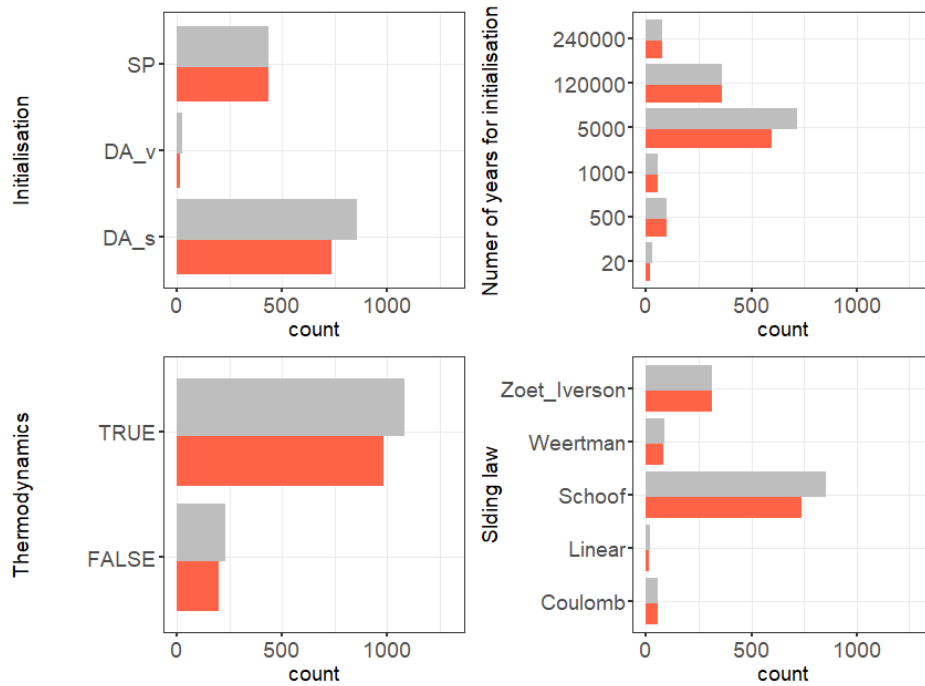
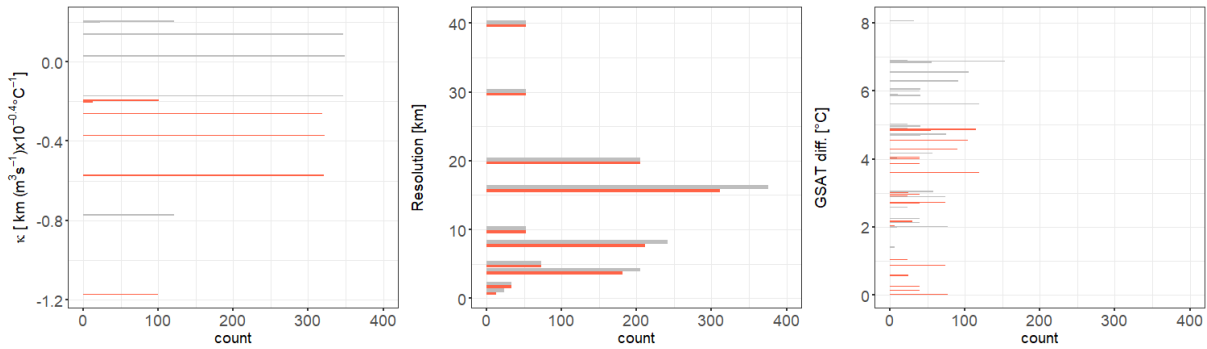


1 **Supplementary Material S1 Complementary analysis for Sect. 2.4**

2 To illustrate the procedure described in Sect. 2.4, Figs. S1 and S2 present the histograms before
 3 and after applying the experiment ‘MAR’.



4
 5 **Figure S1: Count number of the MME members with respect to the different modelling assumptions classified as**
 6 **“categorical” in Table 1 before (grey) and after (red) applying the experiment ‘MAR’.**



7
 8 **Figure S2: Count number of the MME members with respect to the different modelling assumptions classified as**
 9 **“continuous” in Table 1 and after (red) applying the experiment ‘MAR’.**

10 We measure the extent to which the new subset of the MME differs from the original MME, by
 11 assessing, for each modelling assumption, the mean difference in the count numbers between
 12 the perturbed MME and the original one. The overall deviation, denoted D_h , is defined as the
 13 average value (normalised by the total number of experiments n) of the changes assessed across
 14 all considered modelling assumptions.

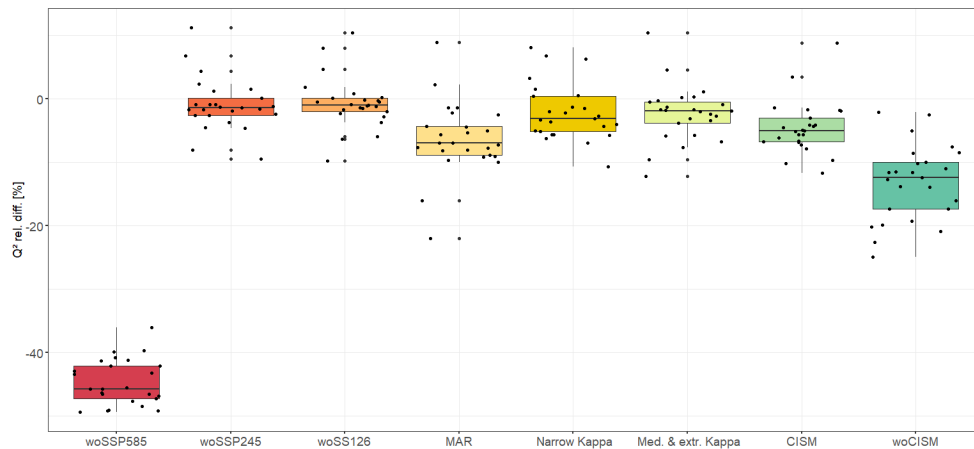
15

1 Supplementary Material S2 use of an alternative performance indicator

2 Figure S3 presents the same type of analysis than Figure 7 but with an alternative indicator of
3 emulator's predictive capability, here chosen as the coefficient of determination Q^2 defined as
4 follows (using notations of Sect. 2.3):

$$5 \quad Q^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (e^{(i)})^2}{\sum_{i=1}^{n_{\text{test}}} (slc^{(i)} - \overline{slc})^2}$$

6 with \overline{slc} the mean value of slc calculated over the test set. The closer Q^2 to one, the higher the
7 emulator's predictive capability.

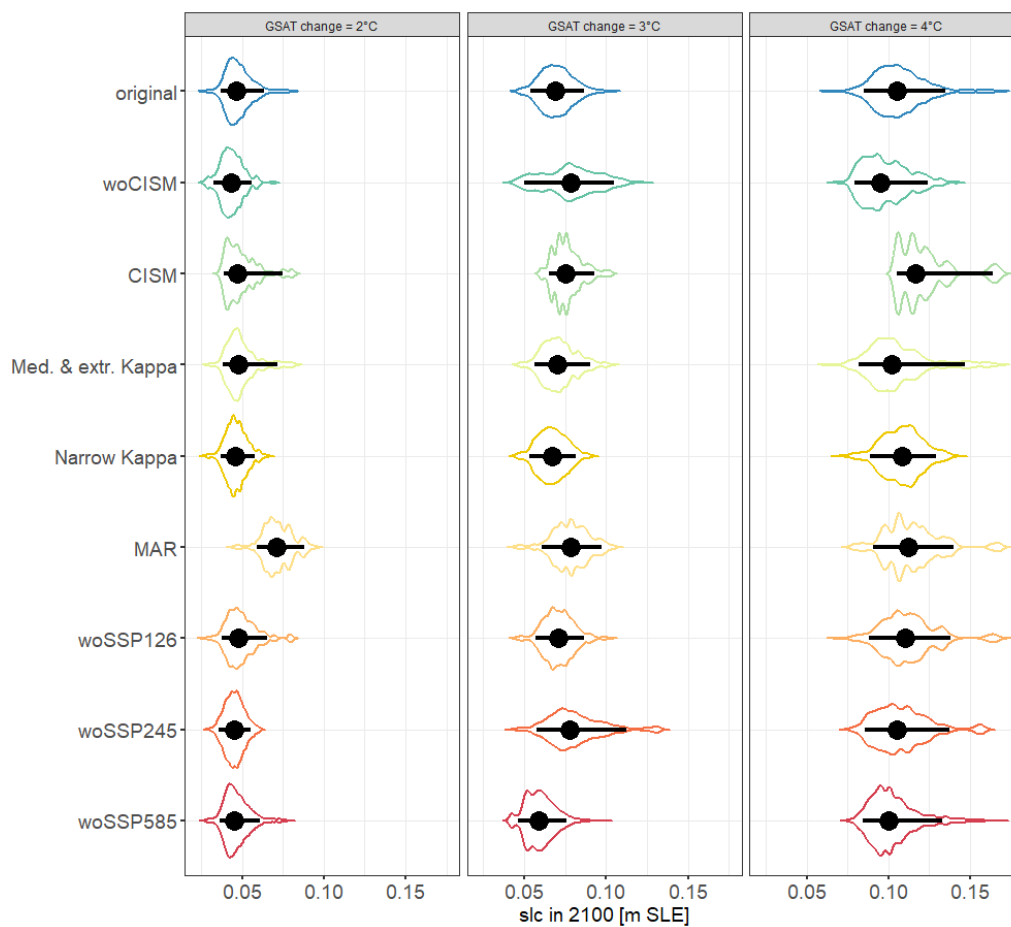


8
9 **Figure S3: Relative difference (in %) for the estimates of RF predictive capability measured by coefficient of**
10 **determination Q^2 , between the RF reference solution and the RF model trained when considering the experiments**
11 **indicated in the x-axis (see Table 2 for full details).**

12

1 Supplementary Material S3 Analysis of the probability distributions

2 We complement the analysis of Fig. 8 with the qualitative inspection of the violin plots (Hintze
3 & Nelson, 1998) in Figure S4. This shows that experiment ‘CISM’ has the largest impact for
4 the highest GSAT change scenario, and ‘MAR’ for the lowest GSAT change scenario (as
5 highlighted by the rightwards shift of the distribution compared to the reference solution
6 outlined in blue). This visual inspection also informs on the influence on the shape and tails of
7 the resulting distributions, by showing that ‘woCISM’ under 3°C GSAT results in a much
8 broader distribution compared to the reference, and CISM and MAR for high GSAT result in a
9 multimodal distribution.



10

11 **Figure S4: Violin plots representing the probabilistic *slc* distributions considering three GSAT changes, 2°C, 3°C and**
12 **4°C using the unperturbed RF emulator (case named ‘original’) and using RF emulators trained by applying the**
13 **experiment described in Table 2. The error bar’s endpoints are defined by the quantile at 5% and at 95% (denoted**
14 **respectively Q5% and Q95% in the manuscript), and the central dot corresponds to the median value.**

15 References

16 Hintze, J.L. and Nelson, R.D.: Violin plots: a box plot-density trace synergism, *Am. Stat.*, 52(2),
17 181–184, 1998.