Phil. Trans. R. Soc. B **371** doi: 10.1098/rstb.2015.0305

SUPPLEMENTARY INFORMATION FOR "Savannas of Asia: evidence for antiquity, current day biogeography and an uncertain future": Figures S1, S2, S3, Table S1, S2, S3 & Appendix 1

Jayashree Ratnam, Kyle W. Tomlinson, Dina N. Rasquinha, Mahesh Sankaran

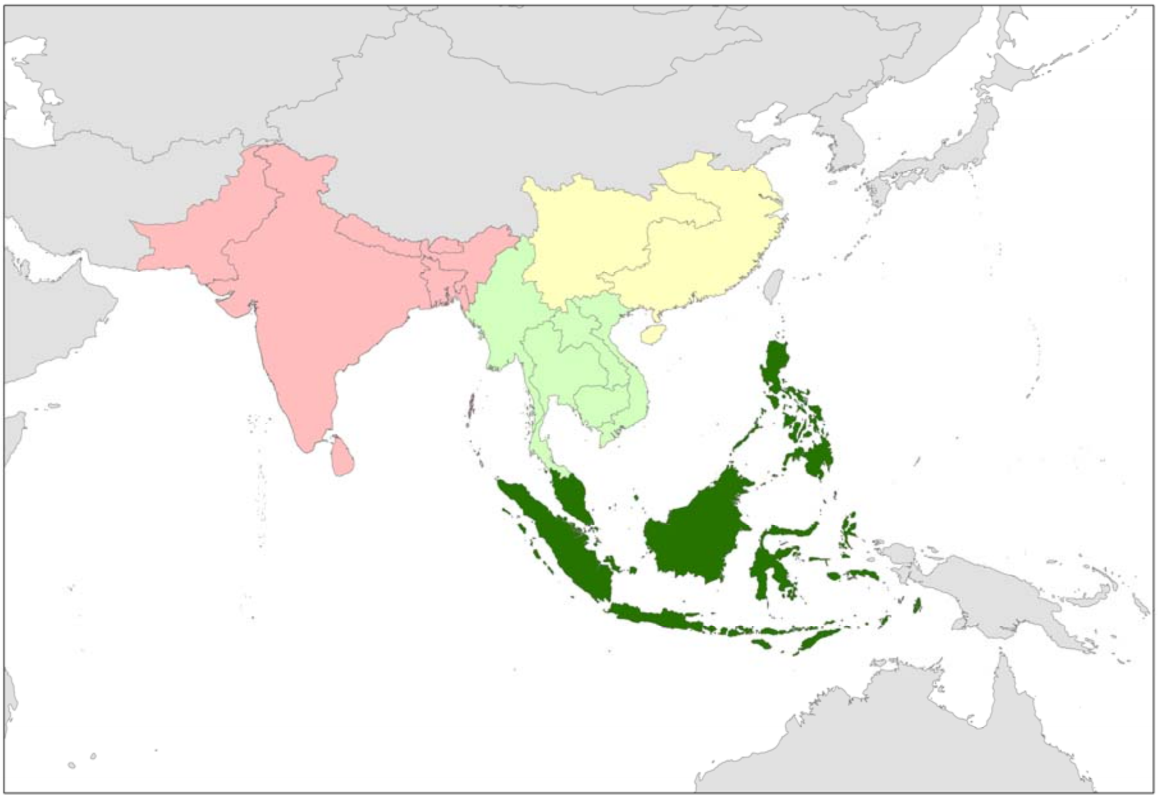

Figure S1 Map of the Asia showing regions covered by this study. Regions shown are those defined by the Taxonomic Databases Working Group (TDWG) level 2 regions (Clayton et al 2016) with the exception of East Asia which is based on TDWG level 3 regions for southern China (Southeast China, Southwest China, Hainan Island). Key to colour codes for regions: Pink = South Asia (Indian subcontinent); Yellow = East Asia (southern China); Pale green = Continental Southeast Asia (Indochina); Dark green = Oceanic Southeast Asia (Malesia).

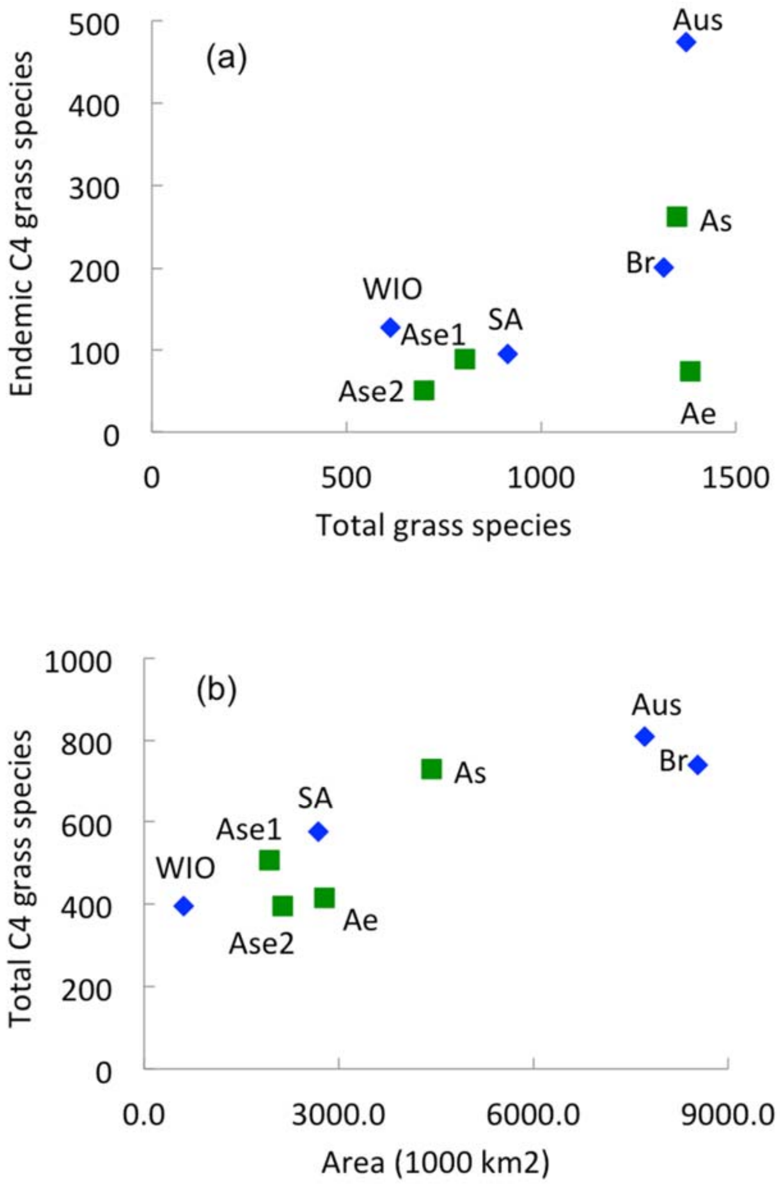

Figure S2 C$_4$ grass endemism and diversity in Asia as compared with other major savanna regions of the world. (a) C4 grass endemism per region versus total grass species per region. (b) Total C4 species per region as a function of region area in 1000s of km^2. Regions are all based on the Taxonomic Databases Working Group (TDWG) level 2 regions (Clayton et al. 2016) (see Fig. S1, Supplementary Materials) with the exception of East Asia which is based on level 3 regions (Southwest China, Southeast China, and Hainan Island). Key to codes for regions: WIO = West Indian Ocean; SA = Southern Africa; Aus = Australia; Br = Brazil; Ase1 = Indochina (continental Southeast Asia); Ase2 = Malesia (oceanic Southeast Asia); Ae = East Asia (southern China); As = India. See Table S1 for further details.

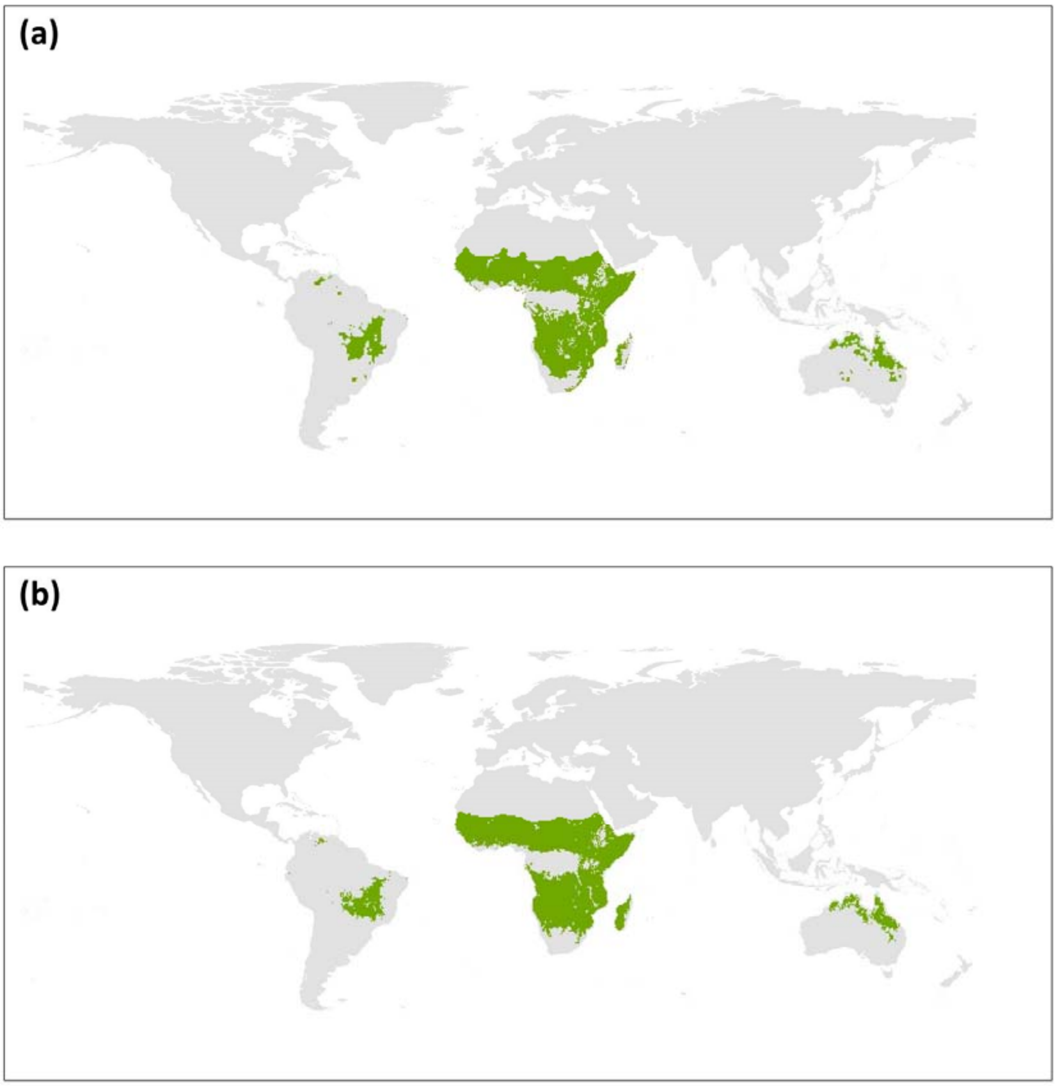

Figure S3 a) Observed distribution of savannas derived from vegetation maps for Africa (White 1983), Australia and South America (Lehmann et al 2011), and b) predicted distribution of savannas in the different continents using stochastic gradient boosting. The distribution of savannas was modeled separately for each continent based on climate, elevation and edaphic parameters, using the observed distribution map for that continent.

Table S1. Grass endemism and diversity as compared with other major savanna regions of the world calculated using TDWG regions (Clayton et al. 2016). Regions are all based on TDWG level 2 regions with the exception of East Asia, which is based on level 3 regions (see text). Size of regions were calculated using Eckert IV projections, following Vorontsova et al. 2016. Endemism of C_4 grasses in the Asian TDWG regions (Fig S1, Supplementary Materials), was estimated using the GrassBase database (Clayton et al. 2016) combined with a database of C_3/ C_4 pathways for grass taxa (Osborne et al. 2014).

Region	**Area**	**Total grass species**						
	(1000 km2)	**Total**	**C3**	**C3-C4**	**C4**	**Unknown**	**Chloridoideae**	**Andropogoneae**
Brazil	8525.2	1316	570	0	740	0	152	112
Australia	7722.0	1373	561	3	809	0	280	162
West Indian Ocean	603.8	612	189	1	396	26	134	95
Southern Africa	2682.0	915	334	3	577	1	228	99
Papuasia	908.2	479	198	1	280	0	54	102
New Zealand	269.4	413	346	0	67	0	17	10
East Asia (South China)	2777.5	1382	964	1	416	1	98	188
South Asia (Indian subcontinent)	4433.6	1350	618	1	731	0	158	371
Southeast Asia 1 (Indochina)	1936.7	804	291	1	508	4	106	254
Southeast Asia 2 (Malesia)	2133.8	699	302	1	396	0	81	178
		Endemic species						
		Total	**C3**	**C3-C4**	**C4**	**Unknown**	**Chloridoideae**	**Andropogoneae**
Brazil		476	275	0	201	0	31	18
Australia		768	293	1	474	0	191	72
West Indian Ocean		279	125	0	128	26	46	23
Southern Africa		267	169	1	96	1	43	3
Papuasia		118	101	0	17	0	2	5
New Zealand		152	149	0	3	0	2	1
East Asia (South China)		680	605	0	75	0	17	33
South Asia (Indian subcontinent)		447	185	0	262	0	39	183
Southeast Asia 1 (Indochina)		202	109	0	90	3	15	62
Southeast Asia 2 (Malesia)		201	150	0	51	0	6	35

Appendix 1: Methods for analysis of climate domains of Asian Savannas

Stochastic gradient boosting is an ensemble method for fitting statistical models that combines the strength of traditional statistical methods (decision trees) with machine learning techniques (boosting; Hastie et al 2001; Friedman et al. 2000, Friedman 2001, 2002). It works by iteratively building a number of small decision trees, each based on a random subset of the data, with each additional tree emphasizing observations poorly modeled by the existing collection of trees (Hastie et al 2001; Friedman et al. 2000, Lawrence et al. 2004, Elith et al. 2008). Finally, observations are assigned a class based on the most common classification amongst the trees (Lawrence et al. 2004, Elith et al. 2008). Gradient boosting is less sensitive to outliers and unbalanced data, is robust against overfitting, and has been shown to outperform many other classifiers (Friedman 2002, Lawrence et al. 2004).

We used existing vegetation maps to model the occurrence of savannas on different continents as a function of climate, elevation and edaphic parameters. For Africa, we used the continent-wide map developed by White (1983) to classify habitats as either savanna or non-savanna (see Sankaran et al. 2005), and for Australia and South America, we used the maps developed by Lehmann et al. (2011) (see Lehmann et al. 2011 for more details). Climate data were derived from the WorldClim climate database (Hijmans et al., 2005, http://www.worldclim.org), which provides data on 19 bioclimatic variables for the time period 1950-2000 (Hijmans et al., 2005), and soil nitrogen and percent clay from the ISRIC-WISE derived soil properties database (Batjes 2012, www.isric.org). Potential evapotranspiration (PET) estimates were obtained from the Global Potential Evapo-Transpiration (Global-PET) database (http://www.cgiar-csi.org/data/global-aridity-and-pet-database; Zomer et al 2007, Zomer et al. 2008). All data were resampled to a resolution of 0.5° for our analysis.

For each continent, we first generated a training dataset by systematically sub-sampling every 4th pixel to account for issues of spatial autocorrelation. These training datasets were used to build continent-specific boosting models, which were then used to predict the distribution of savannas across the entire continent. Our training data set included 2547 pixels for Africa (52.6% of which were savanna), 750 for Australia (15.2% savanna) and 1625 for South America (9.9% savanna). We evaluated the accuracy of the different models based on the fraction of savanna pixels correctly identified. We evaluated two separate models, one including all 19 climate variables along with soil and elevation parameters as predictors, and the second which only included a subset of climatic predictor variables that were largely uncorrelated with one another along with soil and elevation parameters. Model performance did not differ consistently between the two, and we only report results here from the model built using the subset of climatic predictors. Our final set of predictor variables included mean annual temperature, annual temperature range, mean temperature of the driest quarter, mean annual precipitation, precipitation of the driest month, precipitation seasonality, potential evapotranspiration, soil N and clay contents, and elevation. Finally, we used the models developed for each continent to individually predict the potential distribution of savannas in Asia. Although our models were built using data sampled at a resolution of 0.5°, our predictions for Asia were based on climate and soil data sampled at resolution of 5 arc-minutes (approx. 9 km × 9 km). All analyses were carried out using the 'caret' package (Liaw & Wiener, 2002) as implemented in R (R Core Team 2015).

Table S2. Model performance in terms of classification accuracies for Africa, Australia and South America.

	Africa	Australia	South America
Overall Accuracy	0.9028	0.9115	0.9529
Kappa statistic	0.804	0.6671	0.7347
Fraction of non-savanna pixels correctly classified	0.8806	0.9572	0.9792
Fraction of savanna pixels correctly classified	0.9218	0.6821	0.7248
Fraction of predicted non-savanna pixels that were actually non-savanna	0.9064	0.9378	0.9686
Fraction of predicted savanna pixels that were actually savanna	0.8998	0.7611	0.8004

Table S3. Area (in 1000 km^2) of Asia predicted to support savannas as a function of prediction probabilities based on models developed for Africa, Australia and South America.

Predicted probability of being savanna	Area (1000 km^2)		
	Africa	Australia	South America
Medium (0.5 – 0.75)	1686.2	945.9	435.8
High (0.75 – 0.9)	955.3	574.1	224.1
Very high (>0.9)	1118.3	266.4	106.2

References

Batjes, N. H. (2012). ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2). ISRIC.

Clayton, W. D., Vorontsova, M. S., Harman, K. T. & H, W. 2016 World Grass Species: Synonymy.http://www.kew.org/data/grasses-syn.html. [accessed 26 February 2016; 3:00 GMT].

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77: 802 - 813

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. The annals of statistics, 28(2), 337-407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International journal of climatology 25: 1965-1978.

Lawrence, R., Bunn, A., Powell, S., & Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sensing of Environment 90: 331-336.

Lehmann, C. E., Archibald, S. A., Hoffmann, W. A., & Bond, W. J. (2011). Deciphering the distribution of the savanna biome. New Phytologist 191(1): 197-209.

Osborne, C. P., Salomaa, A., Kluyver, T. A., Visser, V., Kellogg, E. A., Morrone, O., Vorontsova, M. S., Clayton, W. D. & Simpson, D. A. 2014 A global database of C4 photosynthesis in grasses. *New Phytol.* **204**, 441–446. (doi:10.1111/nph.12942)

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Sankaran, M et al. (2005). Determinants of woody cover in African savannas. Nature 438: 846-849.

Vorontsova, M. S. et al. 2016 Madagascar's grasses and grasslands: anthropogenic or natural? *Proc. R. Soc. B* (doi:10.1098/rspb.2015.2262)

White, F. (1983). The vegetation of Africa, a descriptive memoir to accompany the UNESCO/AETFAT/UNSO vegetation map of Africa (3 Plates, Northwestern Africa, Northeastern Africa, and Southern Africa, 1: 5,000,000).

Zomer RJ, Bossio DA, Trabucco A, Yuanjie L, Gupta DC & Singh VP, 2007. Trees and Water: Smallholder Agroforestry on Irrigated Lands in Northern India. Colombo, Sri Lanka: International Water Management Institute. pp 45. (IWMI Research Report 122).

Zomer RJ, Trabucco A, Bossio DA, van Straaten O, Verchot LV, 2008. Climate Change Mitigation: A Spatial Analysis of Global Land Suitability for Clean Development Mechanism Afforestation and Reforestation. Agric. Ecosystems and Envir. 126: 67-80.