Using Machine Learning to Link Climate, Phylogeny and Leaf Traits in Eucalypts Through a 50-fold Expansion of Current Leaf Trait Datasets

Karina Guo[1,2], William K. Cornwell[2], Jason G. Bragg[1,2]

[1] Research Centre for Ecosystem Resilience, Botanic Gardens of Sydney, Mrs Macquaries Rd, Sydney, NSW 2000, Australia

[2] School of Biological, Earth, and Environmental Sciences, UNSW Sydney, High St, Kensington, NSW 2052, Australia

karina.guo@botanicgardens.nsw.gov.au

w.cornwell@unsw.edu.au

jason.bragg@botanicgardens.nsw.gov.au

ORCID iD

KG 0009-0008-3420-2950

WKC 0000-0003-4080-4073

JGB 0000-0002-7621-7295

Abstract

- Leaf size varies within and between species, and previous work has linked this variation to the environment and evolutionary history separately. However, many previous studies fail to interlink both factors and are often data limited.

- To address this, our study developed a new workflow using machine learning to automate the extraction of leaf traits (leaf area, largest in-circle area and leaf curvature) from herbarium collections of Australian eucalypts (*Eucalyptus*, *Angophora* and *Corymbia*). Our dataset included 136,599 measurements, expanding existing data on this taxon's leaf area by roughly 50-fold.

- With this dataset, we were able to confirm global positive relationships between leaf area and mean annual temperature and precipitation. Furthermore, we linked this trait-climate relationship to phylogeny, revealing large variation at the within-species level, potentially due to gene flow suppressing local adaptation. At deeper phylogenetic levels, the relationship strengthens and the slope converges towards the overall eucalypt slope, suggesting that the effect of gene flow relaxes just above the species level.

- The strengthening of trait-climate correlations just beyond the intraspecific level may represent a widespread phenomenon across various traits and taxa. Future studies may unveil these relationships with the larger sample sizes of new trait datasets generated through machine learning.

Introduction

As a fundamental unit of photosynthesis, leaf area has impacts across a variety of processes. This has led to an extensive body of research, ranging from regulating carbon flux over vast areas of the earth (Reich 2012), to influencing ecosystem dynamics by affecting the plant's individual growth and survival (Wang et al. 2019, Wright et al. 2017, Leigh et al. 2017). Therefore, an improved understanding of leaf area variation can facilitate better predictions for plant adaptation to changing climates (Wang et al. 2022, Pritzkow et al. 2020). This, in turn, will enable better comprehension of leaf energy balances (Wright et al. 2017) and their relationship with models of forest productivity and plantation growth (Madani et al. 2018, Reich 2012, Battaglia et al. 1998).

The distribution of a plant's traits may be tied to their environment (Li et al. 2020, Souza et al. 2018, Wright et al. 2017, Moles et al. 2014), and this link may manifest in different forms. One potential form of a trait-climate relationship is when variation is constrained by one or more limits that shift with climate. In this case, two limits may form a tight relationship (e.g., Reich 2003), and one limit forms a 'constraint triangle' that contains a probabilistic distribution of traits across the landscape (e.g., Wright et al. 2017, Guo et al. 2000, Cornelissen 1999). For leaf area, mean annual precipitation and temperature are two key environmental drivers that affect this triangle. However, current research suggests that there is a significant constraint on maximum leaf area that shifts with climate, whereas there is no corresponding constraint on minimum leaf area (Wright et al. 2017).

Across climatic gradients, leaf area has been found to increase from dry to wet environments and from colder to hotter climates (Souza et al. 2018, Wright et al. 2017, Moles et al. 2014, Peppe et al. 2011). One proposed explanation is that smaller leaves, particularly leaves with narrow effective widths, possess more effective thermal regulation and reduced water loss through a smaller boundary layer. This layer is a thin space around the leaf with reduced air movement, promoting cooling (Leigh et al. 2017, Nobel 2009). However, the relationship between leaf area and climate is complex. For instance, studies have shown thermal constraints on leaf area to be ineffective in ever-wet conditions (Souza et al. 2018. Wright et

al. 2017). Therefore, while a general relationship exists between leaf area and climate, it is influenced by various factors.

Empirical research at differing geographical and taxonomic scales have yielded varied results on the relative importance of temperature and precipitation in influencing leaf traits; with regional trait-climate correlations possibly being decoupled at local scales (Ackerly et al. 2007). For instance, in Australian eucalypt vegetation stands, Ellis & Hatton (2008) found water availability to play a greater part than temperature in explaining leaf area index. On the other hand, in central Europe, Meier & Leuschner (2008) found leaf expansion of *Fagus sylvatica (L.)* stands primarily controlled by temperature, consistent with a global meta-analysis (Moles et al. 2014). Similarly, leaf area index in *Melaleuca lanceolata (Otto)* in southern Australia was found to have a stronger association to mean maximum temperature than precipitation (Hill et al. 2014). Here, our study aims to clarify this relationship between both climatic variables and leaf traits of Australian eucalypts through a unique workflow. In turn, this can contribute to a better local understanding of ecological processes and improved predictions of trait composition (Peppe et al. 2011, Violle et al. 2007).

When studying the variation in leaf area across climate, it is important to also consider the influence of evolutionary history (e.g., Milla & Reich 2007, McDonald et al. 2003, Ackerly et al. 2002). Varying effects of phylogeny, and contemporary demography (intraspecific gene flow) may result in trait-climate relationships within species being weaker, unrelated, or even following opposite directions to that reported among species (with various potential scenarios illustrated in Fig. 1) (Wilde et al. 2023, An et al. 2021, McDonald et al. 2003, Ackerly et al. 2002). For instance, in Figure 1 Scenario 2, gene flow between populations may prevent adaptation to local environments, counteracting environmental pressures (reviewed at Alexander et al. 2022, Leimu & Fischer 2008). Additionally, an individual's evolutionary history may constrain phenotype and local adaptive capacity (Fig. 1 Scenario 3, An et al. 2021, Leimu & Fischer 2008). This intraspecific trait variation (ITV) has been debated in previous studies. Some have suggested that ITV may obscure general trends (Bastias et al. 2017, Ackerly et al. 2002), while others argue that it does not have such an impact (Westerband et al. 2021, Li et al. 2020, Mudrák et al. 2019). This conflict is potentially due to the limitations of datasets generated using traditional methods (also suggested by Li et al.

2020, Bastias et al. 2017). Regardless, studies of links of leaf traits and climatic variables across varying evolutionary scales, from ITV (e.g., An et al. 2021) to major plant families (e.g., Wilde et al. 2023, Ackerly & Reich 1999), is critical to predicting phenotypic evolution and shifts in traits under a changing climate.

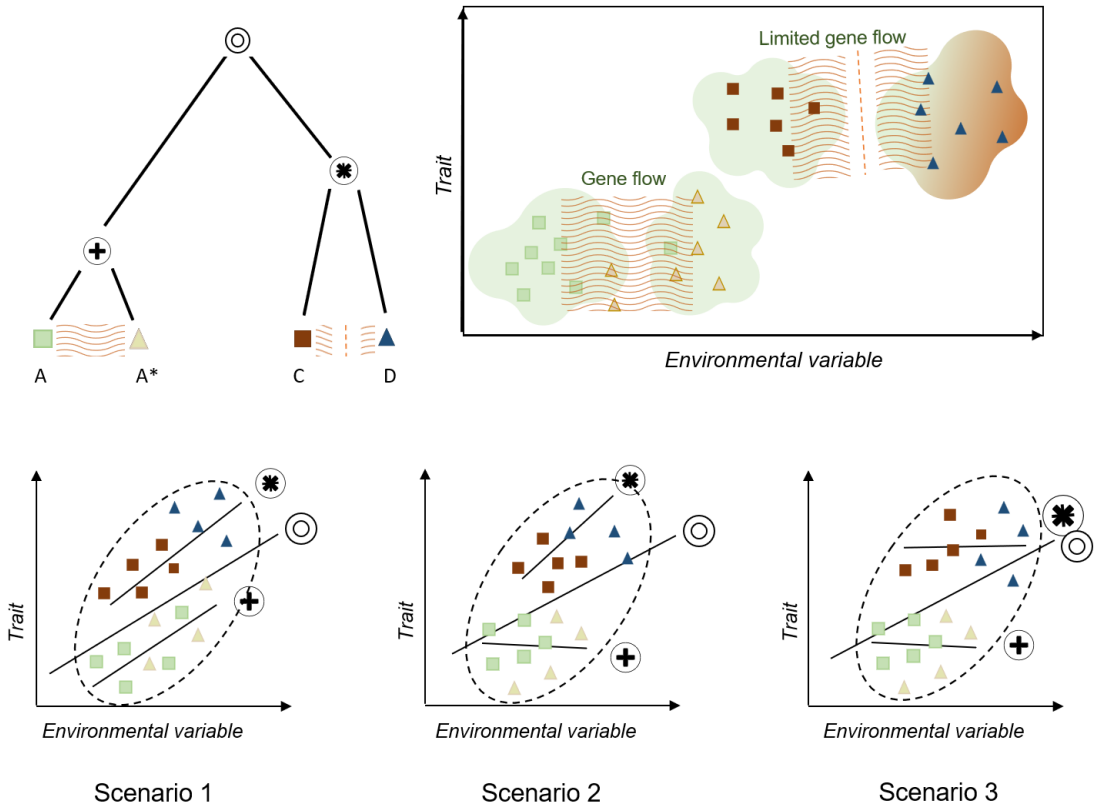

Figure 1. Three scenarios illustrating impacts of evolutionary divergence and intraspecific gene flow on trait-climate relationships. Groups A and A are populations of a species and remain connected by gene flow, while groups C and D are quite recently, but completely, diverged and have limited recent gene flow. The circles represent different internal nodes within the hypothetical phylogenetic tree. In all three scenarios, there is a positive overall trait-climate association.*

In Scenario 1, there is a strong trait-climate relationship within each of the two recently diverged clades, resulting in roughly similar slopes in each clade.

In Scenario 2, gene flow strongly suppresses local adaptation within species, potentially causing divergence from overall trait-climate trends. This effect is however relaxed in recently diverged groups. Therefore, the clade consisting of A and A does not exhibit a trait-climate relationship, and the clade containing groups C and D exhibits a strong trait-climate relationship.*

In Scenario 3, trait evolution is more constrained, so that strong adaptation is observed only among longer diverged groups. Here, there is no trait-climate relationships within the clade containing A and A or C and D, but there is an association overall, reflecting adaptation over longer time scales.*

Understanding evolution of leaf morphology has a recognised importance (Mudrák et al. 2019, Souza et al. 2018, Leimu et al. 2008). Despite this, there is a paucity of research that examines leaf variation in the perspective of phylogeny and ITV simultaneously. One potential reason lies in the laborious and time-intensive nature of data collection (Li et al. 2020, Bastias et al. 2017), which traditionally involve manual measurements of each data point. This makes it difficult to gather datasets with high intraspecific sampling within and across different clades and climates (Li et al. 2020, Bastias et al. 2017). As a consequence, few studies spanning both intraspecific and phylogenetic scales simultaneously have been conducted (see also Wilde et al. 2023, Cutts et al. 2021, Goëau et al. 2020, Pearson et al. 2020, Brenskelle et al. 2020).

This study addresses this by using machine learning (ML) paired with herbarium records. Herbarium specimens are pressed plants of various taxa collected globally. These specimens provide a holistic representation of plant shoots and include both mature and juvenile leaves (Kozlov et al. 2021). As a consequence, trait measurements from these sheets will encompass leaves at different developmental stages, propagating into resulting datasets. Herbarium specimens provide extensive phylogenetic and geographic sampling. However, their potential has remained underutilised due to the impracticality of extracting trait data using traditional methods (Heberling 2022). Thus, we employed ML as a new tool to automate the extraction of trait data from these specimens. Previous studies have used ML to extract leaf traits from digital herbarium specimen images (Hussein et al. 2021, Weaver et al. 2020, Younis et al. 2018). However, to our knowledge, this approach is the first to utilise machine learning operationally in trait ecology, allowing us to create a comprehensive dataset that spans various taxonomic levels across Australia. By pairing this dataset with a fully resolved phylogenetic tree (Thornhill et al. 2019), we could link microevolution to macroevolution, enabling a better observation of the shift in trait-climate relationships across different clades and evolutionary depths.

Overall, leaf morphological traits enable better comprehension of leaf energy balances (Wright et al. 2017), improving our understanding of ecosystem dynamics (Pritzkow et al. 2020) and global vegetation models (Madani et al. 2018, Reich 2012, Battaglia et al. 1998). Despite this, there is a paucity of datasets spanning a wide phylogenetic and spatial range (Moran et al. 2016). Our study proposes a method to address this gap by using ML to bypass traditional trait-collection methods. In particular, we sought to address the following questions:

a) Could ML be used to automatically extract various commonly measured leaf morphological traits, including leaf area, and the largest in-circle area? This will allow us to build a large dataset, unique in its ability in allowing us to answer the following questions simultaneously in the study taxa.

b) How do leaf traits shift across the Australian climate? We hypothesise that leaf area and largest in-circle area will correlate positively with mean annual precipitation and temperature.

c) To what extent does phylogeny shape leaf traits? We hypothesise that gene flow will resolve in large trait variability at a shallow phylogenetic level (within species), which will gradually resolve to a trait-climate relationship at deeper levels (for example, among species).

Our study and its findings help reveal the relationship between traits and their influences, in addition to formulating a more efficient method of trait data collection, applicable to additional taxa and traits in the future.

Method

Study clade and design

This study focused on eucalypts, which are the dominant canopy trees throughout many Australian forests and shrublands (Booth et al. 2015, Govindan 2005). The eucalypt clade consists of three genera, *Eucalyptus (L'Hér.)*, *Angophora (Cav.)*, and *Corymbia (K.D. Hill & L.A.S. Johnson)*. They were selected as the study genera for their wide distribution across Australia's temperature and precipitation range (Fig. 2), the availability of a molecular phylogeny for the clade (Thornhill et al. 2019), and characteristic simple leaves with entire margins. These features allowed us to explore the impact of climate and phylogeny as drivers of leaf trait variation at different evolutionary scales, with the aid of machine learning (ML).

Digital images of herbarium sheets from the National Herbarium of New South Wales (downloaded from https://herbariumnsw-pds.s3-ap-southeast-2.amazonaws.com/images/) were used to capture trait variation across wide spatial and environmental ranges (Fig. 2). This enabled the study of traits in a broader range of lineages and biomes than data collected using observational approaches (Heberling 2022). Herbarium specimens are collected with the aim to record traits present in the population (Kozlov et al. 2021) and thus include both mature and immature leaves. As such, our workflow uses a novel approach of trait sampling that diverges from conventional sampling methods of physiologists, which target fully expanded leaves (e.g., in Wright et al. 2017 and Pérez-Harguindeguy et al. 2013). This distinction is critical within eucalypts due to the significance of ontogeny in leaf morphology, and it is worth noting the important implications it plays in the analysis.

Our project aimed to generate a large dataset of leaf measurements from digital images of eucalypt herbarium specimens and use it to test ecological associations. This dataset would be unusual in its combination of wide spatial distribution (Fig. 2a) and its deep intra- and interspecific sampling. To do this, the method consisted of three separate parts. (i) Develop and refine a leaf masking model, (ii) develop and refine a leaf classification model, (iii) application of models to produce a large trait dataset and carry out quantitative analysis of trait-climate relationships in a phylogenetic framework. An overview of this workflow is found at Figure 4, and relevant data and scripts are available in the Supplementary Information.

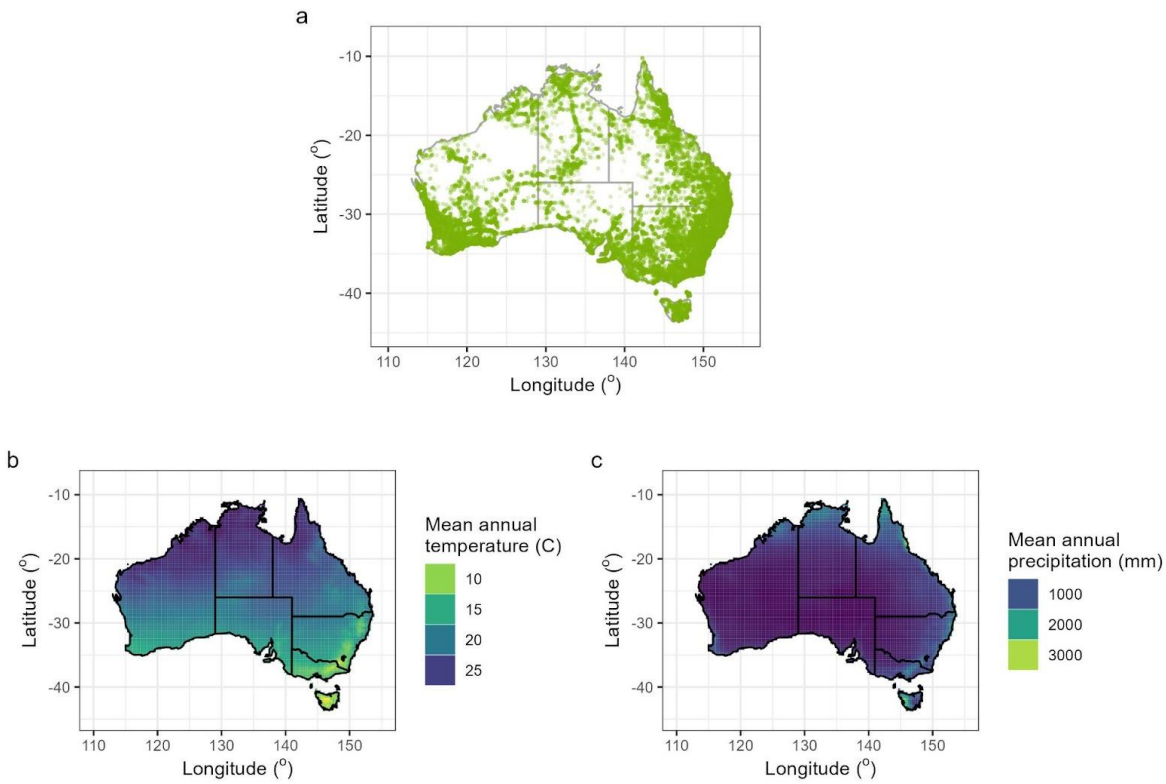

Figure 2. The spatial distribution of sampling. a) The location of each data point of leaf trait measurement. b) The mean annual precipitation across Australia as sourced from WorldClim. c) The mean annual temperature across Australia as sourced from WorldClim, indicating the range of climatic variables the sampling encompasses.

Leaf Masking Model

A convolutional neural network (CNN) model was trained to find leaves and pixels that belonged to each ‘instance’ of a leaf (known as instance segmentation). The CNN model used a ResNet50 architecture (He et al. 2015) and was implemented in Detectron2 (Wu et al. 2019). Transfer learning was performed to reduce the amount of training required. It was conducted from a pretrained model, a Mask R-CNN model with a ResNet50-FPN backbone that was pretrained on the COCO dataset (Lin et al. 2014). Extra details of the model and methods used to train, validate, and test can be found in Supplementary Information A and B. A table of definitions has also been provided in Table 1.

ML models 'learn' patterns through a set of training data that has been manually annotated. In this case, our model is 'learning' to identify pixels of a leaf using annotated images of herbarium specimens. Generating these manual annotations involved creating a polygon around each instance of a leaf following a protocol provided in Supplementary Information B. All annotations were made using the program LabelMe (v 5.01, Wada 2022). In total, 113 manually annotated herbarium sheets were used to train the model, a further 28 were used for validation during training (for adjustment of hyper-parameters by Detectron2) and 20 were used for testing the performance of models after training (for manual adjustment of training parameters).

The final model was refined using an optimisation process. This involved: (i) Training the initial model using the manually annotated training and validation data set, (ii) Predicting leaves for images of the testing data set using the trained model, (iii) Gathering quantitative and qualitative measures of model accuracy from part ii, (iv) altering the model’s training parameters and repeating the cycle at part (i) with a new model. Different iterations of the model are described in Table SA_1.

As part of step iii) of the optimisation process, we calculated a set of standard metrics of model quality, based on the predictions the model made on the test dataset. These metrics were calculated by comparing the masks predicted by the model, to the ground-truth that we manually annotated.

First, Intersection Over Union (IoU) was calculated for each predicted mask generated by the model (P_{mask}) and each ground-truth mask that was labelled (G_{mask}) (Eqn. 1).

$$IoU = \frac{area(P_{mask} \cap G_{mask})}{area\ (P_{mask} \cup G_{mask})} \qquad \textit{(Eqn. 1)}$$

Leaf pairs with an IoU of greater than 70% were regarded as a correct prediction. These were used to calculate precision and recall. Precision is the number of correct predictions compared to all predictions made (Eqn. 2).

$$Precision = \frac{Correct\ predictions}{All\ predictions} \qquad \textit{(Eqn. 2)}$$

Recall is the measure of the number of true positive masks present compared to how many there were actually in the ground-truth (Eqn. 3).

$$Recall = \frac{Correct\ predictions}{All\ groundtruth} \qquad \textit{(Eqn. 3)}$$

The F_1-score combines precision and recall into a single score, allowing it to be evaluated simultaneously (Eqn. 4).

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \qquad \textit{(Eqn. 4)}$$

We used these metrics of accuracy, as well as visual inspections of predictions, to make changes to the model's training parameters and improve performance. We note that we placed greater emphasis on obtaining high levels of precision than recall. This is because we expected that missing real leaves would have a smaller effect on our downstream analyses of leaf area than erroneously including incomplete leaves.

227 *Table 1. A table of definition of commonly used terms*

Phrase	Definition
Convolutional neural network (CNN)	A neural network (algorithms) specifically tailored for image analyses
Instance segmentation	The finding of objects and their segmentation mask, a path that indicates the outline of a polygon that masks the object in question
Annotation	The process of labelling input data to indicate the desired variable. In this case this involved tracing each individual leaf with a polygon
Ground-truth	Ground-truth refers to the correct value of the labels for a given dataset. It is determined through manual annotation and used as a comparison against the model's prediction
Train	The provision of the training dataset to the model's algorithm to allow it to learn the designated task
Validation	The process of evaluating a model's performance and adjusting its hyper-parameters during the training process
Test	Testing the trained model on a testing dataset to evaluate performance
Intersection over union (IoU)	A value that defines how similar the predicted label is to the ground-truth label. Where it is calculated by the intersection of the two labels over the union of the two labels (Eqn. 1). The best value for this measure is 1 or 100%
Recall	The measure of the number of true positives. It is the proportion of actual positive cases that were correctly identified by the model as positive (Eqn. 2)
Precision	The ratio of true positive cases compared to the total number of cases that the model predicted as positive (Eqn. 3)
F_1-score	Also called the harmonic mean of recall and precision. Used to generate a value that balances precision and recall (Eqn. 4)
True positive (TP) True negative (TN)	True positive: Correct prediction of a positive class, for example correctly identifying a valid leaf

False positive (FP) False negative (FN)	True negative: Correct prediction of a negative class, for example correctly identifying an invalid leaf False positive: Incorrect prediction of a positive class False negative: Incorrect prediction of a negative class
Training parameters	May also be called hyper-parameters. Values that are set prior to training by the researcher and defines how the model operates during training.
Hyper-parameters	Values that are changed automatically during the training and validation stage when creating a machine learning model. These include 'weights' that are used to adjust the model's parameters to improve accuracy. These values are not adjusted manually.

Leaf Classification Model

A CNN model of ResNet50 architecture (He et al. 2015), implemented in PyTorch (Paszke et al. 2019) and pretrained on ImageNet data (Deng et al. 2009), was trained to classify images of leaves as valid or invalid. This classifier was applied to the leaves predicted from the leaf masking model as another level of filtration to increase the final precision of our workflow. Here, valid leaves were defined as having more than 90% of the whole blade visible, along with other criteria (Supplementary Information A). Extra details on the model are located in Supplementary Information A.

Digital images of herbarium sheets were used to generate the training, validation, and testing datasets. This was done by first using the leaf masking model, described above, to create predicted leaf masks from herbarium sheets (examples in Fig. 3). Each separate leaf mask was then manually classified as 'valid' or 'invalid', then split into their respective datasets. To prevent an imbalance of training data, the final training dataset was truncated to an equal number of valid and invalid images, totalling to 447 images of each category.

To test the model, we used it to classify the images in the testing dataset. These predictions were then compared to our manual classifications. From this, we generated similar evaluation metrics, calculated using Equations 5. Here, true positives are 'valid' classifications that matched the ground-truth, and true negatives are 'invalid' classifications that matched the ground-truth.

$$Precision = \frac{True\ positives}{True\ positives\ +\ False\ positives}$$

$$Recall = \frac{True\ positives}{True\ positives\ +\ False\ negatives} \qquad \textit{(Eqns. 5)}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Similar to the leaf masking model, we carried out a process of optimisations where we changed different training parameters following the qualitative and quantitative evaluation metrics. All iterative steps in the model generation can be found in Table SA_3, and vary in training epochs, classification criteria, and the volume of training dataset used.

The same testing dataset (i.e., same herbarium sheets) was used in both the leaf masking and leaf classification models. This enabled us to examine how the classifier affected the evaluation metrics of the workflow. This was done by using the classifier to filter out invalid leaves from the leaf masking model's predictions. Precision, recall and, the F_1-score of the results were then recalculated from the ground-truth. These values thus reflected the combination of predictions of the leaf masking model, and filtering by the classification model.

Trait Extraction

From each predicted leaf mask, we extracted three key traits including i) the area of the mask, ii) the area of the largest in-circle within the mask (similar to Leigh et al. 2017), and iii) the curvature of the leaf. Area measurements were calculated by converting the number of pixels

in the mask into cm^2 using the known resolution of the images (561 x 561 dpi). The area of the largest in-circle was calculated using the radius from the Pole of Inaccessibility (from package polylabelr v 0.2.0, Larsson 2020) a geographical point furthest from the edges, correlating to the visual centre of the polygon. Leaf curvature was calculated through a proxy of the ratio between the area of the concave hull : leaf area (more curved leaves have higher values). However, leaf curvature is not a focus of this paper, and all analyses conducted for this trait are reported in Supplementary Information C. All trait extractions and analyses were carried out in R (v 4.2.2, R Core Team 2022) and are further elaborated in Supplementary Information A. The masks used to generate these measurements were predicted by the leaf masking model and classified as valid by the leaf classification model. They were then subject to a 4-connected component analysis. Duplicate predicted masks sometimes occurred and were filtered out by calculating IoU values between predictions of leaf masks on the same herbarium sheet. IoU values greater than 70% between two predicted masks were considered duplicates.

Leaves shrink in size when drying. As leaf area is conventionally measured on fresh leaves, we addressed this by dividing the values for leaf area and largest in-circle area by 0.8973. This value is sourced from the Terrestrial Ecosystem Research Network (TERN), Australia's national land ecosystem observatory, who determined shrinkage to be consistent across *Eucalyptus* leaves (Morgan et al. 2021). We note that the application of a constant multiplier should not affect the slopes or the significance values of any statistical analyses.

Trait values for each leaf were then aggregated into a final dataset through joining the metadata (located at https://herbariumnsw-pds.s3-ap-southeast-2.amazonaws.com/dwca-nsw_avh-v1.0.zip). Metadata fields included the sheet’s genus, specific epithet, decimal latitude, and decimal longitude. Climatic data for each sheet were sourced from WorldClim v2 at resolution 2.5 minutes (Fick & Hijmans 2017) by each sheet’s geolocation. This included the variables Annual Precipitation (BIO12) and Annual Mean Temperature (BIO1) and are referred to as mean annual precipitation and mean annual temperature, respectively.

Analysis

This study aimed to examine the relationship between traits and climate. This was analysed using i) a linear model between the trait and climatic variables (Eqn. 6). ii) A similar linear model, with the mean trait value of each species as a data point (Eqn. 7). This analysis allowed us to account for errors in sampling bias of certain species, improving the generality of the trait-climate relationships. iii) A linear mixed model with the species as a random effect, and herbarium sheet nested within species (Eqn. 8). This examined trait-climate associations while accounting for inter- and intraspecific variation, whilst also appropriately modelling the variation from leaves in the same herbarium sheets, iv) Linear quantile regressions between trait and climatic variables (Eqn. 6 at different quantiles), were used to estimate the limits of the environmental constraint on the trait variables. This method of analysis was as suggested by Guo et al. (2000) to illustrate a 'constraint triangle'.

$$Trait \sim Climate \qquad \textit{(Eqn. 6)}$$

$$Mean\ Trait\ Value \sim Climate \qquad \textit{(Eqn. 7)}$$

$$Trait \sim Climate\ +\ 1|Species/SheetID \qquad \textit{(Eqn. 8)}$$

Prior to all analyses, data points outside Australia were removed using the package CoordinateCleaner (v 2.0-20, Zizka et al. 2019) and right-skewed variables (leaf area, largest in-circle area, and mean annual precipitation) were log-transformed to satisfy the analyses' assumptions. Furthermore, an inclusivity criterion was applied for analyses ii) and iii), where species with fewer than 10 data points were removed from the dataset. This was done to ensure that the model was based on groups with sufficient sample sizes.

Across analyses i) to iv), comparisons of slope, R-squared, and standard error, were made to larger global datasets including that used in Wright et al.'s (2017) analysis and eucalypts in the AusTraits (Falster et al. 2021) database. These comparisons revealed how eucalypt's trait-climate relationship shifted in comparison to global taxa, and the implications of the study's trait sampling method.

An additional set of analyses was performed to ask whether trait-climate relationships were consistent at different evolutionary scales (Figure 1). To do this, our phylogenetic analyses used the dated maximum likelihood (ML2) tree from Thornhill et al. (2019), pruned to contain only the species present within our trait dataset. We first investigated whether phylogeny impacted leaf trait variation through determining the phylogenetic signal. This was carried out using the function phylosig, from the package phytools (v 1.5-1, Revell 2012), which measured how closely the traits reflect a taxa's evolutionary history. This avenue was further explored through observing how the trait-climate relationship altered throughout the taxonomic levels. To do this, a linear model was conducted where the groups at each respective level were designated as random effects. A final novel analysis was carried out. Thornhill et al.'s phylogeny was split at 20 evenly spaced time intervals along the entire length of the tree. At each time interval, tips that had split prior to the point were kept as individual unique lineages, while those that had split after the time interval was merged by common ancestry into a single 'lineage'. For instance, the 1st interval was at 0 million years ago and included every tip of the tree as a lineage (418 lineages). Whereas, at the 3rd interval, 8.57 million years ago, 77 lineages were present. These included groups comprising of individual species and others containing multiple species aggregated into one lineage. A mixed model analysis with each lineage as a random effect, was then performed at each of the 20 intervals. This was done to estimate the average slope of the trait-climate relationships within these lineages.

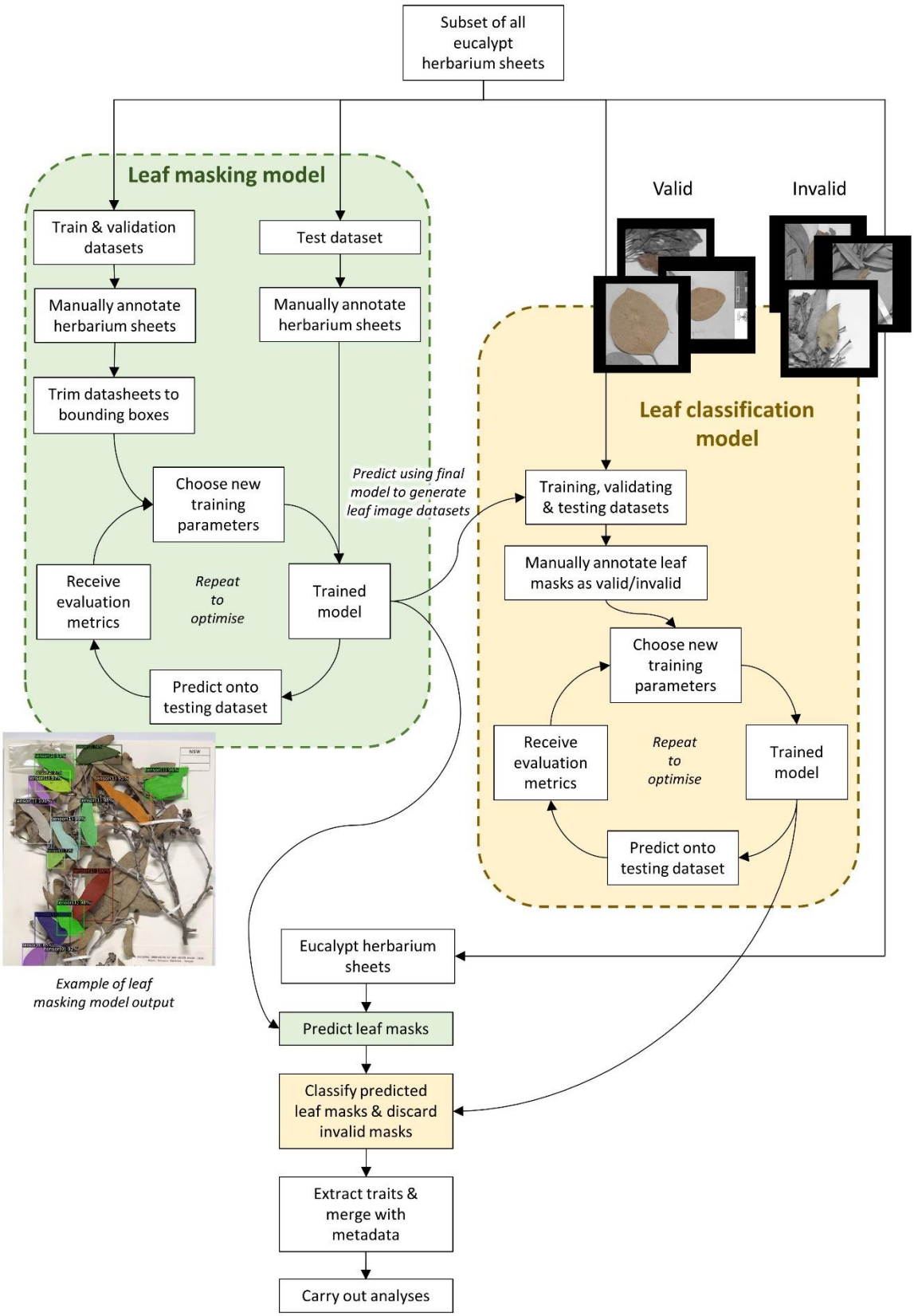

Figure 3. Workflow of the process to create the trained models and the subsequent dataset. Illustrating the generation of the two key models, a leaf masking model and a leaf classification model, followed by their application.

Results

Machine learning produces a large leaf trait dataset with high precision

Our workflow generated a large leaf trait dataset of eucalypts across Australia. Here, we first describe the dataset, including the validity and accuracy of our workflow, before exploring the analyses performed on our dataset. Error validation and extra analyses, including those based on conventional methods used by physiologists, are located in Supplementary Information C and D.

The final leaf trait dataset contained 139,599 measurements across 1,534 separate taxa (including species, hybrids, subspecies, and collector identifications). The number of leaves detected in a species ranged from 1 to 2,430, before the inclusivity criterion was applied. Examples of leaf masks are shown in Figure 4. A comparison of the distribution and volume of our leaf area against AusTraits and Wright et al. (2017) for several exemplar species has been illustrated in Figure 5. This study's sampling method resulted in greater variation in leaf area measurements and a greater representation of smaller leaves (Fig. 5b). This is further reinforced with the quantile regressions explored later.

The leaf masking model had a precision value of 77%, meaning this percentage of leaves predicted onto the testing dataset were valid leaves. When the leaf classifier, with a precision of 67%, was applied to the outcomes of these predictions, the overall workflow's precision increased to 82%. The recall value indicates the percentage of valid leaves that were identified. Of the leaf masking model, this was initially at 68% and the leaf classification model at 63%. When applied together, the workflow's recall reduced to 34%. F_1-scores were 73% for the leaf masking model, 65% for the leaf classifier model and together the overall workflow's score was 48%.

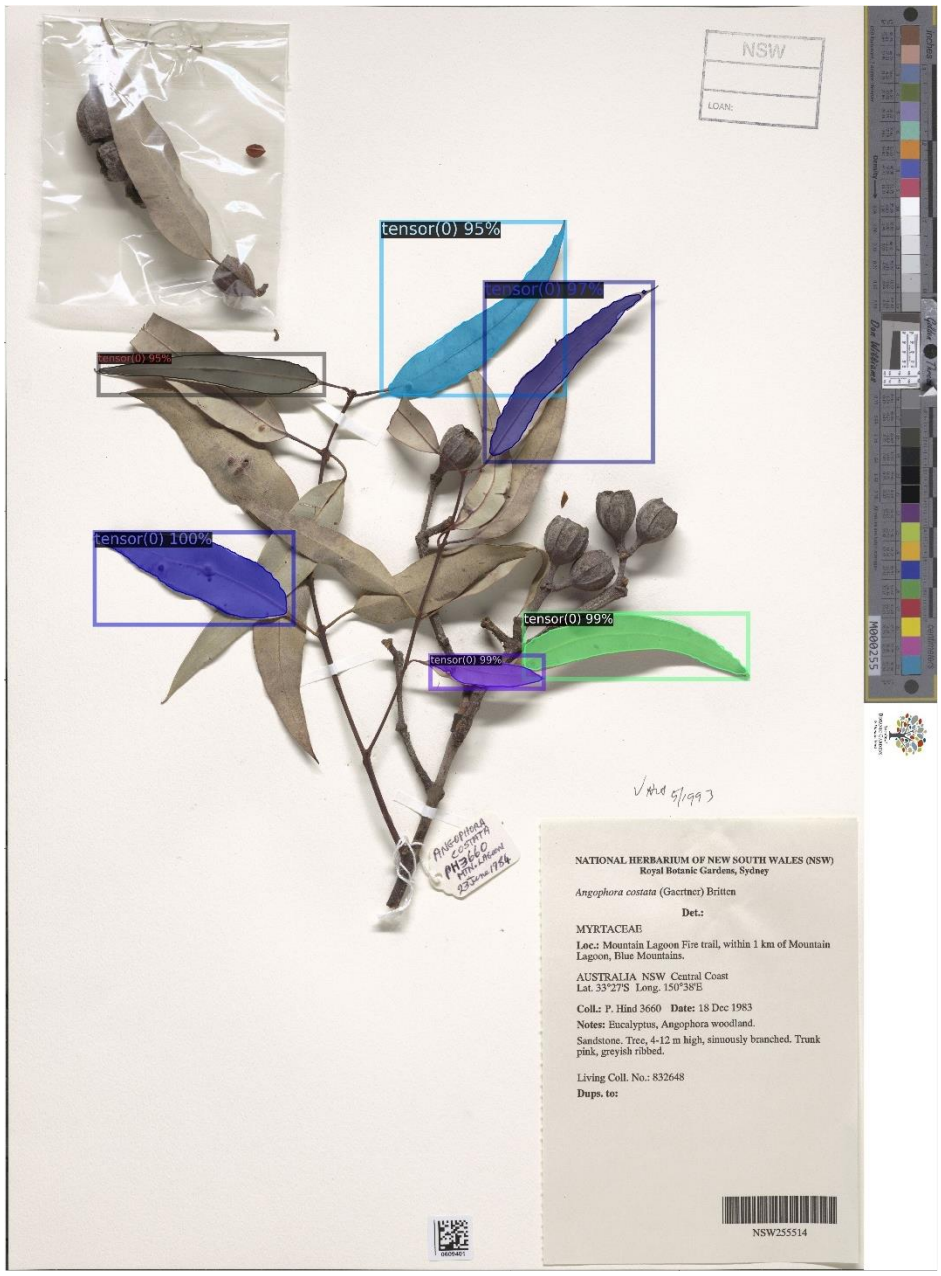

Figure 4. Example of predicted leaves on an herbarium sheet carried out by the leaf masking model. An example of a juvenile leaf being masked can be seen in purple.

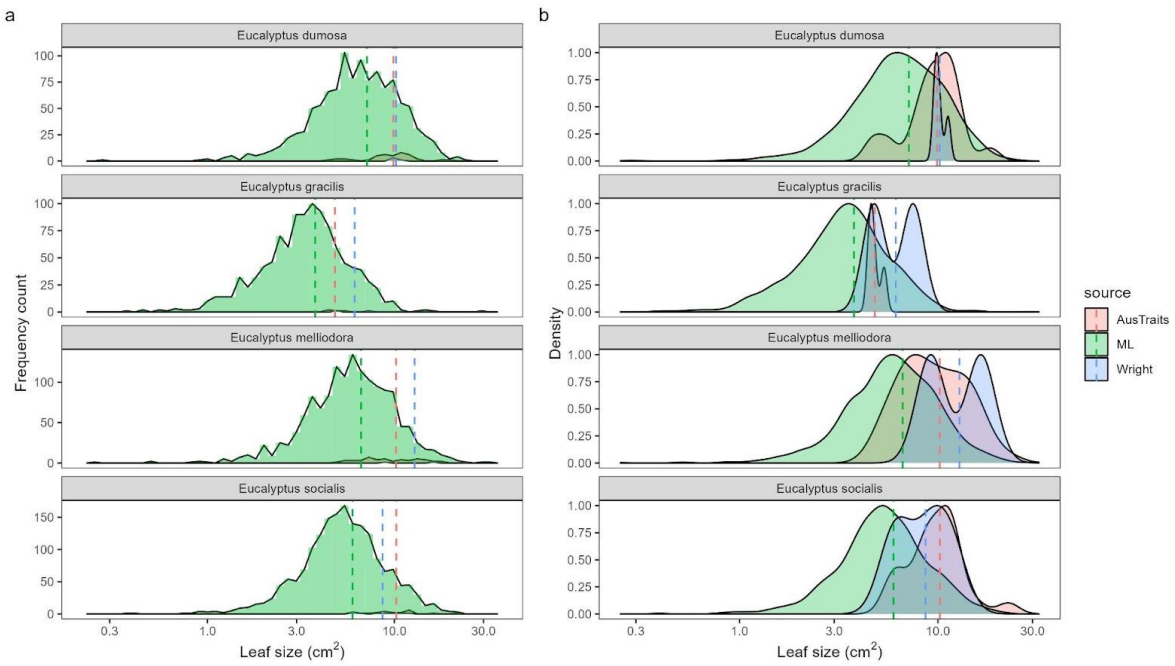

Figure 5. a) Frequency histograms of leaf area measurements for the four most sampled species that are shared in all three datasets (this study (ML), AusTraits, Wright et al. 2017 (Wright)). The present study generated a much greater number of measurements for each species. b) Density frequency distributions for the same species, illustrating the greater representation of smaller leaves in the present study. The dashed lines represent the mean value of the dataset.

Leaf area is positively associated with precipitation and temperature among eucalypts

Leaf area and largest in-circle area were positively associated with mean annual temperature and precipitation in the present study, however, only weakly with the former (Fig. 6). The R^2 value and slopes for the linear models were smaller in comparison to Wright et al.'s (2017) and AusTraits eucalypt datasets (Tbl. 2). Additionally, each of the three eucalypt datasets had a shallower slope than the results for the global study of Wright et al. (2017) (Tbl. 2).

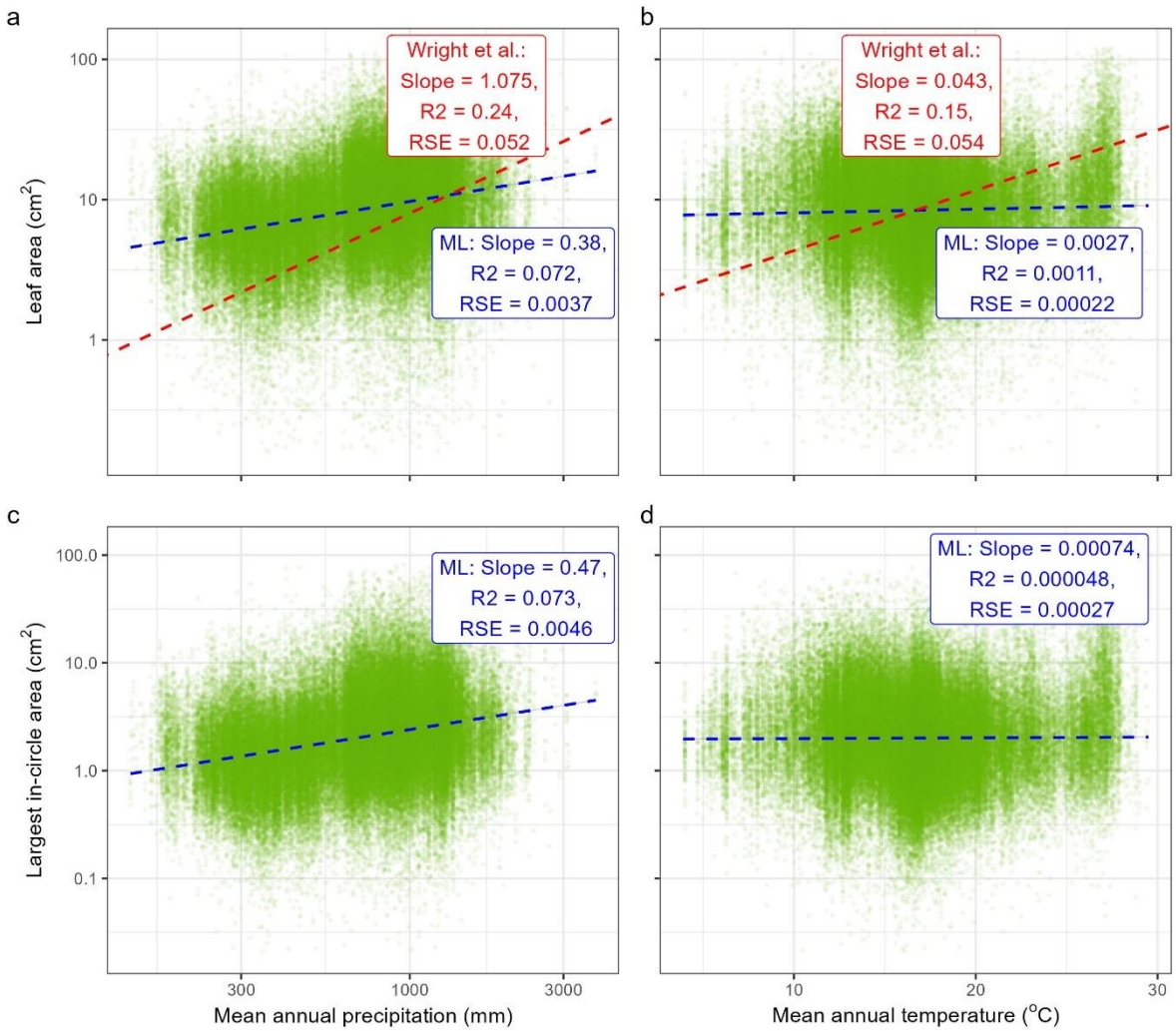

Figure 6. Relationships between the climatic variables (log mean annual precipitation and temperature) against trait values (log leaf area and log largest in-circle area). Plots a and c are log-log relationship plots, while plots b and d are semi-log relationship plots. The blue dashed lines represent the linear model results. The red dashed lines represent the results found in the Wright et al. (2017) analysis of global leaf traits. Values reported are for their respective linear model, where 'Wright et al.' corresponds to the red dashed line and 'ML' corresponds to the blue dashed line, our machine learning dataset.

Table 2. Coefficients of models for log leaf area against log mean annual precipitation and mean annual temperature in comparison to other datasets. Coefficients for Wright's data were sourced from the supplementary information of Wright et al. (2017), which used a mixed regression model.

	Slope	R-squared	Relative standard error
Log leaf area ~ log mean annual precipitation			
Overall	0.38	0.072	0.0037
AusTraits eucalypts	0.69	0.27	0.021
Wright et al.'s eucalypts	0.45	0.25	0.066
Wright et al.'s all taxa	1.08	0.24	0.052
Log leaf area ~ mean annual temp			
Overall	0.0027	0.0011	0.00022
AusTraits eucalypts	0.011	0.022	0.0011
Wright et al.'s eucalypts	0.011	0.0069	0.0078
Wright et al.'s all taxa	0.043	0.15	0.054

For conciseness, the following sections are focused on the associations between leaf area and precipitation. Further leaf trait results are presented in Supplementary Information C. The focus on leaf area will allow for comparison to other datasets (Wright et al. 2017 and AusTraits). Though we note here that leaf area and largest in-circle area were strongly and positively associated, and that largest in-circle area exhibited similar associations with climate to leaf area. Likewise, results for mean annual temperature are also located in Supplementary Information C, however not presented here due to the weak correlation found.

When the relationship between leaf area and mean annual precipitation was examined with quantile regression analyses, the slope increased from the 1st quantile (0.17 ± 0.027) to the 99th quantile (0.53 ± 0.0013) (Fig. 7). At the largest quantiles, the regression slopes were similar to the slopes estimated for AusTraits (0.69 ± 0.27) and Wright et al.'s (2017) eucalypt datasets (0.45 ± 0.023).

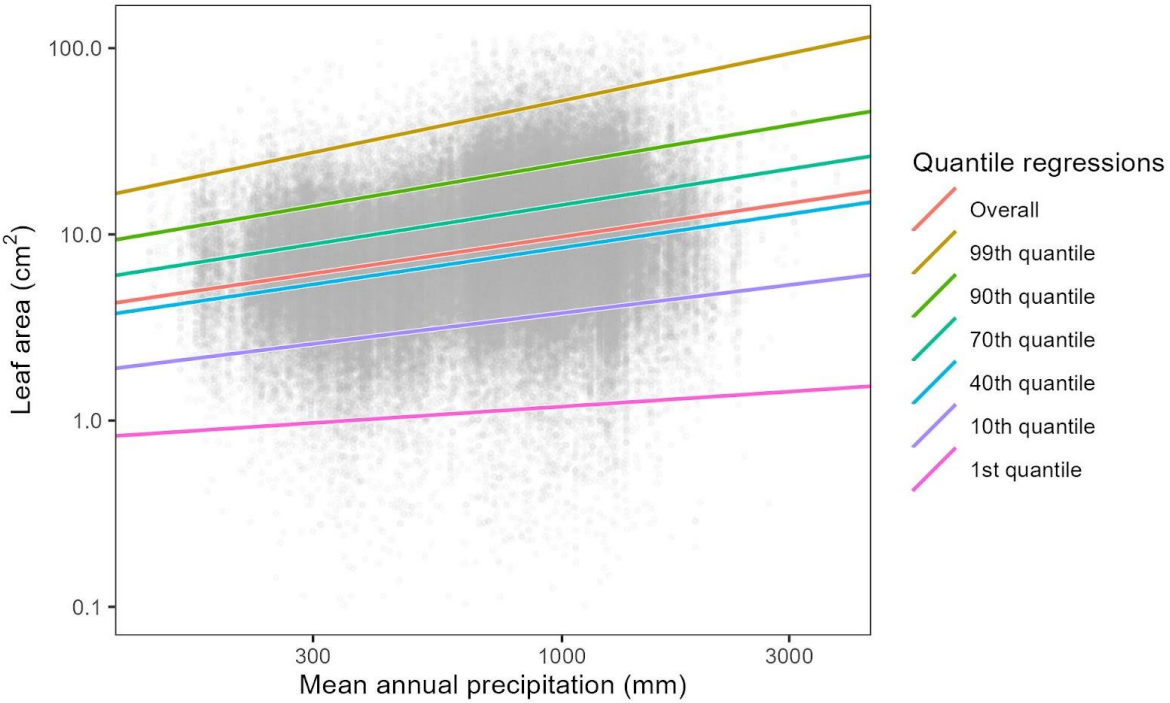

Figure 7. Quantile regression analysis model results. An increase in slope steepness from the 1st to the 99th quantile show a lower range of leaf area variation in drier than wetter conditions as observed.

<u>Eucalypt's leaf trait-climate relationship is constrained by evolutionary history</u>

We next performed several different analyses to consider the effects of taxonomy or phylogeny on the relationship between leaf area and precipitation. When a linear model was fit using the mean trait values of species, the slope was greater and better reflected those of other datasets (Tbl. 3). Additionally, a mixed model with species as a random effect, resulted in an estimate for the mean slope within species. This exhibited a slope smaller than the overall linear model and Wright et al.'s (2017) dataset (Tbl. 3).

Table 3. Coefficients of log leaf area and log mean annual precipitation following Equation 6-8. An overall linear model, a linear model using average species mean, a mixed model with species as a random effect and each herbarium sheet nested within, and Wright et al.'s (2017) *results.*

	Slope	R-squared	Relative standard error
Overall	0.38	0.072	0.0037
Mean species model	0.47	0.2	0.031
Mixed model	0.17	N/A	N/A
Wright et al.'s all taxa	1.08	0.24	0.052

Mixed models were also used to examine the relationship between leaf area and precipitation within groups at levels of taxonomic classifications greater than species. The mean slope for the relationship between leaf area and precipitation was greater within subgenera than within species, and similar to the slopes observed within genera and in the overall model (Fig. 8b). A phylogeny was then used to test the mean slope of the relationship between leaf area and precipitation within lineages at different levels of evolutionary depth in the eucalypts (Fig. 8a). Significant phylogenetic signal, based on the tree estimated by Thornhill et al. (2019), was exhibited for both leaf area (K=0.0021, P=0.001) and mean annual precipitation (K=0.0030, P=0.001). From the shallowest depths of this phylogeny to the deepest, there was an overall increase in the mean slope of the association between leaf area and precipitation within lineages (Fig. 8). At 8.5 MY, the slope drastically increased to a value comparable to that at deeper levels.

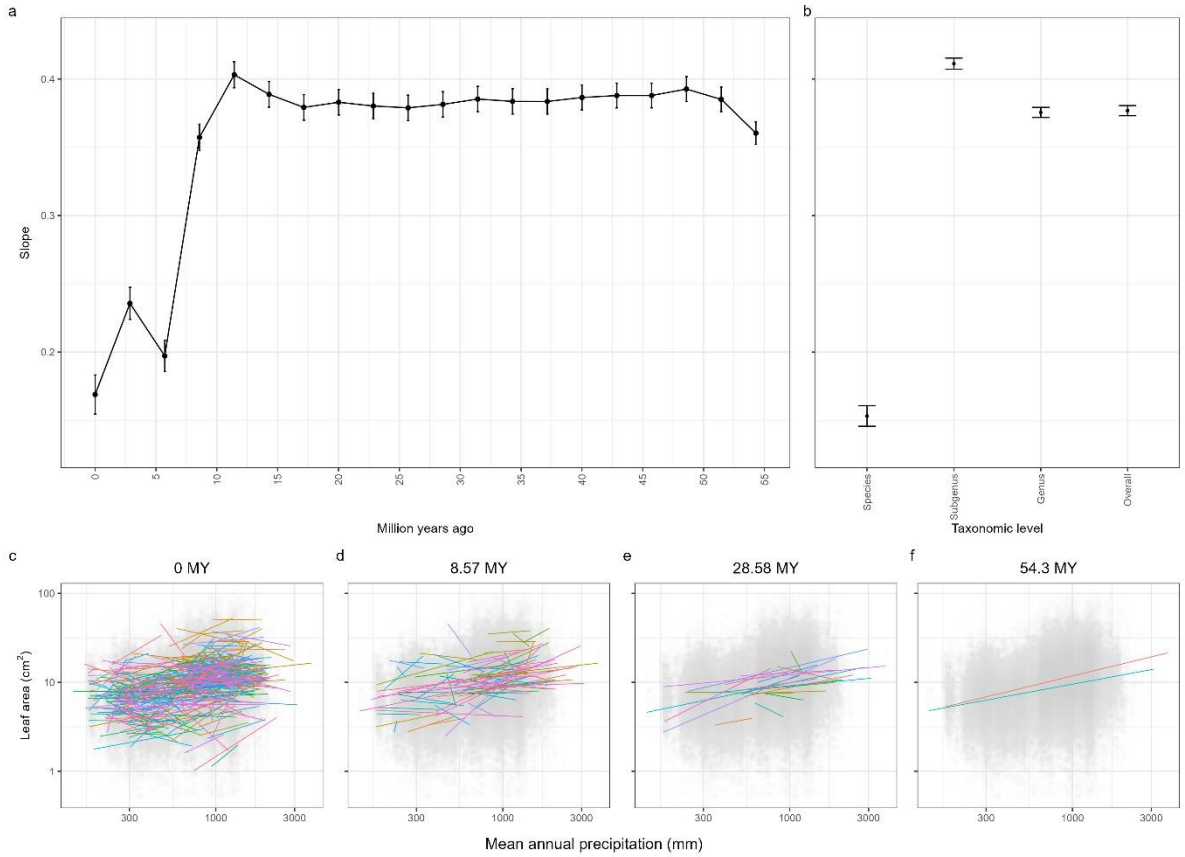

Figure 8. a) The phylogenetic tree was split at 20 intervals at evenly spaced time periods. The mean slope within the lineages at each time point was calculated. For example, 0 MY had each species as a random effect, whereas 55 MY had two groups of species, corresponding to the deepest branch among the eucalypts. A convergence towards an approximate average slope was observed roughly 8.5 MY. b) The average slope and standard error where the respective taxonomic level was used as the random effect in a mixed model. The 'overall' model has no random effect. Species: 0.15 ± 0.0077. Subgenus: 0.41 ± 0.0042. Genus: 0.38 ± 0.0038. Overall: 0.38 ± 0.0037. c-f) Each lineage's linear models at four different intervals (0 MY, 8.57 MY, 28.58 MY, 54.3 MY) are illustrated. Where each colour represents a lineage.

Discussion

Here, we developed, tested, and applied a machine learning (ML) workflow to generate 136,599 leaf trait measurements spanning all species of eucalypts across Australia (Fig. 2). A dataset of this magnitude would not have been feasible using traditional sampling methods, demonstrating the potential of ML in trait ecology. This dataset enabled the analysis of leaf dimensions shaped by climate and phylogeny. We observed a positive relationship between leaf area and both mean annual rainfall and mean annual temperature, which was broadly consistent with previous global observations (Wright et al. 2017, Moles et al. 2014), albeit shallower in eucalypts. Two additional observations offer useful perspectives on this relationship. First, quantile regression models suggest the link between leaf area and precipitation forms a constraint triangle (also seen in Guo et al. 2000). Second, the examination of this trait-climate relationship at different evolutionary scales suggests that, on average, this relationship was not observed within species, but within subgenera and higher taxonomic levels. When examined in relation to phylogenetic depth rather than taxonomy, an association within groups having an age of around 8.5 MY (or between 5 and 10 MY) was found, but not within groups at shallower scales (including within species). Overall, our dataset provides a unique opportunity to study the link between leaf traits and evolutionary history at a scale rarely done in previous studies.

Analysis of this large eucalypt dataset found associations with climate that were largely consistent with previous studies (An et al. 2021, Wang et al. 2019, Souza et al. 2018, Leigh et al. 2017, Wright et al. 2017). We found eucalypt leaf area to have a stronger association with mean annual precipitation than with mean annual temperature, supporting the findings of Ellis & Hatton (2008). The relationship between leaf area and mean annual precipitation formed a constraint triangle (see Fig. 7). This triangle is in contrast to the alternative outcome of a linear relationship between the trait and climatic variables (seen in Cornelissen 1999, Guo et al. 1998), and is broadly consistent with Wright et al.'s (2017) observation that maximum leaf size is associated with rainfall. This likely suggests that larger leaves are excluded from dry environments, but in wet environments smaller leaves are not necessarily disadvantaged. In addition, we note that the machine learning approach used a more comprehensive trait sampling method for leaf selection, as opposed to the traditional approach by physiologists

(Pérez-Harguindeguy et al. 2013). As a result of this, the greater representation of juvenile leaves may have potentially contributed to the triangular shape of the association. However, despite this difference, our analyses support a similar conclusion to Wright et al. (2017). This is further reinforced by the quantitative agreement of our trait-climate relationships to that of other databases, especially at higher quantiles (Fig. 7). Further implications of this sampling method are explored later. Overall, our analysis confirms the association between leaf traits and climate, and the novel workflow and sampling approach offer potentially new perspectives on these relationships.

Our study also revealed the link between traits and climate from both a macroevolutionary and microevolutionary scale. This corresponds to a recent review of Anderegg (2023), which stresses the importance of trait-climate analyses that aim to improve our understanding of the influences of physiology and evolution across different scales. Our dataset's unique characteristic of vast intra- and interspecific sampling, paired with the availability of a fully resolved phylogeny (Thornhill et al. 2019), made it possible to examine evolutionary processes at both of these scales. In particular, the association between leaf area and mean annual precipitation at the broadest scale in our study was not on average replicated within eucalypt species (Fig. 8), consistent with recent observations in *Syzygium* and *Ficus* (Wilde et al. 2023). This raised the question of where, from the deepest to the shallowest evolutionary scales, does the association between leaf area and precipitation weaken? This change in association occurs rather abruptly in analyses within young lineages of approximately 8.5 million years of age, indicating that the absence of association between leaf area and precipitation is mostly confined to the intraspecific analyses (Fig. 8). This observation was consistent with findings in other taxa, which suggest that community-level relationships are predominantly driven by weak intraspecific relationships (Mudrák et al. 2019, McDonald et al. 2003, Ackerly et al. 2002, Guo et al. 2000). This validates the notion that the effects of gene flow in the homogenising of traits, reduces the capacity to adapt locally to climate (Alexander et al. 2022, Leimu & Fischer 2008, Kirkpatrick & Barton 1997). As such, the hypothesis proposed in Figure 1 Scenario 2 is supported, as trait-climate relationships with similar slopes to the whole eucalypt clade is observed within groups of samples that include recently diverged lineages. These groups presumably have much less gene flow between populations in contrasting climate conditions (Fig. 8c). This is unlike Scenario 1 in which each lineage, including single species, reflect the overall trend. It is also unlike Scenario 3,

where no lineages are locally adapted, and the association between trait and climate only manifests among more deeply diverged groups. Understanding the phylogenetic constraint in a trait is critical to improve the conclusions of the numerous trait-environment studies that have been carried without a phylogenetic framework. This highlights the importance of studying a range of taxa to determine general ecological trends, and the need for more datasets with wide scopes across time, space, and phylogeny.

Our dataset was generated using a novel approach for trait measurement and sampling from herbarium specimens. This had several important consequences for downstream analyses of the data. First, the great size of the dataset was expected to provide a robust buffer against uncontrollable stochastic variation that arises when working with herbarium sheets, as recognised in other studies (Goëau et al. 2020, Willis et al. 2017). These included trait and spatial biases in the biological sampling, shrinkage effects of dried leaf material, size limitations of herbarium sheets, and innate errors in measurements (Heberling 2022, Daru et al. 2018) (see Supplementary Information D for analyses for further error validation). Second, we observed that for several species, our approach resulted in leaf area data with greater numbers of smaller leaves, relative to other datasets (Fig 5b). Our model's high level of precision suggests this was likely due to differences in trait sampling methods, rather than measurement errors. In particular, this may be attributed to our leaf masking model being trained to measure all leaves of an herbarium specimen, whereas conventional plant trait ecology protocols target fully expanded leaves (Pérez-Harguindeguy et al. 2013). The use of quantile regressions supported this idea. We observed a convergence of our slopes to other datasets that employed traditional sampling methods, in the higher quantiles of our quantile regression analyses (Fig. 7) and by temporarily filtering out the bottom 50% of results per species (results in Supplementary Information C, suggested by Corney et al. 2012). These analyses explain the quantitative discrepancies between results from our dataset and others and suggest a useful approach for downstream analyses of data generated in this way. It would be worthwhile for future studies to isolate the degree of influence of ecological constraints and ontogeny on these triangular associations. In conclusion, our approach can be perceived as a feasible prototype that can be extended and modified to create large datasets in different taxa and traits.

More generally, the rapid development of technology has opened up a new avenue of information extraction, facilitating the gathering of large volumes of relatively unfiltered data. This study highlights what we predict will be a recurring theme in the use of ML. Contrary to traditional collections that often use high levels of selectivity during data collection (Pérez-Harguindeguy et al. 2013), new collections using ML approaches will have little selection as the data source is typically untailored to the workflow. As a result, we predict that ML generated data may not always be used interchangeably with data collected by traditional methods. We recommend careful validation prior to use, and the adoption of clear definitions in databases (e.g., TRY (Kattge et al. 2020), AusTraits (Falster et al. 2021)) that will potentially include records generated by both traditional methods and approaches based on ML.

In summary, our workflow has linked three key factors: plant traits, climate, and the effects of evolutionary depth. As one of the first operational studies of ML in trait ecology, our workflow represents an exciting advancement. Here, it examined how leaf traits in eucalypts shifted across precipitation and temperature and found associations that confirmed relevant global analyses (Wright et al. 2017, Moles et al. 2014). Our study also extends our understanding of these relationships, suggesting they are underpinned by turnover among species across environments, including recently diverged species, but with little evidence of adaptation to climate among populations still connected by gene flow. Given eucalypt's uniquely low levels of genetic differentiation and high gene flow across geographically distant populations (Jordan et al. 2023, Fahey et al. 2022, Supple et al. 2018), a valuable future development would involve exploring the generality of these observations in other major taxa using the abundance of data available in herbaria. This will allow researchers to create datasets that span different patterns of population genetic variation, as well as wide phylogenetic scopes and multiple traits. In turn, with these new datasets, we may reveal a widespread phenomenon of intraspecific variation within trait-environment correlations similar to ours, across various taxa.

Acknowledgements

We thank the individuals at the University of New South Wales and the Research Centre for Ecosystem Resilience for their support and assistance needed to complete this project, both in funding and the software/hardware used.

Funding

This project was supported from the funding of the University of New South Wales, and with the technology provided by the Research Centre for Ecosystem Resilience of the Royal Botanic Gardens of Sydney.

Author Contributions

KG, JGB and WKC conceived and designed the study. JGB contributed the initial structure of the machine learning model, and KG refined the model. KG wrote the manuscript, JGB and WKC edited, and all authors approved of the final manuscript.

References

Ackerly, D., Knight, C., Weiss, S., Barton, K., & Starmer, K. (2002). Leaf size, specific leaf area and microhabitat distribution of chaparral woody plants: contrasting patterns in species level and community level analyses. *Oecologia*, *130*(3), 449–457. https://doi.org/10.1007/s004420100805

Ackerly, D. D., & Cornwell, W. K. (2007). A trait-based approach to community assembly: Partitioning of species trait values into within- and among-community components. *Ecology Letters*, *10*(2). https://doi.org/10.1111/j.1461-0248.2006.01006.x

Ackerly, D. D., & Reich, P. B. (1999). Convergence and correlations among leaf size and function in seed plants: a comparative test using independent contrasts. *American Journal of Botany*, *86*(9), 1272–1281. https://doi.org/10.2307/2656775

Alexander, J. M., Atwater, D. Z., Colautti, R. I., & Hargreaves, A. L. (2022). Effects of species interactions on the potential for evolution at species' range limits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1848). https://doi.org/10.1098/rstb.2021.0020

An, N., Lu, N., Fu, B., Wang, M., & He, N. (2021). Distinct Responses of Leaf Traits to Environment and Phylogeny Between Herbaceous and Woody Angiosperm Species in China. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.799401

Anderegg, L. D. L. (2023). Why can't we predict traits from the environment? *New Phytologist*, *237*(6), 1998–2004. https://doi.org/10.1111/nph.18586

Bastias, C. C., Fortunel, C., Valladares, F., Baraloto, C., Benavides, R., Cornwell, W., Markesteijn, L., de Oliveira, A. A., Sansevero, J. B. B., Vaz, M. C., & Kraft, N. J. B. (2017). Intraspecific leaf trait variability along a boreal-to-tropical community diversity gradient. *PLOS ONE*, *12*(2), e0172495. https://doi.org/10.1371/journal.pone.0172495

Battaglia, M., Cherry, M. L., Beadle, C. L., Sands, P. J., & Hingston, A. (1998). Prediction of leaf area index in eucalypt plantations: effects of water stress and temperature. *Tree Physiology*, *18*(8–9), 521–528. https://doi.org/10.1093/treephys/18.8-9.521

Booth, T. H., Broadhurst, L. M., Pinkard, E., Prober, S. M., Dillon, S. K., Bush, D., Pinyopusarerk, K., Doran, J. C., Ivkovich, M., & Young, A. G. (2015). Native forests and climate change: Lessons from eucalypts. *Forest Ecology and Management*, *347*, 18–29. https://doi.org/10.1016/j.foreco.2015.03.002

Brenskelle, L., Guralnick, R. P., Denslow, M., & Stucky, B. J. (2020). Maximizing human effort for analyzing scientific images: A case study using digitized herbarium sheets. *Applications in Plant Sciences*, *8*(6), e11370. https://doi.org/10.1002/aps3.11370

Cornelissen, J. H. C. (1999). A triangular relationship between leaf size and seed size among woody species: allometry, ontogeny, ecology and taxonomy. *Oecologia*, *118*(2), 248–255. https://doi.org/10.1007/s004420050725

Corney, D. P. A., Clark, J. Y., Tang, H. L., & Wilkin, P. (2012). Automatic extraction of leaf characters from herbarium specimens. *TAXON*, *61*(1), 231–244. https://doi.org/10.1002/tax.611016

Cutts, V., Hanz, D. M., Barajas-Barbosa, M. P., Algar, A. C., Steinbauer, M. J., Irl, S. D. H., Kreft, H., Weigelt, P., Fernandez Palacios, J. M., & Field, R. (2021). Scientific floras can be reliable sources for some trait data in a system with poor coverage in global trait databases. *Journal of Vegetation Science*, *32*(3). https://doi.org/10.1111/jvs.12996

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M., & Davis, C. C. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, *217*(2), 939–955. https://doi.org/10.1111/nph.14855

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Ellis, T. W., & Hatton, T. J. (2008). Relating leaf area index of natural eucalypt vegetation to climate variables in southern Australia. *Agricultural Water Management*, *95*(6), 743–747. https://doi.org/10.1016/j.agwat.2008.02.007

Fahey, P. S., Udovicic, F., Cantrill, D. J., Nicolle, D., McLay, T. G. B., & Bayly, M. J. (2022). A phylogenetic investigation of the taxonomically problematic. *Australian Systematic Botany*, *35*(5), 403–435. https://doi.org/10.1071/SB21029

Falster, D., Gallagher, R., Wenk, E. H., Wright, I. J., Indiarto, D., Andrew, S. C., Baxter, C., Lawson, J., Allen, S., Fuchs, A., Monro, A., Kar, F., Adams, M. A., Ahrens, C. W., Alfonzetti, M., Angevin, T., Apgaua, D. M. G., Arndt, S., Atkin, O. K., … Ziemińska, K. (2021). AusTraits, a curated plant trait database for the Australian flora. *Scientific Data*, *8*(1), 254. https://doi.org/10.1038/s41597-021-01006-6

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Goëau, H., Mora-Fallas, A., Champ, J., Love, N. L. R., Mazer, S. J., Mata-Montero, E., Joly, A., & Bonnet, P. (2020). A new fine-grained method for automated visual analysis of herbarium specimens: A case study for phenological data extraction. *Applications in Plant Sciences*, *8*(6). https://doi.org/10.1002/aps3.11368

Gombin, J., Vaidyanathan, R., & Agafonkin, V. (2020). *concaveman: A Very Fast 2D Concave Hull Algorithm*. https://cran.r-project.org/package=concaveman

Govindan, M. (2005). Eucalyptus: the Genus Eucalyptus. In J. J. Coppen (Ed.), *Journal of Natural Products*. Natural Resources Institute, University of Greenwich, UK. https://doi.org/10.1021/np0307789

Guo, Q., Brown, J. H., & Enquist, B. J. (1998). Using Constraint Lines to Characterize Plant Performance. *Oikos*, *83*(2), 237. https://doi.org/10.2307/3546835

Guo, Q., Brown, J. H., Valone, T. J., & Kachman, S. D. (2000). Constraints of Seed Size on Plant Distribution and Abundance. *Ecology*, *81*(8), 2149. https://doi.org/10.2307/177103

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. http://arxiv.org/abs/1512.03385

Heberling, J. M. (2022). Herbaria as Big Data Sources of Plant Traits. *International Journal of Plant Sciences*, *183*(2), 87–118. https://doi.org/10.1086/717623

Hijmans, R. J. (2023). *raster: Geographic Data Analysis and Modeling*. https://cran.r-project.org/package=raster

Hill, K. E., Hill, R. S., & Watling, J. R. (2014). Do CO2, temperature, rainfall and elevation influence stomatal traits and leaf width in Melaleuca lanceolata across southern Australia? *Australian Journal of Botany*, *62*(8), 666. https://doi.org/10.1071/BT14300

Hussein, B. R., Malik, O. A., Ong, W. H., & Slik, J. W. F. (2021). Automated extraction of phenotypic leaf traits of individual intact herbarium leaves from herbarium specimen

images using deep learning based semantic segmentation. *Sensors*, *21*(13). https://doi.org/10.3390/s21134549

Jordan, R., Prober, S. M., Andrew, R., Freeman, J., Kerr, R., Steane, D., Vaillancourt, R., & Potts, B. (2023). *Population Genomics of Eucalypts*. https://doi.org/10.1007/13836_2023_107

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Tautenhahn, S., Werner, G. D. A., Aakala, T., Abedi, M., Acosta, A. T. R., Adamidis, G. C., Adamson, K., Aiba, M., Albert, C. H., Alcántara, J. M., Alcázar C, C., Aleixo, I., Ali, H., … Wirth, C. (2020). TRY plant trait database – enhanced coverage and open access. *Global Change Biology*, *26*(1), 119–188. https://doi.org/10.1111/gcb.14904

Kirkpatrick, M., & Barton, N. H. (1997). Evolution of a Species' Range. *The American Naturalist*, *150*(1), 1–23. https://doi.org/10.1086/286054

Kozlov, M. V, Sokolova, I. V, Zverev, V., & Zvereva, E. L. (2021). Changes in plant collection practices from the 16th to 21st centuries: implications for the use of herbarium specimens in global change research. *Annals of Botany*, *127*(7), 865–873. https://doi.org/10.1093/aob/mcab016

Larsson, J. (2020). *{polylabelr}: find the pole of inaccessibility (visual center) of a polygon*. https://cran.r-project.org/package=polylabelr

Leigh, A., Sevanto, S., Close, J. D., & Nicotra, A. B. (2017). The influence of leaf size and shape on leaf thermal dynamics: does theory hold up under natural conditions? *Plant Cell and Environment*, *40*(2), 237–248. https://doi.org/10.1111/pce.12857

Leimu, R., & Fischer, M. (2008). A Meta-Analysis of Local Adaptation in Plants. *PLoS ONE*, *3*(12), e4010. https://doi.org/10.1371/journal.pone.0004010

Li, Y., Zou, D., Shrestha, N., Xu, X., Wang, Q., Jia, W., & Wang, Z. (2020). Spatiotemporal variation in leaf size and shape in response to climate. *Journal of Plant Ecology*, *13*(1), 87–96. https://doi.org/10.1093/jpe/rtz053

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). *Microsoft COCO: Common Objects in Context* (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48

Madani, N., Kimball, J. S., Ballantyne, A. P., Affleck, D. L. R., van Bodegom, P. M., Reich, P. B., Kattge, J., Sala, A., Nazeri, M., Jones, M. O., Zhao, M., & Running, S. W. (2018). Future global productivity will be affected by plant trait response to climate. *Scientific Reports*, *8*(1), 2870. https://doi.org/10.1038/s41598-018-21172-9

McDonald, P. G., Fonseca, C. R., Overton, J. M., & Westoby, M. (2003). Leaf-size divergence along rainfall and soil-nutrient gradients: is the method of size reduction common among clades? *Functional Ecology*, *17*(1), 50–57. https://doi.org/10.1046/j.1365-2435.2003.00698.x

Meier, I. C., & Leuschner, C. (2008). Leaf Size and Leaf Area Index in Fagus sylvatica Forests: Competing Effects of Precipitation, Temperature, and Nitrogen Availability. *Ecosystems*, *11*(5), 655–669. https://doi.org/10.1007/s10021-008-9135-2

Milla, R., & Reich, P. B. (2011). Multi-trait interactions, not phylogeny, fine-tune leaf size reduction with increasing altitude. *Annals of Botany*, *107*(3), 455–465. https://doi.org/10.1093/aob/mcq261

Moles, A. T., Perkins, S. E., Laffan, S. W., Flores-Moreno, H., Awasthy, M., Tindall, M. L., Sack, L., Pitman, A., Kattge, J., Aarssen, L. W., Anand, M., Bahn, M., Blonder, B., Cavender-Bares, J., Cornelissen, J. H. C., Cornwell, W. K., Díaz, S., Dickie, J. B., Freschet, G. T., … Bonser, S. P. (2014). Which is a better predictor of plant traits: temperature or precipitation? *Journal of Vegetation Science*, *25*(5), 1167–1180. https://doi.org/10.1111/jvs.12190

Moran, E. V., Hartig, F., & Bell, D. M. (2016). Intraspecific trait variation across scales: implications for understanding global change responses. *Global Change Biology*, *22*(1), 137–150. https://doi.org/10.1111/gcb.13000

Morgan, R., Martín-Forés, I., Leitch, E., & Guerin, G. (2021). Assessment of protocols and variance for specific leaf area (SLA) in 10 Eucalyptus species to inform functional trait

sampling strategies for TERN Ausplots. *Dataset, The University of Adelaide*. https://doi.org/https://doi.org/10.25909/14197298

Mudrák, O., Doležal, J., Vítová, A., & Lepš, J. (2019). Variation in plant functional traits is best explained by the species identity: Stability of trait-based species ranking across meadow management regimes. *Functional Ecology*, *33*(4), 746–755. https://doi.org/10.1111/1365-2435.13287

Nicolle, D. (2022). *Classification of the eucalypts (Angophora, Corymbia and Eucalyptus) Version 6*. http://www.dn.com.au/Classification-Of-The-Eucalypts.pdf

Nobel, P. S. (2009). Physicochemical and Environmental Plant Physiology. In *Physicochemical and Environmental Plant Physiology, Fourth Edition*. Academic Press. https://doi.org/10.1016/B978-0-12-374143-1.X0001-4

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pearson, K. D., Nelson, G., Aronson, M. F. J., Bonnet, P., Brenskelle, L., Davis, C. C., Denny, E. G., Ellwood, E. R., Goëau, H., Heberling, J. M., Joly, A., Lorieul, T., Mazer, S. J., Meineke, E. K., Stucky, B. J., Sweeney, P., White, A. E., & Soltis, P. S. (2020). Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. *BioScience*, *70*(7), 610–620. https://doi.org/10.1093/biosci/biaa044

Peppe, D. J., Royer, D. L., Cariglino, B., Oliver, S. Y., Newman, S., Leight, E., Enikolopov, G., Fernandez-Burgos, M., Herrera, F., Adams, J. M., Correa, E., Currano, E. D., Erickson, J. M., Hinojosa, L. F., Hoganson, J. W., Iglesias, A., Jaramillo, C. A., Johnson, K. R., Jordan, G. J., … Wright, I. J. (2011). Sensitivity of leaf size and shape to climate: global patterns and paleoclimatic applications. *New Phytologist*, *190*(3), 724–739. https://doi.org/10.1111/j.1469-8137.2010.03615.x

Pérez-Harguindeguy, N., Díaz, S., Garnier, E., Lavorel, S., Poorter, H., Jaureguiberry, P., Bret-Harte, M. S., Cornwell, W. K., Craine, J. M., Gurvich, D. E., Urcelay, C., Veneklaas, E. J., Reich, P. B., Poorter, L., Wright, I. J., Ray, P., Enrico, L., Pausas, J. G., de Vos, A. C., … Cornelissen, J. H. C. (2013). New handbook for standardised measurement of plant functional traits worldwide. *Australian Journal of Botany*, *61*(3), 167. https://doi.org/10.1071/BT12225

Pritzkow, C., Szota, C., Williamson, V. G., & Arndt, S. K. (2020). Phenotypic Plasticity of Drought Tolerance Traits in a Widespread Eucalypt (Eucalyptus obliqua). *Forests*, *11*(12), 1371. https://doi.org/10.3390/f11121371

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Reich, P. B., Wright, I. J., Cavender-Bares, J., Craine, J. M., Oleksyn, J., Westoby, M., & Walters, M. B. (2003). The Evolution of Plant Functional Variation: Traits, Spectra, and Strategies. *International Journal of Plant Sciences*, *164*(S3), S143–S164. https://doi.org/10.1086/374368

Reich, P. B. (2012). Key canopy traits drive forest productivity. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1736), 2128–2134. https://doi.org/10.1098/rspb.2011.2270

Revell, L. J. (2012). phytools: An {R} package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Slee, A., Brooker, M., Duffy, S., & West, J. (2020). *EUCLID: Eucalypts of Australia, Fourth Edition*. Centre for Australian National Biodiversity Research. https://apps.lucidcentral.org/euclid

Souza, M. L., Duarte, A. A., Lovato, M. B., Fagundes, M., Valladares, F., & Lemos-Filho, J. P. (2018). Climatic factors shaping intraspecific leaf trait variation of a neotropical tree along a rainfall gradient. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0208512

Supple, M. A., Bragg, J. G., Broadhurst, L. M., Nicotra, A. B., Byrne, M., Andrew, R. L., Widdup, A., Aitken, N. C., & Borevitz, J. O. (2018). Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change. *ELife*, *7*. https://doi.org/10.7554/eLife.31835

Thornhill, A. H., Crisp, M. D., Külheim, C., Lam, K. E., Nelson, L. A., Yeates, D. K., & Miller, J. T. (2019). A dated molecular perspective of eucalypt taxonomy, evolution and diversification. *Australian Systematic Botany*, *13*(1), 149. https://doi.org/10.1071/SB18015

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, *116*(5), 882–892. https://doi.org/10.1111/j.0030-1299.2007.15559.x

Wada, K. (2022). *Labelme: Image Polygonal Annotation with Python*. https://github.com/wkentaro/labelme

Wang, C., He, J., Zhao, T.-H., Cao, Y., Wang, G., Sun, B., Yan, X., Guo, W., & Li, M.-H. (2019). The Smaller the Leaf Is, the Faster the Leaf Water Loses in a Temperate Forest. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00058

Wang, H., Wang, R., Harrison, S. P., & Prentice, I. C. (2022). Leaf morphological traits as adaptations to multiple climate gradients. *Journal of Ecology*. https://doi.org/10.1111/1365-2745.13873

Weaver, W. N., Ng, J., & Laport, R. G. (2020). LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Applications in Plant Sciences*, *8*(6). https://doi.org/10.1002/aps3.11367

Westerband, A. C., Funk, J. L., & Barton, K. E. (2021). Intraspecific trait variation in plants: a renewed focus on its role in ecological processes. In *Annals of botany* (Vol. 127, Issue 4, pp. 397–410). NLM (Medline). https://doi.org/10.1093/aob/mcab011

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019).

Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilde, B.C., Bragg, J. G. & Cornwell, W. (2023), Trait-climate relationships within and among taxa using machine learning and herbarium specimens. *American Journal of Botany*. *In Press*. https://doi.org/10.1002/ajb2.16167

Willis, C. G., Ellwood, E. R., Primack, R. B., Davis, C. C., Pearson, K. D., Gallinat, A. S., Yost, J. M., Nelson, G., Mazer, S. J., Rossington, N. L., Sparks, T. H., & Soltis, P. S. (2017). Old Plants, New Tricks: Phenological Research Using Herbarium Specimens. *Trends in Ecology & Evolution*, *32*(7), 531–546. https://doi.org/10.1016/j.tree.2017.03.015

Wright, I. J., Dong, N., Maire, V., Prentice, I. C., Westoby, M., Díaz, S., Gallagher, R. V., Jacobs, B. F., Kooyman, R., Law, E. A., Leishman, M. R., Niinemets, Ü., Reich, P. B., Sack, L., Villar, R., Wang, H., & Wilf, P. (2017). Global climatic drivers of leaf size. *Science*, *357*(6354), 917–921. https://doi.org/10.1126/science.aal4760

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2

Younis, S., Weiland, C., Hoehndorf, R., Dressler, S., Hickler, T., Seeger, B., & Schmidt, M. (2018). Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. *Botany Letters*, *165*(3–4), 377–383. https://doi.org/10.1080/23818107.2018.144635

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svanteson, S., Wengstrom, N., Zizka, V., & Antonelli, A. (2019). CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*, 7. https://doi.org/10.1111/2041-210X.13152

Supplementary Information

The following Supporting Information is available for this article:

Supplementary Information A - Model Creation

Table SA_1. *Iterations carried out in the optimisation cycle to generate the final leaf masking model.*

Figure SA_2. *Examples of classifier's invalid leaves*

Table SA_3. *Iterations carried out in the optimisation cycle to generate the final leaf classification model.*

Supplementary Information B - Leaf masking model's labelling protocol

Figure SB_1. *An example of a manually annotated herbarium sheet.*

Supplementary Information C - Extra Analyses

Leaf area x Temperature

Figure SC_1. *Quantile regression analysis model.*

Table SC_2. *Coefficients for the overall linear model, and the different levels of regression quantiles.*

Table SC_3. *Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within, and Wright et al.'s (2017) results.*

Figure SC_4. *Phylogenetic analysis split at 20 intervals, and by taxonomic levels.*

Reflecting Physiologist Sampling - Leaf area

Figure SC_5. *Relationships between the climatic variables (log mean annual precipitation and temperature) against log leaf area.*

Table SC_6. *Coefficients of models for log leaf area against log mean annual precipitation and mean annual temperature in comparison to other datasets.*

Reflecting Physiologist Sampling - Leaf area x Precipitation

Figure SC_7. *Quantile regression analysis model results.*

Table SC_8. *Coefficients for the overall linear model, and the different levels of regression quantiles.*

Table SC_9. *Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each*

herbarium sheet nested within, and Wright et al.'s (2017) results.

Reflecting Physiologist Sampling - Leaf area x Temperature

Figure SC_10. *Quantile regression analysis model results.*

Table SC_11. *Coefficients for the overall linear model, and the different levels of regression quantiles.*

Table SC_12. *Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within, and Wright et al.'s (2017) results.*

Largest in-circle area

Figure SC_13. *Comparison between leaf area to largest in-circle area.*

Table SC_14. *Phylogenetic signals for log largest in-circle area.*

Largest in-circle area x Precipitation

Figure SC_15. *Quantile regression analysis model results.*

Table SC_16. *Coefficients for the overall linear model, and the different levels of regression quantiles.*

Table SC_17. *Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

Figure SC_18. *Phylogenetic analysis split at 20 intervals, and by taxonomic levels.*

Largest in-circle area x Temperature

Figure SC_19. *Quantile regression analysis model results.*

Table SC_20. *Coefficients for the overall linear model, and the different levels of regression quantiles.*

Table SC_21. *Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

Figure SC_22. *Phylogenetic analysis split at 20 intervals, and by taxonomic levels.*

Leaf curvature

Figure SC_22. *Relationships between the climatic variables against leaf curvature.*

Table SC_23. *Phylogenetic signals for leaf curvature*

Leaf curvature x Precipitation

Figure SC_27. *Quantile regression analysis model results.*

Table SC_28. *Coefficients for the overall linear model, and the different levels of regression quantiles.*

Figure SC_29. *Phylogenetic analysis split at 20 intervals, and by taxonomic levels.*

Table SC_30. *Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

Leaf curvature x Temperature

Figure SC_31*. Quantile regression analysis model results.*

Table SC_32*. Coefficients for the overall linear model, and the different levels of regression quantiles.*

Figure SC_33*. Phylogenetic analysis split at 20 intervals, and by taxonomic levels.*

Table SC_34*. Coefficients following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

Supplementary Information D - Error Validation

Figure SD_1*. Residual plot of log leaf area and log mean annual precipitation.*

Figure SD_2*. Plotting the normalised frequency count across the different leaf areas.*

Figure SD_3*. Comparing the mean leaf area across databases, where each data point is a species.*

Supplementary Information A - Model creation

Training, validating, and testing our models

Determining datasets for both models

Three key datasets were created for training, validation, and testing. Sampling for all datasets was done by first separating the whole image dataset into their different genera and for *Eucalyptus*, dividing further by subgenera. Separation of species into their taxonomic grouping first followed Nicolle 2022, then Slee et al. 2020 then Thornhill et al. 2019. Hybrid *Eucalyptus* specimens were placed into a 'Hybrid' subgenus, with a number of *Eucalyptus* species left as 'NA' subgenus if no data could be located. Random sampling using the function slice_sample from tidyverse (v 2.0.0, Wickham et al. 2019) was then carried out within these groups (genera and subgenera), with the number of sheets reflective of the size of groups. This method allowed a vast representation of different forms of eucalypts.

Leaf masking model

Model

The leaf masking model used a ResNet50 architecture (He et al. 2015) and was implemented via Detectron2 (Wu et al. 2019). ResNet is a deep convolutional neural network developed explicitly for image classification tasks, and Detectron2 is an open-source machine learning library developed by Facebook's AI Research team.

ResNet50 is constructed of 50 layers - 48 convolutional layers, 1 MaxPool layer and 1 average pool layer (He et al. 2015 for a detailed description of ResNet's architecture). Each convolutional layer undergoes a batch normalisation to reduce overfitting and improve generalisation. ResNet50 was selected due to 1) ResNet's focus on image detection and 2) the number of layers were selected to balance between the task's complexity and limiting overfitting

Manual annotation of datasets

The labelling of data for training, manual annotation, used the graphic program LabelMe (v 5.01, Wada 2022), and followed the protocol outlined in SI. The use of a bounding box here was suggested by preliminary trials and allowed us to create a pseudo-image of a whole sheet. This in turn enabled a greater range of different leaf types to be used for training. This dataset was later supplemented with manually annotated full sheets, as suggested by improved performances during the cycles of model optimisation undertaken.

Training

Optimisation of the model was carried out to determine the final selection of the training parameters that gave the best performance in terms of prediction and testing. Different training parameters allow altering a model's training. These include the model's base learning rate, max iterations, batch size, and the number of classes, and are defined in Table 1.

The overall steps for optimisation were as followed: (i) Train the initial model using the manually annotated training and validation data set, (ii) Predict the leaves onto the testing data set, (iii) Gather quantitative and qualitative measures of model accuracy from part (ii), (iv) alter the training parameters and repeat the cycle at part (i) with the new model.

Part ii and iii of the optimisation cycle involved carrying out a testing process. This included using the current iteration's model to predict onto the 20 full sheets in the testing dataset. From these predictions, we noted i) the area of the predicted bitmask and how it compared to the ground-truth mask, ii) the number of correct predictions made, iii) the number of incorrect predictions made, iv) a visual check for biases. These were then used to generate evaluation metrics standardised in this field, and include Intersection Over Union (IoU), precision, recall and the harmonic mean of precision and recall (F1-score), defined in Table 1. Using these metrics, we repeated the process of optimisation to improve the model quality.

Iterations:

Our initial model was trained on 7 different classes including Leaf100, Leaf100B, Leaf100UM, Leaf90, Leaf90UM, Leaf50, Leaf50UM, with each class representing a leaf of different coverage and age. Definitions of these categories are found in the protocol. Through the iterations of the optimisation process, we reduced the number of classes to just one, where it joined Leaf90, Leaf100, Leaf100B, Leaf100UM and Leaf90UM labels, and excluded Leaf50 and Leaf50UM. This selection was done based on a balance between data accuracy and volume of leaves detected. The merging of Leaf90 and Leaf100 leaves were executed as the accuracy of the predicted masks had an innate 10% error. Thus, merging allowed a significantly increased number of leaves detected with what we saw as an acceptable rate of error, especially in light of the total volume of leaves detected. As we progressed through the iterations, we increased the training data until we were satisfied with our model's performance.

Table SA_1. *Iterations carried out in the optimisation cycle to generate the final leaf masking model.*

Test Number	Batch Size	Learning Rate	Number of iterations	Number of classes	Number of sheets	Evaluation Metrics					Process
						IoU	Precision	Recall	F_1 Score	Visual Notes	
1	12	0.0001	8000	8 Leaf100, Leaf100B, Leaf90, Leaf50, Leaf100UM, Leaf100BUM, Leaf90UM, Leaf50UM	Training: 43 Validation: 20	Not conducted				High visual IoU however assigned categories incorrectly, high proportion of labels assigned as L100UM even if L50 (increases error)	Decrease number of categories
2	8	0.0001	8000	4 Leaf100, Leaf100B, Leaf90, Leaf50 Where UM classes were merged into their respective categories	Training: 43 Validation: 20 Testing: 20	0.43	0.92**	0.31	0.46	Similar error, where categories were not correctly assigned.	Decrease number of categories

3	8	0.0001	8000	3 Leaf100, Leaf90, Leaf50 Where Leaf100B was merged into Leaf100, and UM classes were removed	Training: 43 Validation: 20 Testing: 20	0.43	0.96**	0.32	0.48	Similar error, where categories were not correctly assigned. Increased number of leaves detected but more visual errors in incorrect masks	Removal of 'UM' classes
4	8	0.0001	8000	3 Leaf100, Leaf90, Leaf50	Training: 43 Validation: 20 Testing: 20	0.87	0.87**	0.29	0.44	Similar error, where categories were not correctly assigned.	Change batch size to see difference
5	15	0.0001	8000	3 Leaf100, Leaf90, Leaf50	Training: 43 Validation: 20 Testing: 20	0.44	0.93**	0.33	0.49	Increased number of leaves detected. Large number of L50 detected were incorrectly labelled as L100/90.	Remove label L50. Decision to accept both L90 and L100 as valid results as within margin of error.

6	15	0.0001	8000	2 Leaf100, Leaf90	Training: 43 Validation: 20 Testing: 20	0.71	0.75	0.46	0.57	Broken and highly overlapping leaves were often wrongly masked	Add extra training sheets to improve detection
7	20	0.0001	8000	2 Leaf100, Leaf90	Training: 96* Validation: 20 Testing: 20	0.68	0.72	0.70	0.71	Reduced number of leaves detected. Less false positives detected	
8	20	0.0001	8000	1 Leaf100	Training: 96 Validation: 20 Testing: 20	0.68	0.19	0.23	0.21	Removing L90 significantly reduced detection rate	Reintroduction of L90
9	20	0.0001	8000	1 Where Leaf90 was merged into Leaf100	Training: 96 Validation: 20 Testing: 20	0.63	0.76	0.71	0.68	Merged category increased detection	Increasing training dataset

10	20	0.0001	8000	1 Where Leaf90 was merged into Leaf100	Training: 113* Validation: 28 Testing: 20	0.64	0.78	0.68	0.73	Reduced detection of half leaves	

* Included training data that were full sheets, instead of sheets restricted by a bounding box

** Precision was not calculated with categories, only whether the predicted leaf mask matched a ground-truth mask. As such, the high precision was a result of the inclusion of the L50 category. This meant most leaves were true positives. However, due to the incorrect assignment of categories, precision was not a reliable indication of model quality

Leaf classification

The leaf classification model was used to separate the predicted masks of the previous model into valid and invalid leaves. It used a ResNet50 architecture (He et al. 2015), implemented in PyTorch (Paszke et al. 2019) and pretrained on ImageNet data (Deng et al. 2009). PyTorch was used to build our models. It was developed by Facebook's AI research group and was selected from a balance of its ease of use and quality of output.

Manual annotation of datasets

The datasets were manually classified into valid and invalid leaves. The final iteration's criteria of valid leaves were classed according to the criteria below:

i) Leaves with the base or tip of the leaf were completely visible and not overlapped by an object,
ii) Less than 5% of the leaf mask was missing from the true leaf,
iii) Broken tips or folded sections (<5% of true leaf) was acceptable if they were rounded,
iv) Edge divots were acceptable if radius was <50% of the shortest distance from the edge to the midrib (<5% of the total volume),
v) Warped leaves (i.e. due to a gall) with rounded edge were accepted,
vi) More than 5% of the total volume exceeding the leaf edge is not acceptable, especially if it is a prominently protruding mask.

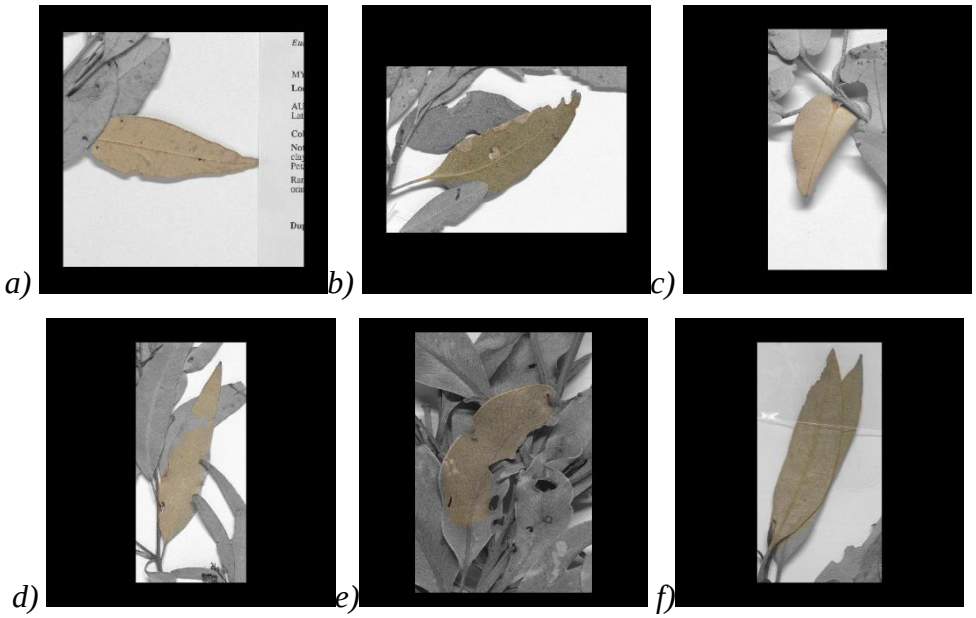

Figure SA_2. *Examples of invalid leaves. a) Example of i. b) Example of ii. c) Example of iii. d) Example of iv. e) Example of v. f) Example of vi.*

Training

This model was trained and validated on leaves that used the leaf masking model described above on a separate set of herbarium sheets, selected using the method detailed prior. To create the leaf masks for the training and validating dataset, we first carried out a connected component analysis (Otsu thresholding with a connection level of 4) to remove pixels disconnected from the main leaf mask. The herbarium sheet images were then cropped to the area of the predicted mask and its colour converted to indicate the predicted mask (coloured) and background (greyscale) (Fig. SA_2). These leaves were then manually annotated and separated into valid and invalid leaves. Once the valid and invalid datasets were balanced to a similar number of data points, they were fed into the model to train the classification model. The trained model was then tested on the same testing dataset as the leaf masking model, allowing us to see the change in evaluation metrics over both processes.

Table SA_3. *Iterations carried out in the optimisation cycle to generate the final leaf classification model.*

Test Number	Number of iterations	Number of leaf masks for training and validation	Evaluation Metrics				Notes	Process
			Average accuracy of last epoch	Precision	Recall	F_1 Score		
Model 1) Criteria accepting only L100	42	Y:151 N:155	0.64	0.68	0.9	0.78	Model only included L100 as valid leaves. Resulted in a high precision but excluded a large number of leaves from the dataset	Include Both L90 and L100 as 'valid' leaves
Model 2) Criteria accepting both L100 and L90	42	Y:325 N:321	0.64	0.59	0.77	0.67	Model included both L100 and L90 as valid leaves	
Model 3) Criteria accepting both L100 and L90 Used less data to compare against Model 1)	42	Y:151 N:157	0.68	0.24	0.57	0.34	Model same as above, but included roughly the same amount of training data to see the impact on evaluation criteria	Model was deemed worse than L100 but may have been due to the larger variety in valid leaves
Model 4.1)	42	Y:221 N:221	0.63	0.28	0.52	0.36		

Criteria accepting only L100 More training data was used Same number of valid and invalid leaves								
Model 4.2) Criteria accepting only L100 More invalid leaves than valid	42	Y:221 N:330	0.7	0.33	0.81	0.47	Model same as above, but included more invalid leaves in the training dataset than valid leaves. Resulted in higher recall	Rebalance the amount of training data in both categories
Model 4.3) Criteria accepting only L100 Same number of valid and invalid leaves Increase number of training epochs	63	Y:221 N:221	0.76	0.22	0.76	0.34		
Model 5) Criteria accepting both L100 and L90 More data	63	Y:566 N:566	0.6	0.67	0.63	0.65	Model was chosen due to high precision. Chosen over Model_L100 due to the high recall metric in the latter.	Adding extra data to improve classification

Model 6) New criteria accepting some L90	63	Y:447 N:447	0.68	0.54	0.72	0.62		

* Few true positives due to selection criteria

Trait extraction

The digitised herbarium sheets had standardised resolution which enabled the conversion from pixels to centimetres squared. Leaf area was calculated from the number of pixels in the predicted mask. The area of the largest in-circle was calculated using R (v 4.2.2, R Core Team 2022). To do this, the package concaveman (v 1.1.0, Gombin 2020) was used to create an outline of the leaf. This was then converted into a polygon to find the Pole of Inaccessibility from package polylabelr (v 0.2.0, Larsson 2020), a geographical point the furthest from the edges correlating to the visual centre of the polygon. The shortest distance to the edge from this point represented the radius of the circle and thus the area. This was done using the function pointDistance from the package raster (v 3.6-14, Hijmans 2023). Curvature was represented through calculating the convex hull of the leaf area and comparing it through a ratio of area to the leaf mask area. Here the convex hull was calculated with chull from base R.

Supplementary Information B - Leaf masking model's labelling protocol

Set up

<u>Data setup</u> (*Eucalyptus* only)

- Records of *Eucalyptus* sp. were assigned their respective subgenus according to the species in question following the classification of Nicolle (2022), Slee et al. (2020), Thornhill et al. (2019).
- The number of distinct species in each subgenus of *Eucalyptus* was counted. If there were less than 10 distinct species, the subgenus would be classified as "small", if there were more, it would be classified as "big".

Labelling

- LabelMe (v 5.01, Wada 2022) was used to label the sheets under the respective labels below using the 'Create Polygons' function.
- As we were aiming for measurements on solely the leaf blades, the petioles were excluded. However, for eucalypts it was difficult to define where one segment started and the other ended thus an approximation was used.
- Bounding boxes were first drawn to include at least one Leaf100 when possible. All leaves with an area of greater than 50% were labelled with the labels below.
- It is to be noted that if the leaf was covered completely across by any object, the labelling would not go around that object.

Labels used in various model iterations:

- It is to be noted that the protocol illustrates the categories used for the first iteration of the leaf masking model. Subsequent iterations of the model merged/removed categories following SI.
- **BB –** Bounding box. This was selected to contain at least one Leaf100/Leaf100B when possible. A suitable size was selected based on leaf area of specimen, with an average of 6 total labelled leaves per sheet.
- **Leaf100** – Complete leaves. No abnormal indentation that indicated herbivory, and no part of the leaf was covered by another. An example can be seen in Figure SB_1 below.
- **Leaf100B** – Complete leaves, blemished. Minor abnormal indentation observed that indicate herbivory or cracks.
- **Leaf90** – Partial leaves. Leaves that had more than 90% of the blade visible, the remaining 10% may be from herbivory, coverage or bending of the leaf tip.
- **Leaf50** – Partial leaves. Leaves that had less than 90%, but more than 50%, of the blade visible.
- **Leaf##UM** – Leaves that were the juvenile version of their respective groups.

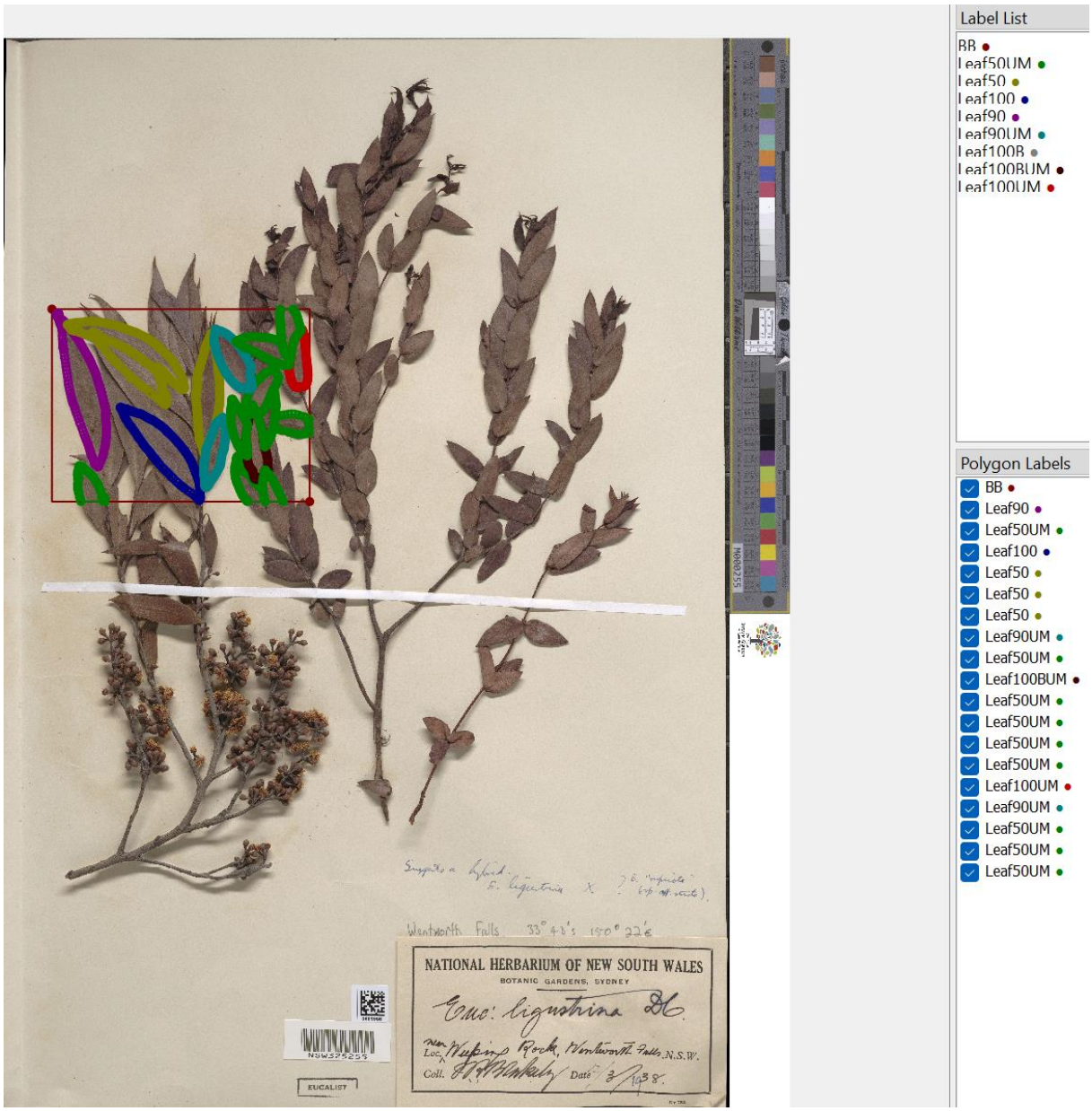

Figure SB_1. *An example of a manually annotated herbarium sheet*

Supplementary Information C - Extra Analyses

Leaf area

This analysis follows on from the main body text's analysis. It supplements the analyses of leaf area in its relationship to mean annual temperature, rather than mean annual precipitation.

Temperature

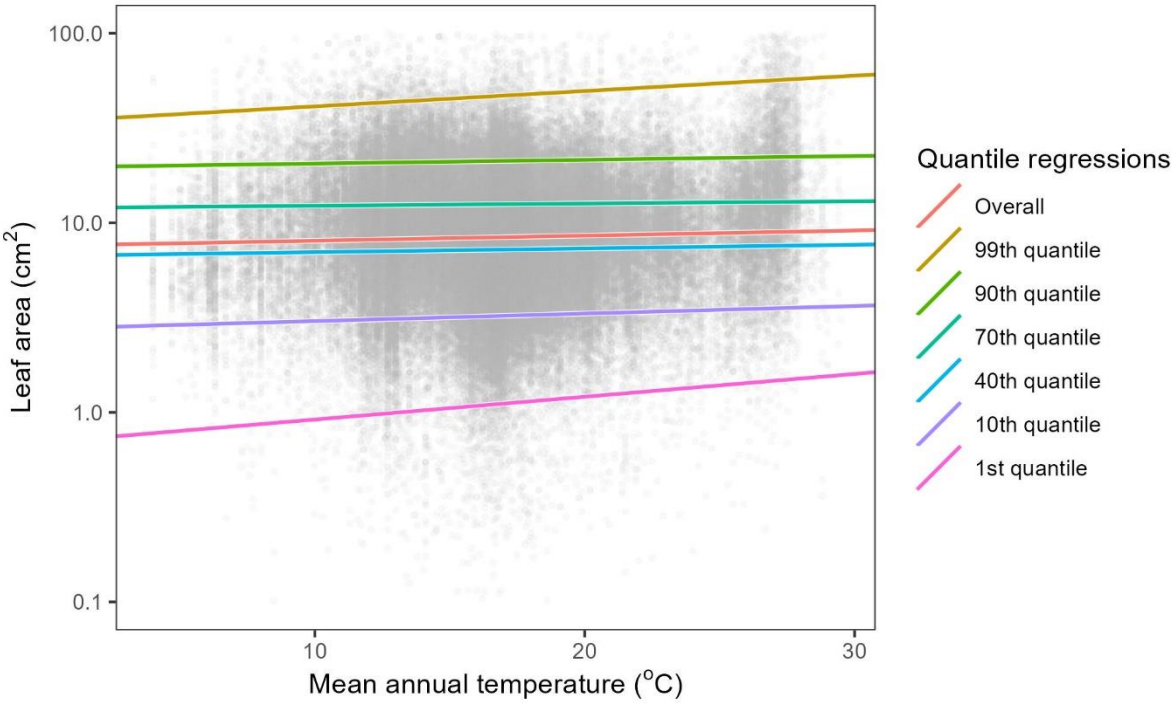

Figure SC_1. *Quantile regression analysis model results between leaf area and mean annual temperature.*

Table SC_2. *Coefficients for the overall linear model between leaf area and mean annual temperature, and the different levels of regression quantiles.*

	Slope	Relative standard error
Overall	0.00269	0.000216
99th quantile	0.0081	0.000888
90th quantile	0.002	0.00035
70th quantile	0.00119	0.000249
40th quantile	0.00196	0.000259
10th quantile	0.00397	0.000396
1st quantile	0.012	0.00124
AusTraits eucalypts	0.0108	0.00135
Wright et al.'s eucalypts	0.011	0.00783

Table SC_3. *Coefficients of log leaf area and mean annual temperature following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within, and Wright et al.'s (2017) results.*

	Slope	R-squared	Relative standard error
Overall	0.00269	0.00113	0.000216
Mean species model	0.278	0.0145	0.0718
Mixed model	0.0059	N/A	N/A
Wright's all taxa	0.041	0.15	0.003

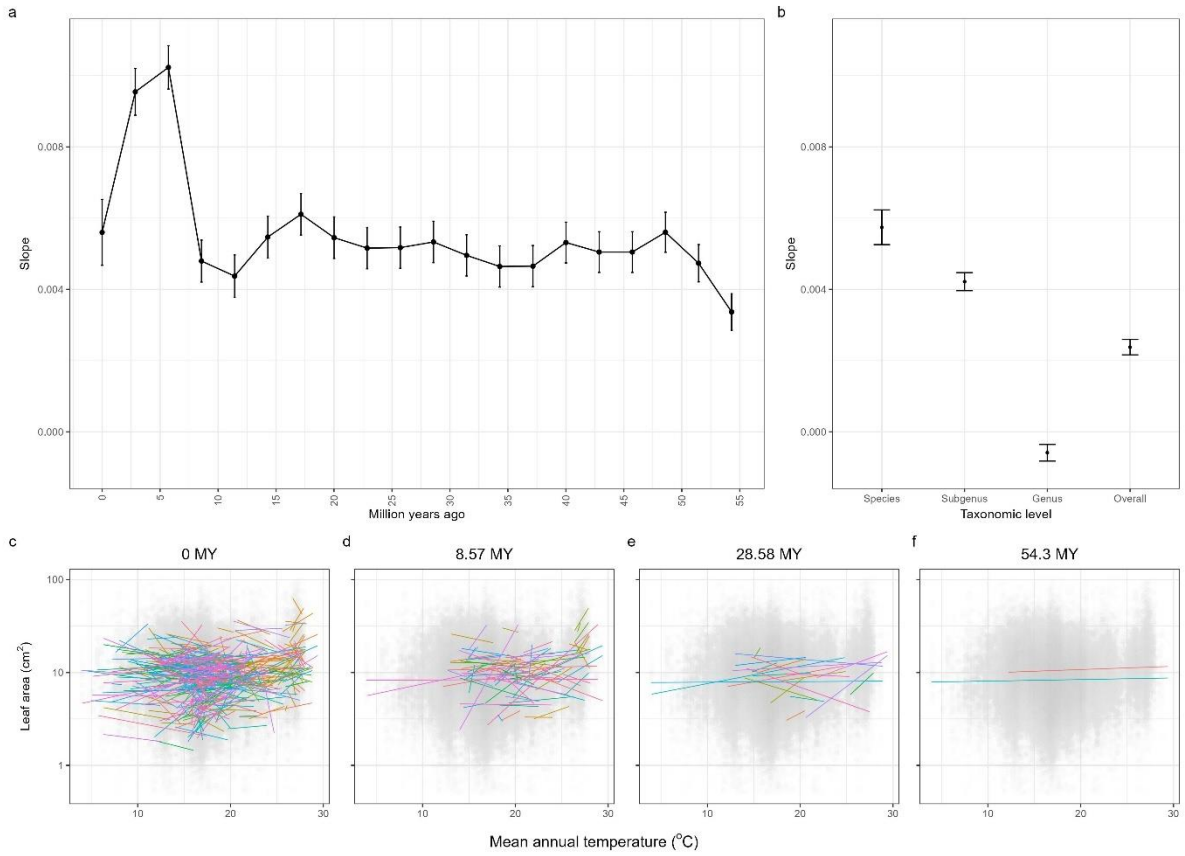

Figure SC_4. *a) The phylogenetic tree was split at 20 intervals at evenly spaced time periods. The mean slope within the clades formed at each time point was calculated. For example, 0 MY had each species as a random effect, whereas 55 MY had two groups of species, corresponding to the deepest branch among the eucalypts. A convergence towards an approximate average slope was observed roughly 8.5 MY. b) The average slope and standard error where the respective taxonomic level was used as the random effect in a mixed model. The 'overall' model has no random effect. Species: 0.00574 ± 0.000487. Subgenus: 0.00421 ± 0.000252. Genus: -0.000585 ± 0.000234. Overall: 0.00269 ± 0.000216. c-f) Each lineage's linear models at four different intervals (0 MY, 8.57 MY, 28.58 MY, 54.3 MY) are illustrated. Where each colour represents a lineage.*

Reflecting Physiologist Sampling - Leaf area

Our dataset was also analysed in a way that reflects conventional sampling methods used by physiologists. This was done through removing the bottom 50% of leaves by species, as suggested by Corney et al. (2012). From these analyses, a significant increase in slopes between leaf area and precipitation was observed. Furthermore, the physiologist sampling method resulted in the loss of the constraint triangle, and a shift towards a more linear relationship between leaf area and climate variables. These results illustrate how the trait sampling method can significantly alter the outcome of the analyses. This set of analyses also further reinforce the validity of our method as our trait-climate results converge to relationships of other datasets.

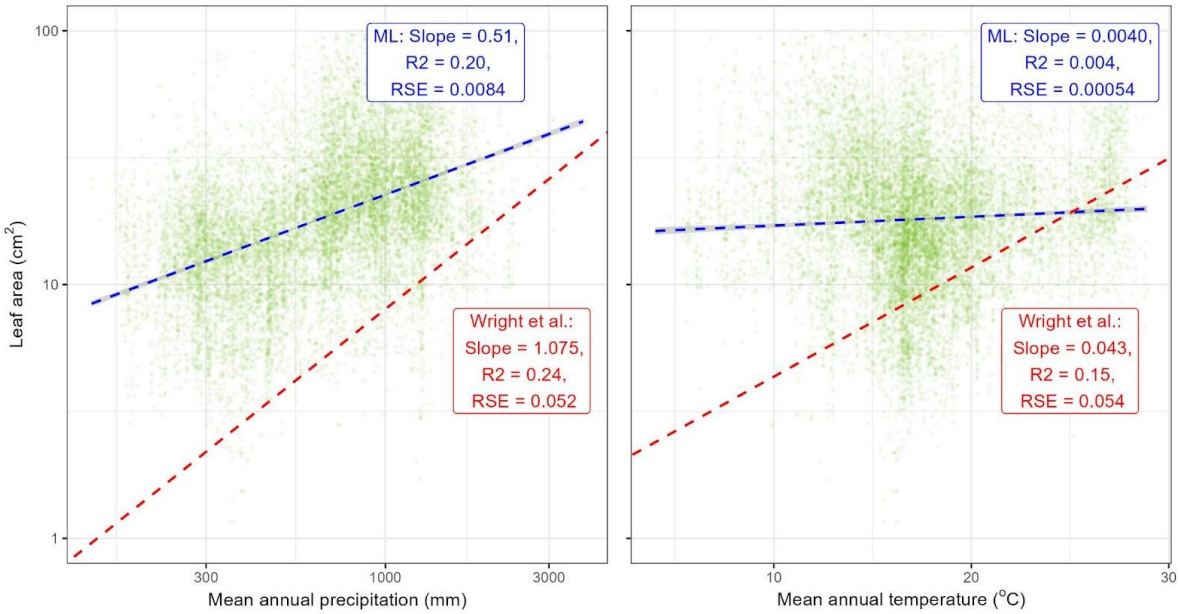

Figure SC_5. *Relationships between log leaf area and the climatic variables (log mean annual precipitation and temperature). The blue dashed lines represent the linear model results. The red dashed lines represent the results found in the Wright et al. (2017) analysis of global leaf traits. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

Table SC_6. *Coefficients of models for log leaf area against log mean annual precipitation and mean annual temperature in comparison to other datasets. Coefficients for Wright's data was sourced from the supplementary information of Wright et al. (2017), which used a mixed regression model. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

	Slope	R-squared	Relative standard error
Log leaf area ~ log mean annual precipitation			
Overall	0.510	0.203	0.00843
AusTraits eucalypts	0.685	0.268	0.0214
Wright et al.'s eucalypts	0.446	0.245	0.0655
Wright et al.'s all taxa	1.08	0.24	0.052
Log leaf area ~ mean annual temp			
Overall	0.00402	0.00376	0.000542
AusTraits eucalypts	0.0108	0.0218	0.00114
Wright et al.'s eucalypts	0.0110	0.00685	0.00783
Wright et al.'s all taxa	0.043	0.15	0.054

Precipitation

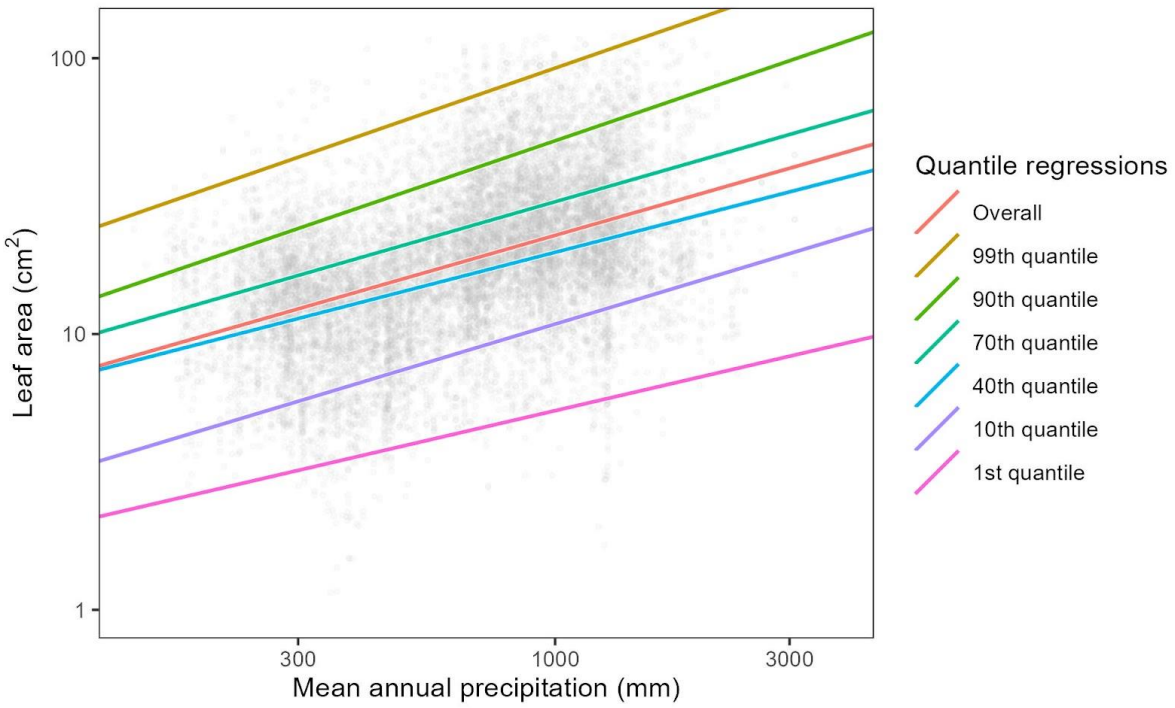

Figure SC_7. *Quantile regression analysis model results. A linear relationship between log leaf area and log mean annual precipitation is observed. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

Table SC_8. *Coefficients for the overall linear model between log leaf area and log mean annual precipitation, and the different levels of regression quantiles. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

	Slope	Relative standard error
Overall	0.510	0.00843
99th quantile	0.618	0.0246
90th quantile	0.608	0.0159
70th quantile	0.511	0.0103
40th quantile	0.459	0.00883
10th quantile	0.536	0.0182
1st quantile	0.414	0.0327
AusTraits eucalypts	0.685	0.0214
Wright et al.'s eucalypts	0.446	0.0655

Table SC_9. *Coefficients of log leaf area and log mean annual precipitation following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within, and Wright et al.'s (2017) results. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

	Slope	R-squared	Relative standard error
Overall	0.510	0.203	0.00843
Mean species model	0.576	0.261	0.0462
Mixed model	0.0300	N/A	N/A
Wright et al.'s all taxa	1.08	0.24	0.052

Temperature

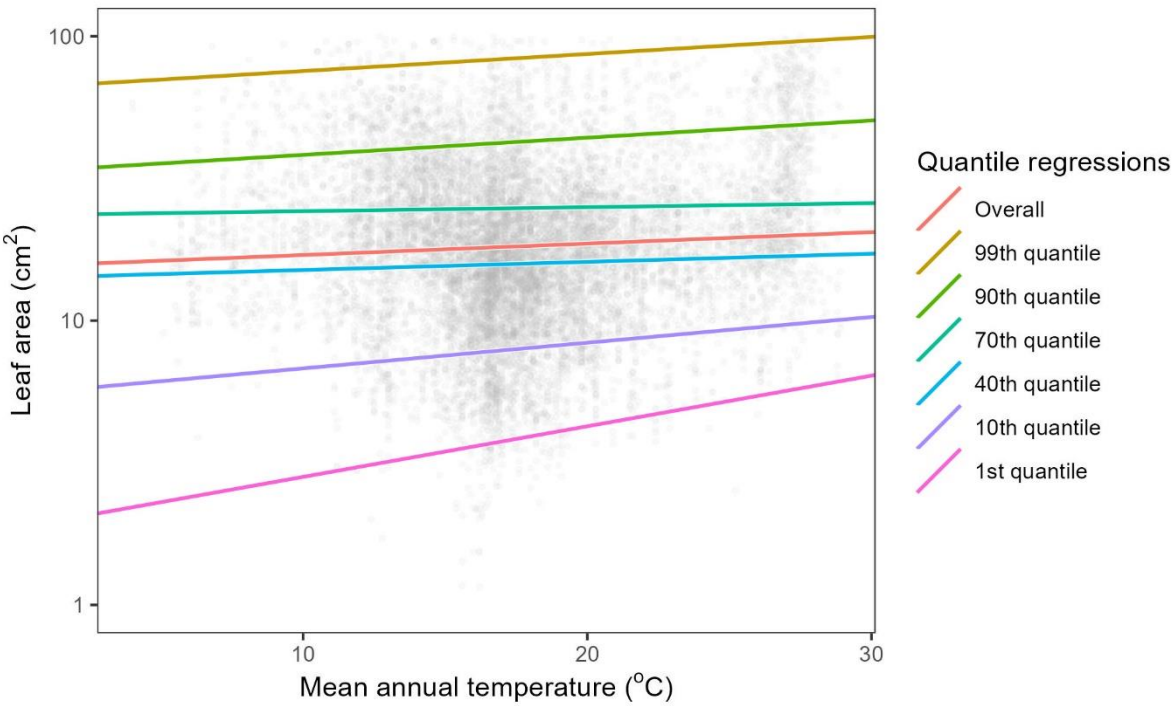

Figure SC_10. *Quantile regression analysis model between log leaf area and mean annual temperature. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

Table SC_11. *Coefficients for the overall linear model between log leaf area and mean annual temperature, and the different levels of regression quantiles. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

	Slope	Relative standard error
Overall	0.00402	0.000542
99th quantile	0.00600	0.00169
90th quantile	0.00605	0.00112
70th quantile	0.00141	0.000672
40th quantile	0.00287	0.000586
10th quantile	0.00903	0.00103
1st quantile	0.0178	0.00142
AusTraits eucalypts	0.0108	0.00135
Wright et al.'s eucalypts	0.0110	0.00783

Table SC_12. *Coefficients of log leaf area and mean annual temperature following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within, and Wright et al.'s (2017) results. The dataset used for this analysis has undergone a filtering of the bottom 50% leaves by species.*

	Slope	R-squared	Relative standard error
Overall	0.00402	0.00376	0.000542
Mean species model	0.113	-0.0000951	0.115
Mixed model	0.000515	N/A	N/A
Wright's all taxa	0.041	0.15	0.003

Largest in-circle area

The analysis was also repeated on measurements of other leaf traits collected in our dataset. This includes the area of the largest circle able to be drawn within the leaf mask, similar to Leigh et al. 2017.

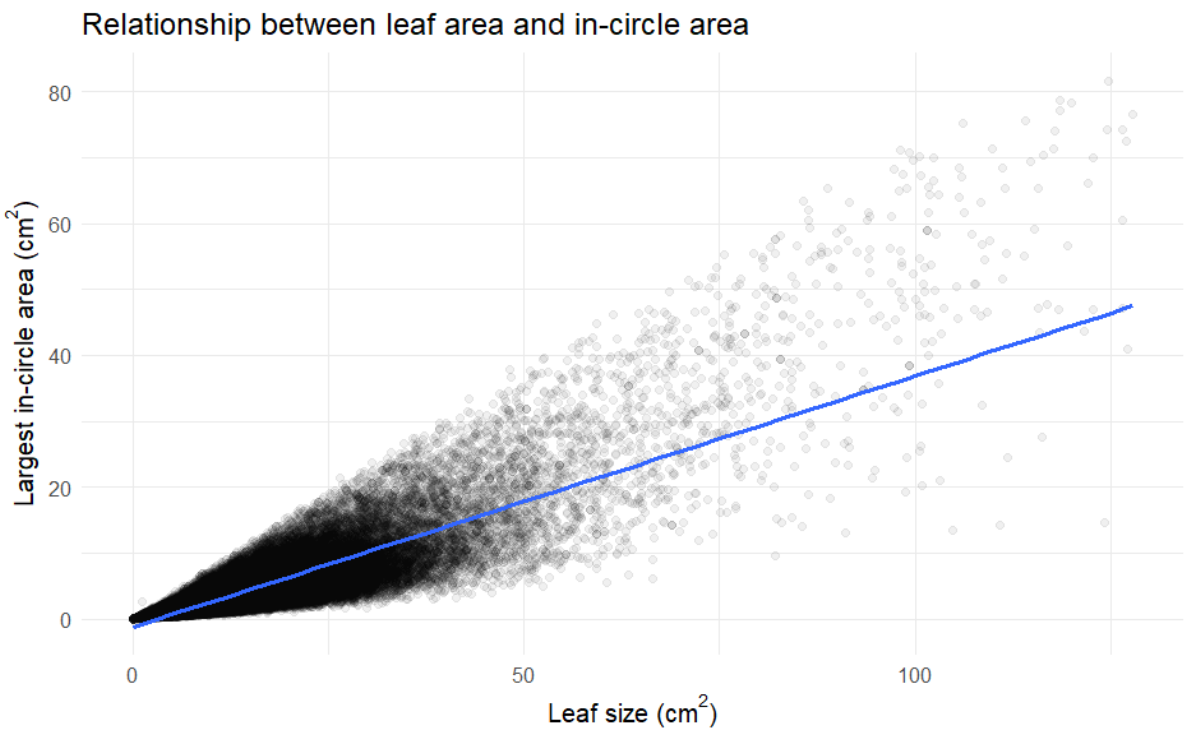

Figure SC_13. *Comparison between leaf area to largest in-circle area. A linear relationship between the two variables is present.*

Table SC_14. *Phylogenetic signal for log largest in-circle area against the ML2 phylogeny estimated by Thornhill et al. (2019).*

	K-value	P-value (1000 randomisations)
Log largest in-circle area	0.0227	0.001

Precipitation

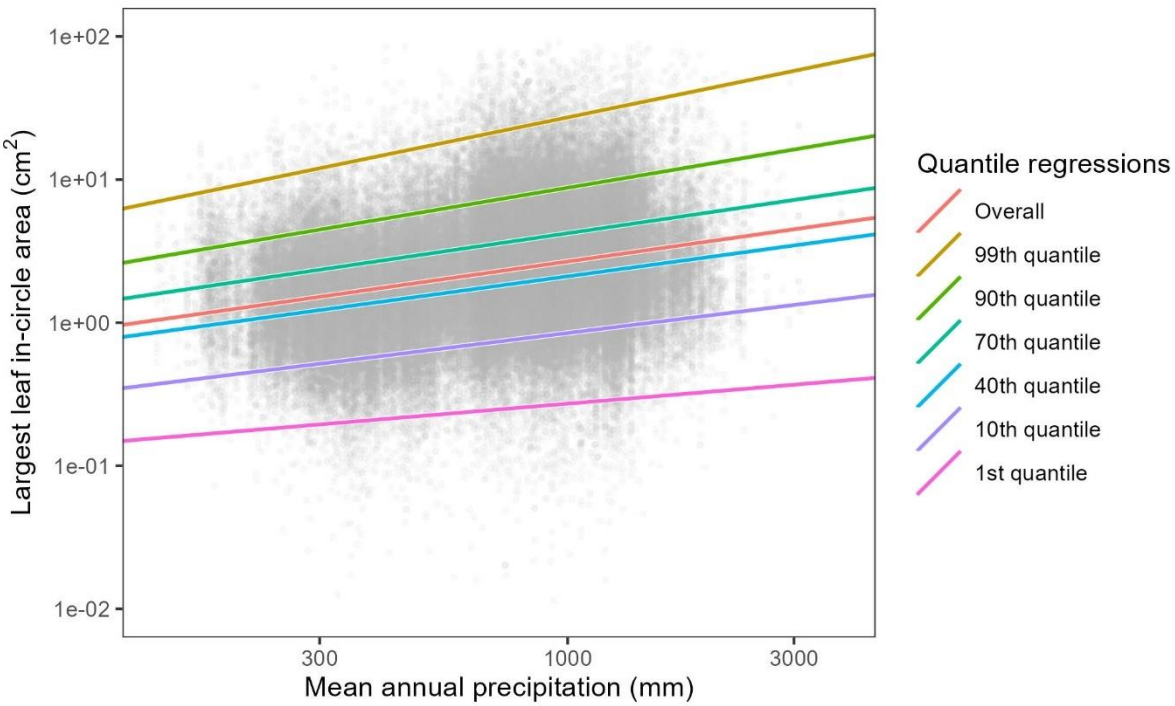

Figure SC_15. *Quantile regression analysis model results of log largest in-circle area and log mean annual precipitation. An increase in slope steepness from the 1st to the 99th quantile, with a lower range of leaf area variation in drier conditions than wetter is observed.*

Table SC_16. *Coefficients for the overall linear model between log largest in-circle area and log mean annual precipitation, and the different levels of regression quantiles.*

	Slope	Relative standard error
Overall	0.472	0.00455
99th quantile	0.68	0.0191
90th quantile	0.56	0.008
70th quantile	0.488	0.00565
40th quantile	0.452	0.00535
10th quantile	0.411	0.00747
1st quantile	0.278	0.0225

Table SC_17. *Coefficients of log largest in-circle area and log mean annual precipitation following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

	Slope	R-squared	Relative standard error
Overall	0.472	0.0731	0.00455
Mean species model	0.609	0.173	0.0432
Mixed model	0.213	N/A	N/A

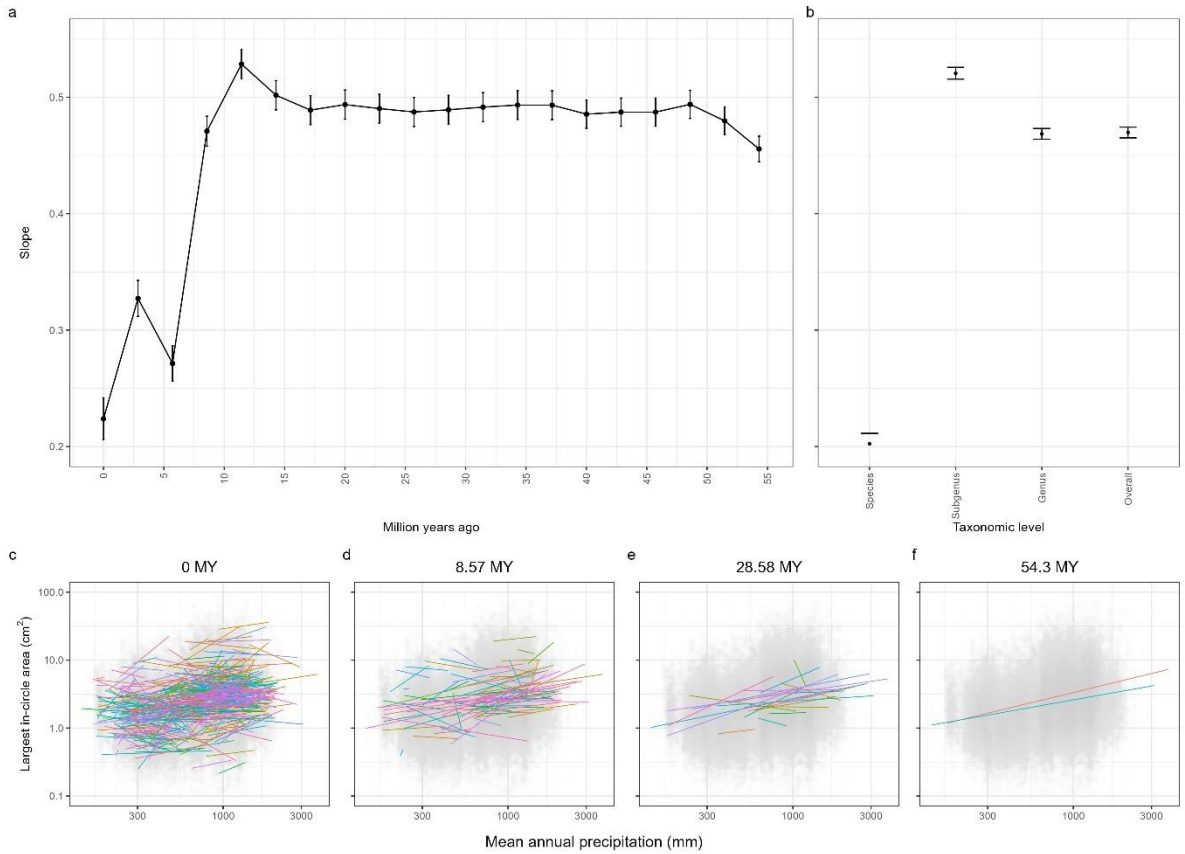

Figure SC_18. *a) The phylogenetic tree was split at 20 intervals at evenly spaced time periods. The mean slope within the clades formed at each time point was calculated. For example, 0 MY had each species as a random effect, whereas 55 MY had two groups of species, corresponding to the deepest branch among the eucalypts. A convergence towards an approximate average slope was observed roughly 12.5 MY. b) The average slope and standard error where the respective taxonomic level was used as the random effect in a mixed model. The 'overall' model has no random effect. Species: 0.202 ± 0.00892. Subgenus: 0.521 E-01 ± 0.00506. Genus: 0.469 ± 0.00459. Overall: 0.470 ± 0.004.58. c-f) Each lineage's linear models at four different intervals (0 MY, 8.57 MY, 28.58 MY, 54.3 MY) are illustrated. Where each colour represents a lineage.*

Temperature

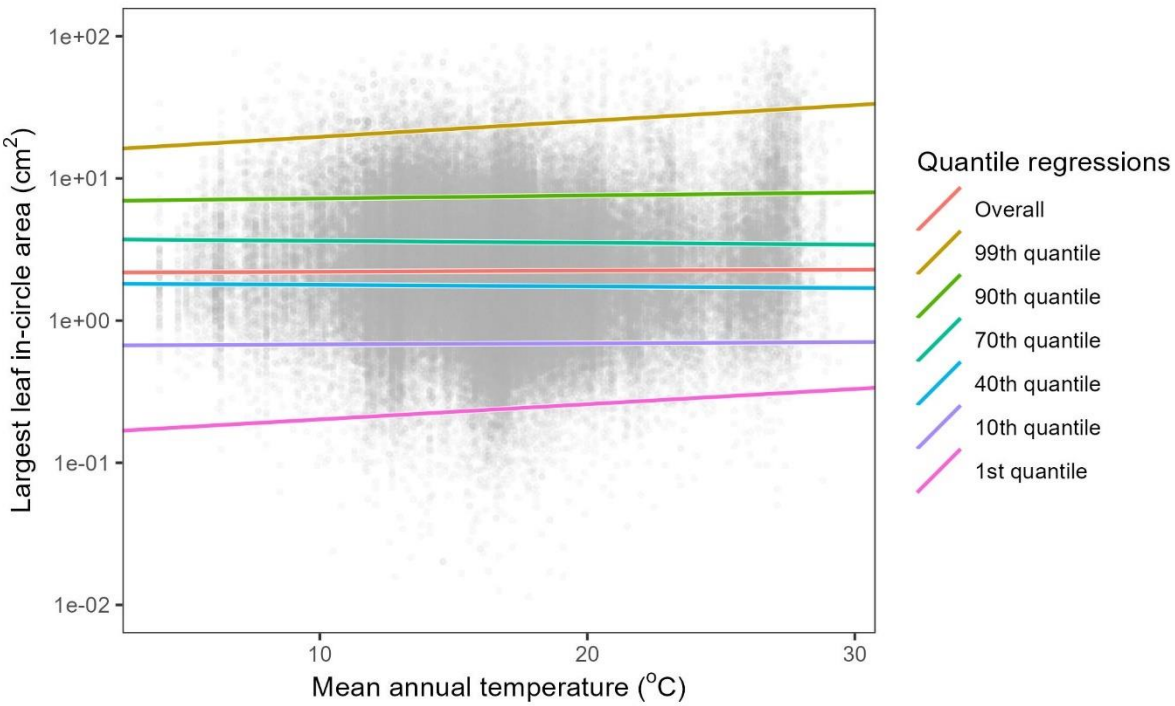

Figure SC_19. *Quantile regression analysis model results of log largest in-circle area and mean annual temperature. An increase in slope steepness from the 1st to the 99th quantile, with a lower range of leaf area variation in drier conditions than wetter is observed.*

Table SC_20. *Coefficients for the overall linear model between log largest in-circle area and mean annual temperature, and the different levels of regression quantiles.*

	Slope	Relative standard error
Overall	0.00269	0.000268
99th quantile	0.0081	0.000888
90th quantile	0.002	0.00035
70th quantile	0.00119	0.000249
40th quantile	0.00196	0.000259
10th quantile	0.00397	0.000396
1st quantile	0.012	0.000124

Table SC_21. *Coefficients of log largest in-circle area and mean annual temperature following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

	Slope	R-squared	Relative standard error
Overall	0.000737	0.0000481	0.000268
Mean species model	0.0105	0.017	0.00253
Mixed model	0.00395	N/A	N/A

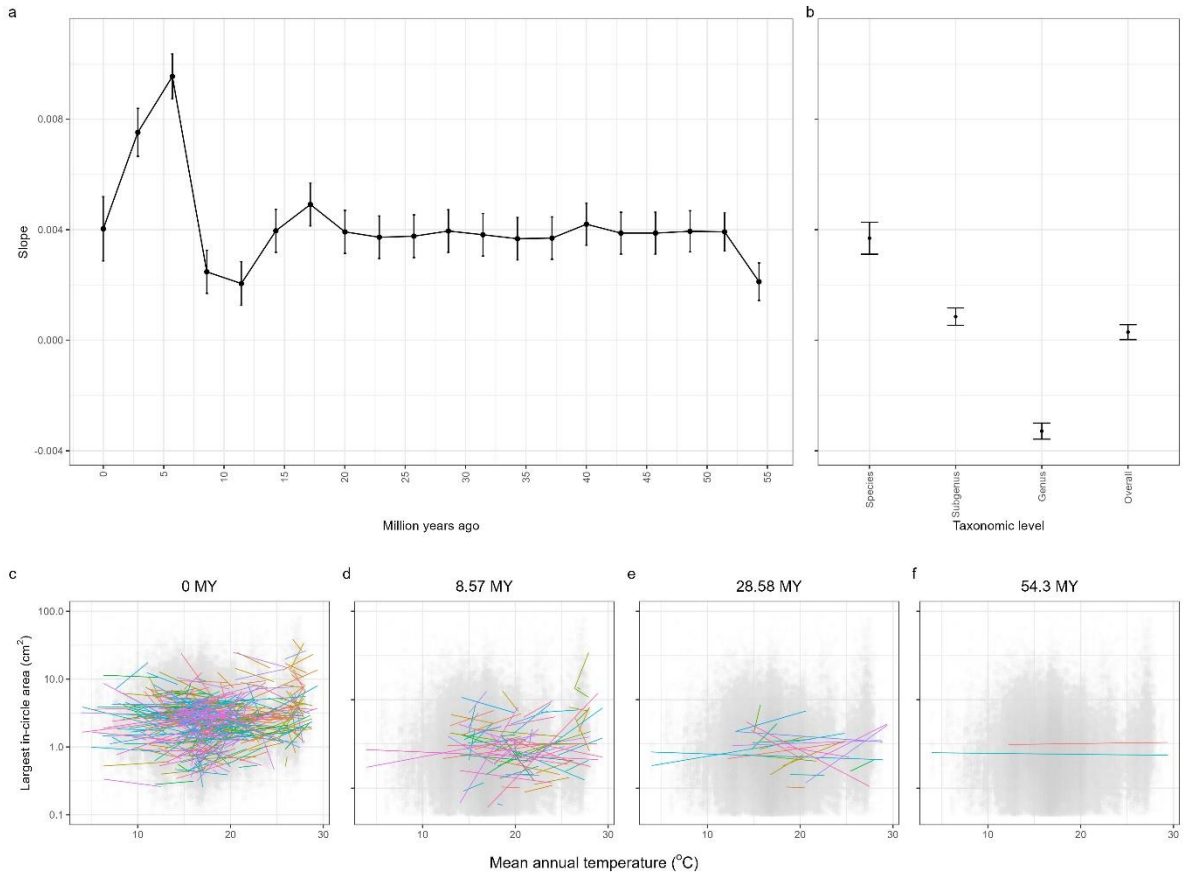

Figure SC_22. *a) The phylogenetic tree was split at 20 intervals at evenly spaced time periods. The mean slope within the clades formed at each time point was calculated. For example, 0 MY had each species as a random effect, whereas 55 MY had two groups of species, corresponding to the deepest branch among the eucalypts. A convergence towards an approximate average slope was observed roughly 12.5 MY. b) The average slope and standard error where the respective taxonomic level was used as the random effect in a mixed model. The 'overall' model has no random effect. Species: 0.00369 ± 0.000578. Subgenus: 0.000853 ± 0.000313. Genus: -0.00329 ± 0.000290. Overall: 0.000291 ± 0.000268. c-f) Each lineage's linear models at four different intervals (0 MY, 8.57 MY, 28.58 MY, 54.3 MY) are illustrated. Where each colour represents a lineage.*

Leaf curvature

The analysis was also repeated on measurements of other leaf traits collected in our dataset. This includes the leaf curvature, which is represented by the area of a convex hull of the leaf over the total leaf mask area.

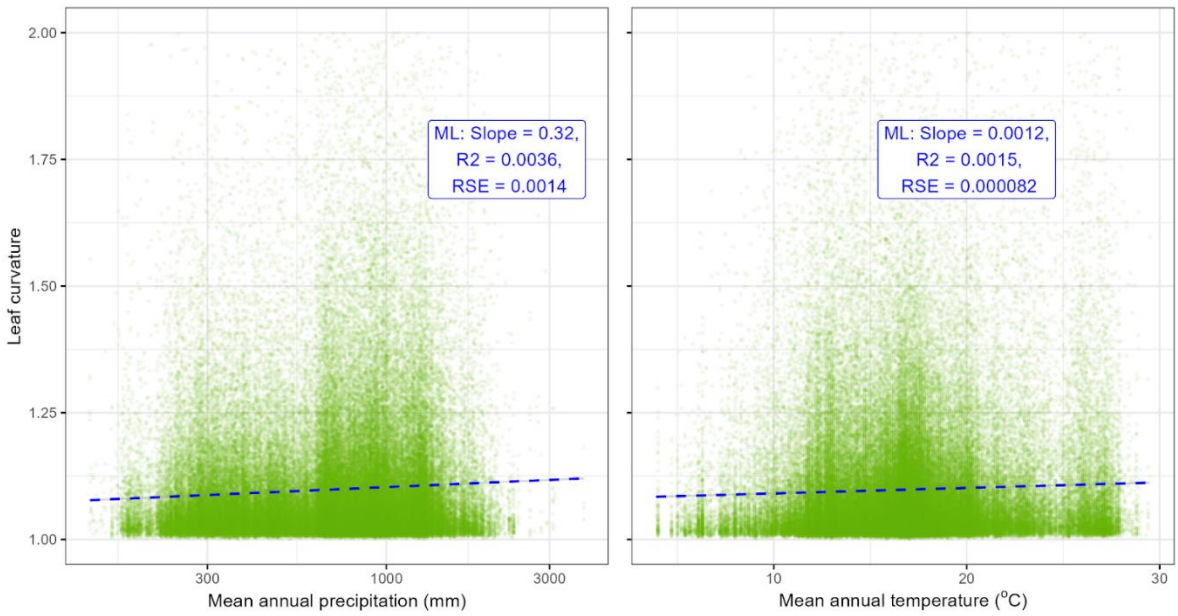

Figure SC_22. *Relationships between the climatic variables (mean annual temperature and precipitation) against leaf curvature. Where the blue dashed line represents a linear model.*

Table SC_23. *Phylogenetic signal for leaf curvature against the ML2 phylogeny estimated by Thornhill et al. (2019)*

	K-value	P-value (1000 randomisations)
Curvature ratio	0.0158	0.203

Precipitation

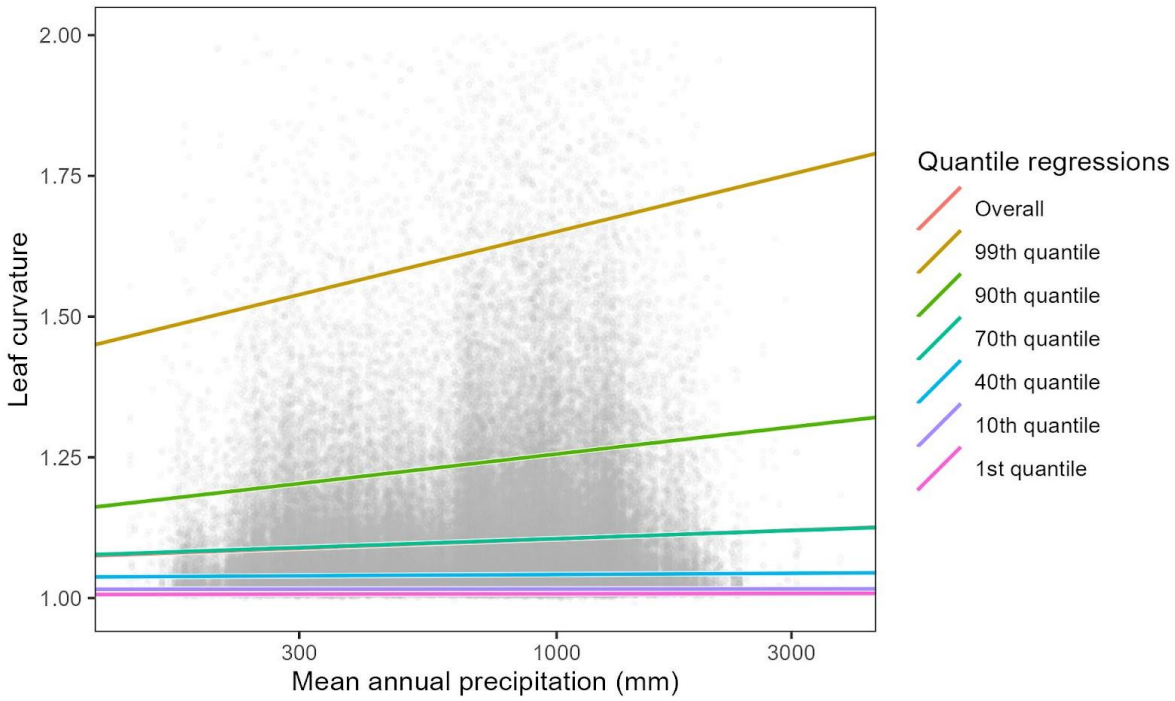

Figure SC_27. *Quantile regression analysis model results of leaf curvature and log mean annual precipitation. An increase in slope steepness from the 1st to the 99th quantile is observed.*

Table SC_28. *Coefficients for the overall linear model between leaf curvature and log mean annual precipitation, and the different levels of regression quantiles.*

	Slope	Relative standard error
Overall	0.0319	0.00144
99th quantile	0.214	0.0196
90th quantile	0.1	0.00474
70th quantile	0.0303	0.00172
40th quantile	0.00454	0.000601
10th quantile	0.000417	0.000251
1st quantile	0.00116	0.000242

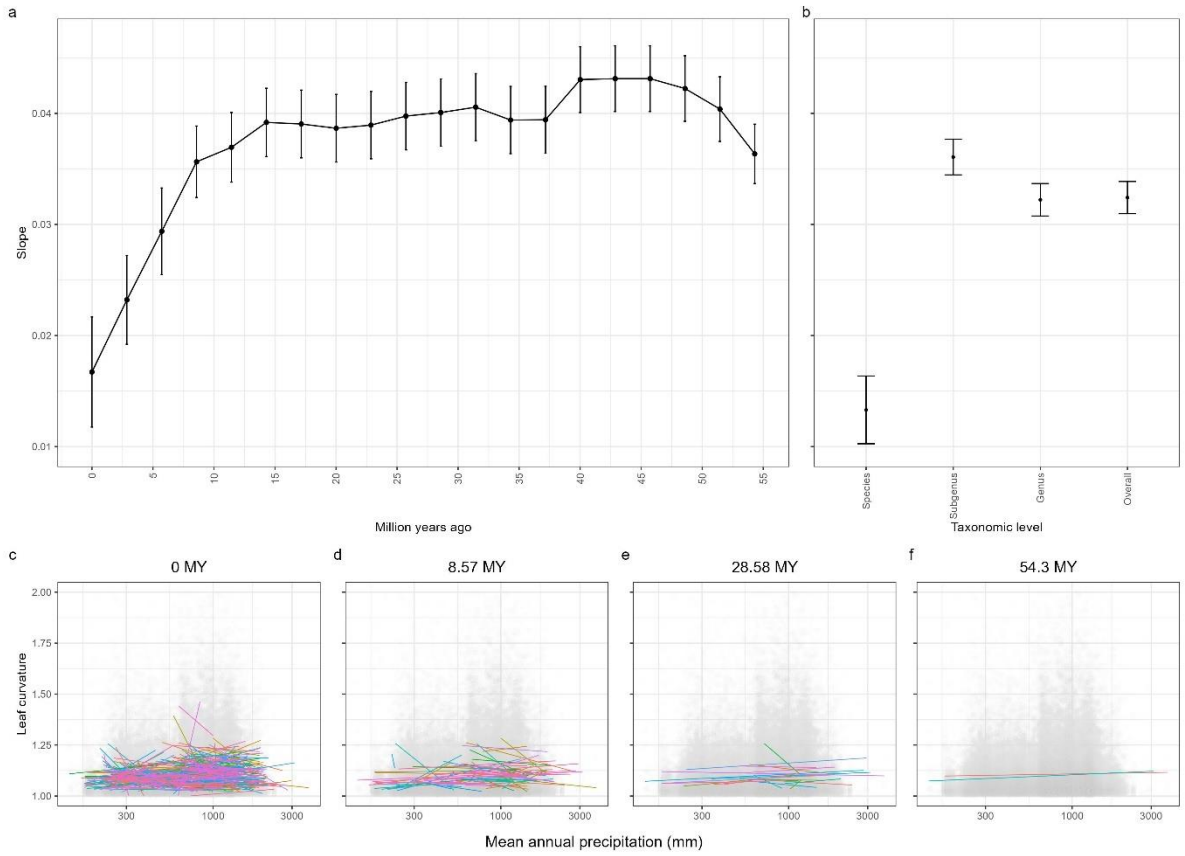

Figure SC_29. *a) The phylogenetic tree was split at 20 intervals at evenly spaced time periods. The mean slope within the clades formed at each time point was calculated. For example, 0 MY had each species as a random effect, whereas 55 MY had two groups of species, corresponding to the deepest branch among the eucalypts. A convergence towards an approximate average slope was observed roughly 15 MY. b) The average slope and standard error where the respective taxonomic level was used as the random effect in a mixed model. The 'overall' model has no random effect. Species: 0.0133 ± 0.00305. Subgenus: 0.0361 ± 0.00161. Genus: 0.0322 ± 0.00146. Overall: 0.0319 ± 0.126. c-f) Each lineage's linear models at four different intervals (0 MY, 8.57 MY, 28.58 MY, 54.3 MY) are illustrated. Where each colour represents a lineage.*

Table SC_30. *Coefficients of leaf curvature and log mean annual precipitation following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

	Slope	R-squared	Relative standard error
Overall	0.0319	0.00357	0.00144
Mean species model	0.0315	0.0286	0.00586
Mixed model	0.0159	N/A	N/A

Temperature

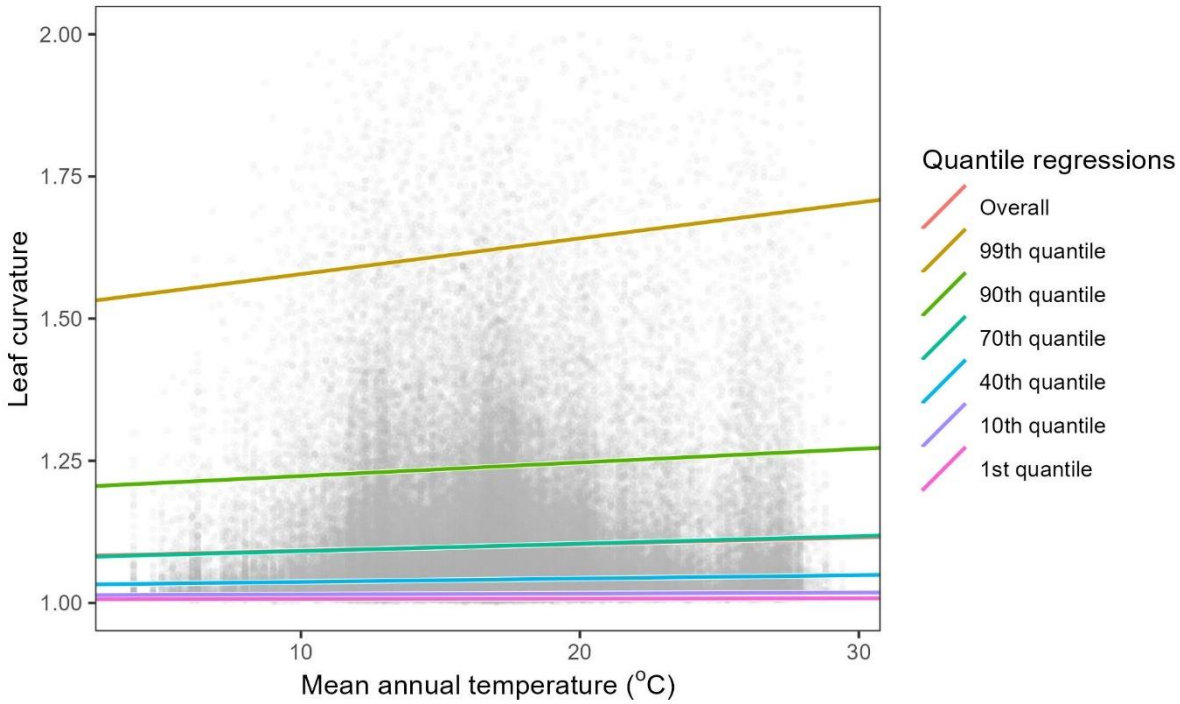

Figure SC_31*. Quantile regression analysis model results of leaf curvature and mean annual temperature. An increase in slope steepness from the 1st to the 99th quantile is observed.*

***Table SC_32**. Coefficients for the overall linear model between leaf curvature and mean annual temperature, and the different levels of regression quantiles.*

	Slope	Relative standard error
Overall	0.00117	0.0126
99th quantile	0.00629	0.0012
90th quantile	0.00239	0.000282
70th quantile	0.00131	0.000103
40th quantile	0.00059	0.0000354
10th quantile	0.000162	0.0000143
1st quantile	0.0000526	0.0000131

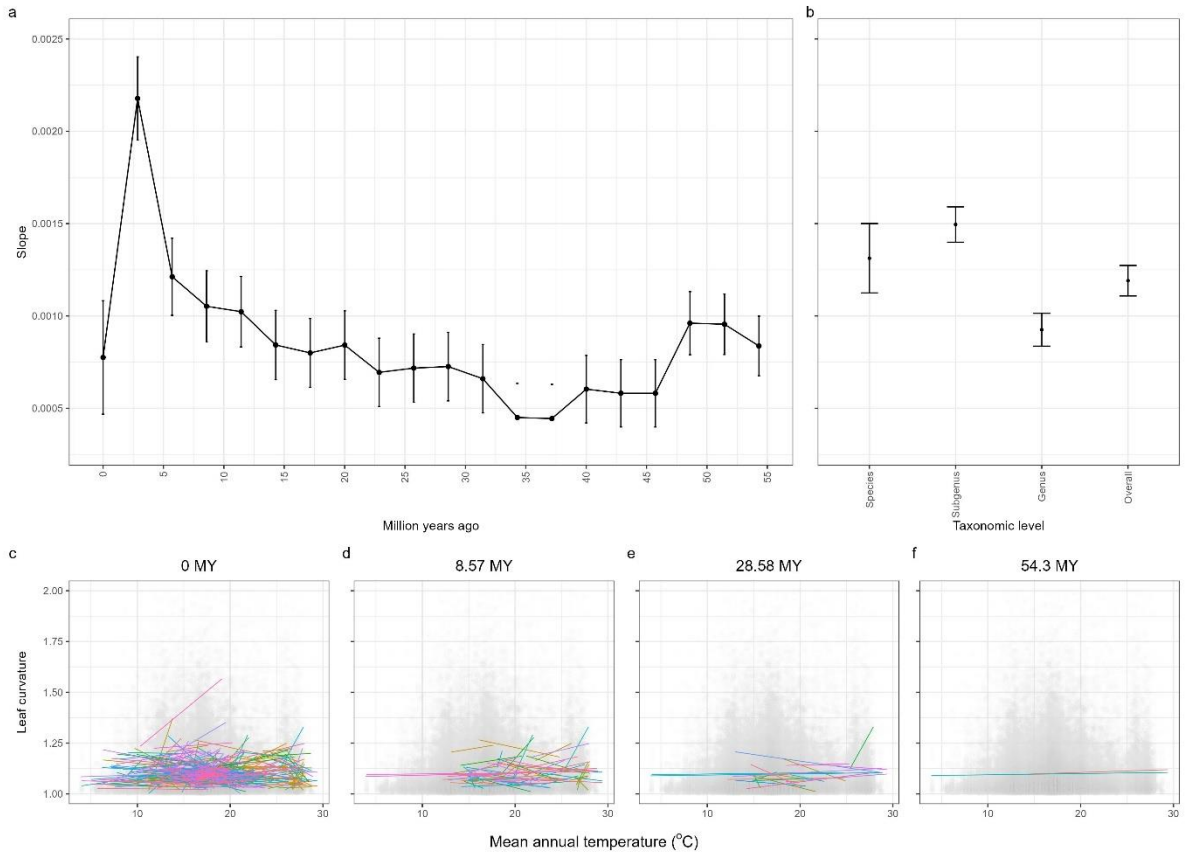

Figure SC_33*. a) The phylogenetic tree was split at 20 intervals at evenly spaced time periods. The mean slope within the clades formed at each time point was calculated. For example, 0 MY had each species as a random effect, whereas 55 MY had two groups of species, corresponding to the deepest branch among the eucalypts. b) The average slope and standard error where the respective taxonomic level was used as the random effect in a mixed model. The 'overall' model has no random effect. Species: 0.00131 ± 0.000188. Subgenus: 0.00150 ± 0.0000958. Genus: 0.000925 ± 0.00008.91. Overall: 0.00117 ± 0.00008.17. c-f) Each lineage's linear models at four different intervals (0 MY, 8.57 MY, 28.58 MY, 54.3 MY) are illustrated. Where each colour represents a lineage.*

Table SC_34*. Coefficients of leaf curvature and mean annual temperature following Equation 6-8. An overall linear model, a linear model using average species mean, and a mixed model with species as a random effect and each herbarium sheet nested within.*

	Slope	R-squared	Relative standard error
Overall	0.00117	0.00149	0.0000817
Mean species model	0.00105	0.0103	0.000319
Mixed model	0.00127	N/A	N/A

Supplementary Information D - Error Validation

The study conducted a linear regression analysis between log leaf area and log mean annual precipitation. Prior to this, residuals were checked to ensure the assumptions of linearity were met. Through this we determined residuals were normally distributed and linearly related.

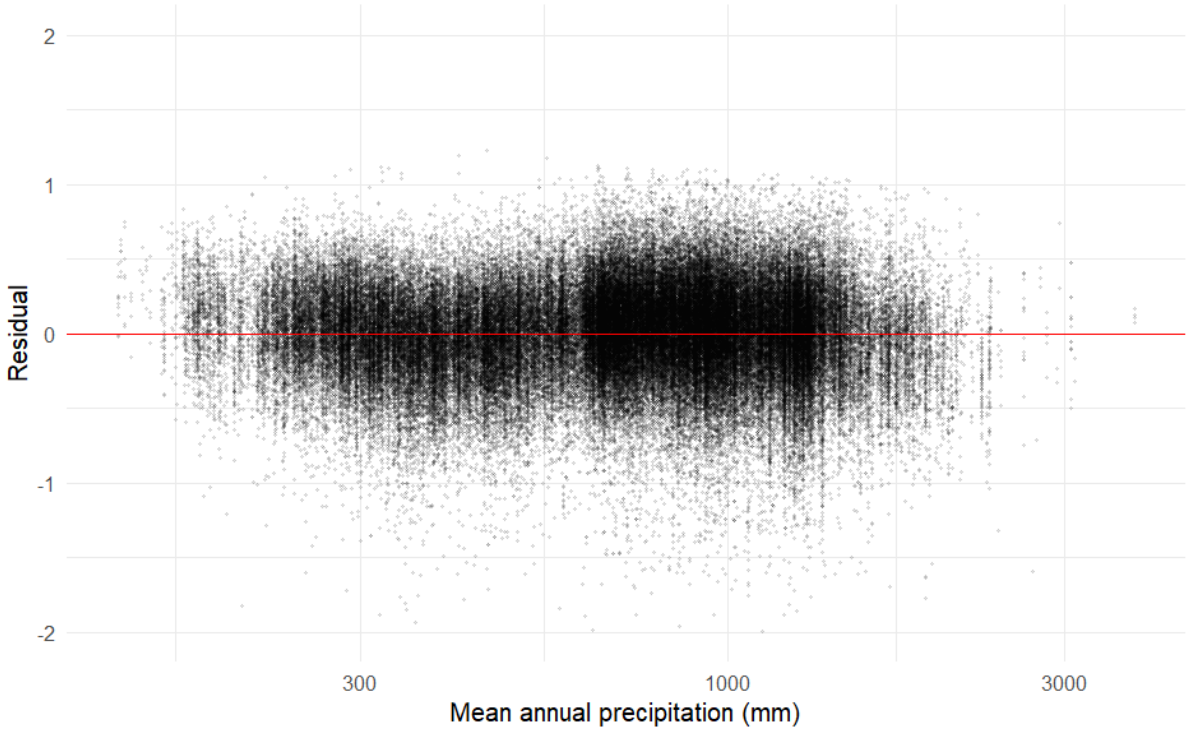

Figure SD_1. *Residual plot of log leaf area and log mean annual precipitation, showing homogeneity of variance and linearity.*

Quality control was also conducted. This data was sourced from retrieving 100 random leaves that passed the classifier model. These 100 leaves were manually sorted into valid and invalid leaves to see whether there was a bias in error towards bigger or smaller leaves. We determined that no significant bias in leaf area was present, as indicated by the significant overlap in leaf area of both categories.

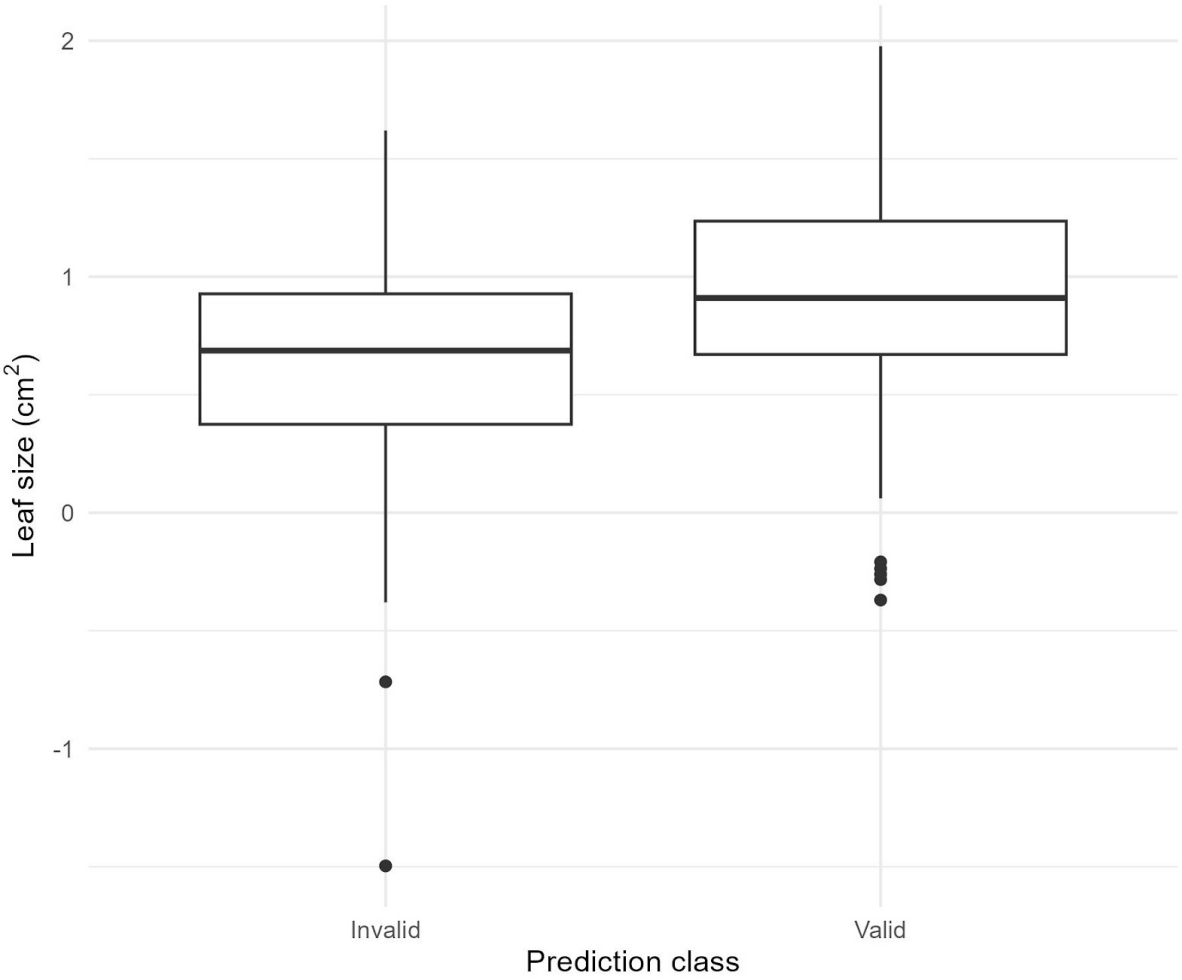

***Figure SD_2**. Examining the presence of bias in leaf areas of invalid/valid leaves. Plotting the normalised frequency count across the different leaf areas.*

The fitted slope between the two datasets (Wright et al.'s 2017 and AusTraits) compared to this study's, was observed to be shallower than the one-to-one relationship. This was attributed to our dataset's more complete sampling method.

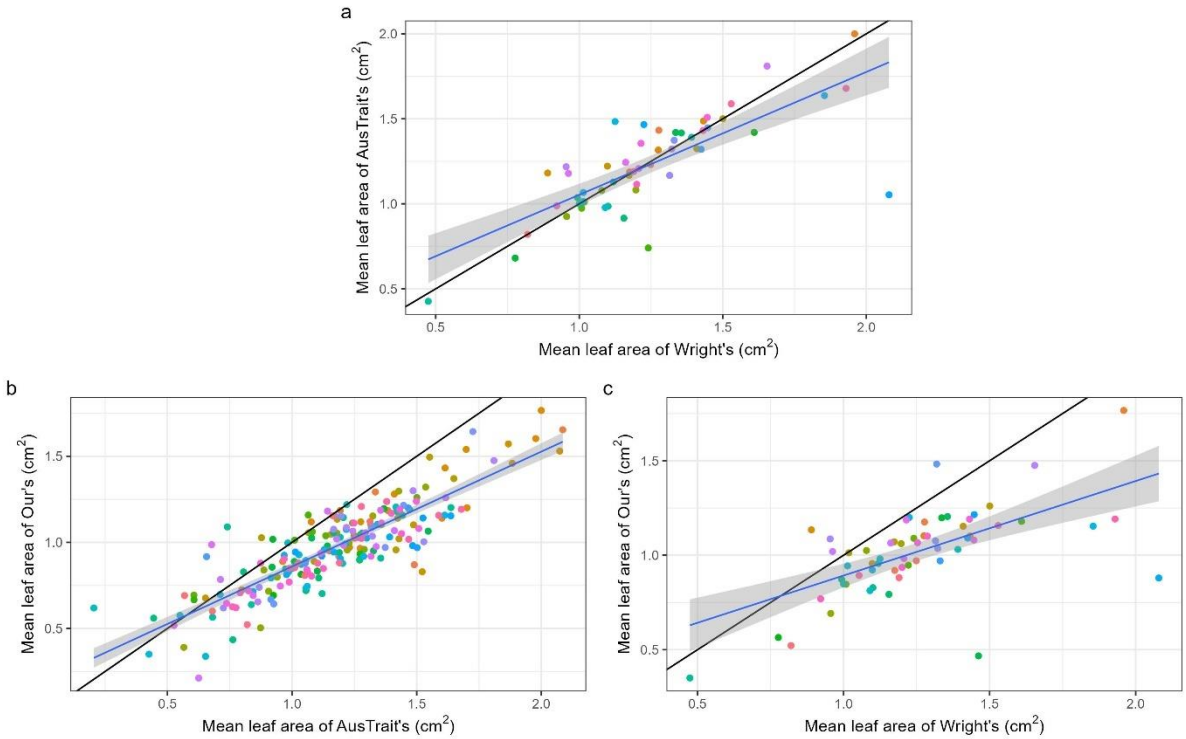

Figure SD_3*. Plotting the mean leaf area of shared species across databases, where each data point is a species. The black line is a one-to-one relationship between the two datasets (indicating an identical species mean). Whereas the blue line is the linear relationship between the two datasets. a) Plotting shared eucalypt species of AusTraits and Wright. b) Plotting shared eucalypt species of our dataset and AusTraits. c) Plotting shared eucalypt species of our dataset and Wright et al.'s (2017).*