method.....
By aggregating many regression trees, using methods like bagging, random forests, and boosting, the predictive performance of decision trees can be substantially improved and this will be shown in this study.

2.2.4.3 Random Forests

Bagged trees with randomness added.

2.2.4.4 Gradient Boosting Machines

Another tree ensemble method but uses the gradient steepest descent method.

The tree based methods were evaluated and compared using the adjusted R squared and RMSE.

2.2.5 Evaluation Metrics

2.2.5.1 Adjusted $R^2$

The coefficient of determination $R^2$ indicates how well the data fits a model. In other words, how well is the response variable explained by the explanatory variable. $R^2$ is a dangerous criterion for model comparisons and adding model terms will increase
Adjusted $R^2$ takes sample size and the number of parameters into account. The adjusted $R^2$ indicates that the proportion of the total variation in the level that is explained by the variables. However, adjusted $R^2$ is a very soft criteria for model selection.

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

where $R^2=$ is the sample coefficient of determination

\[ p = \text{number of predictors} \]
\[ N = \text{total sample size}. \]

### 2.2.5.2 Root Mean Squared Error

The root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values actually observed. Basically, the RMSE represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. RMSE is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent. The RMSE of predicted values $\hat{y}_t$ for times $t$ of a regression’s dependent variable $y$ is computed for $n$ different predictions as the square root of the mean of the squares of the deviations:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}}$$

### 2.3 Results

Regression trees generally do not have the same level of predictive accuracy as other approaches. A small change in the data can cause a large change in the final estimated tree. Random forests performed the best when comparing the root mean squared error (RMSE) for both sites, both seasons (wet and dry) and both response variables (carbon and water flux).
2.3.1 Alice Springs
"predicted vs actual figure" "table"

2.3.2 Howard Springs
"predicted vs actual figure" "table"

2.4 Discussion
Hamza & Larocque (2005) found random forest to be better than boosting, as what our study has shown too, but Gashler et al. (2008) showed that random forest can perform poorly if there are irrelevant variables in the data. Dietterich (2000) reviewed ensemble methods in the computer science literature. "table of pros and cons for each method"

2.5 Conclusions
Random forests are the best method to predict fluxes out of the tree-based methods based on RMSE and $R^2$, despite what season or site it is. We explore in detail the accuracy and variability of the predictions in Chapter 4.
Chapter 3

Identifying the group of driving factors of Carbon and Water fluxes at Alice Springs using Regression Trees

3.1 Introduction

Regression trees are mainly used to predict values based on a set of covariates but can also be used to see the main driving factors per terminal node. The basic idea is to partition the covariate space and to compute simple statistics of the dependent variable inside each cell. By partitioning the covariate space, we can use the variable importance output to indicate to us which are the driving factors of carbon and water fluxes respectively. By doing this, we will help further understand the dynamics in the environmental variables over Australian savannas and suggest us to partition our data first appropriately before feeding into the existing land surface models that try to predict carbon and water fluxes.

3.2 Methods

The model begins with the entire data set, and searches every distinct value of every input variable to find the predictor and split value that partitions the data into two regions such that the overall sums of squares error are minimized, which is also used in regression. This algorithm is a greedy top down algorithm, stopping only when each terminal node has fewer than some minimum number of observations.

Parameters of a regression tree include: Complexity parameter is a tuning parameter, that balances the depth of the tree and its goodness of fit to the training
data. Minsplit is the smallest number of observations in the parent node that could be split further. Maxdepth is the maximum depth a tree can be.

3.3 Results

Figure 3.1 shows as the output for a regression tree. The first split is partitioned by Radiation at 142. Figure 3.2 tells us NEE is first partitioned into day time and night time/early morning time dynamics.

Figure 3.1: Regression Tree with cp=0.001 for Alice Springs Wet Season with NEE (Net Ecosystem Exchange as proxy for carbon flux) as the response variable. The first split is partitioned by Radiation at 142. Figure 3.2 tells us NEE is first partitioned into day time and night time/early morning time dynamics.
Figure 3.2: First split at Radiation=142. Tells us NEE is first partitioned into day time and night time/early morning time dynamics. NEE below 0 represents a carbon sink and above 0 means carbon source.
Figure 3.3: Second split at LAI=0.52. Tells us NEE is secondly partitioned into understory dominance and overstorey dominance.
3.4 Discussion

Radiation was the most important driver of the fluxes by first partitioning the dataset into daytime and night time dynamics or carbon sink and carbon source dynamics. Leaf area index was the second important driver of the fluxes by partitioning understory and overstory contributions. The following drivers further explain the drivers of each cohort (branch).

3.5 Conclusion

Regression trees help us understand the dynamics of the drivers by the cohorts it produces. Can help reduce the uncertainty of future predictions by partitioning the datasets into these cohorts.
CHAPTER 4

Comparison of Carbon and Water Fluxes between two contrasting semi arid sites using Random Forests: Alice Springs and Howard Springs

4.1 Introduction

Models are built on all available knowledge gained by statistical analyses of data, physical-chemical-ecological knowledge and so on (Jorgensen & Fath, 2011). In order to address the research aim to develop new model parameterisations for missing processes and improve current model limitations, the OZFlux data should be statistically analysed to identify patterns and the potential explanations of the interactive impacts of climatic variables that predict the carbon and water fluxes. We want to know what are the most important drivers of savanna vegetation that should be incorporated in these models to predict the fluxes more accurately? Figure 4.1 shows the schematic flow of the research methodology on addressing the research aims.

4.2 Methods

The findings of Chapter 1 concluded that out of the tree based methods, Random Forests were the best method to predict carbon and water flux based on the relative Root Mean Square error criterion. We use this method in this chapter to train a model based on a specific period 2010 - 2016. Then use 2017 - 2018 data as the test set to see how well will this method predict future carbon and water predictions. No statistical model will be perfect in replicating results, but if a model is designed appropriately, it can be used to predict future outcomes. As a famous statistician
George E. Box once wrote in his 1987 book, Empirical Model-Building and Response Surfaces, "Essentially all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind...”.

4.3 Results

![Figure 4.1: Howard Springs NEE predictions where training set was 2012 to 2016 and test set is 2016 to 2017](image)

If we zoom in the predictions we can see that the model generally replicates the pattern well but does not predict the extremes well.
4.4 Discussion

The advantages of the random forest is that even the bias remains the same as that of a single decision tree, the variance decreases and thus we decrease the chances of overfitting. With random forests you don’t have to worry much about the assumptions of the model or linearity in the dataset.

The disadvantages is that: 1. There is a problem of interpretability with random forest. You can’t see or understand the relationship between the response and the independent variables. Understand that a random forest is a predictive tool and not
a descriptive tool. You get variable importance but this may not suffice in many analysis of interests where the objective might be to see the relationship between response and the independent features.

2. The time taken to train random forests may sometimes be too huge as you train multiple decision trees.

3. In case of a regression tree problem, the range of values the response variable can take is determined by the range of response values already available in the training dataset. Unlike linear regression, decision trees and hence random forest cannot take values outside the training data (extrapolation).

4.5 Conclusions

Random forests are a quick and dirty technique to make predictions about the response variable without worrying much about the assumptions or linearity of the dataset.
Chapter 5

Identifying environmental drivers in carbon and water fluxes over Australian Savannas using Principal Component Analysis vs. wavelet based Principal Component Analysis

5.1 Introduction

5.2 Methods

5.2.1 Multivariate Analysis

Majority of real world datasets are multivariate (simultaneous statistical analysis of a collection of variables). Multivariate analysis includes methods that are essentially exploratory and others that can be used for statistical inference. This research will generally need the methods to allow the detection of possibly unanticipated patterns in the data, opening up a wide range of competing explanations (Everitt and Hothorn, 2011). The first step is essentially exploratory and is a classification tool. Deconstructing the drivers of large-scale vegetation is critical to predicting productivity and managing land use changes that will affect regional vegetation cover in savannas. Cluster and Factor Analysis is a multivariate time-series dimension reduction technique to be used to identify the most important physical drivers of regional vegetation scale. Both methods are the main task of statistical data mining and a common technique for pattern recognition. Further development of the research methodology will be investigated during Stage 2 and the statistical R program will be used to analyse the data.
5.2.1.1 Principal Component Analysis

5.3 Results
Show results for PCA for all the sites on NATT Show results for wavelet PCA for all the sites on NATT

5.4 Discussion

5.5 Conclusions
Chapter 6

Discussion and Conclusion

6.1 Thesis Aim

Public awareness of the significance of CO$_2$ sinks has grown since the passage of the Kyoto Protocol - an international treaty that sets binding obligations on industrialised countries to reduce emissions of greenhouse gases by issuing tradable carbon credits (the amount in which that country is permitted to emit CO$_2$ which is based on the amount in which that country have absorbed CO$_2$ via carbon removal or carbon sink activity) amongst other industrialised countries. The Kyoto Protocol promotes the use of carbon sinks as a form of carbon offset.

6.2 Implications for vegetation modelling

6.2.1 Disturbance Effects on the Carbon Cycle

Savannas are subject to frequent natural and anthropogenic disturbances, causing substantial changes in carbon cycling processes over regions. One disturbance event can cause significant impact in savanna productivity. Disturbances include global climate change, agriculture, fire, windstorms, cyclones, grazing, termites invasives, insect epidemics, drought and floods. Anthropegeic land-use conversion from forests/grasslands to croplands, pastures and urban areas not only results in the net release of carbon to the atmosphere but also reduces ecosystem carbon residence time due to the elimination of carbon pools in plant wood biomass, and physical disturbance of long-term soil carbon pools (Hastings and Gross, 2012).
As a result of the frequent burning regimes (Beringer and Hutley et al., 2007), ecosystem photosynthetic capacity may be reduced by removed biomass and therefore has implications for their annual contribution to global productivity and subsequent carbon sequestration capacity (Grace et al., 2006).

Differences in Leaf Area Index (LAI), as well as ecophysiological properties (i.e., leaf stomatal conductance, light and water-use efficiency), will determine the amount of water transpired to the atmosphere, and changes in these factors will lead to changes in the surface-energy balance. LAI can be modified by disturbances such as fire, which occurs frequently in savannas.

There is potential research in the stochastic modelling of fire disturbances in savannas which have not been explored yet. Modelling and theoretical analysis of disturbance effects on the ecosystem carbon cycle is still an infant stage challenging research area. Berlinger and Hutley et. al. (2007) have previously studied about the impact of savanna fires on net ecosystem productivity. Natural disturbances, global change and human intervention may trigger the state changes, resulting in major impacts on the ecosystem carbon cycle. We need innovative methods to examine conditions and processes leading to the state changes (Hastings and Gross, 2012).

### 6.2.2 Constraints in Vegetation

Increasing atmospheric CO$_2$ will have differential effects on wood plants (C3 pathway) vs grasses (C4 pathway) that comprise savanna vegetation types (Donohue et al., 2009) and it is critical to account for these distribution in models since most do not separate over-and understory vegetation or deal with varying tree and grass composition spatially.
6.2.3 Coupling Interactions of Complex Models with Multiple Time Scales

Working with complex Earth system models require not just high computational power that will expedite and enable our modelling activities but require knowledge expertise in land surface and carbon cycle modelling, plant phenology, model evaluation and complex systems science. In terrestrial ecosystems, the efficiency is estimated by models that include many factors regulating productivity and the models of primary production in ecosystems cannot rely on a single limiting factor but instead must consider multiple potentially limiting processes. All measurements have a significant scale associated with it and they are not comparable to one another without a significant effort at spatial and temporal scaling. The studies of different scales and processes provide a richer view of primary production (Weathers et al., 2012).

6.3 Concluding remarks

Original contribution to knowledge...
References


