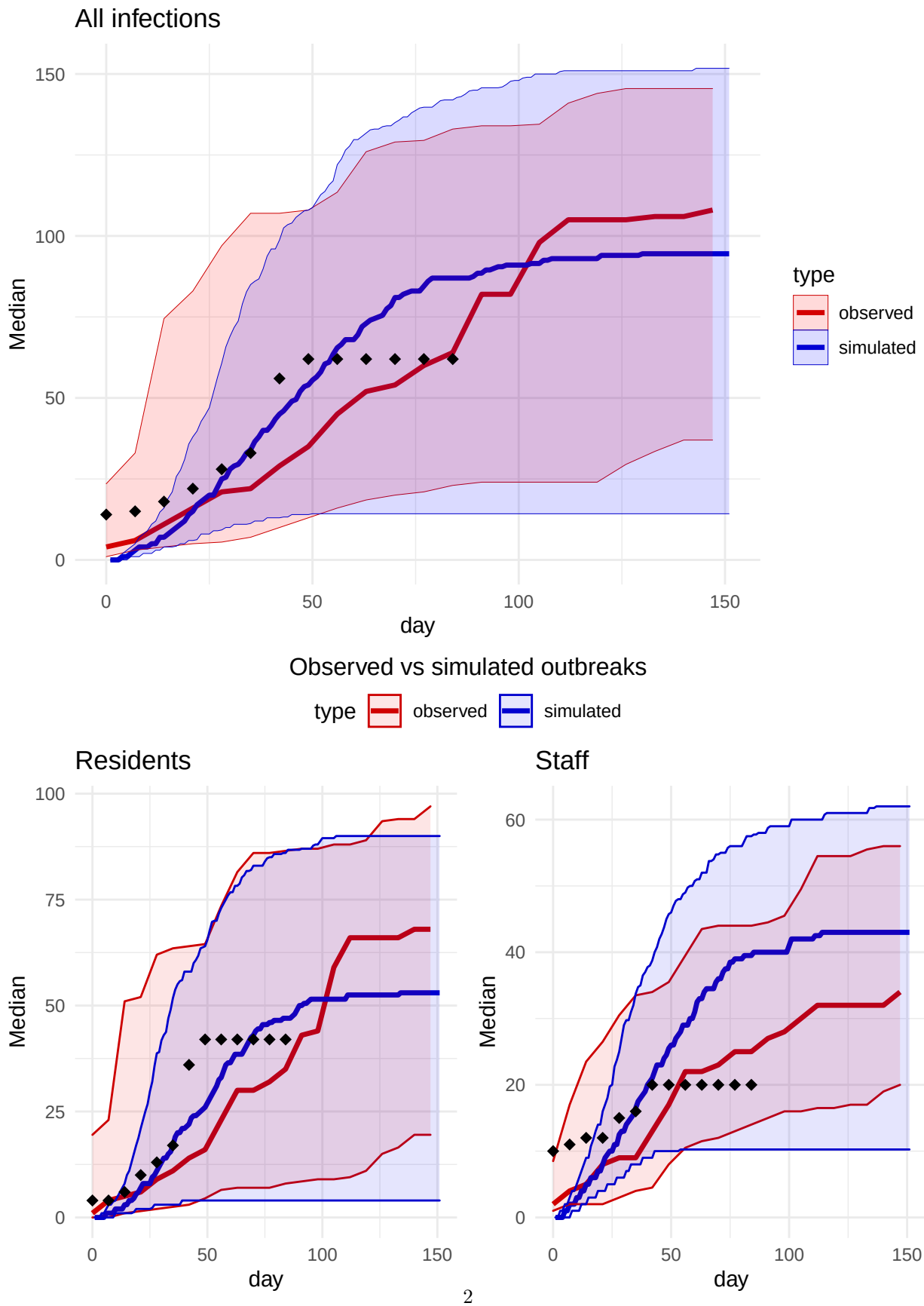
Validation

Observed vs expected plot



Correlation of observed and expected

Using the median

This approach uses only the median and we weould be loosing some information here, but I think is the easiest to interpret.

I calculated the correlation between the median number of infected for the simulated and observed cases, and used a simple linear regression where: $Observed \sim Simulated$ and obtained the R^2.

Model	ρ	R^2
Total Infected	0.9221291	0.8503221
Infected Residents	0.908348	0.8250961
Infected Staff	0.9804931	0.9613668

Using a simple linear regression fro individual observations.

In the following two approaches we used all the observations (simulated and observed), so this could provide us a better insight, but I am not so sure if the model specification I am doing is correct.

To calculate the R^2 we used model

$$Y_i \sim X_1 + X_2 + X_3$$

Where:

- Y_i represent the cumulative number of infected.
- X_1 if its from an observed or simulated outbreak.
- X_2 the day since the outbreak.
- X_3 the number of simulation or outbreak ID.

We did a model for the cumulative number of residents infected, cumulative number of staff infected, and total cumulative number of infected (residents+staff).

Y_i	R^2
Total Infected	0.7963542
Infected residents	0.7992968
Infected Staff	0.7910985

Using a LMM to calculate a pseudo R^2

We used the package `MuMIn` which can be used to estimate a pseudo R^2[@REF].
This approach calculate a marginal R^2, which represents the variance explained by the fixed effects, and a conditional R^2, which represents the variance explained by the random effect.

Y_i	Marginal R^2	Conditional R^2
Total Infected	2.6913266×10^{-4}	0.9336034
Infected residents	5.7619384×10^{-4}	0.9319771
Infected Staff	1.5795679×10^{-5}	0.9366012

My interpretation of this would be:

A low marginal R^2 indicates that there is not much variation being explained by the fixed effects, which in our model is the type of outbreak (either simulated or observed), indicating that the model can not tell the difference when an outbreak is simulated or observed. This will indicate that our simulation model is good when replicating observed outbreaks.
A high conditional R^2 indicates that most of the variation is being explained by the random effect.