Supporting Information S1

Fixed or random? On the reliability of mixed-effect models for a small number of levels in grouping variables

1. Mixed-effect model implementations in R, default settings and convergence issues

The most used packages in R to fit mixed-effect models are lme4 (Bates et al. 2015) and glmmTMB (Brooks et al. 2017). These packages differ in their optimization routines and the calculation of p-values for linear mixed-effect models (LMMs). While lme4 uses standard optimizers, glmmTMB relies on automatic differentiation implemented in TMB package. Another difference is that glmmTMB offers to fit linear and generalized linear models with both the maximum likelihood (MLE) and the restricted maximum likelihood estimation (REML), while lme4 offers only REML for LMMs but not for GLMMs. Due to these different optimization routines glmmTMB and lme4 results could be slightly different, which we analyze in the following.

1.1 Standard deviation estimates and singular fits

With glmmTMB using REML the estimates of the standard deviations are bimodal with one peak at zero and one peak around the correct value (Fig. S1). When excluding values which presented standard deviation estimates smaller than 10^{-3}, the peak around zero vanishes

(Fig. S2). These estimates correspond to singular fits obtained with lme4. Using REML for generalized mixed-effect models led to a peak around zero and a peak slightly higher than the correct value (Fig S1b).

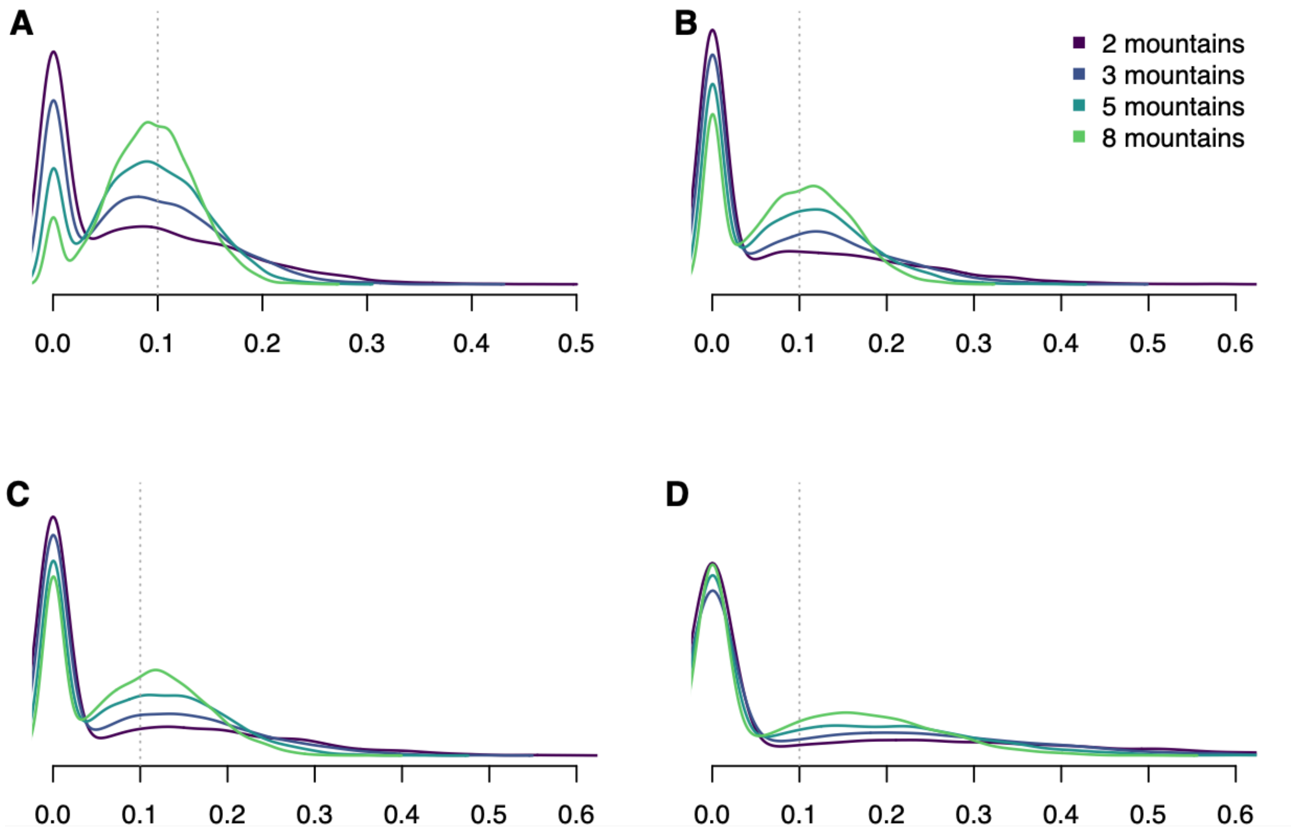

Figure S1: Standard deviation estimates of the random intercepts (A, C) and random slopes (B, D) for linear mixed-effect models (A, B) and generalized linar mixed-effect models (C,D) from the correctly specified mixed-effect model (Table 1. Eq. 10) in Scenario B, fitted with glmmTMB package using REML to simulated data with 2-8 mountains.. For each scenario, 5,000 simulations and models were tested. The grey lines represent the true standard deviation used in the simulation (0.1).

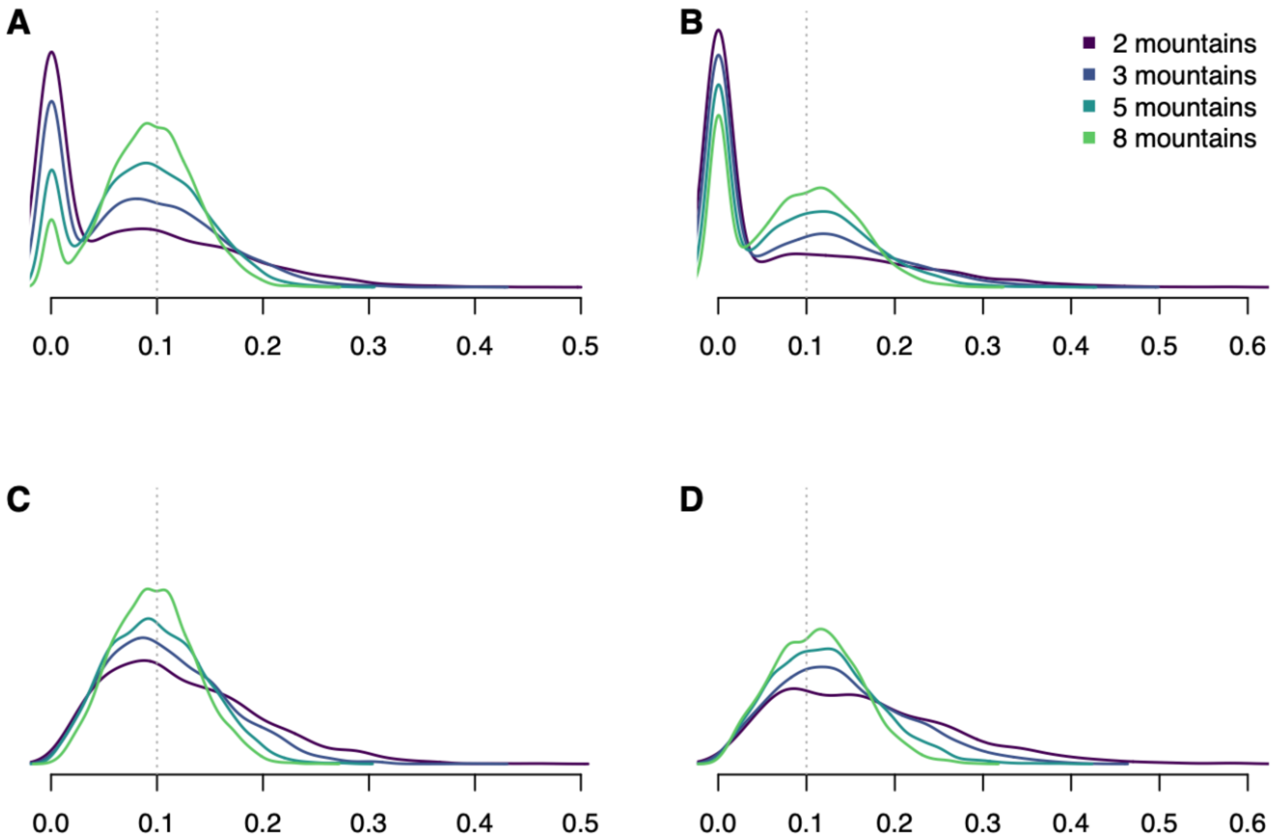

Figure S2: Standard deviation estimates of the random slopes and random intercepts (model 4) for linear mixed-effect models (LMM) fitted with glmmTMB using REML to simulated data with 2-8 mountains. A and B show the results for all models (with and without singular fits), C and D show the results for the models without singular fits (standard deviations of < 10^{-3} were assumed to be singular fits). For each scenario, 5,000 simulations and models were tested. The grey line represents the true standard deviation used in the simulation (0.1).

One difference between *lme4* and *glmmTMB* packages that users should be aware of is the default fitting algorithm for LMMs and GLMMs. By default, *lme4* uses restricted maximum likelihood (REML) for LMMs (*lmer* function) and unrestricted maximum likelihood (MLE) for GLMMs (*glmer* function), while *glmmTMB* uses MLE by default for any distribution. Below, we compare the distributions of the standard deviations of the random effects (random intercept and slope) of the correctly specified model in scenario B between REML and MLE using the package glmmTMB, due to its flexibility in fitting models with both algorithms.

Table S1: Proportion of models ran in **lme4** that presented singular fit convergence problem when using maximum likelihood (MLE) and restricted maximum likelihood (REML) fitting algorithms. Notice that for GLMMs in lme4, REML is not implemented.

	LMM		GLMM
Number of groups	**REML**	**MLE**	**MLE**
2	76%	89%	93%
3	61%	75%	86%
4	51%	64%	81%
5	44%	56%	78%
6	38%	48%	76%
7	34%	43%	72%
8	29%	38%	71%

Additional to the number of singular fits a direct comparison of REML and MLE with respect to their estimates of the standard deviations is necessary to compare their performance. Irrespective of the specific package (lme4 Fig. S3a or glmmTMB Fig S3b) using MLE for linear mixed-effect models lead to estimates, which are biased towards zero, while REML produces estimates, that are around unbiased. For generalized mixed-effect models the same applies (Fig. S8).

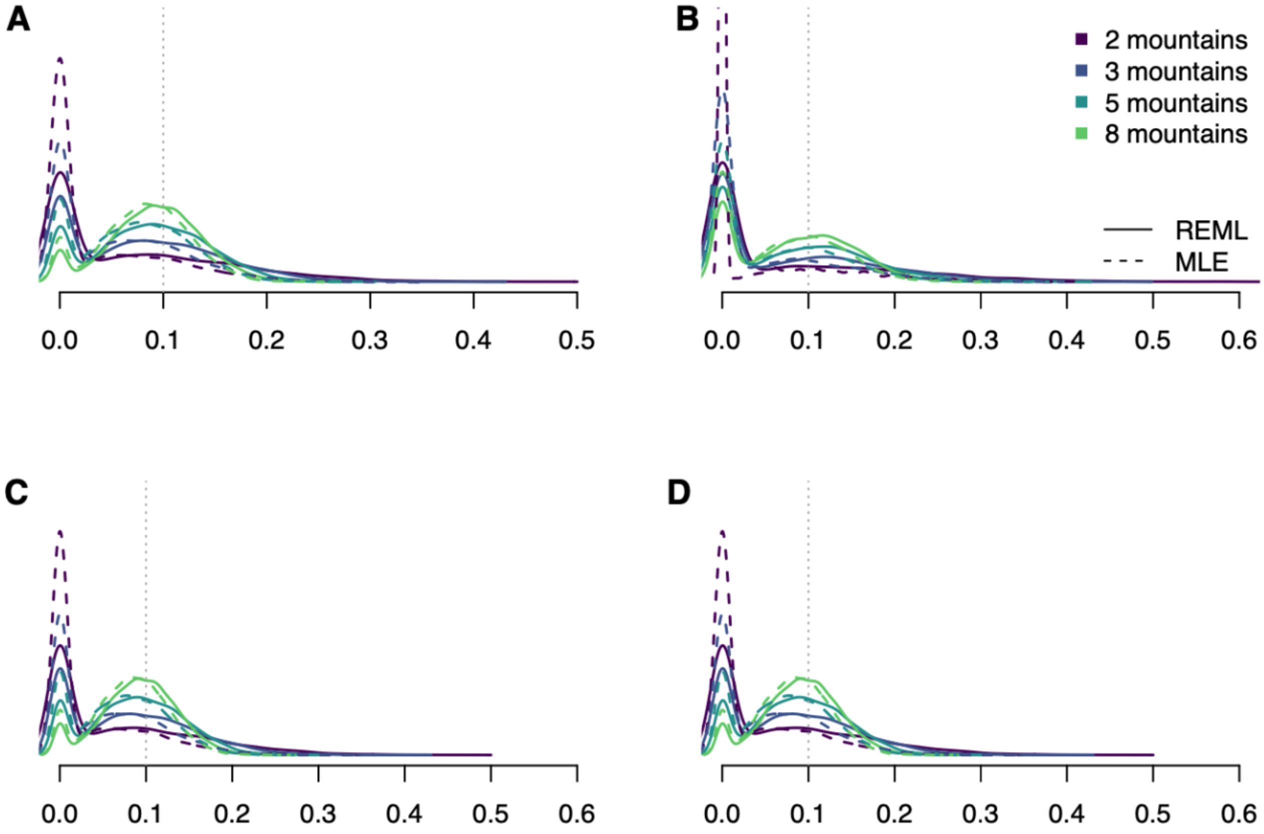

Figure S3: Standard deviation estimates of the random intercepts (A, C) and random slopes (B, D) for **linear mixed-effect models** (LMM) fitted to simulated data with 2-8 numbers of artificial mountain ranges. For each scenario, 5,000 simulations and models were tested. The grey line represents the true standard deviation used in the simulation (0.1). A and B show the results for linear mixed-effect models fitted with the *lme4* R package, C and D show the results for linear mixed effects models fitted with the glmmTMB package. The continuous line shows the results for the models fitted by restricted maximum likelihood estimation (REML) and the dotted line shows results for the models fitted by maximum likelihood estimation (MLE).

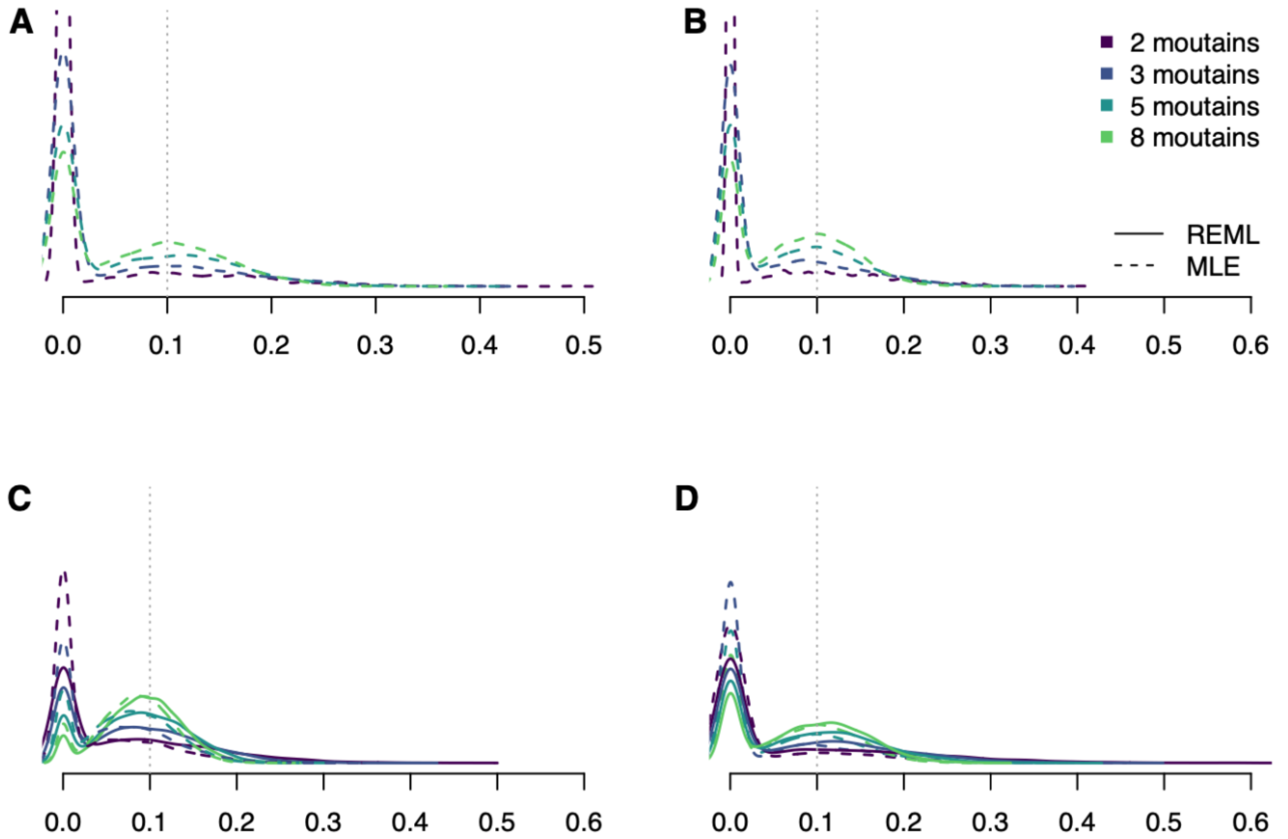

Figure S4: Standard deviation estimates of the random intercepts (A, C) and random slopes (B, D) for **generalized linear mixed-effect models** (GLMM) and linear regression models fitted to simulated data with 2-8 numbers of artificial mountain ranges. For each scenario, 5,000 simulations and models were tested. The grey line represents the true standard deviation used in the simulation (0.1). A and B show the results for GLMMs fitted by the lme4 R package, C and D show the results for GLMMs fitted by the glmmTMB package. The continuous line represents the results for models fitted by restricted maximum likelihood estimation (REML) and the dotted line shows the results for models fitted by maximum likelihood estimation (REML).

1.2. p-value calculations for mixed effect models

There is also a difference in the calculation of p-values between the two packages. While lme4 uses a Satterhwaite approximation to calculate degrees of freedom which then are fed into t-statistics, glmmTMB uses z-statistics and thus avoids the calculation of degrees of freedom. For generalized linear models, however, both use z-statistics and do not calculate degrees of freedom. Z-statistics are the asymptotic limits of t-statistics when having infinite data, however, these two differ in the low data limit and p-values calculated using z-statistics are overconfident. We see this for H2 in Scenario A (Fig. S5a), for which t-statistics can be calculated analytically, and thus t-statistics lead to values around the nominal type I error rates, while z-statistics lead to increased type I error rates. This also translates into power, where using z-statistic causes higher, but probably too high power compared to t-statistic (Fig S5a). For generalized linear mixed-effect models, both packages use z-values and, thus, present the same statistical properties (Fig. S5b). We speculate the reason why glmmTMB is using z-statistics also for LMMs is that t-statistics cannot be calculated analytically for GLMMs and thus have to be approximated anyways. However, when interpreting the results, it is important to keep in mind that z-statistics are only approximations in the low data limit. We believe that the power of GLMMs would be similar to power of LMMs when t-values would be used instead of z-values.

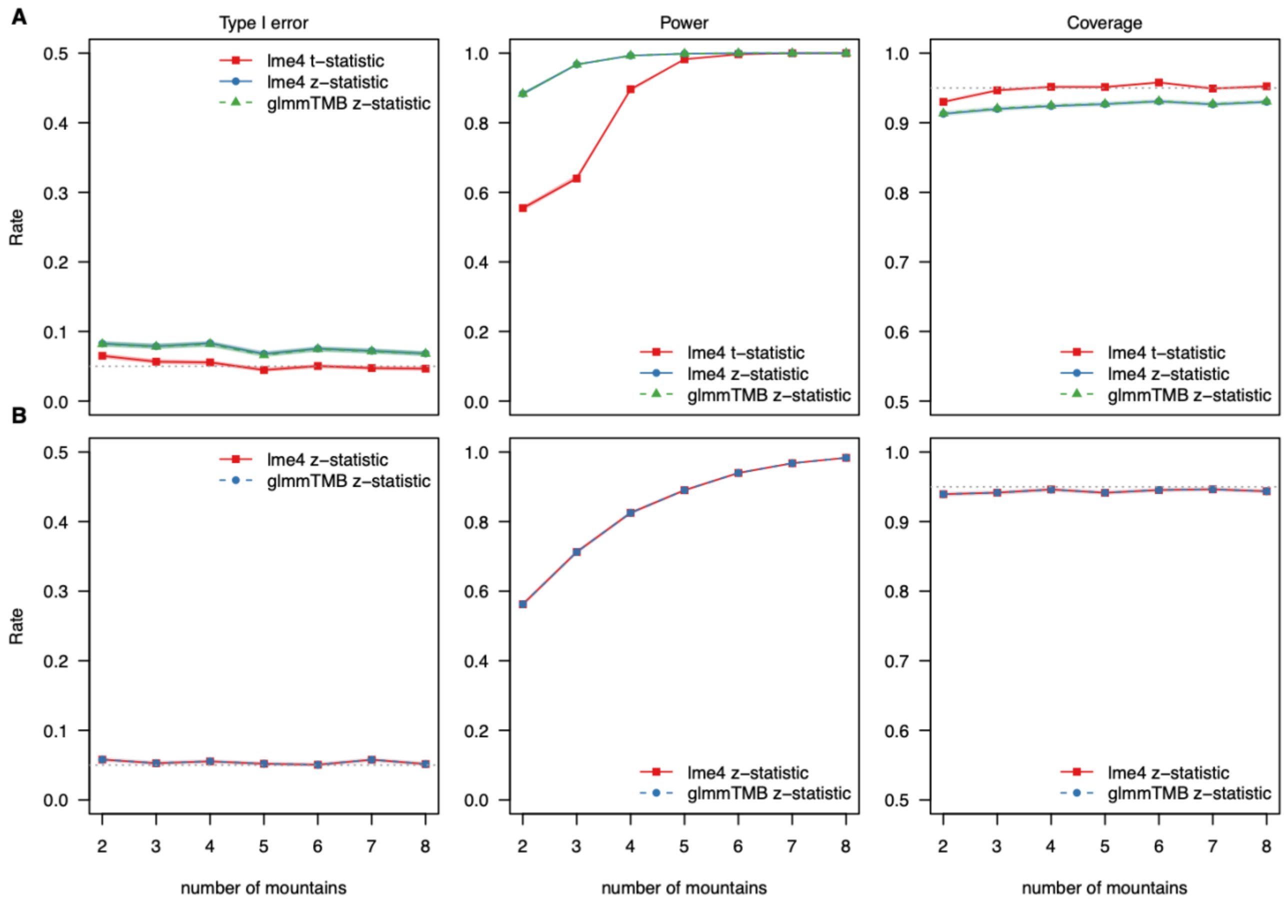

Figure S5: Type I error rates, power, and coverage for linear mixed-effect models (A) and generalized linear mixed-effect models (B), fitted to simulated data with 2-8 mountains for scenario B (random intercept and slope for each mountain) and 50 (lmm) and 200 (glmm) observations per mountain range. For each scenario, 5,000 simulations and models were tested.

1.3. Calculation of the mean temperature effect in fixed-effect models with interaction

As fixed-effect models with interactions estimate the effect of one level and its contrasts to the other levels, the 'grand mean' effect of temperature itself is not estimated.

To calculate the grand mean and its significance, we estimate the grand mean as the weighted mean X of the individual level effect estimates X_i $(i = 1, \dots, I\ mountains)$.

$$\bar{X} = w_i \ * \ X_i$$

with $w_i = \left(\sigma_i^2 + \sum_{j=1}^{k} \sigma_{ij}^2\right) / \left(\sum_{i=1}^{k} \sigma_i^2 + \sum_{i=1}^{k} \sum_{j\,(j \neq i\,)} \sigma_{ij} \right)$,

where σ_{ij} are the respective components of the covariance matrix of the interaction terms.

Since each individual mountain effect is uncertain, it is more difficult to estimate the standard error for the grand mean temperature effect, but it can be done via uncertainty propagation. With this technique, the variance of the mean effect is composed of two parts. The first part, accounts for the uncertainty in the estimators of the individual levels and following the rules of uncertainty propagation (Hughes & Hase 2010) takes the form:

$$\sigma_{\bar{X}}^2 = (\sum_{i=1}^{k} w_i^2 \ \sigma_i^2 + \sum_{i=1}^{k} \ \sum_{j\,(j \neq i)} w_i w_j \sigma_{ij}^2 \ )$$

The second part, which is the averaging of the individual effect estimates has the standard form of the standard deviation:

$$\sigma_{\overline{SD}}^2 = \sum_{i=1}^{k} w_i \frac{(x_i - \bar{x})^2}{k-1}$$

Summing up these two uncertainty contributions, we can calculate the standard error and thus the p-value for the grand mean temperature effect:

$$SE_{\bar{X}} \ = \ \sqrt{\sigma_{\bar{X}}^2 + \sigma_{\overline{SD}}^2}$$

2. Different standard deviations for the random effects for linear models

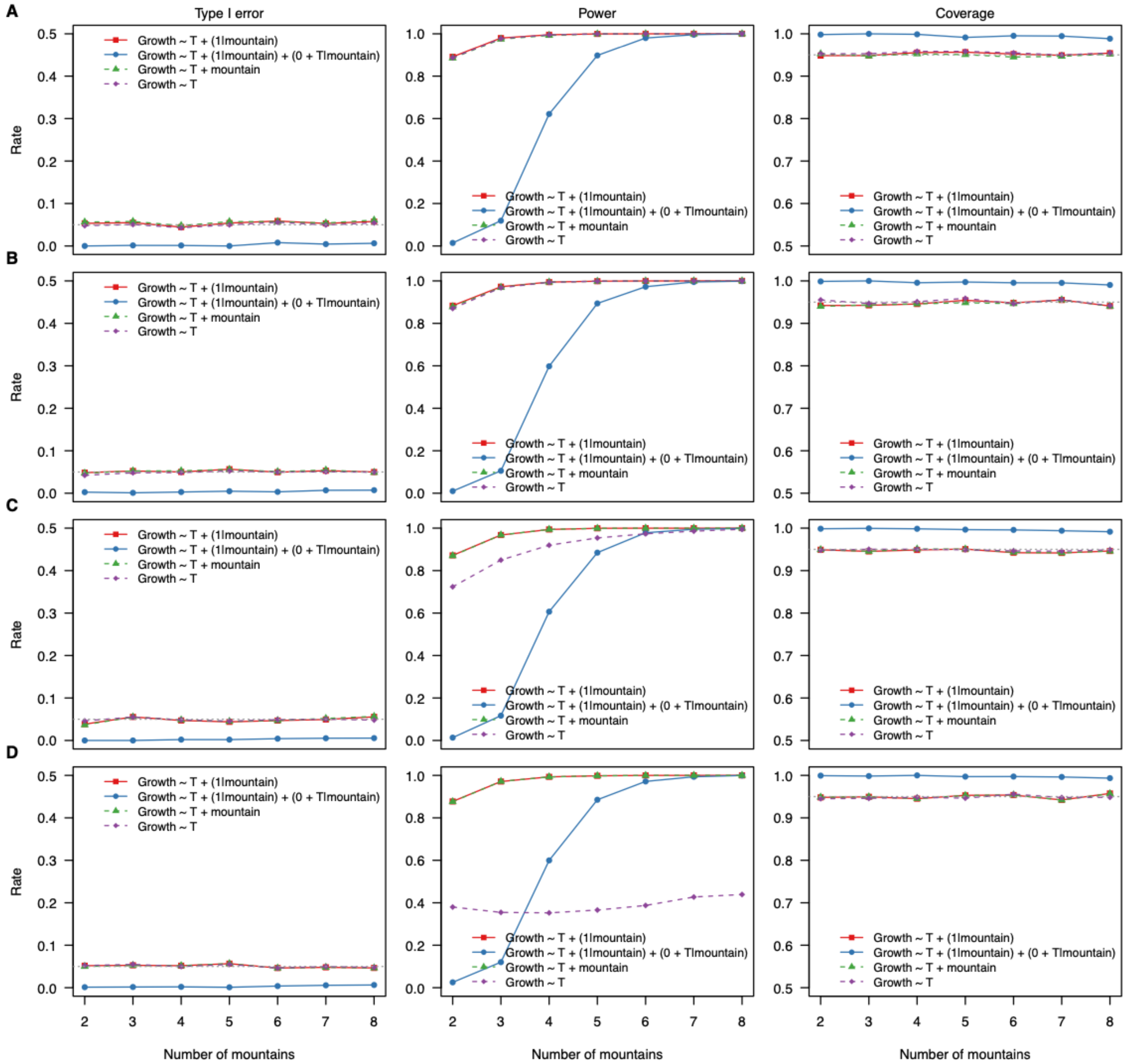

Figure S6: Type I error rates, power, and coverage for linear (mixed effect) models fitted with lme4 to simulated data with 2-8 mountains for scenario A (random intercept for each mountain) and 50 observations per mountain range. A-D show the results for different standard deviations of the random effect (0.01,0.1, 0.5 and 2.0). For each scenario, 5,000 simulations and models were tested.

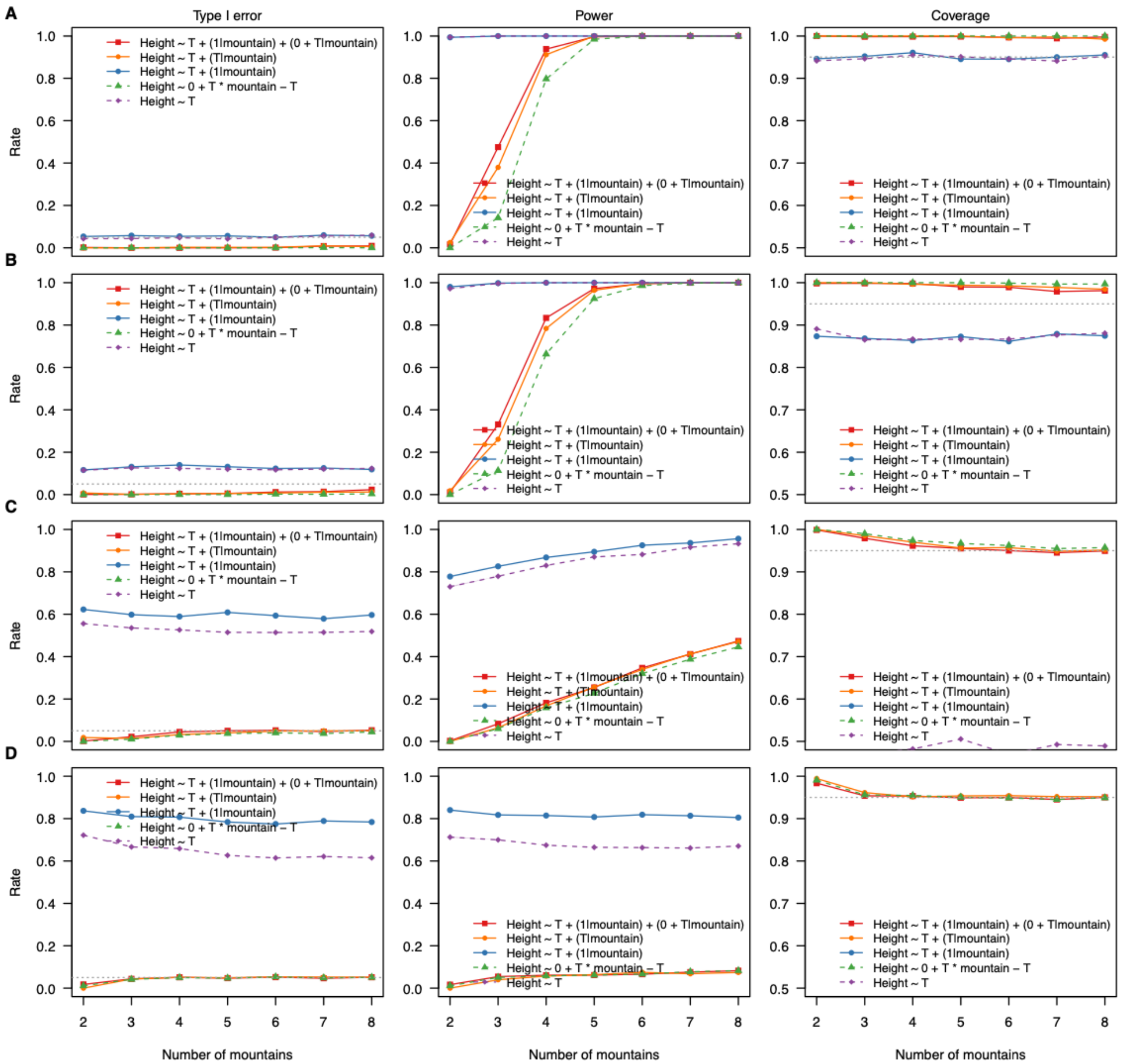

Figure S7: Type I error rates, power, and coverage for generalized linear (mixed effect) models fitted with lme4 to simulated data with 2-8 mountains for scenario B (random intercept and random slope for each mountain) and 50 observations per mountain range. A-D show the results for different standard deviations (0.01,0.1, 0.5 and 2.0). For each scenario, 5,000 simulations and models were tested.

3. Influence of sample size on statistical properties

Generalized linear models (mixed or not) with non-gaussian distributions require much more data to have a high probability to truly detect a significant effect (power). We ran additional data with different sample sizes per level in the grouping variable (25, 50,100, 200) and compared type I, power, and coverage for the binomial models (Hypothesis 1, Box 1) in both scenarios A (Figure S8) and B (Figure S9).

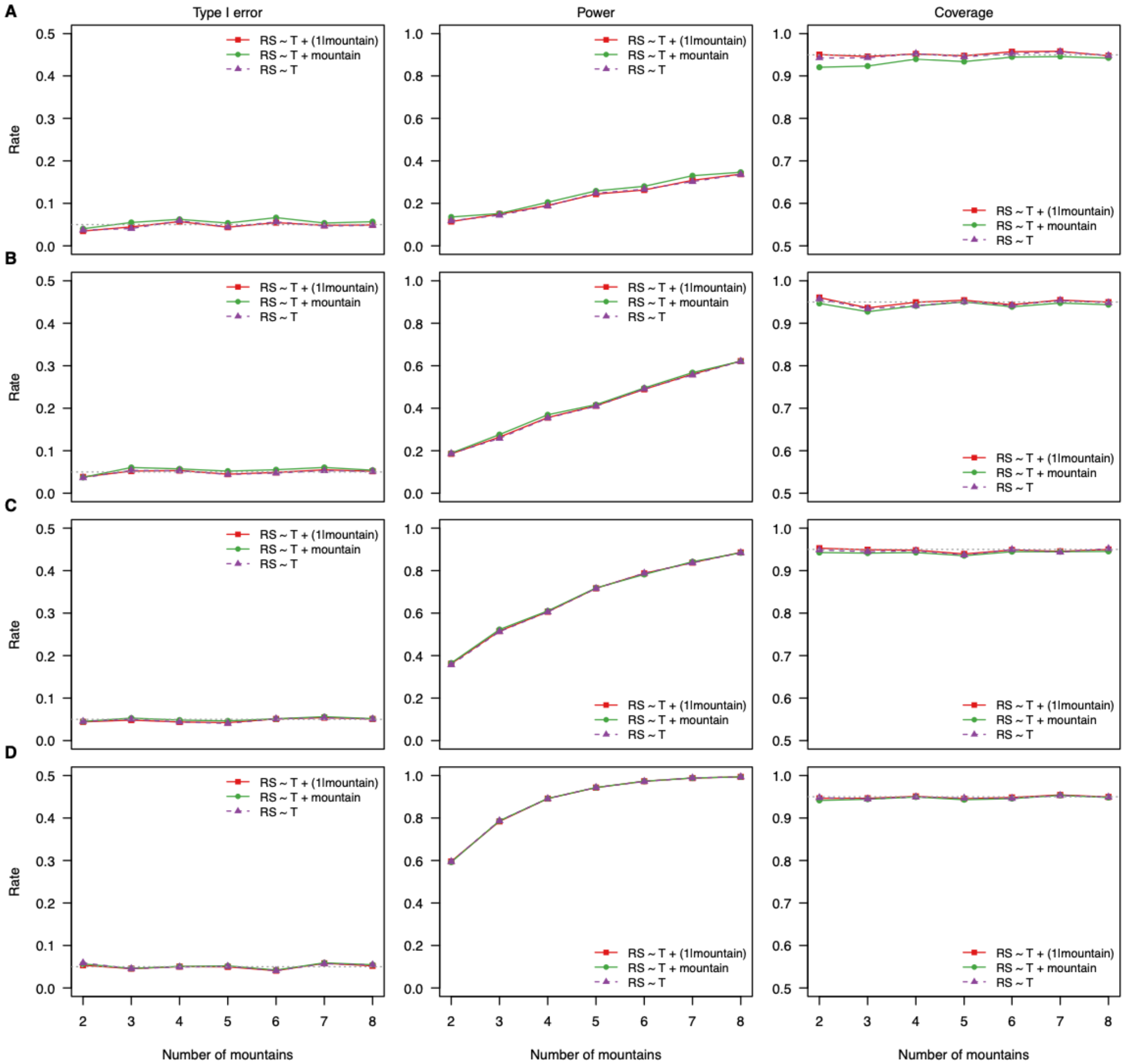

Figure S8: Type I error rates, power, and coverage for generalized linear (mixed effect) models fitted with lme4 to simulated data with 2-8 mountains for scenario A(random intercept for each mountain). A-D show the results for different numbers of observations for each mountain (25, 50, 100, and 200). For each scenario, 5,000 simulations and models were tested.

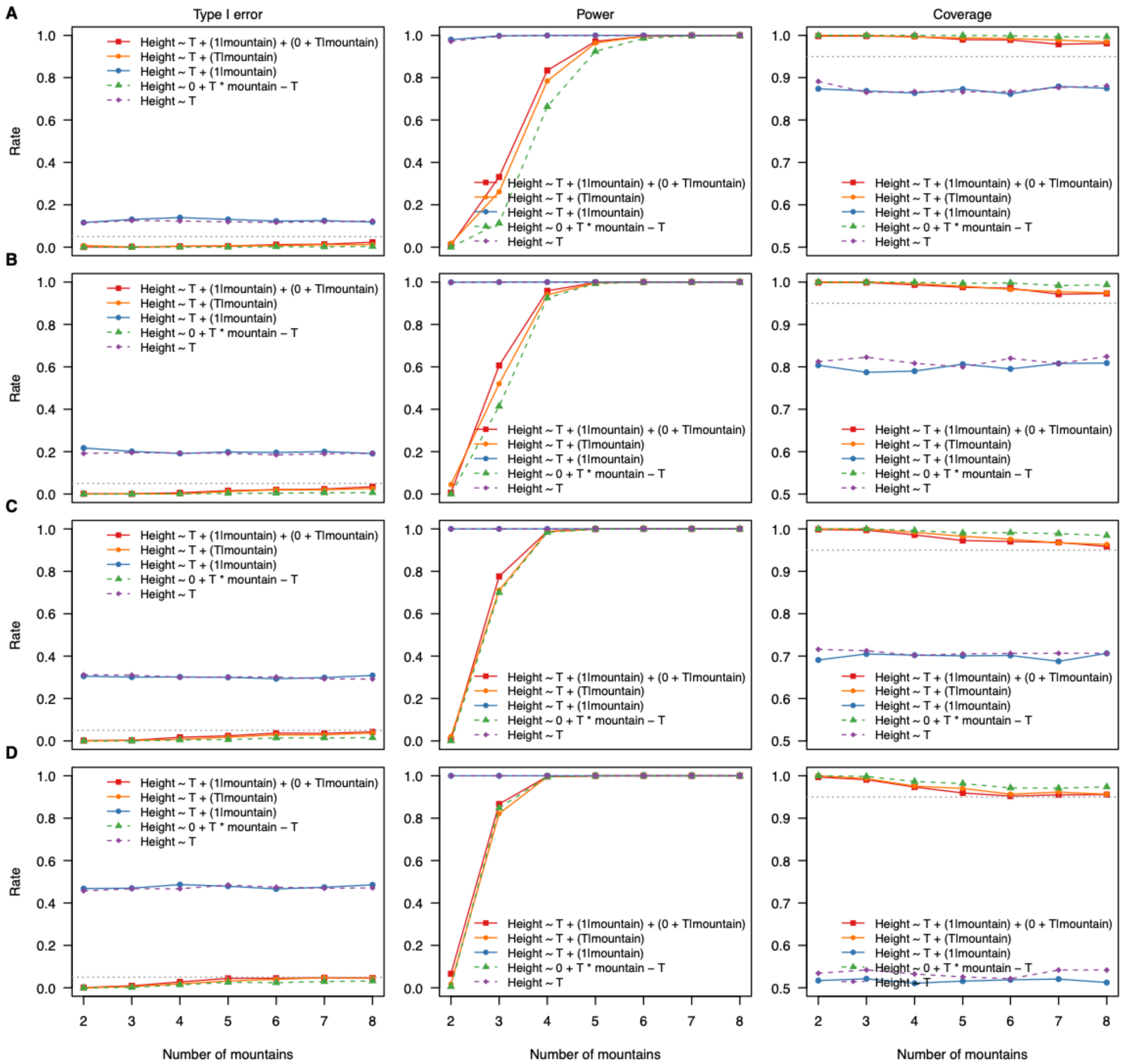

Figure S9: Type I error rates, power, and coverage for linear (mixed effect) models fitted with lme4 to simulated data with 2-8 mountains for scenario B (random intercept and slope for each mountain). A-D show the results for different numbers of observations for each mountain (50, 100, 200, and 500). For each scenario, 5,000 simulations and models were tested.

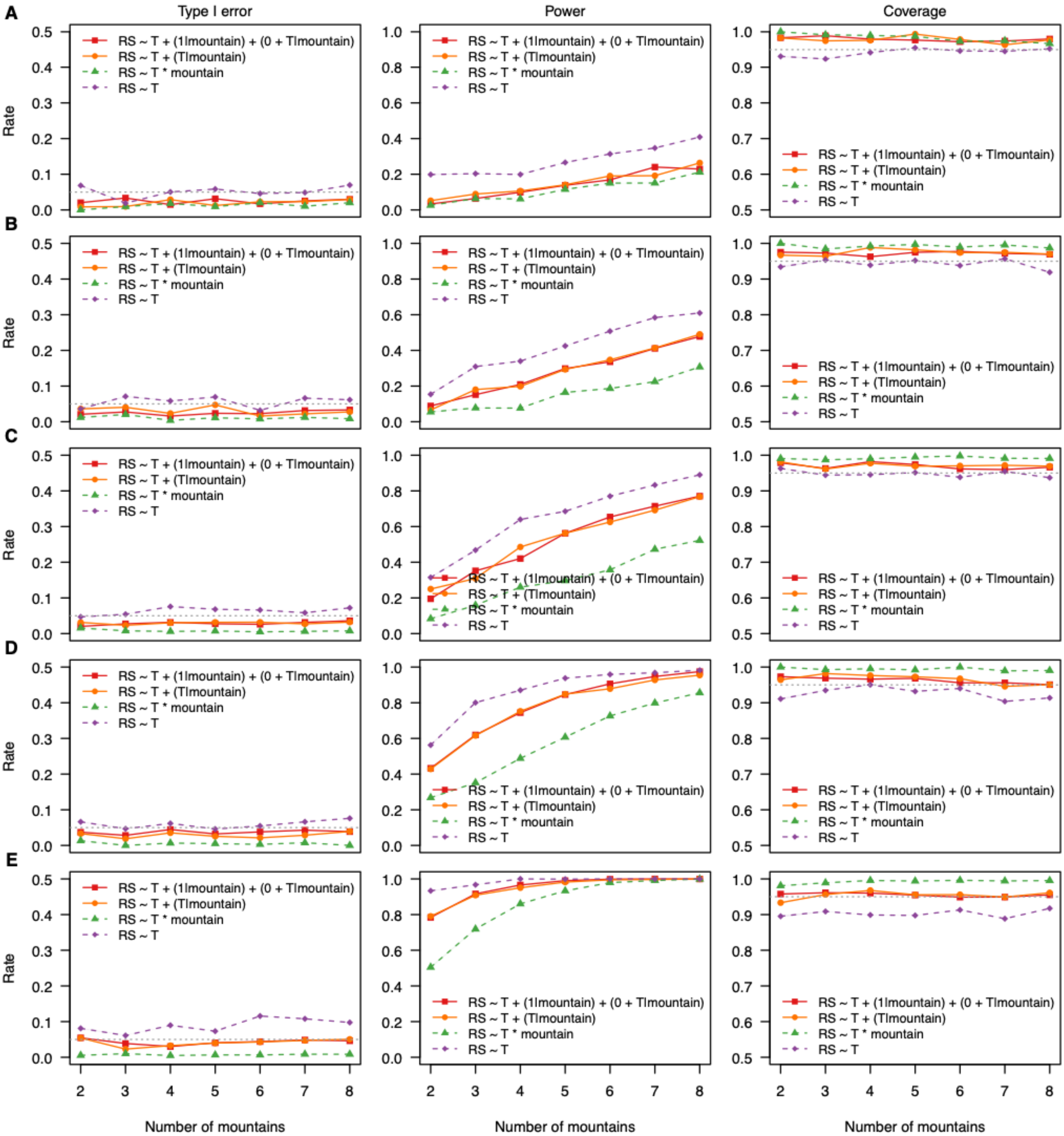

Figure S10: Type I error rates, power, and coverage for generalized linear (mixed effect) models fitted with lme4 to simulated data with 2-8 mountains for scenario B(random intercept and slope for each). A-E show the results for different numbers of observations for each mountain (25, 50, 100, 200, and 500). For each scenario, 5,000 simulations and models were tested.