Supplementary Information for:

A novel spatiophylogenetic modelling approach reveals evolutionary and ecological contributions to extinction risk in a diverse hotspot plant genus,

Russell Dinnage, Alex Skeels and Marcel Cardillo

Statistical Model Details

Mathematically, we can write out the equation of our full model as:

$$y_i \sim \text{Binomial}\left(p_i\right)$$

$$\text{Logit}\left(p_i\right) \ = \ \alpha \ + \ \beta_1 x_1[i] \ + \ \beta_2 x_2[i] \ + \ \beta_3 x_3[i] \ + \ ... \ + \ \beta_k x_k[i] \ + \ \mu_{\text{phylo}}[i] \ + \frac{1}{N_i}\sum_{j=1}^{N_i} f\left(\boldsymbol{s}_j[i]\right)$$

$$\mu_{\text{phylo}}[i] \sim \text{MVN}\left(\mathbf{0}, \sigma^2 \boldsymbol{V}_{\text{phylo}}\right)$$

where α is an intercept term, β_l is the lth fixed effect coefficient describing the effect of species-level predictor l of species i on the response, $\mathbf{0}$ is a vector of zeroes, σ^2 is a phylogenetic scaling factor, $\boldsymbol{V}_{\text{phylo}}$ is the standardised phylogenetic covariance matrix for all species in the model, $f\left(\boldsymbol{s}_j[i]\right)$ is a function describing the spatial effect for the jth longitude/latitude occurrence coordinates of species i: $\boldsymbol{s}_j[i]$, and N_i is the total number of observed occurrences of species i.
The spatial function can be expanded as:

$$f\left(\boldsymbol{s}_j\right) = \gamma_1 z_1[\boldsymbol{s}_j] + \gamma_2 z_2[\boldsymbol{s}_j] + \gamma_3 z_3[\boldsymbol{s}_j] + ... + \gamma_m z_m[\boldsymbol{s}_j] + \Psi\left(\boldsymbol{s}_j\right)$$

Where γ_m represents a regression coefficient determining the linear effect of environmental variable $z_m[\boldsymbol{s}_j]$, measured at coordinate $\boldsymbol{s}_j$, and $\Psi\left(\boldsymbol{s}_j\right)$ is a function describing the spatial random effect. The γ parameters can be moved out of $f\left(\boldsymbol{s}_j\right)$ because they are simple linear terms, and so the mean of their sum is the same as the sum of their means. So, if we set $\overline{z_l}[i] = \frac{1}{N_i}\sum_{j=1}^{N_i} z_l[\boldsymbol{s}_j[i]]$, we can simplify the model to:

$$p_i = \alpha + \beta_1 x_1[i] + \beta_2 x_2[i] + \beta_3 x_3[i] + ... + \beta_k x_k[i] + \gamma_1 \overline{z_1}[i] + \gamma_1 \overline{z_2}[i] + ... + \gamma_1 \overline{z_m}[i] + \mu_{\text{phylo}}[i] + \frac{1}{N_i}\sum_{j=1}^{N_i} \Psi\left(\boldsymbol{s}_j[i]\right)$$

Such that the γ terms are now fixed effect coefficients on the mean environmental variables for each species as a whole (e.g. the mean of the variables across all of a species' occurrence points). The $\Psi\left(\boldsymbol{s}_j\right)$ function is a basis function that approximates the spatial Matérn covariance function across a spatial mesh, using a stochastic partial differential equation approach (SPDE: Lindgren, Rue & Lindström 2011). Lindgren et al. (2011) has mathematical details of the SPDE approach, which we will not reproduce here.

We generated a mesh using the meshbuilder function in the INLA package. This runs a Shiny app (Chang *et al.* 2018) that allows the user to generate different meshes based on a set of parameters that can be modified interactively. We chose a mesh (Figure 1) that gave good coverage across the *Hakea* occurrence points, had good statistical diagnostics, and was large enough to avoid spatial overfitting (e.g was not too small for the choice of prior – see below).

INLA uses a set of weights to calculate the spatial random effects at the coordinates of the data. The weights interpolate between the three closest mesh points, and are encapsulated in a matrix (the $\boldsymbol{A}_{\text{spatial}}$ matrix) with N_{point} rows (where N_{point} is the number of occurrence points), and N_{mesh} columns (where N_{mesh} is the number of mesh points), where the weights in each row sum to 1. In order to calculate the average spatial effect across each species' occurrence points (e.g. $\frac{1}{N_i}\sum_{j=1}^{N_i}\Psi\left(\boldsymbol{s}_j[i]\right)$) we calculated a new $\boldsymbol{A}_{\text{species:spatial}}$ matrix with N_i rows and N_{mesh} columns, where each row was simply the mean of the rows in $\boldsymbol{A}_{\text{spatial}}$ corresponding to the occurrence points of each species i. Again, weights in all rows sum to 1 in this new $\boldsymbol{A}$ matrix.

Choosing Bayesian priors

INLA requires priors on all parameters to be specified. For fixed parameters we used the default INLA prior, a wide gaussian prior with mean = 0 and variance = 100. For the phylogenetic and spatial random effects we used weakly informative priors, as recommended by Simpson *et al.* (2017) and Gelman *et al.* (2008). For the phylogenetic scaling parameter we used the 'pcprior' distribution in INLA with parameters 1 and 0.1, corresponding to an exponential distribution with about 10% of its probability distribution >1. To test the sensitivity of our analysis to the choice of prior, we ran the full model with several different prior parameters ([1, 0.01], [1, 0.1], [1, 0.5], [1, 0.99]) representing distributions with increasingly heavy tails, the last of which approximates an uninformative uniform distribution. The choice of prior had very little effect on any other parameter estimates, and all qualitative results were identical, so for all subsequent analyses we used [1, 0.1].

The priors on the spatial random effect (which include the range and σ parameter of the INLA implementation of the (Rasmussen & Williams 2006) were chosen as follows. Illian et al. (2012) recommend choosing priors on the range parameter (representing the spatial range over which the covariance decays to almost zero) that avoid spatial overfitting, by placing most of the prior density on range values greater than the apparent covariance range of the environmental factors used in the model. Allowing values much less than this can result in a spatial random effect that overfits on a very fine spatial scale, which will explain away any environmental factors, and lead to poor predictions for new data. By choosing a prior that enforces a similar spatial covariance in the spatial random effect and fixed environmental factors, we allow the model to more appropriately compare between them and choose the most parsimonious decomposition of the effects. We chose a prior through trial and error by fitting only the spatial random effect to our data, then comparing a map of the result to maps of our environmental factors until we found a set of priors where the random effect map showed a similar spatial covariance to the environmental factor maps. For the range parameter, a 'pcprior' with 10% of its density <2 decimal degrees resulted in appropriate covariance structure. Any value >2 in the prior resulted in very similar results that avoided overfitting (as the estimated range in the model was considerably greater than 2; see

Results). On the σ parameter we used a ‘pcprior’ distribution with values [1, 0.01]. The α parameter, which controls the ‘smoothness’ of the Matérn covariance function was fixed to 2, which is a standard choice in spatial modelling.

Concluding notes on spatiophylogenetic modelling

The Bayesian spatiophylogenetic method developed here solves a general problem in comparative analysis and so may be useful for addressing a variety of different questions. Many questions in comparative biology are potentially influenced by the spatial arrangement of species, yet spatial effects are still not routinely incorporated into analytical models in the way that phylogenetic effects are. This is likely due to a scarcity of detailed spatial data (at least until fairly recently) and a lack of appropriate methods to incorporate both spatial and phylogenetic effects into comparative analyses. Our approach allows both kinds of effects to be modelled simultaneously, and also offers a potential solution to another major issue for spatially explicit comparative methods, the mismatch in the levels of measurement of spatial variables and species variables. Here we have shown how the method can show spatially-explicit insights into extinction threats not possible otherwise, and allow for a more nuanced and careful exploration of comparative data on species threat status.

Supplementary Figures

Figure S1. Bayesian marginal posterior distributions of model parameters when run using alternative method of calculating ED ("equal splits"). The top panel shows the estimated standard deviation of the two random effects as calculated at the species-level. The bottom panel shows the fixed effect linear regression coefficients. Because all fixed effect variables were standardised, these represent standardised coefficients, and are comparable to one another. Posterior distributions that were categorised as substantial effects, (95% credible intervals do not overlap zero) are plotted as blue; otherwise red.The mean of the posteriors are plotted as points along with errorbars representing the 95% credible interval.

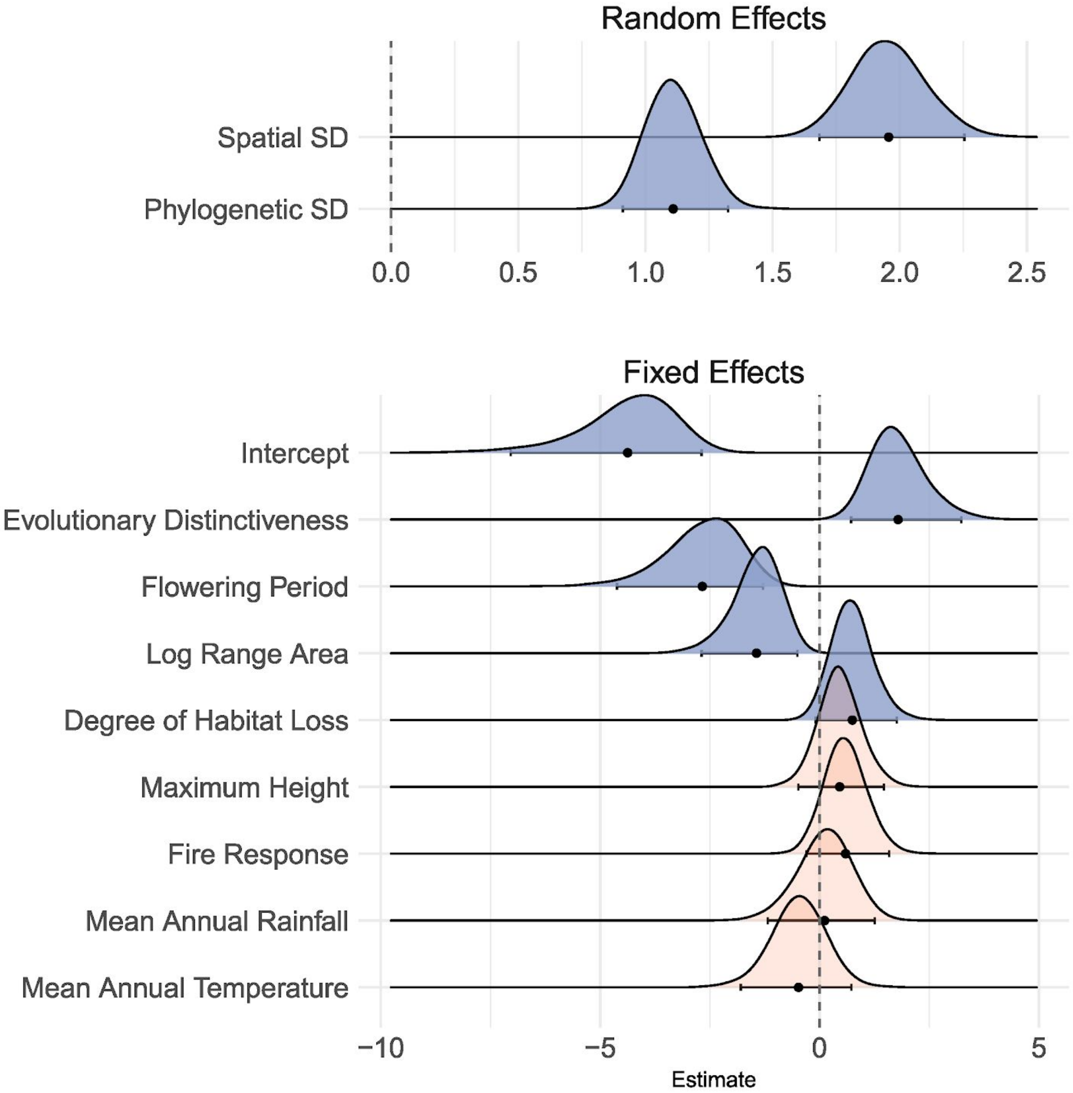


Figure S2. Comparison of model coefficients between our Spatiophylogenetic model, but with spatial effects removed (using the mode of the posterior distribution as the estimate), and a standard phylogenetic binomial regression (in this case implemented in the R package phyr, using maximum likelihood estimates). This shows our model collapses to a standard phylogenetic model when we remove the spatial component, which we argue suggests it should share properties of this well-tested class of models. Dotted line is a fitted line through the coefficient estimates of the two models.

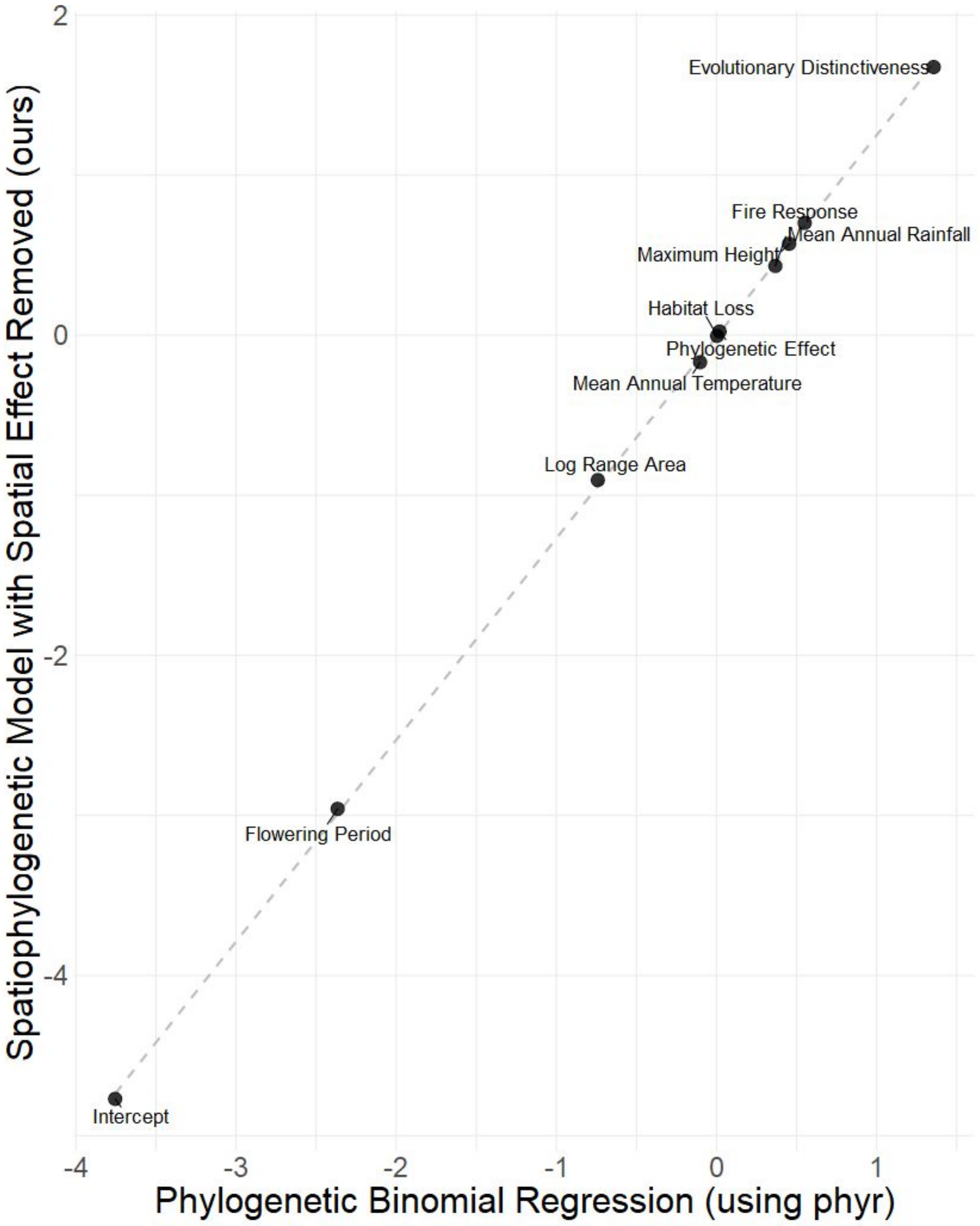


Figure S3. The distribution of model coefficients across models run for each of 200 generated phylogenies, where for each phylogeny all of the 15 missing species were placed randomly within a clade that corresponds to its taxonomy. Specifically, each species was placed randomly within the clade that subtends the node corresponding to the most recent common ancestor of all other species found in same taxonomic group as the missing species (according to Barker *et al.* (1999)). Coefficients were summarised for each model by the median of its marginal posterior distribution.

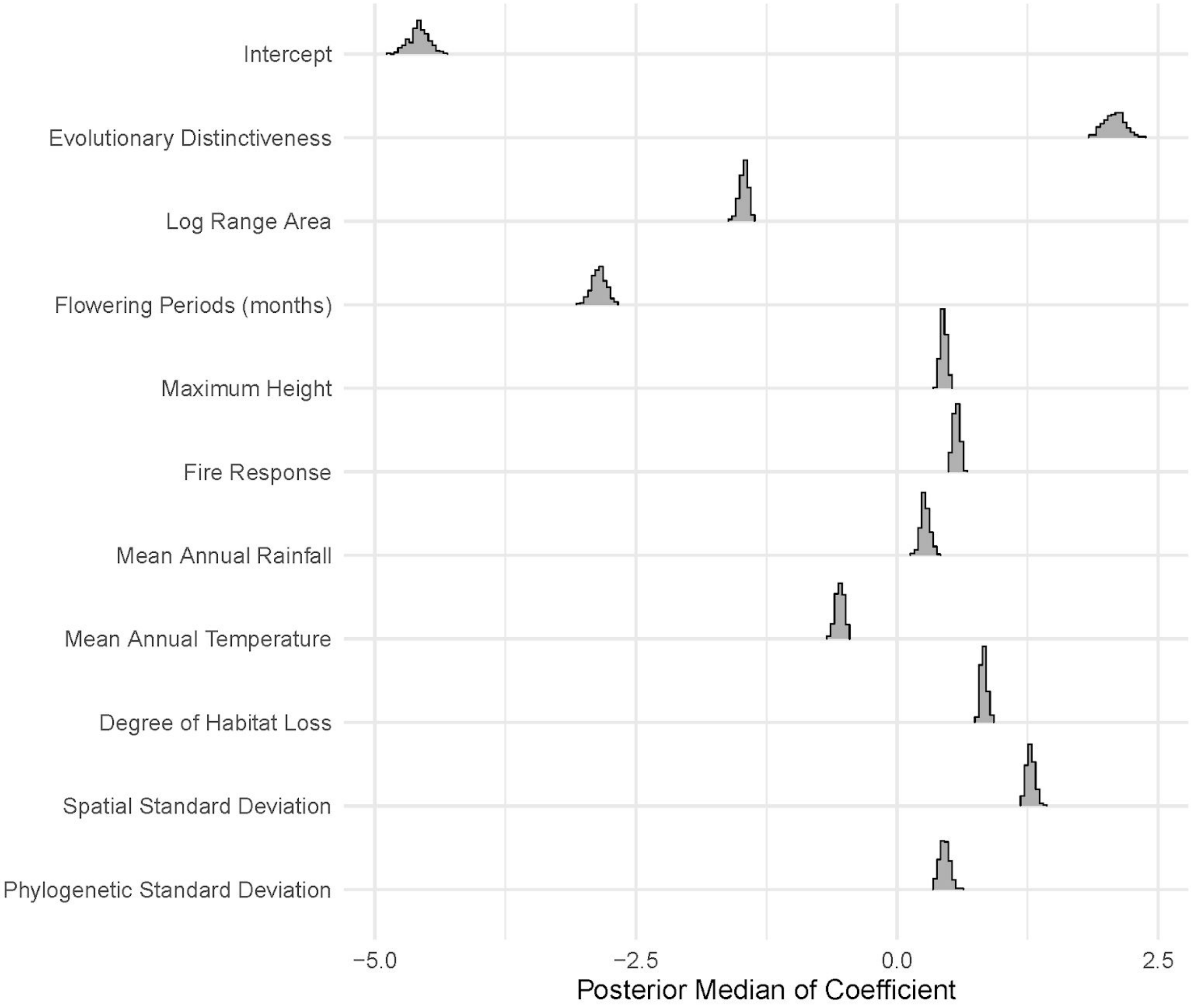


Figure S4. Predicted threat probability according to the spatiophylogenetic model described in this study (X axis: model with space) vs. predicted threat probability according to the phylogenetic only model (Y axis: model without space). Points with predicted probabilities greater than 0.3 according to the model with space are labelled with species names (this is where the two models differ most). Thick black line is the 1:1 line, where predictions from both models are the same. The dotted line is a fitted line through the predictions. The model without space underpredicts relative to model with space, especially at higher predicted probabilities.

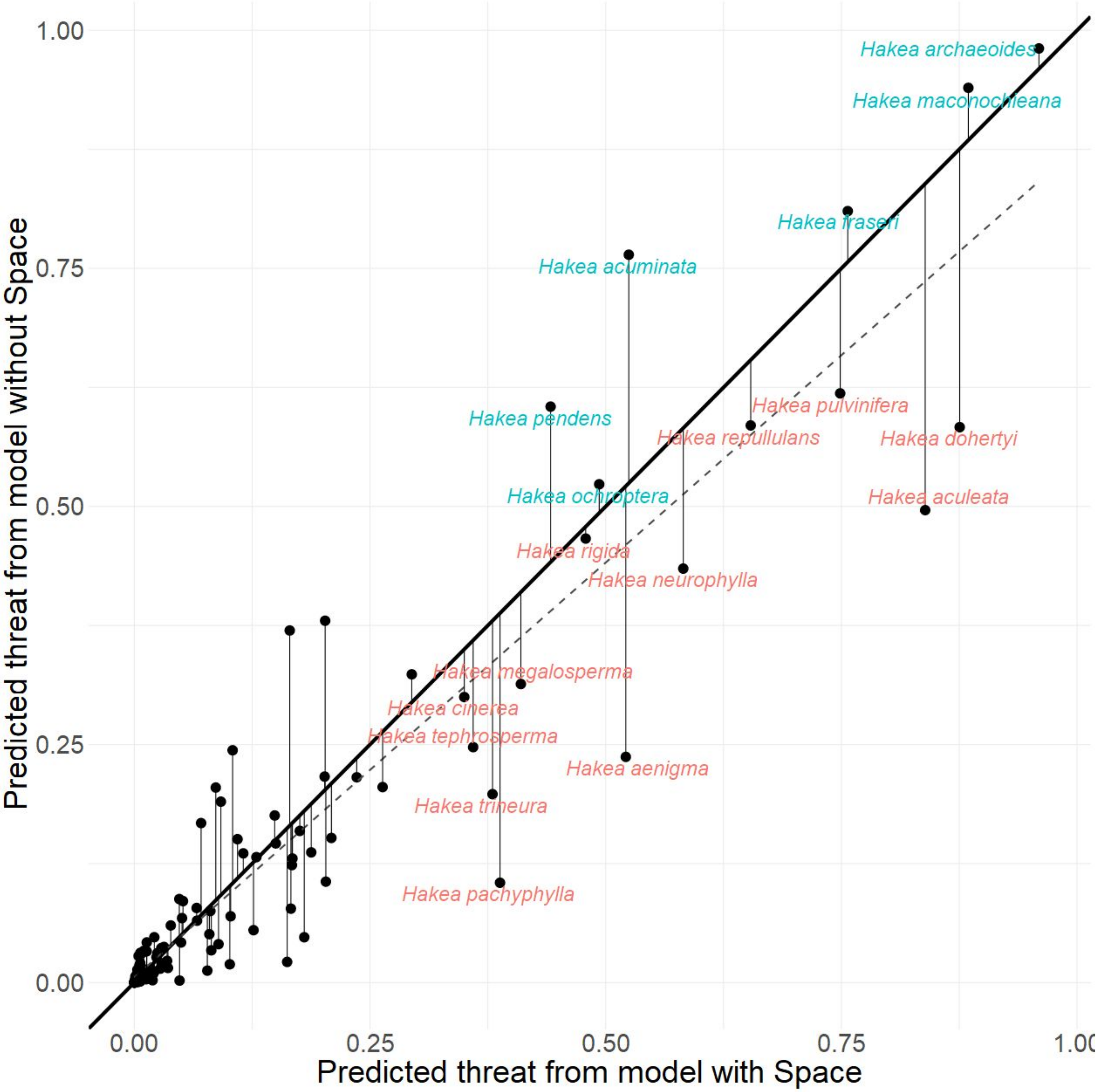


Figure S5. Comparison of the spatiophylogenetic model presented here with a non-spatial phylogenetic model. We used the pglmm.compare function in the R package phyr to fit the model. Points represent approximate effects sizes for each fixed effect. Effect size were calculated as the estimate divided by the standard error for the maximum likelihood based phylogenetic model, and the mean of the posterior distribution divided by the standard deviation of the posterior distribution for the Bayesian spatiophylogenetic model. Arrows point from the non-spatial model to the model with space. Opaque points have an effect size larger than an arbitrary 90% confidence according to a Z distribution. All effects are very similar apart from degree of habitat loss, and log range area. Factors with known spatial structure also change somewhat more than others (e.g. mean annual temperature and mean annual rainfall).

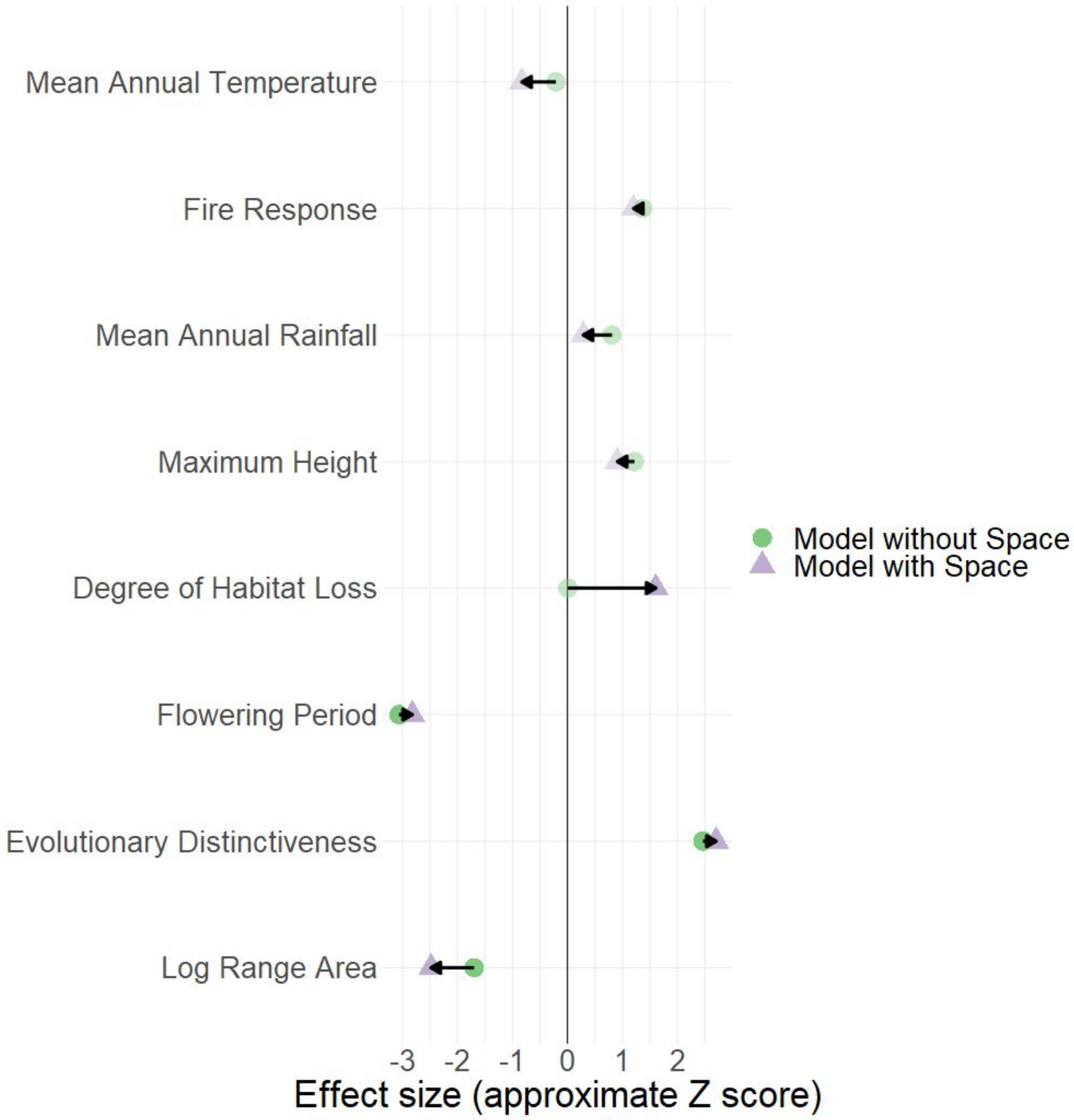


Figure S6. Species occurrence points for a) species currently “threatened”, and b) species the model classified as “of concern” – (not currently threatened but with predicted risk >0.20). Each occurrence point is coloured according to the species’ predicted risk.

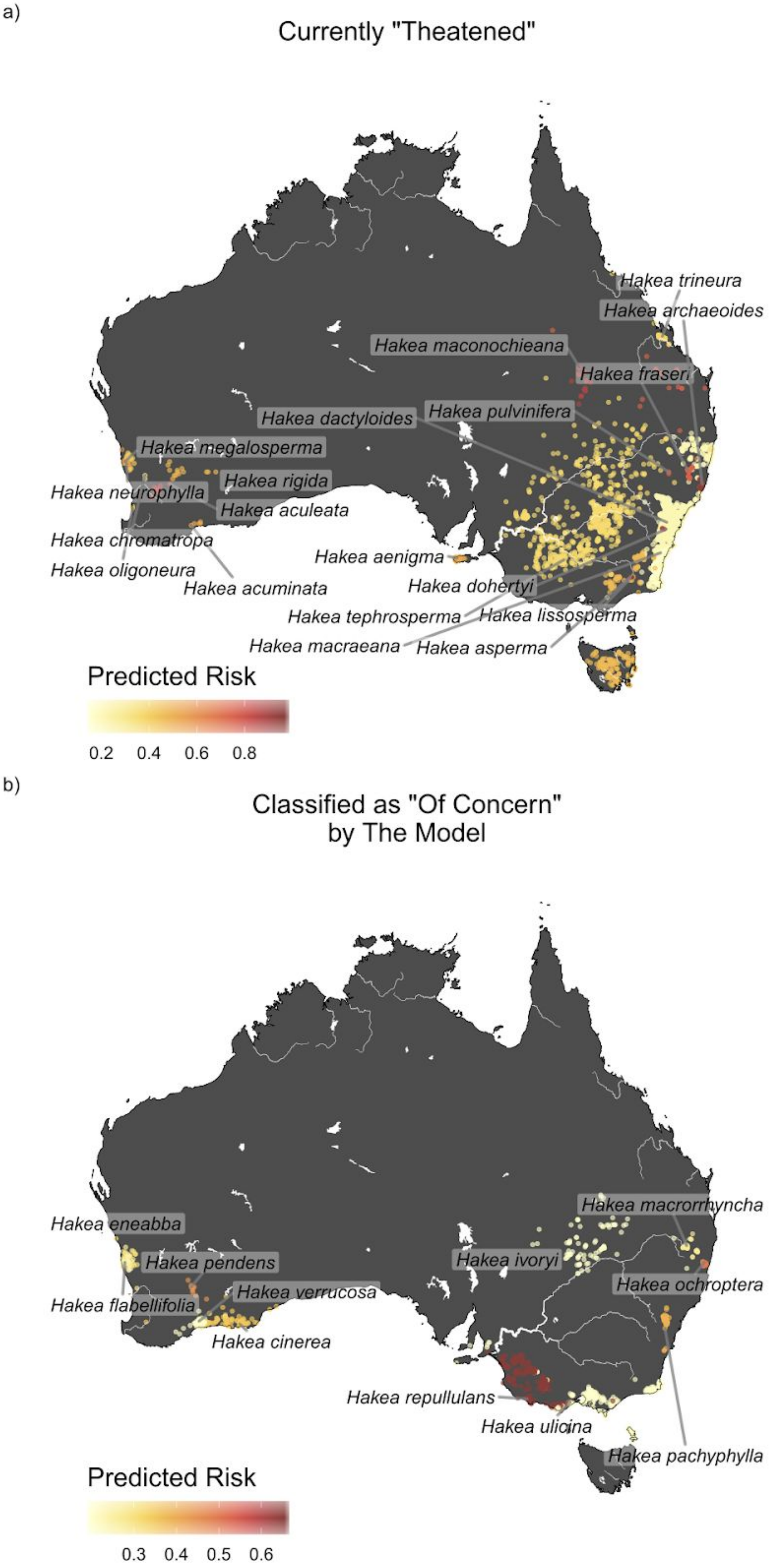

a)
Currently "Theatened"
Hakea trineura
Hakea archaeoides
Hakea maconochieana
Hakea fraseri
Hakea pulvinifera
Hakea dactyloides
Hakea megalosperma
Hakea rigida
Hakea neurophylla
Hakea aculeata
Hakea chromatropa
Hakea oligoneura
Hakea acuminata
Hakea aenigma
Hakea dohertyi
Hakea tephrosperma
Hakea lissosperma
Hakea macraeana
Hakea asperma
Predicted Risk
0.2
0.4
0.6
0.8
b)
Classified as "Of Concern"
by The Model
Hakea macrorrhyncha
Hakea eneabba
Hakea ivoryi
Hakea pendens
Hakea ochroptera
Hakea verrucosa
Hakea flabellifolia
Hakea cinerea
Hakea repullulans
Hakea ulicina
Hakea pachyphylla
Predicted Risk
0.3
0.4
0.5
0.6

Figure S7. Predicted threat of all *Hakea* species from a spatiophylogenetic model, broken down into different factors determined to be substantial by the model. The panels from left to right are: 1) the *Hakea* species and their phylogeny, 2) the decomposition of the predicted threat into contributions from the different “substantial” factors. Predicted Threat is calculated as the predicted deviation from the average threat for each species when taking into account each risk factor independently, whilst holding other factors in the model constant (setting them to zero in the linear predictor). Error bars represent 95% credible intervals. Threat decomposition in the right panel is split by positive and negative effects, such that stacked bars below the zero line show the relative contribution to deviations below the average, and those above the zero line show the contributions to deviations above the average. The overall predicted mean deviation from average threat is shown by a white bar.

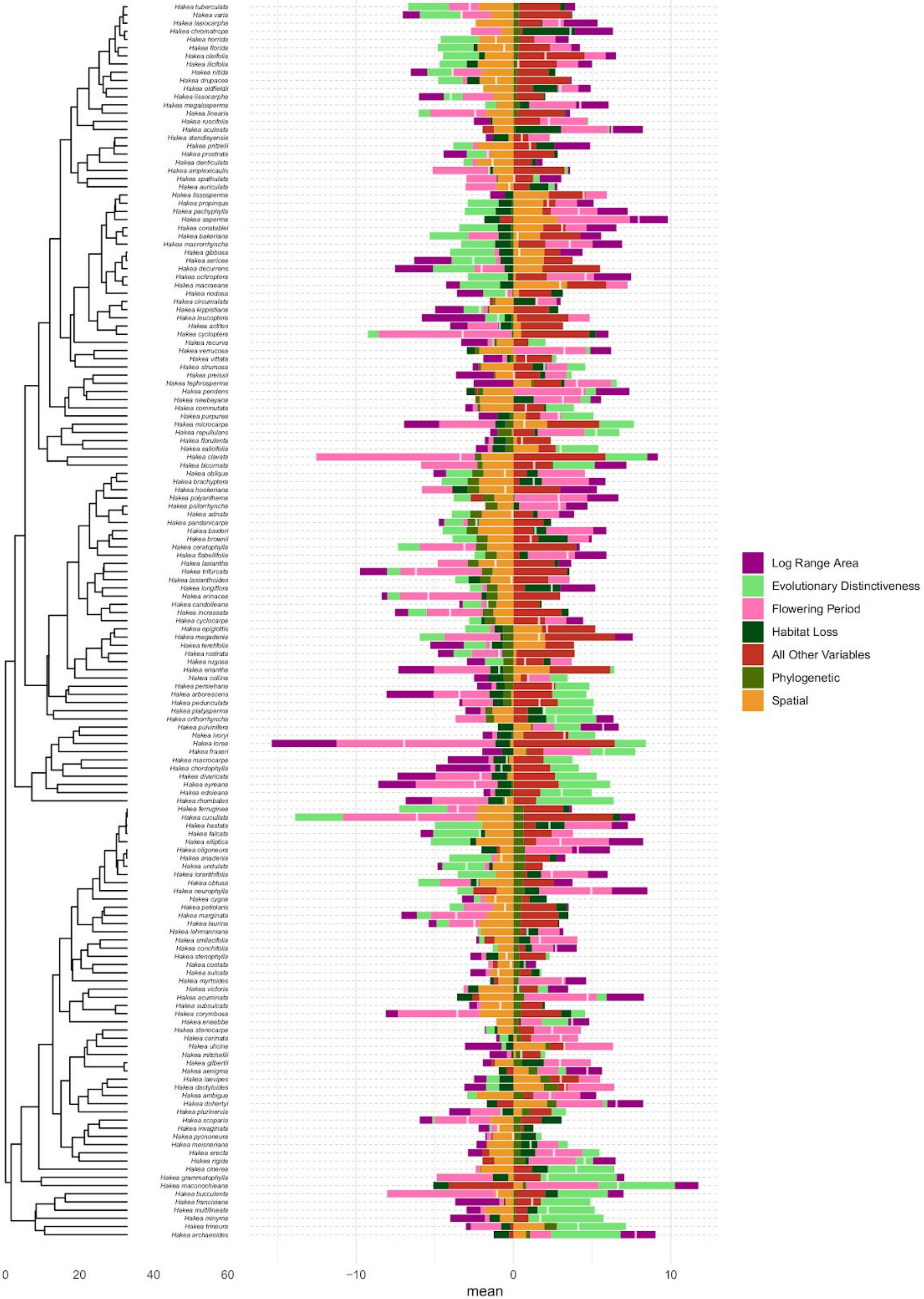

Hakea tuberculata
Hakea varia
Hakea lasiocarpha
Hakea chromatropa
Hakea horrida
Hakea florida
Hakea oleifolia
Hakea ilicifolia
Hakea nitida
Hakea drupacea
Hakea oldfieldii
Hakea lissocarpha
Hakea megalosperma
Hakea linearis
Hakea ruscifolia
Hakea aculeata
Hakea standleyensis
Hakea pritzelii
Hakea prostrata
Hakea denticulata
Hakea amplexicaulis
Hakea spathulata
Hakea auriculata
Hakea lissosperma
Hakea propinqua
Hakea pachyphylla
Hakea asperma
Hakea constablei
Hakea bakeriana
Hakea macrorrhyncha
Hakea gibbosa
Hakea sericea
Hakea decurrens
Hakea ochroptera
Hakea macraeana
Hakea nodosa
Hakea circumalata
Hakea kippistiana
Hakea leucoptera
Hakea actites
Hakea cycloptera
Hakea recurva
Hakea verrucosa
Hakea vittata
Hakea strumosa
Hakea preissii
Hakea tephrosperma
Hakea pendens
Hakea newbeyana
Hakea commutata
Hakea purpurea
Hakea microcarpa
Hakea repullulans
Hakea florulenta
Hakea salicifolia
Hakea clavata
Hakea bicornata
Hakea obliqua
Hakea brachyptera
Hakea hookeriana
Hakea polyanthema
Hakea psilorrhyncha
Hakea adnata
Hakea pandanicarpa
Hakea baxteri
Hakea brownii
Hakea ceratophylla
Hakea flabellifolia
Hakea lasiantha
Hakea trifurcata
Hakea lasianthoides
Hakea longiflora
Hakea erinacea
Hakea candolleana
Hakea incrassata
Hakea cyclocarpa
Hakea epiglottis
Hakea megadenia
Hakea teretifolia
Hakea rostrata
Hakea rugosa
Hakea eriantha
Hakea collina
Hakea persiehana
Hakea arborescens
Hakea pedunculata
Hakea platysperma
Hakea orthorrhyncha
Hakea pulvinifera
Hakea ivoryi
Hakea lorea
Hakea fraseri
Hakea macrocarpa
Hakea chordophylla
Hakea divaricata
Hakea eyreana
Hakea ednieana
Hakea rhombales
Hakea ferruginea
Hakea cucullata
Hakea hastata
Hakea falcata
Hakea elliptica
Hakea oligoneura
Hakea anadenia
Hakea undulata
Hakea loranthifolia
Hakea obtusa
Hakea neurophylla
Hakea cygna
Hakea petiolaris
Hakea marginata
Hakea laurina
Hakea lehmanniana
Hakea smilacifolia
Hakea conchifolia
Hakea stenophylla
Hakea costata
Hakea sulcata
Hakea myrtoides
Hakea victoria
Hakea acuminata
Hakea subsulcata
Hakea corymbosa
Hakea eneabba
Hakea stenocarpa
Hakea carinata
Hakea ulicina
Hakea mitchellii
Hakea gilbertii
Hakea aenigma
Hakea laevipes
Hakea dactyloides
Hakea ambigua
Hakea dohertyi
Hakea plurinervia
Hakea scoparia
Hakea invaginata
Hakea pycnoneura
Hakea meisneriana
Hakea erecta
Hakea rigida
Hakea cinerea
Hakea grammatophylla
Hakea maconochieana
Hakea bucculenta
Hakea francisiana
Hakea multilineata
Hakea minyma
Hakea trineura
Hakea archaeoides
0
20
40
60
−10
0
10
mean
Log Range Area
Evolutionary Distinctiveness
Flowering Period
Habitat Loss
All Other Variables
Phylogenetic
Spatial

Figure S8. Maps of "non-spatial" independent risk factors. These maps plot the spatial random effects estimated on the spatial mesh for Australia (Figure 1b in main text), for the independent predicted risks associated with different risk factors. See methods for details on the calculations. The maps can be interpreted as a continuous approximation of how each risk factor is distributed across Australia, where risk factors associated with species traits or phylogeny is derived from the distributions of the species that they affect. For example, the phylogenetic based risk map shows where species in the most high risk clades tend to occur in Australia. The result for Evolutionary Distinctiveness (ED) requires some additional explanation, as it would appear to suggest high risk in central Australia, where there are currently no at risk *Hakea* species. There are two things to keep in mind when interpreting this result. 1) Recall that predictions based on individual factors are made while *holding all other factors constant.* Given this we would not necessarily expect predictions from every factor to match well with the observed data, because in reality, the factors are not independent. Even though ED-based risk is high in central Australia, all other risk factors predict low risk in central Australia, which is why overall there is low risk in central Australia. If some of these other factors were to change in central Australia, however, the ED-based effect, if real, could become "unmasked", and so this result is still important. 2) This plots the mean of the posterior predictions, and so does not show model uncertainty. In fact, uncertainty around the central Australia effect of ED is quite high (see figure S4 for a plot of uncertainty in each risk factor's independent predictions). This uncertainty is likely driven by the fact that there are only a few species in central Australia and so their high ED-based risk estimates are likely the result of extrapolation from a trend seen in more abundant lower-ED species.

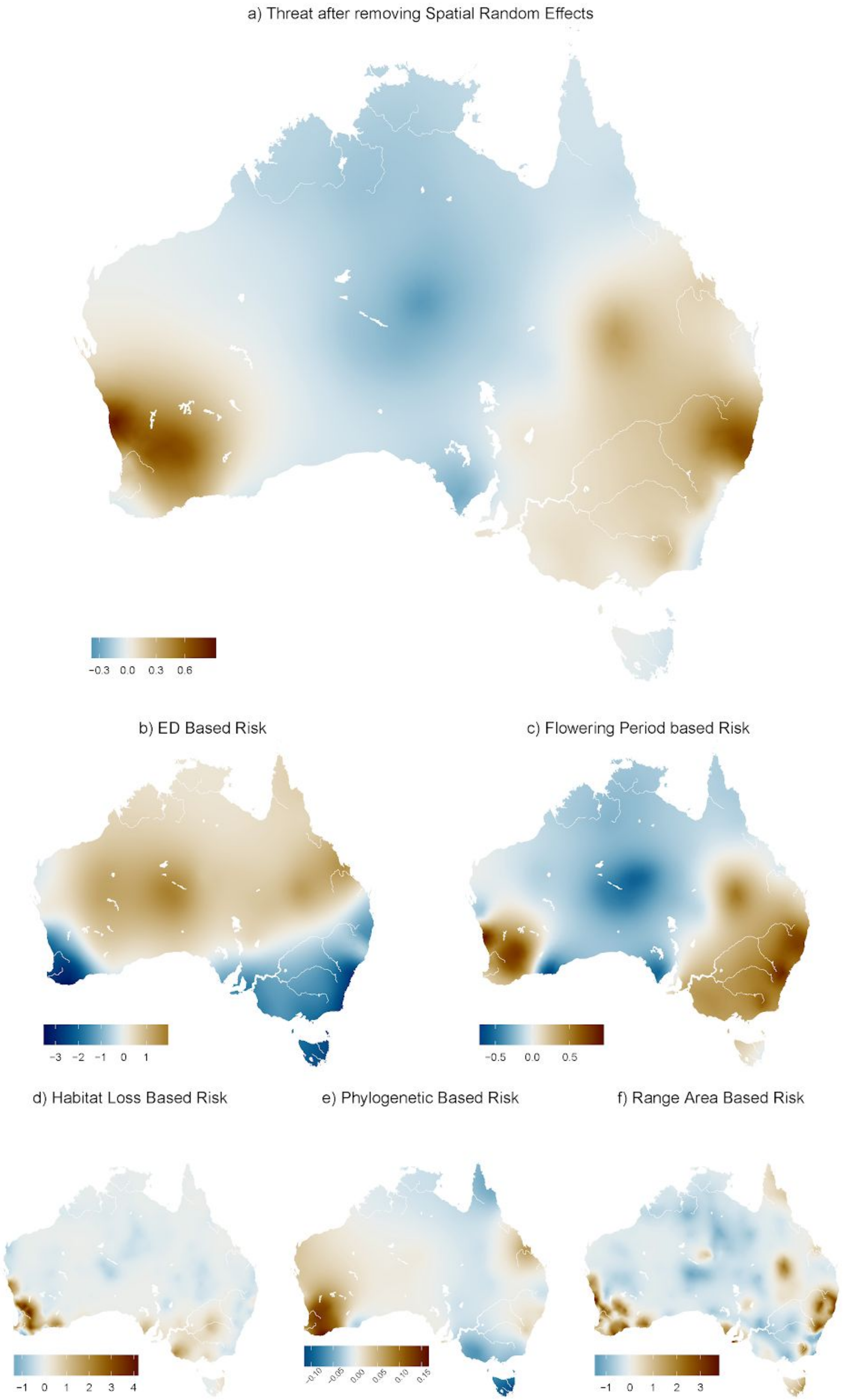
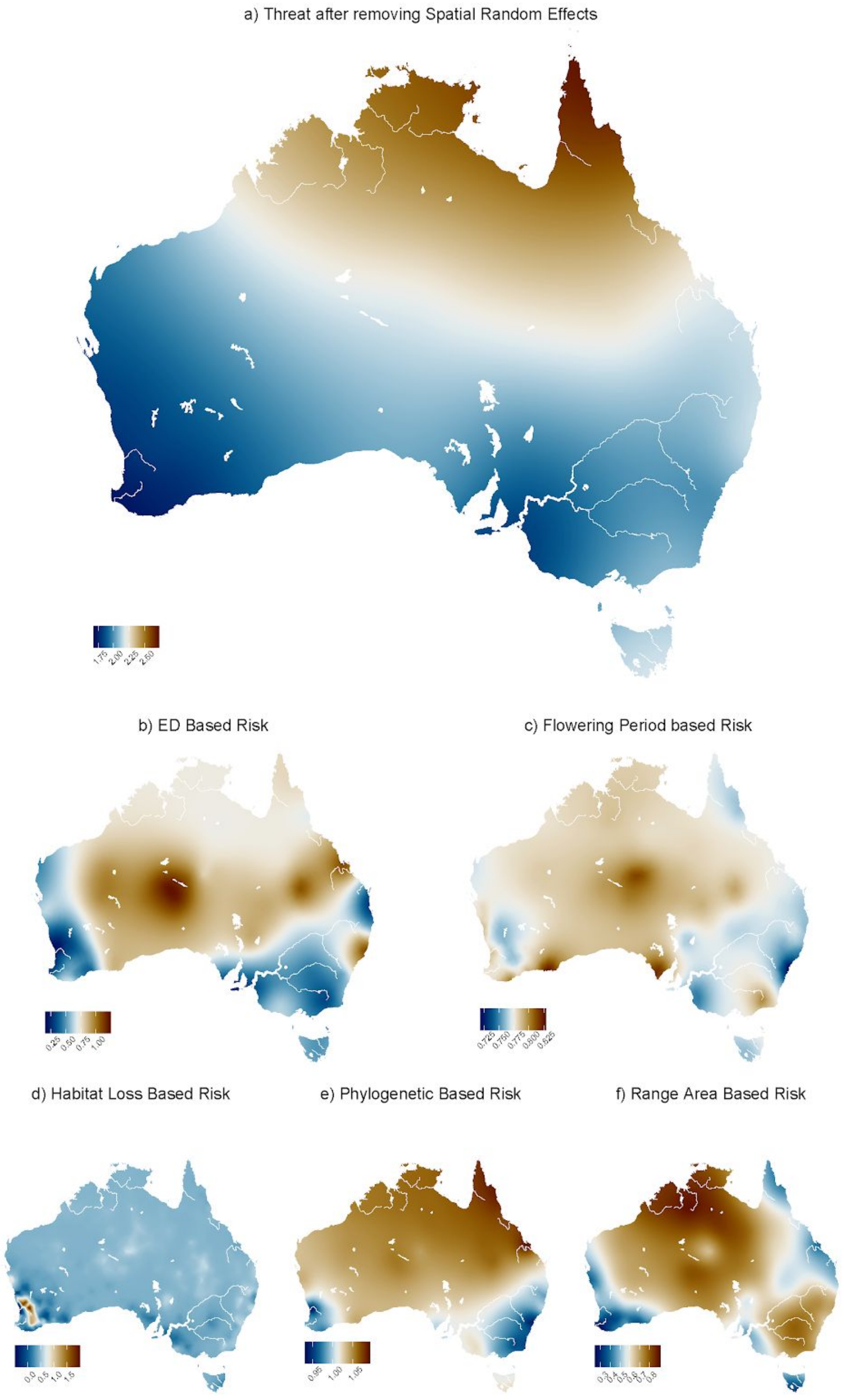

a) Threat after removing Spatial Random Effects
−0.3 0.0 0.3 0.6
b) ED Based Risk
−3 −2 −1 0 1
c) Flowering Period based Risk
−0.5 0.0 0.5
d) Habitat Loss Based Risk
−1 0 1 2 3 4
e) Phylogenetic Based Risk
−0.10 −0.05 0.00 0.05 0.10 0.15
f) Range Area Based Risk
−1 0 1 2 3

Figure S9. Maps of "non-spatial" risk factor uncertainty. These maps plot the uncertainty of the spatial random effects estimated on the spatial mesh for Australia (Figure 1b in main text), for the same risk factors shown in Figure S3. The maps can be interpreted as a continuous approximation of how the uncertainty in each risk factor is distributed across Australia, where uncertainty is measured as the standard deviation of the posterior predictions.



a) Threat after removing Spatial Random Effects

b) ED Based Risk

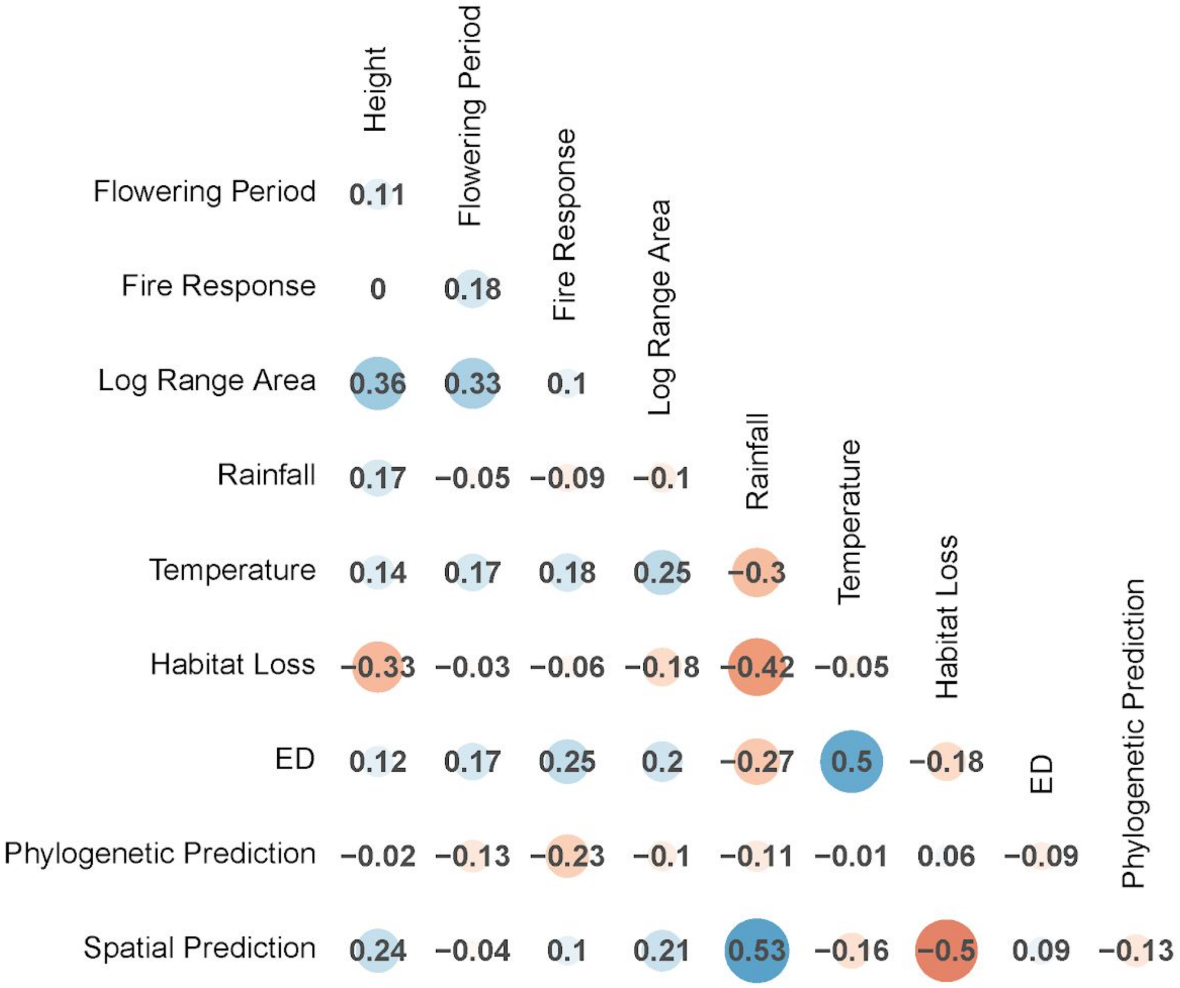
c) Flowering Period based Risk

d) Habitat Loss Based Risk

e) Phylogenetic Based Risk

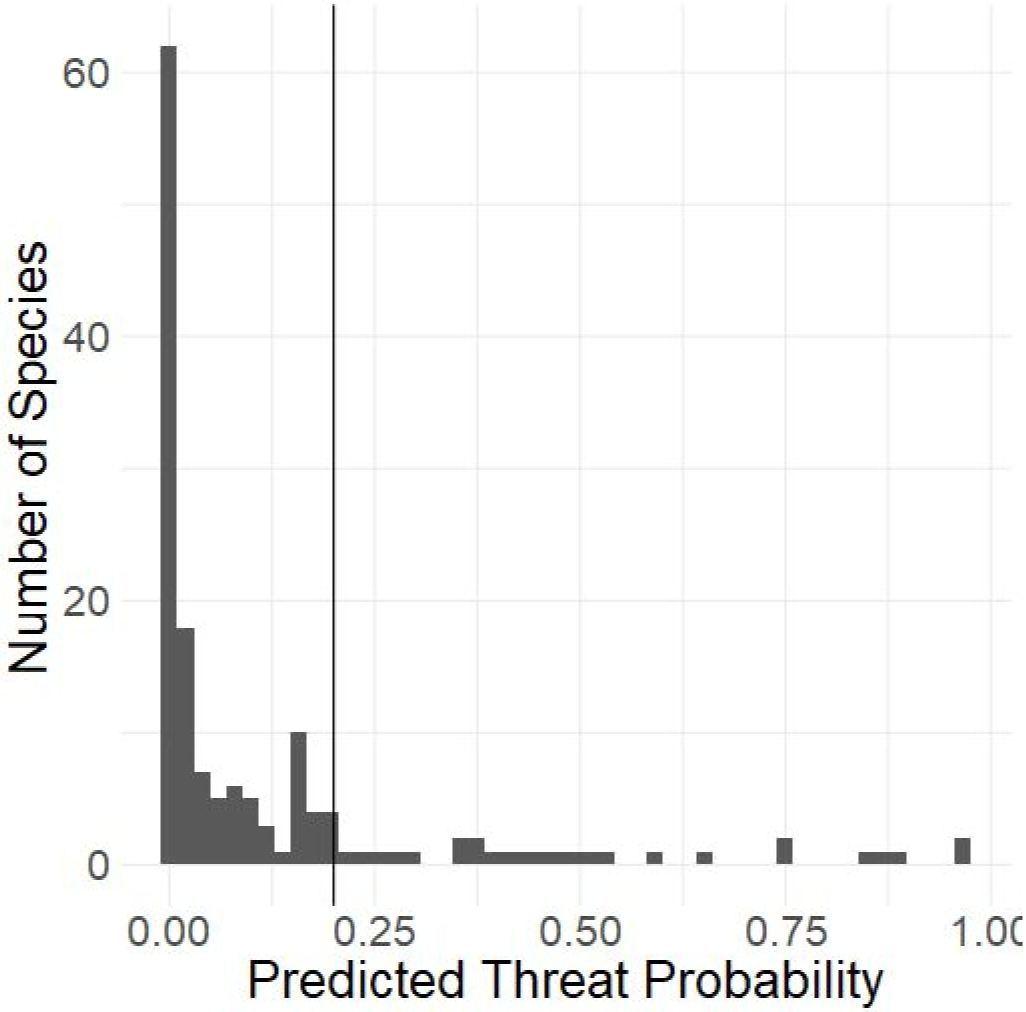
f) Range Area Based Risk

Figure S10. Correlation matrix of all predictors in the model. Additionally included are the phylogenetic and spatial predictions from the model.



	Height	Flowering Period	Fire Response	Log Range Area	Rainfall	Temperature	Habitat Loss	ED	Phylogenetic Prediction
Flowering Period	0.11								
Fire Response	0	0.18							
Log Range Area	0.36	0.33	0.1						
Rainfall	0.17	−0.05	−0.09	−0.1					
Temperature	0.14	0.17	0.18	0.25	−0.3				
Habitat Loss	−0.33	−0.03	−0.06	−0.18	−0.42	−0.05			
ED	0.12	0.17	0.25	0.2	−0.27	0.5	−0.18		
Phylogenetic Prediction	−0.02	−0.13	−0.23	−0.1	−0.11	−0.01	0.06	−0.09	
Spatial Prediction	0.24	−0.04	0.1	0.21	0.53	−0.16	−0.5	0.09	−0.13

Figure S11. Histogram of predicted probabilities for all Hakea species based on our model. Vertical line shows the cutoff of 0.2 over which we used to determine if a species is "of concern" in a conservation context (if it was not already known to be threatened).



References

Chang, W., Cheng, J., Allaire, J.J., Xie, Y. & McPherson, J. (2018) shiny: Web Application Framework for R.

Gelman, A., Jakulin, A., Pittau, M.G. & Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, **2**, 1360–1383.

Illian, J.B., Sørbye, S.H. & Rue, H. (2012) A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The annals of applied statistics*, **6**, 1499–1530.

Lindgren, F., Rue, H. & Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Rasmussen, C.E. & Williams, C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass.

Simpson, D., Rue, H., Riebler, A., Martins, T.G. & Sørbye, S.H. (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, **32**, 1–28.