S1 Metropolis-Hastings Supplement

The Adaptive Metropolis algorithm used here contains three key concepts explained in the following sections. The adaptive random walk, allowing the algorithm to learn features of the target distribution; transformed proposals, providing a natural way of including limits on the parameters; and tempering of the target distribution, to reduce the effects of local maxima and allowing better exploration of the target.

S1.1 Adaptive random walk

A standard Metropolis-Hastings (MH) Gaussian random walk proposal can be written as

$$\widehat{x} = x_t + \epsilon, \qquad \epsilon \sim N(0, \lambda\Sigma), \tag{1}$$

where a mean-zero normal random variable is added to the current value, x_t, to generate a proposal, $\widehat{x}$. The new value is accepted with probability that compares the likelihood of the new and old sample

$$\alpha = \min\left(1, \frac{p(\widehat{x}|Y)}{p(x_t|Y)}\right) = \min\left(1, \exp(-J(\widehat{x}) + J(x_t))\right). \tag{2}$$

If the value is accepted, set $x_{t+1} = \widehat{x}$ otherwise keep the previous value, $x_{t+1} = x_t$.

It has been shown (Gelman, Roberts, & Gilks, 1996 ; Roberts, Gelman, & Gilks, 1997) that optimal behaviour of the MH-algorithm is obtained when about 20% to 30% of samples are accepted. This is achieved when Σ corresponds to the covariance matrix of target distribution, $p(x|Y)$, and the scaling is $\lambda = 2.38^2/d$, where d is the number of parameters in x. The key idea of the adaptive algorithms suggested in Andrieu & Thoms, 2008 is to recursively estimate both Σ and λ from previous samples.

To limit the effect of initial values, we first ran 5 000 steps of the chain without adaptation and taking Σ as 10^{-3} times an identity matrix (i.e. very small initial steps). Thereafter a covariance matrix Σ was estimated based on the initial samples and the chain run for a further 20 000 steps adapting only λ and not Σ. Finally, for the last 100 000 steps both λ and Σ were updated using the Rao-Black-wellised Adaptive Metropolis and Global Adaptive Scaling Metropolis algorithms from (Andrieu & Thoms, 2008).

S1.2 Transformed proposals

The standard proposals in Equation 1 does not include limits on the parameters, e.g. parameters that have to be positive or are constrained to a known intervall. To handle parameter limits we transformed the parameters, resulting in an adjusted random walk proposal

$$\widehat{x} = g\left(g^{-1}(x_t) + \epsilon\right), \qquad \epsilon \sim N(0, \lambda\Sigma), \tag{3}$$

where a list of possible limits and corresponding functions are given in Tabel 1. Note that different functions can be applied to each parameter in x.

Having transformed proposal requires an adjustment of the acceptance probability (Hastings, 1970) in Equation 2 to

$$\alpha = \min\left(1, \frac{p(\widehat{x}|Y)}{p(x_t|Y)} \prod_i \frac{q\left(x_t^{(i)} \mid \widehat{x}^{(i)}\right)}{q\left(\widehat{x}^{(i)} \mid x_t^{(i)}\right)}\right) = \tag{4}$$

$$= \min\left(1, \exp(-J(\widehat{x}) + J(x_t) + \sum_i \log\left(\frac{q\left(x_t^{(i)} \mid \widehat{x}^{(i)}\right)}{q\left(\widehat{x}^{(i)} \mid x_t^{(i)}\right)}\right)\right). \tag{5}$$

Here $x_t^{(i)}$ denotes the i^{th} parameter in the x_t-vector and the q-terms are given in Tabel 1. Note that each transformation results in one adjustment for that parameter and that all adjustment have to be multiplied together.

S1.3 Tempering the target distribution

Table 1: Summary of transformations and corresponding adjustments to the acceptance probability for the three cases of variables with low-limit, upper-limit, and variables constrained to an intervall.

Constrain	Functions		Acceptance
	$g(x)$	$g^{-1}(x)$	$q\left(x_t \mid \widehat{x}\right) / q\left(\widehat{x} \mid x_t\right)$
$x > a$	$\exp(x) + a$	$\log(x - a)$	$(\widehat{x} - a)/(x_t - a)$
$x < a$	$a - \exp(x)$	$\log(a - x)$	$(a - \widehat{x})/(a - x_t)$
$x \in [a, b]$	$\frac{b \exp(x) + a}{\exp(x) + 1}$	$\log(x - a) - \log(b - x)$	$\frac{(\widehat{x} - a)(b - \widehat{x})}{(x_t - a)(b - x_t)}$

For large amounts of data the loss function $J(x)$ can have very deep local minima causing the MH-algorithm to get stuck, even with the two already outlined adjustments. To reduce the scale of the loss function we temper the target distribution (Jennison, 1993)

$$\widetilde{p}(x|Y) = p(x|Y)^{1/T} = \exp(-J(x)/T), \tag{6}$$

where T is a suitably large value.

Having run a MH-algorithm for N samples, the first N_b samples are discarded as burn-in and expectations or variances can be computed as averages of the remaining samples:

$$E(x|Y) \approx \frac{1}{N - N_b} \sum_{i=N_b}^{N} x_i,$$
$$V(x|Y) \approx \frac{1}{N - N_b} \sum_{i=N_b}^{N} (x_i - E(x|Y))^2.$$

However, with a tempered target distribution we have samples from $\widetilde{p}(x|Y)$ and need to use importance sampling to adjust for the difference in distributions (Jennison, 1993).

$$E(x|Y) \approx \frac{1}{N - N_b} \sum_{i=N_b}^{N} w_i x_i,$$
$$V(x|Y) \approx \frac{1}{N - N_b} \sum_{i=N_b}^{N} w_i (x_i - E(x|Y))^2,$$

where the weights are given by

$$w_i = \frac{\exp\left(-\frac{T-1}{T}(J(x_i) - J_{\min})\right)}{\sum_{i=N_b}^{N} \exp\left(-\frac{T-1}{T}(J(x_i) - J_{\min})\right)}.$$

and $J_{\min} = \min_i J(x_i)$. The subtraction by $J_{\min}$ is for numerical stability to avoid cases of $0/0$ when all $J(x_i)$ are very large.

S2 Twin experiments: Parameter convergence

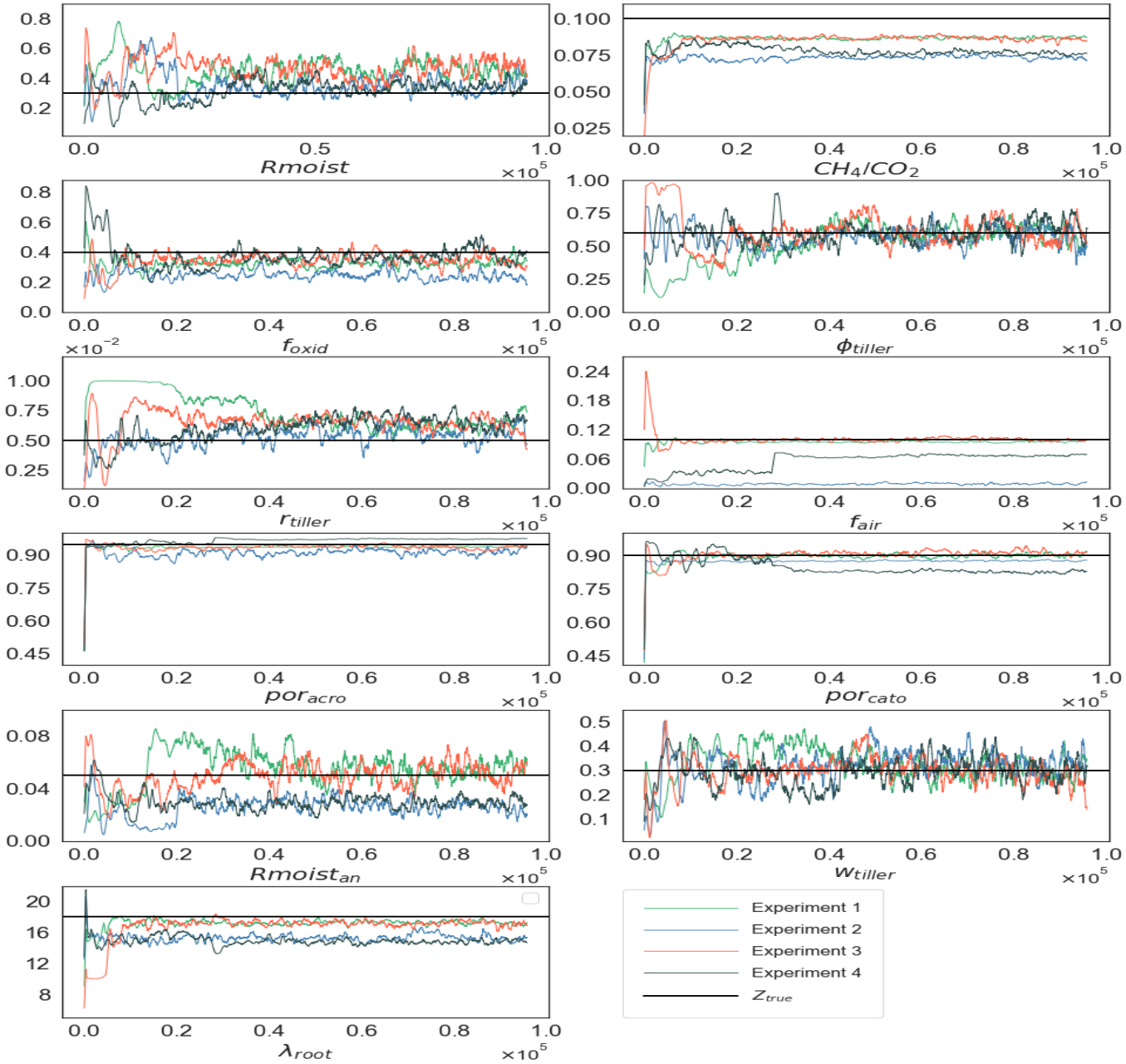

Figure 1: Thinned trace plots resulted from the scenario 1. The black vertical lines shows the z_{true} values used to generate the twin observation and the trace plots in each color represents the different experiments.

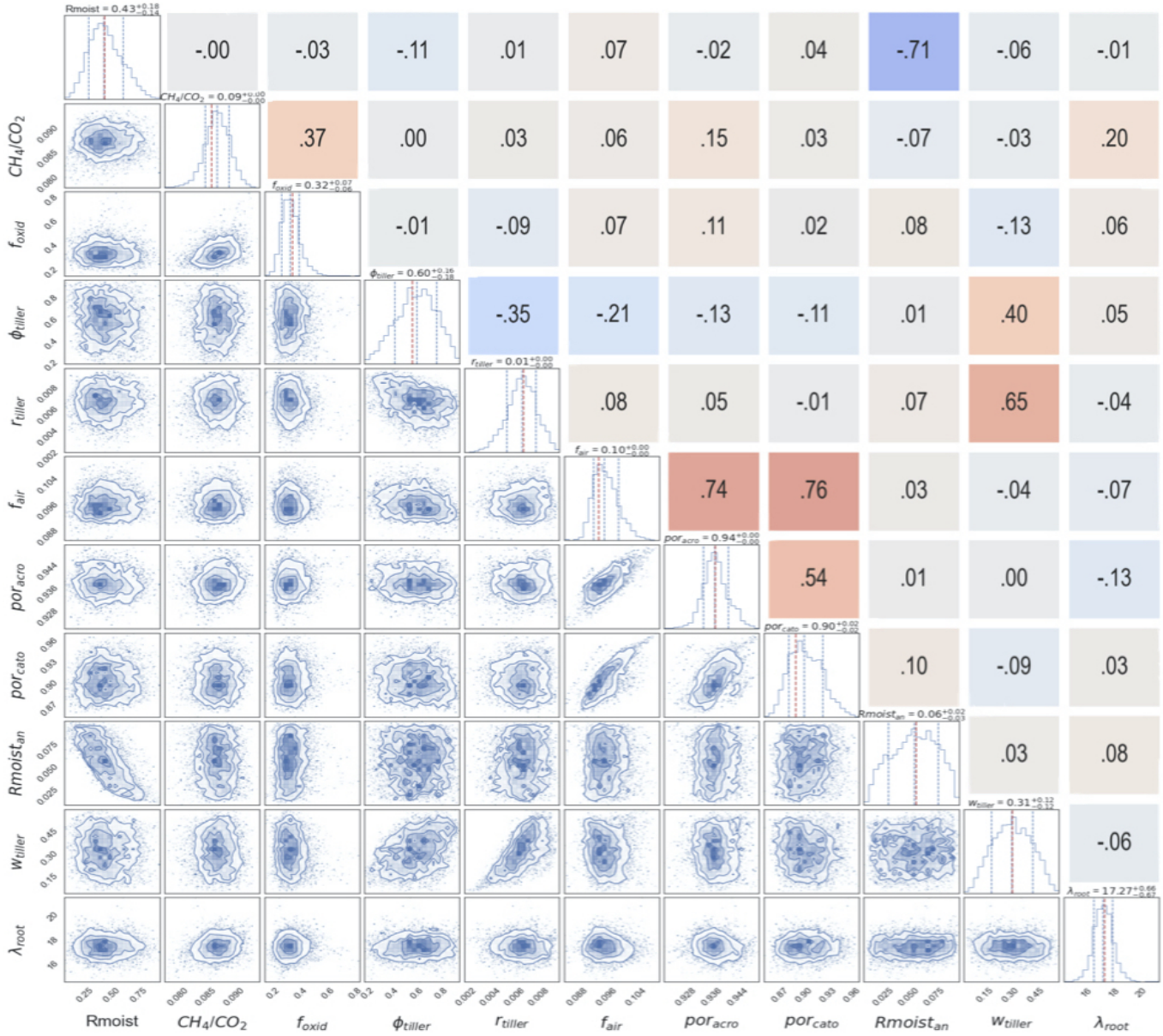

Figure 2: A posteriori correlations between the parameters from the G-RB AM twin optimisation. Blue and red color in the upper triangle represents the strong negative and positive correlations respectively. The numerical labels on the the upper triangle are the values of Pearson's correlation coefficient. The panels on the diagonal show the 1-D histogram for each model parameters with a dashed red vertical line to indicate the best-fit value. The vertical blue lines are the 0.16, 0.5 and 0.84 quantiles of the distributions respectively. On top of each 1D histogram the mode of the distribution and the interval of the 0.16 and 0.84 quantiles are indicated. The lower triangle represents the two-dimensional marginal distributions of each parameter with contours to indicate 1σ, 2σ and and 3σ confidence levels and the points in the plots are the values of G-RB AM chain after the 'burn-in' (blue dots). Ranges of the distributions are labeled on the left and bottom of the figure.

References

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and computing*, *18*(4), 343–373.

Gelman, A., Roberts, G., & Gilks, W. (1996). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 5* (pp. 599–607). Oxford University Press.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Jennison, C. (1993). Discussion on the meeting on the gibbs sampler and other markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B*, *55*, 54–56.

Roberts, G., Gelman, A., & Gilks, W. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Probability*, *7*, 110–120.