A model inter-comparison study to examine limiting factors in modelling Australian tropical savannas

Authors: Rhys Whitley[1], Jason Beringer[2], Lindsay Hutley[3], Gabriel Abramowitz[4], Martin G. De Kauwe[1], Remko Duursma[5], Bradley Evans[6], Vanessa Haverd[7], Longhui Li[8], Youngryel Ryu[9], Benjamin Smith[10], Ying-Ping Wang[11], Mathew Williams[12], Qiang Yu[7]

Institutions:

[1]Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

[2]School of Earth and Environment, University of Western Australia, Crawley, WA 6009, Australia

[3]School of Environment, Charles Darwin University, Casuarina, NT 0810, Australia

[4]Climate Change Research Centre, University of New South Wales, Kensington, NSW 2033, Australia

[5]Hawkesbury Institute for the Environment, University of Western Sydney, Penrith, New South Wales 2751, Australia

[6]Faculty of Agriculture and Environment, University of Sydney, Eveleigh, NSW 2015, Australia

[7]CSIRO Ocean and Atmosphere, Canberra 2601, Australia

[8]School of Life Sciences, University of Technology Sydney, Ultimo, NSW 2007, Australia

[9]Department of Landscape Architecture and Rural Systems Engineering, Seoul National University, Seoul, South Korea

[10]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden

[11]CSIRO Ocean and Atmosphere, Aspendale, Victoria 3195, Australia

[12]School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom

Corresponding author:

Name: Rhys Whitley

Email: rhys.whitley@mq.edu.au

Address:
Department of Biological Sciences
Macquarie University
North Ryde, NSW
2109, Australia

Abstract:

Savanna ecosystems are one of the most dominant and complex terrestrial biomes that derives from a distinct vegetative surface comprised of co-dominant tree and grass populations. While these two vegetation types co-exist functionally, demographically they are not static, but are dynamically changing in response to environmental forces such as annual fire events and rainfall variability. Modelling savanna environments with the current generation of terrestrial biosphere models (TBMs) has presented many problems, particularly describing fire frequency and intensity, phenology, leaf biochemistry of C_3 and C_4 photosynthesis vegetation, and root water uptake. In order to better understand why TBMs perform so poorly in savannas, we conducted a model inter-comparison of 6 TBMs and assessed their performance at simulating latent energy (LE) and gross primary productivity (GPP) for five savanna sites along a rainfall gradient in northern Australia. Performance in predicting LE and GPP was measured using an empirical benchmarking system, which ranks models by their ability to utilise meteorological driving information to predict the fluxes. On average, the TBMs performed as well as a multi-linear regression of the fluxes against solar radiation, temperature and vapour pressure deficit, but were outperformed by a more complicated nonlinear response model that also included the leaf area index (LAI). This identified that the TBMs are not fully utilising their input information effectively in determining savanna LE and GPP, and highlights that savanna dynamics cannot be calibrated into models and that there are problems in underlying model processes. We identified key weaknesses in a model's ability to simulate savanna fluxes and their seasonal variation, related to the representation of vegetation by the models and root water uptake. We underline these weaknesses in terms of three critical areas for development. First, prescribed tree-rooting depths must be deep enough, enabling the extraction of deep soil water stores to maintain photosynthesis and transpiration during the dry season. Second, models must treat grasses as a co-dominant interface for water and carbon exchange, rather than a secondary one to trees. Third, models need a dynamic representation of LAI that encompasses the dynamic phenology of savanna vegetation and its response to rainfall interannual variability. We believe this study is the first to assess how well TBMs simulate savanna ecosystems, and that these results will be used to improve the representation of savannas ecosystems in future global climate model studies.

Introduction

Savanna ecosystems are a diverse and important biome that play a significant role in global land-surface processes (van der Werf et al., 2008). Globally, they occupy regions around the wet-dry tropical to sub-tropical equatorial zone, covering approximately 15 to 20% of the terrestrial surface and contribute ~30% to global net primary production (Grace et al., 2006; Lehmann et al., 2014). Savannas are water-limited ecosystems where rainfall is often seasonal or monsoonal, and have a spatial extent that can cover an area with annual rainfall in the range of 500 to 2000 mm (Bond, 2008; Kanniah et al., 2010; Sankaran et al., 2005). The variability in the amount and timing of annual rainfall, coupled with local topo-edaphic properties, and the frequency and intensity of seasonal fires strongly influences the structure and function of savanna vegetation (Beringer et al., 2007; Kanniah et al., 2010; Ma et al., 2013; Sankaran et al., 2005). Savannas are characterised by a multi-layer stratum of vegetation, where an open and discontinuous canopy overstorey is seasonally dominated by understorey grasses (Scholes and Archer, 1997). These tree and grass layers are distinctly and functionally different, fixing carbon using different photosynthetic pathways, C_3 and C_4 photosynthesis respectively (Bond, 2008; Scholes and Archer, 1997; Williams et al., 1996b). The canopy overstorey can be either evergreen or deciduous (depending on the evolutionary history), while the grass understorey is annual: active only in the wet season and senescing at the end of this period (Williams et al., 1996b). Consequently, water, carbon and nutrient cycling in savannas is largely determined from the balance and co-existence of these two life forms (Lehmann et al., 2009; Sankaran et al., 2005).

Given the complex nature of savannas, modelling the land surface exchange and vegetation dynamics for this biome is challenging for terrestrial biosphere models (TBMs). Here we define TBMs to broadly encompass stand, land-surface, and dynamic global vegetation models (Pitman, 2003). Most land surface schemes that feed into larger earth system models use simplistic representations of vegetation, and these will have difficulty describing the complex structure of savannas ecosystems. Such issues may be: simplistic assumptions in relation to rooting depth and inadequate responses to drought (De Kauwe et al., 2015; Li et al., 2012); ignoring the multilayered nature of savannas and the differing structural (including radiation), functional (including different plant functional types) and phenological differences (Whitley et al., 2011); and in some cases neglecting the C_4 photosynthetic pathway entirely (Parton et al., 1983; Schymanski et al., 2007) It is therefore critical that TBMs meet the challenges that savanna dynamics present if water and carbon exchange are to be correctly simulated in response to global change.

Despite these issues, there have been significant advances in modelling savanna dynamics in recent years, and this has been focused on integrating important features specific to savanna ecosystems, namely frequent fire and tree-grass competitive interactions, processes that shape savanna structure and function (Haverd et al., 2015; Higgins and Scheiter, 2012; Scheiter and Higgins, 2007; Scheiter et al., 2014; Simioni et al., 2003). Nevertheless, little work has been undertaken to critically evaluate the performance and processes of TBMs when used to capture water and carbon cycling in savannas, most notably in west Africa (Simioni et al., 2000) and Australia (Schymanski et al., 2007, 2008, 2009; Whitley et al., 2011). Many global ecosystem models moreover use broad plant functional types (PFTs) with single parameter values to describe whole biomes (Pitman, 2003), making them unable to represent changing vegetation structure (tree:grass ratio) in the continuum of grassland to woodland savanna. Approaches have been developed that can account for savanna dynamics, such as using mixed tiles, whereby trees and grasses are simulated as separate surfaces that are then aggregated together (Kowalczyk et al., 2006). However, this approach fails to capture the competition between trees and grasses for light, water and nutrient resources.

In this study, we take 6 TBMs of distinctly different conceptual frameworks, and assess their ability at simulating savanna water and carbon exchange along the North Australian Tropical Transect (NATT) that is defined by a strong rainfall gradient. Australian tropical savannas can be considered largely intact compared to South American and African savannas, and provide a 'living laboratory' to understand the links between vegetation structure and function and how it responds to environmental change (Hutley et al., 2011). We challenge the models by evaluating them along the rainfall gradient, which extends over a broad biogeographical extent and strong interannual variability in climate (Koch et al., 1995). The aim of this study is to highlight critical processes that may be missing in current TBMs and are required to adequately simulate savanna ecosystems. Specifically, we examine whether a TBM's structural framework, such as the representation of the understorey grasses (C_4 photosynthesis), tree rooting depth, and description of phenology (prescribed vs. dynamic) can adequately replicate observed carbon and water fluxes. To achieve this we measure the performance of each TBM by comparing its predictions to a set of empirical benchmarks that describe *a priori* expected levels of model performance. We identify regions of low performance among sites and seasons, to diagnose under what climate conditions reduced model performance occurs. We then infer what processes (present or missing) may be the cause for reduced performance when applied to savannas ecosystems. Our intention is that these results can be used to flag high priorities for future development by the terrestrial biosphere modelling community.

2. Methodology

2.1 Observational data

The North Australian Tropical Transect (NATT) is a sub-continental rainfall gradient in the wet-dry tropical climate zone of Northern Australia, which encompasses a distance of approximately 1000 km over a latitudinal range of -12 to -23 °S and a decline in mean annual precipitation (MAP) from 1700 mm to 300 mm (Hutley et al., 2011). It is one of three savanna transects established in the mid 1990's, forming part of the International Geosphere Biosphere Program (IGBP) along with the SAvannas in the Long Term (SALT) transect in West Africa and the Kalahari Transect (KT) in Southern Africa (Koch et al., 1995). Soils range from sand dominated red Kandosols to black, cracking clay soils that are more extensive in the southern end of the NATT that are limiting to woody plant growth (Hutley et al., 2011; Williams et al., 1996b). Kandosols are ancient and weathered, such that they have been leached of nutrients by the large monsoonal rainfall (McKenzie et al., 2004). Close to the northern coastline, vegetation is comprised primarily of evergreen *Eucalyptus* and *Corymbia* tree species that overly an understorey of C_4 *Sorghum* and *Heteropogon* spp. grasses. Inland, tree biomass, leaf area index (LAI) and cover tends to decline and by -18 °S savanna vegetation transitions to less dense *Acacia* woodlands, shrublands and grasslands that are dominated by *Astrebla* grass species (Hutley et al., 2011). Fires occur regularly in these environments, increasing in frequency with higher rainfall (MAP > 1000 mm), and are fuelled by the accumulation of understorey C_4 grasses that cure in the dry season (Beringer et al., 2014; Russell-Smith and Edwards, 2006).

The five flux tower sites along the NATT used in this study are outlined in Table 1, and describes stand soil and vegetation characteristics, as well as a summary of local meteorology (Hutley et al., 2011). These sites represent a sampling of savanna environments covering a wide range of MAP and a much smaller range of mean annual temperature (MAT) (Fig. 1). At each site, an eddy covariance system was used to measure the ecosystem-atmosphere exchange of radiation, heat, water and CO_2. Quality assurance and control (QA/QC) and corrections on the fluxes were carried out on the 30 minute dataset using the OzFlux QC/QA protocol (v2.8.5), developed by the OzFlux community under creative commons licensing (www.ozflux.org.au) (see Eamus et al., 2013). Missing or rejected data were gap-filled using the DINGO (Dynamic INtegrated Gap filling and partitioning for Ozflux) system (see Moore *et al.* this issue). Gross primary productivity (GPP) was not observed but determined from the difference between measured net ecosystem exchange (NEE) and modelled ecosystem respiration (Re). Values of Re were determined by assuming nocturnal NEE equals Re under the conditions for sufficient turbulent transport. Values that meet these requirements are then used to make daytime predictions of Re, using an artificial neural network (ANN), with soil moisture and temperature, air temperature, and the normalised difference vegetation index (NDVI) used as predictors. Additionally, the effect of fire on the water and carbon fluxes are quantified and incorporated into the datasets accounting for the nonlinear response in productivity (becoming a carbon source) during the post-fire recovery period (Beringer et al., 2007). Because the TBMs used here do not attempt to simulate stochastic fire events (and other disturbance regimes), these post-fire recovery periods were removed when determining the benchmarks and model performance as described below.

Finally, we use the definitions for water and carbon exchange as outlined by Chapin et al. (2006), whereby the sub-daily rate of GPP is expressed in μmol m^{-2} s^{-1} and uses a negative sign (-) to denote the removal of CO_2 from the atmosphere. Similarly, LE is expressed in terms of energy as W m^{-2} and uses a positive sign to denote the addition of H_2O to the atmosphere.

2.2 Terrestrial biosphere models

The 6 TBMs used in this study cover a wide spectrum of characteristics of operation, scale and function, and include differences in operational time-step (30min vs. daily), scope of simulated processes (soil hydrology, static or dynamic vegetation, multi-layer or big leaf description of the canopy) and intended operational use (coupled to earth system models, offline prediction, driven by remote sensing products). These characteristics along with what we define as a model ‘functional class’ are given in Table 2 and are defined as follows. Stand models (SMs) give detailed multi-layer descriptions of canopy and soil processes for a particular point, operating at a sub-daily time-step

(Soil-Plant-Atmosphere model: SPA, and MAESPA). Land-Surface models (LSMs) operate at the same temporal resolution as SMs, but adopt a simpler representation of canopy processes, allowing them to be applied spatially (Community Atmosphere Biosphere Land Exchange model; CABLE, and BIOS2; a modified version of CABLE). Dynamic Global Vegetation Models (DGVMs) simulate water, carbon much like the other models, but simulate dynamic rather than static vegetation that changes in response to climate and disturbance (Lund-Potsdam-Jena General Ecosystem Simulator; LPJGUESS). Lastly, Remote Sensing models (RSMs) are driven by remotely sensed atmospheric products, and infer water-stress of vegetation through changes in fractional cover rather than detailed soil hydrological processes (Breathing Earth System Simulator; BESS). Some of the TBMs share similar structural frameworks in parts: for example, both SPA and MAESPA use similar below-ground soil hydrology and root-water uptake schemes, while BIOS2 is fine-spatial-resolution (0.05 degree), offline modelling environment for Australia, in which predictions of CABLE (with alternate parameterisations of drought response and soil hydrology) are constrained by multiple observation types (see Haverd *et al.* 2013). Although these similarities reduces the number of truly, functionally, independent models used in the experiment, the presence of such overlap can be useful in identifying if particular frameworks are the cause for model success or failure.

2.3 Experimental protocol

All TBMs were parameterised for each of the five savanna sites using standardised information on vegetation and soil profile characteristics (Table 1). For TBMs that required them, parameter values pertaining to leaf biochemistry, such as maximum Rubisco activity (V_{cmax}) and leaf nitrogen content per leaf area (N_{area}), were assigned from Cernusak et al. (2011), who undertook a physiological measurement campaign during the SPECIAL program (Beringer et al. 2011). Parameters relating to soil sand and clay content were taken from the Australian Soil Classification (Isbell, 2002), while root profile information was sourced from Chen et al. (2003) and Eamus et al. (2002). Each TBM was setup to describe a C_3 evergreen overstorey with an underlying C_4 grass understorey, and conforms well with the characteristics of savannas in Northern Australia (Bowman and Prior, 2005). All TBMs (excluding LPJGUESS) prescribed LAI as an input, to characterise the phenology of vegetation at each site. In these cases LAI was determined from MODIS derived approximations that were well matched to ground-based estimations of LAI at the SPECIAL sites (Sea et al., 2011). The fraction of C_3 to C_4

vegetation was handled differently by each model and was determined for each as follows. For MAESPA and SPA, the models allowed for time-varying tree and grass fractions to be assigned as direct inputs, and these time-varying fractions were determined using the method of Donohue et al. (2009). BIOS2 similarly used the same method to extract time-varying fractions, while CABLE used a static fraction that did not change. The BESS model derived the C_3:C_4 fraction from the C_3 and C_4 distribution map of Still et al. (2003), while for LPJGUESS this fraction is a prognostic determination resulting from the competition between trees and grasses (see Smith et al., 2001). Model simulations were driven using observations of solar radiation, air temperature, relative humidity (or vapour pressure deficit; VPD), rainfall, atmospheric CO_2 concentration and LAI (if prescribed), and included a spin-up period of 5 years to allow internal states, such as the soil water balance and soil temperature to reach equilibrium. The exception to the above was the BIOS2 model, which was run using gridded meteorological inputs and had its model parameters optimised through a model-data fusion process (see Haverd et al., 2013).

Simulations for each savanna site covered a period of 2 to 10 years depending on the availability of data from each flux site (Table 1) and results were standardised to the ALMA (Assistance for Land-surface Modelling Activities) convention. Model predictions of LE and GPP were then evaluated against local observations at each site from the eddy covariance datasets and benchmarked following the methodology proposed by the PALS Land SUrface Model Benchmarking Evaluation PRoject (PLUMBER) (Abramowitz, 2012; Best et al., 2015) as described below.

2.4 Empirical benchmarking

The paradigm for model assessment set out in the Protocol for the Analysis of Land-surface models (PALS) (Abramowitz, 2012) suggests that model assessment is more meaningful when *a priori* expectations of performance in any given metric can be defined. Such benchmarks can be created using simple empirical models, built on statistical relationships between the fluxes and drivers, and establish the degree to which models utilise the information available in their driving data about the fluxes they aim to predict. Additionally, these empirical models are simple in the sense that they are purely instantaneous response to time-varying meteorological forcing and contain no internal states or expression of ecophysiological processes. This is in comparison to

TBMs that are complex, having some 20+ soil and vegetation parameters, internal states, partitioning of light, as well as soil and vegetation, carbon and nitrogen pools (Abramowitz et al., 2008).

We created a set of 3 empirical models of increasing complexity following the procedure of Abramowitz (2012), which we compared with the TBMs. The first benchmark (emp1) is simply a linear relationship between a turbulent flux (LE or GPP) and downward short-wave radiation (R_s). The second benchmark (emp2) is slightly more complex, and is a multi-linear regression between a flux and R_s, air temperature (T_a), and vapour pressure deficit (VPD). Finally, the third benchmark (emp3) is the most complex and is a nonlinear regression of the fluxes against R_s, T_a, VPD and LAI, determined from an ANN. This benchmark is constructed using a self-organising linear output map that clusters the four covariates into 10^2 distinct nodes and performs a multi-linear regression between the fluxes and the 4 covariates at each node, resulting in a nonlinear (piece-wise linear) response to the meteorological forcing data (Abramowitz et al., 2008; Hsu, 2002). In a departure from Abramowitz (2012), we include LAI as an additional covariate, as the seasonal variance of savanna water and carbon exchange is strongly coupled to the phenology of the grasses and to the deciduous and semi-deciduous woody species (Moore et al., *this issue*). The seasonal behaviour of the empirical benchmark drivers along the NATT can be referred to in the supplementary information. Empirical benchmarks are created for each of the five flux sites using non gap-filled data, and are parameterised *out-of-sample*, such that they use data from all sites except the one in question. For example, the Howard Springs empirical benchmark models would use information from Adelaide River, Daly Uncleared, Dry River and Sturt Plains to establish their parameter values, but would exclude Howard Springs itself. Constructing the benchmarks out-of-sample results in what is effectively a generalised response to an independent dataset. Once the empirical models were calibrated for each site, benchmarks were then created for both fluxes using the same meteorological forcing used to run the TBMs.

Finally, we assess ecosystem model performance in terms of a ranking system, following the PLUMBER methodology of Best et al. (2015). The performance of each individual ecosystem model in predicting both LE and GPP at each site was determined using four statistical metrics that describe the mean and variability of a model compared to the observations. These metrics included the correlation coefficient (r), standard deviation (sd), normalised mean error (NME), and mean bias error (MBE) (see Table B1). Similarly, the same metrics were determined for each of the 3 benchmarks at each savanna site. Each TBM was then ranked against the benchmarks (independently of the other models) for each of the metrics listed above., where the ranking is between 1 and 4 (1 model + 3 benchmarks) and the best performing model for a given metric is ranked as 1. An average ranking is then determined across all metrics for each TBM and all benchmarks to give a final ranking of performance for each savanna site. The ranks denote the number of metrics being met by the models and are not a measure of the smallest absolute error. In determining the average ranks, the metrics were evaluated at the daily time scale, as this was the lowest temporal resolution common amongst the 6 TBMs. Additionally, days where either driver or flux had been gap-filled were removed in the determination of these metrics. Herewith we use the term *performance* to relate to how well the TBMs compare to the benchmarks as expressed by the ranks.

3. Results

3.1 Model predictions

Figure 2 shows the daily time-course of LE and GPP from the flux tower, models, and benchmarks at each of the five savanna sites. Models, benchmarks and observations are represented as a smoothed time-series (7-day running mean) and have been aggregated into an ensemble year to express the typical seasonality of savanna water and carbon exchange. Visually, the TBMs showed varying levels of performance across the rainfall gradient. None of the models showed a clear consistency in simulating either flux and responded differently to the meteorological drivers across sites. Additionally, some of the models, such as CABLE and LPJGUESS, showed difficulty in simulating the seasonality of the fluxes across the transect, particularly GPP. Differences among model simulated LE and GPP were larger in the wet season than the dry season. However, modelled LE and GPP appeared to co-vary quite strongly; overall both fluxes were underestimated across sites by most models. Simulations by SPA and MAESPA were the exception to this, broadly capturing tower GPP despite consistently underestimating LE across sites.

Figure 3 shows the probability density functions (PDFs) for the wet (Nov – Apr) and dry season (May – Oct) fluxes at each savanna site. Tower and model PDFs were determined by binning each flux into the respective seasons and using kernel density estimation (Bashtannyk and Hyndman, 2001) to determine smoothed distributions. The shape and spread of the distributions highlight possible biases in the models (over- or underestimating the tower fluxes), as well as their ability to capture the spread of values. The PDFs for the tower fluxes tended to shift to low values and became narrower as annual rainfall declined, and this was most prominent in the dry season. By contrast, the PDFs of the model simulations did not replicate this trend, being mostly stationary (i.e. density of values do not shift) across sites, especially for wet season GPP. For example, the LPJGUESS and MAESPA models showed distributions with higher densities at low LE (20 – 40 W m^{-2}) across all sites and seasons, whereas during the wet season almost all models except MAESPA and SPA had higher densities of GPP at low values (3 – 4 μmol m^{-2} s^{-1}). In models that did replicate the observed spread and shape of the tower fluxes (CABLE, LPJGUESS), PDFs were biased towards lower values particularly for LE, where there was little overlap. A consistent pattern emerged among the models, suggesting that drier conditions resulted in PDFs with higher densities at low values for both fluxes.

The benchmarks showed that there was enough information in R_s (emp1) to predict the magnitude and daily time-course of the tower fluxes (data not shown). However, additional information contained in T_a and VPD (emp2) was needed to capture the water and carbon flux seasonality. Importantly additional phenological information (provided as LAI in emp3) was required in order to fully capture the seasonality of the savannas fluxes and provide predictions with the lowest absolute error. This indicates that in order for models to achieve the best possible performance simulating savanna ecosystems they will require correct information on phenology.

3.2 Residual analysis

Figure 4 shows how model structure (internal processes) is affecting predictions of savanna fluxes across the rainfall gradient. To do this we examine the standardised model residuals from each TBM where the scale is expressed in terms of standard deviations. Figure 4 shows the LE and GPP residuals at each savanna site represented in two ways: (i) against time, and (ii) against the flux prediction itself. Plotting the residuals against time provides an effective way of examining how a model responds to progressive changes in the environment, while plotting the residuals against the model prediction gives a measure of a model’s bias (Medlyn et al., 2005). A linear regression has been applied to the residuals versus flux prediction scatter plots in order to better visualise the degree of bias.

Model residuals for both LE and GPP illustrated a larger tendency towards underestimation in the wet than in the dry season. The LE residuals showed a larger between-day variance than the GPP residuals in the wet season. A possible explanation for this is that LE is the sum of multiple evaporation components: vegetation transpiration, soil evaporation, and wet canopy evaporation, i.e. 3 sources for potential error. However, the error and variance of the LE residuals reduced into the dry season, and as the site itself became drier. It is likely that this reduced error and variance in LE was a combination of the increasing contribution of woody transpiration to bulk land-surface LE as the grass transpiration and soil evaporation declined (via senescence and low surface soil moisture respectively). Despite this reduction in error, there was still an underestimation of dry season LE that was larger at wetter sites, suggesting that processes affecting simulated plant hydrology (e.g. root-water uptake, plant conductance, rooting depth) are inadequate for some of the TBMs (e.g. LPJGUESS, CABLE). The GPP residuals broadly showed a similar response over time, with the exception that the between-day variance was lower and did not change between seasons. Again, moving into the dry season where the grasses become inactive, the residuals go to zero. This suggests that tree transpiration and photosynthesis are in general being adequately modelled, albeit with some underestimation. A possible explanation for this is too shallow tree rooting depths at the wetter sites. The large error in the wet season that is visible across the transect points to a structural issue with how the grasses are modelled.

All models show different biases in their predictions (indicated by the slope), and these biases in almost all cases change with site and are not consistent among the models. The biasing for the CABLE and BIOS2 models is similar and positive for LE, yet divergent for GPP. For the LPJGUESS, MAESPA and SPA models, there is significant negative bias in predicted LE, while for the GPP predictions the bias is smaller. The BESS model shows both negative and positive biasing, but is small and somewhat consistent among sites.

3.3 Model performance

Figure 5 shows a comparison of individual TBM performance ordered by site from wettest (Howard Springs) to driest (Sturt Plains) and in terms of their annual, wet and dry season predictions for each flux. Despite differences in model complexity (Table 1), the TBMs showed a similar performance across sites and seasons. For almost all sites, the TBMs outperformed the emp1 benchmark for annual flux predictions (Fig. 5a). However, there were some exceptions to this, and good performance in one flux did not necessarily result in good performance in the other. For example, MAESPA was unable to beat the emp1 benchmark for LE at sites where MAP > 1000 mm, but performed better than the emp2 benchmark for GPP. In general, there was a slight pattern of increased model performance as annual rainfall declined, though with a degree of site-to-site variability in the rankings for some of the TBMs.

In order to examine how seasonal changes affect model performance, we additionally determined the metrics and rankings for the wet and dry season periods (Fig. 5b-c). Seasonal differences were immediately obvious. Model performance for wet season LE and GPP was low to moderate, and the majority of the TBMs showed a performance that ranged between the emp1 and emp2 benchmarks. In contrast, there were noticeable improvements to dry season model performance amongst the TBMs. For dry season LE, half the models (BIOS2, BESS, and SPA) were able to consistently outperform the emp2 benchmark, and come close to meeting the same number of metrics as the emp3 benchmark particularly at the drier sites. In comparison, predicted dry season GPP saw a larger enhancement in model performance, with TBMS more frequently outperforming the emp2 benchmark and even some outperforming the emp3 benchmark (LPJGUESS, BESS, and SPA at the Daly Uncleared site). The exception to all this was the CABLE model, which showed surprisingly little loss or gain in performance despite the season. The results give an indication that as a whole, input information was better utilised by each TBM at drier sites and in the dry season, suggesting that there are problems in wet season processes.

4. Discussion

The NATT, which covers a marked rainfall gradient, presents a natural 'living laboratory' with which a models ability to simulate fluxes in savanna ecosystems may be assessed. Our results have highlighted that there is a clear failure of the models to adequately perform at predicting wet season dynamics, as compared to the dry season, and suggests that modelled processes relating to the C_4 grass understorey are insufficient. This highlights a key weakness of this group of TBMs, which likely extends to other models outside of this study. The inability of these TBMs to capture wet season dynamics is highlighted by the benchmarking, where the performance for many of the models was at best equivalent to that of a multi-linear regression against R_s, T_a and VPD (emp2) and in some cases no better than a linear regression against R_s (emp1). Given that this subset of TBMs are sophisticated process-based models that represent our best understanding of land-surface, atmospheric exchange processes, we would expect them to perform as well as a neural network prediction (emp3). Consequently there is an evident underutilisation of the driving information (i.e. a failure to describe the underlying relationships in the data) impeding the performance of these models when predicting savanna fluxes. However, there were instances where some of the TBMs were able to reach similar levels of performance with the emp3 benchmark, and strongly suggests that each of these models is capable of replicating savanna dynamics under certain conditions (e.g. during the dry season).

Our results suggest that errors among models are likely to be systematic, rather than related to calibration of existing parameters. For example, BIOS2 had previously optimised model parameters for Australian vegetation (see Haverd et al.2013), but was still unable to out-perform the emp3 benchmark in most cases, although performed better than an un-calibrated CABLE, to which it is functionally similar. Similarly, MAESPA and SPA, which used considerable site characteristic information to parameterise their simulations, did not significantly outperform un-calibrated models (e.g. CABLE). Additionally, despite these models using the same leaf, root and soil parameterisations, both SPA and MAESPA displayed markedly different performances in predicting LE. Consequently, improving how models represent key processes that drive savanna dynamics is critical to improving model performance across this ecosystem.

There is certainly enough information in the time-varying model inputs to be able to adequately simulate wet and dry season dynamics, as is evidenced by the benchmarks. We therefore consider the implications of our results, and present possible reasons below for why this group of TBMs is failing to capture water and carbon exchange along the NATT, and make suggestions as to how this could be improved.

4.1 Water access and tree rooting depth

During the late dry season surface soil moisture in the sandy soils declines to less than 3% volumetric water content, with an equivalent matric potential of 3 to 4 MPa (Prior et al., 1997). During this seasonal phase, the grass understorey becomes inactive and LE can be considered as equivalent to tree transpiration, such that it is the only active component during this period (O'Grady et al., 1999). Using this equivalence, one can infer the relative effect that rooting depth has on LE during this period. Previous studies have shown that for these savanna sites along the NATT, tree transpiration is maintained throughout the dry season by deep root systems that access deep soil-water stores, which in turn are recharged over the wet season (Eamus et al., 2000; Hutley et al., 2001; Kelley et al., 2007; O'Grady et al., 1999). In order for models to perform well they will need to set adequate rooting depths and distributions, along with root water uptake process, to enable a model response to such seasonal variation. Examining performance across the models, we can infer this to be a key deficiency. As expected, TBMs that prescribed shallow rooting depths (e.g. LPJGUESS) did not simulate this process well, and underestimated dry season LE at 3 of the 5 savanna sites by up to 30 to 40%. The two sites at Adelaide River and Sturt Plains were an exception to this with the TBMs displaying a low residual error, which is likely to be a consequence of heavier textured soils and trees at these sites having shallow root profiles. At Adelaide River shallow root profiles are a consequence of shallow, heavier textured soils, however dry season transpiration is sustained due to the presence of saturated yellow hydrosol soils. Sturt Plains is a grassland (end member of the savanna continuum) where C_4 grasses dominate and no trees are present such that transpiration is close to zero in the dry season. The few small shrubs that have established have shallow root profiles that have adapted to isolated rainfall events driven by convective storms (Eamus et al., 2001; Hutley et al., 2001, 2011). Consequently, the TBMs would be expected to perform better at these sites, as water and carbon exchange will be modulated by the soil-water status of the sub-surface soil layers. For the other sites, models which assumed a root depth > 5 m (BIOS2, SPA and MAESPA), showed the most consistent performance in predicting dry season LE, and we suggest for seasonally water-limited ecosystems, such as savanna, that deeper soil water access is critical. Our results highlight the need for data with which to derive more mechanistic approaches to setting rooting depth, such as that of Schymanski et al. (2009).

Interestingly, a low residual error for LE in the dry season, did not translate as good performance in the overall model ranking. This suggests that other processes along the soil-vegetation-atmosphere continuum need to be considered to improve simulated woody transpiration. Such processes may include root-water uptake (distribution of roots and how water is extracted), and the effect of water stress and increased atmospheric demand at the leaf-level (adjustment of stomatal conductance due to changes in leaf water potential). More detailed model experiments that examine how each TBM simulates these processes would help identify how they can be improved.

An exception to the above is the BESS model, which forgoes simulating belowground processes of soil hydrology and root-water uptake entirely. Rather, this model assumes that the effects of soil-moisture stress on water and carbon exchange is expressed through changes in LAI (and by extension V_{cmax}), which acts as a proxy for changes in soil moisture content (Ryu et al., 2011). The fact that BESS performed moderately well along the NATT, coupled with the fact that tree transpiration continues through the dry season suggests that there may be enough active green material for remote sensing proxies of water-stress to generally work rather well for savanna ecosystems. It is notable that BESS overestimated both GPP and ET in dry season at the driest site, Sturt Plains (Fig 2e), implying that greenness detected by satellite remote sensing might not capture carbon and water dynamics well in such a dry site.

4.2 Savanna wet season dynamics

The relative performance of the TBMs at predicting LE was much poorer in the wet season compared to the dry season. The reason for this difference is that wet season LE is the sum of woody and herbaceous transpiration (E_{veg}) as well as soil and wet-surface evaporation (E_{soil}); in contrast dry season LE is predominantly woody transpiration as described previously. During the wet season, up to 75% of total LE arises from understorey herbaceous transpiration and soil evaporation (Eamus et al., 2001; Hutley et al., 2000; Moore et al., *this issue*) and of this fraction the C_4 understorey contributes a significant daily amount (Hutley et al., 2000). In the absence of observations of understory LE it can be difficult to determine whether grass transpiration is being simulated correctly. However, separating out the components of wet season LE into soil and vegetation can help identify which of these components are causes for error.

Separating the outputs of simulated E_{veg} and E_{soil} from each TBM (excluding BESS which did not determine these as outputs during the study) shows that simulated wet season E_{veg} was particularly low for a lot of the models, despite high LAI and non-limiting soil-water conditions (Figure 6). A previous study at Howard Springs by Hutley et al. (2000) observed that during the wet season, the grass understorey could transpire ~2.8 mm d^{-1}, while the tree canopy transpired only 0.9 mm d^{-1} (E_{veg} = 3.7 mm d^{-1}). Of the 6 TBMs at

Howard Springs, only CABLE and SPA were able to predict an E_{veg} close to this level, while the other models predicted values closer to tree transpiration (i.e. an under-estimate). This pattern is similar for other NATT sites, where predicted wet season E_{veg} remained low and was dominated by E_{soil} at the southern end of the NATT. An underestimation of wet season LE could be due to underestimated E_{soil} in some of the models. Conversely, CABLE and BIOS2 predicted a higher E_{soil} than the other models, and this could be a reason for their higher LE performance during the wet season. Although E_{soil} has been reported to reach as high as 2.8 mm d^{-1} at Howard Springs (Hutley et al., 2000), predicted E_{soil} by these models may still be overestimated, given that vegetation cover during this period is at a seasonal peak (limiting energy available at the soil surface) and transpiration is only limited by available energy not water (Hutley et al., 2000; Ma et al., 2013; Schymanski et al., 2009; Whitley et al., 2011). Given the limited data for E_{soil} along the NATT, it is difficult to determine how large E_{soil} should be. However, the ratios displayed by the TBMs appear to be reasonable though, with vegetation acting as the predominant pathway for surface water flux.

Grass transpiration is thus clearly being under-represented by most of TBMs, and reasons for this could be due to multiple factors that we discuss below. The evolution of C_4 grasses to fix carbon under low light, low CO_2 concentrations and high temperatures has resulted in a gas-exchange process that is highly water-use efficient (von Caemmerer and Furbank, 1999). Consequently, this life form is abundant in tropical, water-limited ecosystems, where it can contribute to more than 50% of total LAI (2.0 to 2.5), particularly at high rainfall sites (Sea et al., 2011). The annual strategy of the C_4 grasses at these sites is to indiscriminately expend all available resources to maximise productivity during the monsoon period, for growth and to increase leaf area. This therefore allows grass transpiration to exceed tree transpiration during the peak wet season as evergreen trees will be more conservative in their water-use, allowing them to remain active in the dry season (Eamus et al., 2001; Hutley et al., 2000; Scholes and Archer, 1997). Following this logic, our results suggest that the TBMs are either: i) incorrectly ascribing leaf area to the understorey (i.e. the C_4 fractional cover is too low), ii) incorrectly describing the C_4 leaf-gas exchange physiology, iii) incorrectly describing the understory micro climatic environment (R_s, T_a, VPD), or iv) a combination of these causes. Furthermore, it should be noted that the TBMs used in this study are not truly modelling grasses, but approximating them. Grasses are effectively simulated as 'stem-less' trees, and the distinction between the two life forms is reliant on different parameter sets (e.g. V_{cmax}, height, etc.) and slight modifications of the same process (e.g.

rate of assimilation, respiration, etc.). While our results and the tower data do not allow us to directly determine how C_4 grasses may be misrepresented in these TBMs, they clearly indicate that future development and evaluation should be focused on these issues. Eddy covariance studies of understorey savanna vegetation as conducted by Moore et al. (*this issue*) will be critical to this process.

4.3 Savanna phenology

The results from this study have shown that to simulate savanna fluxes, TBMs must be able to simulate the dynamics of savanna phenology, expressed by LAI. This was highlighted by the empirical benchmarks, where the results showed that while R_s, T_a and VPD were important drivers, LAI was required to capture the seasonality and magnitude of the fluxes to achieve good performance. LAI integrates the observed structural changes of the savanna as annual rainfall declines with reduced woody stem density; driving water and carbon exchange as a result (Kanniah et al., 2010; Ma et al., 2013; Sea et al., 2011). If LAI is prescribed in a model, it is important that leaf area is partitioned correctly between the trees and grass layers to describe their respective phenology. This partitioning is important, as the C_4 grass understorey explains most of the seasonal variation in LAI, and is a consequence of an annual phenology that exhibits rapid growth at the onset of the wet season and senescence at the onset of the dry (Williams et al., 1996b). By contrast the evergreen eucalypt canopy shows modest reductions in canopy leaf area during the dry season, especially as mean annual rainfall declines (Bowman and Prior, 2005; Kelley et al., 2007). The strong seasonal dynamics of the grasses result in large changes in LAI, with levels varying between 0.7 and 2.5 at high rainfall sites (Sea et al., 2011). The phenological strategy of the C_4 grasses also changes with rainfall interannual variability, with the onset of the greening period becoming progressively delayed as sites become drier, to become eventually rain-pulse driven as the monsoonal influence weakens (Ma et al., 2013).

With the exception of LPJGUESS, all models prescribed LAI as an input driver. Prescribing LAI can be problematic depending on the time-scale and how it is partitioned between trees and grass layers. At large time-steps (months) it will fail to capture the rapidly changing dynamics of vegetation during the transition periods, and this is particularly true for the onset of the wet season (Sep-Nov) especially at drier sites that are subject to larger interannual rainfall variability (Hutley et al., 2011). Additionally, as the sites become drier the tree:grass ratio will become smaller and this dynamic can be difficult to predict, although methods do exist (see Donohue et al. 2009). From the results, we infer that TBMs that prescribe LAI and allow for a dynamic representation of tree and grass ratios are better able to capture the changing dynamics of the savanna system. This is a possible explanation for the better performance of the BIOS2, MAESPA and SPA models in simulating GPP as these models dynamically partition leaf area between trees and grasses at the sub-monthly time-scale, rather than using a bulk value. However, there are limitations to using prescribed LAI, predominantly in that it describes a stable system, of which savannas are typically not; having a large sensitivity to changes in climate; particularly rainfall variability and disturbance (Sankaran et al., 2005). DGVMs that consider dynamic vegetation and use a prognostic LAI can simulate the feedback between the climate and the relative cover of trees and grasses, which shapes the savanna continuum. This feedback allows the simulated savanna structure to potentially shift to alternate states (e.g. grassland or forest) in response to changes in annual rainfall and fire severity (Scheiter and Higgins, 2007, 2009). While LPJGUESS was the only TBM to use a prognostic LAI in our study, it achieved only moderate performance, and this may be due to how carbon is allocated from the pool on an annual time step, such that it is not as dynamic as it could be. However, its capability to simulate the feedback between climate and LAI is critical for simulating how savanna dynamics may change from year to year. There may also be issues with how phenology is simulated, particularly as it is determined from empirical formulations, which are: i) not specifically developed for savanna environments and ii) calculated before the growing season begins. Such formulations are therefore not mechanistic, and do not respond to actual season dynamics (e.g. limiting soil water), but are empirically determined (Richardson et al., 2013).

5. Conclusions

This study set out to assess how well a set of functionally different, state-of-the-art TBMs perform at predicting the bulk exchanges of carbon and water over savanna land surfaces. Our model inter-comparison has identified key weaknesses in the assumptions of biosphere-atmosphere processes, which do not hold for savanna environments. Our benchmarking has identified low model performance by TBMs is likely a result of incorrect assumptions related to: i) deep soil water access, ii) a systematic under-estimation of the contribution of the grass understorey in the wet season, and iii) the use of static phenology to represent dynamic vegetation. Our results showed that these assumptions, as they currently exist in TBMs, are not wholly supported by 'observations' of savanna water and carbon exchange and need to be addressed if more reliable projections are to be made on how savannas respond to environmental change. Despite this, our benchmarking has shown that all TBMs could potentially operate well for savanna ecosystems, provided that the above issues are developed on. We suggest that further work investigates how particular processes in the models may be affecting overall predicted water and carbon fluxes, and may include testing variable rooting depths, alternate root-water uptake schemes and how these might affect leaf-level outputs (e.g. stomatal conductance, leaf water potential) among TBMs, and different phenology schemes. The issues highlighted here also have scope beyond savanna environments, and are relevant to other water-limited ecosystems. The results from this study provide a foundation for improving how savanna ecosystem dynamics are simulated.

Acknowledgements

This study was conducted as part of the 'Australian Savanna Landscapes: Past, Present and Future' project funded by the Australian Research Council (DP130101566). The support, collection and utilization of data were provided by the OzFlux network (www.ozflux.org.au) and Terrestrial Ecosystem Research Network (TERN) (www.tern.org.au), and funded by the ARC (DP0344744, DP0772981 and DP130101566). PALS was partly funded by the TERN ecosystem Modelling and Scaling infrAStructure (eMAST) facility under the National Collaborative Research Infrastructure Strategy (NCRIS) 2013-2014 budget initiative of the Australian Government Department of Industry. Rhys Whitley was supported through the ARC Discovery Grant (DP130101566). Jason Beringer is funded under an ARC FT (FT110100602). Vanessa Haverd's contribution was supported by the Australian Climate Change Science Program. We acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028).

References:

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., 5(3), 819–827, doi:10.5194/gmd-5-819-2012, 2012.

Abramowitz, G., Leuning, R., Clark, M. and Pitman, A.: Evaluating the Performance of Land Surface Models, J. Clim., 21(21), 5468–5481, doi:10.1175/2008JCLI2378.1, 2008.

Ball, J. T., Woodrow, I. E. and Berry, J. A.: A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions., in Progress in Photosynthesis Research, pp. 221–224, Martinus-Nijhoff Publishers, Dordrecht, the Netherlands., 1987.

Bashtannyk, D. M. and Hyndman, R. J.: Bandwidth selection for kernel conditional density estimation, Comput. Stat. Data Anal., 36(3), 279–298, doi:10.1016/S0167-9473(00)00046-3, 2001.

Beringer, J., Hutley, L. B., Tapper, N. J. and Cernusak, L. a.: Savanna fires and their impact on net ecosystem productivity in North Australia, Glob. Chang. Biol., 13(5), 990–1004, doi:10.1111/j.1365-2486.2007.01334.x, 2007.

Beringer, J., Hutley, L. B., Abramson, D., Arndt, S. K., Briggs, P., Bristow, M., Canadell, J. G., Cernusak, L. a, Eamus, D., Evans, B. J., Fest, B., Goergen, K., Grover, S. P., Hacker, J., Haverd, V., Kanniah, K., Livesley, S. J., Lynch, A., Maier, S., Moore, C., Raupach, M., Russell-Smith, J., Scheiter, S., Tapper, N. J. and Uotila, P.: Fire in Australian Savannas: from leaf to landscape., Glob. Chang. Biol., 11, 6641, doi:10.1111/gcb.12686, 2014.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello Jr, J. A., Stevens, L. and Vuichard, N.: The plumbing of land surface models: benchmarking model performance, J. Hydrometeorol., 16, 1425–1442, 2015.

Bond, W. J.: What Limits Trees in C4 Grasslands and Savannas?, Annu. Rev. Ecol. Evol. Syst., 39(1), 641–659, doi:10.1146/annurev.ecolsys.39.110707.173411, 2008.

Bowman, D. M. J. S. and Prior, L. D.: Why do evergreen trees dominate the Australian seasonal tropics?, Aust. J. Bot., 53(5), 379–399, doi:10.1071/BT05022, 2005.

von Caemmerer, S. and Furbank, R. T.: Modeling C4 Photosynthesis, in C4 Plant Biology, edited by R. F. Sage and R. K. Monson, pp. 173–211, Academic Press, Toronto., 1999.

Cernusak, L. A., Hutley, L. B., Beringer, J., Holtum, J. A. M. and Turner, B. L.: Photosynthetic physiology of eucalypts along a sub-continental rainfall gradient in northern Australia, Agric. For. Meteorol., 151(11), 1462–1470, doi:10.1016/j.agrformet.2011.01.006, 2011.

Chapin, F. S., Woodwell, G. M., Randerson, J. T., Rastetter, E. B., Lovett, G. M., Baldocchi, D. D., Clark, D. A., Harmon, M. E., Schimel, D. S., Valentini, R., Wirth, C., Aber, J. D., Cole, J. J., Goulden, M. L., Harden, J. W., Heimann, M., Howarth, R. W., Matson, P. A., McGuire, A. D., Melillo, J. M., Mooney, H. A., Neff, J. C., Houghton, R. A., Pace, M. L., Ryan, M. G., Running, S. W., Sala, O. E., Schlesinger, W. H. and Schulze, E. D.: Reconciling carbon-cycle concepts, terminology, and methods, Ecosystems, 9(7), 1041–1050, doi:10.1007/s10021-005-0105-7, 2006.

Chen, X., Hutley, L. B. and Eamus, D.: Carbon balance of a tropical savanna of northern Australia., Oecologia, 137(3), 405–16, doi:10.1007/s00442-003-1358-5, 2003.

Collatz, G. J., Ball, J. T., Grivet, C. and Berry, J. A.: Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer, Agric. For. Meteorol., 54, 107–136, 1991.

Collatz, G. J., Ribas-Carbo, M. and Berry, J. A.: Coupled photosynthesis-stomatal conductance model for leaves of C4 plants, Funct. Plant Biol., 19(5), 519 – 538, 1992.

Donohue, R. J., Mcvicar, T. R. and Roderick, M. L.: Climate-related trends in Australian vegetation cover as inferred from satellite observations, 1981-2006, Glob. Chang. Biol., 15, 1025–1039, doi:10.1111/j.1365-2486.2008.01746.x, 2009.

Duursma, R. A. and Medlyn, B. E.: MAESPA: A model to study interactions between water limitation, environmental drivers and vegetation function at tree and stand levels, with an example application to [CO2] ?? drought interactions, Geosci. Model Dev., 5(4), 919–940, doi:10.5194/gmd-5-919-2012, 2012.

Eamus, D., O'Grady, A. P. O. and Hutley, L. B.: Dry season conditions determine wet season water use in the wet-tropical savannas of northern Australia., Tree Physiol., 20(18), 1219–1226, 2000.

Eamus, D., Hutley, L. B. and O'Grady, A. P. O.: Daily and seasonal patterns of carbon and water fluxes above a north Australian savanna., Tree Physiol., 21(12-13), 977–88, 2001.

Eamus, D., Chen, X., Kelley, G. and Hutley, L. B.: Root biomass and root fractal analyses of an open Eucalyptus forest in a savanna of north Australia, Aust. J. Bot., 50, 31–41, 2002.

Eamus, D., Cleverly, J., Boulain, N., Grant, N., Faux, R. and Villalobos-Vega, R.: Carbon and water fluxes in an arid-zone Acacia savanna woodland: An analyses of seasonal patterns and responses to rainfall events, Agric. For. Meteorol., 182-183, 225–238, doi:10.1016/j.agrformet.2013.04.020, 2013.

Farquhar, G. D., von Caemmerer, S. and Berry, J. A.: A Biochemical Model of Photosynthetic CO2 Assimilation in Leaves of C3 species, Planta, 149, 78–90, 1980.

Grace, J., Jose, J. S., Meir, P., Miranda, H. S. and Montes, R. A.: Productivity and carbon fluxes of tropical savannas, J. Biogeogr., 33(3), 387–400, doi:10.1111/j.1365-2699.2005.01448.x, 2006.

Haverd, V., Raupach, M. R., Briggs, P. R., Canadell, J. G., Isaac, P., Pickett-Heaps, C., Roxburgh, S. H., Van Gorsel, E., Viscarra Rossel, R. a. and Wang, Z.: Multiple observation types reduce uncertainty in Australia's terrestrial carbon and water cycles, Biogeosciences, 10(3), 2011–2040, doi:10.5194/bg-10-2011-2013, 2013.

Haverd, V., Smith, B., Raupach, M. R., Briggs, P. R., Nieradzik, L., Beringer, J., Hutley, L. B., Trudinger, C. M. and Cleverly, J. R.: Coupling carbon allocation with leaf and root phenology predicts tree-grass partitioning along a savanna rainfall gradient. Vanessa, Biogeosciences, 2015.

Haxeltine, A. and Prentice, I. C.: A general model for the light-use efficiency of primary production, Funct. Ecol., 10(5), 551–561, doi:10.2307/2390165, 1996.

Higgins, S. I. and Scheiter, S.: Atmospheric CO2 forces abrupt vegetation shifts locally, but not globally., Nature, 488, 209–212, doi:10.1038/nature11238, 2012.

Hsu, K.: Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis, Water Resour. Res., 38(12), 1–17, doi:10.1029/2001WR000795, 2002.

Huntingford, C. and Monteith, J. L.: The behaviour of a mixed-layer model of the convective boundary layer coupled to a big leaf model of surface energy partitioning, Boundary-Layer Meteorol., 88(1), 87–101, doi:10.1023/A:1001110819090, 1998.

Hutchinson, M. F. and Xu, T.: ANUCLIM v6.1, Fenner School of Environment and Society, Australian National University, Canberra, ACT., 2010.

Hutley, L. B., O'Grady, A. P. O. and Eamus, D.: Evapotranspiration from Eucalypt open-forest savanna of Northern Australia, Funct. Ecol., 14, 183–194, 2000.

Hutley, L. B., Grady, A. P. O. and Eamus, D.: Monsoonal influences on evapotranspiration of savanna vegetation of northern Australia, Oecologia, 126(3), 434–443, doi:10.1007/s004420000539, 2001.

Hutley, L. B., Beringer, J., Isaac, P. R., Hacker, J. M. and Cernusak, L. A.: A sub-continental scale living laboratory: Spatial patterns of savanna vegetation over a rainfall gradient in northern Australia, Agric. For. Meteorol., 151(11), 1417–1428, doi:10.1016/j.agrformet.2011.03.002, 2011.

Isbell, R.: The Australian Soil Classification, Revised Ed., CSIRO Publishing, Collingwood, Victoria., 2002.

Kanniah, K. D., Beringer, J. and Hutley, L. B.: The comparative role of key environmental factors in determining savanna productivity and carbon fluxes: A review, with special reference to northern Australia, Prog. Phys. Geogr., 34(4), 459–490, doi:10.1177/0309133310364933, 2010.

De Kauwe, M. G., Zhou, S.-X., Medlyn, B. E., Pitman, A. J., Wang, Y. P., Duursma, R. A. and Prentice, I. C.: Do land surface models need to include differential plant species responses to drought? Examining model predictions across a latitudinal gradient in Europe, Biogeosciences Discuss., 12(15), 12349–12393, doi:10.5194/bgd-12-12349-2015, 2015.

Kelley, G., O'Grady, A. P., Hutley, L. B. and Eamus, D.: A comparison of tree water use in two contiguous vegetation communities of the seasonally dry tropics of northern Australia: The importance of site water budget to tree hydraulics, Aust. J. Bot., 55(7), 700–708, doi:10.1071/BT07021, 2007.

Koch, G. W., Vitousek, P. M., Steffen, W. L. and Walker, B. H.: Terrestrial transects for global change research, Vegetation, 121(1-2), 53–65, doi:10.1007/BF00044672, 1995.

Kowalczyk, E. A., Wang, Y. P. and Law, R. M.: The CSIRO Atmosphere Biosphere Land Exchange (CABLE) model for use in climate models and as an offline model, Aspendale, Victoria., 2006.

Lehmann, C. E. R., Prior, L. D. and Bowman, D. M. J. S.: Decadal dynamics of tree cover in an Australian tropical savanna, , 601–612, doi:10.1111/j.1442-9993.2009.01964.x, 2009.

Lehmann, C. E. R., Anderson, T. M., Sankaran, M., Higgins, S. I., Archibald, S., Hoffmann, W. A., Hanan, N. P., Williams, R. J., Fensham, R. J., Felfili, J., Hutley, L. B., Ratnam, J., San Jose, J., Montes, R., Franklin, D., Russell-Smith, J., Ryan, C. M., Durigan, G., Hiernaux, P., Haidar, R., Bowman, D. M. J. S. and Bond, W. J.: Savanna vegetation-fire-climate relationships differ among continents., Science (80-.)., 343, 548–552, doi:10.1126/science.1247355, 2014.

Leuning, R.: A critical appraisal of a combined stomatal-photosynthesis model for C3 plants, Plant, Cell Environ., 18(4), 339–355, 1995.

Li, L., Wang, Y. P., Yu, Q., Pak, B., Eamus, D., Yan, J., Van Gorsel, E. and Baker, I. T.: Improving the responses of the Australian community land surface model (CABLE) to seasonal drought, J. Geophys. Res. Biogeosciences, 117(4), 1–16, doi:10.1029/2012JG002038, 2012.

Ma, X., Huete, A., Yu, Q., Coupe, N. R., Davies, K., Broich, M., Ratana, P., Beringer, J., Hutley, L. B., Cleverly, J., Boulain, N. and Eamus, D.: Spatial patterns and temporal dynamics in savanna vegetation phenology across the North Australian Tropical Transect, Remote Sens. Environ., 139, 97–115, doi:10.1016/j.rse.2013.07.030, 2013.

McKenzie, N. N., Jacquier, D., Isbell, R. and Brown, K.: Australian soils and landscapes : an illustrated compendium, CSIRO Publishing, Collingwood, Victoria., 2004.

Medlyn, B. E., Robinson, A. P., Clement, R. and McMurtrie, R. E.: On the validation of models of forest CO2 exchange using eddy covariance data: some perils and pitfalls, Tree Physiol., 25, 839–857, 2005.

Medlyn, B. E., Duursma, R. A., Eamus, D., Ellsworth, D. S., Prentice, I. C., Barton, C. V. M., Crous, K. Y., De Angelis, P., Freeman, M. and Wingate, L.: Reconciling the optimal and empirical approaches to modelling stomatal conductance, Glob. Chang. Biol., 17(6), 2134–2144, doi:10.1111/j.1365-2486.2010.02375.x, 2011.

Moore, C. E., Beringer, J., Evans, B., Hutley, L. B., McHugh, I. and Tapper, N. J.: The contribution of trees and grasses to productivity of an Australian tropical savanna, Biogeosciences, doi:10.5194/bgd-12-19307-2015, this issue.

O'Grady, A. P. O., Eamus, D. and Hutley, L. B.: Transpiration increases during the dry season: patterns of tree water use in eucalypt open-forests of northern Australia., Tree Physiol., 19(9), 591–597, 1999.

Parton, W. J., Anderson, D. W., Cole, C. V and Stewart, J. W. B.: Simulation of soil organic matter formation and mineralization in semiarid agroecosystems, in Nutrient Cycling In Agricultural Ecosystems, vol. 23, pp. 533–550., 1983.

Pitman, A. J.: The evolution of, and revolution in, land surface schemes designed for climate models, Int. J. Climatol., 23, 479–510, doi:10.1002/joc.893, 2003.

Prior, L. D., Eamus, D. and Duff, G. A.: Seasonal Trends in Carbon Assimilation, Stomatal Conductance, Pre-dawn Leaf Water Potential and Growth in Terminalia ferdinandiana, a Deciduous Tree of Northern Australian Savannas, Aust. J. Bot., 45, 53–69, 1997.

Richardson, A. D., Keenan, T. F., Migliavacca, M., Ryu, Y., Sonnentag, O. and Toomey, M.: Climate change, phenology, and phenological control of vegetation feedbacks to the climate system, Agric. For. Meteorol., 169, 156–173, doi:10.1016/j.agrformet.2012.09.012, 2013.

Russell-Smith, J. and Edwards, A. C.: Seasonality and fire severity in savanna landscapes of monsoonal northern Australia, Int. J. Wildl. Fire, 15(4), 541–550, doi:10.1071/WF05111, 2006.

Ryu, Y., Baldocchi, D. D., Kobayashi, H., van Ingen, C., Li, J., Black, T. A., Beringer, J., van Gorsel, E., Knohl, A., Law, B. E. and Roupsard, O.: Integration of MODIS land and atmosphere products with a coupled-process model to estimate gross primary productivity and evapotranspiration from 1 km to global scales, Global Biogeochem. Cycles, 25(GB4017), doi:10.1029/2011GB004053, 2011.

Ryu, Y., Baldocchi, D. D., Black, T. A., Detto, M., Law, B. E., Leuning, R., Miyata, A., Reichstein, M., Vargas, R., Ammann, C., Beringer, J., Flanagan, L. B., Gu, L., Hutley, L. B., Kim, J., McCaughey, H., Moors, E. J., Rambal, S. and Vesala, T.: On the temporal upscaling of evapotranspiration from instantaneous remote sensing measurements to 8-day mean daily-sums, Agric. For. Meteorol., 152(1), 212–222, doi:10.1016/j.agrformet.2011.09.010, 2012.

Sankaran, M., Hanan, N. P., Scholes, R. J., Ratnam, J., Augustine, D. J., Cade, B. S., Gignoux, J., Higgins, S. I., Le Roux, X., Ludwig, F., Ardo, J., Banyikwa, F., Bronn, A., Bucini, G., Caylor, K. K., Coughenour, M. B., Diouf, A., Ekaya, W., Feral, C. J., February, E. C., Frost, P. G. H., Hiernaux, P., Hrabar, H., Metzger, K. L., Prins, H. H. T., Ringrose, S., Sea, W., Tews, J., Worden, J. and Zambatis, N.: Determinants of woody cover in African savannas., Nature, 438, 846–849, doi:10.1038/nature04070, 2005.

Scheiter, S. and Higgins, S. I.: Partitioning of root and shoot competition and the stability of savannas., Am. Nat., 170(4), 587–601, doi:10.1086/521317, 2007.

Scheiter, S. and Higgins, S. I.: Impacts of climate change on the vegetation of Africa: an adaptive dynamic vegetation modelling approach, Glob. Chang. Biol., 15(9), 2224–2246, doi:10.1111/j.1365-2486.2008.01838.x, 2009.

Scheiter, S., Higgins, S. I., Beringer, J. and Hutley, L. B.: Climate change and long-term fire management impacts on Australian savanna, Glob. Chang. Biol., Submitted , 1211–1226, doi:10.1111/nph.13130, 2014.

Scholes, R. J. and Archer, S. R.: Tree-grass interactions in savannas, Annu. Rev. Ecol. Syst., 28, 517–544, doi:10.1146/annurev.ecolsys.28.1.517, 1997.

Schymanski, S. J., Roderick, M. L., Sivapalan, M., Hutley, L. B. and Beringer, J.: A test of the optimality approach to modelling canopy properties and CO2 uptake by natural vegetation., Plant. Cell Environ., 30(12), 1586–98, doi:10.1111/j.1365-3040.2007.01728.x, 2007.

Schymanski, S. J., Sivapalan, M., Roderick, M. L., Beringer, J. and Hutley, L. B.: An optimality-based model of the coupled soil moisture and root dynamics, Hydrol. Earth Syst. Sci., 12(3), 913–932, doi:10.5194/hess-12-913-2008, 2008.

Schymanski, S. J., Sivapalan, M., Roderick, M. L., Hutley, L. B. and Beringer, J.: An optimality-based model of the dynamic feedbacks between natural vegetation and the water balance, Water Resour. Res., 45(1), doi:10.1029/2008WR006841, 2009.

Sea, W. B., Choler, P., Beringer, J., Weinmann, R. a., Hutley, L. B. and Leuning, R.: Documenting improvement in leaf area index estimates from MODIS using hemispherical photos for Australian savannas, Agric. For. Meteorol., 151(11), 1453–1461, doi:10.1016/j.agrformet.2010.12.006, 2011.

Simioni, G., Roux, X. Le, Gignoux, J. and Sinoquet, H.: Treegrass: a 3D, process-based model for simulating plant interactions in tree–grass ecosystems, Ecol. Modell., 131, 47–63, 2000.

Simioni, G., Gignoux, J. and Le Roux, X.: Tree layer spatial structure can affect savanna production and water budget: Results of a 3-D model, Ecology, 84(7), 1879–1894, doi:10.1890/0012-9658(2003)084[1879:TLSSCA]2.0.CO;2, 2003.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J., Levis, S., Lucht, W., Sykes, M., Thonicke, K. and Venevski, S.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic vegetation model, Glob. Chang. Biol., 9, 161–185, 2003.

Smith, B., Prentice, I. C. and Sykes, M. T.: Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space, Glob. Ecol. Biogeogr., 10(6), 621–637, doi:10.1046/j.1466-822X.2001.t01-1-00256.x, 2001.

Still, C. J., Berry, J. a., Collatz, G. J. and DeFries, R. S.: Global distribution of C 3 and C 4 vegetation: Carbon cycle implications, Global Biogeochem. Cycles, 17(1), 6–1–6–14, doi:10.1029/2001GB001807, 2003.

Wang, Y. P., Kowalczyk, E., Leuning, R., Abramowitz, G., Raupach, M. R., Pak, B., Van Gorsel, E. and Luhar, A.: Diagnosing errors in a land surface model (CABLE) in the time and frequency domains, J. Geophys. Res. Biogeosciences, 116, 1–18, doi:10.1029/2010JG001385, 2011.

van der Werf, G. R., Randerson, J. T., Giglio, L., Gobron, N. and Dolman, a. J.: Climate controls on the variability of fires in the tropics and subtropics, Global Biogeochem. Cycles, 22(3), 1–13, doi:10.1029/2007GB003122, 2008.

Whitley, R. J., Macinnis-Ng, C. M. O., Hutley, L. B., Beringer, J., Zeppel, M., Williams, M., Taylor, D. and Eamus, D.: Is productivity of mesic savannas light limited or water limited? Results of a simulation study, Glob. Chang. Biol., 17(10), 3130–3149, doi:10.1111/j.1365-2486.2011.02425.x, 2011.

Williams, M., Rastetter, E. B., Fernandes, D. N., Goulden, M. L., Wofsy, S. C., Shaver, G. R., Melillo, J. M., Munger, J. W., Fan, S.-M. and Nadelhoffer, K. J.: Modelling the soil-plant-atmosphere continuum in a Quercus Acer stand at Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties, Plant, Cell Environ., 19, 911–927, 1996a.

Williams, R. J., Duff, G. A., Bowman, D. M. J. S. and Cook, G. D.: Variation in the composition and structure of tropical savannas as a function of rainfall and soil texture along a large-scale climatic gradient in the Northern Territory, Australia, J. Biogeogr., 23(6), 747–756, doi:10.1111/j.1365-2699.1996.tb00036.x, 1996b.

		Howard Springs[a]	Adelaide River[b]	Daly Uncleared[c]	Dry River[d]	Sturt Plains[e]
	Years (inclusive)	2001 – 2012	2007 – 2009	2008 – 2012	2008 – 2012	2008 – 2012
	Co-ordinates	12°29'39.12'' S 131°09'09'' E	13°04'36.84'' S 131°07'04.08'' E	14°09'33.12'' S 131°23'17.16'' E	15°15'31.62'' S 132°22'14.04'' E	17°09'02.76'' S 133°21'01.14'' E
	Elevation (m)	64	90	110	175	250
	[f]Meteorology					
	Annual Rainfall (mm)	1714	1460	1170	850	535
	Min/Max Daily Temperature (°C)	22.0/33.0	21.8/35.3	20.8/35.0	20.0/34.8	19.0/34.2
	Min/Max Absolute Humidity (g m^{-3})	11.0/18.5	8.9/17.7	8.6/15.1	7.8/12.3	6.1/9.0
	Min/Max Soil Moisture (m^{3} m^{-3})	0.06/0.1	0.09/0.14	0.03/0.06	0.03/0.05	0.04/0.1
	Soil Temperature (°C)	32.7	35.7	32.8	*n.a.*	30.2
	Solar Radiation (W m^{-2})	256.5	258.1	270.6	266.5	269.7
	Bowen Ratio	1.7	3.1	3.2	4.6	15.8
	[f]Vegetation					
	Overstorey species	*Eu. Miniata* *Eu. tetrodonta* *Er. chlorostachys*	*Eu. tectifica* *Pl. careya* *Co. latifolia*	*Te. grandiflora* *Eu. tetrodonta* *Co. latifolia*	*Eu. tetrodonta* *Co. terminalis* *Eu. dichromophloia*	*n.a.*
	Understorey species	*Sorghum* spp. *He. triticeus*	*Sorghum* spp. *Ch. fallax*	*Sorghum* spp. *He. triticeus*	*Sorghum intrans* *Th. Tiandra* *Ch. fallax*	*Astrabla* spp.
	Basal Area (m^{2} ha^{-1})	9.7	5.1	8.3	5.4	*n.a.*
	Canopy Height (m)	18.9	12.5	16.4	12.3	0.2
	LAI (m^{2} m^{-2})	1.04 ± 0.07	0.68 ± 0.07	0.80 ± 0.12	0.58 ± 0.11	0.39 ± 0.11
	Total Leaf Nitrogen (g m^{-3})	1.42 ± 0.20	1.27 ± 0.18	1.35 ± 0.19	1.97 ± 0.15	2.37 ± 0.17
	[g]Soil					
	Type	Red kandosol	Yellow hydrosol	Red kandosol	Red kandosol	Grey vertosol
A Horizon	Texture	Sandy loam	Sandy loam	Loam	Clay	loam
	Clay PSD (%)	15	20	20	50	20
	Sand PSD (%)	60	50	40	25	40
	Thickness (m)	0.30	0.30	0.20	0.15	0.20
	Bulk Density (Mg m^{-3})	1.29	1.60	1.39	1.20	1.39
	Hydraulic Conductivity (mm hr^{-1})	9	7	9	3	9
	Field Capacity (mm m^{-1})	156	132	147	140	147
B Horizon	Texture	Clay loam	Clay	Clay loam	Clay	Clay loam
	Clay PSD (%)	40	55	35	55	35
	Sand PSD (%)	30	20	30	20	30
	Thickness (m)	1.20	0.60	0.69	1.29	0.69
	Bulk Density (Mg m^{-3})	1.39	1.70	1.39	1.39	1.39
	Hydraulic Conductivity (mm hr^{-1})	8	5	7	2	7
	Field Capacity (mm m^{-1})	146	31	146	107	146

Table 1: Summarised dataset information for each of the five savanna sites used in this study. This includes site descriptions pertaining to local meteorology, vegetation and below ground soil characteristics. Where data were not available, the abbreviation n.a. is used. Definitions for the species genus mentioned in the table are as follows: *Eucalytpus* (*Eu.*), *Erythrophleum* (*Er.*), *Terminalia* (*Te.*), *Corymbia* (*Co.*), *Planchonia* (*Pl.*), *Buchanania* (*Bu.*), *Themda* (*Th.*), *Hetropogan* (*He.*), and *Chrysopogon* (*Ch.*). Eddy covariance datasets relating to each of the 5 sites here can be download from www.ozflux.org.au and hdl references are given by order of column (Jason Beringer (2013) – [a]hdl: 102.100.100/14228, [b]hdl: 102.100.100/14239, [c]hdl: 102.100.100/14229, [d]hdl: 102.100.100/14234, [e]hdl: 102.100.100/14230). Site meteorology is given as 30 year averages with values taken from [f]Hutley, et al. (2011). Soil descriptions are taken from the Digital Atlas of Australian Soils (www.asris.csiro.au) [g]Isbell, (2002).

Model Name	SPA	MAESPA	CABLE	BIOS2	BESS	LPJGUESS
Model definition	Soil-Plant-Atmosphere Model	MAESTRA-SPA	Community Atmosphere Biosphere Land-surface Exchange Model	Modified CABLE (CABLE + SLI + CASA-CNP)	Breathing Earth System Simulator	Lund-Potsdam-Jena General Ecosystem Simulator
Version	1.0	1.0	2.0	2.0	1.0	2.1
Reference	Williams et al. (1996a)	Duursma & Medlyn (2012)	Kowalyzck et al. (2006), Wang et al. (2011)	Haverd et al. (2013)	Ryu et al. (2011, 2012)	Smith et al. (2001)
Temporal resolution	30-min	30-min	30-min	Daily (30-min time-steps are generated from daily time-series)	Snap shot with MODIS overpass, then up-scaled to a daily and 8-day time series	Daily
Spatial resolution	Point	Point	0.05° (5 km)	0.05° (5 km)	0.05° (5 km)	Patch (c. 0.1 ha)
Functional class	Stand model	Individual Plant or Stand Model	Land-Surface Model	Land-Surface Model	Remote Sensing Model	Dynamic Global Vegetation Model
Canopy Description						
C_3 Assimilation	Farquhar et al. (1980)	Farquhar et al. (1980)	Farquhar et al. (1980)	Farquhar et al. (1980)	Farquhar et al. (1980)	Collatz et al. (1991)
C_4 Assimilation	Collatz et al. (1992)	Collatz et al. (1992)	Collatz et al. (1992)	Collatz et al. (1992)	Collatz et al. (1992)	Collatz et al. (1992)
Stomatal conductance	Williams et al. (1996a)	Medlyn et al. (2011)	Leuning (1995)	Leuning (1995)	Ball et al. (1987)	Haxeltine & Prentice (1996)
Transpiration	Penman-Monteith calculated at leaf-scale accounting for g_b and limitation of soil-water supply via Ψ_l	Penman-Monteith calculated at the leaf scale	Penman-Monteith	Penman-Monteith	Penman-Monteith	Haxeltine & Prentice (1996)
Boundary layer resistance	*f(wind speed, leaf width, air temperature)*	*f(wind speed, leaf width, air temperature and atmospheric pressure)*	*f(wind speed, leaf width, air temperature*	*f(wind speed, leaf width, air temperature*	*Not Modelled*	Huntingford & Monteith (1998)
Aerodynamic resistance	*f(wind speed, canopy height)*	Not calculated unless transpiration is calculated at the canopy scale, in which case g_b above isn't calculated.	*f(wind speed, canopy height)*	*f(wind speed, canopy height)*	*f(wind speed, canopy height)*	Huntingford & Monteith (1998)
Leaf area index	Prescribed (MODIS)	Prescribed (MODIS)	Prescribed (MODIS)	Prescribed (MODIS)	Prescribed (MODIS)	Prognostic (C allocation)
Canopy structure	Canopy + understorey divided into 10 layers	Individual plant crowns, spatially explicit locations and uniform understorey	2 (tree/grass) big leaf (sunlit/shaded)	2 (tree/grass) big leaf (sunlit/shaded)	2 (tree/grass) big leaf (sunlit/shaded)	5-year age/size cohorts for trees, single-layer grass understorey
C_3:C_4 fraction	Dynamic ratio variable with time. Compete for water and light.	Dynamic ratio variable with time. Compete for water and light.	Simulated as independent layers	Dynamic ratio variable with time. Compete for water not light.	Still et al. (2003) Ratio changes 70:30 to 10:90 down transect	Prognostic, determined as the outcome of the competition with trees
Canopy interception	YES	YES	YES	YES	NO	YES
Simulates growth	NO	NO	NO	NO	NO	YES
Soil Profile Description						
Soil profile structure	Profile divided into N layers (prescribed - 20 in this case.)	Profile divided into N layers (prescribed - 20 in this case.)	Profile divided into 6 layers	Profile divided into 12 layers (adjustable)	*Not Modelled*	2 layers (0-0.5, 0.5-2 m) with 10 cm evaporation sub-layer
Soil hydraulic properties	Function of sand and clay particle size distributions	Function of sand and clay particle size distributions	Prescribed	Australian Soils Resource Information System (ASRIS)	*Not Modelled*	Sitch et al. (2003)
Soil depth	6.5 m	5.0 m	4.5 m	10.0 m	*Not Modelled*	2 m
Root depth	6.5 m	5.0 m	4.5 m	0.5 m (grasses), 5.0 m (trees)	*Not Modelled*	2 m
Root distribution	Prescribed; exponential decay as a function of surface biomass and the total root biomass of the column	Prescribed; exponential decay as a function of surface biomass and the total root biomass of the column	Prescribed; exponential decay	Prescribed; exponential decay	*Not Modelled*	PFT-specific, trees have deeper roots on average
Soil-water stress modifier	E_t via g_s is increased to meet atmospheric demand while Ψ_l remains above a critical threshold	Maximum transpiration rate calculated from hydraulic conductance (soil-to-leaf) sets limit on actual transpiration, OR uses the Tuzet et al. (2003) model of stomatal conductance	Supply/Demand	g_s scaled by a soil moisture limitation function related to extractible water accessible by roots	Assumes LAI and seasonal variation of V_{cmax} reflect soil water stress	Supply/Demand
Hydraulic pathway resistance	$R_{soil} + R_{plant}$	$R_{soil} + R_{plant}$	*Not Modelled*	*Not Modelled*	*Not Modelled*	Not explicit, min(supply, demand) determines sapflow

Table 2: Summary table of the ecosystem models used in the experiment; highlighting differences and similarities in model structure and shared processes. Information is broken down into how each model describes aboveground canopy and belowground soil processes.

Statistical Metric	Definition
Correlation coefficient (*r*)	$\frac{n\sum_{i=1}^{n}(O_iM_i)-\sum_{i=1}^{n}O_i\sum_{i=1}^{n}M_i}{\sqrt{\left(n\sum_{i=1}^{n}O_i^2-\left(\sum_{i=1}^{n}O_i\right)^2\right)\left(n\sum_{i=1}^{n}M_i^2-\left(\sum_{i=1}^{n}M_i\right)^2\right)}}$
Standard Deviation (sd)	$\left\lvert 1-\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(M_i-\bar{M}\right)^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(O_i-\bar{O}\right)^2}}\right\rvert$
Normalised mean error (NME)	$\frac{\sum_{i=1}^{n}\lvert M_i-O_i\rvert}{\sum_{i=1}^{n}\lvert\bar{O}-O_i\rvert}$
Normalised mean bias (MBE)	$\frac{1}{n}\sum_{i=1}^{n}\left(M_i-O_i\right)$

Table A1: Definition of common metrics used to determine ranks against the empirical benchmarks. The terms *M* and *O* stand for model and observations respectively, while *n* denotes the length of the data, and *i* is the datum.

Figure Captions

Figure 1: The Northern Territory of Australia and the North Australian Tropical Transect (NATT) showing (a) the flux site locations with an accompanying 30-year (1970 to 2000) expression of the average meteorological conditions for (b) mean annual temperature, and (c) total annual precipitation derived from ANUCLIM v6.1 climate surfaces (Hutchinson and Xu, 2010).

Figure 2: Time-series of daily mean latent heat (LE) flux and gross primary productivity (GPP) depicting an average year for each of the 5 savanna sites using a smoothed, 7-day moving average. The sites are ordered from wettest to driest; (a) Howard Springs, (b) Adelaide River, (c) Daly River, (d) Dry River and (e) Sturt Plains. The joined, black dots are the tower flux time-series, while the grey lines are the performance benchmarks (emp1, emp2, emp3). Predictions of LE and GPP for each of the six terrestrial biosphere models are given by a spectrum of colours described in the legend.

Figure 3: Probability densities (expressed in scientific notation) of daily mean latent heat (LE) flux and gross primary productivity (GPP) at each of the 5 savanna sites, where the distributions for each flux are partitioned into wet and dry seasons. The order of the sites are from wettest to driest; (a) Howard Springs, (b) Adelaide River, (c) Daly River, (d) Dry River and (e) Sturt Plains. The grey region is the tower flux, while the dotted lines are the empirical benchmarks. Predicted LE and GPP probability densities from each of the six process-based models are given by a spectrum of colours described in the legend.

Figure 4: Standardised model residuals for latent energy (LE) and gross primary productivity (GPP) expressed in units of standard deviations (sd) [(modelled flux — observed flux)/sd(observed flux)]. Residuals are presented for each model: (a) CABLE, (b) BIOS2, (c) LPJGUESS, (d) MAESPA, (d) BESS and (e) SPA, where each flux site is represented by a blue-green-yellow gradient. For both fluxes, the residuals are plotted against time (ensemble average year) and against the flux prediction (bias).

Figure 5: Average rank plot showing the performance of the terrestrial biosphere models for all sites across the North Australian Tropical Transect (NATT) ordered in terms of annual rainfall as follows: Howard Springs (HowSpr), Adelaide River (AdrRiv), Daly Uncleared (DalUnc), Dry River (DryRiv), and Sturt Plains (StuPla). Models are individually ranked against the benchmarks in order of 1 to 4 (1 model + 3 benchmarks) and express the amount of metrics the models are meeting listed in Table S1. The rankings are determined individually for latent energy (LE) and gross primary productivity (GPP). The coloured lines represent each of the 6 models in the study, while the grey

lines represent the empirical benchmarks. The average ranking for each model was determined for (a) a complete year, (b) the wet season and (c) the dry season.

Figure 6: Average year outputs of vegetation transpiration (grass + trees) and soil evaporation, as well as their percentage contributions to total latent energy (LE) for each of the 6 terrestrial biosphere models at each of the 5 savanna sites.

Figure S1: A smoothed (7-day moving average) representation of the environmental drivers used to construct the empirical benchmarks at each of the 5 savanna sites, and are shown from wettest to driest; (a) Howard Springs, (b) Adelaide River, (c) Daly River, (d) Dry River and (e) Sturt Plains. The time-series represents the seasonality over an average year for mean daily solar radiation (R_s), mean daily air temperature (T_a), mean daily vapour pressure deficit (VPD) and leaf area index (LAI).

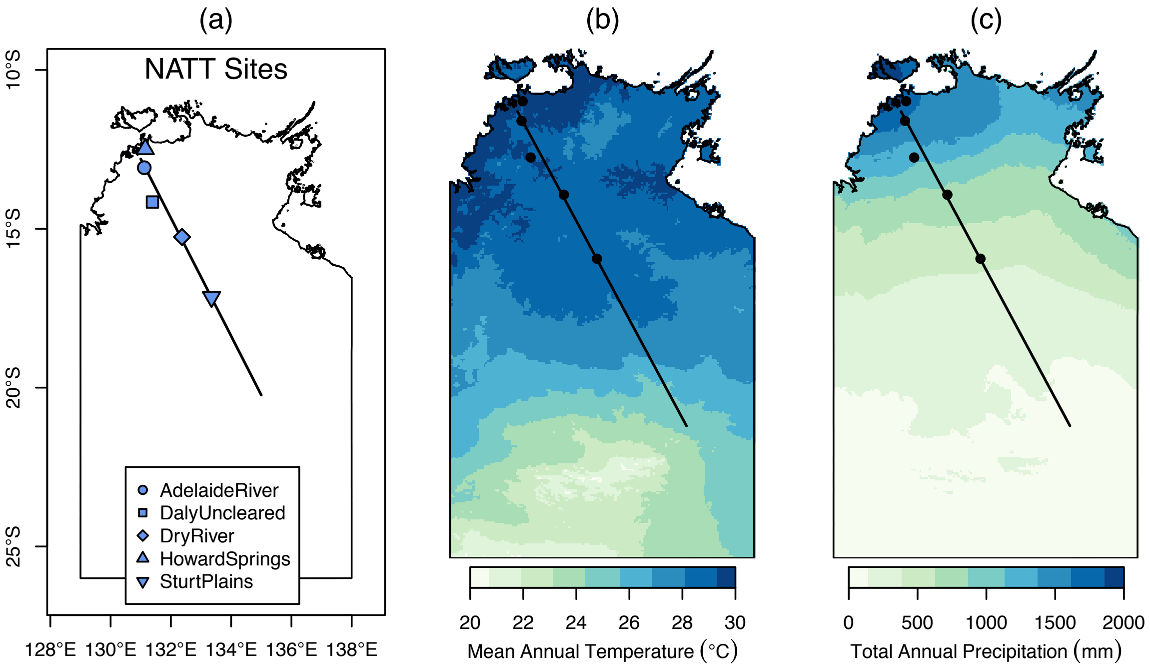


Figure 1: The Northern Territory of Australia and the North Australian Tropical Transect (NATT) showing (a) the flux site locations with an accompanying 30-year (1970 to 2000) expression of the average meteorological conditions for (b) mean annual temperature, and (c) total annual precipitation derived from ANUCLIM v6.1 climate surfaces (Hutchinson and Xu, 2010).

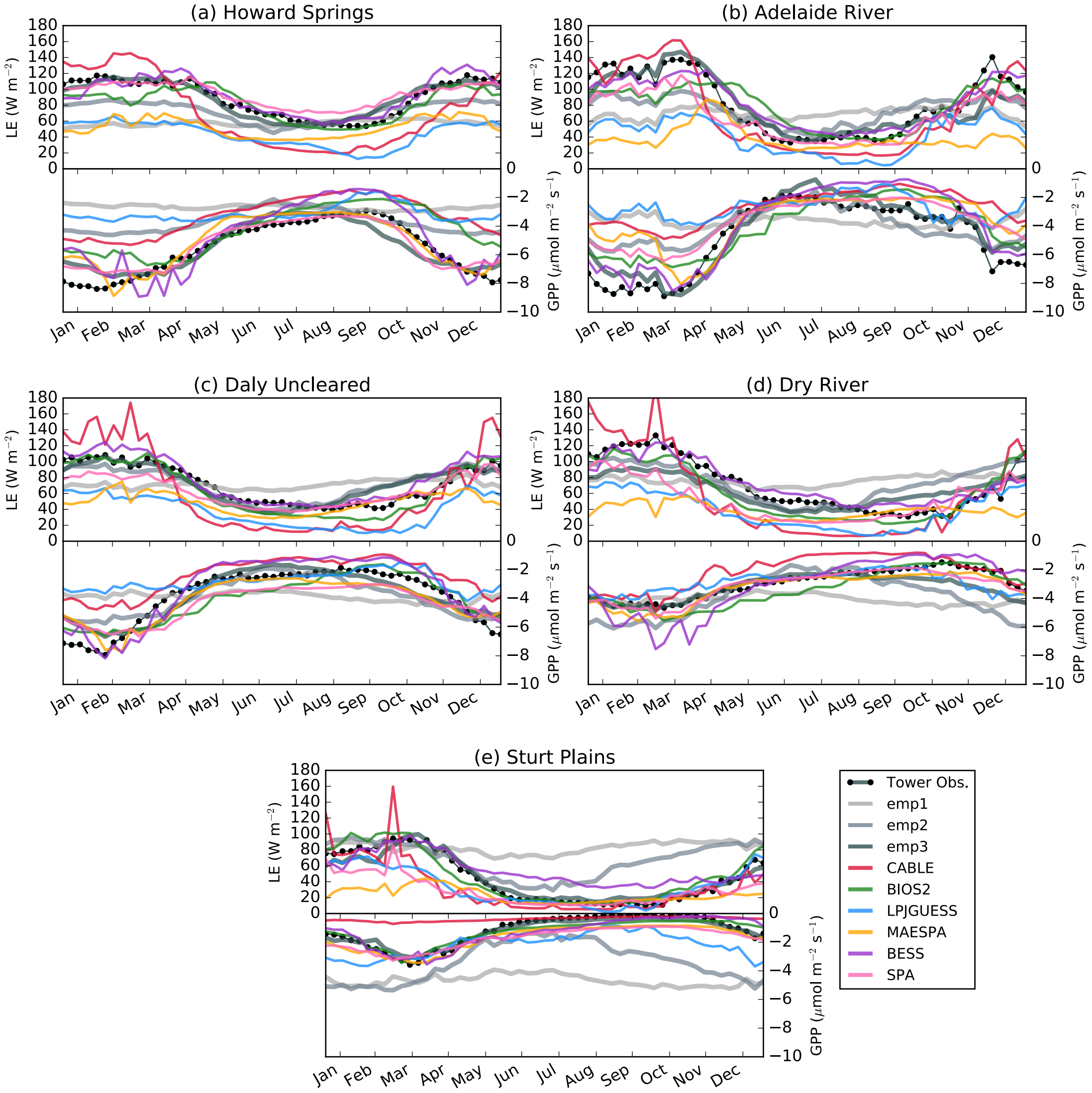


Figure 2: Time-series of daily mean latent heat (LE) flux and gross primary productivity (GPP) depicting an average year for each of the 5 savanna sites using a smoothed, 7-day moving average. The sites are ordered from wettest to driest; (a) Howard Springs, (b) Adelaide River, (c) Daly River, (d) Dry River and (e) Sturt Plains. The joined, black dots are the tower flux time-series, while the grey lines are the performance benchmarks (emp1, emp2, emp3). Predictions of LE and GPP for each of the six terrestrial biosphere models are given by a spectrum of colours described in the legend.

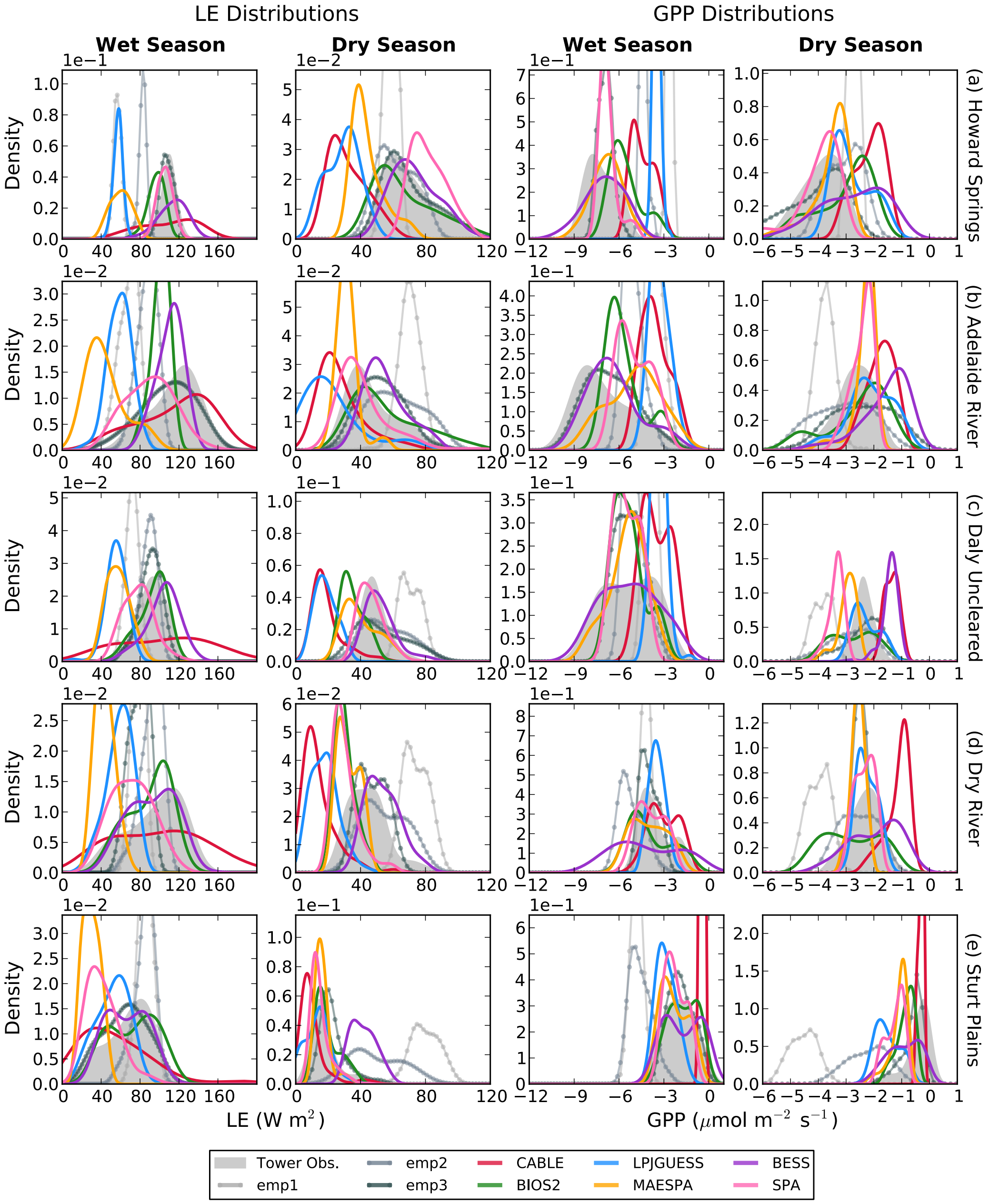


Figure 3: Probability densities (expressed in scientific notation) of daily mean latent heat (LE) flux and gross primary productivity (GPP) at each of the 5 savanna sites, where the distributions for each flux are partitioned into wet and dry seasons. The order of the sites are from wettest to driest; (a) Howard Springs, (b) Adelaide River, (c) Daly River, (d) Dry River and (e) Sturt Plains. The grey region is the tower flux, while the dotted lines are the empirical benchmarks. Predicted LE and GPP probability densities from each of the six process-based models are given by a spectrum of colours described in the legend.

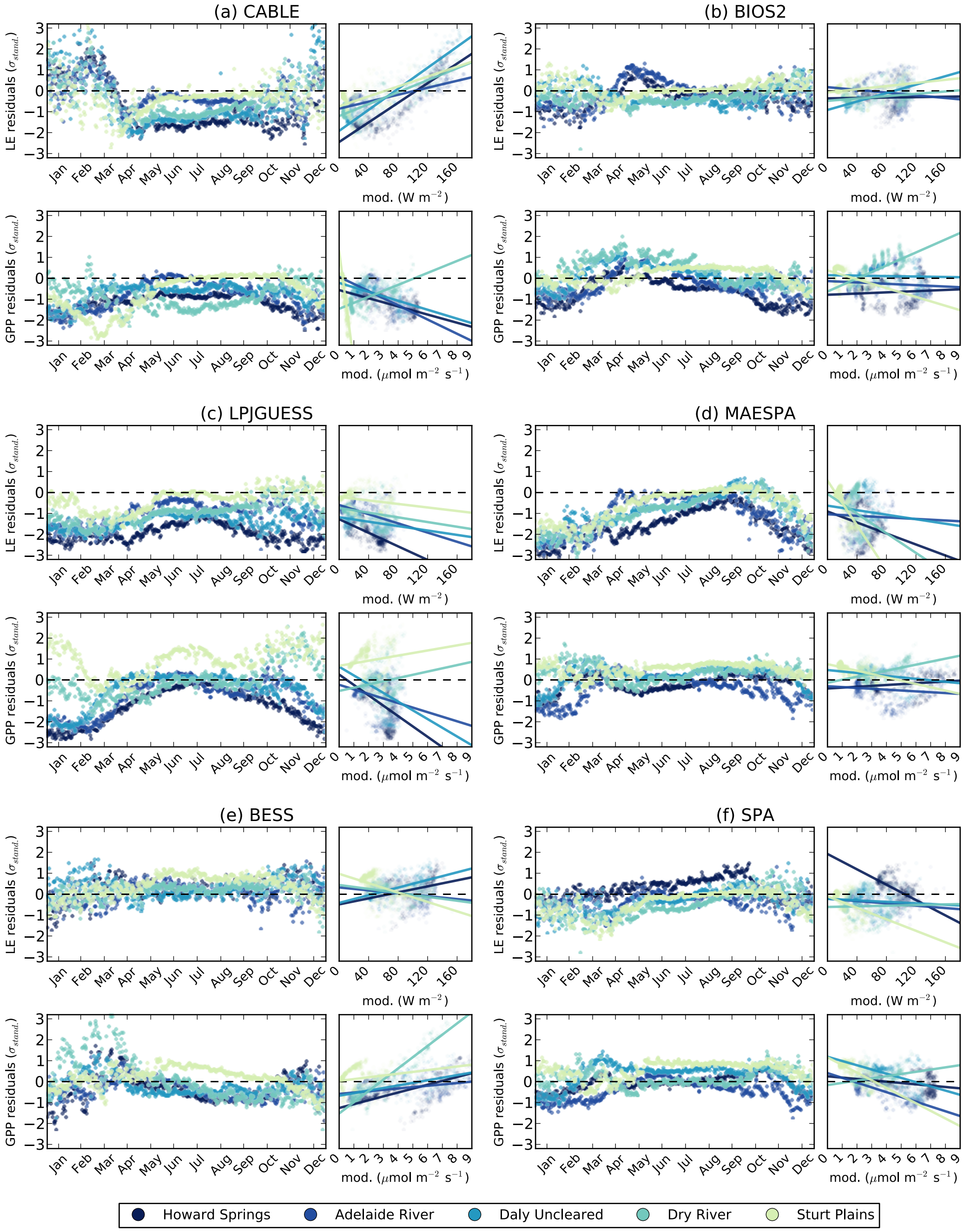
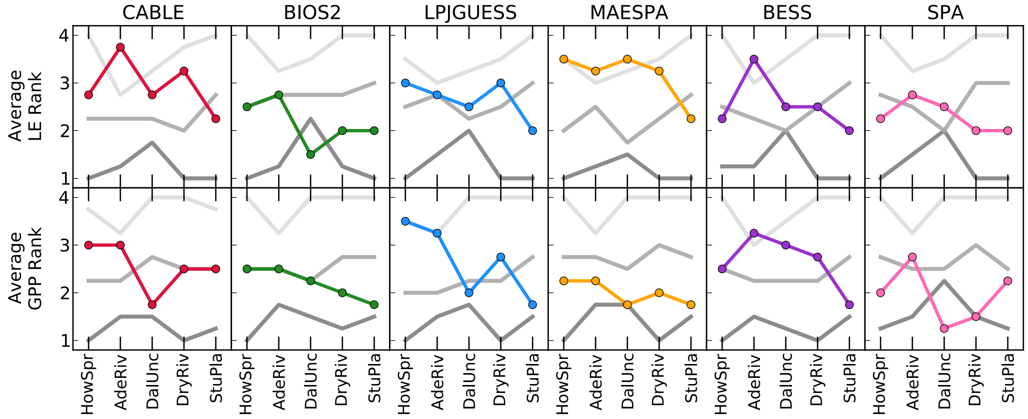
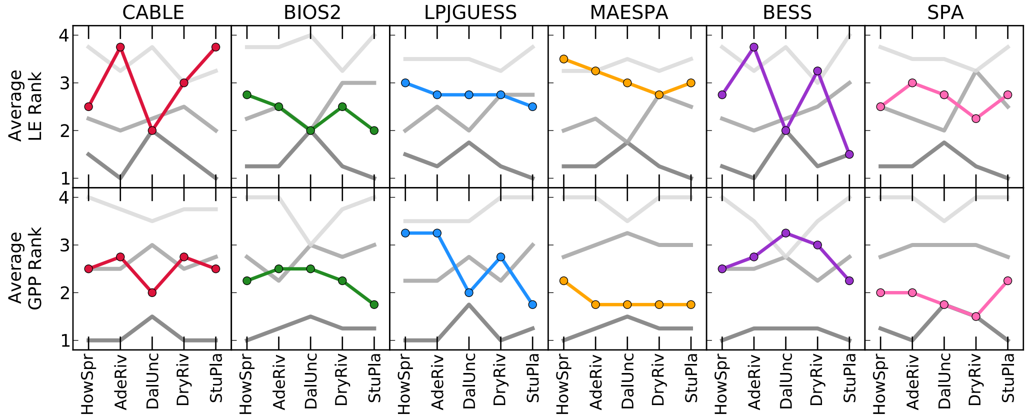


Figure 4: Standardised model residuals for latent energy (LE) and gross primary productivity (GPP) expressed in units of standard deviations (sd) [(modelled flux — observed flux)/sd(observed flux)]. Residuals are presented for each model: (a) CABLE, (b) BIOS2, (c) LPJGUESS, (d) MAESPA, (d) BESS and (e) SPA, where each flux site is represented by a blue-green-yellow gradient. For both fluxes, the residuals are plotted against time (an average year) and against the flux prediction (bias).

(a) Annual Performance



(b) Wet Season Performance



(c) Dry Season Performance

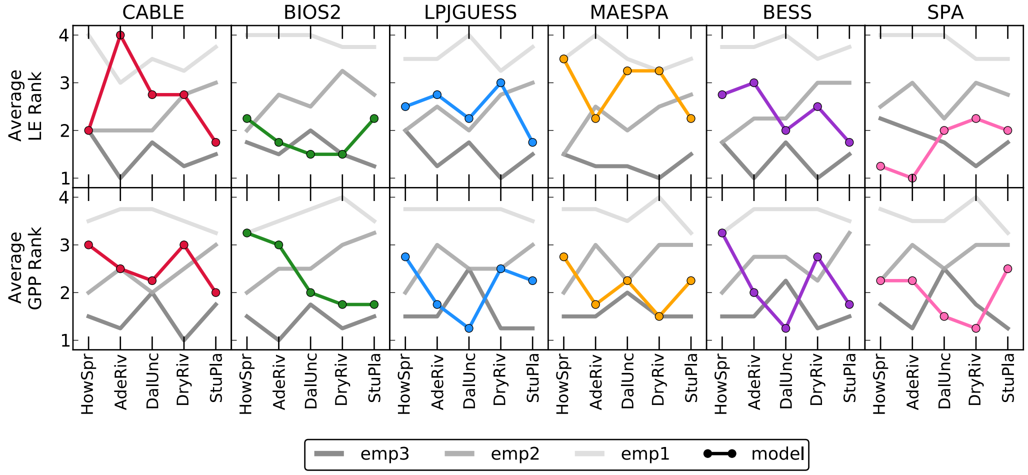


Figure 5: Average rank plot showing the performance of the ecosystem models for all sites across the North Australian Tropical Transect (NATT) ordered in terms of annual rainfall as follows: Howard Springs (HowSpr), Adelaide River (AdrRiv), Daly Uncleared (DalUnc), Dry River (DryRiv), and Sturt Plains (StuPla). Models are individually ranked against the benchmarks in order of 1 to 4 (1 model + 3 benchmarks) and express the amount of metrics the models are meeting listed in Table B2. The rankings are determined individually for latent energy (LE) and gross primary productivity (GPP). The coloured lines represent each of the 6 models in the study, while the grey lines represent the empirical benchmarks. The average ranking for each model was determined for (a-b) a complete year, (c-d) the wet season and (e-f) the dry season.

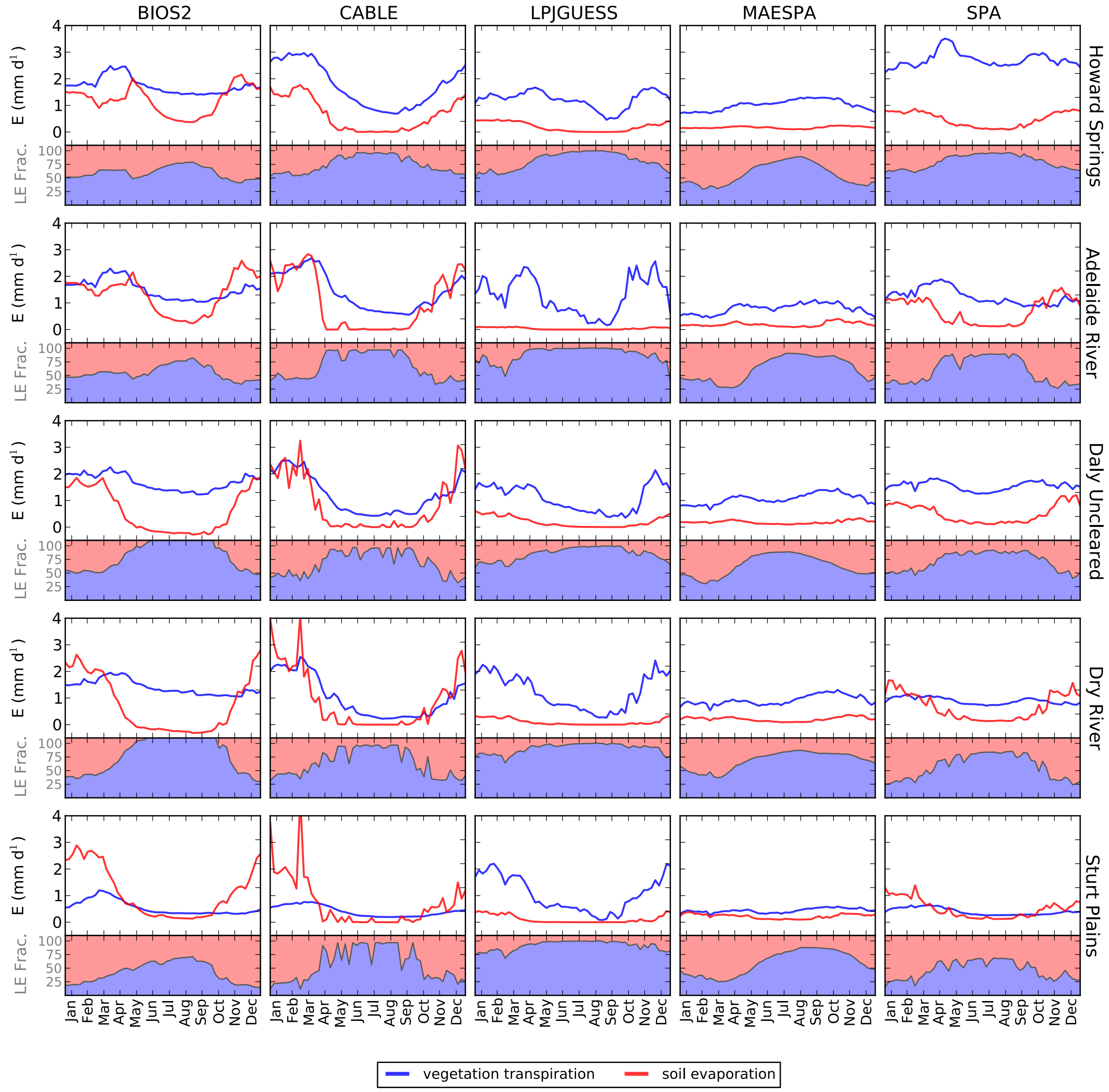


Figure 6: Average year outputs of vegetation transpiration (grass + trees) and soil evaporation, as well as their percentage contributions to total latent energy (LE) for each of the 6 terrestrial biosphere models at each of the 5 savanna sites.

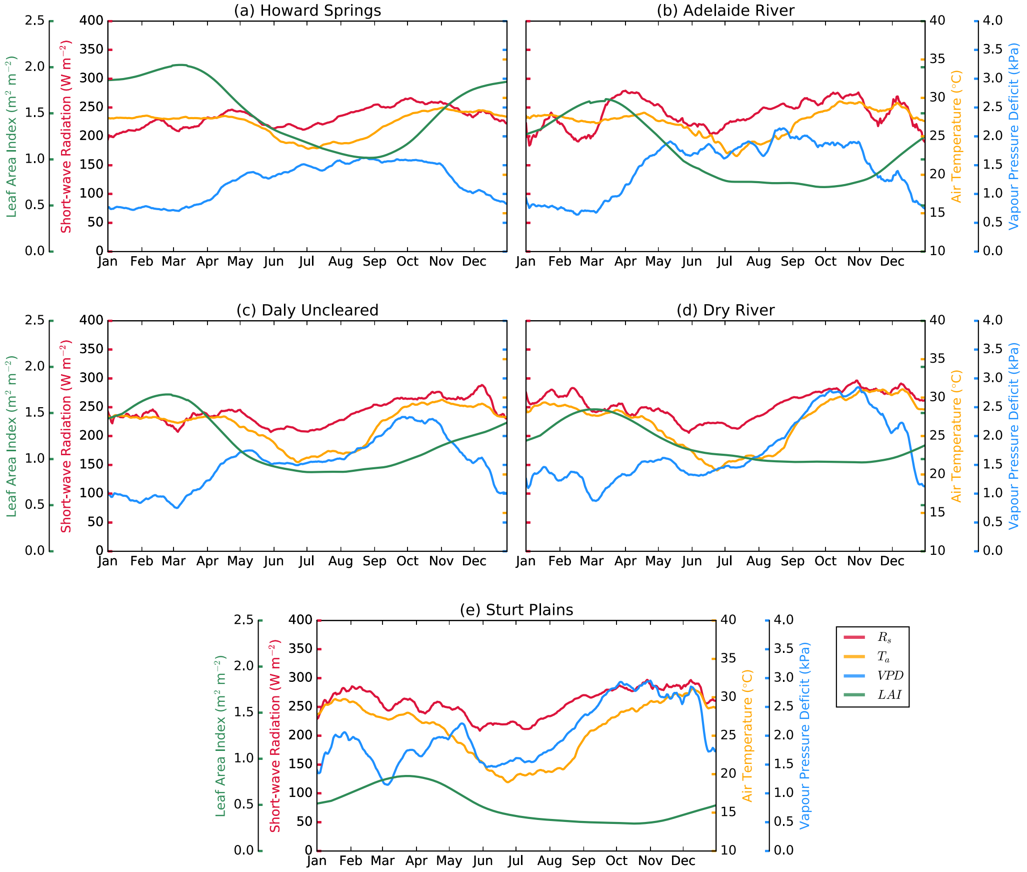


Figure S1: A smoothed (7-day moving average) representation of the environmental drivers used to construct the empirical benchmarks at each of the 5 savanna sites, and are shown from wettest to driest; (a) Howard Springs, (b) Adelaide River, (c) Daly River, (d) Dry River and (e) Sturt Plains. The time-series represents the seasonality over an average year for mean daily solar radiation (R_s), mean daily air temperature (T_a), mean daily vapour pressure deficit (VPD) and leaf area index (LAI).