Dear Dr. Carsten Dormann and Colleagues,

Thank you for your comments on Liang et al. 2016. It is always stimulating when someone is discussing our findings. There are many interesting questions you raise, and others neither you nor we have yet wrestled with fully. Please find, below, our response to your comments as numbered on Page 1.

(1) *The authors computed "relative tree species richness" in such a way that it represents a gradient from boreal to tropical plots, rather than in local species richness. When instead computing species richness relative to the maximum value in the region the effect of species richness on productivity is dramatically reduced.*

Response: Thank you for your suggestion in your first sentence. However, confining our analysis strictly at the ecoregion level would render us unable to derive a true global biodiversity-productivity relationship (BPR) which should account for both intra- and inter-ecoregion variability. There are likely a variety of different ways of assessing this; ours and yours are just two. Considering mounting concerns on the delineation of ecoregion boundaries (e.g. Jepson and Whittaker 2002), an ecoregion-level study would create substantial problems of its own. Thus we believe both options (yours and ours) have strengths and weaknesses, and address the same overall question but from different angles. There are many other issues that could be, and should be addressed, in grappling with how best to do this. This includes whether productivity should be standardized (i.e. the issues raised for richness might also apply in some way for productivity); and how best to standardize either richness or productivity (as there a number of ways of doing this). We are working on delving further into these issues.

Regarding the point in the second sentence, we disagree that the BPR relationship is dramatically reduced when examined at eco-regional scales. We will demonstrate below that even when we use *relative tree species richness* at an ecoregion-level, the trendline and standard error bands are similar to the global trend as reported by Liang et al. 2016.

For this demonstration, we selected the three grassland biomes (i.e. Montane Grasslands and Shrublands, Flooded Grasslands and Savannas, and Temperate Grasslands, Savannas and Shrublands), because your graphs in Page 32 suggest that these biomes do not conform to the global trend of Liang et al. 2016. For this analysis, we combined the three biomes together, because there are less than 2000 plots for Montane Grasslands and Shrublands and Flooded Grasslands and Savannas together, and almost a half of the plots within these two ecoregions are monocultures

The combined grassland biomes have a total of 23,133 plots (including ~3000 monoculture plots). For simplicity, we ignored the spatial autocorrelation, and the result from a robust bootstrapping estimation (Efron and Tibshirani, 1993) is quite consistent with the global trend of Liang et al. 2016 (Fig. B1) (see the **Appendix** for the R script for estimating BPR for the grassland biomes). This is also generally true for most of the other ecoregions (not shown), as long as there are a sufficient number of plots and a sufficient number of mixed-species plots. In fact, the theta values we have produced to date across regions don't systematically differ from the global one, although we are still working on making sure we are doing these appropriately. So we are unclear how you arrived at the values you did. Additionally, we also think that perhaps we (and you or anyone else working with these data) should eventually re-run everything at a forest type-level, as ecoregion is a poor way to delineate forests. For instance, we have plots from the desert ecoregion with near zero productivity, which nonetheless are identified by local foresters as forests.

We acknowledge that performing an ecoregion-level study would be a good supplement to Liang et al. 2016. We would be glad to collaborate with you or anyone else on this idea. Additionally we believe that examining alternative approaches, including non-parametric models, and different ways of standardizing either or both productivity and richness, to the global relationship would be worth doing.

We also note that we have some residual questions about your approach. We are unable to understand how a global line like yours (your left panel, Figure 1) could average and max out around 2.5 for productivity when so many of the Ecoregions with most of the data have means so much higher than that? Additionally, you call the x-axis of your first panel in Figure 1 ”relative local species richness” which confuses us. If your draws were across all data, then the ‘relative’ value is not ‘local’ even if you used the maximum values of each draw rather than the global max as we did (but we are not entirely sure what you did). If the maximum richness was from each draw, should your x-axis be “sample max” not “local max”. Are we misinterpreting what you did or is this just unclearly labeled?

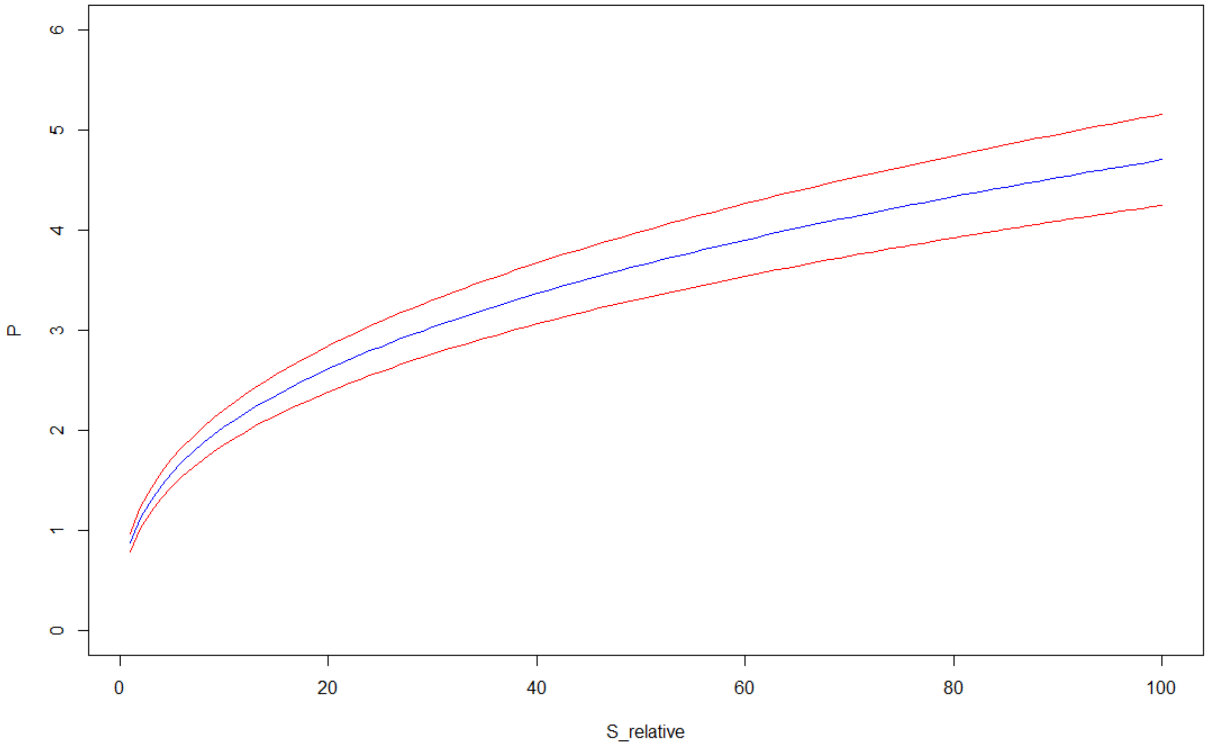

*Figure B1. Estimated BPR curve (with 95% confidence interval bands), using an ordinary least squares (OLS) model, based on the three grassland biomes (i.e. Montane Grasslands and Shrublands, Flooded Grasslands and Savannas, and Temperate Grasslands, Savannas and Shrublands). We converted species richness (S) to relative species richness (S_hat): S_hat = S *100 / 271.*

(2) *Plots are overwhelmingly from temperate forest; indeed only some 2500 plots are from the tropics (equivalent to 0.4%), despite these forests representing around 30% of the world's forest. Stratifying the plots accordingly weakens the TSR-P-relationship.*

Response: Thanks for the concern raised in your first sentence. We are well aware of that problem, and have even discussed it in our paper. Of course this is just one more case of a general trend of under-documentation of all species (not just trees) from developing countries. This is problem all researchers from developed countries should at least be aware of and try to mend as best we can; we at the GFBi are doing our part and currently trying to collect more samples from the tropics for future research studies.

Regarding your second point, we recognize that stratifying the plots may make the results more robust, but its effect would be limited and will not alter the overall global trend, because you already stated in your comments (Page 20) that "the (stratification) effect is moderate, with slightly lower values than the original non-stratified approach. This result suggests that also with non-stratified sampling always some tropical plots with high species richness are drawn, making the original Š robust to unrepresentative sampling."

Additionally, because the data are overwhelmingly temperate, roughly 3% boreal, and <1% tropical, and draws in Liang et al 2016 were random across the globe, most of the 500-stand draws in our original 2016 paper were likely to have most data from non-tropical sites, so the influence of tropical high diversity, high productivity sites were likely modest, unless they had extremely high influence per datum on the overall fitted function because of their position in data space (which is possible). This is relevant to your concern (above) about our global result being influenced by the sharp gradient in boreal to tropical forests in both productivity and richness. Similarly, boreal stands would have shown up not very often; maybe 15 or so times on average in each 500 stand draw, with tropical stands drawn twice or so on average out of each 500 draw. In contrast, if our data had hypothetically been roughly representative equally of boreal, temperate and tropical forests, the global relationship might have been much more influenced by the gradient from low diversity, low productivity boreal to high-high tropical. In other words, our original data and fits were likely strongly temperate in flavor, despite our concerns about the undue influence of the boreal-tropical gradient. It may be in fact that we should have a different concern; not that boreal-tropical gradient exerted too much influence on our published global fitted relationship of productivity-richness, but that our global analysis 'undercounted' the impact of tropical and boreal forests on the global relationship, given that the vast majority of stands in each 500-lot draw were temperate. We are not yet sure how best to check these issues.

(3) *In the spatial regression model, distances between plots were computed without taking the spherical nature of earth into account. This had little effect on the slope estimate of the TSRP-relationship.*

Response: Thank you for sharing your insight into and findings about this. We appreciate it. We recognize that calculating distances between plots by taking the spherical nature of earth into account may slightly improve the accuracy of our estimated BPR. The magnitude of such improvement is yet to be determined by future research.

(4) *The computational burden of the spatial model required subsampling the data to 500 data points. The authors did not correctly compute confidence intervals for this approach, wrongly interpreting subsampling as bootstrapping and additionally incorrectly computing bootstrap standard errors. A correct subsampling-based estimation led to approximate trippling of the reported confidence interval.*

Response:

Thank you for raising this concern. Bootstrapping is only efficient at depicting a global trend if the re-sampling size is close to the global sampling size (Efron and Tibshirani 1993). However, for our study, the 500-plot subsample is far from our global sampling size (>700,000). Considering that you used a minimalism approach, in which "while Liang et al. (2016) run 10000 bootstraps, we only do 50," (p.8) your suggested global results only represent, in fact, ~ 50*500=25000 plots or approximately 3 percent of the global sample. In other words, there is a 97% information loss in your approach.

In the textbook description of the bootstrapping by Efron and Tibshirani (1993), echoed by many (e.g. Hesterberg 2015), it is outlined that the bootstrap sample should be equal in size as the original sample, and that any smaller re-sampling sizes would lead to a biased estimate of standard error. This is also the main reason why you did not find a significant global BPR as it should have been.

Allow us to demonstrate, with R-code (in blue) and outputs, how we have derived our results. While there is well-established literature regarding the validity of the subsampling method we have taken, less is known about an appropriate choice of the size of a subsample and the number of subsamples. With a global sample size over 600,000, we have chosen the subsample size to be 500 and a total of 10,000 subsamples out of consideration for computational feasibility and adequate representation of the global sample. Our approach leading to these choices is indeed ad-hoc and the standard errors are at best approximations. We welcome ideas and possible collaboration to establish more rigorous approaches. On the other hand, with a large amount of data and thus information, statistical significance is not tenuous to attain.

1. For each random subset of 500 plots, we consider this subset a separate study unit (one can regard this as equivalent to a subregion). In the Geospatial Random Forests model, we calibrate one biodiversity-productivity relationship (BPR) curve based on this subset. With a global sampling size of >700,000, we find that it takes more than 2,000 subsets of 500 samples in our global BPR analysis, so that any single plot would have been accounted for at least once in the analysis. To be safe, we used 10,000 subsets (i.e. iterations) (**Fig. 1**);

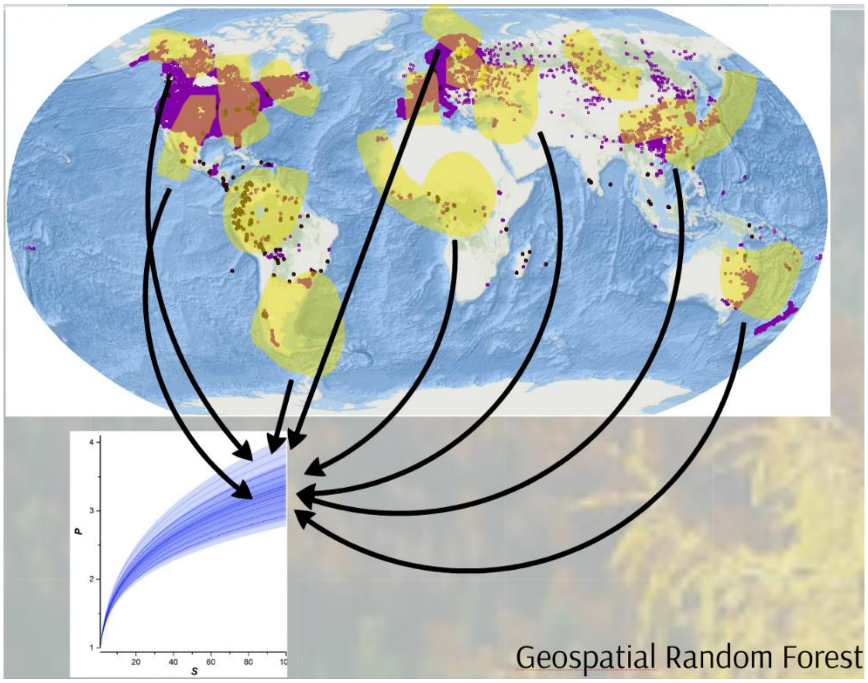

*Figure 1. A graphic demonstration of the Geospatial random forests model. We randomly select 500 plots from across the world as one study unit or "subregion" (yellow), calibrate one biodiversity-productivity relationship (BPR) using the model, and draw a ceteris paribus BPR curve. Repeating this 10,000 times provide a sufficient global coverage as each plot has on average been covered for ~7 times (500*10000/720000≈7). Note that actual subregions can be spatially discontinuous depending on the randomization.*

A major strength of this approach is that it does not require any *a priori* assumption on the population distribution or any *a priori* delineation of forest type units across the world, within which forests have similar conditions. This is especially useful because there is no universally accepted forest type delineation across the world (FAO 2015).

2. Load the global data set, note that we did remove plots with extreme species richness or productivity values (i.e. those beyond 99.996th percentile), and plots with zero species richness or productivity.

```
# Load packages
library(nlme)

# Load plot-level data
# Download GFB1_data_figshare.xlxs from Figshare and convert to a csv file

data<- read.csv("GFB1_data_figshare.csv")
data <- subset(data, P>0)
data <- subset(data, S>0)

quantile(data$S,0.99996)
quantile(data$P,0.99996)

data1 <- subset(data,data$S<=270 & data$P<=533 & data$S >0 & data$P>0)   # removed 894 plots with 0 or extreme S and P values
```

3. For each subset of 500 plots (without replacement), we consider this subset a separate study unit (one can regard this as equivalent to a subregion). We draw one BPR curve based on this subset, using our geospatial random forests, by keeping other variables constant at their sample mean, only increasing species richness from 1 to 271 (the global maximum).

```
##############################################################
################## Derive Global GeoRF Estimation ###################
##############################################################
logP <- log(data1$P)
# jig coordinates to avoid duplicated values
Lon1 <- data1$Lon+ runif(length(data1$Lon),-0.0001,0.0001)
Lat1 <- data1$Lat+ runif(length(data1$Lat),-0.0001,0.0001)
data1 <- cbind.data.frame(data1, logP, Lat1, Lon1)
########### Loop #################
coef <- matrix(0, nrow=10000, ncol=20)  # Coef Matrix

for(i in 1: 10000) {
  tryCatch({
        training <- data1[sample(1:nrow(data1), 500, replace=FALSE),]   # turn 'replace' off to maximize
inclusion of new plots
        logS <- log(training$S)
        training <- cbind.data.frame(training, logS)
        gls1 <- gls(logP~ logS + G + T3 + C1 + C3 + PET + IAA + E, data=training, method="ML", corr=
corSpher(form = ~ Lon1 + Lat1, nugget = TRUE), control=glsControl(singular.ok=TRUE))
        coef[i,3] <- i
        coef[i,4] <- logLik (gls1)
        coef[i,5] <- AIC (gls1)
        coef[i,6]<- BIC (gls1)
                #Generalized coefficient of determination
                gls0 <- gls(logP~ 1, data=training, method="ML")
                R2   <- 1-exp(logLik(gls0)-logLik(gls1))^(2/500)
                coef[i,7]<- R2
        coef[i,8]  <- coef(gls1)[1]
        coef[i,9]  <- coef(gls1)[2]
        coef[i,10] <- coef(gls1)[3]
        coef[i,11] <- coef(gls1)[4]
        coef[i,12] <- coef(gls1)[5]
        coef[i,13] <- coef(gls1)[6]
        coef[i,14] <- coef(gls1)[7]
        coef[i,15] <- coef(gls1)[8]
        coef[i,16] <- coef(gls1)[9]
        coef[i,17] <- 0
        # Baseline (S=1) productivity
        # logS + B1 + T3 + C1 + C3 + PET + IAA + E
        newdata <- data.frame(logS=0, G=mean(training$G), T3=mean(training$T3),
C1=mean(training$C1), C3=mean(training$C3),PET=mean(training$PET), IAA=mean(training$IAA),
E=mean(training$E))
        coef[i,20]  <- exp(predict(gls1,newdata))
        #counter
        cat(i, " of ", 1000, date(),"Theta=",coef(gls1)[2], "R2=", R2, "\n" )
        #remove files
        rm(training, newdata, gls1, R2)
  }, error=function(e){})
            }
coef_df <- as.data.frame(coef)

names(coef_df) <- c("0", "0", "i", "Loglik", "AIC", "BIC", "R2","const","theta", "B", "T3", "C1", "C3", "PET",
"IAA", "E", "0", "0", "0", "P_1")

write.csv(coef_df, "global_estimates.csv")
```

4. Repeating the foregoing step 10,000 times, we get a combined subregions that cover the entire global forest range. Meanwhile, we have 10,000 curves (green in the following **Fig. 2**) that represent possible BPR's across the world. Treating each region as an independent study unit, instead of a bootstrapping re-sample, we can calculate and plot the mean and standard error (SE) of the predicted BPR curves across the world as shown in the figure below (mean: black line, with red curves representing 95% C.I.)

```
######################################################################
#### Draw estimated Biodiversity-Productivity Relationship (BPR) curves #########
######################################################################

data<- read.csv("global_estimates.csv")

theta <- data$theta
mean(theta)
P_base <- mean(data$P_1)

# Predict P over an increased S from 1 to global max (271), which corresponds to S_hat from 100/271 to
100
S     <- seq(1,271,1)
S_hat <- S*100/271

P_est <- data.frame(matrix(0, 10000, ncol =273))
P_est[,1] <- P_base
P_est[,2] <- theta

for (i in 1:10000){
  P_est[i,3:273] <- P_est[i,1] * S ^ P_est[i,2]
  }

# demosntration plot only shows the first 18 iterations
plot(S_hat,colMeans(P_est[,3:273]), ylim=c(0,20), type="l",col = "blue", ylab="P")
for (i in 1:18){
  P_est[i,3:273] <- P_est[i,1] * S ^ P_est[i,2]
  lines(S_hat,P_est[i,3:273],col = "green")
}
# Confidence intervals
lines(S_hat,colMeans(P_est[,3:273])+1.96*apply(P_est[,3:273], 2, sd)/sqrt(10000), ylim=c(0,20),
type="l",col = "red")
lines(S_hat,colMeans(P_est[,3:273])-1.96*apply(P_est[,3:273], 2, sd)/sqrt(10000), ylim=c(0,20),
type="l",col = "red")
```

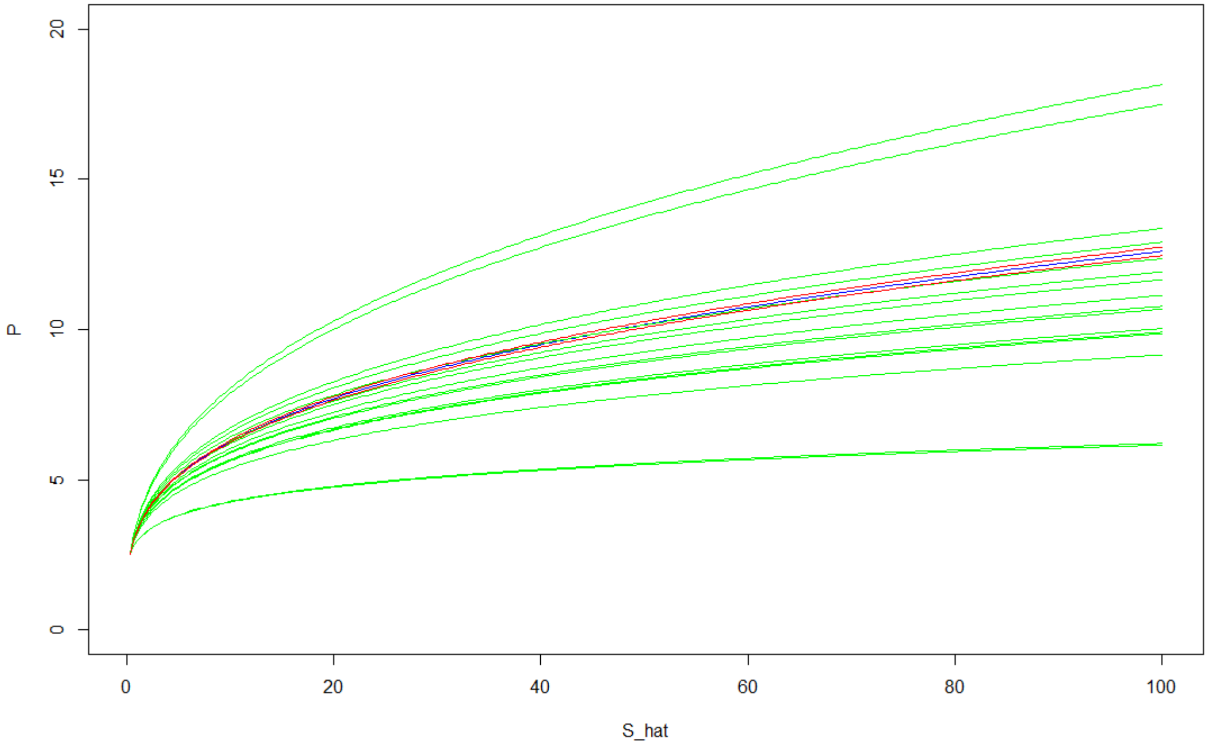

*Figure 2. Sample BPR curves from the 10,000 estimated curves from across the world. The figure is nearly identical to **Fig. 3A** of Liang et al. 2016, with some minor differences due to the random process. For easy comparison across the world, we set the base value of P as 2.5m^{3}ha^{-1}yr^{-1}, and convert species richness (S) to relative species richness (S_hat): S_hat = S *100 / 271.*

5. To demonstrate that this estimated global mean and confidence interval from our Geospatial random forests model (Fig. 2) is a good proxy of the true global BPR trend, we compare this result with an outcome from an ordinary least squares model (OLS), of which the estimates are based on the entire sample (with >700,000 plots).

```
## A comparison with OLS model  ##
data   <- read.csv("GFB1_data_figshare.csv")
data1 <- subset(data,data$S<=270 & data$P<=533 & data$S >0 & data$P>0)   # removed 894

logS <- log(data1$S)
ols1 <- lm(logP~ logS + G + T3 + C1 + C3 + PET + IAA + E, data=data1)

theta <- coef(ols1)[2]
summary(ols1)
se_theta <- 2.100e-03

S     <- seq(1,271,1)
S_hat <- S*100/271
P_base <- 2.5

P_est_ols <- P_base * S ^ theta                      # mean predicted BPR
P_est_ols_ub <- P_base * S ^ (theta+1.96* se_theta)   # upper bound of 95% CI
```

```
P_est_ols_lb <- P_base * S ^ (theta-1.96* se_theta)   # lower bound of 95% CI

plot(S_hat, P_est_ols, ylim=c(0,20), type="l",col = "blue", ylab="P")
# Confidence intervals
lines(S_hat,P_est_ols_ub, ylim=c(0,20), type="l",col = "red")
lines(S_hat,P_est_ols_lb , ylim=c(0,20), type="l",col = "red")
```

The corresponding line plot is printed below. According to this graph, the BPR has the same curvature, but estimated productivity (P) is in general 10-20% lower than the estimated values from the Geospatial random forests, presumably due to the fact that spatial autocorrelation is not accounted for in the OLS model. Nevertheless, the confidence interval from the OLS model generally matches the confidence interval from the Geospatial random forests (**Fig. 2**).

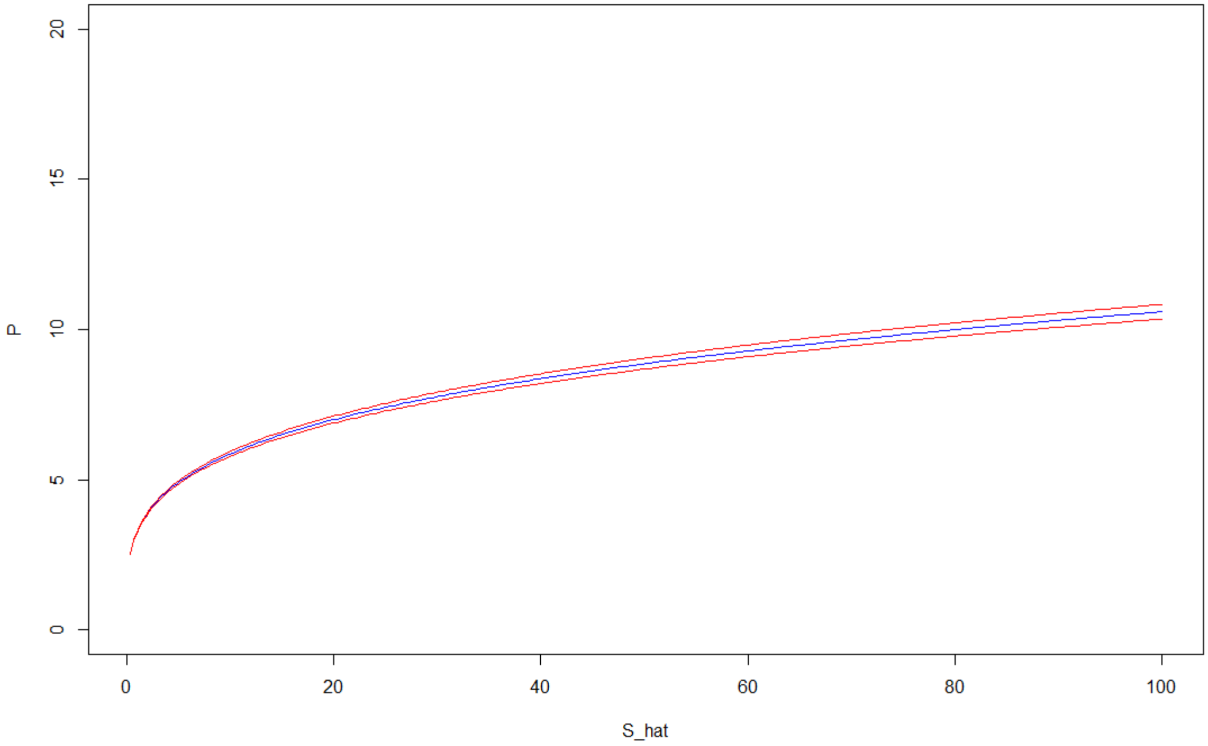

*Figure 3 Estimated BPR curve (with 95% confidence interval bands), using an ordinary least squares (OLS) model, based on the entire GFB sample with >700,000 plots. For easy comparison across the world, we set the base value of P as 2.5m^{3}ha^{-1}yr^{-1}, and convert species richness (S) to relative species richness (S_hat): S_hat = S *100 / 271.*

(5) *As noted earlier (Schulze et al., 2018), some 4% of the plots had productivity values (far) beyond what is biologically plausible (Stape et al., 2010). The likely reason is that small plots with large inventory errors in the productivity may lead to erratically high values. Not taking this into account in the analysis, e.g. by down-weighting plots with productivities above 30 m2ha□1y□1 at least indicates an unre-ected use of data.*

Response:
Thank you for your concern. As shown in the R-code above, we have removed extremely high productivity values, above the top 0.004 percent quantile (P<=533). It is admittedly a difficult task to filter out the potentially biased values from such a large sample, but we are working with data scientists and data contributors to further improve the accuracy of our data.

References

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman & Hall, New York.

FAO. 2015. Global Forest Resources Assessment 2015 - How are the world's forests changing? , Food and Agriculture Organization of the United Nations, Rome, Italy.

Hesterberg, T. C. 2015. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. The American Statistician **69**:371-386.

Jepson, P., and R. J. Whittaker. 2002. Ecoregions in Context: A Critique with Special Reference to Indonesia. Conservation Biology **16**:42-57.

Appendix: R script for estimating BPR for the grassland biomes

```
# Estimate BPR curves by ecoregion
# (C) Jingjing Liang 2018

library(nlme)

# Load plot-level data
# Download GFB1_data_figshare.xlxs from Figshare and convert to a csv file

data<- read.csv("GFB1_data_figshare.csv")
data <- subset(data, P>0)
data <- subset(data, S>0)
attach(data)


## Montane Grass and shrubs ##

data1 <- subset(data, data$Ecoregion==10 | data$ Ecoregion ==9 | data$Ecoregion ==8)
data1 <- subset(data1,data1$P<=quantile(data1$P,0.999))


############### BPR Estimation ######################

############ Bootstrapping ##################
coef <- matrix(0, nrow=50, ncol=101)          # Coef Matrix


for(i in 1: 50) {
  tryCatch({

            training <- data1[sample(1:nrow(data1), 23133, replace=TRUE),]
            logP <- log(training$P)

            Lat1 <- training$Lat + rnorm(length(training$Lat))
            Lon1 <- training$Lon + rnorm(length(training$Lon))
            training <- cbind(training, logP, Lat1, Lon1)

            S_max <- max(training$S)
            SR <- training$S/S_max*100

            logS <- log(SR)
            training <- cbind(training, logS)

            lm1 <- lm(logP~ logS + G + T3 + C1 + C3 + PET + IAA + E, data=training)

            # Derive ceteris paribus BPR curve
            newdata <- data.frame(logS=log(seq(1,100,1)),
G=mean(training$G),T3=mean(training$T3), C1=mean(training$C1),
C3=mean(training$C3),PET=mean(training$PET), IAA=mean(training$IAA),
E=mean(training$E))
```

```
        coef[i,1]  <-coef(lm1)[2]             #theta
        coef[i,2:101]   <- exp(predict(lm1,newdata))
 plot(coef[i,])
        #counter
        cat(i, " of ", 50, date(), "\n" )

        #remove files
        rm(training, newdata, gls1)

 }, error=function(e){})
        }

coef_df <- as.data.frame(coef)

write.csv(coef_df, "Ecoregion_Grasslands_BPR.csv")

# Plot mean and 95% CI of bootstrapping
plot(seq(1,100,1),colMeans(coef_df[,2:101]), ylim=c(0,6), type="l",col = "blue",
ylab="P",xlab="S_relative")
# Confidence interval
lines(seq(1,100,1),colMeans(coef_df[,2:101])+1.96*apply(coef_df[,2:101], 2, sd), ylim=c(5,8),
type="l",col = "red")
lines(seq(1,100,1),colMeans(coef_df[,2:101])-1.96*apply(coef_df[,2:101], 2, sd), ylim=c(5,8),
type="l",col = "red")


# End of the code
```