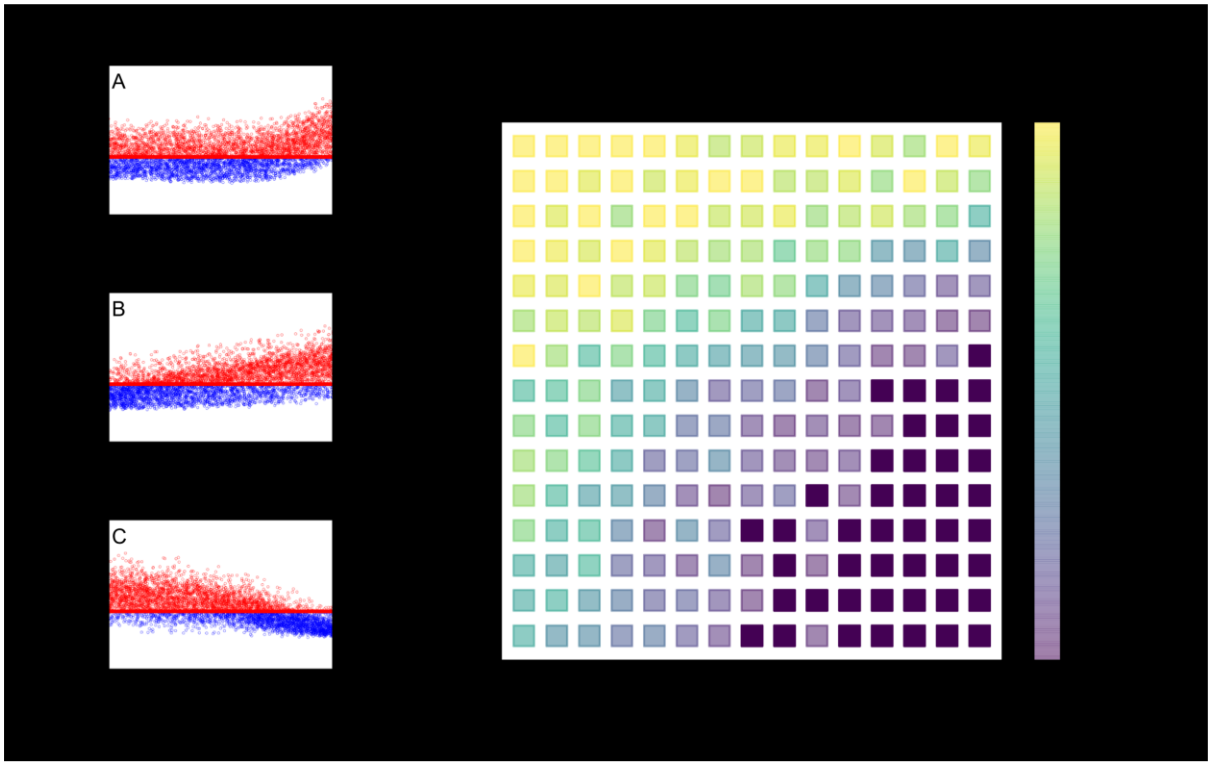
Appendix: Comparison with the method of history matching. It is possible to provide frequentist confidence sets on parameters using a method inspired by history matching. Below, we describe an adaptation of the version of the method employed in Johnson et al (2020). Principally, the typical use of history matching in that work and others is to exclude implausible parameter values in a systematic way, not to produce a plausible constraint set with fixed probabilistic properties.

One difference between the method of Johnson et al (2020) and our PLAUSIBILITYTEST pipe of Section 2.4 is the choice of implausibility statistic. Johnson et al (2020) use the statistic

$$I_{\mathrm{HM},N}(u) = \left\{ \frac{|\,\mathbb{E}[\eta_x(u)|D_{\mathrm{train}}] - z(x)\,|}{\sqrt{\mathrm{Var}[\eta_x(u)|D_{\mathrm{train}}] + \mathrm{Var}[\varepsilon_{\mathrm{meas},x}] + \mathrm{Var}[\varepsilon_{\mathrm{other},x}]}} : x \in \mathcal{M}^* \right\}_{(N)},$$

Figure 4. *Constraints resulting from an adaptation of the history matching method at* 95% *confidence level. Referring to the implausibility statistic given by Eq.* (3)*, we use* $q = 0.25$.



where $S_{(N)}$ denotes the N th largest element of a set S. For instance,

$$I_{\text{HM},1}(u) = \max\left\{ \frac{|\mathbb{E}[\eta_x(u)|D_{\text{train}}] - z(x)|}{\sqrt{\text{Var}[\eta_x(u)|D_{\text{train}}] + \text{Var}[\varepsilon_{\text{meas},x}] + \text{Var}[\varepsilon_{\text{other},x}]}} : x \in \mathcal{M}^* \right\}.$$

The authors then tune the "tolerance level" (which corresponds to the number $(|\mathcal{M}^*| - N)$ in our expression for the statistic) and the "exceedence threshold" (which, by analogy to our test, corresponds to an implausibility cutoff or critical value; in most cases set to 3.5) until a fixed proportion (say, 40%) of test parameters u are retained as plausible under the statistic. By this account, a constraint on the parameters is necessarily yielded, but the confidence level at which this constraint holds is undetermined.

Alternatively, we can set a confidence level first and compare the resulting constraints obtained by the strict bounds approach or this history matching-inspired approach, where the former approach yields the constraints shown in Figure 3. In particular, consider the implausibility statistic

$$I_{1-q}(u) = \text{quantile}_{1-q}\left\{ \frac{|\mathbb{E}[\eta_x(u)|D_{\text{train}}] - z(x)|}{\sqrt{\text{Var}[\eta_x(u)|D_{\text{train}}] + \text{Var}[\varepsilon_{\text{meas},x}] + \hat{\delta}^2_{\text{MLE}}}} : x \in \mathcal{M}^* \right\}, \tag{3}$$

where quantile_{1-q} returns the $(1-q)$ th quantile of the given set. For example, $I_1(u)$ returns the maximum absolute normalized discrepancy for parameter u and $I_{0.5}(u)$ is the median. This is a close analog to the statistic seen in Johnson et al (2020) when $q \approx N/|\mathcal{M}^*|$. Naturally the approximate null distribution for this new statistic is that of the $(1-q)$ th percentile of a sample of $|\mathcal{M}^*|$ half-normal random variables. A critical value for the plausibility test at the 5% significance level can therefore be estimated quickly by simulating a collection of samples of $|\mathcal{M}^*|$ half-normal random variables, drawing the $(1-q)$ th percentile from each sample, and then selecting the 95th percentile from that collection. In Figure 4, we show the 95% confidence level constraints on the aerosol parameters that are obtained from the above-described history matching method using parameter $q = 0.25$.

As this example illustrates, similar non-trivial and principled constraints on aerosol parameters are possible. However, the history matching-inspired approach requires choosing the tuning parameter q which substantially affects the final constraints. In Figure 4, we chose q to obtain constraints similar to those given by our method in Figure 3. If we choose $q = 0.5$ (i.e., use the median), we find that the constraints (not shown) become looser than those provided by our method. On the other hand, if we choose q close to zero, we find that this history matching approach becomes sensitive to non-Gaussianities in the tails of the error distributions, leading to overly discriminating plausiblity test results. If history matching was calibrated like this to obtain confidence sets at a prescribed confidence level (which, we emphasize, is not currently done), it seems difficult to choose q optimally to balance the power of the tests with robustness to mismodeling of the error distribution tails.