Dear Niklas and co-authors,

This project tackles an important issue in the study of phenotypic evolution using fossil data: that the structure of the stratigraphic record confounds the direct application of evolutionary models to these data. This is underexplored (and somewhat ignored) but is exactly the type of work paleobiologists need to be doing. Please find below three reviews from researchers who are very familiar with the evolutionary modeling side of things and/or the sedimentary modeling side of things. I think these comments indicate both the merit of the work but also the need for some revisions if this is going to be as impactful as it could be.

To this end, I was also particularly surprised by the performance of the model selection without stratigraphic biases. One reviewer's comments about the need for the continuous-time expansion may be relevant here. But I also wonder if the results that you are getting are due at least in part to using model implementations for testing (those from the paleoTS package) which are different from the implementations used to simulate the data (those original to this study). To quickly assess this, I ran some simulations using the functions available in the paleoTS package. I found some interesting patterns and they are not completely different from what you found. For example, short time series (N=5) simulated under the stasis model in paleoTS are hard to distinguish from unbiased random walks (and this is expected given the smaller number of parameters in the unbiased random walk model). Similarly there is increasing support for OU as time series length increases. However, stasis never "loses" to the same degree as shown in your Figure 10. In fact, stasis is the best model in the majority of cases across all time series lengths*. It is also worth noting that the results are sensitive to the amount of variance (stasis is more strongly supported for time series simulated with smaller variance). I have included the R script and figure here.

*Although rarely with an AICc > 0.9. One reviewer commented on the stringency (and interpretation) of relying on AICc values that high, which I think is worth considering. Another reviewer wondered about using AICc instead of AIC. Based on my knowledge of the paleoTS package, you must have used AICc -- so this should be clarified in the text.

I look forward to seeing the revised manuscript!

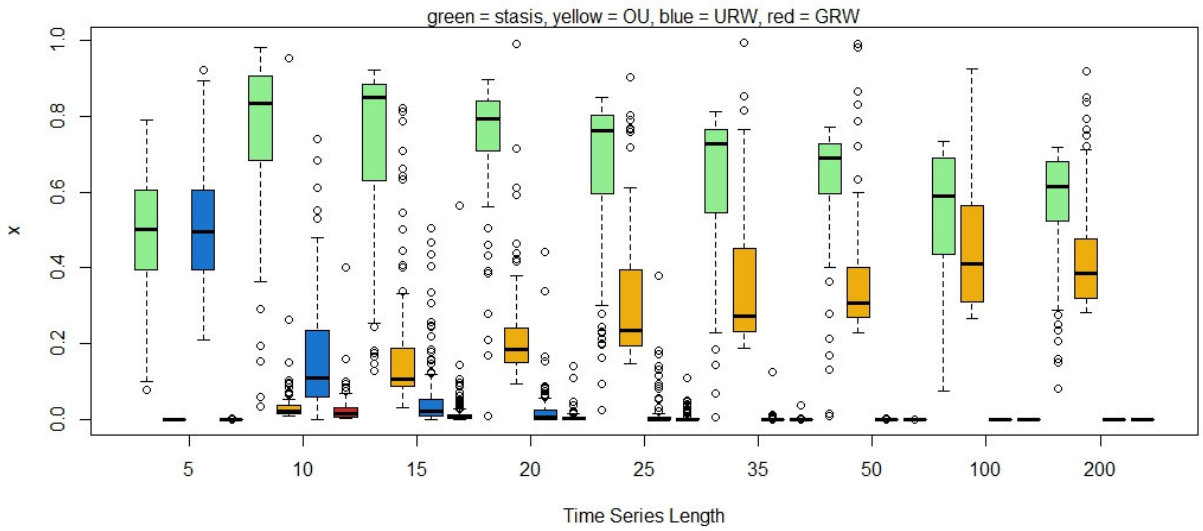
```
##This script was compiled to see how poorly model selection performed
#for simulated data using just functions from the paleoTS package #0.5.3
#Melanie J Hopkins 2/26/24.

compile<-array(dim = c(4,100,9))
ns<-c(5,10,15,20,25,35,50,100,200)

for (j in 1:9){
      for (i in 1:100){
            test<-sim.Stasis(ns=ns[j],omega = 0.5)
            fts<-fitSimple(test,model = 'Stasis')
            fto<-fitSimple(test,model = 'OU')
            ftb<-fitSimple(test,model = 'URW')
            ftt<-fitSimple(test,model = 'GRW')
            compile[,i,j]<-compareModels(fts,fto,ftb,ftt,silent = TRUE)$modelFits$Akaike.wt
      }
}

compile.group<-data.frame(
        x=matrix(compile,ncol=1),
        y=c(rep(c('Stasis','OU','URW','GRW'),900)),
        z=c(rep(5,400),rep(10,400),rep(15,400),rep(20,400),rep(25,400),
                    rep(35,400),rep(50,400),rep(100,400),rep(200,400)),
        stringsAsFactors = FALSE
)
compile.group$y<-ordered(compile.group$y,levels=c('Stasis','OU','URW','GRW'))

boxplot(x~y+z,data = compile.group,
       col=c('lightgreen','darkgoldenrod2','dodgerblue3','firebrick3'),
       xaxt = 'n',
       xlab = 'Time Series Length',
       'AICc weight'
)
axis(1,at=c(2.5,6.5,10.5,14.5,18.5,22.5,26.5,30.5,34.5),
    labels = c(5,10,15,20,25,35,50,100,200))
mtext('green = stasis, yellow = OU, blue = URW, red = GRW',side=3)
```


green = stasis, yellow = OU, blue = URW, red = GRW
x
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
35
50
100
200
Time Series Length